MASARYK UNIVERSITY
Faculty of Science

Habilitation Thesis

Analysis of biomacromolecular structural fragments

Brno 2016 Radka Svobodová Vařeková

To my grandmother, my parents and my daughter Radunka.

Acknowledgements

I would like to thank my past supervisors for their valuable advice, my colleagues and co-authors for our fruitful work together and also my family for their love and support. Last but not least, I thank God for his mercy and patience with my faults.

Contents

Preface

This habilitation thesis is a compilation of scientific publications to which I contributed as the first author, the corresponding author, or as a co-author. The articles were published between 2006 and 2015, and a list of them is given on page 66. A common topic of all the publications is the analysis of biomacromolecular structural fragments such as ligand and metal binding sites, channels and pores, and supersecondary structure elements. Specifically, the articles focus on the development of approaches for the validation, detection, extraction, comparison and characterization of fragments, and also on the application of these approaches for solving important biological questions. The accompanying text highlights the author's contribution to the field of biomacromolecular structural fragment analyses, and it also contains a brief introduction to the topic. Detailed information on the developed approaches and their applications can be found in the enclosed original publications.

1

Introduction

In recent years, the life science research community has strongly benefited from the fact that a vast amount of data is available about various types of molecules. For example, we can obtain the complete human genome of a selected individual in less than 14 days, nearly 90 million various small molecules are described in freely accessible databases (e.g., Pubchem [1], ChEMBL [2], ZINC [3], DrugBank [4]), more than 110 thousand biomacromolecular structures have been determined and published (Protein Data Bank [5]). This richness of data has catalyzed and strengthened the formation of so-called modern life sciences – life science specializations focused on research utilizing some part of this data. The best known modern life sciences are genomics, proteomics, metabolomics, systems biology, bioinformatics, structural bioinformatics, and chemoinformatics. Although these modern life sciences are relatively young, they have provided many key results in basic and applied research – from understanding biomacromolecular functions and mechanisms of their action to the classification of types of disease or the rational development of novel drugs (e.g. [6–12]).

An important and very interesting development in modern life sciences is the interconnection of structural bioinformatics and chemoinformatics. Structural bioinformatics focuses on researching three-dimensional (3D) biomacromolecular structures – e.g., their prediction, comparison and characterization. Chemoinformatics focuses on small organic molecules (primarily drugs or ligands) and studies mainly their 3D structure generation, similarity searches, and predictions of their properties or activities. Chemoinformatics and structural bioinformatics are highly compatible, since we need to discover both biomacromolecules and their targets – ligands. A powerful combination of structural bioinformatics and chemoinformatics enables us to focus on key regions of a biomacromolecule – biomacromolecular structural fragments such as ligand or metal binding sites, channels, pores, or supersecondary structure motifs. Analyses of these fragments can produce very beneficial outputs such as discovering drug design patterns, information for the classification of biomacromolecules, understanding the relationship between structure and function, and predicting their putative functions (e.g. [11–20]). The interconnection of structural bioinformatics and chemoinformatics can therefore bring useful insights into biology, biochemistry and biomolecular chemistry.

Nowadays, when we would like to analyze the biomacromolecular structural fragments, we are in both a great and a challenging situation, caused by the real richness of the available data. I would like to demonstrate these two sides of the coin with a small example. In 1990, there was just one 3D structure of a cytochrome P450cam in Protein Data Bank (found using its UniProt [21] molecular name "p450-cam"). In 2000, 22 structures of these cytochromes were available (with various substrates, including a few mutants). Nowadays, we have 120 such cytochrome structures in Protein Data Bank. Therefore, twenty years ago, analyses focused on some properties of cytochrome fragments (e.g., how cytochrome active sites or the channels leading to them look in general) were purely science fiction. Ten years ago, some modest analyses could be performed, but their statistical significance or the proper coverage of substrates were questionable. Today, we have a supercritical volume of data and we can do really meaningful, useful and reliable analyses. In parallel, the richness of the data introduces a significant problem. In 1990, the analysis of one structure could be done easily intuitively and manually. About ten years ago, the manual analyses were theoretically still applicable (if a researcher had enough students) or some basic methodologies implemented in in-house scripts could be used. Currently, manual processing is clearly nonsense and employing some basic methodologies meets with problems, because they are not robust and general enough to handle a large and heterogeneous set of samples. Therefore, before we can begin the highly interesting analyses and before we can obtain some key results, we must first focus on the methodologies and software tools for performing analyses. Unfortunately, we cannot just find and obtain ready-to-use approaches for individual analyses of biomacromolecular structural fragments. The reason for this is that these methodologies are still under intensive development.

Our deep interest in biomacromolecular structural fragments and in parallel, the lack of approaches for their realization motivated us to concentrate our work on two interconnected goals. The first was the development of methodologies for the analysis of biomacromolecular structural fragments. The second was their application in resolving important biological and chemical questions. Therefore my habilitation thesis is also focused on these two topics.

Important steps within the analysis of protein structural fragments are their validation, detection, extraction, comparison and characterization. The necessity of structure validation became evident when some published structures were found to contain serious errors [22–25]. Reliable and well-established approaches for the validation of standard biomacromolecular building blocks (amino acids and nucleotides) are now available [26–29]. The validation of ligands, which are frequent components of biomacromolecular fragments and the main sources of their errors [30], is a markedly more complex problem and its methodology is still under development. We introduced an extended methodology for the validation of ligand annotation, applicable for any ligands and non-standard residues [MV, VDB].

An essential step in biomacromolecular structural fragment analysis is the collection of all fragment samples. Therefore, the fragment should be first described via a proper molecular language, then detected within the structures and then extracted from them. Several molecular languages for describing various molecular structures were introduced [31–33] and also a few approaches were developed for

the extraction of specialized compounds [11, 12, 16, 17, 20, 34]. Based on them, we developed the molecular language PatternQuery [PTQ], which enables any structural fragment to be determined, and a methodology for the rapid extraction of fragments described this way from the Protein Data Bank. An important class of biomacromolecular fragments are their channels and pores, since they provide a substrate with access to an active site. Few methodologies for their extraction have been published – Caver [35–37], MolAxis [38, 39], MOLEonline [MO] and MOLE 2.0 [MO2]. The last two were developed by ourselves.

A substantial task within structural fragment analysis is also fragment comparison. In general, when we compare similar 3D structures, we need to first pair corresponding atoms of the compared structures and then the structures are superimposed based on this pairing. Many state-of-the-art approaches for comparing organic molecules were developed, and they include implicit [40–44] or sequence alignment-based pairing [45–47]. We created an approach employing so-called combinatorial and subgraph matching pairing [SB], which is appropriate for fragments.

The final step in the process of structural fragment analysis is its characterization, i.e. determining the fragment's characteristic properties. Partial atomic charges are one such property, providing information about the electron distribution within a molecule. The most appropriate approach for their calculation is quantum mechanics (QM), which is unfortunately very demanding in terms of time and computational resources. A faster alternative to QM are empirical methods, of which the Electronegativity Equalization Method (EEM) [48] is the most popular and applicable. EEM has been significantly improved in recent years [49–53], and we also intensively participated in its development. Specifically, we published EEM parameterization for organic molecules [EO], for biomacromolecules [EB] and for ligands [EL]. In parallel, we developed a methodology for calculating EEM charges in large biomacromolecules (a parallel approach [EPM] and an approximative method [EAM]) and in ligands and drug-like molecules [EL]. Other key characteristics of biomacromolecular structural fragments are channel properties. While the channel radius and length were calculated by all current channel detection tools, our approaches MOLEonline [MO] and MOLE 2.0 [MO2] provide a rich set of chemical, geometrical and also physico-chemical channel properties.

The developed methodologies for the analysis of biomolecular structural fragments enabled us to perform several interesting studies and in this way resolve a few important biological and chemical questions. Specifically, we evaluated the quality of structures for different molecular classes [VDB], we used charges for pK_a prediction [PQ,PE,PS] and for researching apoptotic protein Bax activation [BAX], and we discovered the anatomy of enzyme channels [AN].

A quality evaluation of individual ligands or biomacromolecules can be straightforward to obtain [MV]. But we lack the bigger picture -– information about the quality of various molecular classes (e.g., drug molecules, organometals, carbohydrates). Therefore, we performed such an analysis [VDB]. The best quality proved to be drug molecules, probably because markedly more effort is expended on determining their structure in biomacromolecular complexes. Carbohydrates and polycyclic ligands exhibited problems in the chirality of their carbon atoms, as expected. The most problematic ligands are organometals, exhibiting clear validation problems in most validation criteria.

Partial atomic charges proved to correlate with the acid dissociation constant pK_a [54–57]. This constant is an important molecular property and its values are of interest in chemical, biological, environmental and pharmaceutical research [58–60]. At the same time, its measurement is highly demanding and its prediction is still a challenge. Therefore, we focused our research on charge-based approaches to pK_a prediction. We demonstrated that QM charges are highly successful descriptors for the prediction of pK_a via Quantitative Structure-Property Relationship (QSPR) models [61], but a proper charge calculation approach must be used [PQ]. We later demonstrated that empirical charges calculated via EEM are also applicable for pK_a prediction [PE]. We then found that the pK_a predicting QSPR models are sensitive to the 3D structure generation methodology employed for preparation of input molecular 3D structures. Therefore, we evaluated which 3D structure sources are applicable [PS]. We discovered that a workflow for the fast and accurate prediction of pK_a can be as follows: The preparation of 3D structures via the data- and knowledge-based approach (used in software tools CORINA [62] and Omega [63]) with no further optimization, calculation of EEM charges for these structures and then the EEM QSPR prediction of pK_a. Such a workflow can be directly used within the process of *in silico* drug design or incorporated into other chemoinformatics applications.

Charges can also help us to understand the mechanisms of important biological processes -– e.g., the activation of the apoptotic protein Bax [BAX]. Apoptosis is a programmed cell death, and its proper regulation is essential for multi-cellular organisms. Apoptosis includes Bax activation [64, 65], Bax oligomerization and forming pores in the mitochondrial membrane. The Bax activation mechanism is still unclear and it motivated us to investigate changes in the Bax charge profile upon activation. We found [BAX] that charge reorganizations after activator binding mediate the exposure of the functional sites of Bax (i.e., the C-domain and BH3 domain) and as a consequence activate Bax. The affinity of the Bax C-domain for its binding groove is decreased due to the Arg94-mediated abrogation of the Ser184-Asp98 interaction. We further identified a network for charge transfer, which brings the activation information from the activation site, through the hydrophobic core of Bax, to the distant functional sites of Bax. The network was mediated by a hub of three residues on helix 5 of the hydrophobic core of Bax.

Channels play a key role in enzymes – proteins which catalyze reactions changing substrates into products [66–77]. The reason for this is that the enzymatic active site is often buried deep within their structure and connected to the outside by a channel. Therefore, these channels influence the substrate preference of the active site. An important biological question is what properties the enzyme channels have. We performed analyses of the channels in more than 4,000 enzyme structures [AN]. We identified that at least 64% of these enzymes contain on average two channels longer than 15 Å leading to the catalytic site. The longest and the most hydrophobic channels were found in oxidoreductases and the shortest and the most hydrophilic channels in ligases. The composition of channel lining residues differed from the average composition of enzyme structures as well as from the composition of the protein surface. Specifically, aromatic, charged and polar amino acids occur more frequently in channel walls. All of these findings indicate that the active site access channels have a significant biological function, as they are involved in co-determining the enzyme's substrate preferences.

2

Analysis of biomacromolecular structural fragments

2.1 Validation MV, VDB

2.1.1 State of the Art

The validation of biomacromolecular structures obtained by NMR and X-ray crystallography has become a very important topic, because some published structures have been found to contain serious errors [22–25]. The first step in the validation of biomacromolecules and their complexes is checking the standard building blocks (residues), namely standard amino acids and nucleotides. The usual methodology is to evaluate the specific properties of all of these standard residues (e.g., electron density, atom clashes, bond lengths, bond angles, torsion angles). These approaches are embedded into common validation software tools such as OOPS [29], WHAT_CHECK [26], PROCHECK [27], PROCHECK-NMR [78], AQUA [78] and MolProbity [28].

The next key step is the validation of ligands and non-standard residues. This step can be performed in a similar manner as for the standard residues (with focus on electron density and atom clashes). Such a methodology is implemented in several validation software packages: ValLigURL [79], Mogul [80], Coot [81] and PHENIX [82].

A different methodology for validating ligands was developed later – the validation of annotation. The goal of this approach is to evaluate if the ligand or non-standard residue is denoted (annotated) correctly, i.e., if its structure corresponds to the 3-letter code it was assigned in the PDB file format. Specifically, the topology and stereochemistry of the validated molecule are compared to those of a correct molecule (reference molecule, model), and any differences found are reported. The first proposal of such an approach was published by Lütteke et al. and implemented in Pdb-care [83]. It contained basic validation analyses and it was purely focused only on carbohydrates. The demand for more universal and deeper ligand validation analyses and also the end of the availability of Pdb-care motivated us to accept the challenge and focus on the development of ligand validation methodologies based on the validation of annotation.

Specifically, we developed a rich set of validation analyses, which covers the main issues observed in the topology (2D structure) and geometry (3D structure) of ligands and which are important for their correct annotation [MV, VDB]. These analyses can be performed on any ligand from the Protein Data Bank. In parallel we also developed the software tools MotiveValidator [MV] and ValidatorDB [VDB] to facilitate of the analyses. MotiveValidator is a tool that enables individual ligand or sets of ligands to be validated and introduces several advanced validation analyses. ValidatorDB provides a database of weekly updated validation results for all ligands from the Protein Data Bank and introduces further useful validation analyses.

2.1.2 Validation of annotation – validation analyses

Validation analyses can be classified into three categories, namely Completeness, Chirality and Advanced analyses (Figure 2.1) [VDB].

The Completeness analyses attempt to find which atoms are missing (Figure 2.1 B), whether these atoms are part of rings (Figure 2.1 C), or the structure is degenerate, i.e., the molecule contains very severe errors (Figure 2.1 D). These severe errors may refer to residues overlapping in the 3D space, or atoms which are disconnected from the rest of the structure.

The Chirality analyses are only performed on complete structures, and aim to evaluate the chirality of each atom in the validated molecule. We distinguish between several types of chirality errors: on carbon atoms (C chirality, Figure 2.1 E), on metal atoms (Metal chirality, Figure 2.1 F), on atoms with 4 substituents in one plane (Planar chirality, Figure 2.1 G), on atoms connected to at least one substituent by a higher order bond (High order chirality, Figure 2.1 H), and the remaining chirality issues (Other chirality).

The Advanced analyses are focused on issues which are not actual chemical problems, but which can complicate further processing and exploration of the data, and thus should be noted. The Substitution analysis (Figure 2.1 I) reports the replacement of an atom with an atom of a different chemical element. The Foreign atom analysis (Figure 2.1 J) detects atoms which originate from the neighborhood of the validated molecule (i.e., having different PDB residue ID than the majority of the validated molecule), and generally marks sites of inter-molecular linkage. The Different naming analysis (Figure 2.1 K) identifies atoms whose name in PDB format is different from the standard convention for the validated molecule. The Zero RMSD analysis reports molecules whose structure is identical (root mean square deviation = 0 Å) to the model. The Alternate conformations analysis detects the occurrence of alternate conformations in the validated PDB entry.

Basic validation analyses (i.e., missing atoms and wrong chirality) were published by Lütteke et al. [83]. We introduced the detection of missing rings, substitutions, different atom naming and foreign atoms [MV]. Afterwards, we presented a classification of chirality errors and also extended the set of advanced validation analyses (i.e., zero RMSD and Alternate conformation) [VDB].

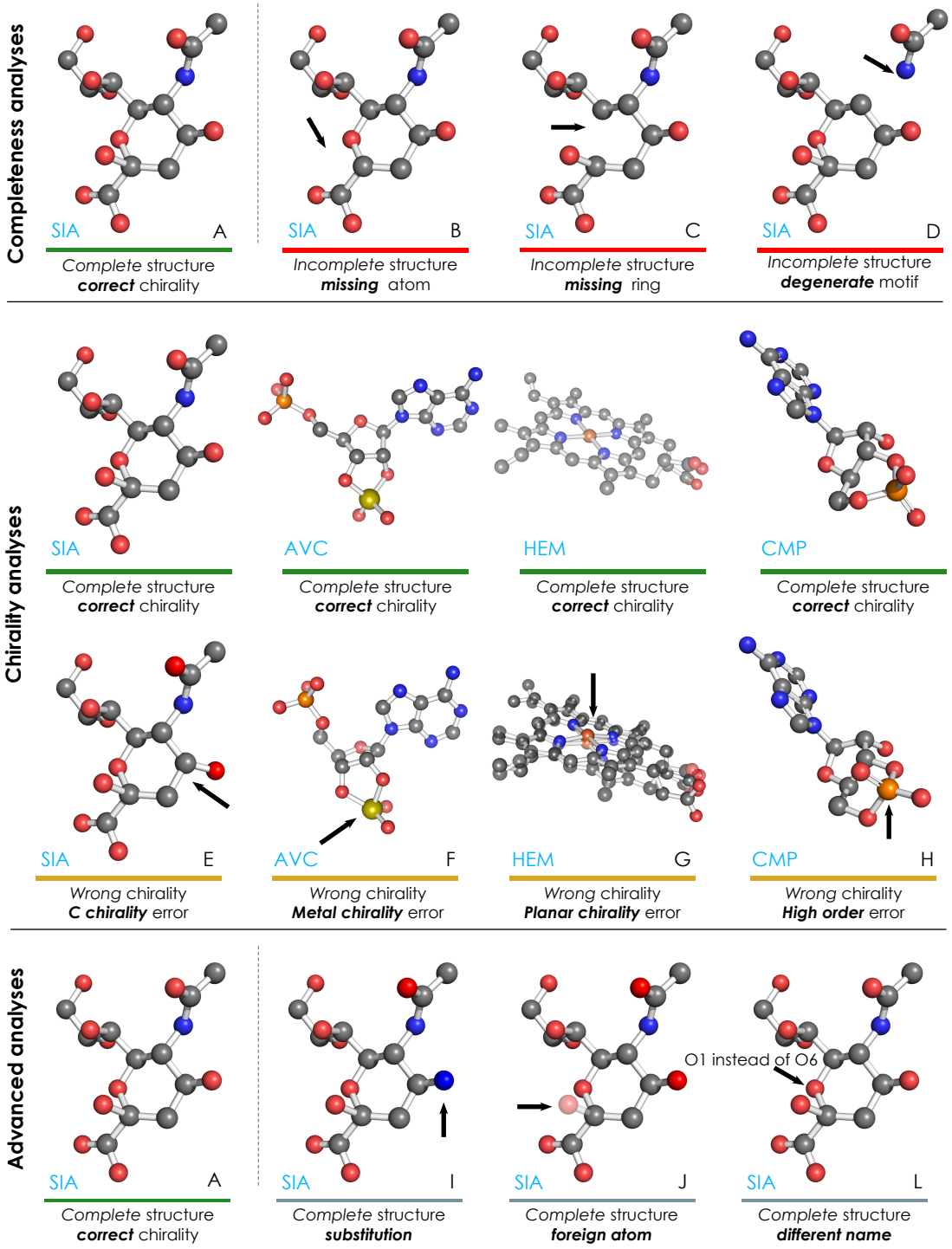


Figure 2.1: Examples of results provided by different validation analyses. ValidatorDB classifies results into three main categories (Completeness, Chirality, Advanced), each referring to several related analyses.

2.1.3 Validation – procedure and software

The ligand is validated based on its annotation (i.e., the 3-letter code of the residue in the PDB file). The ligand itself with its immediate vicinity (2 neighbouring atoms) are taken as an input for validation. The model (reference molecule for

validation) is taken from the wwPDB CCD database [84] and has the same annotation. The validation proceeds by identifying the maximum common subgraph between the input motif and the model. The validated molecule and the model are then superimposed via Sitebinder [SB] in such a way that their root mean square deviation (RMSD) is minimal. The superimposition provides a pairing (bijection) between atoms in the validated molecule and corresponding (chemically equivalent) atoms in the model. This bijection enables various properties of each atom in the validated molecule to be compared, with those of the chemically equivalent atom from the model. All the validation analyses are based on this comparison of atom properties (presence, chirality, element symbol, PDB name, etc.). A scheme of the validation procedure is depicted in [MV].

The above-mentioned procedure is incorporated into two software tools – MotiveValidator [MV] and ValidatorDB [VDB]. MotiveValidator is a web application, which enables the user to validate individual ligands or their sets. It can be found at `http://ncbr.muni.cz/MotiveValidator`. ValidatorDB is a database which summarizes the validation results for all ligands found in the Protein Data Bank. It is updated weekly based on updates to the Protein Data Bank and can be found at `http://ncbr.muni.cz/ValidatorDB`.

2.1.4 Example: Validation of nipah G attachment glycoprotein complexed with ephrin-B3

Nipah virus infection may lead to severe respiratory disease and fatal encephalitis in humans. The Nipah virus relies on the Nipah G attachment glycoprotein for host cell recognition. The crystal structure of the glycoprotein complexed with its receptor ephrin-B3 (PDB ID: 3D12 [85]) contains 30 instances of 11 different carbohydrates, each with one ring and five chiral atoms.

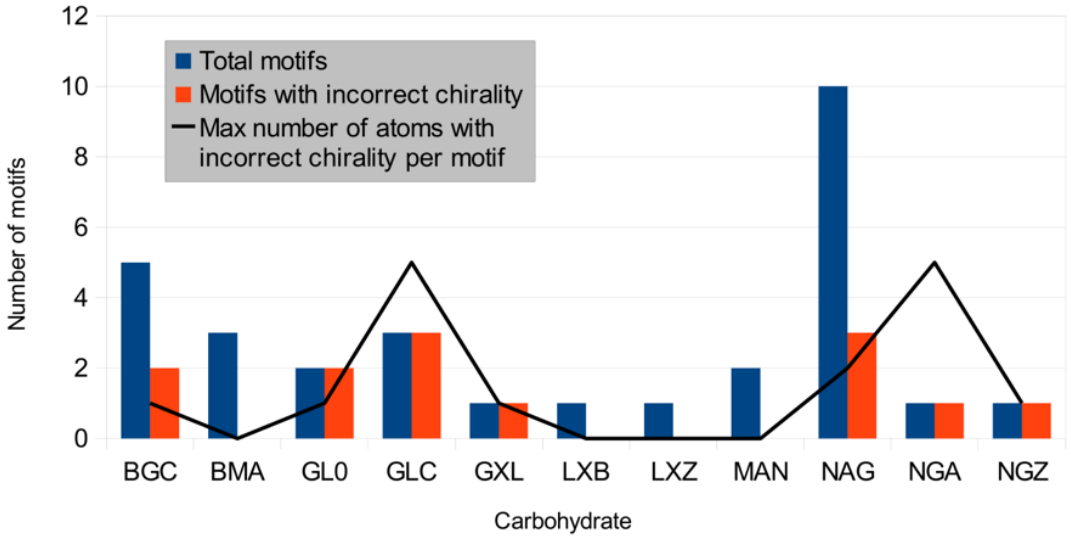


Figure 2.2: Validation of all carbohydrate ligands in Nipah G attachment glycoprotein and its receptor ephrin-B3 (PDB ID: 3D12). 13 out of 30 ligands displayed incorrect chirality of 1 to 5 atoms.

We validated all carbohydrate structures in this biomacromolecular complex using MotiveValidator [MV]. The validation showed that 13 of these ligands had incorrect chirality (see Figure 2.2). In the few cases with GLC or NGA ligands, all 5 chiral atoms exhibited incorrect chirality. Manual inspection of the structure showed further discrepancies in the ligand part (see details in [MV]).

2.2 Description, detection and extraction PTQ, MO, MO2

2.2.1 General biomacromolecular fragments PTQ

2.2.1.1 State of the art

When we need to analyze biomacromolecular structural fragments, it is first necessary to find and obtain individual samples of the fragment among the available structures. In general, such a fragment (also called a pattern) can be any set of atoms – e.g., a protein backbone, ligands or metals together with their binding sites or surroundings, specific amino acid or nucleotide sequences, atoms or residues satisfying given criteria (distance, composition, intramolecular connectivity), etc. A key step is to describe the fragment in a manner that is readable to a computer, and in this way develop a query for detecting the fragment. The fragments can be described based on their 1D structure (sequence), 2D structure (topology) or 3D structure (geometry) or via some combination of all of these aspects.

This description of the fragment requires a molecular language general enough to record all the possible properties of the fragment. At the same time, the language needs to be simple and transparent enough to be usable by the wider research community.

To date, several languages for the description of molecular structure have been designed. One of the most used languages is SMARTS (SMILES arbitrary target specifications language) [31]. It is an extension of the SMILES linear notation (used to describe the 2D structure of a molecule), that uses wild cards (i.e., expressions that can match multiple elements) for atoms and bonds. This enables queries to be specified where only partial information about the structure of the molecule is provided. Other languages describing molecular structures are for example MQL (molecular query language) [32] and SLN (Sybyl line notation) [33, 86]. In parallel, a few software tools were developed to enable the extraction of some structural patterns [11, 12, 16, 17, 20, 34]. Unfortunately, these tools are designed to operate either on a low number of structures, or their functionality is focused on very specific and narrow applications.

Because we had a strong need to automatically process (and therefore also automatically detect) the fragments, we focused on the development of an applicable molecular language. Namely, we developed the language PatternQuery [PTQ], which enables the robust and straightforward description of biomacromolecular structural fragments. We also developed a methodology and software – the PatternQuery web application [PTQ], which detects all samples of the fragment in Protein Data Bank based on its description in the PatternQuery language.

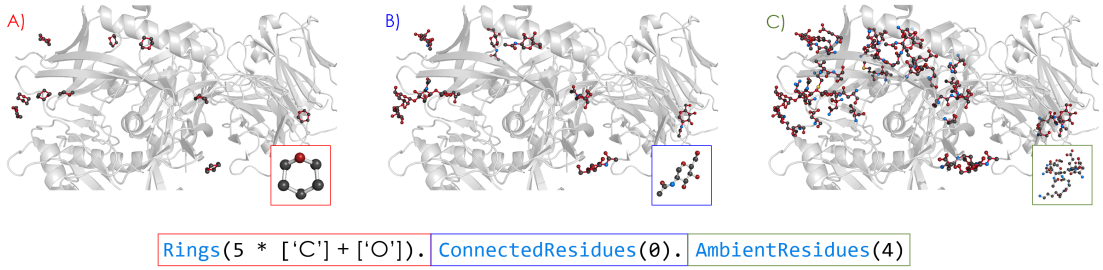


Figure 2.3: The query recognizes the binding pocket of any residue containing a pyranose moiety in the envelope glycoprotein gp160 from Human immunodeficiency virus 1 in complex with Homo sapiens immunoglobulins (PDB ID: 3U7Y). A) First, the query identifies a pyranose moiety (a ring composed of 5 carbons and an oxygen atom). B) Then, all residues which include this pattern in their structure are identified. C) Finally, all the residues that are at most 4 Å from any of the pyranose-containing residues are detected as well.

2.2.1.2 Description – molecular language PatternQuery

PatternQuery [PTQ] is a molecular language based on the Python programming language. This molecular language describes biomacromolecular structural fragments using the nature and relationship between atoms, residues, and other structural elements. The individual fragment descriptions in the language (co called queries) define the composition, topology, connectivity, and geometry of a fragment. Therefore, the PatternQuery language enables us to operate at the 1D, 2D and also 3D structure level. PatternQuery contains 110 keywords, examples of several PatternQuery keywords are given below:

- Atoms(X) returns all atoms with the element symbol X
- Residues(R1, R2) returns all residues with the 3-letter code R1 or R2
- ConnectedAtoms(F, r) returns all atoms within distance r from fragment F
- Authors(F) returns the authors of the structure containing fragment F
- Weight(F) returns the molecular weight of fragment F

The queries can be also combined into more complex ones. Figure 2.3 gives an example of a query that identifies and extracts a fragment made up of a residue containing a pyranose moiety, together with its immediate surroundings.

2.2.1.3 Detection and extraction – procedure and software

The PatternQuery language describes a fragment in such a way that it can be easily translated into a set of rules. In parallel, our methodology for finding fragments represents a biomacromolecular structure as a molecular graph, where atoms are vertexes and bonds are edges. Searching for a fragment is therefore realized as the detection of sets of atoms which meet the criteria defined in the PatternQuery language description.

This methodology is implemented in PatternQuery server [PTQ] – an interactive web application for finding and obtaining a fragment from the whole Protein Data Bank. Depending on the complexity of the defined queries and the amount

of data set entries, running the queries may take from a few seconds, up to approximately one hour (for the whole Protein Data Bank). The application PatternQuery is available at `http://ncbr.muni.cz/PatternQuery`.

2.2.1.4 Example: Detection of LecB sugar binding sites

Pseudomonas aeruginosa is an opportunistic pathogen associated with a number of chronic infections. This pathogen forms a biofilm, enabling it to survive both the response of the host immune system, and antibiotic treatment [87]. One of the cornerstones of biofilm formation is the presence of sugar-binding proteins on the outer cell membrane - LecA (PA-IL) and LecB (PA-IIL). Their inhibition is considered to be a promising approach for anti-pseudomonadal treatment [88].

We employed PatternQuery to discover sugar binding sites of similar geometry to the tetrameric LecB entry in the PDB. Specifically, we searched for 2 calcium ions at most 4 Å apart, and all the residues directly interacting with either of these ions. Furthermore, only the molecular fragments containing a residue with a furan or pyran ring were preserved. The complete query which identifies such fragments is given in [PTQ]. The initial analysis of the PDB archive revealed 355 different fragments originating from 231 PDB entries. However, the majority of the sugar moieties originated from nucleotides. To filter them out, a simple filter was employed, which gave 108 distinct fragments originating from 36 PDB entries of 7 different organisms. The majority of them originated from *P. aeruginosa*, however other pathogens such as *Ralstonia solanacearum, Burkholderia cenocepacia*, or *Chromobacterium violaceum* were identified among the organisms of origin. The sugar-binding domain in 87 fragments is composed of 3x Asp, 2x Asn and Glu and Gly residues, which is the binding site referred to as the sugar binding motif in the literature [89] for a total of 24 PDB entries from 3 organisms. In 12 further fragments a glycine residue was not present due to the fact that the structure stored in the PDB is only the asymmetric unit, rather than the expected biological unit, which is a tetramer. Finally, the remaining 9 fragments originated from 6 different pectate lyase (EC: 4.2.2.2) structures and exhibited a different binding motif to the LecB protein. Details on the analysis can be found in [PTQ]. Information about the structural similarity of the *P. A aeruginosa* LecB sugar binding site with sugar binding sites in *R. solanacearum, B. cenocepacia and C. violaceum* fully agrees with the findings published by Mitchell et al. [90].

2.2.2 Channels and pores [MO, MO2]

2.2.2.1 State of the art

A channel (or tunnel) is a pathway connecting a point inside the biomacromolecule (e.g., an active site) to an exterior one. A pore is a pathway that passes through the biomacromolecule from one point on the surface to another. Channels play significant roles in many biologically relevant systems. For example, the internal pores of ion channel proteins maintain a highly selective ionic balance between the cell interior and exterior [91–95], photosystem II channels are involved in photosynthesis [96,97], ribosomal polypeptide exit channels allow nascent peptides to leave the ribosome during translation [34], and active site access/egress channels

enable the substrate/product to enter/leave the occluded active sites of various enzymes (e.g., cytochrome P450 [75, 98–102], acetylcholinesterase [103–105], etc.). Information about the nature of active site access paths can also be utilized in biotechnology applications aimed at designing more effective and selective enzymes [106–108]. Therefore, the identification and characterization of channels are fundamental to understanding numerous biologically relevant processes and serve as a starting point for rational drug design, protein engineering and biotechnological applications.

Over the last few years, numerous computational approaches have been developed for the detection and characterization of empty spaces in biomacromolecules, particularly in proteins [109]. The main strategies used in the developed algorithms can be grouped into four classes [110]. The first class is comprised of grid-based methods, which project biomacromolecular structures onto a 3D grid, process the void grid voxels and connect them into pockets or tunnels. These methods are used in numerous software tools, such as POCKET [111], 3V [112], LIGSITE [113,114], dxTuber [115], HOLLOW [116], CHUNNEL [117], and CAVER 1.x [35]. Sphere-filling methods belong to the second class. These methods carpet biomacromolecules with spheres layer by layer. A cluster of carpeting spheres is considered a pocket. This method is implemented in SURFNET [118] and PASS [119]. The third class involves slice and optimization methods. These methods split a biomacromolecular structure into slices along a starting vector defined by the user and then optimization methods are used to determine the largest sphere. These approaches are implemented in the software HOLE [120] and PoreWalker [121]. The fourth class represents methods utilizing Voronoi diagrams, in which the shortest path is searched for from a starting point to the biomacromolecular surface. This approach was used in the previous version of MOLE 1.x [106] and it is also utilized in other software tools, e.g., MolAxis [38,39], CAVER 2.0 [36] and CAVER 3.0 [37].

In our work, we introduced a marked improvement and enrichment of the current channel detection methodology, based on Voronoi diagram utilization and found in the software MOLE. This software was originally developed by our research group and it is one of the most used and best known software tools for the detection and characterization of channels and pores. Our improved methodology was published in two steps – the first approach was part of the MOLEonline web service [MO] and further improvements were introduced as a part of the MOLE 2.0 software tool [MO2].

2.2.2.2 Channel detection – procedure and software

The algorithm for finding channels implemented in MOLE (version 2.0 and higher) involves seven steps: i) computation of the Delaunay triangulation/Voronoi diagram of the atomic centers, ii) construction of the molecular surface, iii) identification of cavities, iv) identification of possible channel start points, v) identification of possible channel end points, vi) localization of channels, and vii) filtering of the localized channels (see Figure 2.4).

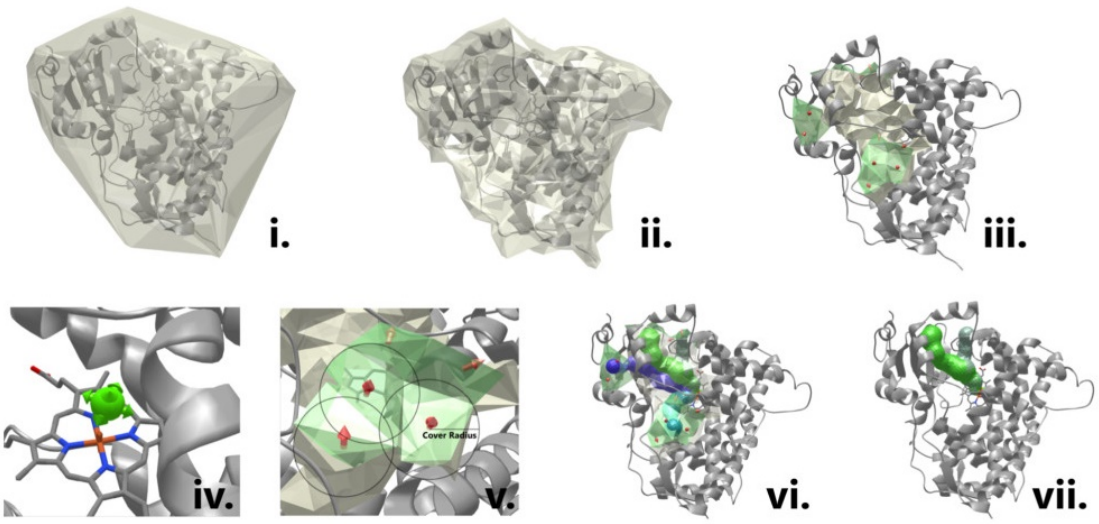


Figure 2.4: Scheme showing the steps involved in the channel calculation algorithm, illustrated for cytochrome P450 3A4 (PDB ID: 1TQN).

Step i: Computing the Delaunay triangulation/Voronoi diagram In the first step, the Delaunay triangulation of the atomic centers is computed. The Voronoi diagram is then constructed as the dual of the Delaunay triangulation.

Steps ii and iii: Approximating the molecular surface and identifying cavities

The molecular surface is approximated by the iterative removal of boundary tetrahedrons from the outermost layers (i.e., tetrahedrons found at the interface between the molecule and the external environment). Boundary tetrahedrons produced by the triangulation are removed in this step if they are sufficiently large to fit a sphere with a given probe radius. The remaining tetrahedrons form one or more connected components. We call the components that contain at least one tetrahedron on the molecular surface cavity diagrams.

Steps iv and v: Identifying possible start and end points of channels

The approach includes two ways to specify potential channel start and end points:

- Computed: Start and end points are defined as the centers of the deepest tetrahedrons in each cavity. The depth of the tetrahedron is defined as the number of Voronoi edges from the closest boundary tetrahedron.
- User-defined: Specified by a user defined 3D point. Next, cavities that have at least one tetrahedron within a defined distance from the user-specified point are found. Finally, for each such cavity, the start point is selected as the circumsphere center of the tetrahedron closest to the original point. Potential channel end points are placed in the centers of particular boundary tetrahedrons in such a way that the distance between the two end points is at least the cover radius. This is achieved by picking the largest boundary tetrahedron and marking it as an exit, then removing all boundary tetrahedrons within the cover radius. This process is repeated until all non-exit boundary tetrahedrons are removed.

Step vi: Computing channels

Once the potential start and end points have been identified, channels are computed as the shortest paths between all pairs of start and end points in the same cavity diagram. To achieve this, Dijkstra's algorithm is used. At this stage, each channel is represented by a sequence of tetrahedrons. The next step is to approximate the channel centerline by a natural cubic spline of the circumsphere centers of the tetrahedrons. Additionally, a "radius spline" is computed that determines the centerline distance to the closest atom van der Waals sphere.

Step vii: Filtering of channels The above-described steps usually generate a large number of channels. However, many of these channels are either too narrow (i.e., have a bottleneck with a small radius) to be considered relevant or are duplicated (i.e., too similar to each other). To obtain the most relevant channels, the methodology contains a filtering of too narrow and too similar channels.

This methodology is implemented in MOLEonline [MO] – an interactive web application for finding channels and pores. In parallel, its implementation is also available in MOLE 2.0 [MO2] – a standalone software package focused on the detection and characterization of channels and pores. Both these tools are available at `http://ncbr.muni.cz/mole`.

2.2.2.3 Example: Detection of channels in cytochrome P450

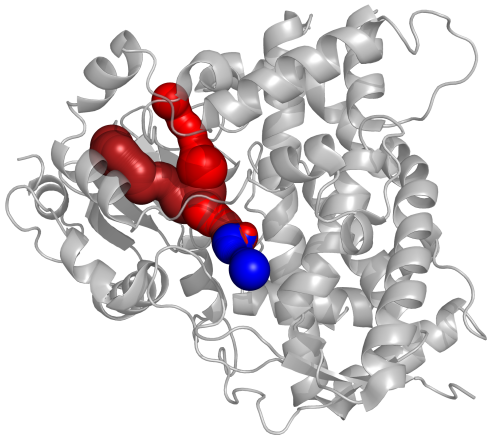


Figure 2.5: Results of channel analysis of Cytochrome P450 3A4 (CYP3A4). Three channels found from a user-specified starting point (calculation started from Glu 308 and Thr 309 according to the CSA]) are shown – the solvent channel (in blue), the channel 2a (in dark red) and the channel 2e (in light red).

Microsomal Cytochrome P450 (CYP) enzymes are important for the metabolism of many endogenous compounds and xenobiotics [122, 123]. CYPs share a buried active site [124], which is connected to the outside environment by various access/egress channels [102]. These channels are responsible for substrate passage

to and product release from the active site, and they are considered to be involved in the substrate preferences of CYP, which has been shown to vary considerably among CYP enzymes [75, 100].

Figure 2.5 shows all the channels connecting the active site of an enzyme CYP 3A4 (PDB ID: 1TQN) with the exterior (detected via MOLEonline). The top-ranked channel found by MOLEonline (blue in Figure 2.5) is the solvent channel (as described in [102]). The solvent channel is 10 Å long and its bottleneck is 1.41 Å wide. These findings are consistent with previous data, which have identified the solvent channel as the main channel responsible for active site solvation [125]. Other two channels are 2x family channels (as described in [102]), which are considered to be involved in hydrophobic substrate binding. Their position and shape fully agree with published information [100, 126].

2.3 Comparison SB

2.3.1 State of the Art

Comparing biomacromolecular structural fragments is a necessary task during the analysis of these fragments. It enables e.g. multiple fragments to be processed as one sample or to classify fragments according their similarity. In addition, knowledge concerning fragment similarity can help us to understand the relationship between a protein's structure and function, to classify biomacromolecules, to identify evolutionary relationships between proteins, and to obtain patterns for drug discovery [11–20].

The comparison of 3D structures is a complex topic that can be divided into several subtopics. We distinguish between methods that compare compounds with identical (or very similar) 2D structure, as opposed to methods dealing with compounds for which the 2D structure differs significantly. In our work, we were mainly focused on a comparison of molecules with identical (or very similar) 2D structures (also called superimposition or superposition), because it is very helpful in processing biomacromolecular fragments. The superimposition consists of two interdependent stages [SB]. First, it is necessary to find the correspondence (atom pairing) between the atoms coming from different structures. In the second step, the sets of paired atoms (3D points) are fitted together as tightly as possible by a geometrical transformation (optimal fitting). There are several heuristics and algorithms to obtain an atom pairing.

Implicit pairing associates atoms with the same index or position (i.e., pairing the i-th atom of the first molecule to the i-th atom of the second molecule). Many state-of-the-art programs that offer the superimposition of organic molecules (e.g. Chimera [40], VMD [41], Gromacs [42], gOpenMol [43], Pymol [44]) use implicit pairing.

Employing sequence alignment is an improvement on the implicit pairing approach. First, the sequence alignment is performed by a selected algorithm (e.g., Needleman and Wunsch alignment [127], ICM ZEGA alignment [128] etc.). Afterwards, the atoms from the aligned residues are paired using an implicit pairing. This approach is only applicable for the superimposition of proteins or protein sequences with a reasonable degree of sequence similarity. Several drug design

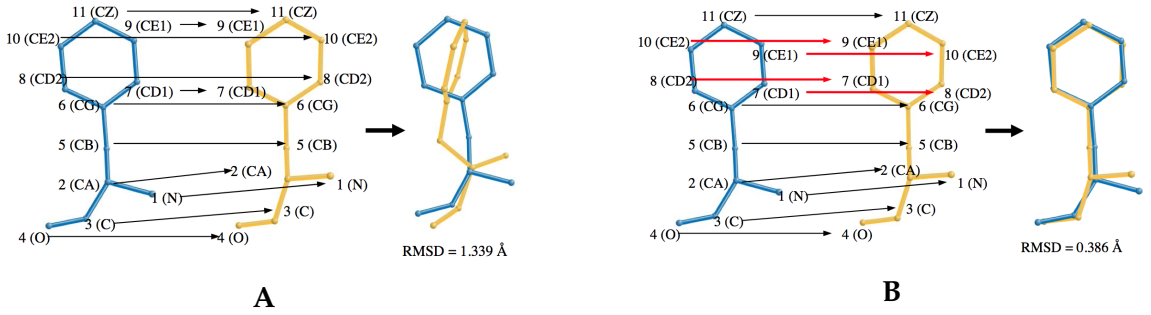


Figure 2.6: A) Implicit pairing between residues PHE 83 (blue) and PHE 91 (orange) from the PDB entry 2WH6, and the superimposition calculated by the program VMD, which uses this pairing. B) The best possible pairing between PHE 83 (blue) and PHE 81 (orange) from 2WH6 and the superimposition calculated by our program SiteBinder, which is able to find this pairing. The differences compared to the implicit pairing are indicated with red arrows. In both A) and B), atoms are denoted by their number in the residue, while their PDB name is in brackets.

packages (e.g., MOE [45], Discovery Studio [46], ICM [47], etc.) implement this approach.

The systematic (combinatorial) approach finds all possible pairings and is therefore very robust, but in parallel it can be nontrivially time-consuming.

Subgraph matching, which was originally developed for processing molecules with different 2D structures, can also be used to find a relevant pairing. This approach identifies the largest possible atom sets that can be superimposed.

When an atom pairing has been found, the paired 3D points are fitted by performing a geometrical transformation. The transformation can be found via an iterative approach [129], via a closed form solution that utilizes rotation matrices [130] or by the application of a quaternion algebra [131–133], which is currently the most frequently used solution.

The comparison of fragments has several specifics. For example, the order of atoms in the fragments can be different, therefore the implicit atom pairing can neither be used for fragments nor for residues. Figure 2.6 demonstrates how sensitive a superimposition is to the choice of atom pairing.

The next challenge in fragment superimposition is the fact that a comparison of many fragments simultaneously (e.g. several thousands) is required and often also a common pattern needs to be detected. This motivated us to develop a methodology tailored to fragment superimposition and capable of handling its difficulties. The methodology is implemented in the web application SiteBinder [SB].

2.3.2 Fragment comparison – procedure and software

Superimposition of two fragments: Our methodology [SB] provides two superimposition approaches – a combinatorial approach and a subgraph matching approach.

The combinatorial approach first generates a set of all chemically meaningful atom pairings. These pairings are generated in such a way, that first the atoms in both fragments are divided into subsets according their properties. Specifically, the subsets can be created according to a residue name, a residue identifier and

an element symbol. Afterwards, all pairings between fragments which connect atoms from the same subsets are generated. Then for each pairing, the optimal fit is performed using a state-of-the-art quaternion algebra approach [134]. Finally, the pairing which provides the closest fit is selected, and the fit calculated using this pairing is taken as the result. This approach can only superimpose fragments containing the same number of atoms for each element symbol.

The subgraph matching approach first detects the largest subgraph of the two fragments. Afterwards, it generates the atom pairings based on the subgraph. For all the pairings, the optimal fit is calculated again using the quaternion algebra approach. The best fit is then taken as the result.

Superimposition of multiple fragments: We used a multiple superimposition approach published by Wang et al. [135], adapted it to biomacromolecular fragments and combined it with our algorithm for the superimposition of two fragments. Our multiple superimposition approach works in two steps. First, each fragment is superimposed onto the first one. Afterwards, we calculate an average fragment as the arithmetic average of the x, y and z coordinates of the corresponding atoms. Next, all the fragments are superimposed onto the average fragment. The new coordinates of all these superimposed fragments are used as an input to the next iteration of the multiple superimposition approach. The iterative superimposition process ends when a further iteration is not able to improve the fit.

2.3.3 Example: Comparison of Zn binding sites in zinc fingers

Cys2His2 zinc fingers are one of the most common structural motifs in eukaryotes - each finger recognizes three to four base pairs of DNA. There is also evidence that some Cys2His2 zinc fingers bind RNA and that others may participate in protein-protein interactions. Individual fingers contain approximately 30 amino acids, and the hallmark of the motif is the presence of two cysteines and two histidines that serve as zinc ligands. It can be defined as the pattern X2-CYS-X2-4-CYS-X12-HIS-X3-5-HIS, where X represents any amino acid residue. We used SiteBinder to determine whether the center of the zinc finger motif (i.e., two CYS, two HIS and a Zn atom) has a conserved geometry. First, we collected all zinc fingers from the Protein Data Bank. Specifically, we found 329 zinc fingers from 205 different Protein Data Bank structures. Afterwards, we performed four superimpositions for our set of zinc finger central motifs. These procedures differed in the number of atoms selected for superimposition (displayed in red in Figure 2.7). The results showed that the part of the motif which closely surrounds Zn has a conserved structure and the conformation of more distant parts of CYS and HIS may differ. We then focused on a special group of zinc finger Cys2His2 motifs, namely those known to bind RNA. Their superimposition showed that the motifs coming from the PDB entry 1ZU1 are markedly different than those in the other investigated RNA binding proteins. This is probably explained by the fact that one of the two HIS residues faces the binding site with the opposite face of the imidazole ring (see Figure in [SB]). This structural dissimilarity correlates with a

functional difference. Specifically, the 1ZU1 zinc-finger motif has evolved to bind double-stranded RNA [136].

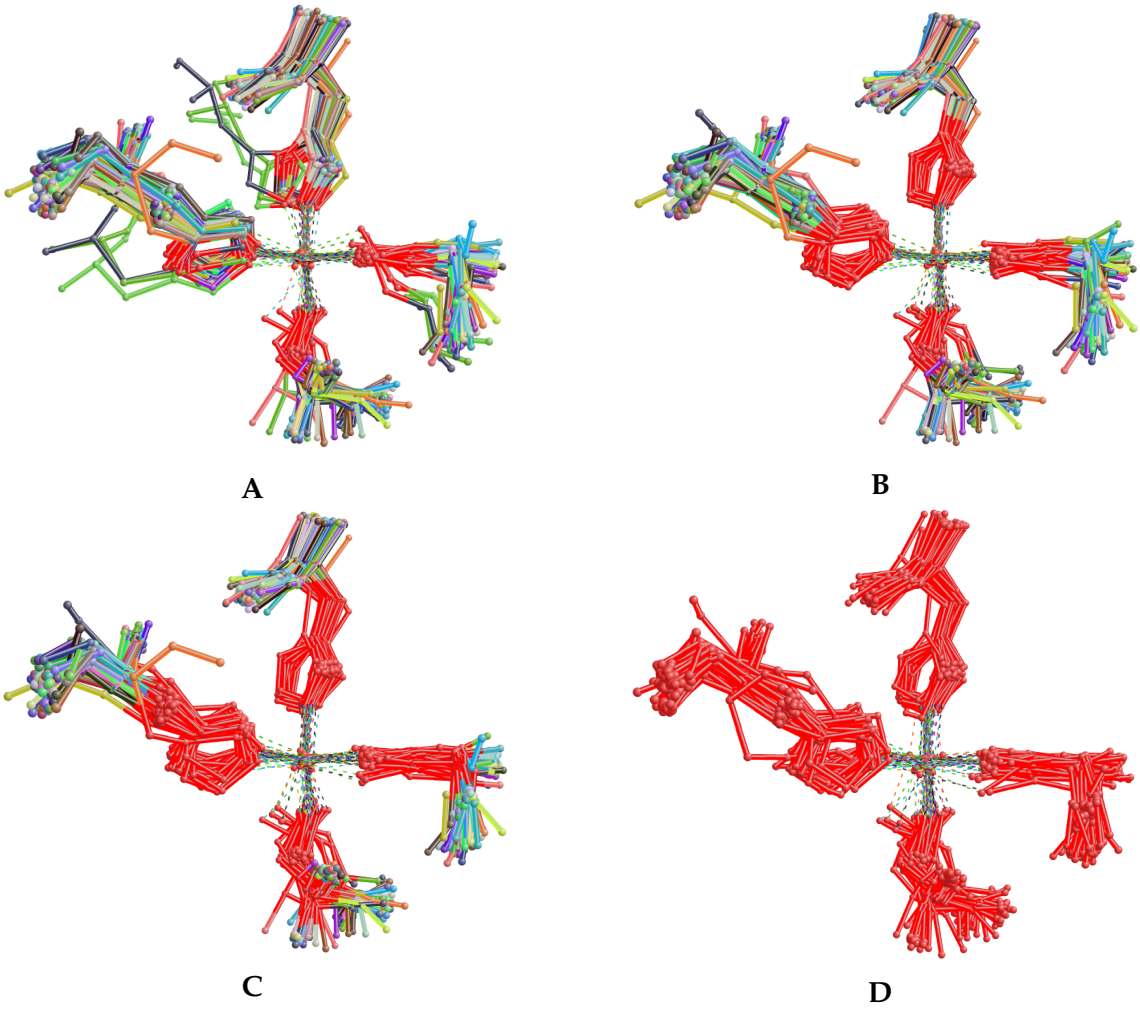


Figure 2.7: Superimposition of 329 zinc finger central motifs. A) 9 superimposed atoms, RMSD 0.501 Å , B) 15 superimposed atoms, RMSD 0.527 Å , C) 19 superimposed atoms, RMSD 0.599 Å , D) 33 superimposed atoms, RMSD 0.840 Å . Atoms used in the superimposition procedure are in red. For ease of visual interpretation, only the first 80 motifs are displayed.

2.4 Characterization

EO, EB, EL, EPM, EAM, BAX, MO, MO2

2.4.1 Partial atomic charges

2.4.1.1 State of the art

Partial atomic charges are real numbers describing the distribution of electron density in a molecule, thus providing clues to the chemical behaviour of the molecules. The concept of charges began to be used in physical chemistry and organic chemistry. Afterwards, partial atomic charges were adopted by computational chemistry and molecular modelling, where they are used to calculate electrostatic interactions, describe the reactivity of the molecule etc. Specifically, they are applied in molecular dynamics, docking, conformational searches, binding site pre-

dictions etc. Recently, partial atomic charges also became popular in chemoinformatics and bioinformatics, as they proved to be informative descriptors for QSAR and QSPR modelling [PQ, PE, PS] [54, 56, 137–140] and for other applications [141–143]; they can be utilized in pharmacophore design [144–146], virtual screening [147–149], similarity searches [150–152], molecular structure comparison [153–155] etc.

Partial atomic charges cannot be determined experimentally or derived straightforwardly from the results of quantum mechanics (QM), and many different methods have been developed for their calculation. The most common method for charge calculation is an application of the QM approach and afterwards the utilization of a charge calculation scheme. Charge calculation schemes can use orbital-based population analysis, wave-function-dependent physical observables or the reproduction of charge-dependent observables. Examples of orbital-based population analyses are Mulliken population analysis (MPA) [156, 157], Löwdin population analysis [158] and Natural population analysis (NPA) [159, 160]. Wave-function-dependent physical observables are applied in the atoms-in-molecules (AIM) approach [161,162], Hirshfeld population analysis [163–165], CHELPG [166] and Merz-Singh-Kollman (MK) [167, 168] method. The reproduction of charge-dependent observables is embedded in the CM1, CM2, CM3, CM4, and CM5 approaches [169, 170].

Unfortunately, QM charge calculation approaches are very time-consuming. A markedly faster alternative is to employ empirical charge calculation approaches, which can also provide highly accurate charges. These approaches can be divided into conformationally-independent, which are based on 2D structure (e.g., Gasteiger's and Marsili's PEOE [171, 172], GDAC [173], KCM [174], DENR [175]) and conformationally-dependent, calculated from the 3D structure (e.g., EEM [48], QEq [176] or SQE [177,178]). Conformationally-dependent charges are considered to be more suitable for the characterization of biomacromolecules and their fragments. The reason is that these charges contain extensive information not only about the chemical surroundings of atoms, i.e., its topology (2D structure based charges), but also the geometry and "chemical quality" of the surroundings. Such information is missing, for example, in force field charges which use averaged atomic charges from large sets of structures. Therefore we focused purely on this category of atomic charges.

EEM (electronegativity equalization method) is the most frequently used conformationally-dependent empirical charge calculation approach. It calculates charges using the following system of linear equations:

$$\begin{pmatrix} B_1 & \frac{\kappa}{R_{1,2}} & \cdots & \frac{\kappa}{R_{1,N}} & -1 \\ \frac{\kappa}{R_{2,1}} & B_2 & \cdots & \frac{\kappa}{R_{2,N}} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\kappa}{R_{N,1}} & \frac{\kappa}{R_{N,2}} & \cdots & B_N & -1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ \bar{\chi} \end{pmatrix} = \begin{pmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_N \\ Q \end{pmatrix} \tag{2.1}$$

where q_i is the charge of atom i; $R_{i,j}$ is the distance between atoms i and j; Q is the total charge of the molecule; N is the number of atoms in the molecule; $\bar{\chi}$ is the molecular electronegativity, and A_i, B_i and are empirical parameters. EEM is not only a rapid charge calculation approach, but it can also provide highly accurate

charges, i.e., they can mimic the QM charges for which EEM has been parameterized. Therefore, many EEM parameter sets for various QM charge calculation approaches were published later or recently [50, 51, 53, 179, 180]. Also, we participated in the EEM parameterization process – we published a parameterization for organic molecules [EO], for biomacromolecules [EB] and for ligands [EL].

In parallel, a few freely available software tools also include an EEM charge calculation method [181, 182]. We also contributed to this field, specifically we developed a methodology for calculating of EEM charges in large biomacromolecules – first a parallel approach [EPM] and then an approximative method [EAM]. We also created an OpenBabel patch for calculating the EEM charges in ligands and drug-like molecules [EL].

2.4.1.2 Parameterization of EEM EO, EB, EL

EEM parameterization for organic molecules We first focused on EEM parameterization more than ten years ago. Our goal was to provide EEM parameters for organic, organohalogen and organometal molecules. For this reason, we developed the following EEM parameterization procedure:

- Selection of set of molecules for parametrization.
- Calculation of QM atomic charges for all molecules in the selected set (via Gaussian).
- Calculation of average electronegativity of molecule:

$$\bar{\chi} = N \left(\sum_{i=1}^{N} \frac{1}{x_i^0} \right)^{-1} \tag{2.2}$$

- Calculation of x and y pairs for each atom:

$$x = q_i \tag{2.3}$$

$$y = \chi_i - \kappa \sum_{j=1\ (j \neq i)}^{N} \frac{q_j}{R_{i,j}} \tag{2.4}$$

  Note: This is calculated for all values of in a defined set.

- Division of x and y pairs into sets according to the atom that the pair belongs to.
- Calculation of parameters A and B for each set of x and y pairs.
- Finding the optimal κ value.

The methodology enabled us to perform parameterizations based on really large sets of organic molecules. Specifically, we did EEM parameterization on 12 training sets selected from a database of predicted 3D structures (NCI DIS) and from a database of crystallographic structures (CSD), where each set contained

2,000 to 6,000 molecules. The results of this parameterization confirmed that the number of molecules in the training set is very important for the quality of the parameters. One result of our EEM parameterization was improved EEM parameters (based on HF/STO-3G/MPA QM charges) for elements that were already parameterized, specifically: C, O, N, H, S, F and Cl. We also calculated new parameters for elements not yet parameterized, specifically for Br, I, Fe and Zn. The results of this parameterization are summarized in the article [EO].

EEM parameterization for biomacromolecules We then focused on EEM parameterization for biomacromolecules. Specifically, we prepared EEM parameters for proteins containing calcium and no other metals nor any ligands. The proteins itself are too large to perform QM charge calculations on them. Therefore, the inputs for our EEM parameterization were protein fragments. These fragments included the Ca atom and its surroundings. Therefore, we adapted the parameterization procedure in the appropriate way:

- Selection of a set of proteins for parameterization.
- Preparation of their fragments.
- Calculation of QM atomic charges for all fragments in the selected set (via Gaussian).
- Common EEM parameterization procedure (the same as for organic molecules).

This procedure enabled us to perform EEM parametrization for large protein fragments as input structures. We calibrated the EEM parameter sets using atomic charges computed by three population analyses (MPA, NPA, iterative Hirshfeld), at the Hartree-Fock level with two basis sets (6-31G*, 6-31G**) and in two environments (gas phase, implicit solvent). Thus, we produced 24 sets of EEM parameters. Afterwards, we did an external validation of these EEM parameters on two reference proteins (insulin and ubiquitin) and we found, that all EEM parameter sets reproduce QM charges very accurately. The results of this parameterization are summarized in the article [EB].

EEM parameterization for ligands We recently found, that even though several EEM parameter sets have been published for organic molecules, none of them were rich enough to cover common drug-like molecules, ligands and consequently also biomacromolecular fragments. The reason for this was that the available parameters only focused on a very limited part of the chemical space. Specifically, they contained only parameters a for few elements, and even for these elements only some bond orders were available. This strongly limits the applications of EEM in fragment characterization and also its usage in chemoinformatics and bioinformatics.

For this reason, we took our EEM parameterization methodology for organic molecules [EO], improved its efficacy (developing the parameterization tool NEEMP [183]) and employed it to obtain universal and accurate EEM parameters for drug-like molecules and ligands. In this way we prepared six EEM parameter sets. They enable the user to calculate EEM charges in a quality comparable to quantum mechanics (QM) charges based on common charge calculation schemes (i.e.,

MPA, NPA and AIM) and a robust QM approach (HF/6-311G, B3LYP/6-311G). The calculated EEM parameters exhibited very good quality on a training set and also on a test set. They were applicable for at least 95% of molecules in key drug databases (Drugbank, ChEMBL, Pubchem and ZINC) compared to less than 60% of the molecules from these databases for which the other EEM parameters can be used. The results of this parameterization are summarized in the article [EL].

2.4.1.3 EEM charges – procedure and software EPM, EAM, EL

Optimized and parallelized EEM charge calculation method (EEM SOLVER) When we started our research in the field of EEM atomic charges and their applications, we found that even though the EEM method is published, there is no available software tool for implementing it. This motivated us to prepare such a tool. Specifically, we developed the software EEM SOLVER, which enables the user to calculate EEM charges based on the inputted EEM parameters, which he/she provides. EEM SOLVER is a command line application and has two versions – a serial version, which is able to calculate EEM charges in small organic molecules, and a parallel version which can be also used for large biomacromolecules. The serial version includes this straightforward methodology:

- Filling the EEM matrix
- Transforming the matrix into row echelon form via Gaussian elimination
- Solving the equation system

The parallel version replaces the Gaussian elimination (the most complex part of the method) with the parallel algorithm WIRS [184] and implemented this algorithm within the PVM environment [185].

We also demonstrated the accuracy and performance of EEM SOLVER with several examples. The parallel version was even applicable for molecules with more than 1,000 atoms. These results are summarized in the article [EPM] and the software EEM SOLVER is available here: `http://ncbr.muni.cz/~svobodova/eem_abeem/`

EEM charges for large biomolecular complexes and drug-like molecules (ACC) Recently, the necessity to process much larger biomacromolecules and also a community demand for web-based software forced us to provide a new software solution for the calculation of EEM charges. Specifically, we developed the web application Atomic Charge Calculator (ACC). ACC embeds all published EEM parameters and also enables the utilization of user-provided EEM parameters. ACC can perform charge calculations for large sets of organic molecules (e.g., ten thousands of molecules or more). In parallel, ACC is also able to calculate EEM charges on really large biomacromolecular systems (e.g., close to a hundred thousand atoms). ACC processes small molecules using the same methodology as EEM SOLVER.

In parallel, it offers two new approaches for EEM calculation on large biomacromolecules – EEM Cutoff and EEM Cover. These approaches work by splitting the EEM matrix into multiple smaller matrices. With the EEM Cutoff approach, for

each atom in the molecule, ACC generates a fragment made up of all atoms within a cutoff radius R of the original atom. Thus, for a molecule containing N atoms, the EEM Cutoff approach solves N smaller EEM matrices describing a set of N overlapping fragments from the original molecule. This markedly reduces the complexity and time demands of the algorithm. EEM Cover provides another streamlining of the calculations. It also splits the EEM matrix into smaller matrices, but it only generates fragments for a subset of atoms in the molecule. Thereore, the number of EEM matrices that need to be solved is reduced by at least 50% compared to EEM Cutoff, while maintaining high accuracy. Both the EEM Cutoff and EEM Cover method are approximative approaches, but their accuracy is very high and in fact comparable to standard EEM.

We demonstrated the performance and applicability of ACC on three case studies – a set of organic molecules, a set of antimicrobial peptides and a large proteosome (more than 80,000 atoms). These results are summarized in the article [EAM] and the software EEM SOLVER is available here: `http://ncbr.muni.cz/ACC`

EEM charges for ligands and drug-like molecules (OpenBabel patch) The EEM method was also implemented in the software tool OpenBabel. A weak point of this implementation was that it only allowed one set of EEM parameters to be used (namely, EEM parameters published by Bultnick et al. [50], based on QM charge calculation approach B3LYP/6-31G*/MPA). These parameters only cover a few elements (i.e., C, O, N, H, F). Unfortunately, OpenBabel replaced the EEM parameters for missing elements with the EEM parameters for some other atom types. This nonstandard approximation caused inaccuracies in the results. For this reason, we provided an OpenBabel patch which introduces our EEM parameters published in [EL]. Closely after its release, the developers of OpenBabel incorporated our solution directly into OpenBabel.

2.4.1.4 Example: Docking of glycerol into ubiquitin based on different charges [EAM]

To show the applicability of EEM charges in proteins, we performed the common docking calculation of glycerol into ubiquitin. Specifically, we used the same ubiquitin molecule which was employed for the external validation of the EEM parameters for proteins. Glycerol was chosen as a potential ligand because it has been found to stabilize the native state of ubiquitin [186]. The ligand's initial conformation was taken from the coordinates of the ideal glycerol model available in Ligand Expo [84] and contained 5 rotatable bonds.

In this experiment, we compared the docking results obtained in the ideal case (therefore via using accurate QM charges) with the results obtained via various empirical charges. The QM charges were calculated by HF/6-31G*/MPA. The empirical charges were the EEM charges (calculated by our parameters mimicking HF/6-31G*/MPA and published in [EB]), the Gasteiger-Marsili charges and the AMBER ff94 charges. The results of the docking are depicted in Figure 2.8.

The figure shows that the docking results obtained using QM charges are well reproduced via the EEM charges given by the above-mentioned EEM parameters. The EEM binding pose differs by 0.07 kcal/mol and an RMSD of 0.131 Å from the

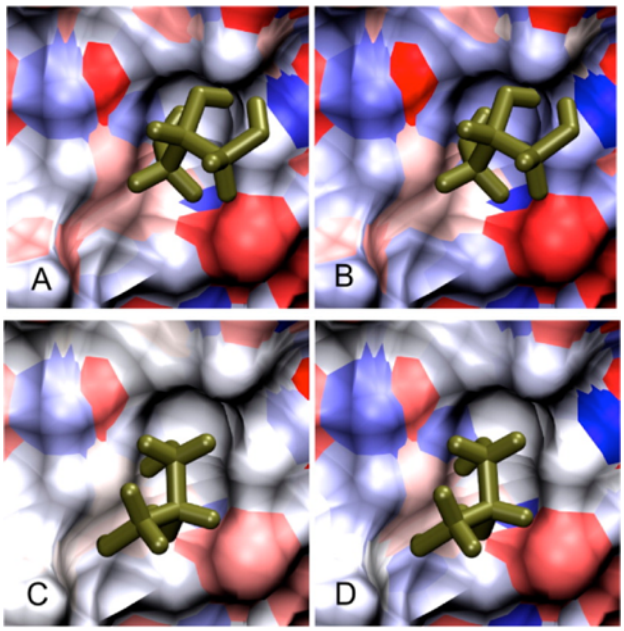


Figure 2.8: Docking of glycerol into ubiquitin via: A) HF/6-31G*/MPA QM charges: estimated binding energy −9.64 kcal/mol. B) HF/6-31G*/MPA mimicking EEM charges: estimated binding energy −9.71 kcal/mol, RMSD 0.132 Å compared to the QM pose. C) Gasteiger-Marsili charges: estimated binding energy -8.65 kcal/mol, RMSD 3.244 Å compared to the QM pose. D) AMBER ff94 charges: estimated binding energy −8.7 kcal/mol, RMSD 3.235 Å compared to the QM pose.

binding pose given using QM charges. For comparison, using Gasteiger-Marsili or AMBER ff9469 charges on ubiquitin produces different binding poses. Specifically, the binding pose given by Gasteiger-Marsili charges differs from that given by QM charges by 0.99 kcal/mol and an RMSD of 3.244 Å . The binding pose given by AMBER ff94 charges differs by 1.57 kcal/mol, and an RMSD of 3.235 Å .

2.4.2 Channel characteristics MO, MO2

2.4.2.1 State of the art

Closely connected to channel detection discovery is research in the field of channel characterization, i.e., the calculation of channel properties. Channel properties can be divided into three classes – geometrical (length, radius, bottleneck radius, etc.), chemical (lining residues of the channel, their neighborhoods etc.), and physicochemical (e.g., polarity, charge, hydropathy). These properties are essential for understanding the channel's chemical and biological role. Based on them we can evaluate which compound (how large, positively or negatively charged, polar or nonpolar) can pass through the channel and how easy or difficult it will be from the energetic point of view. Therefore the channel's properties are a clue to the substrate specificity of a channel.

Basic chemical and geometrical properties (lining residues, channel length and radius) are automatically provided as a side product of all the channel detection methodologies. Therefore even the initial versions of channel detection software (MOLE 1.0 [106], Caver 1.0 [35], MolAxis [38, 39]) returned them as additional data about the detected channels. The computation of more sophisticated geo-

metrical properties (bottleneck radius, local minima radius etc.) was included in channel calculation approaches later, and is provided by recent versions of channel discovery tools [37]. Obtaining the physicochemical properties of the channel is relatively nontrivial. Therefore, despite their high usefulness, channel physicochemical properties only started to become available a few years ago. Specifically, they were introduced by the software tool MOLEonline [MO] and are also provided by MOLE 2.0 [MO2] .

We contributed to this field by developing the methodologies for calculating physicochemical and advanced geometrical properties. These methodologies are incorporated in the software tools MOLEonline [MO] and MOLE 2.0 [MO2].

2.4.2.2 Channel properties

A channel can be viewed as a void volume inside the biomacromolecular structure, and it can be described using the arrangement of residues which surround this empty volume. Highly interesting parts of the channel are its local narrowings, which are referred to as local minima. The global minimum of the channel is then referred to as the bottleneck.

There are three recognized types of channel properties – geometrical, chemical and physicochemical. The chemical properties of the channel are focused on the residues which surround the channel. The best known chemical property is the so-called lining residues, which describes the residues which are found in the channel walls. These chemical properties also include local minima residues, bottleneck residues and various derived criteria such as the second layer of the channel (residues directly adjacent to the lining residues) etc.

The geometrical properties of the channel describe its geometry characteristics. Basic geometrical properties are channel length and the radius of the channel at a specific point. Important points for measuring the radius are the bottleneck and other local minima. Also the 3D position of the centerline (line composed of points in the center of the channel) and a profile of the channel are widely used geometrical properties.

The most complex properties are the physicochemical properties. Nowadays, channel discovery methodologies only provide values for a few of them, i.e., hydropathy, polarity, mutability and charge. Hydrophobicity and hydrophilicity are two extremes of a spectrum, commonly referred to as hydropathy, and describe the tendency of a molecule to interact with water [187]. Polarity is the property of a molecule given by the separation of electric charge, leading to the molecule having electric poles. The mutability (or relative mutability) quantifies the tendency of an amino acid to be substituted (mutated) in a protein's structure. Substitution with similar amino acids generally retains protein function, while substitution with amino acids with different properties may affect the protein's structure or function. Relative mutability is high for easily substitutable amino acids (e.g., small polar residues) and low for amino acids which play a significant role in the protein structure (e.g., amino-acids with substrate binding or catalytic activity). Charge describes the localization of charged residues in the channel.

2.4.2.3 Channel properties – procedure and software

Our channel characterization methodology first computes channel lining residues and a channel centerline. The chemical and geometrical properties of the channel are calculated in a straightforward manner by the application of linear algebra algorithms.

A calculation of physicochemical properties is performed in the following way: the properties are calculated based on tabulated values for unique residues surrounding the channel (article [MO2] contains information about these tabulated values). Specifically, the property values for all lining amino acids, which have a side chain oriented towards the channel, are summarized and then averaged. The calculation of individual properties contains specialized steps. Hydropathy is computed using the Kyte-Doolittle scale [187] and when the amino acid has its main chains oriented towards the channel, tabulated hydropathy values for glycine (Gly) are used. With polarity, the value for asparagine (Asn) is used for amino acids with the main chain oriented towards the channel. Amino acid residues that have their main chains lining the channel are not considered when computing mutability. Charge is calculated by a different procedure – as the sum of the charges of individual amino acid side chains. We use the charge (protonation or deprotonation) at physiological pH, therefore lysine and arginine are positively charged, whereas aspartic and glutamic acids are negatively charged. Despite the protonation state of histidine being dependent on its microenvironment, we treat all histidines as positively charged. This simplification only produces a slight inaccuracy and enables us to perform the calculation faster. This methodology is implemented in the web application MOLEonline [MO] and in the software package MOLE 2.0 [MO2]. Both these tools are available at `http://ncbr.muni.cz/mole`.

2.4.2.4 Example: Physicochemical properties for known channels

We evaluated the physicochemical properties calculated by MOLE 2.0 for several biomacromolecules containing biologically important channels/pores with known functionality and known functions.

Gramicidin D (PDB ID: 1GRM) is known to form a polar pore in membranes (Figure 2.9 A) [188], which agrees with the physico-chemical properties identified using MOLE 2.0.

In the cytochrome c oxidase (PDB ID: 1M56), MOLE 2.0 identified two channels with different polarities (Figure 2.9 B), which may be involved in the transfer process required for the proper functioning of this enzyme [189].

For the nicotinic acetylcholine receptor (PDB ID: 2BG9), MOLE 2.0 computed the central pore (Figure 2.9 C) lined with 18 negatively charged amino acids. This could explain the experimentally observed selectivity for cation permeation [190].

Afterwards, we analyzed carbonic anhydrase (PDB ID: 3EYX), a biomacromolecule which can utilize the inorganic carbon sources CO_2 and HCO_3^- [191]. MOLE 2.0 predicted that the channel is highly polar (Figure 2.9 D), in agreement with expectations.

These results demonstrate that physicochemical properties may provide useful information about the nature of the channel and its biological function.

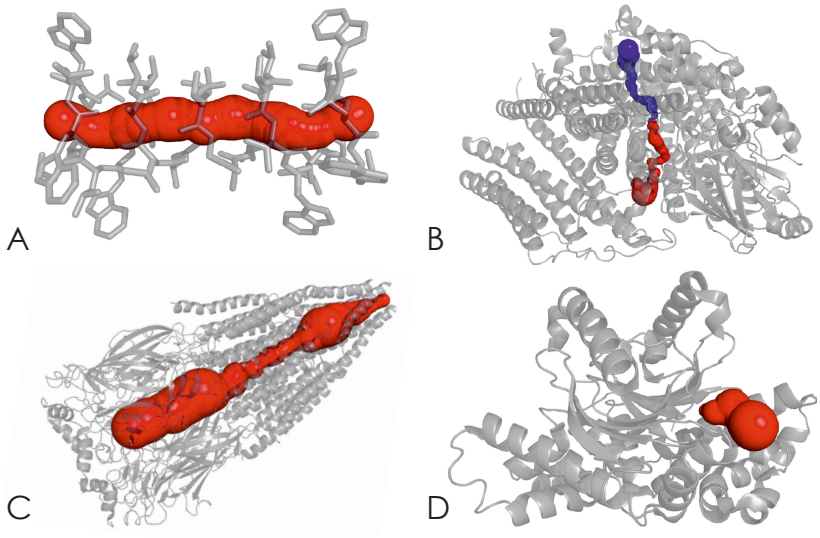


Figure 2.9: Comparison of known properties of biologically important channels/pores and their polarity calculated with MOLE 2.0. A) gramicidin D, B) cytochrome c oxidase, C) nicotinic acetylcholine receptor, D) carbonic anhydrase. The nonpolar channels are in blue and the polar channels in red.

3

Selected applications

3.1 Validation: Evaluation of quality for different classes of molecules [VDB]

3.1.1 Introduction

It is a well-known fact that the validation of biomacromolecular structures is an important issue for the life sciences community [22–25]. Several studies showed that ligands are the most problematic and corrupted parts of biomacromolecules [30]. Thanks to current software tools [26–29], we can evaluate the quality of any individual ligand in a biomacromolecular structure. Moreover, thanks to complete validation information for the ligands from the Protein Data Bank [VDB], we can see a summary of the quality status for all instances of one ligand (e.g., all samples of α-D-mannose) in the Protein Data Bank. Therefore, we can compare whether a particular ligand (i.e., all its instances on average) has a higher or lower quality than another. Or we can recognize ligands with a low or high quality.

This is very helpful, but we are still missing the bigger picture – information about the quality of various molecular classes (e.g., drug molecules, organometals, sugars). Namely, how strongly do the classes of molecules differ in quality? Which classes of molecules are more problematic from the quality point of view and therefore will require more of our attention? And in contrast, which classes of molecules have good quality and we can trust their structures? Additionally, are some types of validation errors common for some classes of molecules? We focused on these questions in our work. Specifically, we selected several important classes of molecules and evaluated their quality.

3.1.2 Data set preparation and methodology

In our analysis, we focused on six classes of molecules, which are important from the scientific point of view and which are in parallel markedly chemically different. Specifically, we selected a class of experimental drugs and a class of approved drugs, because drugs are one of the most frequent targets of research. Afterwards,

we chose a class of carbohydrates, because they are known as molecules containing a lot of errors. Interestingly, their structural errors were so alarming that the first ligand validation tool was focused specifically on carbohydrates [83]. Next, we selected mannose derivatives – a well-defined subclass of carbohydrates – to see how fluctuating quality is inside the carbohydrate class. Then, we also selected a further class that includes highly biologically important molecules (hormones, metabolites etc.) and which can potentially contain more validation issues – polycyclic molecules. Last but not least, we added the class of organometals – a group of molecules highly interesting from the application point of view. The formal definition of the individual classes is the following:

- Experimental drugs: Described in DrugBank [4] as experimental drugs, i.e., have been shown to bind specific proteins in mammals, bacteria, viruses, fungi, or parasites.
- Approved drugs: Described in DrugBank as approved drugs, i.e., have received approval in at least one country.
- Carbohydrates: Contain the pyran or furan ring. Molecules with P (e.g., ATP) were excluded, as their quality is influenced more by the occurrence of phosphate derivatives than by the sugar part.
- Mannose derivatives: A subclass of carbohydrates, i.e., all carbohydrates derived from mannose.
- Polycyclic molecules: Contain 3 or more conjugated rings. The molecules with metals were excluded, as their quality is influenced more by the presence of the metal than by their polycyclic structure.
- Organometals: Contain a metal atom.

For each class of molecules, we collected all their member ligands from the wwPDB Chemical Compound Dictionary [84]. Afterwards, for each member ligand, we extracted all their instances from the Protein Data Bank, validated them and averaged the validation results for all instances belonging to the same molecular class. In this way, we obtained validation results for all six analyzed molecular classes. In parallel, we also calculated the overall ligand validation results (i.e., the averaged validation results for all ligands from the Protein Data Bank). Finally, we compared the validation results for individual classes and the overall ligand validation results.

3.1.3 Overall ligand validation results

The overall ligand validation results (from August 10th 2014, see Table 3.1) showed that about 9% of ligands are incomplete, of which about 6% are missing at least one atom and 2.6% are missing rings. Chirality problems occur in less than 8% of the remaining validated ligands. The frequency of basic chirality errors is even lower – only 2.4% of molecules exhibit chirality errors at a carbon atom, and 1.4% at a metal atom. Other chirality issues are generally reported more frequently -– i.e., 4.3% of molecules have the wrong High order chirality and 1.1% the wrong

Planar chirality. Therefore, about 83% of validated molecules are complete and have the correct chirality. This statement is slightly more optimistic than previous estimates, which are based on the fit to the electron density and 3D structure of the ligands, and place the expected percentage of erroneous molecules between 20 and 30% [83, 192].

Table 3.1: Summarization of validation results for individual molecular classes and comparison with the overall validation results for all ligands in the Protein Data Bank (from August 10^{th} 2014).

	All ligands	Polycyclic	Carbo-hydrates	Mannose derivates	Organo-metals	Experi-mental drugs	Approved drugs
Number of PDB entries analyzed	102364	3568	8752	1534	5216	15307	958
Number of validated molecules	238153	6804	57302	6341	22600	37450	1934
Number of models used as reference	17674	1370	913	53	331	3399	185
Incomplete	8.9%	6.7%	5.9%	3.5%	18.0%	6.1%	3.2%
Missing only atoms	5.9%	3.1%	4.2%	3.0%	6.0%	5.0%	0.9%
Missing rings	2.6%	3.0%	1.5%	0.1%	10.7%	0.6%	2.0%
Degenerate	0.5%	0.6%	0.2%	0.4%	1.4%	0.5%	0.3%
Wrong chirality	7.9%	5.5%	4.0%	7.6%	16.5%	2.1%	2.8%
Wrong C chirality	2.4%	3.5%	4.0%	7.4%	2.5%	1.7%	2.8%
Wrong Metal chirality	1.4%	0.0%	0.0%	0.0%	14.3%	0.0%	0.0%
Wrong High order chirality	4.3%	1.9%	0.0%	0.2%	0.0%	0.4%	0.0%
Wrong Planar chirality	1.1%	0.0%	0.0%	0.0%	10.5%	0.1%	0.8%
Complete	**91.1%**	**93.3%**	**94.1%**	**96.5%**	**82.1%**	**93.9%**	**96.8%**
Complete + Correct chirality	**83.0%**	**87.6%**	**90.1%**	**88.9%**	**64.3%**	**91.8%**	**93.9%**

Legend: The color code refers to the relative difference between the results of each case study and the PDB-wide average for all ligands. Specifically:
> 2 times better, > 30% better, > 30% worse, > 2 times worse

3.1.4 Quality comparison of individual molecular classes

Validation results for individual molecular classes are summarized in Table 3.1. The validation results for experimental drugs demonstrated, that the quality of their structures is clearly much higher than the statistics for all ligands.

For approved drugs, i.e., drugs already on the market, the situation is even better. About 95% of these molecules are complete and have the correct chirality, a consequence of the fact, that markedly more effort is expended on determining their structure in biomacromolecular complexes.

Compared to the statistics for all ligands, carbohydrate molecules have overall a higher quality (higher percentage of molecules with complete structure and correct chirality). Nonetheless, they exhibit more errors in C chirality, probably because they generally contain more chiral atoms.

Mannose derivatives play an important role in cell-cell recognition, a biological function which relies heavily on chirality. Therefore they must have a characteristic structure (determined by chirality) and are also strongly predisposed to have C chirality errors. We found that the percentage of errors in C chirality is over 3 times higher for mannose derivatives than the PDB-wide evaluation for all ligands.

Polycyclic molecules exhibit similar trends to carbohydrate molecules, since their structure is also ring-based. They also exhibit more errors in C chirality than the average, probably due to their more complicated, carbon-based scaffolds. Interestingly, they contain less C chirality errors than carbohydrates, confirming carbohydrates to be highly problematic from the quality point of view.

Organometals seem to have a low quality in general, showing that many challenges remain in the field of organometal structure determination.

3.1.5 Conclusions

The analysis of the quality of all ligands in the Protein Data Bank showed that 8% of ligands are incomplete and a further 9% of ligands have chirality errors. Therefore, more than 80% of ligand structures are correct. Experimental and approved drugs were found to be molecular classes with a markedly higher quality than average ligands. As expected, carbohydrates proved to be problematic from the quality point of view, since they have a markedly higher percentage of C chirality errors. This problem is further accented for mannose derivatives. Polycyclic ligands gave similar results to carbohydrates, only with slightly lower percentage of C chirality. The most problematic ligands are organometals, exhibiting outstanding validation problems in most of the validation criteria.

3.2 Charges in small molecules: Prediction of pK_a PQ, PE, PS

3.2.1 Introduction

The acid dissociation constant, K_a, and its negative logarithm pK_a, are important molecular properties and their values are of interest in chemical, biological, environmental and pharmaceutical research [58–60]. pK_a values have found applications in many areas, such as the evaluation and optimization of candidate drug molecules [193–195], ADME profiling [196, 197], pharmacokinetics [58], understanding protein-ligand interactions [59, 198], etc. Moreover, the key physicochemical properties such as lipophilicity, solubility, and permeability are all pK_a dependent. For these reasons, pK_a values are important for virtual screening. Furthermore, pK_a is often used as a descriptor for QSAR models. Unfortunately, experimental pK_a values are usually unavailable even for compounds from the chemical catalogues. In addition to that, obtaining experimental pK_a values for newly designed molecules (i.e., molecules existing only as a sketch on paper) is a long-term project, because first the synthetic pathway must be discovered. Therefore, both the research community and pharmaceutical companies are interested in the development of reliable and fast methods for pK_a prediction.

Several approaches for pK_a prediction have been developed [198–201], specifically LFER (Linear Free Energy Relationships) methods [202,203], database methods, decision tree methods [204], *ab initio* quantum mechanical calculations [205, 206], QSPR (Quantitative Structure-Property Relationship) modelling [56,137,207]

or ANN (artificial neural networks) methods [208]. However, pK_a value prediction remains a challenge for the research community.

An application of partial atomic charges for calculating the relative acidity or reactivity of organic compounds is a known concept in organic chemistry. The reason for this is that the partial atomic charges concept enables the prediction of relative acidity or reactivity by estimating the extent of charge delocalization based on molecular structure information. Therefore, the correlation between pK_a and relevant atomic charges calculated by different ab initio or semiempirical approaches has been analyzed. For example, Gross et al. [54] calculated QM charges via different population analyses and studied their correlation with pK_a for substituted phenols and anilines. Similarly, Kreye et al. [55] compared the correlations of three different QM charge types (calculated via various levels of theory) for substituted phenols. Dixon et al. calculated pK_a from σ and π partial charges [56], Citra [57] used partial charges and bond order, Xing et al. [209] charges and polarizabilities, Soriano et al. [210] charges and frontier orbital energy and Yangjeh [211] combined charges, polarizability, molecular weight, hydrogen-bond accepting capability and partial-charge weighted topological electronic descriptors. The above-mentioned studies demonstrated that charges are very powerful descriptors for pK_a modelling and they demonstrate a linear dependency on pK_a. This indicates that partial atomic charges are very promising descriptors for QSPR models focused on pK_a prediction.

In our work, we continue to research pK_a prediction methods. Specifically, we focused on charge-based QSPR models for pK_a prediction. Our first publication in this field analysed the influence of QM theory level, basis set and population analysis on the accuracy of QSPR models employing QM charges [PQ]. The second publication answered the question of whether accurate but computationally demanding QM charges can be replaced with markedly rapidly obtainable empirical charges [PE]. Specifically, we used the empirical method EEM (Electronegativity Equalization Method) – an approach mimicking QM charges, for which it was parameterized. EEM is very fast and accurate and moreover, we have published several EEM parameter sets [EO, EP, EL] and a few software tools for EEM calculation [EPM, EAM, EL]. Our third pK_a related publication filled in the final piece of the puzzle. Namely, it discusses how the 3D structure generation methodology influences the accuracy of charge-based QSPR models for pK_a prediction [PS]. All three of our articles together therefore describe a way to quickly and precisely predict pK_a for newly designed molecules.

3.2.2 Quantitative Structure-Property Relationship modeling

QSPR (Quantitative Structure-Property Relationship) models [61] calculate a molecular property as a function of descriptors – values, which are computed directly from the molecular structure. The relationship between a property and its descriptors is predominantly linear and has the following form:

$$property = param_1 \cdot descr_1 + param_2 \cdot descr_2 + \cdots + param_n \cdot descr_n + param_{n+1} \tag{3.1}$$

where $descr_1$, $descr_2$, ..., $descr_n$ are the descriptors mentioned above; $param_1$, $param_2$, ..., $param_{n+1}$ are parameters of the QSPR model and n is the number of descriptors in the model. The parameters are calculated based on experimental values of the property and multiple linear regression (MLR) is the most common parameterization method.

The quality of QSPR models, i.e. the correlation between the experimental property and the property calculated by the model, is evaluated using so-called quality criteria of the QSPR model [61]. The most frequently used quality criteria are the squared Pearson correlation coefficient (R^2), root mean square error (RMSE), average absolute pK_a error ($\bar{\Delta}$), standard deviation of the estimation (s) and Fisher's statistics of the regression (F).

3.2.3 Prediction of pK_a using QM charges PO

Introduction. QM charges are the most accurate partial atomic charges obtainable. Therefore they are the best choice for understanding the correlation between charge descriptors and pK_a. For this reason, we used them to design our first charge-based QSPR models and to recognize which charge descriptors we should use in these models. QM charges can be calculated via several theory levels, basis sets and charge calculation schemes. Different QM charge calculation approaches (i.e. combinations of QM theory level, basis set and charge calculation scheme) are appropriate for different applications. Therefore we also evaluated, which QM charge calculation approaches are suitable for pK_a prediction.

Methods. For our research, we used a dataset containing 124 phenol molecules. The phenols were selected, because they are frequently used to evaluate QSPR models. Their 3D structures, which are necessary for QM charge calculation, were obtained from the DTP NCI database [212]. Experimental pK_a values, required for the parameterization of QSPR models, were taken from the Physprop database [213].

We analyzed QM charges calculated via five theory levels. The first two were the Hartree–Fock (HF) method and second-order Møller–Plesset (MP2) perturbation theory, which includes more sophisticated approximations of the Hamiltonian than HF. The other three were density functional theory methods with BLYP, BP86 and B3LYP functionals. BLYP is a gradient-corrected functional and is denoted according to its authors (Becke, Lee, Yang and Parr). BP86 (Becke Perdew 1986) is similar to BLYP, but uses an older correlation functional (Perdew86). B3LYP (Becke, three-parameter, Lee-Yang-Parr) is a hybrid functional constructed as a linear combination of the HF and BLYP functionals. In parallel, we used three basis sets (the simple basis set STO-3G and the more advanced basis sets 6-31G* and 6-311G) and five charge calculation schemes – natural population analysis (NPA), Mulliken charges (MPA), Löwdin charges, Hirshfeld charges, and Merz-Singh-Kollman charges fitted to the electrostatic potential (MK). Therefore, we evaluated 75 (5*3*5) different QM charge calculation approaches.

In parallel, we tested various charge descriptors expressing charges close to the phenolic OH group (which includes the dissociating hydrogen atom). Specifically, we tested these descriptors: The atomic charge of the hydrogen atom from the phenolic OH group (q_H), the charge on the oxygen atom (q_O), the charge on the

Table 3.2: Squared Pearson coefficients (R^2) between calculated and experimental pK_a .

R^2 for the basis set 6-31G*

Theory level	Charge calculation scheme				
	MK	Hir.	Löw.	MPA	NPA
BLYP	0.813	0.886	0.959	0.953	0.959
BP86	0.813	0.89	0.959	0.954	0.959
B3LYP	0.808	0.897	0.963	0.959	0.961
HF	0.788	0.908	0.966	0.966	0.963
MP2	0.788	0.912	0.966	0.966	0.964

R^2 for the basis set 6-311G

Theory level	Charge calculation scheme				
	MK	Hir.	Löw.	MPA	NPA
BLYP	0.826	0.868	0.926	0.932	0.96
BP86	0.825	0.874	0.931	0.939	0.959
B3LYP	0.822	0.882	0.937	0.938	0.962
HF	0.811	0.907	0.95	0.945	0.961
MP2	0.812	0.91	0.951	0.945	0.961

R^2 for the basis set STO-3G

Theory level	Charge calculation scheme				
	MK	Hir.	Löw.	MPA	NPA
BLYP	0.786	0.886	0.879	0.896	0.877
BP86	0.787	0.882	0.878	0.896	0.876
B3LYP	0.817	0.902	0.895	0.904	0.894
HF	0.867	0.928	0.92	0.92	0.92
MP2	0.869	0.929	0.921	0.922	0.921

Legend

	R^2	*RMSE*	$\overline{\Delta}$
excellent	0.95 – 0.97	0.4 – 0.5	0.32 – 0.38
very good	0.92 – 0.95	0.5 – 0.63	0.38 – 0.51
good	0.9 – 0.92	0.63 – 0.7	0.51 – 0.54
acceptable	0.85 – 0.9	0.7 – 0.8	0.54 – 0.64
weak	0.8 – 0.85	0.8 – 0.97	0.71 – 0.73
very weak	< 0.8	> 0.97	> 0.73

carbon atom binding the OH group (q_{C1}) and charges on all the other carbons from the benzene ring.

Results. The best correlation with experimental pK_a values were given by the descriptors q_H, q_O and q_{C1}, the other descriptors only gave a weak correlation. Therefore, we utilized the following pK_a predicting QSPR model:

$$pK_a = param_H \cdot q_H + param_O \cdot q_O + param_{C1} \cdot q_{C1} + constant \qquad (3.2)$$

where $param_H$, $param_O$, $param_{C1}$ and $constant$ are the parameters of the QSPR model. For each evaluated QM charge calculation approach, we parameterized an individual QSPR model (therefore we had 75 QM QSPR models). The parameterization of the QSPR models was performed via multiple linear regression (MLR). Afterwards, we utilized these QSPR models for pK_a prediction, compared the predicted and the experimental pK_a values and calculated quality criteria (R^2, RMSE, $\bar{\Delta}$) for all of them. The values of the most common quality criterion (R^2) are summarized in Table 3.2. Based on these results, we evaluated which QM theory level, basis set and charge calculation scheme are applicable for pK_a prediction.

In general, the results showed us that QM QSPR models provide a successful approach for pK_a prediction. Specifically, more than 25% of the models had excellent quality (R^2 >0.95) and more than half exhibited very good quality (R^2 >0.9). All five examined theory levels are applicable for pK_a prediction. The best QSPR models are provided by MP2 and HF and their accuracy is comparable. The most appropriate basis set is 6-31G*, the results for the 6-311G basis set are slightly weaker. Surprisingly, the charges calculated via the simple basis set STO-3G also provide acceptable QSPR models. The selection of the charge calculation scheme had the strongest influence on the accuracy of the QM QSPR model. Mulliken, Natural and Löwdin population analyses with all levels of theory and basis sets

provide charges that are appropriate for pK_a prediction. Hirshfeld population analysis is also usable, when a proper theory level is selected. In the contrast, MK charges are not applicable for pK_a prediction via QSPR models, because all the models based on these charges gave a weak quality.

3.2.4 Prediction of pK_a using EEM charges [PE]

Introduction. Although QM charges provide accurate pK_a predicting QSPR models, these charges have one big limitation – their calculation is very time-consuming. Therefore, QM QSPR models cannot be used for pK_a prediction with a large set of molecules. At the same time molecules with molecules with a large number of atoms cannot be treated with QM QSPR models, or at least the prediction is highly time-consuming. For these reasons, QM QSPR models cannot be used in virtual screening or drug design, and also their applicability in chemionformatics is limited. A markedly faster alternative to QM charges is the use of EEM charges (i.e., charges calculated by the Electronegativity Equalization Method). EEM charges have a comparable accuracy to QM charges, for which they were parameterized. These facts motivated us to focus on EEM QSPR models. Specifically, our goal was to design successful EEM QSPR models and to evaluate the influence of EEM parameter set selection on their accuracy.

Methods. The analyses of EEM QSPR models were performed on a data set containing 74 phenol molecules. The robustness and the applicability of the models was afterwards illustrated using a data set with 80 carboxylic acids. The 3D structures of the molecules were taken from DTP NCI and their pK_a values originated from Physprop. For our research, we used all EEM parameters published to date. Specifically, we found 18 different EEM parameter sets, published in 8 different articles [EO] [48,50–53,179,180] and reflecting 8 different QM charge calculation approaches. Two of them contain charge calculation schemes which were not evaluated in [PQ] – namely AIM (atoms in molecules) [161,162], and CHELPG [166]. Basic information about the EEM parameters and corresponding QM parameters used are in Table 3.3 and more details can be found in the article [PE]. We needed to compare the EEM QSPR models with the corresponding QM QSPR models. Therefore we calculated not only EEM charges, but also the corresponding QM charges and we prepared both EEM QSPR models and the corresponding QM QSPR models.

Even though EEM charges are able to mimic the QM charges for which they were parameterized, it is clear that they cannot fully maintain the accuracy of QM charges. Therefore in our work we also focused on improving the QSPR models to compensate for this effect. Specifically, we took inspiration from the article of Dixon et al. [56] and introduced two further pK_a correlating descriptors – charge on the phenoxide O^- from the dissociated molecule (q_{OD}), and the charge on the carbon atom binding this oxygen (q_{C1D}). For this reason, we evaluated not only the three descriptor QSPR (3d QSPR) models mentioned in [PQ] (see equation 3.2 in page 35), but also the extended five descriptor QSPR (5d QSPR) models. This means that we worked with 36 (2*18) EEM QSPR models and 16 (2*8) QM QSPR models. All the QSPR models were parameterized by MLR and afterwards we utilized them for pK_a prediction, compared the predicted and the

experimental pK_a and calculated their quality criteria (R^2, RMSE, $\bar{\Delta}$). Values of R^2 are summarized in Table 3.3.

Table 3.3: Basic information about EEM parameters (first three columns) and R^2 for correlation of experimental pK_a and pK_a predicted via EEM QSPR model based on these EEM parameters.

QM theory level + basis set	Charge calc. scheme	EEM parameter set name	R^2 of QSPR model			
			3d EEM	5d EEM	3d QM	5d QM
HF/STO-3G	MPA	Svob2007_cbeg2	0.8671	0.9179	0.9501	0.9646
		Svob2007_cmet2	0.8663	0.9189		
		Svob2007_chal2	0.8737	0.9203		
		Svob2007_hm2	0.8671	0.9179		
		Baek1991	0.9099	0.9195		
		Mort1986	0.8860	0.9142		
HF/6-31G*	MK	Jir2008_hf	0.8696	0.9154	0.8394	0.8864
B3LYP/6-31G*	MPA	Chavez2006	0.8910	0.9192	0.9670	0.9723
		Bult2002_mul	0.8876	0.9158		
	NPA	Ouy2009	0.8731	0.9094	0.9588	0.9679
		Ouy2009_elem	0.8727	0.9132		
		Ouy2009_elemF	0.8848	0.8866		
		Bult2002_npa	0.9044	0.9180		
	Hir.	Bult2002_hir	0.8415	0.9050	0.9122	0.9477
	MK	Jir2008_mk	0.8696	0.9148	0.8447	0.8960
		Bult2002_mk	0.8639	0.9131		
	Chel.	Bult2002_che	0.8695	0.9057	0.8528	0.9087
	AIM	Bult2004_aim	0.8646	0.9017	0.9609	0.9677

Legend

	R^2	*RMSE*	$\overline{\Delta}$
excellent	0.95 – 0.97	0.4 – 0.5	0.32 – 0.38
very good	0.92 – 0.95	0.5 – 0.63	0.38 – 0.51
good	0.9 – 0.92	0.63 – 0.7	0.51 – 0.54
acceptable	0.85 – 0.9	0.7 – 0.8	0.54 – 0.64
weak	0.8 – 0.85	0.8 – 0.97	0.71 – 0.73

Results. As we expected, QM QSPR models exhibited similar behavior to what we had seen in [PQ]. Specifically, the QM QSPR models based on Mulliken or Natural population analyses were very accurate, models based on Hirshfeld population analysis were acceptable and models employing MK charges again proved to be weak. The CHELPG charges provide comparable results to MK, because they are based on a similar approach. A new finding was that QM QSPR models based on AIM charges are also very accurate. The introduction of new descriptors brought about an improvement to QM QSPR models, especially when the original 3d QM QSPR models were weaker.

EEM QSPR models performed worse than QM QSPR models, but they still gave an acceptable accuracy. In particular 5d EEM QSPR models demonstrated very good quality criteria. An interesting and pleasant fact is that EEM QSPR models are not as sensitive to the choice of EEM parameter set.

When we prepared similar QSPR models for carboxylic acids, these models gave comparable trends and accuracy, and therefore illustrated the robustness and transferability of this pK_a prediction approach.

3.2.5 3D structure sources for pK_a prediction using charges [PS]

Introduction. A necessary input for pK_a prediction using charge-based QSPR models is the 3D structure of the molecule. The experimental 3D structures were only measured for a limited set of molecules. On the other hand, when we design a molecule (or a set of molecules for virtual screening), obtaining its (their) experimental structure is a long-term project, because we first need to synthetize the molecule or find some other source of it. The markedly faster and in reality the only applicable way, to obtain the 3D structure of these molecules is to generate the structure automatically. There are a few methodologies for preparing 3D structures and they are implemented in several software tools. Specifically, the 3D structure can be created via a data- and knowledge-based approach (used in CORINA [62], OpenBabel [182] and Omega [63]), distance geometry approach (Balloon [181], RDKit [214]) or other approaches (Frog2 [215]); e.g, Frog2 first generates a graph of rings and acyclic elements, and afterwards performs a Monte Carlo search. Some of the software tools were used to prepare the 3D structures stored in well-known databases. For example, CORINA was employed in the preparation of DTP NCI database [212] and Omega was used to prepare Pubchem database [1]. An important question is which methodology and software tool can be used to obtain proper 3D structures. This means 3D structures applicable for accurate pK_a prediction via our QM and EEM QSPR models. A primary goal of our work was to answer this question.

Methods. We performed our analyses on three data sets: 60 phenols, 82 carboxylic acids and 48 anilines. Additionally, we tested our results on an independent test set containing 53 phenol molecules. For all of these molecules, we obtained experimental pK_a values from the Physprop database. Afterwards, for each molecule, we obtained its 3D structure from 6 different sources: From the DTP NCI and Pubchem databases and via the software tools Balloon, Frog2, OpenBabel and RDKit. Furthermore, for each of these 3D structures we performed three types of optimization: none, QM (B3LYP/6-31G*) and MM (MMFF94). Additionally, we also performed an optimization via the MM force field UFF (Universal Force Field) for structures prepared with RDkit, because it was recommended by RDKit's developers. For each 3D structure, we calculated QM charges via 4 QM charge calculation approaches (i.e., HF/STO-3G/MPA, B3LYP/6-31G*/MPA, B3LYP/6-31G*/NPA, and B3LYP/6-31G*/AIM) and EEM charges via corresponding parameters. These charge calculation approaches were selected, because they provided high-quality QSPR models [PQ, PE]. Afterwards, we prepared QM and

Table 3.4: R^2 describing correlation between calculated and experimental pK_a for QM QSPR models.

R^2		Class of molecules	Phenols				Carboxylic acids				Anilines				Average
		Charge calculation approach	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	
Source + Optimization	Balloon	none	0.896	0.939	0.908	0.904	0.823	0.720	0.819	0.846	0.836	0.903	0.912	0.805	0.859
		MM	0.917	0.881	0.933	0.891	0.867	0.587	0.805	0.843	0.874	0.953	0.927	0.921	0.867
		QM	0.915	0.871	0.901	0.856	0.890	0.618	0.824	0.807	0.948	0.967	0.933	0.921	0.871
	Frog2	none	0.894	0.912	0.906	0.891	0.896	0.876	0.876	0.884	0.934	0.911	0.924	0.916	0.902
		MM	0.967	0.931	0.907	0.938	0.907	0.830	0.903	0.922	0.958	0.973	0.965	0.926	0.927
		QM	0.969	0.963	0.953	0.939	0.917	0.853	0.906	0.917	0.875	0.973	0.911	0.853	0.919
	NCI	none	0.947	0.971	0.960	0.973	0.931	0.891	0.911	0.910	0.951	0.970	0.966	0.903	0.940
		MM	0.958	0.963	0.959	0.936	0.938	0.889	0.929	0.922	0.954	0.955	0.967	0.914	0.940
		QM	0.891	0.935	0.861	0.902	0.925	0.854	0.903	0.921	0.942	0.959	0.937	0.892	0.910
	OpenBabel	none	0.955	0.961	0.957	0.963	0.869	0.658	0.845	0.876	0.952	0.973	0.966	0.930	0.909
		MM	0.961	0.965	0.959	0.961	0.863	0.665	0.841	0.875	0.958	0.975	0.967	0.927	0.910
		QM	0.955	0.957	0.956	0.936	0.845	0.674	0.804	0.827	0.874	0.974	0.928	0.880	0.884
	PubChem	none	0.960	0.950	0.935	0.900	0.909	0.873	0.891	0.907	0.938	0.939	0.921	0.937	0.922
		MM	0.963	0.911	0.927	0.864	0.916	0.885	0.892	0.916	0.942	0.979	0.966	0.916	0.923
		QM	0.943	0.936	0.922	0.886	0.901	0.871	0.896	0.908	0.934	0.974	0.885	0.828	0.907
	RDKit	none	0.782	0.895	0.796	0.882	0.780	0.723	0.804	0.817	0.853	0.816	0.851	0.796	0.816
		MM-UFF	0.947	0.961	0.941	0.934	0.894	0.821	0.842	0.860	0.965	0.979	0.973	0.980	0.925
		MM	0.931	0.909	0.934	0.950	0.902	0.750	0.797	0.862	0.959	0.976	0.967	0.927	0.905
		QM	0.935	0.944	0.933	0.922	0.861	0.696	0.814	0.855	0.940	0.964	0.927	0.908	0.892
Average			0.931	0.934	0.924	0.917	0.886	0.776	0.858	0.878	0.926	0.953	0.936	0.899	

Legend	$R^2 \geq 0.95$	$R^2 \geq 0.9$	$R^2 \geq 0.866$	$R^2 \geq 0.833$	$R^2 \geq 0.8$	$R^2 \geq 0.7$	$R^2 < 0.7$

Table 3.5: R^2 describing correlation between calculated and experimental pK_a for EEM QSPR models.

R^2		Class of molecules	Phenols				Carboxylic acids				Anilines				Average
		Charge calculation approach	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	
Source + Optimization	Balloon	none	0.873	0.904	0.903	0.888	0.832	0.924	0.888	0.853	0.806	0.847	0.826	0.870	0.868
		MM	0.852	0.906	0.907	0.885	0.800	0.917	0.883	0.837	0.867	0.845	0.855	0.880	0.870
		QM	0.869	0.908	0.906	0.890	0.772	0.917	0.889	0.851	0.953	0.930	0.908	0.945	0.895
	Frog2	none	0.907	0.897	0.898	0.858	0.832	0.875	0.831	0.870	0.894	0.879	0.904	0.887	0.878
		MM	0.918	0.906	0.917	0.868	0.859	0.888	0.860	0.848	0.863	0.857	0.852	0.902	0.878
		QM	0.921	0.907	0.918	0.869	0.841	0.898	0.866	0.874	0.939	0.926	0.907	0.939	0.900
	NCI	none	0.906	0.906	0.899	0.890	0.875	0.926	0.891	0.879	0.870	0.852	0.839	0.882	0.884
		MM	0.891	0.926	0.926	0.916	0.860	0.920	0.888	0.829	0.844	0.834	0.848	0.889	0.881
		QM	0.896	0.924	0.925	0.912	0.821	0.923	0.884	0.834	0.921	0.884	0.869	0.920	0.893
	OpenBabel	none	0.900	0.920	0.912	0.908	0.830	0.898	0.848	0.826	0.860	0.849	0.851	0.899	0.875
		MM	0.900	0.919	0.911	0.907	0.827	0.903	0.849	0.835	0.858	0.851	0.857	0.897	0.876
		QM	0.896	0.917	0.911	0.904	0.807	0.911	0.856	0.851	0.946	0.935	0.939	0.934	0.901
	PubChem	none	0.896	0.918	0.913	0.902	0.888	0.891	0.866	0.873	0.874	0.881	0.874	0.907	0.890
		MM	0.887	0.917	0.915	0.899	0.874	0.902	0.876	0.871	0.886	0.852	0.872	0.900	0.888
		QM	0.898	0.921	0.925	0.899	0.825	0.923	0.894	0.892	0.890	0.905	0.867	0.927	0.897
	RDKit	none	0.894	0.907	0.904	0.885	0.836	0.932	0.889	0.874	0.832	0.842	0.840	0.857	0.874
		MM-UFF	0.923	0.917	0.912	0.895	0.801	0.919	0.866	0.844	0.838	0.845	0.843	0.875	0.873
		MM	0.899	0.908	0.902	0.892	0.823	0.907	0.871	0.852	0.846	0.852	0.854	0.897	0.875
		QM	0.909	0.919	0.916	0.895	0.753	0.915	0.881	0.851	0.933	0.892	0.869	0.923	0.888
Average			0.897	0.913	0.911	0.893	0.829	0.910	0.872	0.855	0.880	0.871	0.867	0.902	

Legend	$R^2 \geq 0.95$	$R^2 \geq 0.9$	$R^2 \geq 0.866$	$R^2 \geq 0.833$	$R^2 \geq 0.8$	$R^2 \geq 0.7$

EEM QSPR models for each combination of molecular type, 3D structure source, optimization, and charge calculation approach. Specifically, we used the extended QSPR models [PE], i.e., QSPR models also including descriptors from dissociated forms of molecules (phenols, carboxylic acids) or associated forms of molecules (anilines). Finally, we predicted pK_a via these QSPR models, compared the results with experimental pK_a values, calculated the quality criteria for the models (R^2, RMSE, $\bar{\Delta}$) and evaluated the models. R^2 values are summarized in Tables 3.4 and 3.5.

Results. QM QSPR models proved to be highly accurate for all four types of QM charge calculation approach, as in [PQ, PE]. EEM QSPR models performed less well than QM QSPR models, but provided acceptable results, as in [PE].

Interestingly, the type of molecule also influences the accuracy of QSPR models. Specifically, these models are weaker (but still acceptable) for carboxylic acids. A reason for this could be that the carboxyl group bound some arbitrary chemical scaffold. In contrast, the –OH group of phenols and $–NH_2$ group of anilines have the same, conserved neighborhood – the phenolic ring, which additionally allows a higher de-localization of electrons.

A major message of our work is, that an appropriate selection of 3D structure source and optimization method is essential for the QSPR modeling of pK_a. The 3D structures from the DTP NCI and Pubchem databases, i.e. structures generated by CORINA and Omega, respectively, exhibited the best performance. These 3D structures provided very accurate QSPR models for all the tested molecular classes and charge calculation approaches, and they do not require optimization. Frog2 also performed very well for all of the tested molecular classes and charge calculation approaches. Other 3D structure sources can be used, but they are not so robust, and an unfortunate combination of molecular class and charge calculation approach can lead to weak QSPR models. Additionally, these structures generally need to be optimized in order to produce high-quality QSPR models. Specifically, the best approach is to apply MM optimization to 3D structures used with QM QSPR models, and QM optimization to 3D structures used with EEM QSPR models.

3.2.6 Conclusions

We showed that QM charges are very successful descriptors for the prediction of pK_a via QSPR models. The accuracy of the prediction is strongly influenced by the selection of the charge calculation scheme. Proper charge calculation schemes were MPA, NPA and AIM.

We later demonstrated that empirical charges calculated via EEM are also applicable for pK_a prediction. The EEM charges should be parameterized based on a proper QM charge calculation scheme. Because EEM charges are less precise than QM charges, EEM QSPR models have to contain more descriptors.

We then found that the pK_a predicting QSPR models are sensitive to the source of molecular 3D structure. The most appropriate 3D structure sources are CORINA and Omega and moreover, the 3D structures generated by these software tools do not require further optimization.

Therefore, a workflow for the rapid and accurate prediction of pK_a can be as follows: The preparation of 3D structures by CORINA or Omega with no further optimization, the calculation of EEM charges for these structures and then the EEM QSPR calculation of pK_a. Such a workflow can be used directly within the process of *in silico* drug design or incorporated into other chemoinformatics applications.

3.3 Charges in proteins: BAX Activation [BAX]

3.3.1 Introduction

Apoptosis is a programmed cell death and is fundamental for development, growth and homeostasis in multi-cellular organisms. Malfunctions in the apoptosis machinery are known to be involved in cancer, autoimmune diseases, neurodegenerative disorders etc. An important process within apoptosis is mitochondrial outer membrane permeabilization. This permeabilization allows the release of apoptotic proteins from the mitochondrial inter-membrane space, which causes the activation of cell death proteases. Afterwards, the proteases cleave the cell's cytoskeleton and genetic material. This permeabilization is executed by the Bcl-2 family proteins Bak and Bax, which are activated during apoptosis and then oligomerize and form pores in the mitochondrial membrane [216–219]. Bak and Bax oligomerisation is controlled by a set of further Bcl-2 proteins [220–223]. The activation of Bak and Bax is performed via a subclass of apoptotic Bcl-2 proteins such as Bim and Bid [224–226]. The activation steps required for Bax oligomerization were extensively investigated [227–231]. They include Bax translocation from the cytosol to the mitochondrial membrane, and changes in Bax conformation. Conformational changes of Bax include the exposure of its C-domain, insertion in this C-domain into the membrane, and exposure of the Bax BH3 domain, one of four homology domains of Bcl-2 proteins. In inactive Bax [229], the C-domain is tightly bound inside a hydrophobic pocket, denoted the 'BH groove' (Figure 3.1 A). This tight binding increases the solubility of Bax and keeps Bax in the cytosol, when apoptosis is not required. Gavathiotis et al. [231] synthesized a helix mimicking the BH3 domain of the activator Bim (denoted Bim-SAHB, because it is a Bim-stabilized α-helix of Bcl-2 domains) and they also resolved a structure of Bim-SAHB activated Bax via NMR (Figure 3.1 B). Interestingly, the suggested Bax activation site and the Bax C-domain are separated by over 25 Å. Additionally, the binding of Bim-SAHB to Bax is weak and transient, and neither significant disturbances in the helical packing, nor covalent modifications have been observed in Bax upon activation. Therefore the mechanism by which the binding of Bim-SHAB into the Bax activation site involves a C-domain and causes its exposure still remains unclear [64, 65].

Charge transfer was discovered to be significant in many biomolecular interactions [232–234]. Therefore we investigated the role of charge transfer during Bax activation. For this analysis, we utilized partial atomic charges calculated via the Electronegativity Equalization Method (EEM) [48] and EEM parameters for biomacromolecules [EB].

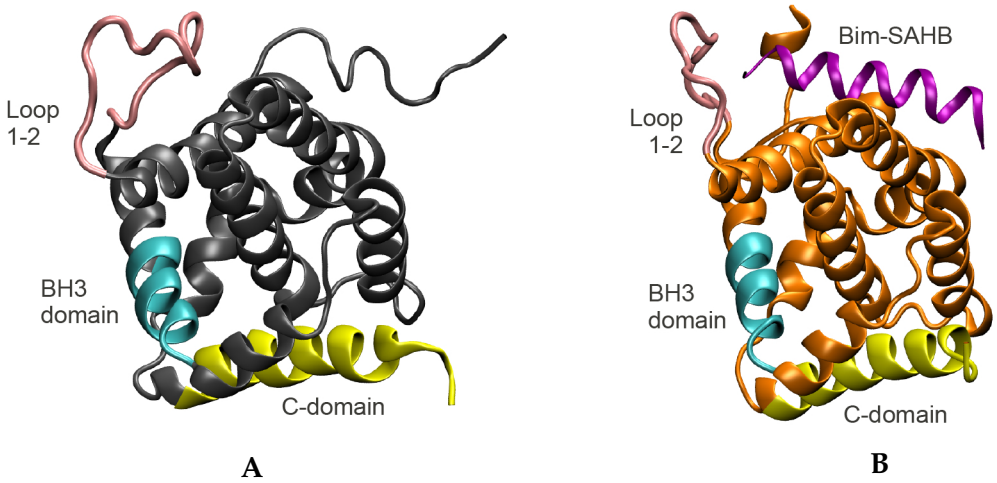


Figure 3.1: A) A structure of inactive Bax (PDB ID: 1F16), with its BH3 domain (cyan), the C-domain (yellow) and loop 1–2 (pink). The rest of the protein is in gray. B) A structure of Bax activated by Bim-SAHB (PDB ID: 2K7W), where Bim-SAHB is shown in purple, other domains have the same coloring as in inactive Bax, and the rest of the protein is in orange.

3.3.2 Analysis of charge transfer within Bax activation

The analysis of the charge transfer was performed on 3D structures obtained from the Protein Data Bank – the inactive Bax structure had the PDB ID 1F16 and active Bax in complex with the activator peptide Bim-SAHB had the PDB ID 2K7W. We computed EEM atomic charges on both structures using an EEM parameter set for biomacromolecules, which was based on the HF/6-31G*/MPA quantum mechanical charge calculation approach [EB]. Afterwards, we calculated the absolute charge transfer per residue and we used this value as a metric for evaluating the overall charge transfer within the Bax. Specifically, we concentrated on residues with a high value of absolute charge transfer. At the same time, we took into account published information about residues which play an important role in Bax activation (e.g., which mutation influences the activation) [229,235]. Based on this information, we analysed the correlation between the high absolute charge transfer value of a residue and its reported influence on Bax activation.

Experimental evidence suggests that in inactive Bax, the C-terminal helix is bound tightly to its hydrophobic pocket (BH-groove). During activation, this binding becomes destabilized. Consequently, the C-domain vacates the BH-groove and inserts into the mitochondrial outer membrane. Mutagenesis studies showed a critical interaction between residues Ser184 and Asp98 at the C-domain-BH-groove interface, whose abrogation can immediately activate Bax [229, 235]. We therefore focused on the changes in charge density distribution in the vicinity of this interaction. Our calculations did not report any change in the charge profile of Ser184. But they showed that Arg94 becomes more positive after activation. This can lead to the recruitment of Asp98, the abrogation of the Asp98-Ser184 interaction, and ultimately the destabilization of the C-domain (see Figure 3.2). This demonstrates that the binding of Bim-SAHB to Bax can activate Bax by destabilizing the interaction between the Bax C-domain and its binding groove.

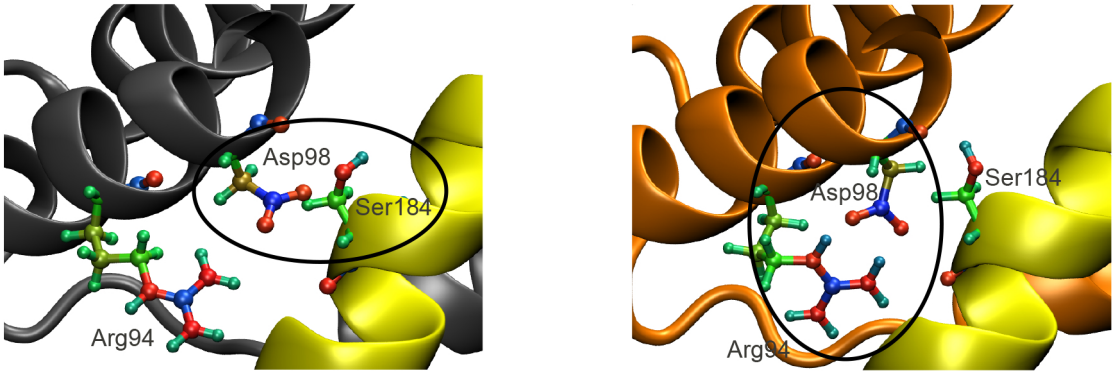


Figure 3.2: A) In inactive Bax, Asp98 interacts with Ser184, which keeps the C-domain in its binding pocket. B) In active Bax, the now more positively charged Arg94 interacts with Asp98, which no longer contributes to the stabilization of the Bax C-domain in its BH-groove.

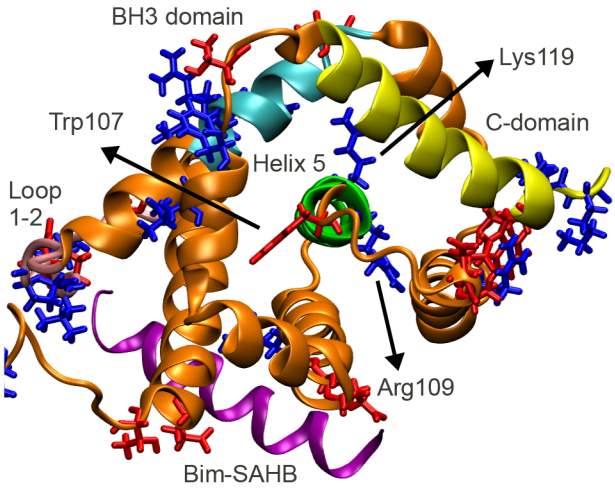


Figure 3.3: Top view of helix 5, showing the organization of residues Trp107, Arg109 and Lys119 inside the hydrophobic core of Bax and also the location of other residues with a high charge transfer. Residues with markedly positive or negative charge transfer upon activation are in blue or red, respectively.

A further uncertainty in Bax activation is the way the BH-groove is influenced by the binding of Bim-SAHB to Bax, because the Bax activation site and the Bax C-domain are separated by over 25 Å. The residues which exhibited a high charge transfer provided a clue to the way the activation information proceeds through the protein. Foremost, significant changes in the net residue charges were found at the Bax activation site, the BH3-domain (required for oligomerization) and the C-domain (required for membrane insertion). Since these are all functional sites of Bax, these changes could expected. At the same time, George et al. [235] found that a triple alanine mutant at residues 63–65 (on the BH3 domain of Bax) prevented Bax oligomerisation and apoptotic activity, which fully correlates with the high charge transfer we found on residues 64 and 65 upon Bax activation. However, in addition to the expected changes, our method surprisingly also identified significant charge transfer on the central helix, inside the hydrophobic core of Bax (residues Trp107, Arg109 and Lys119 on helix 5). The presence of significant charge transfer in a predominantly hydrophobic environment suggests that helix

5 acts as a hub which collects and distributes charge density. This idea is also supported by the spatial organization of residues Trp107, Arg109 and Lys119 inside the hydrophobic core (Figure 3.3), suggesting that the interaction at the Bax activation site is transmitted via a network of charges from the activation site, through the protein core, to the C- and BH3-domains.

3.3.3 Conclusions

We investigated the changes in the Bax charge profile upon activation via a functional peptide of its natural activator protein, Bim. We found that charge reorganizations after activator binding mediate the exposure of the functional sites of Bax (i.e., C-domain and BH3 domain) and consequently activate Bax. The affinity of the Bax C-domain for its binding groove is decreased due to the Arg94-mediated abrogation of the Ser184-Asp98 interaction. We further identified a network for charge transfer, which brings the activation information from the activation site, through the hydrophobic core of Bax, to the distant functional sites of Bax. The network was mediated by a hub of three residues on helix 5 of the hydrophobic core of Bax. Our results suggest that allostery mediated by charge transfer is responsible for the activation of Bax.

3.4 Channels: Enzyme channels anatomy AN

3.4.1 Introduction

Enzymes are proteins which catalyze reactions that change substrates into products. The enzymatic reactions occur in active sites of the enzymes. Based on the type of chemical reaction which enzymes catalyze, they can be categorized into six enzymatic classes, each marked by its Enzyme Commision (EC) number [236]. Therefore, we use the following enzymatic classes: oxidoreductases (EC1), transferases (EC2), hydrolases (EC3), lyases (EC4), isomerases (EC5), and ligases (EC6).

Thanks to the many analyses of enzymatic reactions, we now have a better understanding of how active site chemistry works [237–240] and which amino acids are present in the sites [241]. However, relatively little is known about how the substrates enter the active sites and how the respective products leave them. Some active sites are located on the surface of the protein, in clefts or pockets, other enzymes have active sites deeply buried inside. These buried active sites are connected to the outside by one or more channels. These channels therefore allow the passage of substrates and products to/from the active site [66–77]. It has been shown that mutations in the access channels to enzyme active sites can alter the substrate preferences of enzymes, and may be utilized in the rational design of enzymes [107, 242]. Despite the channels themselves being intensively studied in recent years, there was no in-depth analysis of enzyme channels. For this reason, we focused on performing such an analysis. Our goal was first to recognize, whether the channels are a general property of the majority of enzyme structures and how often they are present. Afterwards, we studied the geometrical properties of the channel, such as their length. Then, we focused on channel's chemical properties and we analyzed their lining residues, bottleneck residues,

residue composition within their parts etc. Last but not least, we concentrated on the channel's physicochemical properties.

3.4.2 Data set preparation and methodology

For performing the channel analyses, we first prepared a dataset of enzyme structures. Specifically, we selected all the enzymes whose active sites were annotated in the Catalytic Site Atlas (CSA) [243] and whose structures were available in the Protein Data Bank [244]. Afterwards, we removed the low quality structures (with a resolution higher than 2.5 Å) and structures sharing at least 90% sequence identity. In this way we obtained a data set containing 4,306 enzymes.

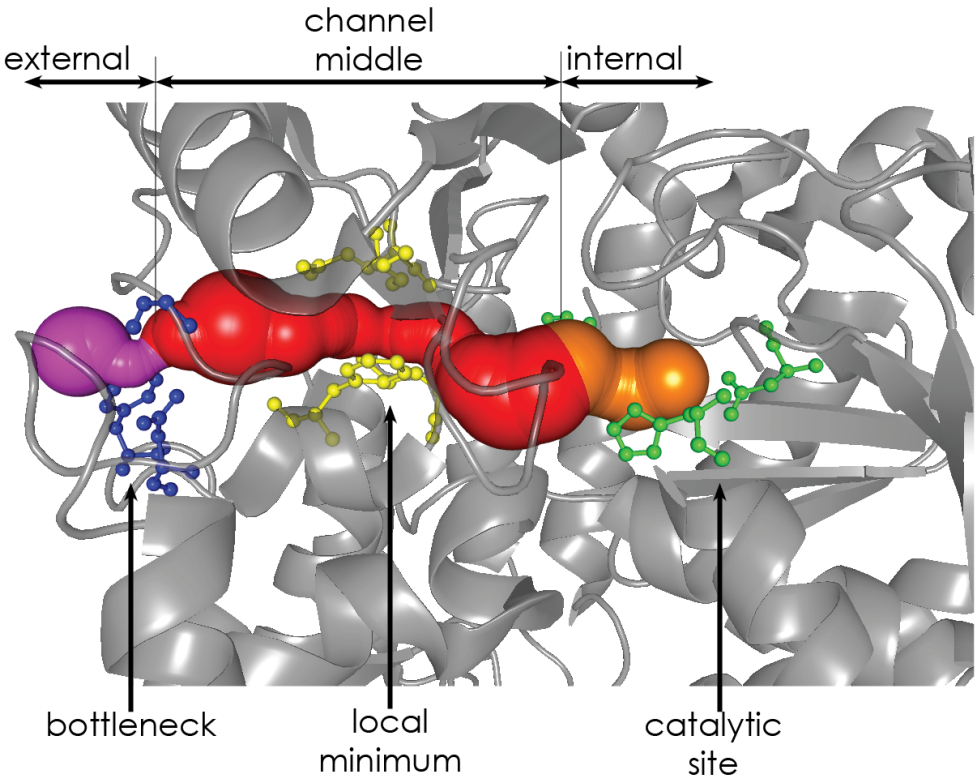


Figure 3.4: Channel parts, depicted on a channel pyridoxal-5'-phosphate-dependent acyl-CoA transferase (PDB ID 3KKI). Active site amino acids (present in the internal part of the channel) are shown in green, amino acids in the middle part forming the wall of a local minimum (channel narrowing) are in yellow, and amino acids in the external part lining the bottleneck are in blue. Internal, middle and external parts are colored orange, red and magenta, respectively.

For all the enzymes from the data set, we calculated channels via MOLE 2.0 software [MO2] and the starting point of the channels was the active site. MOLE 2.0 was also used for calculating the channel's geometrical, chemical and physicochemical properties. For the needs of the analysis, the channel was divided into particular parts – external, middle, and internal (depicted in Figure 3.4). Important areas of the channel are also their local minima (narrowings of the channel) and global minimum (bottleneck of the channel) – see Figure 3.4.

3.4.3 Channel occurrence and geometrical properties

We found that 64% of the enzymes contained channels at least one 15 Å long channel and more than 87% contain channels at least 5 Å long. However, the short channels may correspond to the paths of active sites located in biomacromolecular pockets, therefore in our study we focused purely on channels longer than 15 Å. Channel occurrence varies among the enzymatic classes. The highest percentage (77.8%) of enzymes with channels longer than 15 Å was identified in oxidoreductases, while the lowest percentage (51.8%) applied to hydrolases. The average number of channels in an enzyme was two.

The median channel length was 27.7 Å, 40% of channels were 15–30 Å long and 10% of enzymes contained channels longer than 50 Å. Surprisingly, we found that the number of long channels does not correlate with protein size. The median length of channels in the enzymatic classes was comparable (see Figure 3.5) – oxidoreductases have a median channel length slightly longer (by about 2 Å) than other enzymes, whereas transferases and ligases have a shorter average channel length (also by about 2 Å). An interesting finding was that the channel length only exhibited a low correlation with the enzyme atom count.

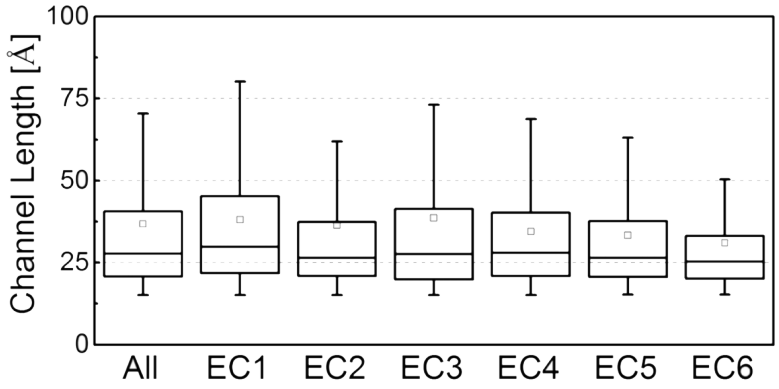


Figure 3.5: Average channel length in individual enzymatic classes in comparison with the overall average channel length.

3.4.4 Channel chemical properties

We calculated the frequencies of amino acids in defined regions of a channel (e.g., lining residues, external, middle and internal part, bottleneck, local minima) and compared this frequency with the frequency of these amino acids in the whole enzyme structure. Specifically, we calculated the ratio of the amino acid frequency in the defined region to its frequency in the whole enzyme. This comparison provided us with direct information about the preference or exclusion of an amino acid (or a group of amino acids) from some regions. The results of this analysis are visualized in Figure 3.6.

The rather bulky and aromatic amino acids (His, Tyr, Trp, Arg), occur over 1.25 times more frequently in the channel lining residues than in the whole enzyme. Additionally, other amino acids (Asn, Phe, Asp, Thr, Met, Ser) also exhibit a slightly higher frequency in the channel lining residues than in the rest of the

protein. On the other hand, nonpolar aliphatic amino acids (Pro, Gly, Ile, Leu, Ala, and Val) are significantly less common in channel lining residues.

The channel bottlenecks reflect the composition of lining residues, but they contain significantly more cysteine (Cys), histidine (His) and tyrosine (Tyr) residues than usual and much fewer small aliphatic amino acids (Pro, Gly and Ala). As histidine (His) and cysteine (Cys) have unique binding properties, it is possible to hypothesize that these binding properties might provide a gate-keeping activity at the channel bottlenecks.

The frequencies of amino acids in active sites, on the protein surface and inside the protein, or in general channels, are markedly different from both the average protein amino acid composition and the composition of the lining residues. The active sites contain significantly more amino acids that could be part of a catalytic cycle (His, Asp, Cys, Glu, Arg, Tyr, Lys) enabling proton and electron shuffling and covalent bond reorganization. Conversely, the frequency of less reactive amino acids (Trp, Thr, Gln, Phe) or amino acids with nonreactive side-chains (Met, Ala, Pro, Ile, Val, Leu) is lower in the active sites. These results are in perfect agreement with data published by Holliday and coworkers [245]. On the other hand, the surface regions contain mainly charged (Lys, Arg, Glu, Asp) and polar residues (Asn, Gln), which facilitate contact with the polar water environment. Also, the surface has a higher than average frequency of prolines (Pro), as these helix-breaker amino acids are common in turns in the protein structures and rigidify the protein fold.

Channel-lining residues are not uniformly distributed along the length of the channel. Internal parts of the channel tend to contain more aromatic residues (His, Trp, Tyr) together with cysteine (Cys).

These trends are similar to the catalytic site propensities. The frequencies of amino acids in the middle regions of the channels correspond to the frequencies in the entire channel with the exception of glycine (Gly) and aromatic amino acids (Trp, Tyr, Phe), which are present more frequently. We may hypothesize that the higher frequency of glycine (Gly) in the middle channel parts is because it facilitates flexibility, which may be important for substrate/product channeling between the active site and protein surface [101] , whereas aromatic amino acids can serve as gate-keepers. External parts of the channel bear more charged residues than any other part (Arg, Lys, Glu, Asp) together with proline (Pro) and glutamine (Gln). The analysis of amino acids leading to active sites, divided according to the six enzymatic groups, shows that amino acid channel propensities correspond to overall channel propensities. However, some differences were identified (see [AN] for details). The main findings were that channels in oxidoreductases have significantly lower frequencies of charged lining amino acids but higher frequencies of aliphatic lining amino acids. On the other hand, channels in ligases contain fewer cysteine, aromatic and aliphatic amino acids and more charged amino acids and glycine.

3.4.5 Channel physicochemical properties

We first focused on channel hydropathy. The average channel hydropathy is -0.92 (the hydropathy of amino acids varies from the -4.5 of Arg to the 4.5 of Ile).

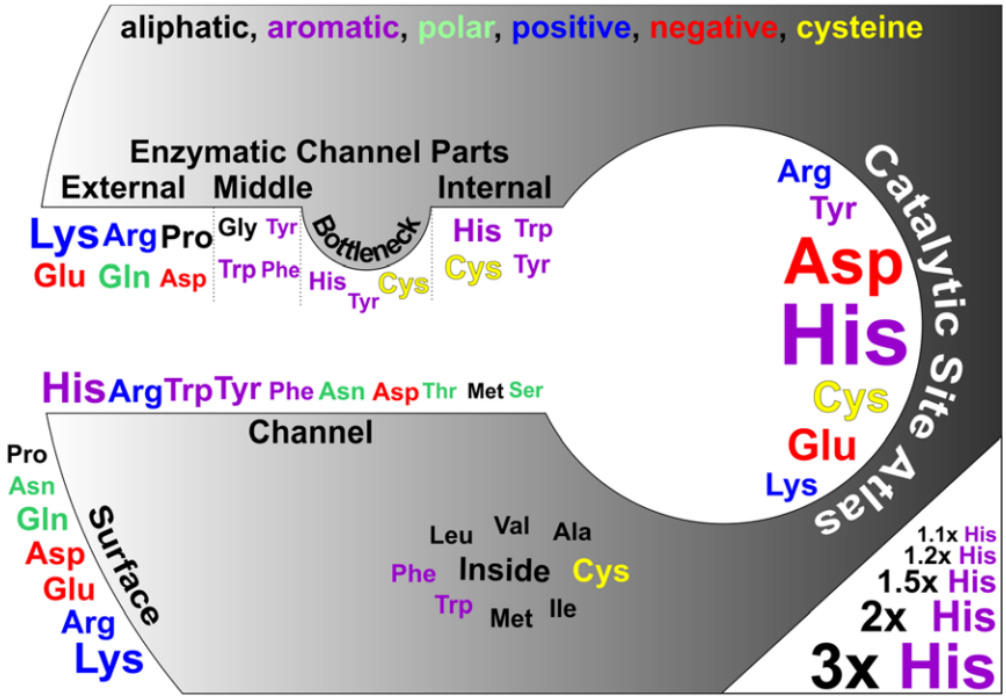


Figure 3.6: Enhancement of amino acid frequency in different parts of the enzyme structure. Amino acids that are found more often than average in different regions of an enzyme's structure.

The distribution plots of hydropathy (Figure 3.7) also indicate that hydrophilic channels are preferred to hydrophobic ones.

We then analyzed channel polarity. We found that the average channel polarity is 16.5 (the polarity varies between 52.0 for highly polar amino acids and 0.0 for nonpolar amino acids). This indicates that the channels are relatively polar.

Taking all this information into account we may conclude that the average channel has a slightly negative hydropathy and higher polarity. However highly hydrophobic and nonpolar, as well as highly hydrophilic and polar, channels were also detected.

We also analyzed the presence of charged amino acid side chains (Asp, Glu, His, Lys and Arg) in channel walls. On average the channel walls are lined with two negative and two positive side chains, resulting in sum neutral channel walls. Despite this, we also identified channels with significant extreme physico-chemical properties (see [AN]).

We also found that enzymatic classes differ in their average physico-chemical properties (Figure 3.7): oxidoreductases (EC1) exhibit the most hydrophobic as well as the least polar channels among the enzyme classes, while ligases (EC6), and to some extent also isomerases (EC5), lyases (EC4) and hydrolases (EC3), exhibit the most hydrophilic as well as the most polar channels. In parallel, we identified that some physico-chemical features differ across the three parts of the channel: internal, middle and external. The polarity of the middle part is always lower than the polarity of both the internal and external parts, respectively. The lower polarity of the middle part of the channel is also reflected in its significantly more hydrophobic behaviour. The charged residues occur mainly in the external parts of enzyme channels, while the internal and middle part contain more aromatic residues.

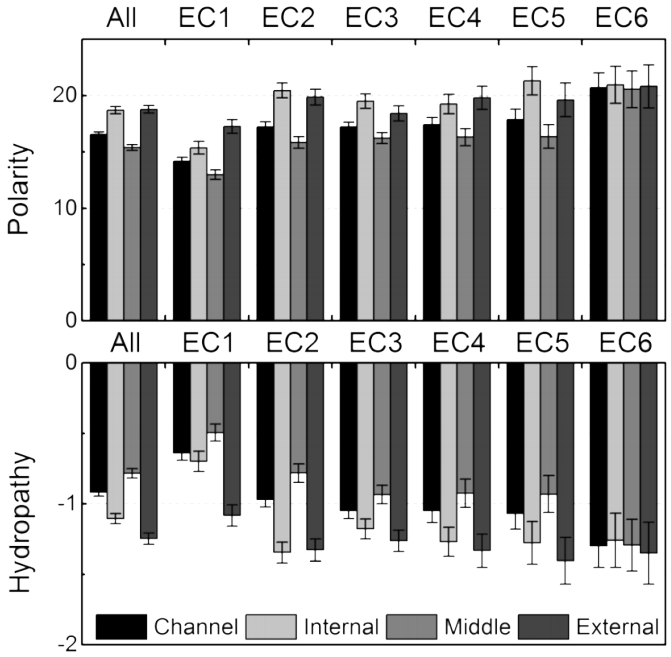


Figure 3.7: Average channel length in individual enzymatic classes in comparison with the overall average channel length.

3.4.6 Conclusions

To summarize, we analyzed channels in 4,306 enzyme structures from the Protein Data Bank, which are annotated in the Catalytic Site Atlas. We identified that at least 64% of these enzymes contain on average two channels longer than 15 Å leading to the catalytic site. Consequently, we can anticipate that these enzymes contain buried active sites. The longest and the most hydrophobic channels were found in oxidoreductases, while the smallest number of channels were detected in hydrolases and the shortest and the most hydrophilic channels in ligases. The composition of channel lining residues differs from the average composition of enzyme structures as well as from the composition of the protein surface. Hydrophobic aliphatic amino acids, which are the most common amino acids found in protein cores, occur in channel walls less frequently, whereas aromatic, charged and polar amino acids occur more frequently in channel walls. All these findings indicate that the active site access channels have a significant biological function as they are involved in co-determining the enzyme's substrate preferences.

4

Future prospects

In the field of biomacromolecular structural fragments analysis, a lot of effort is still invested in the development and improvement of their methodologies. Some analyses merely require an extension of current approaches (e.g., validation, channel detection), others need marked improvements (e.g., accurate calculation of channel physicochemical properties) and many new types of analyses are still under development. A current trend in the performance of structural fragment analyses is the precalculation of their results for available structures and providing these results to the user directly. Another important activity is the integration of key analyses directly into the databases of biomacromolecular structures (e.g., Protein Data Bank). This activity is connected with the need to markedly streamline the analyses and enable their rapid execution on the complete database.

A straightforward application of structural fragment analyses is the prediction of fragment (or biomacromolecule) properties such as acid dissociation constants, activities, partition coefficients etc. Another important utilization is research focused on common aspects or features of selected structural fragments – for example, the typical charge distribution of cytochrome channels, standard amino acid surroundings of a fucose-binding site, etc. This knowledge provides us with a clue to understanding their chemical interactions and some insight into their biological role. Furthermore, based on this information, we can predict the occurrence of individual structural fragments. A highly interesting and at the same time a highly challenging application is the study of the mechanisms and effects connected with certain chemical actions (e.g., the activation of a particular biomacromolecule, the binding of a particular ligand). Last but not least, there are even greater possibilities associated with the analysis of biomacromolecular structural fragments — its application may enable us to predict the influences of individual structural changes within fragments (e.g., the influence of a certain point mutation, a ligand modification or an atom substitution).

5

Summary

Biomacromolecules (e.g, proteins, nucleic acids, polysaccharides) are essential biological entities, since they are responsible for building cell components and ensuring their functionality. The biomacromolecule is a large object containing several thousand atoms. Different parts (fragments) of them play diverse roles – e.g., they create an active site, bind a certain ligand or metal, form a channel or a pore, or are responsible for the proper shape of the molecule. The research of these fragments (especially biologically important fragments) can provide very useful results such as discovering drug design patterns, information for the classification of biomacromolecules, understanding the relationship between their structure and function, discovering their putative functions etc. A key property of these fragments is their three-dimensional structure. At the same time, a vast amount of biomacromolecular structural data is currently available. We can benefit from these resources and focus on analyses of biomacromolecular structural fragments.

In my habilitation thesis, I describe key steps of this analysis – the validation, detection, extraction, comparison and characterization of biomacromolecular structural fragments – and methodologies for their realization. I also summarize here my contribution to the development of these approaches. Afterwards, I focus on particular applications of selected analyses: a quality comparison of different molecular classes, the prediction of acid dissociation constants using partial atomic charges, understanding apoptosis protein activation based on partial atomic charges and the discovery of enzyme channel anatomy.

Appendix

Bibliography

[1] Bolton, E. E., Wang, Y., Thiessen, P. A., and Bryant, S. H. (2008) Pubchem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry*, **4**, 217–241.

[2] Gaulton, A., et al. (2012) Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, **40**, D1100–D1107.

[3] Irwin, J. J., Sterling, T., Mysinger, M. M., Bolstad, E. S., and Coleman, R. G. (2012) Zinc: A free tool to discover chemistry for biology. *Journal of chemical information and modeling*, **52**, 1757–1768, pMID: 22587354.

[4] Law, V., et al. (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research*, **42**, D1091–7.

[5] Berman, H. M., Kleywegt, G. J., Nakamura, H., and Markley, J. L. (2014) The protein data bank archive as an open data resource. *Journal of computer-aided molecular design*, **28**, 1009–1014.

[6] Paulsen, C. E., Armache, J.-P., Gao, Y., Cheng, Y., and Julius, D. (2015) Structure of the TRPA1 ion channel suggests regulatory mechanisms. *Nature*, **520**, 511–517.

[7] Cao, E., Liao, M., Cheng, Y., and Julius, D. (2013) Trpv1 structures in distinct conformations reveal activation mechanisms. *Nature*, **504**, 113–118.

[8] Prota, A. E., Bargsten, K., Zurwerra, D., Field, J. J., Díaz, J. F., Altmann, K.-H., and Steinmetz, M. O. (2013) Molecular mechanism of action of microtubule-stabilizing anticancer agents. *Science*, **339**, 587–590.

[9] Lu, J., et al. (2005) Microrna expression profiles classify human cancers. *Nature*, **435**, 834–838.

[10] Puente, X. S., et al. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature*, **475**, 101–105.

[11] Nayal, M. and Honig, B. (2006) On the nature of cavities on protein surfaces: application to the identification of drug-binding sites. *Proteins*, **63**, 892–906.

[12] Xie, L., Xie, L., and Bourne, P. (2009) A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*, **25**, i305–i312.

[13] Thornton, J. M., Todd, A. E., Milburn, D., Borkakoti, N., and Orengo, C. A. O. A. (2000) From structure to function: approaches and limitations. *Nature structural biology*, **7**, 991–994.

[14] Kinoshita, K. and Nakamura, H. (2003) Protein informatics towards function identification. *Current opinion in structural biology*, **13**, 396–400.

[15] Watson, J. D., Laskowski, R. A., and Thornton, J. M. (2005) Predicting protein function from sequence and structural data. *Current opinion in structural biology*, **15**, 275–284.

[16] Eidhammer, I., Jonassen, I., and Taylor, W. R. (2000) Structure comparison and structure patterns. *Journal of computational biology*, **7**, 685–716.

[17] Chang, Y. S., Gelfand, T. I., Kister, A. E., and Gelfand, I. M. (2007) New classication of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage. *Proteins*, **68**, 915–921.

[18] Via, A., Ferre, F., Brannetti, B., Valencia, A., and Helmer-Citterich, M. (2000) Three-dimensional view of the surface motif associated with the p-loop structure: cis and trans cases of convergent evolution. *Journal of molecular biology*, **303**, 455–465.

[19] Ausiello, G., Peluso, D., Via, A., and Helmer-Citterich, M. (2007) Local comparison of protein structures highlights cases of convergent evolution in analogous functional sites. *BMC Bioinformatics*, **8**, S24.

[20] Gherardini, P. F., Wass, M. N., Helmer-Citterich, M., and Sternberg, M. J. E. (2007) Convergent evolution of enzyme active sites is not a rare phenomenon. *Journal of molecular biology*, **372**, 817–845.

[21] Bairoch, A. M., et al. (2005) The universal protein resource (uniprot). *Nucleic acids research*, **33**, D154–9.

[22] Kleywegt, G. J. (2009) On vital aid: the why, what and how of validation. *Acta crystallographica. Section D, Biological crystallography*, **65**, 134–9.

[23] Matthews, B. W. (2007) Five retracted structure reports: inverted or incorrect? *Protein science*, **16**, 1013–6.

[24] Rupp, B. (2012) Detection and analysis of unusual features in the structural model and structure-factor data of a birch pollen allergen. *Acta crystallographica. Section F, Structural biology and crystallization communications*, **68**, 366–76.

[25] Johnston, C. A., Kimple, A. J., Giguère, P. M., and Siderovski, D. P. (2008) RETRACTED: Structure of the Parathyroid Hormone Receptor C Terminus Bound to the G-Protein Dimer Gβ1γ2. *Structure*, **16**, 1086–1094.

[26] Hooft, R. W., Vriend, G., Sander, C., and Abola, E. E. (1996) Errors in protein structures. *Nature*, **381**, 272.

[27] Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, **26**, 283–291.

[28] Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta crystallographica. Section D, Biological crystallography*, **66**, 12–21.

[29] Kleywegt, G. J. and Jones, T. A. (1996) Efficient rebuilding of protein structures. *Acta crystallographica. Section D, Biological crystallography*, **52**, 829–32.

[30] Read, R. J., et al. (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure*, **19**, 1395 – 1412.

[31] Weininger, D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, **28**, 31–36.

[32] Proschak, E., Wegner, J. K., Schüller, A., Schneider, G., and Fechner, U. (2007) Molecular query language (MQL)–a context-free grammar for substructure matching. *Journal of chemical information and modeling*, **47**, 295–301.

[33] Homer, R. W., Swanson, J., Jilek, R. J., Hurst, T., and Clark, R. D. (2008) SYBYL line notation (SLN): A single notation to represent chemical structures, queries, reactions, and virtual libraries. *Journal of chemical information and modeling*, **48**, 2294–2307.

[34] Voss, N. R., Gerstein, M., Steitz, T. A., and Moore, P. B. (2006) The geometry of the ribosomal polypeptide exit tunnel. *Journal of molecular biology*, **360**, 893–906.

[35] Petrek, M., Otyepka, M., Banás, P., Kosinová, P., Koca, J., and Damborský, J. (2006) CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC bioinformatics*, **7**, 316.

[36] Medek, P., Benes, P., and Sochor, J. (2008) Multicriteria tunnel computation. *Computer graphics and Imaging*, Innsbruck.

[37] Chovancova, E., et al. (2012) CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS computational biology*, **8**, e1002708.

[38] Yaffe, E., Fishelovitch, D., Wolfson, H. J., Halperin, D., and Nussinov, R. (2008) MolAxis: efficient and accurate identification of channels in macromolecules. *Proteins*, **73**, 72–86.

[39] Yaffe, E., Fishelovitch, D., Wolfson, H. J., Halperin, D., and Nussinov, R. (2008) MolAxis: a server for identification of channels in macromolecules. *Nucleic acids research*, **36**, W210–5.

[40] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., and Ferrin, T. E. (2004) Ucsf chimera – a visualization system for exploratory research and analysis. *Journal of computational chemistry*, **25**, 1605–1612.

[41] Humphrey, W., Dalke, A., and Schulten, K. (1996) Vmd - visual molecular dynamics. *Journal of molecular graphics*, **14**, 33–38.

[42] Hess, B., Kutzner, C., Spoel, D., and Lindahl, E. (2008) Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *Journal of chemical theory and computation*, **4**, 435–447.

[43] Laaksonen, L., *gOpenmol*, version 2.0; csc — it center for science ltd.: Espoo, findland, 2001.

[44] *The PyMOL Molecular Graphics System*, version 1.3r1; Schrödinger, LLC: New york, ny, 2010.

[45] *MOE (The Molecular Operating Enviroment)*, version 2005.06; chemical computing group inc.: Montreal, quebec, canada, 2009.

[46] *Discovery Studio*, version 2.5; accelrys software inc.: San diego, ca, 2009.

[47] Abagyan, R., Totrov, M., and Kuznetsov, D. (1994) Icm - a new method for protein modeling and design. application to docking and structure prediction from the distorted native conformation. *Journal of computational chemistry*, **15**, 488–506.

[48] Mortier, W. J., Ghosh, S. K., and Shankar, S. (1986) Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. *Journal of the American Chemical Society*, **108**, 4315–4320.

[49] Yang*, Z.-Z., , and Wang, C.-S. (1997) Atom-bond electronegativity equalization method. 1. calculation of the charge distribution in large molecules. *The journal of physical chemistry A*, **101**, 6315–6321.

[50] Bultinck, P., Langenaeker, W., Lahorte, P., De Proft, F., Geerlings, P., Van Alsenoy, C., and Tollenaere, J. P. (2002) The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. *Journal of physical chemistry A*, **106**, 7895–7901.

[51] Bultinck, P., Vanholme, R., Popelier, P. L. A., De Proft, F., and Geerlings, P. (2004) High-speed Calculation of AIM Charges Through the Electronegativity Equalization Method. *Journal of physical chemistry A*, **108**, 10359–10366.

[52] Chaves, J., Barroso, J. M., Bultinck, P., and Carbo-Dorca, R. (2006) Toward an alternative hardness kernel matrix structure in the electronegativity equalization method (eem). *Journal of chemical information and modeling*, **46**, 1657–1665.

[53] Ouyang, Y., Ye, F., and Liang, Y. (2009) A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules. *Physical chemistry chemical physics*, **11**, 6082–9.

[54] Gross, K. C., Seybold, P. G., and Hadad, C. M. (2002) Comparison of Different Atomic Charge Schemes for Predicting pKa Variations in Substituted Anilines and Phenols. *International journal of quantum chemistry*, **90**, 445–458.

[55] Kreye, W. C. and Seybold, P. G. (2009) Correlations between quantum chemical indices and the pk_as of a diverse set of organic phenols. *International journal of quantum chemistry*, **109**, 3679–3684.

[56] Dixon, S. L. and Jurs, P. C. (1993) Estimation of pKa for Organic Oxyacids Using Calculated Atomic Charges. *Journal of computational chemistry*, **14**, 1460–1467.

[57] Citra, M. J. (1999) Estimating the pk_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods. *Chemosphere*, **1**, 191–206.

[58] Comer, J. and Tam, K. (2001) *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*. Verlag Helvetica Chimica Acta, Postfach, CH-8042 Zürich, Switzerland.

[59] Klebe, G. (2000) Recent developments in structure-based drug design. *Journal of molecular medicine*, **78**, 269–281.

[60] Kim, J. H., Gramatica, P., Kim, M. G., Kim, D., and Tratnyek, P. G. (2007) Qsar modelling of water quality indices of alkylphenol pollutants. *SAR and QSAR in environmental research*, **18**, 729–743.

[61] Gasteiger, J. and Engel, T. (2006) *Chemoinformatics: a textbook*. John Wiley & Sons.

[62] Sadowski, J. and Gasteiger, J. (1993) From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chemical reviews*, **93**, 2567–2581.

[63] Hawkins, P. C. D., Skillman, A. G., Warren, G. L., Ellingson, B. A., and Stahl, M. T. (2010) Conformer generation with omega: Algorithm and validation using high quality structures from the protein databank and cambridge structural database. *Journal of chemical information and modeling*, **50**, 572–584.

[64] Czabotar, P. E., Colman, P. M., and Huang, D. C. S. (2009) Bax activation by Bim? *Cell death and differentiation*, **16**, 1187–1191.

[65] Westphal, D., Dewson, G., Czabotar, P. E., and Kluck, R. M. (2011) Molecular biology of Bax and Bak activation and action. *Biochimica et biophysica acta*, **1813**, 521–531.

[66] Huang, X., Holden, H. M., and Raushel, F. M. (2001) Channeling of substrates and intermediates in enzyme-catalyzed reactions. *Annual review of biochemistry*, **70**, 149–80.

[67] Park, J., Czapla, L., and Amaro, R. E. (2013) Molecular simulations of aromatase reveal new insights into the mechanism of ligand binding. *Journal of chemical information and modeling*, **53**, 2047–56.

[68] Sgrignani, J. and Magistrato, A. (2012) Influence of the membrane lipophilic environment on the structure and on the substrate access/egress routes of the human aromatase enzyme. A computational study. *Journal of chemical information and modeling*, **52**, 1595–606.

[69] Madrona, Y., Hollingsworth, S. A., Khan, B., and Poulos, T. L. (2013) P450cin active site water: implications for substrate binding and solvent accessibility. *Biochemistry*, **52**, 5039–50.

[70] Cui, Y.-L., Zhang, J.-L., Zheng, Q.-C., Niu, R.-J., Xu, Y., Zhang, H.-X., and Sun, C.-C. (2013) Structural and dynamic basis of human cytochrome P450 7B1: a survey of substrate selectivity and major active site access channels. *Chemistry*, **19**, 549–57.

[71] Lee, S. J., McCormick, M. S., Lippard, S. J., and Cho, U.-S. (2013) Control of substrate access to the active site in methane monooxygenase. *Nature*, **494**, 380–4.

[72] Pryor, E. E., Horanyi, P. S., Clark, K. M., Fedoriw, N., Connelly, S. M., Koszelak-Rosenblum, M., Zhu, G., Malkowski, M. G., Wiener, M. C., and Dumont, M. E. (2013) Structure of the integral membrane protein CAAX protease Ste24p. *Science*, **339**, 1600–4.

[73] Xu, S., Mueser, T. C., Marnett, L. J., and Funk, M. O. (2012) Crystal structure of 12-lipoxygenase catalytic-domain-inhibitor complex identifies a substrate-binding channel for catalysis. *Structure*, **20**, 1490–7.

[74] Guskov, A., Nordin, N., Reynaud, A., Engman, H., Lundbäck, A.-K., Jong, A. J. O., Cornvik, T., Phua, T., and Eshaghi, S. (2012) Structural insights into the mechanisms of Mg2+ uptake, transport, and gating by CorA. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 18459–64.

[75] Otyepka, M., Berka, K., and Anzenbacher, P. (2012) Is there a relationship between the substrate preferences and structural flexibility of cytochromes P450? *Current drug metabolism*, **13**, 130–42.

[76] Rengachari, S., Aschauer, P., Schittmayer, M., Mayer, N., Gruber, K., Breinbauer, R., Birner-Gruenberger, R., Dreveny, I., and Oberer, M. (2013) Conformational plasticity and ligand binding of bacterial monoacylglycerol lipase. *Journal of biological chemistry*, **288**, 31093–104.

[77] Salter, M. D., Blouin, G. C., Soman, J., Singleton, E. W., Dewilde, S., Moens, L., Pesce, A., Nardini, M., Bolognesi, M., and Olson, J. S. (2012) Determination of ligand pathways in globins: apolar tunnels versus polar gates. *Journal of biological chemistry*, **287**, 33163–78.

[78] Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) Aqua and procheck-nmr: Programs for checking the quality of protein structures solved by nmr. *Journal of biomolecular NMR*, **8**, 477–486.

[79] Kleywegt, G. J. and Harris, M. R. (2007) ValLigURL: a server for ligand-structure comparison and validation. *Acta crystallographica. Section D, Biological crystallography*, **63**, 935–8.

[80] Bruno, I. J., et al. (2004) Retrieval of crystallographically-derived molecular geometry information. *Journal of Chemical Information and Computer Sciences*, **44**, 2133–2144.

[81] Debreczeni, J. É. and Emsley, P. (2012) Handling ligands with Coot. *Acta crystallographica. Section D, Biological crystallography*, **68**, 425–30.

[82] Adams, P. D., et al. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta crystallographica. Section D, Biological crystallography*, **66**, 213–21.

[83] Lütteke, T. and von der Lieth, C.-W. (2004) pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC bioinformatics*, **5**, 69.

[84] Feng, Z., Chen, L., Maddula, H., Akcan, O., Oughtred, R., Berman, H. M., and Westbrook, J. (2004) Ligand depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, **20**, 2153–2155.

[85] Xu, K., Rajashankar, K. R., Chan, Y.-P., Himanen, J. P., Broder, C. C., and Nikolov, D. B. (2008) Host cell recognition by the henipaviruses: crystal structures of the Nipah G attachment glycoprotein and its complex with ephrin-B3. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 9953–8.

[86] Sheila Ash, ., Cline, M. A., Homer, R. W., Hurst, T., , and Smith, G. B. (1997) Sybyl line notation (sln): A versatile language for chemical structure representation. *Journal of chemical information and computer sciences*, **37**, 71–79.

[87] Hauck, D., Joachim, I., Frommeyer, B., Varrot, A., Philipp, B., Möller, H. M., Imberty, A., Exner, T. E., and Titz, A. (2013) Discovery of two classes of potent glycomimetic inhibitors of Pseudomonas aeruginosa LecB with distinct binding modes. *ACS chemical biology*, **8**, 1775–84.

[88] Ernst, B. and Magnani, J. L. (2009) From carbohydrate leads to glycomimetic drugs. *Nature reviews. Drug discovery*, **8**, 661–677.

[89] Mitchell, E., Houles, C., Sudakevitz, D., Wimmerova, M., Gautier, C., Pérez, S., Wu, A. M., Gilboa-Garber, N., and Imberty, A. (2002) Structural basis for oligosaccharide-mediated adhesion of Pseudomonas aeruginosa in the lungs of cystic fibrosis patients. *Nature structural biology*, **9**, 918–21.

[90] Mitchell, E. P., et al. (2005) High affinity fucose binding of pseudomonas aeruginosa lectin pa-iil: 1.0 Å resolution crystal structure of the complex combined with thermodynamics and computational chemistry approaches. *Proteins: Structure, Function, and Bioinformatics*, **58**, 735–746.

[91] Walz, T., Smith, B. L., Agre, P., and Engel, A. (1994) The three-dimensional structure of human erythrocyte aquaporin CHIP. *The EMBO journal*, **13**, 2985–93.

[92] Jiang, Y., Lee, A., Chen, J., Cadene, M., Chait, B. T., and MacKinnon, R. (2002) Crystal structure and mechanism of a calcium-gated potassium channel. *Nature*, **417**, 515–22.

[93] Doyle, D. A., Morais Cabral, J., Pfuetzner, R. A., Kuo, A., Gulbis, J. M., Cohen, S. L., Chait, B. T., and MacKinnon, R. (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science*, **280**, 69–77.

[94] Alexander, S. P. H., Mathie, A., and Peters, J. A. (2011) Guide to Receptors and Channels (GRAC), 5th edition. *British journal of pharmacology*, **164 Suppl**, S1–324.

[95] MacKinnon, R. (2004) Potassium channels and the atomic basis of selective ion conduction (Nobel Lecture). *Angewandte chemie*, **43**, 4265–77.

[96] Murray, J. W. and Barber, J. (2007) Structural characteristics of channels and pathways in photosystem II including the identification of an oxygen channel. *Journal of structural biology*, **159**, 228–37.

[97] Guskov, A., Kern, J., Gabdulkhakov, A., Broser, M., Zouni, A., and Saenger, W. (2009) Cyanobacterial photosystem II at 2.9-A resolution and the role of quinones, lipids, channels and chloride. *Nature structural & molecular biology*, **16**, 334–42.

[98] Wade, R. C., Winn, P. J., Schlichting, I., and Sudarko (2004) A survey of active site access channels in cytochromes P450. *Journal of inorganic biochemistry*, **98**, 1175–82.

[99] Otyepka, M., Skopalík, J., Anzenbacherová, E., and Anzenbacher, P. (2007) What common structural features and variations of mammalian P450s are known to date? *Biochimica et biophysica acta*, **1770**, 376–89.

[100] Berka, K., Hendrychová, T., Anzenbacher, P., and Otyepka, M. (2011) Membrane position of ibuprofen agrees with suggested access path entrance to cytochrome P450 2C9 active site. *The journal of physical chemistry. A*, **115**, 11248–55.

[101] Hendrychova, T., Berka, K., Navratilova, V., Anzenbacher, P., and Otyepka, M. (2012) Dynamics and hydration of the active sites of mammalian cytochromes P450 probed by molecular dynamics simulations. *Current drug metabolism*, **13**, 177–89.

[102] Cojocaru, V., Winn, P. J., and Wade, R. C. (2007) The ins and outs of cytochrome P450s. *Biochimica et biophysica acta*, **1770**, 390–401.

[103] Gilson, M. K., Straatsma, T. P., McCammon, J. A., Ripoll, D. R., Faerman, C. H., Axelsen, P. H., Silman, I., and Sussman, J. L. (1994) Open "back door" in a molecular dynamics simulation of acetylcholinesterase. *Science*, **263**, 1276–8.

[104] Wiesner, J., Kriz, Z., Kuca, K., Jun, D., and Koca, J. (2007) Acetylcholinesterases–the structural similarities and differences. *Journal of enzyme inhibition and medicinal chemistry*, **22**, 417–24.

[105] Sanson, B., Colletier, J.-P., Xu, Y., Lang, P. T., Jiang, H., Silman, I., Sussman, J. L., and Weik, M. (2011) Backdoor opening mechanism in acetylcholinesterase based on X-ray crystallography and molecular dynamics simulations. *Protein science : a publication of the Protein Society*, **20**, 1114–8.

[106] Petřek, M., Košínová, P., Koča, J., and Otyepka, M. (2007) MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure*, **15**, 1357–63.

[107] Pavlova, M., Klvana, M., Prokop, Z., Chaloupkova, R., Banas, P., Otyepka, M., Wade, R. C., Tsuda, M., Nagata, Y., and Damborsky, J. (2009) Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nature chemical biology*, **5**, 727–33.

[108] Biedermannová, L., Prokop, Z., Gora, A., Chovancová, E., Kovács, M., Damborsky, J., and Wade, R. C. (2012) A single mutation in a tunnel to the active site changes the mechanism and kinetics of product release in haloalkane dehalogenase LinB. *Journal of biological chemistry*, **287**, 29062–74.

[109] Brezovsky, J., Chovancova, E., Gora, A., Pavelka, A., Biedermannova, L., and Damborsky, J. (2013), Software tools for identification, visualization and analysis of protein tunnels and channels.

[110] Lee, P.-H. and Helms, V. (2011) Identifying continuous pores in protein structures with PROPORES by computational repositioning of gating residues. *Proteins*, **80**, 421–432.

[111] Levitt, D. G. and Banaszak, L. J. (1992) POCKET: A computer graphies method for identifying and displaying protein cavities and their surrounding amino acids. *Journal of molecular graphics*, **10**, 229–234.

[112] Voss, N. R. and Gerstein, M. (2010) 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic acids research*, **38**, W555–62.

[113] Hendlich, M., Rippmann, F., and Barnickel, G. (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *Journal of molecular graphics and modelling*, **15**, 359–363.

[114] Huang, B. and Schroeder, M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC structural biology*, **6**, 19.

[115] Raunest, M. and Kandt, C. (2011) dxTuber: detecting protein cavities, tunnels and clefts based on protein and solvent dynamics. *Journal of molecular graphics & modelling*, **29**, 895–905.

[116] Ho, B. K. and Gruswitz, F. (2008) HOLLOW: generating accurate representations of channel and interior surfaces in molecular structures. *BMC structural biology*, **8**, 49.

[117] Coleman, R. G. and Sharp, K. A. (2009) Finding and characterizing tunnels in macromolecules with application to ion channels and pores. *Biophysical journal*, **96**, 632–45.

[118] Laskowski, R. A. (1995) SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions. *Journal of Molecular Graphics*, **13**, 323–330.

[119] Brady, G. P., Stouten, P. F., and Brady Jr., G. P. (2000) Fast prediction and visualization of protein binding pockets with PASS. *Journal of computer-aided molecular design*, **14**, 383–401.

[120] Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A., and Sansom, M. S. (1996) HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *Journal of molecular graphics*, **14**, 354–360.

[121] Pellegrini-Calace, M., Maiwald, T., and Thornton, J. M. (2009) PoreWalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS computational biology*, **5**, e1000440.

[122] Anzenbacher, P. and Anzenbacherová, E. (2001) Cytochromes p450 and metabolism of xenobiotics. *Cellular and Molecular Life Sciences CMLS*, **58**, 737–747.

[123] Guengerich, F. P. (2005) *Cytochrome P450: Structure, Mechanism, and Biochemistry*, chap. Human Cytochrome P450 Enzymes, pp. 377–530. Springer US.

[124] Anzenbacher, P., Anzenbacherová, E., Lange, R., Skopalík, J., and Otyepka, M. (2008) Active sites of cytochromes p450: what are they like? *Acta chimica Slovenica*, **55**, 63.

[125] Skopalík, J., Anzenbacher, P., and Otyepka, M. (2008) Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences. *The journal of physical chemistry. B*, **112**, 8165–73.

[126] Conner, K. P., Woods, C. M., and Atkins, W. M. (2011) Interactions of cytochrome {P450s} with their ligands. *Archives of Biochemistry and Biophysics*, **507**, 56 – 65, {P450} Catalysis Mechanisms.

[127] Needleman, S. and Wunsch, C. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, **48**, 443–453.

[128] Abagyan, R. A. and Batalov, S. (1997) Do aligned sequences share the same fold? *Journal of molecular biology*, **273**, 355–368.

[129] McLachlan, A. D. (1972) A mathematical procedure for superimposing atomic coordinates of proteins. *Acta crystallographica*, **28**, 656–657.

[130] Kabsch, W. (1976) A solution for the best rotation to relate two sets of vectors. *Acta crystallographica*, **32**, 922–923.

[131] Horn, B. K. P. (1987) Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, **4**, 629–642.

[132] Diamond, R. (1988) A note on the rotational superposition problem. *Acta crystallographica*, **44**, 211–216.

[133] Kearsley, S. K. (1989) On the orthogonal transformation used for structural comparisons. *Acta crystallographica*, **45**, 208–210.

[134] Coutsias, E. A., Seok, C., and Dill, K. A. (2004) Using quaternions to calculate rmsd. *Journal of computational chemistry*, **25**, 1849–1857.

[135] Wang, X. and Snoeyink, J. (2008) Defining and computing optimum rmsd for gapped and weighted multiple-structure alignment. *IEEE/ACM transactions on computational biology and bioinformatics*, **5**, 525–533.

[136] Moller, H., Martinez-Yamout, M., Dyson, H., and Wright, P. (2005) Solution structure of the N-terminal zinc fingers of the Xenopus laevis double-stranded RNA-binding protein ZFa. *Journal of molecular biology*, **351**, 718–730.

[137] Zhang, J., Kleinöder, T., and Gasteiger, J. (2006) Prediction of pKa Values for Aliphatic Carboxylic Acids and Alcohols With Empirical Atomic Charge Descriptors. *Journal of chemical information and modeling*, **46**, 2256–2266.

[138] Ghafourian, T. and Dearden, J. (2000) The Use of Atomic Charges and Orbital Energies as Hydrogen-bonding-donor Parameters for QSAR Studies: Comparison of MNDO, AM1 and PM3 Methods. *Journal of pharmacy and pharmacology*, **52**, 603–610.

[139] Dudek, A. Z., Arodz, T., and Gálvez, J. (2006) Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial chemistry and high throughput screening*, **9**, 213–28.

[140] Karelson, M., Lobanov, V. S., and Katritzky, A. R. (1996) Quantum-Chemical Descriptors in QSAR/QSPR Studies. *Chemical reviews*, **96**, 1027–1044.

[141] Todeschini, R. and Consonni, V. (2008) *Handbook of Molecular Descriptors*. Wiley-VCH Verlag GmbH.

[142] Galvez, J., Garcia, R., Salabert, M. T., and Soler, R. (1994) Charge Indexes. New Topological Descriptors. *Journal of chemical information and modeling*, **34**, 520–525.

[143] Stalke, D. (2011) Meaningful structural descriptors from charge density. *Chemistry*, **17**, 9264–78.

[144] Wermuth, C. G. (2006) Pharmacophores: historical perspective and viewpoint from a medicinal chemist. Langer, T. and Hoffmann, R. D. (eds.), *Pharmacophores and Pharmacophore Searches. Volume 32*, Wiley-VCH Verlag GmbH & Co. KGaA.

[145] MacDougall, P. J. and Henze, C. E. (2007) Fleshing-out Pharmacophores with Volume Rendering of the Laplacian of the Charge Density and Hyperwall Visualization Technology. Matta, C. F. and Boyd, R. J. (eds.), *The Quantum Theory of Atoms in Molecules: From Solid State to DNA and Drug Design*, Wiley-VCH Verlag GmbH & Co. KGaA.

[146] Clement, O. O. and Mehl, A. T. (2000) HipHop: pharmacophores based on multiple common-feature alignments. Güner, O. F. (ed.), *Pharmacophore perception, development, and use in drug design*, International University Line.

[147] Lyne, P. D. (2002) Structure-based virtual screening: an overview. *Drug discovery today*, **7**, 1047–1055.

[148] Bissantz, C., Folkers, G., and Rognan, D. (2000) Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *Journal of medicinal chemistry*, **43**, 4759–4767.

[149] Park, H., Lee, J., and Lee, S. (2006) Critical assessment of the automated AutoDock as a new docking tool for virtual screening. *Proteins*, **65**, 549–54.

[150] Kearsley, S., Sallamack, S., Fluder, E., Andose, J., Mosley, R., and Sheridan, R. (1996) Chemical Similarity Using Physiochemical Property Descriptors. *Journal of chemical information and modeling*, **36**, 118–127.

[151] Nikolova, N. and Jaworska, J. (2003) Approaches to Measure Chemical Similarity – a Review. *QSAR & combinatorial science*, **22**, 1006–1026.

[152] Holliday, J. D., Jelfs, S. P., Willett, P., and Gedeck, P. (2003) Calculation of intersubstituent similarity using R-group descriptors. *Journal of chemical information and computer sciences*, **43**, 406–11.

[153] Tervo, A. J., Rönkkö, T., Nyrönen, T. H., and Poso, A. (2005) BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. 1. Alignment and virtual screening applications. *Journal of medicinal chemistry*, **48**, 4076–86.

[154] Vainio, M. J., Puranen, J. S., and Johnson, M. S. (2009) ShaEP: molecular overlay based on shape and electrostatic potential. *Journal of chemical information and modeling*, **49**, 492–502.

[155] Lemmen, C., Lengauer, T., and Klebe, G. (1998) FLEXS: a method for fast flexible ligand superposition. *Journal of medicinal chemistry*, **41**, 4502–20.

[156] Mulliken, R. S. (1955) Electronic Population Analysis on LCAO-MO Molecular Wave Functions. II. Overlap Populations, Bond Orders, and Covalent Bond Energies. *Journal of chemical physics*, **23**, 1841.

[157] Mulliken, R. S. (1955) Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I. *Journal of chemical physics*, **23**, 1833.

[158] Löwdin, P.-O. (1950) On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *Journal of chemical physics*, **18**, 365.

[159] Reed, A. E. and Weinhold, F. (1983) Natural bond orbital analysis of near-Hartree-Fock water dimer. *Journal of chemical physics*, **78**, 4066–4073.

[160] Reed, A. E., Weinstock, R. B., and Weinhold, F. (1985) Natural population analysis. *Journal of chemical physics*, **83**, 735.

[161] Bader, R. F. W. (1985) Atoms in molecules. *Accounts of chemical research*, **18**, 9–15.

[162] Bader, R. F. W. (1991) A quantum theory of molecular structure and its applications. *Chemical reviews*, **91**, 893–928.

[163] Hirshfeld, F. L. (1977) Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta*, **44**, 129–138.

[164] Ritchie, J. P. (1985) Electron density distribution analysis for nitromethane, nitromethide, and nitramide. *Journal of the American Chemical Society*, **107**, 1829–1837.

[165] Ritchie, J. P. and Bachrach, S. M. (1987) Some methods and applications of electron density distribution analysis. *Journal of computational chemistry*, **8**, 499–509.

[166] Breneman, C. M. and Wiberg, K. B. (1990) Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.*, **11**, 361–373.

[167] Singh, U. C. and Kollman, P. A. (1984) An approach to computing electrostatic charges for molecules. *Journal of computational chemistry*, **5**, 129–145.

[168] Besler, B. H., Merz, K. M., and Kollman, P. A. (1990) Atomic charges derived from semiempirical methods. *Journal of computational chemistry*, **11**, 431–439.

[169] Kelly, C. P., Cramer, C. J., and Truhlar, D. G. (2005) Accurate partial atomic charges for high-energy molecules using class IV charge models with the MIDI! basis set. *Theoretical chemistry accounts*, **113**, 133–151.

[170] Marenich, A. V., Cramer, C. J., and Truhlar, D. G. (2009) Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *Journal of physical chemistry B*, **113**, 6378–96.

[171] Gasteiger, J. and Marsili, M. (1978) A new model for calculating atomic charges in molecules. *Tetrahedron letters*, **19**, 3181–3184.

[172] Gasteiger, J. and Marsili, M. (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, **36**, 3219–3228.

[173] Cho, K.-H., Kang, Y. K., No, K. T., and Scheraga, H. A. (2001) A Fast Method for Calculating Geometry-Dependent Net Atomic Charges for Polypeptides. *Journal of physical chemistry B*, **105**, 3624–3634.

[174] Oliferenko, A. A., Pisarev, S. A., Palyulin, V. A., and Zefirov, N. S. (2006) Atomic Charges via Electronegativity Equalization: Generalizations and Perspectives. *Advances in quantum chemistry*, **51**, 139–156.

[175] Shulga, D. A., Oliferenko, A. A., Pisarev, S. A., Palyulin, V. A., and Zefirov, N. S. (2010) Fast tools for calculation of atomic charges well suited for drug design. *SAR QSAR Environ. Res.*, **19**, 153–65.

[176] Rappe, A. K. and Goddard, W. A. (1991) Charge equilibration for molecular dynamics simulations. *Journal of physical chemistry*, **95**, 3358–3363.

[177] Nistor, R. A., Polihronov, J. G., Müser, M. H., and Mosey, N. J. (2006) A generalization of the charge equilibration method for nonmetallic materials. *Journal of chemical physics*, **125**, 094108.

[178] Mathieu, D. (2007) Split charge equilibration method with correct dissociation limits. *Journal of chemical physics*, **127**, 224103.

[179] Baekelandt, B. G., Mortier, W. J., Lievens, J. L., and Schoonheydt, R. A. (1991) Probing the reactivity of different sites within a molecule or solid by direct computation of molecular sensitivities via an extension of the electronegativity equalization method. *Journal of the American Chemical Society*, **113**, 6730–6734.

[180] Jiroušková, Z., Vařeková, R. S., Vaněk, J., and Koča, J. (2009) Electronegativity equalization method: parameterization and validation for organic molecules using the Merz-Kollman-Singh charge distribution scheme. *Journal of computational chemistry*, **30**, 1174–8.

[181] Vainio, M. J. and Johnson, M. S. (2007) Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *Journal of chemical information and modeling*, **47**, 2462–2474.

[182] O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T., and Hutchison, G. (2011) Open Babel: An Open Chemical Toolbox. *Journal of cheminformatics*, **3**, 33–47.

[183] Raček, T., Svobodová Vařeková, R., Křenek, A., and Koča, J., NEEMP – Tool for parameterization of empirical charge calculation method EEM.

[184] Golub, G. H. and Ortega, J. M. (2014) *Scientific computing: an introduction with parallel computing*. Elsevier.

[185] Geist, A., Beguelin, A., Dongarra, J., Jiang, W., Mancheck, R., and Sunderam, V. (1994), Pvm: Parallel virtual machine. scientific and engineering computation.

[186] Page, R. C., Pruneda, J. N., Amick, J., Klevit, R. E., and Misra, S. (2012) Structural insights into the conformation and oligomerization of e2 ubiquitin conjugates. *Biochemistry*, **51**, 4175–4187, pMID: 22551455.

[187] Kyte, J. and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *Journal of molecular Biology*, **157**, 105–132.

[188] Andersen, O., Koeppe, R., and Roux, B. (2005) Gramicidin Channels. *IEEE Transactions on Nanobioscience*, **4**, 10–20.

[189] Brzezinski, P. and Gennis, R. B. (2008) Cytochrome c oxidase: exciting progress and remaining mysteries. *Journal of bioenergetics and biomembranes*, **40**, 521–31.

[190] Unwin, N. (2005) Refined structure of the nicotinic acetylcholine receptor at 4A resolution. *Journal of molecular biology*, **346**, 967–89.

[191] Teng, Y.-B., Jiang, Y.-L., He, Y.-X., He, W.-W., Lian, F.-M., Chen, Y., and Zhou, C.-Z. (2009) Structural insights into the substrate tunnel of Saccharomyces cerevisiae carbonic anhydrase Nce103. *BMC structural biology*, **9**, 67.

[192] Liebeschuetz, J., Hennemann, J., Olsson, T., and Groom, C. R. (2012) The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. *Journal of computer-aided molecular design*, **26**, 169–83.

[193] Ishihama, Y., Nakamura, M., Miwa, T., Kajima, T., and Asakawa, N. (2002) A rapid method for pk_a determination of drugs using pressure-assisted capillary electrophoresis with photodiode array detection in drug discovery. *Journal of pharmaceutical sciences*, **91**, 933–942.

[194] Babić, S., Horvat, A. J., Pavlović, D. M., and Kaštelan-Macan, M. (2007) Determination of pk_a values of active pharmaceutical ingredients. *TrAC*, **26**, 1043–1061.

[195] Manallack, D. (2007) The pk$_a$ distribution of drugs: Application to drug discovery. *Perspectives in medicinal chemistry*, **1**, 25–38.

[196] Wan, H. and Ulander, J. (2006) High-throughput pk_a screening and prediction amenable for adme profiling. *Expert opinion on drug metabolism & toxicology*, **2**, 139–155.

[197] Cruciani, G., Milletti, F., Storchi, L., Sforna, G., and Goracci, L. (2009) *In silico* pK_a prediction and adme profiling. *Chemistry & biodiversity*, **6**, 1812–1821.

[198] Lee, A. C. and Crippen, G. M. (2009) Predicting pk_a. *Journal of chemical information and modeling*, **49**, 2013–2033.

[199] Rupp, M., Körner, R., and Tetko, I. V. (2010) Predicting the pK$_a$ of small molecules. *Combinatorial chemistry and high throughput screening*, **14**, 307–327.

[200] Fraczkiewicz, R. (2006) *In Silico Prediction of Ionization*, vol. 5. Elsevier.

[201] Ho, J. and Coote, M. (2010) A universal approach for continuum solvent pk_a calculations: Are we there yet? *Theoretica chimica acta*, **125**, 3–21.

[202] Clark, J. and Perrin, D. D. (1964) Prediction of the strengths of organic bases. *Quarterly reviews of the Chemical Society*, **18**, 295–320.

[203] Perrin, D. D., Dempsey, B., and Serjeant, E. P. (1981) *pK_a prediction for organic acids and bases*. Chapman and Hall: New York.

[204] Blower, P. E. and Cross, K. P. (2006) Decision tree methods in pharmaceutical research. *Current topics in medicinal chemistry*, **6**, 31–39.

[205] Liptak, M. D., Gross, K. C., Seybold, P. G., Feldgus, S., and Shields, G. (2002) Absolute pk_a determinations for substituted phenols. *Journal of the American Chemical Society*, **124**, 6421–6427.

[206] Toth, A. M., Liptak, M. D., Phillips, D. L., and Shields, G. C. (2001) Accurate relative pk_a calculations for carboxylic acids using complete basis set and gaussian-n models combined with continuum solvation methods. *Journal of chemical physics*, **114**, 4595–4606.

[207] Jelfs, S., Ertl, P., and Selzer, P. (2007) Estimation of pk_a for druglike compounds using semiempirical and information-based descriptors. *Journal of chemical information and modeling*, **47**, 450–459.

[208] Hagan, M. T., Demuth, H. B., and Beale, M. (1996) *In Neural Network Design*. PWS: Boston, MA.

[209] Xing, L. and Glen, R. C. (2002) Novel methods for the prediction of logp, pk_a, and logd. *Journal of chemical information and computer sciences*, **42**, 796–805.

[210] Soriano, E., Cerdán, S., and Ballesteros, P. (2004) Computational determination of pk_a values. a comparison of different theoretical approaches and a novel procedure. *Theochem*, **684**, 121–128.

[211] Habibi-Yangjeh, A. (2006) Application of artificial neural networks for predicting the aqueous acidity of various phenols using qsar. *Journal of molecular modeling*, **12**, 338–347.

[212] Nci open database compounds. Retrieved from http://cactus.nci.nih.gov/ on August 10, 2010.

[213] Howard, P. and Meylan, W., Physical/chemical property database (physprop). Syracuse Research Corporation, Environmental Science Center, North Syracuse NY, 1999.

[214] Landrum, G., Rdkit: Open-source cheminformatics. Retrieved from http://www.rdkit.org on January 10, 2014.

[215] Miteva, M. A., Guyon, F., and Tufféry, P. (2010) Frog2: Efficient 3d conformation ensemble generator for small compounds. *Nucleic acids research*, **38**, W622–W627.

[216] Kroemer, G., Galluzzi, L., and Brenner, C. (2007) Mitochondrial Membrane Permeabilization in Cell Death. *Physiological reviews*, **87**, 99–163.

[217] Wei, M. C. (2001) Proapoptotic BAX and BAK: A Requisite Gateway to Mitochondrial Dysfunction and Death. *Science*, **292**, 727–730.

[218] Kuwana, T. and Newmeyer, D. D. (2003) Bcl-2-family proteins and the role of mitochondria in apoptosis. *Current opinion in cell biology*, **15**, 691–699.

[219] Tait, S. W. G. and Green, D. R. (2010) Mitochondria and cell death: outer membrane permeabilization and beyond. *Nature reviews molecular cell biology*, **11**, 621–632.

[220] Letai, A., Bassik, M. C., Walensky, L. D., Sorcinelli, M. D., Weiler, S., and Korsmeyer, S. J. (2002) Distinct BH3 domains either sensitize or activate mitochondrial apoptosis, serving as prototype cancer therapeutics. *Cancer cell*, **2**, 183–192.

[221] Marsden, V. S. and Strasser, A. (2003) Control of apoptosis in the immune system: Bcl-2, BH3-Only Proteins and More. *Annual review of immunology*, **21**, 71–105.

[222] Leber, B., Lin, J., and Andrews, D. W. (2007) Embedded together: The life and death consequences of interaction of the Bcl-2 family with membranes. *Apoptosis*, **12**, 897–911.

[223] Chipuk, J. E. and Green, D. R. (2008) How do BCL-2 proteins induce mitochondrial outer membrane permeabilization? *Trends in cell biology*, **18**, 157–164.

[224] Eskes, R., Desagher, S., Antonsson, B., and Martinou, J.-C. (2000) Bid Induces the Oligomerization and Insertion of Bax into the Outer Mitochondrial Membrane. *Molecular and cellular biology*, **20**, 929–935.

[225] Kuwana, T., Bouchier-Hayes, L., Chipuk, J. E., Bonzon, C., Sullivan, B. A., Green, D. R., and Newmeyer, D. D. (2005) BH3 domains of BH3-only proteins differentially regulate Bax-mediated mitochondrial membrane permeabilization both directly and indirectly. *Molecular cell*, **17**, 525–35.

[226] Walensky, L. D., Pitter, K., Morash, J., Oh, K. J., Barbuto, S., Fisher, J., Smith, E., Verdine, G. L., and Korsmeyer, S. J. (2006) A Stapled BID BH3 Helix Directly Binds and Activates BAX. *Molecular cell*, **24**, 199–210.

[227] Wolter, K. G., Hsu, Y.-T., Smith, C. L., Nechushtan, A., Xi, X.-G., and Youle, R. J. (1997) Movement of Bax from the Cytosol to Mitochondria during Apoptosis. *Journal of cell biology*, **139**, 1281–1292.

[228] Hsu, Y.-T., Wolter, K. G., and Youle, R. J. (1997) Cytosol-to-membrane redistribution of Bax and Bcl-XL during apoptosis. *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 3668–3672.

[229] Suzuki, M., Youle, R. J., and Tjandra, N. (2000) Structure of Bax: Coregulation of Dimer Formation and Intracellular Localization. *Cell*, **103**, 645–654.

[230] Lovell, J. F., Billen, L. P., Bindner, S., Shamas-Din, A., Fradin, C., Leber, B., and Andrews, D. W. (2008) Membrane Binding by tBid Initiates an Ordered Series of Events Culminating in Membrane Permeabilization by Bax. *Cell*, **135**, 1074–1084.

[231] Gavathiotis, E., et al. (2008) BAX activation is initiated at a novel interaction site. *Nature*, **455**, 1076–1081.

[232] van der Vaart, A., Bursulaya, B. D., Brooks, C. L., and Merz, K. M. (2000) Are Many-Body Effects Important in Protein Folding? *Journal of physical chemistry B*, **104**, 9554–9563.

[233] Cho, A. E., Guallar, V., Berne, B. J., and Friesner, R. (2005) Importance of accurate charges in molecular docking: Quantum mechanical/molecular mechanical (QM/MM) approach. *Journal of computational chemistry*, **26**, 915–931.

[234] Bucher, D., Raugei, S., Guidoni, L., Dal Peraro, M., Rothlisberger, U., Carloni, P., and Klein, M. L. (2006) Polarization effects and charge transfer in the KcsA potassium channel. *Biophysical chemistry*, **124**, 292–301.

[235] George, N. M., Evans, J. J., and Luo, X. (2007) A three-helix homo-oligomerization domain containing BH3 and BH1 is responsible for the apoptotic activity of Bax. *Genes & development*, **21**, 1937–1948.

[236] Webb, E. C. (1992) *Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. Published for the International Union of Biochemistry and Molecular Biology by Academic Press.

[237] Kraut, D. A., Carroll, K. S., and Herschlag, D. (2003) Challenges in enzyme mechanism and energetics. *Annual review of biochemistry*, **72**, 517–71.

[238] Warshel, A., Sharma, P. K., Kato, M., Xiang, Y., Liu, H., and Olsson, M. H. M. (2006) Electrostatic basis for enzyme catalysis. *Chemical reviews*, **106**, 3210–35.

[239] Garcia-Viloca, M., Gao, J., Karplus, M., and Truhlar, D. G. (2004) How enzymes work: analysis by modern rate theory and computer simulations. *Science*, **303**, 186–95.

[240] Benkovic, S. J. and Hammes-Schiffer, S. (2003) A perspective on enzyme catalysis. *Science*, **301**, 1196–202.

[241] Porter, C. T., Bartlett, G. J., and Thornton, J. M. (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic acids research*, **32**, D129–33.

[242] Stepankova, V., et al. (2013) Expansion of access tunnels and active-site cavities influence activity of haloalkane dehalogenases in organic cosolvents. *Chembiochem : a European journal of chemical biology*, **14**, 890–7.

[243] Furnham, N., Holliday, G. L., De Beer, T. A. P., Jacobsen, J. O. B., Pearson, W. R., and Thornton, J. M. (2014) The Catalytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. *Nucleic acids research*, **42**, 1–5.

[244] Gutmanas, A., et al. (2014) PDBe: Protein Data Bank in Europe. *Nucleic acids research*, **42**, D285–91.

[245] Holliday, G. L., Mitchell, J. B. O., and Thornton, J. M. (2009) Understanding the functional roles of amino acid residues in enzyme catalysis. *Journal of molecular biology*, **390**, 560–77.

List of publications included in the habilitation thesis

[VDB] Sehnal, D., Svobodová Vařeková, R., Pravda, L., Ionescu, C.-M., Geidl, S., Horský, V., Jaiswal, D., Wimmerová, M. and Koča, J. (2015) ValidatorDB: database of up-to-date validation results for ligands and non-standard residues from the Protein Data Bank. *Nucleic Acids Res.*, **43**, D369–D375. IF: 9.112 (2014); Authorship: First

[PTQ] Sehnal, D., Pravda, L., Svobodová Vařeková, R., Ionescu, C.-M. and Koča, J. (2015) PatternQuery: web application for fast detection of biomacromolecular structural patterns in the entire Protein Data Bank. *Nucleic Acids Res.*, **43**, W383–W388. IF: 9.112 (2014); Authorship: Co-author

[EL] Geidl, S., Bouchal, T., Raček, T., Svobodová Vařeková, R., Hejret, V., Křenek, A., Abagyan, R. and Koča, J. (2015) High-quality and universal empirical atomic charges for chemoinformatics applications. *J. Cheminform.*, **7**, 59. IF: 4.547 (2014); Authorship: Corresponding

[PS] Geidl, S., Svobodová Vařeková, R., Bendová, V., Petrusek, L., Ionescu, C.-M., Jurka, Z., Abagyan, R. and Koča, J. (2015) How does the methodology of 3D structure preparation influence the quality of pK_a prediction? *J. Chem. Inf. Model.*, **55**, 1088–1097. IF: 3.738 (2014); Authorship: Corresponding

[EAM] Ionescu, C.-M., Sehnal, D., Falginella, F.L., Pant, P., Pravda, L., Bouchal, T., Svobodová Vařeková, R., Geidl, S. and Koča, J. (2015) AtomicChargeCalculator: interactive web-based calculation of atomic charges in large biomolecular complexes and drug-like molecules. *J. Cheminform.*, **7**, 50. IF: 4.547 (2014); Authorship: Co-author

[AN] Pravda, L., Berka, K., Svobodová Vařeková, R., Sehnal, D., Banáš, P., Laskowski, R.A., Koča, J. and Otyepka, M. (2014) Anatomy of enzyme channels. *BMC Bioinformatics*, **15**, 379. IF: 2.576; Authorship: Co-author

[MV] Svobodová Vařeková, R., Jaiswal, D., Sehnal, D., Ionescu, C.-M., Geidl, S., Pravda, L., Horský, V., Wimmerova, M. and Koča, J. (2014) MotiveValidator: interactive web-based validation of ligand and residue structure in biomolecular complexes. *Nucleic Acids Res.*, **42**, W227–W233. IF: 9.112; Authorship: First

[MO2] Sehnal, D., Svobodová Vařeková, R., Berka, K., Pravda, L., Navrátilová, V., Banáš, P., Ionescu, C.-M., Otyepka, M. and Koča, J. (2013) MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *J. Cheminform.*, **5**, 39. IF: 4.540; Authorship: Co-author

[PE] Svobodová Vařeková, R., Geidl, S., Ionescu, C.-M., Skřehota, O., Bouchal, T., Sehnal, D., Abagyan, R. and Koča, J. (2013) Predicting pK_a values from EEM atomic charges. *J. Cheminform.*, **5**, 18. IF: 4.540; Authorship: First

[EB] Ionescu, C.-M., Geidl, S., Svobodová Vařeková, R. and Koča, J. (2013) Rapid calculation of accurate atomic charges for proteins via the Electronegativity Equalization Method. *J. Chem. Inf. Model.*, **53**, 2548–2558. IF: 4.068; Authorship: Corresponding

[BAX] Ionescu, C.-M., Svobodová Vařeková, R., Prehn, J.H.M., Huber, H.J. and Koča,J. (2012) Charge profile analysis reveals that activation of pro-apoptotic regulators Bax and Bak relies on charge transfer mediated allosteric regulation. *PLoS Comput. Biol.*, **8**, e1002565. IF: 4.867; Authorship: Co-author

[MO] Berka, K., Hanák, O., Sehnal, D., Banáš, P., Navratilová, V., Jaiswal, D., Ionescu, C.-M., Svobodová Vařeková, R., Koča, J. and Otyepka, M. (2012) MOLEonline 2.0: interactive web-based analysis of biomacromolecular channels. *Nucleic Acids Res.*, **40**, W222–W227. IF: 8.278; Authorship: Co-author

[SB] Sehnal, D., Svobodová Vařeková, R., Huber, H.J., Geidl, S., Ionescu, C.-M., Wimmerová, M. and Koča, J. (2012) SiteBinder: an improved approach for comparing multiple protein structural motifs. *J. Chem. Inf. Model.*, **52**, 343–359. IF: 4.304; Authorship: Corresponding

[PQ] Svobodová Vařeková, R., Geidl, S., Ionescu, C.-M., Skřehota, O., Kudera, M., Sehnal, D., Bouchal, T., Abagyan, R., Huber, H.J. and Koča, J. (2011) Predicting values of substituted phenols from atomic charges: comparison of different quantum mechanical methods and charge distribution schemes. *J. Chem. Inf. Model.*, **51**, 1795–1806. IF: 4.675; Authorship: First

[EO] Svobodová Vařeková, R., Jiroušková, Z., Vaněk, J., Suchomel, S. and Koča, J. (2007) Electronegativity Equalization Method: parameterization and validation for large sets of organic, organohalogene and organometal molecule. *Int. J. Mol. Sci.*, **8**, 572–582. IF: 0.750; Authorship: First

[EPM] Svobodová Vařeková, R. and Koča, J. (2006) Optimized and parallelized implementation of the Electronegativity Equalization Method and the Atom-Bond Electronegativity Equalization Method. *J. Comput. Chem.*, **27**, 396–405. IF: 4.893; Authorship: First

ValidatorDB: database of up-to-date validation results for ligands and non-standard residues from the Protein Data Bank

Published online 11 November 2014 *Nucleic Acids Research, 2015, Vol. 43, Database issue* ***D369–D375***
doi: 10.1093/nar/gku1118

ValidatorDB: database of up-to-date validation results for ligands and non-standard residues from the Protein Data Bank

David Sehnal[1,2,3,†], Radka Svobodová Vařeková[1,2,†], Lukáš Pravda[1,2], Crina-Maria Ionescu[1], Stanislav Geidl[1,2], Vladimír Horský[3], Deepti Jaiswal[1], Michaela Wimmerová[1,2] and Jaroslav Koča[1,2,*]

[1]CEITEC—Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic, [2]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic and [3]Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic

Received August 29, 2014; Revised October 24, 2014; Accepted October 24, 2014

ABSTRACT

Following the discovery of serious errors in the structure of biomacromolecules, structure validation has become a key topic of research, especially for ligands and non-standard residues. ValidatorDB (freely available at http://ncbr.muni.cz/ValidatorDB) offers a new step in this direction, in the form of a database of validation results for all ligands and non-standard residues from the Protein Data Bank (all molecules with seven or more heavy atoms). Model molecules from the wwPDB Chemical Component Dictionary are used as reference during validation. ValidatorDB covers the main aspects of validation of annotation, and additionally introduces several useful validation analyses. The most significant is the classification of chirality errors, allowing the user to distinguish between serious issues and minor inconsistencies. Other such analyses are able to report, for example, completely erroneous ligands, alternate conformations or complete identity with the model molecules. All results are systematically classified into categories, and statistical evaluations are performed. In addition to detailed validation reports for each molecule, ValidatorDB provides summaries of the validation results for the entire PDB, for sets of molecules sharing the same annotation (three-letter code) or the same PDB entry, and for user-defined selections of annotations or PDB entries.

INTRODUCTION

Validation of biomacromolecular structures has become a very important topic, because some published structures have been found to contain serious errors (1–4). The first step in the validation of biomacromolecules and their complexes is checking the standard building blocks, namely, standard amino acids and nucleotides. The usual procedure is to evaluate specific properties of each residue (e.g. electron density, atom clashes, bond lengths, bond angles, torsion angles, etc.). Various software tools have been developed to perform such analyses, e.g. WHAT_CHECK (5), PROCHECK (6), MolProbity (7) and OOPS (8).

The next key step is the validation of ligands and non-standard residues in biomacromolecular structures, which can be performed in a similar manner as for standard residues (focus on electron density, atom clashes, etc.). An example of software specialized on this type of validation is ValLigURL (9). This approach was also added to several software tools focused on the validation of standard residues (Mogul (10), Coot (11), PHENIX (12)).

A different ligand validation approach, which can be denoted as validation of annotation, was developed later. The goal of this approach is to evaluate if the ligand or non-standard residue is annotated correctly (i.e. if its structure corresponds to the three-letter code it was assigned in the Protein Data Bank (PDB) file format). Specifically, the topology and stereochemistry of the validated molecule are compared to those of a reference molecule (model), and any differences found are reported. The first software tool implementing this methodology has been pdb-care (13), a tool specialized on carbohydrates. The next step has been MotiveValidator (14), which allows validation of all ligands and residues, performs basic validation analyses and reports

*To whom correspondence should be addressed. Tel: +420 54949 4947; Fax: +420 54949 2556; Email: Jaroslav.Koca@ceitec.muni.cz
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

basic warnings (substitutions, foreign atoms, different naming). Because this approach is relatively young, the available tools cover only some of its key topics, leaving many aspects to be explored or improved.

At the same time, with the exponential increase in the size of structural databases, the concept of storing precomputed validation results is becoming increasingly attractive. The first step in this direction was achieved by the PDBREPORT database (5), which is a collection of the outputs from the WHAT_CHECK program. Afterward, the PDB_REDO database (15) of validation results for existing PDB entries was published. Recently, wwPDB included validation reports (16) providing detailed validation results for individual PDB entries directly into their pages.

In our work, we address all challenges described above. We first developed and implemented an improved approach for the validation of annotation, which we subsequently applied to validate all ligands and non-standard residues in the PDB. We then collected all results and built the database ValidatorDB, which offers several advantages over currently available tools (ValLigURL, pdb-care, MotiveValidator):

- ValidatorDB is a database of precomputed validation results for all ligands and non-standard residues in the PDB (except small molecules having fewer than seven heavy atoms).
- ValidatorDB provides summaries of the validation results for the entire PDB, for sets of molecules sharing the same annotation or the same PDB entry, and for user-defined selections of annotations or PDB entries.
- ValidatorDB provides a systematic insight into validation results. The validation analyses are classified into three main categories (Completeness, Chirality and Advanced), each containing several related analyses.
- ValidatorDB classifies the types of chirality errors, enabling the user to distinguish between serious chirality issues and minor inconsistencies.
- ValidatorDB performs novel analyses and can report completely erroneous ligands, alternate conformations, identity with the model molecules, etc. Such analyses can provide information valuable for further data processing.

ValidatorDB obtains correct structures of ligands and non-standard residues from the wwPDB Chemical Components Dictionary (wwPDB CCD) (17), which it uses as reference molecules (models) during validation. ValidatorDB is updated weekly, and is freely available via the Internet at: http://ncbr.muni.cz/ValidatorDB.

VALIDATION ANALYSES

As ValidatorDB implements the approach of validation of annotation, each validated molecule is compared against a model with the same annotation from wwPDB CCD. The validation analyses performed by ValidatorDB cover the main issues which have been observed in the topology (2D structure) and geometry (3D structure) of ligands and non-standard residues, and which are important for their correct annotation. These validation analyses, along with their respective results, can be classified into three categories, namely, Completeness, Chirality and Advanced analyses (Figure 1). If no issues are found during these analyses, the molecule is marked as having complete structure and correct chirality (Figure 1a).

The Completeness analyses attempt to find which atoms are missing (Figure 1b), whether these atoms are part of rings (Figure 1c) or the structure is degenerate, i.e. the molecule contains very severe errors (Figure 1d). These severe errors may refer to residues overlapping in the 3D space, or atoms which are disconnected from the rest of the structure. Validated molecules exhibiting an error in at least one of the Completeness analyses are denoted as incomplete, whereas the remaining molecules are reported as complete.

The Chirality analyses are performed only on complete structures, and aim to evaluate the chirality of each atom in the validated molecule. We distinguish between several types of chirality errors: on carbon atoms (C chirality, Figure 1e), on metal atoms (Metal chirality, Figure 1f), on atoms with four substituents in one plane (Planar chirality, Figure 1g), on atoms connected to at least one substituent by a bond of higher order (High-order chirality, Figure 1h) and the remaining chirality issues (Other chirality). If no issues are detected during the chirality analyses, the validated molecule is marked as having Correct chirality, whereas the remaining molecules are marked as having Wrong chirality. Some types of chirality errors do not constitute real issues, but are artifacts of the automated chirality-determination procedure (i.e. planar chirality and high-order chirality). Therefore, if the validated molecule is found to have these chirality errors, but no other type of chirality issues, the molecule is marked as having 'Correct chirality (tolerant)'.

The Advanced analyses are focused on issues which are not real chemical problems, but which can complicate further processing and exploration of data, and thus should be noted. When issues are found during an advanced analysis, a warning is reported: Substitution, Foreign atom, Different naming, Zero root mean square deviation (RMSD) or Alternate conformations. The Substitution analysis (Figure 1i) reports the replacement of some atom by an atom of a different chemical element. The Foreign atom analysis (Figure 1j) detects atoms which originate from the neighborhood of the validated molecule (i.e. having different PDB residue ID than the majority of the validated molecule), and generally marks sites of intermolecular linkage. The Different naming analysis (Figure 1k) identifies atoms whose name in PDB format are different than the standard convention for the validated molecule. The Zero RMSD analysis reports molecules whose structure is identical (RMSD = 0 Å) to the model from wwPDB CCD. The Alternate conformation analysis informs about the occurrence of alternate conformations in the validated PDB entry.

DATA PREPARATION

Validation procedure for a single molecule

The starting information characterizing the investigated molecule consists of a PDB residue ID, annotation and PDB ID. According to this information, the input motif is

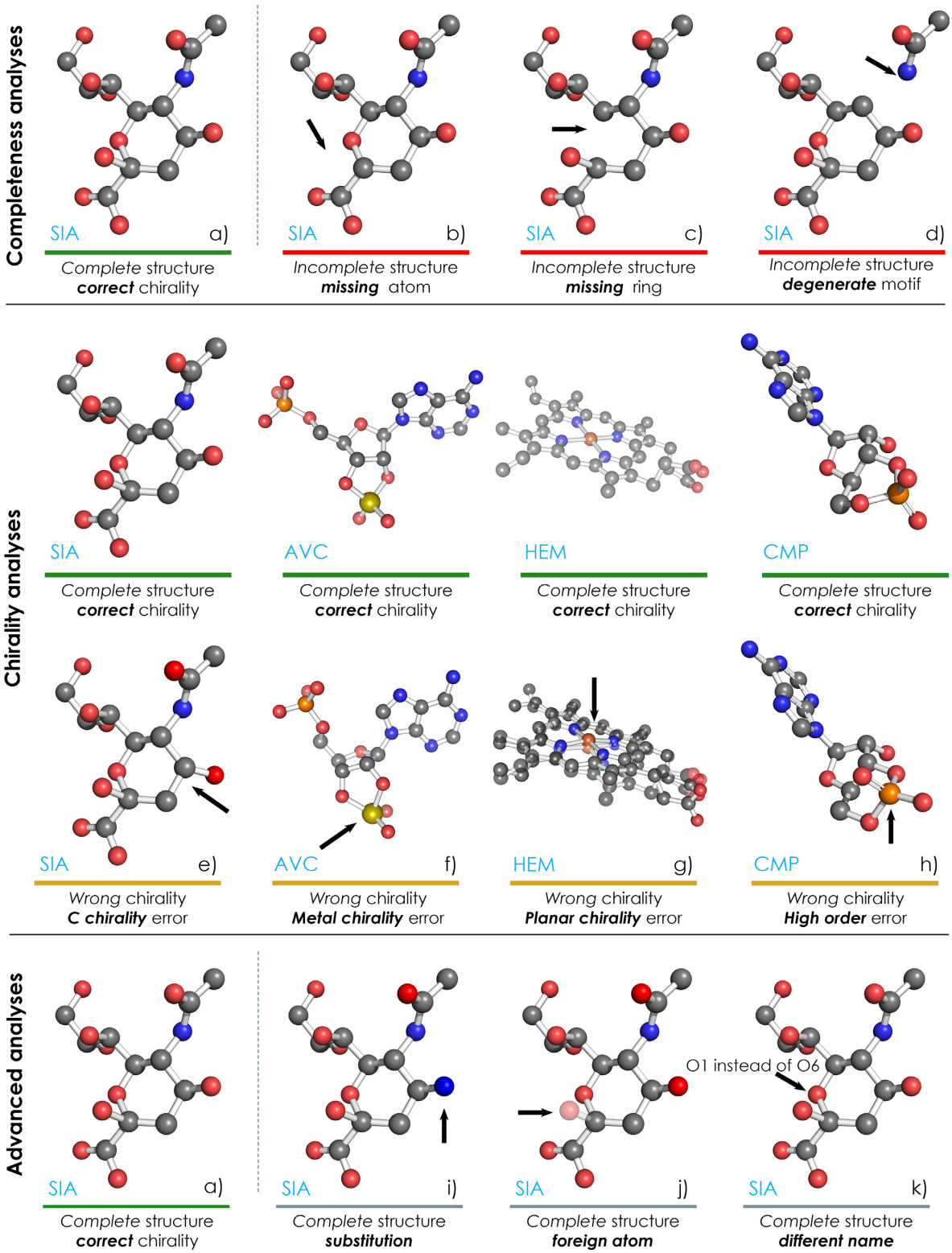


Figure 1. Examples of results provided by different validation analyses. ValidatorDB classifies results into three main categories (Completeness, Chirality, Advanced), each referring to several related analyses. Information about the source of the particular molecules displayed here is given in Supplemental Table S3.

extracted from the PDB entry under investigation. The input motif contains all atoms with the given PDB residue ID, along with their surroundings (atoms within two bonds from any atom of the investigated molecule). The annotation of the molecule is used to identify a suitable model from wwPDB CCD, which then serves as the correct reference structure. The validation proceeds by identifying the maximum common subgraph between the input motif and the model. The atoms of the input motif which belong to this common subgraph make up the validated molecule, which can thus be reliably identified in the PDB entry under investigation. The validated molecule and the model are then superimposed (18) in such a way that their RMSD is minimal. The superimposition provides a pairing (bijection) between atoms in the validated molecule and the corresponding (chemically equivalent) atoms in the model. This bijection allows comparing various properties of each atom in the validated molecule with those of the chemically equivalent atom from the model. All the validation analyses are based on this comparison of atom properties (presence, chirality, element symbol, PDB name, etc.). Other unusual aspects encountered during validation are reported as processing warnings (e.g. which conformer was validated if several conformers were present). A scheme of the validation procedure is depicted in Supplemental Figure S1.

Generation of validation data for all ligands and non-standard residues in the entire Protein Data Bank

The latest versions of the PDB and wwPDB CCD are downloaded once a week, and the following steps ensue.

Obtaining a set of models for validation. Select all models from wwPDB CCD which contain at least seven heavy atoms, excluding the five standard nucleotides and their common deoxy- forms, the 20 standard amino acids and selenomethionine (MSE). ValidatorDB does not focus on the standard building blocks of biomacromolecules because many tools already cover these. Additionally, MSE is also excluded from validation due to its extremely high occurrence in the PDB (markedly higher than other ligands and non-standard residues) and high incidence of circumstantial inclusion in biomacromolecules (to aid X-ray crystallography experiments).

Obtaining validation results for all ligands and non-standard residues in a single PDB entry. For a PDB entry with a given PDB ID, identify the PDB residue IDs of all molecules sharing the annotation with any model obtained in the previous step. Using the procedure described in the first step, detect all validated molecules (via PDB residue ID and corresponding annotation) and compare them to the appropriate models. Collect the validation results for all molecules validated in this PDB entry, and summarize the results of each validation analysis.

Obtaining PDB-wide validation results for each ligand or non-standard residue. For each set of molecules sharing the same annotation, collect validation results from all PDB entries and summarize the results of each validation analysis.

Obtaining a validation overview for the entire PDB. Collect and summarize the results of all types of validation analyses for all validated molecules, irrespective of annotation or PDB entry.

While the algorithm we use for data preparation is generally applicable, highly automated and produces results with straightforward interpretation, it does have limitations. These limitations are described in detail in the Supplementary Material.

DATABASE ORGANIZATION

ValidatorDB provides the user with direct access to a wide range of validation reports, where the results of the validation analyses are organized on several levels. Specifically:

- Validation report for a particular molecule or a set of molecules (accessible via Search → Molecule Identifier), depicted in Figure 2.
- Validation report for a particular PDB entry or a set of PDB entries (accessible via Search → PDB Entry).
- Validation report for a particular annotation or a set of annotations (accessible via Search → Molecule Annotation).
- Table with validation results for all PDB entries (accessible via Details by PDB Entry).
- Table with validation results for all annotations (accessible via Details by Molecule).
- Graph with results of all validation analyses for the entire PDB (accessible via Overview).

A description of the ValidatorDB user interface is provided in the ValidatorDB Wiki Manual.

RESULTS AND DISCUSSION

Validation results for the entire PDB

One of the advantages of ValidatorDB is that it can provide a straightforward overview of the quality of ligands and non-standard residues in the entire Protein Data Bank. The results in Supplemental Table S1 show that currently the PDB (10 August 2014) contains about 9% incomplete ligands and non-standard residues, out of which about 6% miss at least one atom and 2.6% miss rings. Chirality problems occur in less than 8% of the validated molecules. The frequency of basic chirality errors is even lower—only 2.4% of molecules exhibit chirality errors on a carbon atom, and 1.4% on a metal atom. Other chirality issues are generally reported more frequently—i.e. 4.3% of molecules have wrong High-order chirality plus 1.1% wrong Planar chirality, but the majority of these are very probably artifacts (as mentioned in the section Validation Analyses). Therefore, about 83% of validated molecules are complete and have correct chirality. This statement is slightly more optimistic than previous estimations, which are based on the fit to electron density and 3D structure of the ligands and place the expected percentage of erroneous molecules between 20 and 30% (19,20). The situation appears even better if we exclude the chirality errors reported during the Planar and High-order chirality analyses. Specifically, about 88%

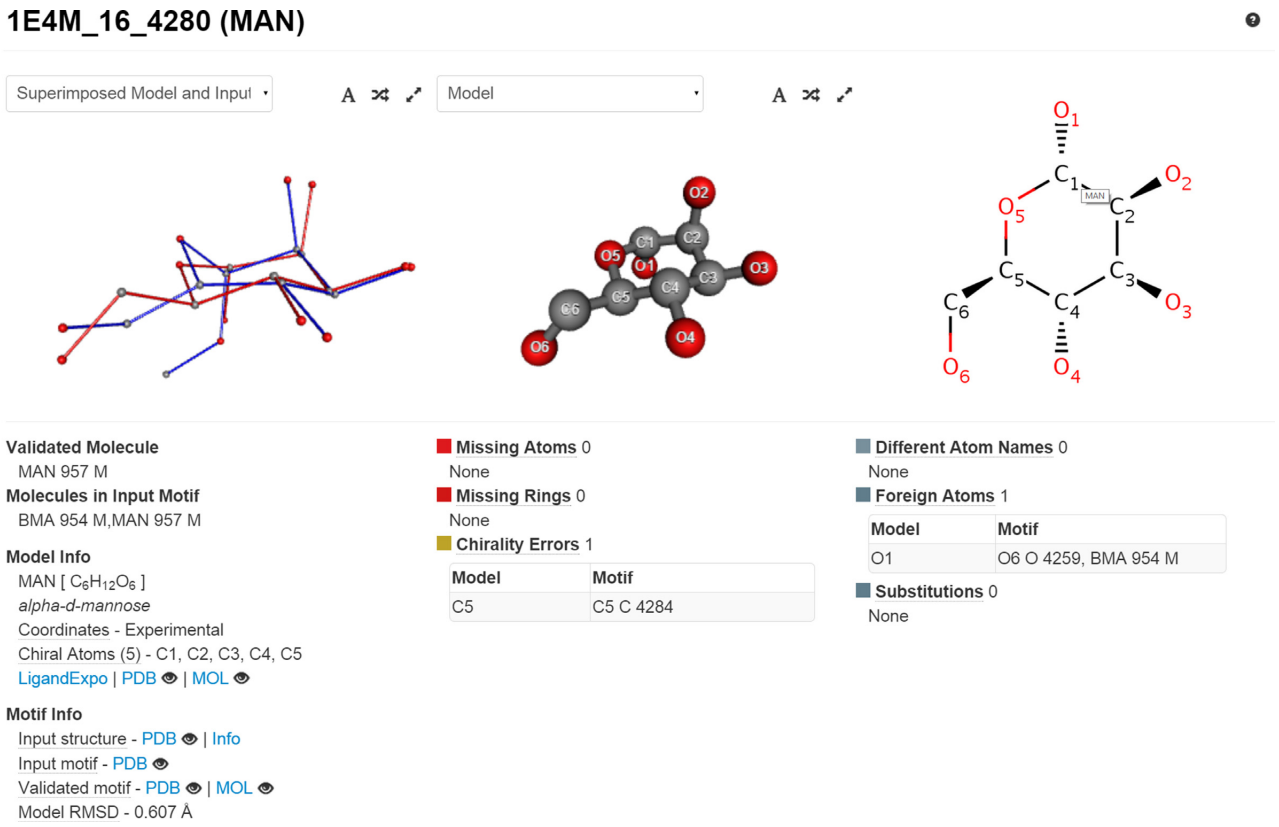


Figure 2. Detailed validation report for the saccharide MAN 957 from PDB entry 1E4M. The structure is complete, but exhibits a chirality error on atom C5. Additionally, the warning of foreign atom at position O1 indicates that this molecule is part of an oligosaccharide chain.

of molecules are complete and have correct chirality for all carbon and metal atoms.

On the other hand, the issues found by the Advanced analyses occur more frequently than completeness and chirality errors. More than 20% of the validated molecules contain substitutions, and about 35% have at least one atom formally located in the neighbor residue. Additionally, 38% contain atoms which are not named in agreement with the standard PDB atom naming convention. Overall, the validation was carried out uneventfully for about 30% of the molecules. While the results of the Advanced analyses have no bearing over the chemical soundness of the validated molecules, they indicate that further, especially automated processing of these structures can be very problematic. Therefore, it is indeed useful to validate the structure of ligands or non-standard residues of interest before performing further investigations, especially where a high degree of automation is involved.

Samples

To show the functionality of ValidatorDB and also the importance of such validation analyses, we selected a few interesting samples and included them in the ValidatorDB web page.

Case studies

One important question is how the quality of the structures varies for different classes of molecules. We have thus designed and conducted several case studies to show how ValidatorDB can answer such questions. We selected the molecules according to a combination of features related to chemical structure, biological function, area of application, availability, etc. The following classes were defined as subsets of models from wwPDB CCD:

- Polycyclic molecules: contain three or more conjugated rings. The molecules containing metals were excluded, as their quality is influenced more by the presence of the metal than by their polycyclic structure.
- Carbohydrates: contain the pyran or furan ring. Molecules containing P (e.g. ATP) were excluded, as their quality is influenced more by the occurrence of phosphate derivatives than by the sugar part.
- Mannose derivatives: subclass of carbohydrates.
- Organometals: contain a metal atom.
- Experimental drugs: described in DrugBank (21) as experimental drugs, i.e. have been shown to bind specific proteins in mammals, bacteria, viruses, fungi or parasites.
- Approved drugs: described in DrugBank as approved drugs, i.e. have received approval in at least one country.

A list of the annotations of the molecules from each class can be found in the Supplementary Material. Summaries of

the validation results for each class are given in Supplemental Table S2.

Compared to the PDB-wide statistics for all ligands and non-standard residues (see above), polycyclic molecules have overall higher quality (higher percentage of molecules with complete structure and correct chirality). Nonetheless, they exhibit more errors in C chirality, probably due to their more complicated, carbon-based scaffolds. Carbohydrate molecules show similar trends as polycyclic molecules, since their structure is also ring-based. However, they exhibit a higher rate of errors in C chirality, a consequence of the fact that they generally contain more chiral atoms. Mannose derivatives play an important role in cell–cell recognition, a biological function which relies heavily on chirality. Therefore, they must have a characteristic structure (determined by chirality) and are also strongly predisposed to have C chirality errors. We found that the percentage of errors in C chirality is over three times higher for mannose derivatives than the PDB-wide evaluation for all ligands and non-standard residues.

Organometals seem to have overall lower quality. Part of the errors is artifacts of our validation algorithm, as such molecules can have very complicated scaffolds (see algorithm limitations in the Supplementary Material). However, the majority of the reported errors are significant, proving that many challenges remain in the field of structure determination for organometals.

On the other hand, the overall quality of the structure of experimental drugs is clearly much higher than the PDB-wide statistics for all ligands and non-standard residues. For approved drugs, i.e. drugs already on the market, the situation is even better. About 95% of these molecules are complete and have correct chirality, a consequence of the fact that markedly more effort is expended in the determination of their structure in biomacromolecular complexes.

CONCLUSIONS

In this article we introduced ValidatorDB, a database of up-to-date validation results for all ligands and non-standard residues from the Protein Data Bank (all molecules with seven or more heavy atoms). The validation of annotation approach implemented here employs correct reference molecules in the form of models from the wwPDB CCD. ValidatorDB offers analyses which cover the main aspects of validation of annotation, by systematically evaluating the completeness, chirality and other features of the validated molecules. ValidatorDB is the only validation tool able to report several types of chirality errors, which allows distinguishing between serious chirality issues and formal inconsistencies. ValidatorDB can further report completely erroneous ligands, alternate conformations, identity with the model, etc. The validation results are organized systematically, from detailed reports for single molecules, to a PDB-wide general summary, and fully customized reports. All results are available in interactive graphical and tabular form via the web interface, and can be readily downloaded in convenient formats.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was funded by the Ministry of Education, Youth and Sports of the Czech Republic [LH13055], the CEITEC - Central European Institute of Technology [CZ.1.05/1.1.00/02.0068] from the European Regional Development Fund, the "Capacities" specific program [286154] and by INBIOR [CZ.1.07/2.3.00/20.0042] from the European Social Fund and the state budget of the Czech Republic. Additional support was provided by the project "Employment of Newly Graduated Doctors of Science for Scientific Excellence" [CZ.1.07/2.3.00/30.0009] co-financed from the European Social Fund and the state budget of the Czech Republic. Funding for open access charge: European Social Fund and the State Budget of the Czech Republic [CZ.1.07/2.3.00/20.0042].

Conflict of interest statement. None declared.

REFERENCES

1. Kleywegt,G.J. (2009) On vital aid: the why, what and how of validation. *Acta Crystallogr. D. Biol. Crystallogr.*, **65**, 134–139.
2. Matthews,B.W. (2007) Five retracted structure reports: inverted or incorrect? *Protein Sci.*, **16**, 1013–1016.
3. Rupp,B. (2012) Detection and analysis of unusual features in the structural model and structure-factor data of a birch pollen allergen. *Acta Crystallogr. Sect. F. Struct. Biol. Cryst. Commun.*, **68**, 366–376.
4. Johnston,C.A., Kimple,A.J., Giguère,P.M. and Siderovski,D.P. (2008) Structure of the parathyroid hormone receptor C terminus bound to the G-protein dimer Gbeta1gamma2. *Structure*, **16**, 1086–1094.
5. Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
6. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
7. Chen,V.B., Arendall,W.B., Headd,J.J., Keedy,D.A., Immormino,R.M., Kapral,G.J., Murray,L.W., Richardson,J.S. and Richardson,D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 12–21.
8. Kleywegt,G.J. and Jones,T.A. (1996) Efficient rebuilding of protein structures. *Acta Crystallogr. D. Biol. Crystallogr.*, **52**, 829–832.
9. Kleywegt,G.J. and Harris,M.R. (2007) ValLigURL: a server for ligand-structure comparison and validation. *Acta Crystallogr. D. Biol. Crystallogr.*, **63**, 935–938.
10. Bruno,I.J., Cole,J.C., Kessler,M., Luo,J., Motherwell,W.D.S., Purkis,L.H., Smith,B.R., Taylor,R., Cooper,R.I., Harris,S.E. *et al.* Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.*, **44**, 2133–2144.
11. Debreczeni,J.É. and Emsley,P. (2012) Handling ligands with Coot. *Acta Crystallogr. D. Biol. Crystallogr.*, **68**, 425–430.
12. Adams,P.D., Afonine,P.V., Bunkóczi,G., Chen,V.B., Davis,I.W., Echols,N., Headd,J.J., Hung,L.-W., Kapral,G.J., Grosse-Kunstleve,R.W. *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 213–221.
13. Lütteke,T. and von der Lieth,C.-W. (2004) pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics*, **5**, 69–74.
14. Vařeková,R.S., Jaiswal,D., Sehnal,D., Ionescu,C.-M., Geidl,S., Pravda,L., Horský,V., Wimmerová,M. and Koča,J. (2014) MotiveValidator: interactive web-based validation of ligand and residue structure in biomolecular complexes. *Nucleic Acids Res.*, **42**, W227–W233.
15. Joosten,R.P., te Beek,T.A.H., Krieger,E., Hekkelman,M.L., Hooft,R.W.W., Schneider,R., Sander,C. and Vriend,G. (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Res.*, **39**, D411–D419.

16. Berman,H.M., Kleywegt,G.J., Nakamura,H. and Markley,J.L. (2014) The Protein Data Bank archive as an open data resource. *J. Comput. Aided Mol. Des.*, **28**, 1009–1014.
17. Feng,Z., Chen,L., Maddula,H., Akcan,O., Oughtred,R., Berman,H.M. and Westbrook,J. (2004) Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, **20**, 2153–2155.
18. Sehnal,D., Vařeková,R.S., Huber,H.J., Geidl,S., Ionescu,C.-M., Wimmerová,M. and Koča,J. (2012) SiteBinder: an improved approach for comparing multiple protein structural motifs. *J. Chem. Inf. Model.*, **52**, 343–359.
19. Lütteke,T., Frank,M. and von der Lieth,C.-W. (2004) Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr. Res.*, **339**, 1015–1020.
20. Liebeschuetz,J., Hennemann,J., Olsson,T. and Groom,C.R. (2012) The good, the bad and the twisted: a survey of ligand geometry in protein crystal structures. *J. Comput. Aided Mol. Des.*, **26**, 169–183.
21. Law,V., Knox,C., Djoumbou,Y., Jewison,T., Guo,A.C., Liu,Y., Maciejewski,A., Arndt,D., Wilson,M., Neveu,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1097.

PatternQuery: web application for fast detection of biomacromolecular structural patterns in the entire Protein Data Bank

Nucleic Acids Research Advance Access published May 26, 2015

doi: 10.1093/nar/gkv561

PatternQuery: web application for fast detection of biomacromolecular structural patterns in the entire Protein Data Bank

David Sehnal[1,2,3,†], Lukáš Pravda[1,2,†], Radka Svobodová Vařeková[1,2], Crina-Maria Ionescu[1] and Jaroslav Koča[1,2,*]

[1]CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic, [2]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic and [3]Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic

Received March 13, 2015; Revised May 01, 2015; Accepted May 17, 2015

ABSTRACT

Well defined biomacromolecular patterns such as binding sites, catalytic sites, specific protein or nucleic acid sequences, etc. precisely modulate many important biological phenomena. We introduce PatternQuery, a web-based application designed for detection and fast extraction of such patterns. The application uses a unique query language with Python-like syntax to define the patterns that will be extracted from datasets provided by the user, or from the entire Protein Data Bank (PDB). Moreover, the database-wide search can be restricted using a variety of criteria, such as PDB ID, resolution, and organism of origin, to provide only relevant data. The extraction generally takes a few seconds for several hundreds of entries, up to approximately one hour for the whole PDB. The detected patterns are made available for download to enable further processing, as well as presented in a clear tabular and graphical form directly in the browser. The unique design of the language and the provided service could pave the way towards novel PDB-wide analyses, which were either difficult or unfeasible in the past. The application is available free of charge at http://ncbr.muni.cz/PatternQuery.

INTRODUCTION

In the past years an overwhelming volume of biomacromolecular structures have been deposited in the worldwide deposition system Protein Data Bank (PDB) (1). The amount of data which was available 20 years ago is nowadays released every week, and this rapid pace is maintained. Small high-resolution protein structures are deposited, as well as extensive ribosomes or viral capsids. The whole scientific community can benefit from this abundance of biomacromolecular structures, being enabled to carry out experiments and analyses which were not feasible before (2–4). Such richness of 3D data accents the immense need for structural bioinformatics tools and services to help in reasoning out a variety of structural properties, which often go hand in hand with biological function.

Presently, various computational tools and frameworks exist for the definition of molecular (sub)structure, such as SMILES (5), MQL (6), or SLN (7), which are mainly focused on small organic compounds. There are also tools that enable the definition and analysis of more general structural patterns, some of which rely on an internal molecular language (8–14). A structural pattern can, in principle, be any part of a biomacromolecule, i.e. protein backbone, ligands or metals together with their binding sites or surroundings, specific amino acids or nucleotide sequences, and sets of atoms or residues satisfying given criteria (distance, composition, intramolecular connectivity, etc.). Nevertheless, these tools are designed to operate either on a low number of structures, or their functionality is focused on very specific and narrow applications. Furthermore, some of the most popular services and databases use structure information for defining or inferring structure-function relationships (15,16). Even critical interaction sites are defined at the primary and secondary structure level (17,18), mainly because of the large structural variation of biomacromolecules. Ultimately, to our knowledge, there is no tool available for the general and systematic description and extraction of 3D structural patterns from biomacromolecules tailored for the mining of structural databases.

In this article, we address the general philosophy of describing 3D structural patterns, and present an approach

*To whom correspondence should be addressed. Tel: +420 54949 4947; Fax: +420 54949 2556; Email: Jaroslav.Koca@ceitec.muni.cz
†These authors contributed equally to the paper as first authors.

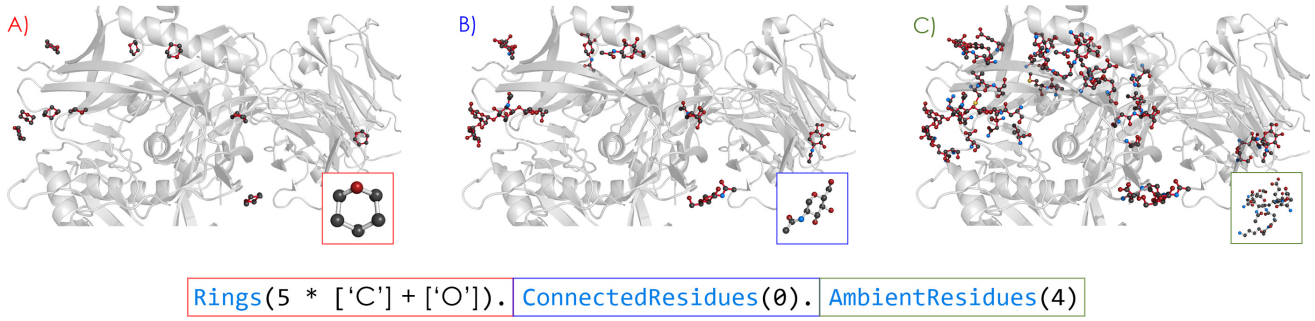


Figure 1. The query recognizes the binding pocket of any residue containing a pyranose moiety in the envelope glycoprotein gp160 from *Human immunodeficiency virus 1* in complex with *Homo sapiens* immunoglobulins (3u7y). One of the recognized patterns is highlighted in the box. (**A**) First, the query identifies a pyranose moiety (a ring composed of 5 carbons and an oxygen atom). (**B**) Then, all residues which include this pattern in their structure are identified. (**C**) Finally, all the residues that are at most 4Å from any of the pyranose containing residues are detected as well. This ensures all the potential coordination partners are recognized properly. The molecules were visualized using PyMOL.

for their effective identification and extraction from individual biomacromolecules, as well as from the PDB archive. This approach is implemented as the user-friendly web service PatternQuery (PQ). The service is built on a simple yet powerful language for the description of any molecular structural patterns based on the nature and relationship between atoms, residues and other structural elements. The unique design of PQ allows the user to simultaneously operate at the primary, secondary and tertiary level of biomacromolecular structure.

The results provided by PQ can serve as a source of input data in further analyses, such as structural and functional assignment of uncharacterized proteins, analysis of newly determined structures, comparative structural analysis, design and engineering of novel functional sites, etc.

DESCRIPTION OF THE TOOL

PatternQuery is an interactive web application for the optimal definition of biomacromolecular structural patterns, followed by their fast detection and extraction from the entire PDB or user defined datasets. These patterns are described by unique expressions based on the Python programming language, which are designed to define biomacromolecular structural patterns based on the nature and relationship between atoms, residues and other structural elements. These expressions define the composition, topology, connectivity, and 3D structure of a pattern. By composing these expressions into a query, 3D structural patterns can be identified inside biomacromolecules. Figure 1 gives the PQ query example that identifies and extracts a 3D pattern made up of a residue containing a pyranose moiety, together with its immediate surroundings.

The PatternQuery application can be used in two modes. The PQ Explorer mode (Figure 2) is useful for real-time investigation of smaller datasets (either user-uploaded or a small subset of the PDB), and tuning the queries prior to searching the whole PDB. The PQ Service mode (Supplementary Figure S1) is optimized for querying the entire PDB archive. Finally, a command-line version of the PQ application is available for processing in-house databases of 3D biomacromolecular structure data.

The PQ web pages contain several interactive guides, which explain the features and give an easy walkthrough the application, along with plenty of tips. Rich documentation is provided as well, in the form of a Wiki user manual with many examples.

PatternQuery workflow

The procedure of using the PatternQuery application involves four steps: (i) query definition; (ii) input data specification; (iii) running the PQ query; (iv) visualization and analysis of retrieved patterns.

(i) Query definition

First, it is necessary to build a query that optimally describes the structural pattern(s) of interest. The PatternQuery language is well documented, and its usage is richly illustrated on many examples and several case studies. Detailed knowledge of the language is not required, since the integrated high performance coding editor (ACE, http://ace.c9.io) provides syntax suggestions and relevant query examples. Multiple queries can be defined for a single run.

(ii) Input data specification

Second, the queried data has to be specified. Small subsets of the PDB, or the user's own datasets can be queried in the PQ Explorer mode. Large custom databases can be queried using the command line version of PatternQuery. In the PQ Service mode, the default queried dataset is a weekly updated mirror of the latest release of the PDB stored in the PDBx/mmCIF file format. Alternatively, a subset of the PDB can be specified based on a list of PDB entry IDs, or on various metadata criteria. By specifying a subset of the PDB as input, it is ensured that only patterns from relevant structures are retrieved, and the query can be executed in a more time efficient manner. For example one may restrict the search only to biomacromolecules including a DNA chain from *Homo sapiens,* determined by X-ray diffraction of resolution better than 2Å and published in the past 3 years.

Optionally, all the patterns identified while running PQ may be subjected to the structural validation of annotation (19). During this process, which is briefly described in the supplementary materials in the section SI Structure Validation, all ligands and non-standard residues larger than six heavy atoms are inspected for their completeness and chirality correctness. Possible discrepancies or structural inconsistencies are highlighted. This may aid further processing of the results by discarding low-quality patterns.

(iii) Running the PQ query

After setup, the specified data set is queried with all the defined PQ expressions. This process involves generating the structure's internal representation, together with proper bond identification based on the intramolecular atomic distances, and then attempting to match the PQ query with any suitable substructure. The theoretical framework behind this process is given in the supplementary materials (Theoretical Background section).
Depending on the complexity of the defined queries and the number of dataset entries, running the queries may take from a few seconds (for a few hundred small to medium-sized entries), up to approximately one hour for 100 000 PDB entries. Most types of queries have $O(N)$ or $O(N \log N)$ time complexity (where N is the number of atoms in the structure), meaning that doubling the number of structures being processed will roughly double the running time. A benchmark of the application is available in the supplementary materials (section Performance Overview).

(iv) Visualization and analysis of retrieved patterns

The PQ results consist of structure files with the patterns, and statistics about their origin and composition. All the results are made available for inspection or download under a unique web address for at least a month, in both the PQ Explorer and PQ Service modes.

The PatternQuery output provides a straightforward and rich report in both tabular and graphical form, including summary and detailed information about each pattern identified. The summary includes the number of detected patterns and PDB entries that the patterns were extracted from, together with possible errors and warnings, often caused by discrepancies either in the biomacromolecular structure, or in the file format. The detailed report provides a pattern view, focused on each individual pattern identified, and a PDB entry view, focused on each PDB entry queried. Additionally, in the PDB entry view, the results for all patterns identified in that particular PDB entry can be accessed together.

Useful statistics in the form of the atom and residue composition are given for each extracted pattern, along with all the metadata from the parent data set entry (PDB entry). These can serve for further filtering of interesting results. Each extracted pattern can be visualized interactively (ChemDoodle, http://www.chemdoodle.com). Optionally, the validation report can be readily accessed.

Limitations

The setup of the PatternQuery web application, particularly in the PQ Service mode, is limited to 10 queries to be executed during a single run. The maximum number of results that can be returned by a single query execution on our server is one million patterns or ten million atoms, whichever is reached first. This limitation is not present in the command line tool. Additional limitations are discussed in detail in the supplementary materials (Limitations section).

RESULTS AND DISCUSSION

We provide two case studies, which demonstrate the possible usage of the PatternQuery web application. Additional biologically relevant examples, together with the corresponding PQ queries, are available on our wiki pages. All the queries used in the case studies can be found in the supplementary materials.

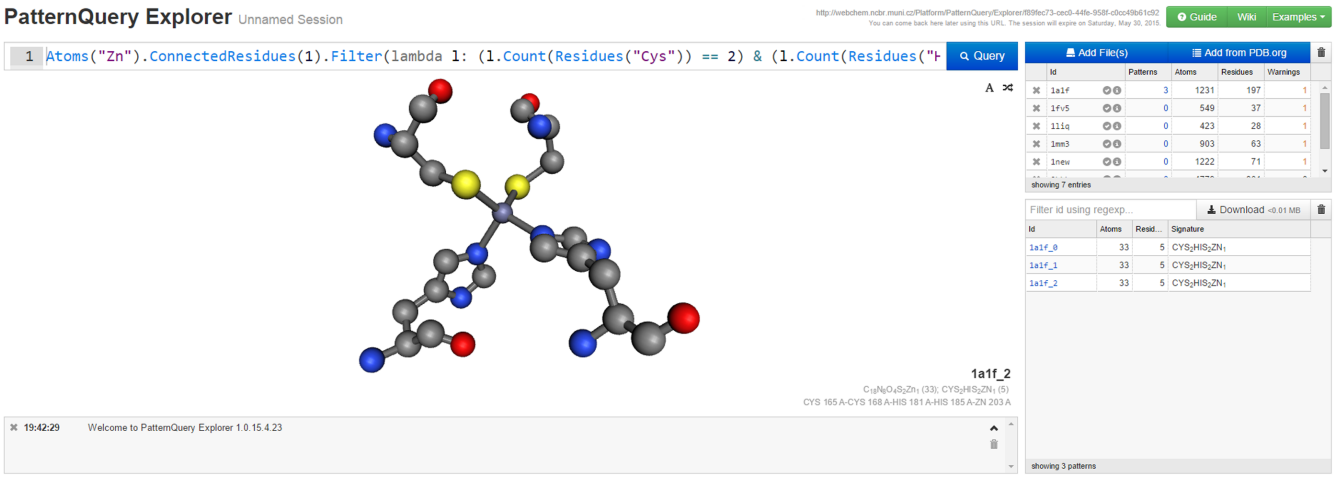


Figure 2. The PatternQuery Explorer mode is tailored for querying smaller user-defined datasets (up to 100 entries) uploaded in one of the supported formats. Additionally, a subset of the PDB archive can be queried as well, based on PDB ID or a variety of metadata.

Case study I - LecB sugar binding sites

Pseudomonas aeruginosa is an opportunistic pathogen associated with a number of chronic infections. This pathogen forms a biofilm enabling it to survive both the response of the host immune system, and antibiotic treatment (20). One of the cornerstones of biofilm formation, in the case of *P. aeruginosa,* is the presence of sugar-binding proteins on the outer cell membrane — LecA (PA-IL) and LecB (PA-IIL). Their inhibition is considered to be a promising approach for anti-pseudomonadal treatment (21).

LecB binds with the highest affinity to L-fucosides and D-mannosides (22), however, other monosaccharides are recognized as well (23). The sugar-binding domain is calcium dependent, with two calcium ions stabilizing the binding site. We employed PQ in the discovery of sugar binding sites of similar geometry as the tetrameric LecB entry in the PDB. Specifically, we have searched for 2 calcium ions at most 4Å apart, and all the residues with direct interaction with either of these ions. Furthermore, just the molecular patterns containing a residue with a furan or pyran ring were preserved. The complete PQ query which identifies such patterns is given as SI Query 1. Due to the fact that the sugar-binding domain is calcium dependent, we were able to restrict the search only to the biomacromolecules having a calcium ion in their structure, and containing a pyranose or furanose moiety (3074 PDB entries as of 25.4.2015), which tremendously reduced query-running time. The initial analysis of the PDB archive revealed 355 different patterns originating from 231 PDB entries. However, the majority of the sugar moieties originated from nucleotides. To filter them out, a simple filter was employed (SI Query 2), which provided 108 distinct patterns originating from 36 PDB entries of 7 different organisms. The majority of them originated from *P. aeruginosa*, however other pathogens such as *R. solanacearum*, *B. cenocepacia* or *C. violaceum* were identified among the organisms of origin. The sugar-binding domain in 87 of the patterns are composed of 3x Asp, 2x Asn and Glu and Gly residues, which is the binding site referred to as the sugar binding motif in the literature (24) for a total of 24 PDB entries from 3 organisms. In 12 further patterns a glycine residue was not present due to the fact that the structure stored in the PDB is only the asymmetric unit, rather than the expected biological unit, which is a tetramer. Finally, the remaining 9 patterns, originating from 6 different pectate lyase (EC: 4.2.2.2) structures, exhibited a different binding motif in comparison to the LecB protein. These patterns contained α-D-galactopyranuronic acid and its derivatives rather than a fucose or mannose derivative. A detailed list of these sugar ligands is given in the Supplementary Tables S1 and S2.

Finally, the quality of the 3D structure of the patterns was examined. A total of 9 patterns originating from 3 PDB entries exhibited a serious structural issue, i.e. half of the α-L-fucose ligands in complex with the 1oxc PDB entry exhibit incorrect chirality at the C1 carbon atom. The details of this analysis can be found in the supplementary materials (SI Query Validation 1).

Case study II - C_2H_2 zinc fingers

The class of zinc finger DNA-binding proteins is the most abundant across all biology (25). They fulfill a remarkable range of diverse functions, including DNA recognition, transcriptional activation, regulation of apoptosis or lipid binding (25). Due to their specificity and modular architecture, they often serve as a rational engineering target for binding a wide range of DNA sequences to activate, repress, cut or paste genes (26). The classical C_2H_2 zinc finger domain is composed of a simple ββα fold, which is stabilized by a zinc ion coordinated by two histidine and two cysteine residues. The fold is often described by the pattern of X_2-C-X_{2-4}-C-X_{12}-H-X_{3-5}-H, where X stands for any amino acid, C is cysteine and H is histidine. Nevertheless, atypical variations also exist, which differ from the consensus profile (27) (e.g. UniProt ID (28): P47043). The X_{12} region of the consensus profile is usually further decomposed into the sequence X_3-[F|Y]-X_5-Ψ-X_2, where [F|Y] represents either a phenylalanine or tyrosine residue, and Ψ denotes a hydrophobic residue (29).

We have queried the whole PDB archive (access date 25.4.2015) using several different PQ queries. At first, we searched just for patterns with primary sequences which satisfy the basic consensus profile of the typical C_2H_2 zinc finger domain, without further specification of the X_{12} region (SI Query 3). We identified 595 patterns in 342 different PDB entries. The results of such a query will inevitably be plagued by a number of false positive hits, i.e. patterns satisfying the primary sequence criteria, but which are not zinc fingers. This is due to the fact that no further checks for the presence of a zinc ion or stabilizing residues were included in the query. Closer inspection of the results revealed that the above-defined primary sequence corresponds not only to the C_2H_2 zinc finger fold, but also to a variety of fumarate reductases and hydrolases. In order to filter out these false positive hits, we adjusted the query so that the pattern must contain a zinc ion stabilized by two cysteine and two histidine residues from the consensus profile (SI Query 4). This final query resulted in 461 different patterns originating from 278 PDB entries. The majority of the results (356 patterns in 239 PDB entries) also satisfied the special pattern of the X_{12} region between the second cysteine and the first histidine (SI Query 5). The largest number of structures was isolated from Eukaryotes, mainly *Homo sapiens,* and determined by solution nuclear magnetic resonance spectroscopy. However, a few structures originating from viruses and bacteria were found as well. No residues relevant for validation were detected inside the input patterns, and therefore validated.

Furthermore, it has been reported that the zinc finger fold may also be stabilized by other metals (30). We have modified the query so that possible substitutions of zinc with other metals can also be considered (SI Query 6). Running this query returned five additional patterns from two PDB entries, where the zinc ion was substituted by cobalt (31) and cadmium (32). Although the cobalt-binding protein contains 5 zinc finger domains, just 4 patterns were identified, due to the alternate primary sequence in one of the patterns. These primary sequence modifications can be ac-

counted for by modifying the regular expression in the PQ query.

CONCLUSION

In this article, we presented PatternQuery, a novel web application for rapid definition and extraction of 3D structural patterns from the entire PDB. The web application is easy to use and platform-independent. Results are presented in a clear graphical and tabular form. Rich documentation regarding both the underlying language and the features of the web application, along with several biologically relevant case studies are available at http://ncbr.muni.cz/PatternQuery.

The innovative approach described in the present study enables mining large databases (entire PDB or in-house structural databases), a task which was unfeasible in the past, or was difficult for patterns with more complex structure.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

This work was funded by the Ministry of Education,Youth and Sports of the Czech Republic [contract number LH13055], the European Community's Seventh Framework Programme [CZ.1.05/1.1.00/02.0068] from the European Regional Development Fund, and the Grant Agency of the Czech Republic [14-29577S]. Support to CMI was provided via the project "Employment of Newly Graduated Doctors of Science for Scientific Excellence" [CZ.1.07/2.3.00/30.0009], co-financed from the European Social Fund and the state budget of the Czech Republic. Funding for open access charge: Institutional budget of the National Centre for Biomolecular Research, Masaryk University, Czech Republic.

Conflict of interest statement. None declared.

REFERENCES

1. Gutmanas,A., Alhroub,Y., Battle,G.M., Berrisford,J.M., Bochet,E., Conroy,M.J., Dana,J.M., Fernandez Montecelo,M.A., van Ginkel,G., Gore,S.P. *et al.* (2014) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*, **42**, D285–D291.
2. Smith,K.P., Gifford,K.M., Waitzman,J.S. and Rice,S.E. (2014) Survey of phosphorylation near drug binding sites in the Protein Data Bank (PDB) and their effects.*Proteins Struct. Funct. Bioinforma.*, **83**, 25–36.
3. Gavenonis,J., Sheneman,B.A., Siegert,T.R., Eshelman,M.R. and Kritzer,J.A. (2014) Comprehensive analysis of loops at protein-protein interfaces for macrocycle design. *Nat. Chem. Biol.*, **10**, 1–8.
4. Steinkellner,G., Gruber,C.C., Pavkov-Keller,T., Binter,A., Steiner,K., Winkler,C., Lyskowski,A., Schwamberger,O., Oberer,M., Schwab,H. *et al.* (2014) Identification of promiscuous ene-reductase activity by mining structural databases using active site constellations. *Nat. Commun.*, **5**, 4150.
5. Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
6. Proschak,E., Wegner,J.K., Schüller,A., Schneider,G. and Fechner,U. (2007) Molecular query language (MQL)–a context-free grammar for substructure matching. *J. Chem. Inf. Model.*, **47**, 295–301.
7. Homer,R.W., Swanson,J., Jilek,R.J., Hurst,T. and Clark,R.D. (2008) SYBYL line notation (SLN): A single notation to represent chemical structures, queries, reactions, and virtual libraries. *J. Chem. Inf. Model.*, **48**, 2294–2307.
8. Humphrey,W., Dalke,A. and Schulten,K. (1996) VMD: Visual molecular dynamics. *J. Mol. Graph.*, **14**, 33–38.
9. Biasini,M., Schmidt,T., Bienert,S., Mariani,V., Studer,G., Haas,J., Johner,N., Schenk,A.D., Philippsen,A. and Schwede,T. (2013) OpenStructure: an integrated software framework for computational structural biology. *Acta Crystallogr. D. Biol. Crystallogr.*, **69**, 701–709.
10. Kalev,I., Mechelke,M., Kopec,K.O., Holder,T., Carstens,S. and Habeck,M. (2012) CSB: a Python framework for structural bioinformatics. *Bioinformatics*, **28**, 2996–2997.
11. The PyMOL Molecular Graphics System. Version 1.7.4, Schrödinger, LLC.
12. Täubig,H., Buchner,A. and Griebsch,J. (2006) PAST: fast structure-based searching in the PDB. *Nucleic Acids Res.*, **34**, W20–W23.
13. Nadzirin,N., Willett,P., Artymiuk,P.J. and Firdaus-Raih,M. (2013) IMAAAGINE: a webserver for searching hypothetical 3D amino acid side chain arrangements in the Protein Data Bank. *Nucleic Acids Res.*, **41**, W432–W440.
14. Samson,A.O. and Levitt,M. (2009) Protein segment finder: an online search engine for segment motifs in the PDB. *Nucleic Acids Res.*, **37**, D224–D228 .
15. Andreeva,A., Howorth,D., Chothia,C., Kulesha,E. and Murzin,A.G. (2014) SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.*, **42**, D310–D314.
16. Sillitoe,I., Lewis,T.E., Cuff,A., Das,S., Ashford,P., Dawson,N.L., Furnham,N., Laskowski,R.A., Lee,D., Lees,J.G. *et al.* (2014) CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.*, **43**, D376–D381.
17. Furnham,N., Holliday,G.L., De Beer,T.A.P., Jacobsen,J.O.B., Pearson,W.R. and Thornton,J.M. (2014) The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.*, **42**, 1–5.
18. Higurashi,M., Ishida,T. and Kinoshita,K. (2009) PiSite: a database of protein interaction sites using multiple binding states in the PDB. *Nucleic Acids Res.*, **37**, D360–D364.
19. Sehnal,D., Svobodová Vařeková,R., Pravda,L., Ionescu,C.-M., Geidl,S., Horský,V., Jaiswal,D., Wimmerová,M. and Koča,J. (2015) ValidatorDB: database of up-to-date validation results for ligands and non-standard residues from the Protein Data Bank. *Nucleic Acids Res.*, **43**, D369–D375.
20. Hauck,D., Joachim,I., Frommeyer,B., Varrot,A., Philipp,B., Möller,H.M., Imberty,A., Exner,T.E. and Titz,A. (2013) Discovery of two classes of potent glycomimetic inhibitors of Pseudomonas aeruginosa LecB with distinct binding modes. *ACS Chem. Biol.*, **8**, 1775–1784.
21. Ernst,B. and Magnani,J.L. (2009) From carbohydrate leads to glycomimetic drugs. *Nat. Rev. Drug Discov.*, **8**, 661–677.
22. Winzer,K., Falconer,C., Garber,N.C., Diggle,S.P., Camara,M. and Williams,P. (2000) The Pseudomonas aeruginosa lectins PA-IL and PA-IIL are controlled by quorum sensing and by RpoS. *J. Bacteriol.*, **182**, 6401–6411.
23. Sabin,C., Mitchell,E.P., Pokorná,M., Gautier,C., Utille,J.-P., Wimmerová,M. and Imberty,A. (2006) Binding of different monosaccharides by lectin PA-IIL from Pseudomonas aeruginosa: thermodynamics data correlated with X-ray structures. *FEBS Lett.*, **580**, 982–987.
24. Mitchell,E., Houles,C., Sudakevitz,D., Wimmerova,M., Gautier,C., Pérez,S., Wu,A.M., Gilboa-Garber,N. and Imberty,A. (2002) Structural basis for oligosaccharide-mediated adhesion of Pseudomonas aeruginosa in the lungs of cystic fibrosis patients. *Nat. Struct. Biol.*, **9**, 918–921.
25. Laity,J.H., Lee,B.M. and Wright,P.E. (2001) Zinc finger proteins: new insights into structural and functional diversity. *Curr. Opin. Struct. Biol.*, **11**, 39–46.
26. Gersbach,C.A., Gaj,T. and Barbas,C.F. (2014) Synthetic zinc finger proteins: the advent of targeted gene regulation and genome modification technologies. *Acc. Chem. Res.*, **47**, 2309–2318.
27. Wang,Z., Feng,L.S., Matskevich,V., Venkataraman,K., Parasuram,P. and Laity,J.H. (2006) Solution structure of a Zap1 zinc-responsive

domain provides insights into metalloregulatory transcriptional repression in Saccharomyces cerevisiae. *J. Mol. Biol.*, **357**, 1167–1183.
28. Activities at the Universal Protein Resource (UniProt). (2014) *Nucleic Acids Res.*, **42**, D191–D198.
29. Pabo,C.O., Peisach,E. and Grant,R.A. (2001) Design and selection of novel Cys2His2 zinc finger proteins. *Annu. Rev. Biochem.*, **70**, 313–340.
30. Hartwig,A. (2001) Zinc finger proteins as potential targets for toxic metal ions: differential effects on structure and function. *Antioxid. Redox Signal.*, **3**, 625–634.
31. Pavletich,N.P. and Pabo,C.O. (1993) Crystal structure of a five-finger GLI-DNA complex: new perspectives on zinc fingers. *Science*, **261**, 1701–1707.
32. Malgieri,G., Zaccaro,L., Leone,M., Bucci,E., Esposito,S., Baglivo,I., Del Gatto,A., Russo,L., Scandurra,R., Pedone,P.V. *et al.* (2011) Zinc to cadmium replacement in the A. thaliana SUPERMAN Cys 2His 2 zinc finger induces structural rearrangements of typical DNA base determinant positions. *Biopolymers*, **95**, 801–810.

High-quality and universal empirical atomic charges for chemoinformatics applications

Geidl *et al. J Cheminform (2015) 7:59*
DOI 10.1186/s13321-015-0107-1

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

High-quality and universal empirical atomic charges for chemoinformatics applications

Stanislav Geidl[1†], Tomáš Bouchal[1†], Tomáš Raček[1,2†], Radka Svobodová Vařeková[1*], Václav Hejret[1], Aleš Křenek[3], Ruben Abagyan[4] and Jaroslav Koča[1*]

Abstract

Background: Partial atomic charges describe the distribution of electron density in a molecule and therefore provide clues to the chemical behaviour of molecules. Recently, these charges have become popular in chemoinformatics, as they are informative descriptors that can be utilised in pharmacophore design, virtual screening, similarity searches etc. Especially conformationally-dependent charges perform very successfully. In particular, their fast and accurate calculation via the Electronegativity Equalization Method (EEM) seems very promising for chemoinformatics applications. Unfortunately, published EEM parameter sets include only parameters for basic atom types and they often miss parameters for halogens, phosphorus, sulphur, triple bonded carbon etc. Therefore their applicability for drug-like molecules is limited.

Results: We have prepared six EEM parameter sets which enable the user to calculate EEM charges in a quality comparable to quantum mechanics (QM) charges based on the most common charge calculation schemes (i.e., MPA, NPA and AIM) and a robust QM approach (HF/6-311G, B3LYP/6-311G). The calculated EEM parameters exhibited very good quality on a training set ($R^2 > 0.9$) and also on a test set ($R^2 > 0.93$). They are applicable for at least 95 % of molecules in key drug databases (DrugBank, ChEMBL, Pubchem and ZINC) compared to less than 60 % of the molecules from these databases for which currently used EEM parameters are applicable.

Conclusions: We developed EEM parameters enabling the fast calculation of high-quality partial atomic charges for almost all drug-like molecules. In parallel, we provide a software solution for their easy computation (http://ncbr.muni.cz/eem_parameters). It enables the direct application of EEM in chemoinformatics.

Keywords: Partial atomic charges, Electronegativity Equalization Method, EEM, Quantum mechanics, QM, Drug-like molecules

Background

Partial atomic charges are real numbers describing the distribution of electron density in a molecule, thus providing clues as to the chemical behaviour of molecules. The concept of charges began to be used in physical chemistry and organic chemistry. Afterwards, partial atomic charges were adopted by computational chemistry and molecular modelling, where they serve for calculating electrostatic interactions, describe the reactivity of the molecule etc. Specifically, they are applied in molecular dynamics, docking, conformational searches, binding site predictions etc. Recently, partial atomic charges also became popular in chemoinformatics, as they proved to be informative descriptors for QSAR and QSPR modelling [1–9] and for other applications [10–12]; they can be utilised in pharmacophore design [13–15], virtual

*Correspondence: radka.svobodova@ceitec.muni.cz; jkoca@chemi.muni.cz
†Stanislav Geidl, Tomáš Bouchal and Tomáš Raček are joint first authors
[1] National Centre for Biomolecular Research, Faculty of Science and CEITEC, Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic
Full list of author information is available at the end of the article

screening [16–18], similarity searches [19–21], molecular structure comparison [22–24] etc.

The partial atomic charges cannot be determined experimentally or derived straightforwardly from the results of quantum mechanics (QM), and many different methods have been developed for their calculation. The most common method for charge calculation is an application of the QM approach and afterwards the utilisation of a charge calculation scheme. Charge calculation schemes can be based on orbital-based population analysis, on wave-function-dependent physical observables or on reproducing charge-dependent observables. Examples of orbital-based population analyses are Mulliken population analysis (MPA) [25, 26], Löwdin population analysis [27] and Natural population analysis (NPA) [28, 29]. Wave-function-dependent physical observables are used in the atoms-in-molecules (AIM) approach [30, 31], Hirshfeld population analysis [32–34], CHELPG [35] and Merz-Singh-Kollman (MK) [36, 37] method. The reproduction of charge-dependent observables is applied in the CM1, CM2, CM3, CM4, and CM5 approaches [38, 39].

Unfortunately, QM charge calculation approaches are very time-consuming. A markedly faster alternative is to employ empirical charge calculation approaches, which can also provide high-quality charges. These approaches can be divided into conformationally-independent, which are based on 2D structure (e.g., Gasteiger's and Marsili's PEOE [40, 41], GDAC [42], KCM [43], DENR [44]) and conformationally-dependent, calculated from 3D structure (e.g., EEM [45], QEq [46] or SQE [47, 48]). We would like to highlight that conformationally-dependent charges are considered to be more suitable for chemoinformatics applications [1–3, 7, 12, 20]. The reason is that these charges contain extensive information not only about chemical surrounding of atoms, i.e., its topology (2D structure based charges) but also geometry and "chemical quality" of the surrounding. Such information is missing, for example, in force field charges which use averaged atomic charges from large sets of structures. Therefore we only focus on conformationally-dependent atomic charges.

Electronegativity equalization method (EEM) is the most frequently used conformationally-dependent empirical charge calculation approach. It calculates charges using the following system of linear equations:

$$\begin{pmatrix} B_1 & \frac{\kappa}{R_{1,2}} & \cdots & \frac{\kappa}{R_{1,N}} & -1 \\ \frac{\kappa}{R_{2,1}} & B_2 & \cdots & \frac{\kappa}{R_{2,N}} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\kappa}{R_{N,1}} & \frac{\kappa}{R_{N,2}} & \cdots & B_N & -1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ \bar{\chi} \end{pmatrix} = \begin{pmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_N \\ Q \end{pmatrix} \quad (1)$$

where q_i is the charge of an atom i; $R_{i,j}$ is the distance between atoms i and j; Q is the total charge of the molecule; N is the number of atoms in the molecule; κ is the molecular electronegativity, and A_i, B_i and κ are empirical parameters. The parameters A_i and B_i vary for individual atom types, where atom type is a combination of element type and maximal bond order of the atom i. For example, atom type C2 means that the atom is carbon and it creates at least one double bond with its neighbors. An atom X in the aromatic ring is therefore also included into X2 atom type. The parameters A_i, B_i and κ are molecule independent and they are calculated from QM atomic charges by a process of EEM parameterization [49]. EEM is not only a fast charge calculation approach, but it can also provide highly accurate charges, i.e., they can mimic the QM charges for which EEM has been parameterized. On the other hand, EEM charges can be outperformed in certain situations. Specifically, QEq showed better agreement with experimental dipole moments [46] and SQE is presented as an extension of the EEM to obtain the correct size-dependence of the molecular polarizability [47]. But this drawback is compensated by a fact that the quality of EEM charges was documented by many successful applications [2, 3, 50–55] and they are clearly the most cited empirical conformationally-dependent charges.

Therefore, many EEM parameter sets for various QM charge calculation approaches were published later or recently (see Table 1). In parallel, a few freely available software tools also include an EEM charge calculation method (see Table 2).

EEM recently began to be also used in chemoinformatics, giving very promising results [1–3, 64, 65]. Because of their rapid calculation, they can be easily computed for large sets of molecules (e.g., drug-like compounds). Unfortunately, a broader utilisation of EEM charges in chemoinformatics is now limited by the fact that available EEM parameter sets can only cover part of common organic molecules, as they only contain the parameters for some elements and certain bond orders (Table 1). For the above reasons, our aim with this work is to provide EEM parameter sets that cover most of the drug-like molecules and with accuracy comparable to QM charges. Specifically, we have parameterized EEM for frequently used charge calculation schemes, high enough QM theory levels and a large basis set. Afterwards, we compared the coverage and quality of our EEM parameter sets with previously published EEM parameter sets (see Table 1) and with EEM parameter sets embedded in software tools (see Table 2). Additionally, we have prepared a software solution, enabling the user to easily calculate EEM charges via our EEM parameters.

Table 1 Summary information about published EEM parameters evaluated in this study

QM theory Level + basis set	Charge calc. scheme	EEM parameter set name	Published by	Elements and bond orders included[†]
HF/STO-3G	MPA	Baek1991	Baekelandt et al. [56]	C, O, N, H, P, Al, Si
		Svob2007_cbeg2	Svobodova et al. [49]	C1, C2, O, N1, N2, H, S1
		Svob2007_cmet2	Svobodova et al. [49]	C1, C2, O, N1, N2, H, S1, Fe, Zn
		Svob2007_chal2	Svobodova et al. [49]	C1, C2, O, N1, N2, H, S1, Br, Cl, F, I
		Svob2007_hm2	Svobodova et al. [49]	C1, C2, O, N1, N2, H, S1, F, Cl, Br, I, Fe, Zn
HF/6-31G*	MK	Jir2008_hf	Jirouskova et al. [57]	C1, C2, O, N1, N2, H, S1, F, Cl, Br, Zn
B3LYP/6-31G*	MPA	Bult2002_mpa	Bultinck et al. [58]	C, O, N, H, F
	NPA	Bult2002_npa	Bultinck et al. [58]	C, O, N, H, F
		Ouy2009[‡]	Ouyang et al. [59]	C, O, N, H
		Ouy2009_elem	Ouyang et al. [59]	C, O, N, H
	Hir.	Bult2002_hir	Bultinck et al. [58]	C, O, N, H, F
	MK	Bult2002_mk	Bultinck et al. [58]	C, O, N, H, F
		Jir2008_mk	Jirouskova et al. [57]	C1, C2, O, N1, N2, H, S1, F, Cl, Br, Zn
	CHELPG	Bult2002_che	Bultinck et al. [58]	C, O, N, H, F
	AIM	Bult2004_aim	Bultinck et al. [60]	C, O, N, H, F

[†] An element symbol with no further information (e.g., C) means that the EEM parameters are available for this element bound by all possible bond orders. The element symbol followed by a number (e.g., C1) means that the EEM parameters are only available for this element bound by a bond with an order described using this number

[‡] For this parameter set, C1 represents sp^3 hybridization, C2 sp^2 hybridization, C3 sp hybridization, etc.

Table 2 Information about freely available software tools enabling EEM charge calculation

Software	EEM parameters used by a software
OpenBabel [61]	It contains the embedded EEM parameter set Bult2002_mpa, which was parameterized for B3LYP/6-31G*/MPA charges. It does not allow any other EEM parameter set to be used
Balloon [23]	It contains an embedded EEM parameter set published by Puranen et al. [62], which was calculated by fitting to the MEP field. Balloon's developers claim that the EEM charges calculated via Balloon should be comparable to B3LYP/cc-pVTZ/MPA. It does not allow any other EEM parameter set to be used
EEM SOLVER [63]	It allows the use of any input EEM parameter sets provided by the user. It does not contain any embedded EEM parameter sets

Methods

EEM parameterization (step 1)

All the steps performed during our work are depicted in Fig. 1a. The most challenging part of our work was the EEM parameterization. This step required several tasks (see Fig. 1b) and the quality of the calculated EEM parameters sets depends on the proper accomplishment of all these tasks.

EEM parameterization: selection of atom types to be parameterized

Our goal is to provide EEM parameter sets applicable for most common drug-like molecules. Therefore, we provide EEM parameters for the majority of atom types occurring in these molecules. These atom types are summarized in Table 3 (columns 1–3).

EEM parameterization: preparation of the training set

Our training set contains the 3D structures of 4475 distinct small organic molecules. The molecules were obtained from the DTP NCI database [66] and their 3D structures were generated with CORINA 3.60 [67], without any further geometry optimization. The DTP NCI database collects compounds tested as anticancer drugs (with positive or negative results), therefore it is a database of common drug-like molecules. The training set was created in such a way that each selected atom type is contained in at least 100 molecules. The occurrences of individual atom types in the training set are summarized in Table 3. The list of training set molecules, including their NSC numbers and summary formulas, can be found in (Additional file 1: Table S1).

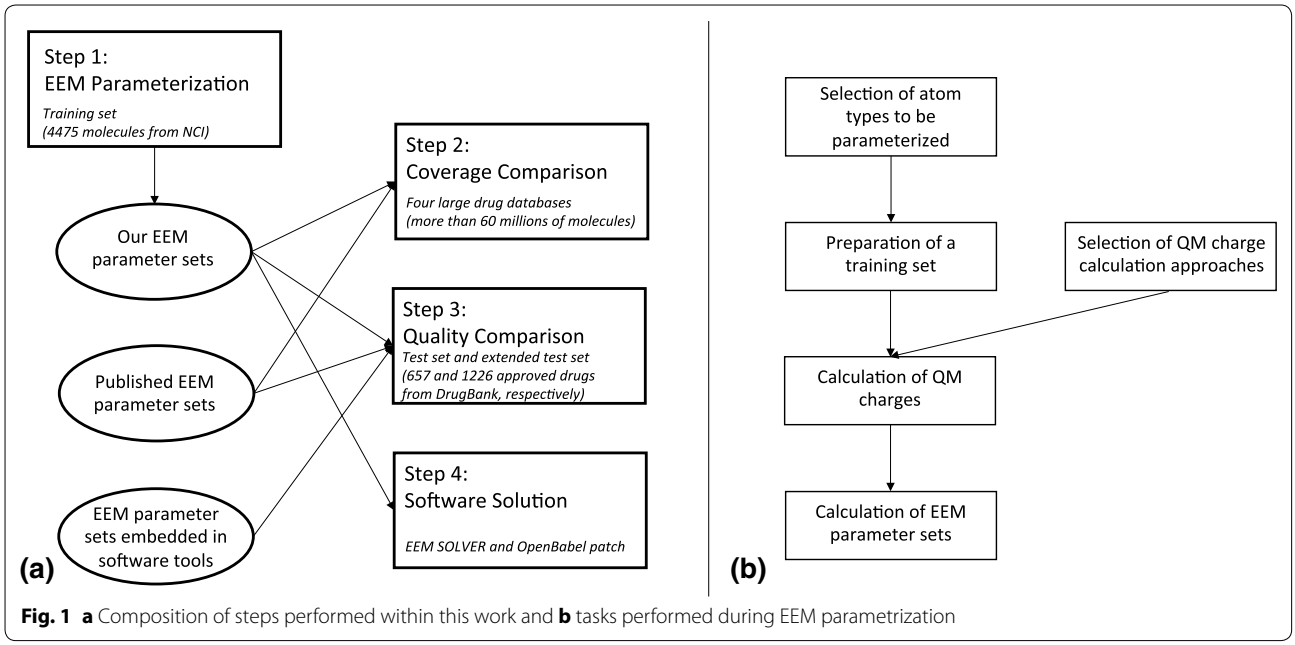


Fig. 1 **a** Composition of steps performed within this work and **b** tasks performed during EEM parametrization

Table 3 Occurrence of atom types in the training set

Denotation of atom type	Element symbol	Maximal bond order	Number of atoms with this atom type in the training set	Number of molecules containing this atom type in the training set
H1	H	1	57,119	4442
C1	C	1	15,220	3447
C2		2	38,097	4149
C3		3	345	266
N1	N	1	4151	2483
N2		2	3383	1879
N3		3	345	266
O1	O	1	5016	2525
O2		2	5793	3069
F1	F	1	938	395
P1	P	1	153	143
P2		2	251	213
S1	S	1	1034	770
S2		2	1391	1211
Cl1	Cl	1	1084	676
Br1	Br	1	336	261
I1	I	1	1734	1365
Total	–	–	136,390	4475

EEM parameterization: selection of QM charge calculation approach

We performed the EEM parameterization for two QM theory levels (B3LYP and HF), one basis set (6-311G) and three charge calculation schemes (MPA, NPA and AIM). We provide the EEM parameters for all combinations of these theory levels, the basis sets and the charge calculation schemes (see Table 4). Theory levels HF and B3LYP were selected, because they are very often used for QM charge calculation and were also successfully used for EEM parameterization several times [49, 56–60]. The basis set 6-311G was used, because it is robust, also covers iodine and moreover, Pople basis sets are very suitable for EEM parameterization. MPA and NPA

Table 4 Quality criteria of our EEM parameter sets

EEM parameter set name	Relevant QM charges	R^2	RMSD	$\overline{\Delta}$
Cheminf_b3lyp_mpa	B3LYP/6-311G/MPA	0.9007	0.1038	0.0727
Cheminf_b3lyp_npa	B3LYP/6-311G/NPA	0.9651	0.0746	0.0540
Cheminf_b3lyp_aim	B3LYP/6-311G/AIM	0.9499	0.0785	0.0558
Cheminf_hf_mpa	HF/6-311G/MPA	0.9178	0.1125	0.0776
Cheminf_hf_npa	HF/6-311G/NPA	0.9633	0.0805	0.0574
Cheminf_hf_aim	HF/6-311G/AIM	0.9441	0.0919	0.0651

Table 5 Size of database, used for comparison of EEM parameter set coverages

Database	Number of compounds
DrugBank	6874
ChEMBL	1,456,020
PubChem	63,676,639
ZINC	21,957,378

population analyses were employed, because they are the most known charge calculation schemes and additionally, EEM is able to mimic MPA and NPA charges very successfully [49, 58, 59]. AIM was selected, because it is based on a different principle from the other two, and EEM can also mimic AIM charges very efficiently [60]. Note that we do not provide EEM parameters for ESP and RESP charges, because it is known that EEM does not mimic these charges well [2, 58].

EEM parameterization: calculation of QM charges

For each molecule from the training set, six sets of QM charges were calculated via the above-mentioned six QM charge calculation approaches. The calculations of QM charges were carried out using Gaussian09 [68]. With the AIM population analysis, the output from Gaussian03 was further processed with the software package AIMAll [69].

EEM parameterization: calculation of EEM parameter sets

For each set of QM charges, the EEM parameterization was performed and the values of the parameters are provided in (Additional file 2: EEM parameters). The software NEEMP [70] was used for the parameterization. This software implements the parameterization methodology described by [49] and introduces several marked improvements into it. NEEMP provides EEM parameter sets together with their quality criteria, i.e., squared Pearson correlation coefficient (R^2), root mean square deviation (RMSD), and average absolute error ($\overline{\Delta}$), calculated via Eqs. (2), (3) and (4), respectively

$$R^2 = \frac{\left(\sum_{i=1}^{N}\left(q_i^{EEM}-\overline{q}^{EEM}\right)\left(q_i^{QM}-\overline{q}^{QM}\right)\right)^2}{\sum_{i=1}^{N}\left(q_i^{EEM}-\overline{q}^{EEM}\right)^2\sum_{i=1}^{N}\left(q_i^{QM}-\overline{q}^{QM}\right)^2} \quad (2)$$

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}\left(q_i^{EEM}-q_i^{QM}\right)^2}{N}} \quad (3)$$

$$\overline{\Delta} = \frac{\sum_{i=1}^{N}\left|q_i^{EEM}-q_i^{QM}\right|}{N} \quad (4)$$

where q_i^{EEM} is the EEM charge of an atom i; q_i^{QM} is the QM charge of an atom i; $\overline{q}^{EEM}$ is an average of all EEM charges; $\overline{q}^{QM}$ is an average of all QM charges, N is the number of atoms in the molecule.

Coverage comparison (step 2)

For comparison, we used our six EEM parameter sets and 15 published EEM parameter sets, described in Table 1 (all 21 of these EEM parameter sets will be below referred to as the tested EEM parameter sets). The coverage comparison was done on four very well-known databases of drug-like chemical compounds: DrugBank [71, 72], ChEMBL [73], PubChem [74], and ZINC [75]. The number of compounds in all these databases (from 10th February 2015) are summarized in Table 5. For each tested EEM parameter set, we analysed how many compounds from the four databases can be covered by them (i.e., contains only atom types present in the tested EEM parameter sets). This coverage analysis was done using NEEMP.

Quality comparison (step 3)

This evaluation was done for the 21 above-mentioned tested EEM parameter sets and was performed on two data sets—a test set (657 molecules) and an extended test set (1226 molecules). The extended test set contained all approved drugs (i.e., drugs which have received approval in at least one country) from the DrugBank database (downloaded 10th February 2015), for which it was possible to calculate all QM charges necessary for testing. The test set was a subset of the extended test set, which contained only molecules covered by all the tested EEM parameter sets. The 2D structures of all molecules were obtained from DrugBank. The lists of molecules from the test set and the extended test set, including their DrugBank IDs and summary formulas, can be found in (Additional file 3: Table S2a; Additional file 4: Table S2b, respectively). The 3D structures of all the molecules were

generated with CORINA 2.6 [67], without any further geometry optimization. For all the molecules, we calculated all the types of QM charges which corresponded to the tested EEM parameters. This means we used the 8 QM charge calculation approaches mentioned in Table 1 and the six QM charge calculation approaches employed for calculating our EEM parameter sets. The calculations of QM charges were done with Gaussian09 and the AIMAll software package was used for AIM charges. We compared the quality of the tested EEM parameter set on both the test set and the extended test set. The comparison was done using NEEMP, which provided quality criteria for all the tested EEM parameter sets. In the extended test set, some molecules were not covered by certain EEM parameter set(s). Therefore, we calculated quality criteria based purely on the covered molecules and in parallel, we also computed the coverage.

Quality comparison: EEM parameter sets embedded in software tools

The calculation of EEM charges can be done with a few software tools, e.g., EEM SOLVER, OpenBabel or Balloon. The software tools OpenBabel and Balloon contain embedded EEM parameter sets (see Table 2). Therefore, we also evaluated the quality of these embedded EEM parameter sets. This evaluation was done for the same data sets and via the same procedure as with the tested EEM parameter sets. The only difference was that the EEM charges were not calculated with NEEMP, but with OpenBabel and Balloon. Afterwards, these EEM charges were compared with the relevant QM charges using R statistical software [76], which provided their quality criteria.

Software solution (step 4)

We provide the user two such solutions, the first based on EEM SOLVER and the second on OpenBabel.

Results and discussion

EEM parameterization (step 1)

EEM parameterization was performed for six QM charge calculation approaches, and a training set containing 4475 drug-like molecules was used. Squared Pearson correlation coefficient (R^2), root mean square deviation (RMSD) and average absolute error ($\overline{\Delta}$) of the obtained EEM parameter sets, calculated for the training set, are summarized in Table 4. These quality criteria describe the correlation between QM charges and the corresponding EEM charges and they were calculated using NEEMP software.

These results show that the quality of our EEM parameter sets is very high, i.e., all the R^2 values are higher or equal to 0.9. Table 4 also illustrates that QM theory levels B3LYP and HF are both applicable for EEM parameterization, and EEM charges based on them have similar accuracy. From this table, we can also see that the quality of EEM parameters based on NPA and AIM population analysis is slightly better than for MPA.

Coverage comparison (step 2)

Information about the coverages of published EEM parameter sets and our EEM parameter sets are summarized in Table 6. The coverages were computed on four well-known databases of drug-like molecules—DrugBank, ChEMBL, PubChem and ZINC. Table 6 shows that the coverages of the published EEM parameter sets are low (<60 %). The only exception are the EEM parameter sets published by Svobodova et al. and Jirouskova et al., which have coverage between 70 and 80 %. In contrast, our EEM parameter sets have very high coverage—about 95 % or more for all the databases. The not covered molecules include atom types rare for drug-like molecules, e.g., metals or boron. An interesting fact is that the coverages are very similar for all four analyzed databases. Therefore, low EEM parameter set coverage is not merely an isolated issue related to one database, but a general problem.

Quality comparison (step 3)

Table 6 summarizes the main quality criteria (i.e., R^2 values) of all tested EEM parameter sets for the test set, which contained 657 approved drugs from DrugBank. Other quality criteria (RMSD and $\overline{\Delta}$) can be found in (Additional file 5: Table S3) and all values of partial atomic charges (represented as tables and as graphs) are in (Additional file 6). The table shows that our EEM parameter sets are among the best performing EEM parameter sets to have been published so far. The table also illustrates that the quality of EEM parameters is strongly influenced by the selection of QM charge calculation scheme. Specifically, EEM parameters based on MPA, NPA and AIM charges are very high quality, and EEM parameters based on Hirshfeld charges are still acceptable. EEM parameters based on MK and CHELPG charges are very low quality, which is in agreement with published data [2, 58]. Both theory levels (HF and B3LYP) and all three basis sets used (STO-3G, 6-31G* and 6-311G) are applicable for EEM parameterization. These results also confirm that our selection of QM theory level, basis set and charge calculation schemes is appropriate.

For the extended test set, the quality criteria exhibit similar trends (see Additional file 7: Table S4). In parallel, the coverages for this data set are slightly higher than for the complete DrugBank database. An interesting fact is that even for such common compounds as approved drugs, the

Table 6 Summary information about coverage and quality of all tested EEM parameters (see below for meaning of colours)

Relevant QM charges		EEM parameter set name	Coverage comparison				Quality comparison
QM theory level + basis set	Charge calc. scheme		Coverage [%]				R^2 Test set
			DrugBank	ChEMBL	PubChem	ZINC	
HF/STO-3G	MPA	Baek1991	58.1	42.3	40.5	40.1	0.8981
		Svob2007_cbeg2	55.0	49.5	47.3	51.9	0.9758
		Svob2007_chal2	71.7	75.2	77.2	80.2	0.9668
		Svob2007_chm2	72.2	75.2	77.3	80.2	0.9623
		Svob2007_cmet2	55.5	49.5	47.3	51.9	0.9676
HF/6-31G*	MK	Jir2008_hf	70.8	74.7	76.5	79.8	0.6872
B3LYP/6-31G*	MPA	Bult2002_mpa	55.4	49.4	48.2	49.6	0.9658
	NPA	Bult2002_npa	55.4	49.4	48.2	49.6	0.8131
		Ouy2009	49.0	41.1	39.1	40.0	0.9655
		Ouy2009_elem	50.0	41.2	39.1	40.0	0.9633
	Hirshfeld	Bult2002_hir	55.4	49.4	48.2	49.6	0.9061
	MK	Bult2002_mk	55.4	49.4	48.2	49.6	0.7844
		Jir2008_mk	70.8	74.7	76.5	79.8	0.7022
	CHELPG	Bult2002_che	55.4	49.4	48.2	49.6	0.7803
	AIM	Bult2004_aim	55.4	49.4	48.2	49.6	0.9739
HF/6-311G	MPA	Cheminf_hf_mpa	94.6	95.7	96.9	100.0	0.9606
	NPA	Cheminf_hf_npa					0.9713
	AIM	Cheminf_hf_aim					0.9791
B3LYP/6-311G	MPA	Cheminf_b3lyp_mpa					0.9552
	NPA	Cheminf_b3lyp_npa					0.9695
	AIM	Cheminf_b3lyp_aim					0.9800

Coverage	> 90%	> 80%	> 70%	> 60%	< 60%
R^2	> 0.95	> 0.9	> 0.85	> 0.8	< 0.8

coverages of published EEM parameter sets are low. Specifically, most published EEM parameter sets have coverages between 55 and 65 %. Further remarkable fact is that quality criteria of our EEM parameters are better for the test set than for the training set. The reason is that the training set is much larger and heterogeneous than the test set.

Quality comparison: EEM parameter sets embedded in software tools

EEM charges produced with OpenBabel were compared with QM charges calculated with B3LYP/6-31G*/MPA. The quality criteria for the test set were the same as for the EEM parameters Bult2002_mpa (i.e., R^2 about 0.97). This was expected, because OpenBabel uses Bult2002_mpa as its embedded EEM parameters. Very surprising was the behavior of OpenBabel on the extended set. The coverage was 100 %, but the quality criteria were markedly lower (e.g., R^2 about 0.82). The reason for this is that OpenBabel replaces the EEM parameters for atom types which are not provided in Bult2002_mpa with the EEM parameters for some other atom types. Unfortunately, this approach is not very reliable, i.e., the quality criteria for molecules which are in the extended test set but are not in the test set are very low ($R^2 = 0.66$). Additionally, this approach is relatively tricky. The user does not know whether the correct or the estimated EEM parameters are used and, therefore, whether the resulting EEM charges will be of a good quality.

The EEM charges produced by Balloon were compared with the QM charges calculated by the B3LYP/cc-pVTZ/MPA approach. The coverage was close to 100 %, but the correlation was also low ($R^2 < 0.8$). On the other hand, the Balloon developers mentioned that the EEM charges provided by Balloon do not correspond directly to some particular QM charges, and they should only be close to B3LYP/cc-pVTZ/MPA charges.

All the quality criteria and coverages for EEM parameter sets embedded in OpenBabel and Balloon are summarized in (Additional file 8: Table S5).

Coverage comparison and quality comparison combined

To date, there have been no EEM parameter sets available which would provide both high coverage and high-quality EEM charges (see Table 6). On the other hand, the EEM parameter sets calculated in this paper solve this problem, because they exhibit coverage close to 100 % and excellent quality criteria. Therefore, they can be used for chemoinformatics applications.

Software solution (step 4)

For the actual applicability of EEM in chemoinformatics, the user doesn't just need EEM parameter sets that are high quality and cover almost all molecules. They also need a software package that embeds these EEM parameter sets and calculates EEM charges based on them. We provide the user with two such solutions. First, we provide our EEM parameter sets in a format that can be directly used in EEM SOLVER (Additional file 2: EEM parameter sets). Second, we provide an OpenBabel patch which allows our EEM parameter sets to be used directly in OpenBabel (Additional file 9: OpenBabel patch). All the information including documentation is also accessible on the web: http://ncbr.muni.cz/eem_parameters. The parameters are also accessible via ACC web application [77].

Conclusion

We provide here six EEM parameter sets which enable the user to calculate EEM charges with quality comparable to frequently used QM charges computed by well-known charge calculation schemes (i.e., MPA, NPA and AIM) and based on a robust QM approach (HF/6-311G, B3LYP/6-311G). The training set for EEM parameterization contained more than 4000 molecules from the DTP NCI drug database, and all six calculated EEM parameter sets exhibited a very good quality on this training set ($R^2 > 0.9$).

The coverage of these computed EEM parameter sets was then compared with the coverages of 15 EEM parameter sets published in the past. This comparison was done on four key databases of drug-like molecules—DrugBank, ChEMBL, Pubchem and ZINC. The comparison showed that our EEM parameter sets enable us to calculate EEM charges for almost all molecules in these databases.

We then compared the quality of computed and published EEM parameter sets on two test data sets composed of approved drugs from DrugBank. This comparison also included EEM parameter sets embedded in the software tools OpenBabel and Balloon. The comparison showed that our EEM parameter sets are among the best performing EEM parameter sets published to date ($R^2 > 0.93$).

To summarize, charge calculation methodology suitable for chemoinformatics applications like virtual screening or QSAR should be fast, conformationally-dependent and accurate. EEM fulfils all these requirements. However, EEM parameter sets that would exhibit high coverage of drug-like molecule databases and provide high quality charges have not been available to date. The EEM parameters calculated in this paper solve this problem. They exhibit coverage close to 100 % and excellent quality criteria, therefore they are applicable in chemoinformatics.

Last but not least, we provide a software solution for the easy computing of EEM charges based on these EEM parameter sets—input files for EEM SOLVER and OpenBabel patch.

Additional files

Additional file 1: Table S1. List of training set molecules, including their NSC numbers and summary formulas.

Additional file 2: EEM parameters. Values of EEM parameter sets for these six charge calculation approaches (i.e. B3LYP/6-311G/MPA,B3LYP/6-311G/NPA, B3LYP/6-311G/AIM, HF/6-311G/MPA, HF/6-311G/NPA, and HF/6-311G/AIM). These EEM parameter sets are in a format which can be used as an input file for EEM SOLVER.

Additional file 3: Table S2a. A list of molecules from the test set including their DrugBank IDs and summary formulas.

Additional file 4: Table S2b. A list of molecules from the extended test set including their DrugBank IDs and summary formulas.

Additional file 5: Table S3. RMSD and $\overline{\Delta}$ values of all tested EEM parameter sets on the test set.

Additional file 6: Charge details. Values of partial atomic charges (represented as tables and as graphs) for all tested EEM parameter sets on the testset.

Additional file 7: Table S4. R^2, RMSD, $\overline{\Delta}$ and coverage values of all tested EEM parameter sets on the extended test set.

Additional file 8: Table S5. RMSD and $\overline{\Delta}$ values for OpenBabel and Balloon on the test set and extended test set.

Additional file 9: OpenBabel patch. A patch for OpenBabel, which enables it to use the EEM parameter sets calculated in this paper.

Authors' contributions

The concept of the study originated from JK and was reviewed and extended by RA, while the design was put together by RSV and SG and reviewed by JK and RA. TB and SG prepared the input data (molecules and published EEM parameters). TB, SG and VH performed QM charge calculation. TR updated and extended NEEMP software. TB and TR performed EEM parameterizations, EEM charges validation and calculation of statistical data. VH prepared an automatic workflow, which is able to reproduce all steps preformed in the article. AK reviewed, corrected and improved this workflow. TR wrote the OpenBabel patch. The data were analyzed and interpreted by RSV, SG and JK. The manuscript was written by RSV in cooperation with JK, and reviewed by all authors. All authors read and approved the final manuscript.

Author details
[1] National Centre for Biomolecular Research, Faculty of Science and CEITEC, Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic. [2] Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic. [3] Institute of Computer Science, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic. [4] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, 9500 Gilman Drive, San Diego, MC 0657, USA.

Acknowledgements
This work was supported by the Grant Agency of the Czech Republic [13-25401S]; the European Community's Seventh Framework Programme (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund; and by the European Social Fund and the state budget of the Czech Republic (CZ.1.07/2.3.00/20.0042, CZ.1.07/2.3.00/30.0009).

This work was also supported in part by NIH Grants R01 GM071872, U01 GM094612, and U54 GM094618 to R.A.. The access to MetaCentrum supercomputing facilities provided under research intent MSM6383917201 is greatly appreciated.

Authors' information
Stanislav Geidl, Tomáš Bouchal and Tomáš Raček wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors. Radka Svobodová Vařeková and Jaroslav Koča wish it to be known that, in their opinion, they should be regarded as joint Corresponding Authors.

Competing interests
The authors declare that they have no competing interests.

Received: 7 July 2015 Accepted: 16 November 2015
Published online: 02 December 2015

References

1. Svobodová Vařeková R, Geidl S, Ionescu C-M, Skřehota O, Kudera M, Sehnal D, Bouchal T, Abagyan R, Huber HJ, Koča J (2011) Predicting pKa values of substituted phenols from atomic charges: comparison of different quantum mechanical methods and charge distribution schemes. J Chem Inf Model 51(8):1795–1806
2. Svobodová Vařeková R, Geidl S, Ionescu C-M, Skřehota O, Bouchal T, Sehnal D, Abagyan R, Koča J (2013) Predicting pKa values from EEM atomic charges. J Chem Inf 5(1):18
3. Geidl S, Svobodová Vařeková R, Bendová V, Petrusek L, Ionescu C-M, Jurka Z, Abagyan R, Koča J (2015) How does the methodology of 3D structure preparation influence the quality of pKa prediction? J Chem Inf Model 55(6):1088–1097
4. Dixon SL, Jurs PC (1993) Estimation of pKa for organic oxyacids using calculated atomic charges. J Comput Chem 14:1460–1467
5. Zhang J, Kleinöder T, Gasteiger J (2006) Prediction of pKa values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. J Chem Inf Model 46:2256–2256
6. Gross KC, Seybold PG, Hadad CM (2002) Comparison of different atomic charge schemes for predicting pKa variations in substituted anilines and phenols. Int J Quantum Chem 90:445–58
7. Ghafourian T, Dearden JC (2000) The use of atomic charges and orbital energies as hydrogen-bonding-donor parameters for QSAR studies: comparison of MNDO, AM1 and PM3 methods. J Pharm Pharmacol 52(6):603–610
8. Dudek AZ, Arodz T, Gálvez J (2006) Computational methods in developing quantitative structure-activity relationships (QSAR): a review. Comb Chem High Throughput Screen 9(3):213–228
9. Karelson M, Lobanov VS, Katritzky AR (1996) Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev 96(3):1027–1044
10. Todeschini R, Consonni V (2008) Handbook of molecular descriptors. Wiley-VCH Verlag GmbH, Weinheim
11. Galvez J, Garcia R, Salabert MT, Soler R (1994) Charge indexes. New topological descriptors. J Chem Inf Model 34(3):520–525
12. Stalke D (2011) Meaningful structural descriptors from charge density. Chemistry 17(34):9264–9278
13. Wermuth CG (2006) Pharmacophores: historical perspective and viewpoint from a medicinal chemist. In: Langer T, Hoffmann RD (eds) Pharmacophores and pharmacophore searches, vol 32. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim
14. MacDougall PJ, Henze CE (2007) Fleshing-out pharmacophores with volume rendering of the Laplacian of the charge density and hyperwall visualization technology. In: Matta CF, Boyd RJ (eds) The quantum theory of atoms in molecules: from solid state to DNA and drug design. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, pp 499–514
15. Clement OO, Mehl AT (2000) HipHop: pharmacophores based on multiple common-feature alignments. In: Güner OF (ed) Pharmacophore perception, development, and use in drug design. International University Line, La Jolla, pp 69–84
16. Lyne PD (2002) Structure-based virtual screening: an overview. Drug Discov Today 7(20):1047–1055
17. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43(25):4759–4767
18. Park H, Lee J, Lee S (2006) Critical assessment of the automated AutoDock as a new docking tool for virtual screening. Proteins 65(3):549–554
19. Kearsley SK, Sallamack S, Fluder EM, Andose JD, Mosley RT, Sheridan RP (1996) Chemical similarity using physiochemical property descriptors. J Chem Inf Model 36(1):118–127
20. Nikolova N, Jaworska J (2003) Approaches to measure chemical similarity—a review. QSAR Comb Sci 22(910):1006–1006
21. Holliday JD, Jelfs SP, Willett P, Gedeck P (2003) Calculation of intersubstituent similarity using R-group descriptors. J Chem Inf Comput Sci 43(2):406–411
22. Tervo AJ, Rönkkö T, Nyrönen TH, Poso A (2005) BRUTUS: optimization of a grid-based similarity function for rigid-body molecular superposition. 1. Alignment and virtual screening applications. J Med Chem 48(12):4076–4086
23. Vainio MJ, Johnson MS (2007) Generating conformer ensembles using a multiobjective genetic algorithm. J Chem Inf Model 47(6):2462–2474
24. Lemmen C, Lengauer T, Klebe G (1998) FLEXS: a method for fast flexible ligand superposition. J Med Chem 41(23):4502–4520
25. Mulliken RS (1955) Electronic Population Analysis on LCAO-MO Molecular Wave Functions. I. J Chem Phys 23(10):1833
26. Mulliken RS (1955) Electronic population analysis on LCAO-MO molecular wave functions. II. Overlap populations, bond orders, and covalent bond energies. J Chem Phys 23(10):1841
27. Löwdin P-O (1950) On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. J Chem Phys 18(3):365
28. Reed AE, Weinhold F (1983) Natural bond orbital analysis of near-Hartree-Fock water dimer. J Chem Phys 78(6):4066–4073
29. Reed AE, Weinstock RB, Weinhold F (1985) Natural population analysis. J Chem Phys 83(2):735
30. Bader RFW (1985) Atoms in molecules. Acc Chem Res 18(1):9–15
31. Bader RFW (1991) A quantum theory of molecular structure and its applications. Chem Rev 91(5):893–928
32. Hirshfeld FL (1977) Bonded-atom fragments for describing molecular charge densities. Theor Chim Acta 44(2):129–138
33. Ritchie JP (1985) Electron density distribution analysis for nitromethane, nitromethide, and nitramide. J Am Chem Soc 107(7):1829–1837
34. Ritchie JP, Bachrach SM (1987) Some methods and applications of electron density distribution analysis. J Comput Chem 8(4):499–509
35. Breneman CM, Wiberg KB (1990) Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. J Comput Chem 11(3):361–373
36. Singh UC, Kollman PA (1984) An approach to computing electrostatic charges for molecules. J Comput Chem 5(2):129–145
37. Besler BH, Merz KM, Kollman PA (1990) Atomic charges derived from semiempirical methods. J Comput Chem 11(4):431–439
38. Kelly CP, Cramer CJ, Truhlar DG (2005) Accurate partial atomic charges for high-energy molecules using class IV charge models with the MIDI! basis set. Theor Chem Acc 113(3):133–151
39. Marenich AV, Cramer CJ, Truhlar DG (2009) Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. J Phys Chem B 113(18):6378–6396

40. Gasteiger J, Marsili M (1978) A new model for calculating atomic charges in molecules. Tetrahedron Lett 19(34):3181–3184
41. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. Tetrahedron 36(22):3219–3228
42. Cho K-H, Kang YK, No KT, Scheraga HA (2001) A fast method for calculating geometry-dependent net atomic charges for polypeptides. J Phys Chem B 105(17):3624–3624
43. Oliferenko AA, Pisarev SA, Palyulin VA, Zefirov NS (2006) Atomic charges via electronegativity equalization: generalizations and perspectives. Adv Quantum Chem 51:139–156
44. Shulga DA, Oliferenko AA, Pisarev SA, Palyulin VA, Zefirov NS (2010) Fast tools for calculation of atomic charges well suited for drug design. SAR QSAR Environ Res 19(1–2):153–165
45. Mortier WJ, Ghosh SK, Shankar S (1986) Electronegativity equalization method for the calculation of atomic charges in molecules. J Am Chem Soc 108:4315–4320
46. Rappe AK, Goddard WA (1991) Charge equilibration for molecular dynamics simulations. J Phys Chem 95(8):3358–3363
47. Nistor RA, Polihronov JG, Müser MH, Mosey NJ (2006) A generalization of the charge equilibration method for nonmetallic materials. J Chem Phys 125(9):094108
48. Mathieu D (2007) Split charge equilibration method with correct dissociation limits. J Chem Phys 127(22):224103
49. Svobodová Vařeková R, Jiroušková Z, Vaněk J, Suchomel S, Koča J (2007) Electronegativity equalization method: parameterization and validation for large sets of organic, organohalogene and organometal molecule. Int J Mol Sci 8:572–572
50. Janssens GOA, Baekelandt BG, Toufar H, Mortier WJ, Schoonheydt RA (1995) Comparison of cluster and infinite crystal calculations on zeolites with the electronegativity equalization method (EEM). J Phys Chem 99(10):3251–3258
51. Heidler R, Janssens GOA, Mortier WJ, Schoonheydt RA (1996) Charge sensitivity analysis of intrinsic basicity of Faujasite-type zeolites using the electronegativity equalization method (EEM). J Phys Chem 100(50):19728–19734
52. Sorich MJ, McKinnon RA, Miners JO, Winkler DA, Smith PA (2004) Rapid prediction of chemical metabolism by human UDP-glucuronosyltransferase isoforms using quantum chemical descriptors derived with the electronegativity equalization method. J Med Chem 47(21):5311–5317
53. Bultinck P, Langenaeker W, Carbó-Dorca R, Tollenaere JP (2003) Fast calculation of quantum chemical molecular descriptors from the electronegativity equalization method. J Chem Inf Comput Sci 43(2):422–428
54. Smirnov KS, van de Graaf B (1996) Consistent implementation of the electronegativity equalization method in molecular mechanics and molecular dynamics. J Chem Soc Faraday Trans 92(13):2469
55. Ionescu C-M, Geidl S, Svobodová Vařeková R, Koča J (2013) Rapid calculation of accurate atomic charges for proteins via the electronegativity equalization method. J Chem Inf Model 53(10):2548–2548
56. Baekelandt BG, Mortier WJ, Lievens JL, Schoonheydt RA (1991) Probing the reactivity of different sites within a molecule or solid by direct computation of molecular sensitivities via an extension of the electronegativity equalization method. J Am Chem Soc 113(18):6730–6734
57. Jiroušková Z, Vařeková RS, Vaněk J, Koča J (2009) Electronegativity equalization method: parameterization and validation for organic molecules using the Merz-Kollman-Singh charge distribution scheme. J Comput Chem 30(7):1174–1178
58. Bultinck P, Langenaeker W, Lahorte P, De Proft F, Geerlings P, Van Alsenoy C, Tollenaere JP (2002) The electronegativity equalization method II: applicability of different atomic charge schemes. J Phys Chem A 106(34):7895–7901
59. Ouyang Y, Ye F, Liang Y (2009) A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules. Phys Chem Chem Phys 11(29):6082–6089
60. Bultinck P, Vanholme R, Popelier PLA, De Proft F, Geerlings P (2004) High-speed calculation of AIM charges through the electronegativity equalization method. J Phys Chem A 108(46):10359–10366
61. O'Boyle N, Banck M, James C, Morley C, Vandermeersch T, Hutchison G (2011) Open Babel: an open chemical toolbox. J Chem Inf 3(1):33–47
62. Puranen JS, Vainio MJ, Johnson MS (2010) Accurate conformation-dependent molecular electrostatic potentials for high-throughput in silico drug discovery. J Comput Chem 31(8):1722–1732
63. Svobodová Vařeková R, Koča J (2006) Optimized and parallelized implementation of the electronegativity equalization method and the atom-bond electronegativity equalization method. J Comput Chem 3:396–405
64. Bultinck P, Carbó-Dorca R, Langenaeker W (2003) Negative Fukui functions: new insights based on electronegativity equalization. J Chem Phys 118(10):4349
65. Burden FR, Polley MJ, Winkler DA (2009) Toward novel universal descriptors: charge fingerprints. J Chem Inf Model 49(3):710–715
66. Open NCI Database (2012) Release 4. http://cactus.nci.nih.gov/download/nci/
67. Sadowski J, Gasteiger J (1993) From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. Chem Rev 93:2567–2581
68. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA Jr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA. Gaussian 09, Revision E.01. http://www.gaussian.com
69. Todd A Keith (2015) AIMAll 15.05.18. http://aim.tkgristmill.com
70. Raček T, Svobodová Vařeková R, Křenek A, Koča J NEEMP—tool for parameterization of empirical charge calculation method EEM. http://ncbr.muni.cz/neemp/
71. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M(2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36(Database issue):901–906
72. Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS (2004) DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 42(Database issue):1091–1097
73. Bento AP, Gaulton A, Hersey A, Bellis LJ, Chambers J, Davies M, Krüger FA, Light Y, Mak L, McGlinchey S, Nowotka M, Papadatos G, Santos R, Overington JP (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res 42(Database issue):1083–1090
74. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) PubChem: integrated platform of small molecules and biological activities. In: Wheeler R, Spellmeyer D (eds) Annual Reports in Computational Chemistry, vol. 4, Chap 12. Elsevier, Oxford
75. Irwin JJ, Sterling T, Mysinger MM, Bolstad ES, Coleman RG (2012) ZINC: a free tool to discover chemistry for biology. J Chem Inf Model 52(7):1757–1768
76. R Core Team R: A Language and Environment for Statistical Computing. http://www.r-project.org/
77. Ionescu CM, Sehnal D, Falginella FL, Pant P, Pravda L, Bouchal T, Svobodová Vařeková R, Geidl S, Koča J (2015) AtomicChargeCalculator: interactive web-based calculation of atomic charges in large biomolecular complexes and drug-like molecules. J Cheminf 7(1):50

How does the methodology of 3D Structure preparation influence the quality of pK_{a} prediction?

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

How Does the Methodology of 3D Structure Preparation Influence the Quality of pK_a Prediction?

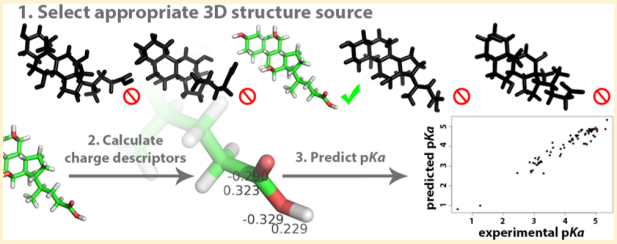
Stanislav Geidl,[†] Radka Svobodová Vařeková,*[,†] Veronika Bendová,[†] Lukáš Petrusek,[†] Crina-Maria Ionescu,[†] Zdeněk Jurka,[†] Ruben Abagyan,[‡] and Jaroslav Koča*[,†]

[†]National Centre for Biomolecular Research, Faculty of Science, and CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic

[‡]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, 9500 Gilman Drive, MC 0657, San Diego, California 92161, United States

Supporting Information

ABSTRACT: The acid dissociation constant is an important molecular property, and it can be successfully predicted by Quantitative Structure−Property Relationship (QSPR) models, even for *in silico* designed molecules. We analyzed how the methodology of *in silico* 3D structure preparation influences the quality of QSPR models. Specifically, we evaluated and compared QSPR models based on six different 3D structure sources (DTP NCI, Pubchem, Balloon, Frog2, OpenBabel, and RDKit) combined with four different types of optimization. These analyses were performed for three classes of molecules (phenols, carboxylic acids, anilines), and the QSPR model descriptors were quantum mechanical (QM) and empirical partial atomic charges. Specifically, we developed 516 QSPR models and afterward systematically analyzed the influence of the 3D structure source and other factors on their quality. Our results confirmed that QSPR models based on partial atomic charges are able to predict pK_a with high accuracy. We also confirmed that *ab initio* and semiempirical QM charges provide very accurate QSPR models and using empirical charges based on electronegativity equalization is also acceptable, as well as advantageous, because their calculation is very fast. On the other hand, Gasteiger-Marsili empirical charges are not applicable for pK_a prediction. We later found that QSPR models for some classes of molecules (carboxylic acids) are less accurate. In this context, we compared the influence of different 3D structure sources. We found that an appropriate selection of 3D structure source and optimization method is essential for the successful QSPR modeling of pK_a. Specifically, the 3D structures from the DTP NCI and Pubchem databases performed the best, as they provided very accurate QSPR models for all the tested molecular classes and charge calculation approaches, and they do not require optimization. Also, Frog2 performed very well. Other 3D structure sources can also be used but are not so robust, and an unfortunate combination of molecular class and charge calculation approach can produce weak QSPR models. Additionally, these 3D structures generally need optimization in order to produce good quality QSPR models.



INTRODUCTION

The acid dissociation constant, K_a, and its logarithmic version, pK_a, are important molecular properties and their values are of interest in chemical, biological, environmental, and pharmaceutical research.[1−3] Experimental pK_a values are usually unavailable for all compounds from the chemical catalogues. Therefore, they cannot be used for example in virtual screening, which requires predictions of physicochemical properties for large sets of *in silico* designed molecules. Several pK_a prediction methodologies have been published to date. They are summarized in review articles,[4−7] but reliable and accurate pK_a prediction is still a challenge and a topic of intensive research.[8−10]

A popular and frequently used pK_a prediction approach is based on the QSPR (Quantitative Structure−Property Relationship) methodology.[11−13] Various types of input values (so-called descriptors) can be used for the calculation of pK_a via QSPR models. Partial atomic charges are definitely relevant descriptors for pK_a calculations[12,14−17] and can be calculated directly from the 3D structure of the molecule. The partial atomic charges cannot be determined experimentally or derived from the results of quantum mechanics (QM) in a straightforward manner. For this reason, many different methods have been developed for their calculation. The most common method for charge calculation is using a quantum mechanical approach (a combination of a theory level and a basis set) and the subsequent application of a charge calculation scheme. For example, for pK_a prediction via QSPR models, *ab initio* QM charges calculated via HF or B3LYP theory levels and STO-3G or 6-31G* basis sets proved suitable. The most appropriate charge calculation schemes for these purposes seem

Received: December 21, 2014
Published: May 26, 2015

DOI: 10.1021/ci500758w
J. Chem. Inf. Model. 2015, 55, 1088−1097

to be MPA (Mulliken population analysis), NPA (natural population analysis), and AIM (atoms in molecules).[8,15,17] Semiempirical QM charges have also been employed in QSPR models for pK_a prediction (e.g., AM1, PM3, or PM6 theory levels in combination with MPA).[11,14,17−19] A major drawback of the QM charges is the computational effort required for the calculation of the wave function. For this reason, the computational complexity of obtaining QM charges is at least $\theta(B^4)$, where B is the number of basis functions. Therefore, the calculation of *ab initio* QM charges is very time consuming, while the calculation of semiempirical QM charges is also relatively slow. The Electronegativity Equalization Method[20] is an empirical charge calculation approach that presents a faster alternative to the QM methods. EEM is able to provide partial atomic charges with comparable accuracy to QM charges, and it is markedly less time consuming than QM charge calculation approaches. EEM is even able to mimic a certain QM charge calculation approach (i.e., the combination of a theory level, a basis set, and a charge calculation scheme) because it includes parameters based on the QM charges. EEM charges also proved applicable for pK_a prediction via QSPR.[8] Last but not least, pK_a predicting QSPR models based on conformationally independent empirical charges (so-called topological charges, e.g., Gasteiger-Marsili charges) have also been evaluated.[13,19]

Therefore, in principle, we can prepare a straightforward and time-efficient workflow for obtaining pK_a values for molecules designed *in silico*: Use the 3D structures of molecules prepared *in silico*, calculate partial atomic charges for them, employ the charges as descriptors in QSPR models, and predict the required pK_a values. Such a workflow can be applied in virtual screening. We can also design similar workflows for other biologically important properties such as log*P*, biodegradability, dioxin-like activity, etc.

Nonetheless, before implementing the workflow, we need to answer a key question: How does the methodology of *in silico* 3D structure preparation influence the quality of QSPR models for pK_a prediction? In previous works focused on pK_a prediction via QSPR,[8,17,19,21,22] 3D structures were mainly obtained from the DTP NCI database[23] (which uses CORINA to generate the 3D structures) or directly designed by CORINA.[24] But there are other tools and databases that are often used as sources of 3D structures, for example, the database Pubchem[25] (employing the software Omega[26]) or software tools such as Balloon,[27] Frog2,[28] OpenBabel,[29] or RDKit.[30] These tools create 3D structures via a data or knowledge-based approach (CORINA, OpenBabel, Omega), distance geometry approach (Balloon, RDKit), or other approaches (Frog2). Specifically, Frog2 first generates a graph of rings and acyclic elements and afterward performs a Monte Carlo search. Can we use any of these 3D structure sources for the QSPR modeling of pK_a? Or is it that only some methodologies for 3D structure preparation provide acceptable QSPR models? In parallel, another important question is whether the 3D structures need to be optimized before they can be used in QSPR models or not. Some articles on this topic use optimization,[14,15,22,31,32] while some provide accurate models even without it.[8,11,17]

In this study, we addressed the above questions. Specifically, we evaluated and compared QSPR models based on six different 3D structure sources combined with four different types of optimization. The 3D structure sources were the databases DTP NCI and Pubchem and the software tools Balloon, Frog2, OpenBabel, and RDKit. The optimization was either skipped or done by molecular mechanics (MMFF94 for all 3D structure sources, MM-UFF for RDKit) or quantum mechanics (B3LYP/6-31G*). These analyses were performed for three classes of molecules (phenols, carboxylic acids, anilines). We mainly focused on *ab initio* QM charges, which provide the most accurate pK_a predicting QSPR models, and on empirical EEM charges, which are a faster and comparably accurate alternative to *ab initio* QM charges. Specifically, we used four types of QM charges (HF/STO-3G/MPA, B3LYP/6-31G*/MPA, B3LYP/6-31G*/NPA, and B3LYP/6-31G*/AIM) and four corresponding types of EEM charges. To create a complete overview, we provide also QSPR models based on semiempirical charges (i.e., PM6 charges) and on conformationally independent empirical charges (i.e., Gasteiger-Marsili charges). Thus, we developed 516 QSPR models and afterward systematically analyzed the influence of the 3D structure source and other factors on their quality.

METHODS

Data Sets. Our training data set is composed of three classes of molecules (i.e., phenols, anilines, and carboxylic acids), which represent common classes of organic molecules. These types of molecules are also frequently used for the evaluation of QSPR models.[8,11,14−17,19,22,31] The data set contains 190 molecules: 60 phenols, 82 carboxylic acids, and 48 anilines. Additionally, we used a test data set containing 53 phenols that were not included in the training data set. The list of molecules including their figures, NCS numbers, and CAS numbers can be found in the Supporting Information (Table S1).

pK_a Values. The experimental pK_a values were taken from the Physprop database.[33] The pK_a values of all molecules can be found in the Supporting Information (Table S1).

2D Structure of Molecules. Information about the 2D structure of individual molecules was obtained from the DTP NCI database. The 2D structures were described in SMILES format. The SMILES of all molecules are given in the Supporting Information.

Sources of 3D Structure of Molecules. For each molecule, the 3D structure was obtained from six different sources. Specifically, the structure was obtained from two databases (Pubchem, DTP NCI) and in parallel generated by four different freely available software tools (Balloon, Frog2, OpenBabel, and RDKit). These sources were selected because they appear to be the most popular, and they also represent the main approaches for 3D structure preparation.

Optimization. Each molecule was thus associated with six different 3D structures, obtained by the six approaches described above. Afterward, each 3D structure was processed in two different ways. Specifically, two types of optimization were performed: optimization via quantum mechanics (QM) and optimization via molecular mechanics (MM). The QM optimization was performed by Gaussian 09[34] using B3LYP/6-31G*, and the MM optimization was done with RDKit using MMFF94. These approaches were selected because they are common and frequently used representatives of QM and MM optimization. Additionally, we also performed an optimization via the MM force field UFF (Universal Force Field) for structures prepared with RDKit. The reason is that the RDKit developers recommend applying this particular force field for the structures generated with RDKit.

3D Structures in Training and Test Data Sets. Each molecule in our training data set was associated with 19

DOI: 10.1021/ci500758w
J. Chem. Inf. Model. 2015, 55, 1088−1097

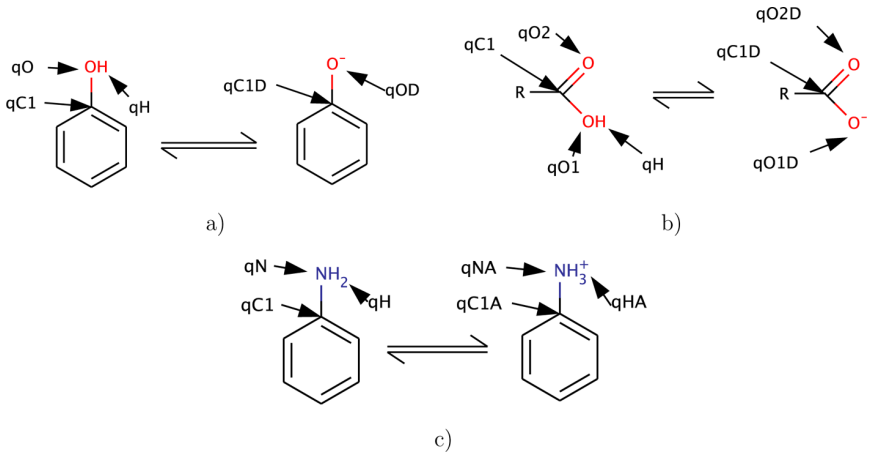


Figure 1. (a) Dissociation of phenols. (b) Dissociation of carboxylic acids. (c) Association of anilines. The particular atomic charges used in our QSPR models are marked by their denotations.

different structures because there were six sources of 3D structure and three types of optimization for each (no optimization, QM optimization, and MM optimization) plus an additional UFF optimization for RDKit. The test data set contained only phenol molecules. Each molecule was associated with two different structures because we selected two sources of 3D structure (i.e., DTP NCI and RDKit) and one type of optimization for each (no optimization).

In our QSPR models, we used neutral forms of all the molecules and also dissociated forms of phenols and carboxylic acids and associated forms of anilines (Figure 1). The dissociated forms of molecules were created by removing the hydrogen atom of the dissociating group. The associated forms of anilines were created by adding one hydrogen atom to the amino group. The adding of the atom was done via an in-house script that applies the Bioshell library,[35,36] and a detailed description of the procedure is given in the Supporting Information.

In this way, our training data set contained 19 (6 × 3 + 1) different structures for each molecule, and 7220 (= 19 × 190 × 2) structures in total. In parallel, our test data set included two different structures for each molecule, therefore, 212 (= 2 × 53 × 2) structures in total.

QM Charges. For each of the 7220 structures from the training set, we calculated *ab initio* QM partial atomic charges via four QM charge calculation approaches (i.e., HF/STO-3G/MPA, B3LYP/6-31G*/MPA, B3LYP/6-31G*/NPA, and B3LYP/6-31G*/AIM) and semiempirical QM charges using PM6. These approaches were selected because they represent the main types of charge calculation approaches that have been reported as successful for pK_a prediction via QSPR.[8,15,17] The second reason for selection of the *ab initio* QM approaches was that corresponding EEM parameters are available for them. For each of the 212 structures from the test set, we calculated *ab initio* QM charges via B3LYP/6-31G*/NPA. This charge calculation approach was selected based on the results obtained on the training set. All the *ab initio* and semiempirical QM charges were calculated by Gaussian 09.[34]

EEM Charges. For each of the 7220 structures in our data set, the EEM charges were calculated by the program EEM SOLVER[37] using the four EEM parameter sets described in Table 1. EEM charges calculated using these parameter sets should mimic QM charges calculated by the relevant QM charge calculation approaches.

Table 1. Summary Information about EEM Parameter Sets Used in This Study

parameter set name	QM charge calculation approach	published by
Svob2007_chal2	HF/STO-3G/MPA	Svobodova et al.[38]
Chaves2006	B3LYP/6-31G*/MPA	Chaves et al.[39]
Bult2002_npa	B3LYP/6-31G*/NPA	Bultinck et al.[40]
Bult2004_aim	B3LYP/6-31G*/AIM	Bultinck et al.[41]

Gasteiger-Marsili Charges. We calculated also empirical Gasteiger-Marsili charges for all the molecules from the training set, including their dissociated or associated forms, therefore for 380 (= 2 × 190) molecules. Gasteiger-Marsili charges are based on 2D structure; therefore, they do not depend on the source of 3D structure and on the optimization. All these charges were calculated by RDKit.[30]

Descriptors and QSPR Models. The descriptors used for QSPR modeling were partial atomic charges from atoms that are close to the dissociation or association site. We employed both charges from neutral and from dissociated (or associated) molecules. The linear model is justified by the linear relationship between pK_a and the electrostatic potential at the protonation site combined with the linear dependence of the potential on the surrounding charges. The distance dependences are absorbed by the p coefficients derived from the experimental data.

Thus, the QSPR model employed in this study for phenol molecules has the following equation:

$$\mathrm{p}K_a = p_{\mathrm{p(H)}} \cdot q_{\mathrm{H}} + p_{\mathrm{p(O)}} \cdot q_{\mathrm{O}} + p_{\mathrm{p(C1)}} \cdot q_{\mathrm{C1}} + p_{\mathrm{p(OD)}} \cdot q_{\mathrm{OD}} + p_{\mathrm{p(C1D)}} \cdot q_{\mathrm{C1D}} + p_{\mathrm{p}} \quad (1)$$

where q_{H} is the atomic charge of the hydrogen atom from the phenolic OH group of the neutral molecule; q_{O} is the charge on the oxygen atom from the phenolic OH group of the neutral molecule; q_{C1} is the charge on the carbon atom binding the phenolic OH group of the neutral molecule; q_{OD} is the charge on the phenoxide O^- from the dissociated molecule; and q_{C1D} is the charge on the carbon atom binding this oxygen in the dissociated molecule (Figure 1a). The symbols $p_{\mathrm{p(H)}}$, $p_{\mathrm{p(O)}}$,

DOI: 10.1021/ci500758w
J. Chem. Inf. Model. 2015, 55, 1088−1097

Table 2. R^2 Describing the Correlation between Calculated and Experimental pK_a for QM QSPR Models

R^2		Class of molecules	Phenols				Carboxylic acids				Anilines				Average
		Charge calculation approach	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	
Source + Optimization	Balloon	none	0.896	0.939	0.908	0.904	0.823	0.720	0.819	0.846	0.836	0.903	0.912	0.805	0.859
		MM	0.917	0.881	0.933	0.891	0.867	0.587	0.805	0.843	0.874	0.953	0.927	0.921	0.867
		QM	0.915	0.871	0.901	0.856	0.890	0.618	0.824	0.807	0.948	0.967	0.933	0.921	0.871
	Frog2	none	0.894	0.912	0.906	0.891	0.896	0.876	0.876	0.884	0.934	0.911	0.924	0.916	0.902
		MM	0.967	0.931	0.907	0.938	0.907	0.830	0.903	0.922	0.958	0.973	0.965	0.926	0.927
		QM	0.969	0.963	0.953	0.939	0.917	0.853	0.906	0.917	0.875	0.973	0.911	0.853	0.919
	NCI	none	0.947	0.971	0.960	0.973	0.931	0.891	0.911	0.910	0.951	0.970	0.966	0.903	0.940
		MM	0.958	0.963	0.959	0.936	0.938	0.889	0.929	0.922	0.954	0.955	0.967	0.914	0.940
		QM	0.891	0.935	0.861	0.902	0.925	0.854	0.903	0.921	0.942	0.959	0.937	0.892	0.910
	OpenBabel	none	0.955	0.961	0.957	0.963	0.869	0.658	0.845	0.876	0.952	0.973	0.966	0.930	0.909
		MM	0.961	0.965	0.959	0.961	0.863	0.665	0.841	0.875	0.958	0.975	0.967	0.927	0.910
		QM	0.955	0.957	0.956	0.936	0.845	0.674	0.804	0.827	0.874	0.974	0.928	0.880	0.884
	PubChem	none	0.960	0.950	0.935	0.900	0.909	0.873	0.891	0.907	0.938	0.939	0.921	0.937	0.922
		MM	0.963	0.911	0.927	0.864	0.916	0.885	0.892	0.916	0.942	0.979	0.966	0.916	0.923
		QM	0.943	0.936	0.922	0.886	0.901	0.871	0.896	0.908	0.934	0.974	0.885	0.828	0.907
	RDKit	none	0.782	0.895	0.796	0.882	0.780	0.723	0.804	0.817	0.853	0.816	0.851	0.796	0.816
		MM-UFF	0.947	0.961	0.941	0.934	0.894	0.821	0.842	0.860	0.965	0.979	0.973	0.980	0.925
		MM	0.931	0.909	0.934	0.950	0.902	0.750	0.797	0.862	0.959	0.976	0.967	0.927	0.905
		QM	0.935	0.944	0.933	0.922	0.861	0.696	0.814	0.855	0.940	0.964	0.927	0.908	0.892
Average			0.931	0.934	0.924	0.917	0.886	0.776	0.858	0.878	0.926	0.953	0.936	0.899	

Legend	$R^2 \geq 0.95$	$R^2 \geq 0.9$	$R^2 \geq 0.866$	$R^2 \geq 0.833$	$R^2 \geq 0.8$	$R^2 \geq 0.7$	$R^2 < 0.7$

$p_{p(C1)}$, $p_{p(OD)}$, $p_{p(C1D)}$, and p_p are parameters of the QSPR model.

The QSPR model employed in this study for carboxylic acids uses the following equation:

$$pK_a = p_{c(H)} \cdot q_H + p_{c(O1)} \cdot q_{O1} + p_{c(O2)} \cdot q_{O2} + p_{c(C1)} \cdot q_{C1} + p_{c(O1D)} \cdot q_{O1D} + p_{c(O2D)} \cdot q_{O2D} + p_{c(C1D)} \cdot q_{C1D} + p_c \quad (2)$$

where q_H and q_{O1} are the atomic charge of the hydrogen and oxygen atoms from the OH group of the neutral molecule, respectively; q_{O2} is the charge on the oxygen atom from the carbonyl group of the neutral molecule; q_{C1} is the charge on the carbon atom binding in the COOH group of the neutral molecule; q_{O1D} is the charge on the O^- oxygen from the dissociated molecule; q_{O2D} is the charge on the oxygen atom from the carbonyl group of the dissociated molecule; and q_{C1D} is the charge on the carbon atom in the carboxyl group of the dissociated molecule (Figure 1b). Because the structures of dissociated carboxylic acid molecules were created by removing the H atom with no further correction of the structure, the values q_{O1D}, q_{O2D}, and q_{C1D} describe charge distribution immediately after removing of this hydrogen atom. The symbols $p_{c(H)}$, $p_{c(O1)}$, $p_{c(O2)}$, $p_{c(C1)}$, $p_{c(O1D)}$, $p_{c(O2D)}$, $p_{c(C1D)}$, and p_c are parameters of the QSPR model.

The QSPR model employed in this study for anilines is based on the following equation:

$$pK_a = p_{a(H)} \cdot q_H + p_{a(N)} \cdot q_N + p_{a(C1)} \cdot q_{C1} + p_{a(HA)} \cdot q_{HA} + p_{a(NA)} \cdot q_{NA} + p_{a(C1A)} \cdot q_{C1A} + p_a \quad (3)$$

where q_H is the average of charges located on both hydrogens in the amino group of the neutral molecule; q_N is the charge of the nitrogen from the amino group of the neutral molecule; q_{C1} is the charge on the carbon atom binding the amino group in the neutral molecule; q_{HA} is the average of charges located on the three hydrogens in the amino group of the associated molecule; q_{NA} is the charge on the nitrogen from the amino group of the associated molecule and q_{C1A} is the charge on the carbon atom binding the amino group in the associated molecule (Figure 1c). The symbols $p_{a(H)}$, $p_{a(N)}$, $p_{a(C1)}$, $p_{a(HA)}$, $p_{a(NA)}$, $p_{a(C1A)}$, and p_a are parameters of the QSPR model.

The QSPR model eqs 1 and 2 were published by Svobodová and Geidl et al.,[8] and they proved useful for pK_a prediction based on QM and EEM charges. Equation 3 was inspired by these two equations.

In this way, we created one QSPR model for each of our three classes of molecules (phenols, carboxylic acids, anilines), 19 types of structures (six sources of 3D structures × three methods of optimization + RDKit with MM-UFF), and nine types of charges (five types of QM charges and four types of EEM charges). For each class of molecules, we additionally created one QSPR model based on Gasteiger-Marsili charges. Thus, we created 516 (= 3 × 19 × 9 + 3) QSPR models. Specifically, 228 QSPR models based on *ab initio* QM charges (denoted QM QSPR models), 57 models based on semi-empirical charges (denoted semiempirical QM QSPR models), 228 models based on EEM charges (denoted EEM QSPR models), and three models based on Gasteiger-Marsili charges (GM QSPR models). The parametrization of the QSPR models was done by multiple linear regression (MLR) using the software QSPR Designer.[42]

Cross-Validation. The robustness of all 516 QSPR models was tested by cross-validation. The *k*-fold cross-validation procedure was used,[43,44] where $k = 5$. Specifically, for each QSPR model, its training data set was divided into five parts (each contained 20% of the molecules). This division was done randomly and included stratification by pK_a value. Afterward, five cross-validation steps were performed. In the first step, the first part was selected as a test set, and the remaining four parts were taken together as the training set. The test and training sets for the other cross-validation steps were prepared in a similar manner.

DOI: 10.1021/ci500758w
J. Chem. Inf. Model. 2015, 55, 1088−1097

RESULTS AND DISCUSSION

The quality of the QSPR models, i.e., the correlation between experimental pK_a and the pK_a calculated by each model, was evaluated using the squared Pearson correlation coefficient (R^2), root-mean-square error (RMSE), and average absolute pK_a error ($\overline{\Delta}$), while the statistical criteria were the standard deviation of the estimation (s) and Fisher's statistics of the regression (F).

Tables 2 and 11 and Table S2 in Supporting Information summarize the squared Pearson correlation coefficients for all QM QSPR models, EEM QSPR models, and semiempirical QM QSPR models, respectively. Table S3 in the Supporting Information contains all the quality criteria (R^2, RMSE, $\overline{\Delta}$) and statistical criteria (s and F) for all the QSPR models analyzed. All these models are statistically significant at $p = 0.01$. Because our data sets contained 60 phenols, 82 carboxylic acids, and 48 anilines, the appropriate F values to consider were those for 60 samples, 80 samples, and 50 samples, respectively. The QSPR models for phenols, carboxylic acids, and anilines contained 5, 7, and 6 descriptors, respectively. Thus, the QSPR models for phenols are statistically significant (at $p = 0.01$) when $F > 3.34$, the QSPR models for carboxylic acids when $F > 2.87$, and the QSPR models for anilines when $F > 3.19$.

The parameters of the QSPR models are summarized in the Supporting Information (Table S4).

Quality of QM QSPR Models: General Summary. The results summarized in Tables 2 and 3 confirmed that the QSPR models based on QM charges are able to predict pK_a with high accuracy. Specifically, about 24% of the models have excellent quality ($R^2 \geq 0.95$), close to 40% have very good quality ($R^2 \geq 0.9$), 30% have lower quality but are still applicable ($R^2 \geq 0.8$), and only about 6% have low quality ($R^2 < 0.8$).

Table 3. Number and Percentage of QM QSPR Models with R^2 Higher than a Defined Limit

R^2	≥0.95	(0.95, 0.9>	(0.9, 0.8>	<0.8
number of models	55	90	69	14
percentage of models	24%	39%	30%	6%

Predictivity of QM QSPR Models. In general, the predictivity of QSPR models calculating pK_a based on charges was shown in the literature[11−13]. Additionally, high quality of QM QSPR models based on the same charge descriptors as our models was shown by Svobodová Vařeková et al.[17] To confirm the predictivity, we did a cross-validation for all our QSPR models. Cross-validation results for selected QSPR models are in Table 4 (i.e., based on B3LYP/6-31G*/NPA charges and nonoptimized OpenBabel 3D structures, which show average quality in comparison with other QM QSPR models). All the cross-validation results can be found in the Supporting Information (Table S5). These results showed that the values of R^2 are similar for the test set, the training set, and the complete set; therefore, the models are stable.

For further confirmation of our QSPR models predictivity, we tested selected QSPR models on an independent test data set prepared only for testing purposes, with a size comparable to that of the training data set. Specifically, the test data set includes 53 phenol molecules, and we used it for testing two selected QM QSPR models for phenols, namely, one of the best quality models (B3LYP/6-31G*/NPA charges and nonoptimized 3D structures from NCI) and one of the worst quality models (HF/STO-3G/MPA charges and nonoptimized 3D structures from RDKit). The quality criteria for the test set and the training set are in Table 5. These results demonstrate that the QSPR models perform comparably for the test set and the training set.

Table 4. R^2 Values for Cross-Validation of Selected QM QSPR Models

QSPR model description: phenols. Charges: B3LYP/6-31G*/NPA. 3D structure: OpenBabel with no optimization.					
cross-validation step	1	2	3	4	5
R^2 for training set	0.955	0.956	0.964	0.959	0.957
R^2 for test set	0.956	0.967	0.939	0.952	0.957
R^2 for complete set			0.957		
QSPR model description: carboxylic acids. Charges: B3LYP/6-31G*/NPA. 3D structure: OpenBabel with no optimization.					
cross-validation step	1	2	3	4	5
R^2 for training set	0.818	0.825	0.889	0.863	0.852
R^2 for test set	0.928	0.785	0.609	0.850	0.816
R^2 for complete set			0.845		
QSPR model description: anilines. Charges: B3LYP/6-31G*/NPA. 3D structure: OpenBabel with no optimization.					
cross-validation step	1	2	3	4	5
R^2 for training set	0.966	0.965	0.973	0.963	0.970
R^2 for test set	0.937	0.925	0.910	0.988	0.932
R^2 for complete set			0.966		

Table 5. Quality Criteria for Testing of Selected QM QSPR Models

QSPR model description: phenols. Charges: B3LYP/6-31G*/NPA. 3D structure: NCI with no optimization.			
quality criteria	R^2	RMSE	$\overline{\Delta}$
training set	0.960	0.415	0.333
test set	0.948	0.532	0.437
QSPR model description: phenols. Charges: HF/STO-3G/MPA. 3D structure: RDKit with no optimization.			
quality criteria	R^2	RMSE	$\overline{\Delta}$
training set	0.782	1.067	0.896
test set	0.715	0.421	0.328

Influence of *ab initio* QM Charge Calculation Approach. The results (Tables 2 and 6) show that all four of the *ab initio* QM charge calculation approaches tested here provide a comparable quality of pK_a prediction. These results therefore confirmed that all the selected charge calculation approaches are suitable for the QSPR prediction of pK_a. Additionally, all the charge calculation approaches are applicable for all three classes of molecules. Specifically, for each class of molecules, any *ab initio* QM charge calculation

Table 6. Number and Percentage of QM QSPR Models with R^2 Higher than a Defined Limit for Individual Charge Calculation Approaches

	R^2			
QM charge calculation approach	≥0.9	(0.9, 0.8>	<0.8	R^2_{chrg} *
HF/STO-3G/MPA	67%	30%	4%	0.914
B3LYP/6-31G*/MPA	60%	25%	16%	0.888
B3LYP/6-31G*/NPA	68%	28%	4%	0.906
B3LYP/6-31G*/AIM	60%	39%	2%	0.898

*R^2_{chrg} is the average value of R^2 for all QSPR models, which use charges calculated by a given QM charge calculation approach.

DOI: 10.1021/ci500758w
J. Chem. Inf. Model. 2015, 55, 1088−1097

approach provides good quality QSPR models (R^2 close to 0.9) at least for some sources of 3D structures. An interesting finding is that the suitability of a certain charge calculation approach strongly depends on the class of molecules. For example, B3LYP/6-31G*/MPA charges work very well for anilines and markedly poorer for carboxylic acids. The next interesting finding is that the charge calculation approach HF/STO-3G/MPA, which uses the smallest basis set (STO-3G) and the simplest population analysis (MPA), performs very well.

Influence of the Class of Molecules. It is shown in Tables 2 and 7 that some classes of molecules are more easily handled by QSPR modeling, while some are more challenging. Specifically, QSPR models work very well for anilines and phenols. These models have high R^2 for all charge calculation approaches and for most of the 3D structure sources. On the other hand, QSPR models provide markedly weaker pK_a predictions for carboxylic acids. Namely, only a few 3D structure sources are applicable for QSPR modeling for carboxylic acids. One reason for the lower quality of QSPR models for the carboxylic acids is that the carboxyl group bound some arbitrary chemical scaffold. In contrast, the −OH group of phenols and $-NH_2$ group of anilines have the same conserved neighborhood—the phenolic ring. In parallel, the phenolic ring also allows higher delocalization of electrons, which is better suited for the calculation of QM descriptors than the more rigid electron localization in carboxylic acids.

Table 7. Number and Percentage of QM QSPR Models with R^2 Higher than a Defined Limit for Individual Classes of Molecules

	R^2			
class of molecules	≥0.9	(0.9, 0.8>	<0.8	R^2_{mol} *
phenols	32%	49%	17%	0.927
carboxylic acids	0%	29%	57%	0.849
anilines	41%	41%	17%	0.929

*R^2_{mol} is the average value of R^2 for all QSPR models, which were built for a given class of molecules.

Influence of 3D Structure Preparation Methodology on Quality of the QM QSPR Model. Tables 2, 8, and 9 show that an appropriate selection of 3D structure source and optimization method is essential for the QSPR modeling of pK_a.

These results imply that the most appropriate 3D structures were obtained from the DTP NCI and Pubchem databases (i.e., structures prepared with the tools CORINA and Omega, respectively). The QSPR models based on these structures are very accurate, and these 3D structures do not require optimization. A great feature of these 3D structures was that they performed very well for all the tested QM charge calculation approaches and classes of molecules. An interesting finding is that the QM optimization of such 3D structures can markedly decrease the accuracy of the models.

Frog2 also seems to be applicable. QSPR models based on 3D structures from Frog2 are accurate even when the structures were not optimized, and the MM optimization of these structures mainly improves the models. They can be successfully used for all the classes of molecules and all the QM charge calculation approaches tested here.

RDKit, OpenBabel, and Balloon are slightly troublesome sources of 3D structures. They can provide accurate QSPR models (R^2 > 0.9) for some classes of molecules. In this case, the MM optimization of 3D structures improves the models. But when we process other classes of molecules (carboxylic acids), the QSPR models are weak (R^2 ∼ 0.85) for most of the charge calculation approaches. For certain charge calculation approaches, the QSPR models can even be unsatisfactory (R^2 < 0.7). An interesting fact is that the structures generated by RDKit with no optimization provide the worst-performing QSPR models of the whole study. The explanation is clear. These 3D structures are just the raw results of RDKit, and as mentioned in its manual, they need to be optimized by RDKit's internal force field UFF. This case study shows how weak QSPR models can be when based on problematic structures.

Particular geometrical properties, which are incorrectly modeled in certain 3D structure preparation methodologies and which cause worse performance of QSPR models, are summarized in the Supporting Information.

Table 8. Percentage of QM QSPR Models with Given R^2 for Individual 3D Structure Sources[a]

		R^2				
source	optimization	≥0.95	(0.95, 0.9>	(0.9, 0.85>	(0.85, 0.8>	<0.8
Balloon	none	0%	42%	8%	42%	8%
	MM	8%	33%	33%	17%	8%
	QM	**8%**	**42%**	**25%**	**17%**	**8%**
Frog2	none	0%	50%	50%	0%	0%
	MM	**33%**	**58%**	**0%**	**8%**	**0%**
	QM	33%	42%	25%	0%	0%
NCI	**none**	**50%**	**42%**	**8%**	**0%**	**0%**
	MM	**50%**	**42%**	**8%**	**0%**	**0%**
	QM	8%	58%	33%	0%	0%
OpenBabel	**none**	**58%**	**8%**	**17%**	**8%**	**8%**
	MM	**58%**	**8%**	**17%**	**8%**	**8%**
	QM	33%	17%	17%	25%	8%
PubChem	**none**	**8%**	**75%**	**17%**	**0%**	**0%**
	MM	**25%**	**50%**	**25%**	**0%**	**0%**
	QM	8%	50%	33%	8%	0%
RDKit	none	0%	0%	33%	25%	42%
	UFF	**42%**	**25%**	**17%**	**17%**	**0%**
	MM	25%	50%	8%	0%	17%
	QM	8%	58%	17%	8%	8%

[a]Optimization procedures that produce the best QSPR models for each source of 3D structures are marked in bold font.

Table 9. Sensitivity of 3D Structure Source to Change of Molecular Class[a]

	percent of insensitive QSPR models					
Optimization	source					
	Balloon	Frog2	NCI	OpenBabel	PubChem	RDKit
none	50%	100%	25%	0%	75%	75%
MM	50%	25%	75%	0%	50%	0%
QM	25%	50%	75%	0%	75%	25%
UFF	–	–	–	–	–	25%
total	42%	58%	58%	0%	67%	31%

[a]Sensitivity of a particular QSPR model to a change of molecular class was analyzed via a statistical test, which compared the correlation coefficient of three independent populations (i.e., molecular classes), employed Fisher's z-transformation, and used the significance level 0.05. Detailed information about this statistical test is in the Supporting Information.

DOI: 10.1021/ci500758w
J. Chem. Inf. Model. 2015, 55, 1088−1097

Semiempirical QM QSPR Models: Quality, Predictivity, and Influences. The results summarized in Table 10 and Table 2 of the Supporting Information show that the quality of these models is comparable to the quality of QSPR models based on ab initio QM charges, just slightly lower for phenols and anilines and slightly better for carboxylic acids. The cross-validation results (Supporting Information, Table S5) confirmed the robustness of the semiempirical QM QSPR models. When we evaluated the influence of the class of molecules and the 3D structure preparation methodology, we saw the same trends as for the *ab initio* QM QSPR models (Table 10 and Table S2, Supporting Information).

Table 10. Number and Percentage of Semiempirical QM QSPR Models with R^2 Higher than a Defined Limit

R^2	≥0.95	(0.95, 0.9>	(0.9, 0.8>	<0.8
number of models	15	25	17	0
percentage of models	26%	44%	30%	0%

Quality of EEM QSPR models: General Summary. The results summarized in Tables 11 and 12 show that the quality of EEM QSPR models is in general lower than for QM QSPR models but still sufficient. Specifically, about 36% of the models are very good quality ($R^2 \geq 0.9$), most of the models are acceptable quality (R^2 between 0.9 and 0.8), and only about 2% are low quality ($R^2 < 0.8$). On the other hand, the number of weak models is lower than for QM QSPR models, and there are no models with ($R^2 < 0.75$).

Predictivity of EEM QSPR Models. A high quality of EEM QSPR models based on the same charge descriptors as our models was shown in ref 8. We tested the predictivity of our EEM QSPR models the same way as we did for the QM QSPR models—by cross-validation and by testing on a larger set of independent molecules. These results are summarized in the Supporting Information (Table S5 and S6, respectively) and confirm that our EEM QSPR models are robust and can handle molecules outside the training set.

Table 12. Number and Percentage of EEM QSPR Models with R^2 Higher than a Defined Limit

R^2	≥0.95	(0.95, 0.9>	(0.9, 0.8>	<0.8
number of models	82	106	38	2
percentage of models	36%	46%	17%	1%

Influence of EEM Parameter Set. The results in Table 11 and Table S7 of the Supporting Information show that all four EEM parameter sets tested here are applicable for pK_a prediction. The quality of the QSPR models obtained by all the EEM parameter sets is comparable. The parameter set Chaves2006 (mimicking B3LYP/6-31G*/MPA charges) performed slightly better than the remaining sets.

Influence of the Class of Molecules. As with QM charges, some classes of molecules are more challenging for the QSPR modeling of pK_a (i.e., carboxylic acids), see Table 11 and Table S8, Supporting Information. Nonetheless, the differences between the quality of EEM QSPR models for various classes of molecules are markedly smaller than for the QM QSPR models.

Influence of 3D Structure Preparation Methodology on Quality of the EEM QSPR Model. Table 8 and Table S6 of the Supporting Information show that EEM QSPR models are markedly less sensitive to the selection of 3D structure source and optimization method.

As with QM QSPR models, 3D structures from DTP NCI and Pubchem can be successfully used for all of the tested molecular classes and all EEM parameter sets, even without optimization (i.e., more than 90% of EEM QSPR models based on nonoptimized NCI 3D structures and all EEM QSPR models based on nonoptimized Pubchem 3D structures have $R^2 > 0.85$).

Frog2 also preforms very well. More than 80% of EEM QSPR models based on nonoptimized Frog2 3D structures have $R^2 > 0.85$. Additionally, these models seem to be

Table 11. R^2 Describing the Correlation between Calculated and Experimental pK_a for EEM QSPR Models

R^2		Class of molecules	Phenols				Carboxylic acids				Anilines				Average
		Charge calculation approach	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	HF, STO-3G, MPA	B3LYP, 6-31G*, MPA	B3LYP, 6-31G*, NPA	B3LYP, 6-31G*, AIM	
Source + Optimization	Balloon	none	0.873	0.904	0.903	0.888	0.832	0.924	0.888	0.853	0.806	0.847	0.826	0.870	0.868
		MM	0.852	0.906	0.907	0.885	0.800	0.917	0.883	0.837	0.867	0.845	0.855	0.880	0.870
		QM	0.869	0.908	0.906	0.890	0.772	0.917	0.889	0.851	0.953	0.930	0.908	0.945	0.895
	Frog2	none	0.907	0.897	0.898	0.858	0.832	0.875	0.831	0.870	0.894	0.879	0.904	0.887	0.878
		MM	0.918	0.906	0.917	0.868	0.859	0.888	0.860	0.848	0.863	0.857	0.852	0.902	0.878
		QM	0.921	0.907	0.918	0.869	0.841	0.898	0.866	0.874	0.939	0.926	0.907	0.939	0.900
	NCI	none	0.906	0.906	0.899	0.890	0.875	0.926	0.891	0.879	0.870	0.852	0.839	0.882	0.884
		MM	0.891	0.926	0.926	0.916	0.860	0.920	0.888	0.829	0.844	0.834	0.848	0.889	0.881
		QM	0.896	0.924	0.925	0.912	0.821	0.923	0.884	0.834	0.921	0.884	0.869	0.920	0.893
	OpenBabel	none	0.900	0.920	0.912	0.908	0.830	0.898	0.848	0.826	0.860	0.849	0.851	0.899	0.875
		MM	0.900	0.919	0.911	0.907	0.827	0.903	0.849	0.835	0.858	0.851	0.857	0.897	0.876
		QM	0.896	0.917	0.911	0.904	0.807	0.911	0.856	0.851	0.946	0.935	0.939	0.934	0.901
	PubChem	none	0.896	0.918	0.913	0.902	0.888	0.891	0.866	0.873	0.874	0.881	0.874	0.907	0.890
		MM	0.887	0.917	0.915	0.899	0.874	0.902	0.876	0.871	0.886	0.852	0.872	0.900	0.888
		QM	0.898	0.921	0.925	0.899	0.825	0.923	0.894	0.892	0.890	0.905	0.867	0.927	0.897
	RDKit	none	0.894	0.907	0.904	0.885	0.836	0.932	0.889	0.874	0.832	0.842	0.840	0.857	0.874
		MM-UFF	0.923	0.917	0.912	0.895	0.801	0.919	0.866	0.844	0.838	0.845	0.843	0.875	0.873
		MM	0.899	0.908	0.902	0.892	0.823	0.907	0.871	0.852	0.846	0.852	0.854	0.897	0.875
		QM	0.909	0.919	0.916	0.895	0.753	0.915	0.881	0.851	0.933	0.892	0.869	0.923	0.888
Average			0.897	0.913	0.911	0.893	0.829	0.910	0.872	0.855	0.880	0.871	0.867	0.902	

Legend	$R^2 \geq 0.95$	$R^2 \geq 0.9$	$R^2 \geq 0.866$	$R^2 \geq 0.833$	$R^2 \geq 0.8$	$R^2 \geq 0.7$

DOI: 10.1021/ci500758w
J. Chem. Inf. Model. 2015, 55, 1088−1097

applicable for all molecular classes and all EEM parameter sets tested here.

For the other four tools, the accuracy of EEM QSPR models depends on the molecular class and EEM parameter set, as certain combinations of these can produce lower accuracy QSPR models.

For all six sources of 3D structures tested in this study, QM optimization produces an improvement in the EEM QSPR models in most cases.

Quality of GM QSPR Models. Gasteiger-Marsili charges do not depend on the 3D structure of molecules; therefore, we prepared only one QSPR model for each class of molecules. The R^2 values of these models are given in Table 13, and

Table 13. R^2 Describing Correlation between Calculated and Experimental pK_a for GM QSPR Models

class of molecules	phenols	carboxylic acids	anilines
R^2	0.747	0.737	0.870

further quality criteria are available in Table S3 of the Supporting Information. These results show that GM QSPR models are markedly less accurate than EEM QSPR models, and therefore, GM charges are not applicable for pK_a prediction. These conclusions are in agreement with results published in the past.[15]

CONCLUSION

Our results confirmed that QSPR models based on QM and EEM partial atomic charges are able to predict pK_a with high accuracy. Specifically, more than 60% of *ab initio* and semiempirical QM QSPR models and nearly 40% of EEM QSPR models are very good quality ($R^2 \geq 0.9$). We also confirmed that *ab initio* and semiempirical QM charges provide very accurate QSPR models and using EEM charges is also acceptable and moreover advantageous because their calculation is very fast. Afterward, we evaluated the predictivity of our QM, semiempirical QM, and EEM QSPR models via cross-validation and via testing on an independent test data set. This way, we verified that all the types of *ab initio* and semiempirical and EEM charges used are applicable for QSPR modeling. On the contrary, QSPR models based on empirical Gasteiger-Marsili charges showed low quality, suggesting that Gasteiger-Marsili charges are not suitable descriptors for the prediction of pK_a.

We then focused on the influence of molecular class. We found that some molecular classes are more amenable to QSPR modeling (phenols and anilines), while some are more challenging (carboxylic acids).

In this context, we compared the influence of the different 3D structure sources. We found that the selection of 3D structure source and optimization method can strongly influence the quality of QSPR models for pK_a prediction. The 3D structures from the DTP NCI and Pubchem databases, i.e., structures generated by CORINA and Omega, respectively, exhibited the best performance. These 3D structures provided very accurate QSPR models for all the tested molecular classes and charge calculation approaches, and they do not require optimization. Frog2 also performed very well for all of the tested molecular classes and charge calculation approaches. Other 3D structure sources can also be used, but they are not so robust. An unlucky combination of molecular class and charge calculation approach can lead to weak QSPR models. Additionally, these structures generally need to be optimized in order to produce high quality QSPR models. Specifically, the best approach is to apply MM optimization to 3D structures used with QM QSPR models and QM optimization to 3D structures used with EEM QSPR models.

The main point of this article is that a workflow for the fast and accurate prediction of pK_a or other important properties for *in silico* designed molecules can be as follows: Preparation of 3D structures by CORINA or Omega (with no further optimization), calculation of EEM charges for these structures, and then the EEM QSPR calculation of pK_a.

ASSOCIATED CONTENT

Supporting Information

List of molecules, including their figures, NCS numbers, CAS numbers, SMILES and experimental pK_a values (Table S1); file with the SMILES of the molecules; description of the procedure, which was used for the creation of dissociated and associated forms of molecules; table summarizing R^2 values of the correlation between calculated and experimental pK_a for semiempirical QM QSPR models (Table S2); quality criteria and statistical criteria of all the QSPR models (Table S3); parameters of all the QSPR models (Table S4); cross-validation results for all QSPR models (Table S5); quality criteria for testing of selected EEM QSPR models (Table S6); number and percentage of EEM QSPR models with R^2 higher than a defined limit for individual charge calculation approaches (Table S7); number and percentage of EEM QSPR models with R^2 higher than a defined limit for individual classes of molecules (Table S8); percentage of QM QSPR models with given R^2 for individual 3D structure sources (Table S9); information about geometrical properties, which are incorrectly modeled in certain 3D structure preparation methodologies; information about the statistical test used for analysis of a 3D structure source sensitivity to a change of molecular class; and information about QSPR models limitations. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/ci500758w.

AUTHOR INFORMATION

Corresponding Authors

*Phone: +420 549 494 860. E-mail: svobodova@chemi.muni.cz (R.S.V.).

*Phone: +420 549 494 947. E-mail: jkoca@chemi.muni.cz (J.K.).

Author Contributions

The authors wish it to be known that, in their opinion, S. Geidl and R. Svobodová Vařeková should be regarded as joint first authors.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (LH13055); the European Community's Seventh Framework Programme (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund and the Capacities specific program (286154); and the European Social Fund and the state budget of the Czech Republic (CZ.1.07/2.3.00/20.0042, CZ.1.07/2.3.00/30.0009). This work was also supported in part by NIH Grants R01 GM071872, U01 GM094612, and U54 GM094618 to R.A. The

DOI: 10.1021/ci500758w
J. Chem. Inf. Model. 2015, 55, 1088−1097

access to MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is greatly appreciated.

REFERENCES

(1) Comer, J.; Tam, K. Lipophilicity Profiles: Theory and Measurement. In *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*; Testa, B., van de Waterbeemd, H., Folkers, G., Guy, R., Eds.; Verlag Helvetica Chimica Acta: Zürich, Switzerland, 2001; pp 275−304.

(2) Klebe, G. Recent Developments in Structure-Based Drug Design. *J. Mol. Med.* **2000**, *78*, 269−281.

(3) Kim, J. H.; Gramatica, P.; Kim, M. G.; Kim, D.; Tratnyek, P. G. QSAR Modelling of Water Quality Indices of Alkylphenol Pollutants. *SAR QSAR Environ. Res.* **2007**, *18*, 729−743.

(4) Lee, A. C.; Crippen, G. M. Predicting pK_a. *J. Chem. Inf. Model.* **2009**, *49*, 2013−2033.

(5) Rupp, M.; Körner, R.; Tetko, I. V. Predicting the pK_a of Small Molecules. *Comb. Chem. High Throughput Screening* **2010**, *14*, 307−327.

(6) Ho, J. Predicting pKa in Implicit Solvents: Current Status and Future Directions. *Aust. J. Chem.* **2014**, *67*, 1441−1460.

(7) Balogh, G. T.; Tarcsay, Á; Keserű, G. M. Comparative Evaluation of pK_a Prediction Tools on a Drug Discovery Dataset. *J. Pharm. Biomed. Anal.* **2012**, *67*−*68*, 63−70.

(8) Svobodová Vařeková, R.; Geidl, S.; Ionescu, C.-M.; Skřehota, O.; Bouchal, T.; Sehnal, D.; Abagyan, R.; Koča, J. Predicting pK_a Values From EEM Atomic Charges. *J. Cheminf.* **2013**, *5*, 18−34.

(9) Fraczkiewicz, R.; Lobell, M.; Göller, A. H.; Krenz, U.; Schoenneis, R.; Clark, R. D.; Hillisch, A. Best of Both Worlds: Combining Pharma Data and State of the Art Modeling Technology To Improve in silico pK_a Prediction. *J. Chem. Inf. Model.* **2015**, *55*, 389−397.

(10) Settimo, L.; Bellman, K.; Knegtel, R. M. A. Comparison of the Accuracy of Experimental and Predicted pK_a Values of Basic and Acidic Compounds. *Pharm. Res.* **2014**, *31*, 1082−1095.

(11) Jelfs, S.; Ertl, P.; Selzer, P. Estimation of pK_a for Druglike Compounds Using Semiempirical and Information-Based Descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 450−459.

(12) Dixon, S. L.; Jurs, P. C. Estimation of pK_a for Organic Oxyacids Using Calculated Atomic Charges. *J. Comput. Chem.* **1993**, *14*, 1460−1467.

(13) Zhang, J.; Kleinöder, T.; Gasteiger, J. Prediction of pK_a Values for Aliphatic Carboxylic Acids and Alcohols with Empirical Atomic Charge Descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 2256−2266.

(14) Citra, M. J. Estimating the pK_a of Phenols, Carboxylic Acids and Alcohols From Semi-empirical Quantum Chemical Methods. *Chemosphere* **1999**, *1*, 191−206.

(15) Gross, K. C.; Seybold, P. G.; Hadad, C. M. Comparison of Different Atomic Charge Schemes for Predicting pK_a Variations in Substituted Anilines and Phenols. *Int. J. Quantum Chem.* **2002**, *90*, 445−458.

(16) Kreye, W. C.; Seybold, P. G. Correlations Between Quantum Chemical Indices and the pK_as of a Diverse Set of Organic Phenols. *Int. J. Quantum Chem.* **2009**, *109*, 3679−3684.

(17) Svobodová Vařeková, R.; Geidl, S.; Ionescu, C.-M.; Skřehota, O.; Kudera, M.; Sehnal, D.; Bouchal, T.; Abagyan, R.; Huber, H. J.; Koča, J. Predicting pK_a Values of Substituted Phenols from Atomic Charges: Comparison of Different Quantum Mechanical Methods and Charge Distribution Schemes. *J. Chem. Inf. Model.* **2011**, *51*, 1795−1806.

(18) Rayne, S.; Forest, K.; Friesen, K. Examining the PM6 Semiempirical Method for pK_a Prediction Across a Wide Range of Oxyacids. *Nat. Precedings* **2009**, http://hdl.handle.net/10101/npre.2009.2981.1.

(19) Gieleciak, R.; Polanski, J. Modeling Robust QSAR. 2. Iterative Variable Elimination Schemes for CoMSA: Application for Modeling Benzoic Acid pK_a Values. *J. Chem. Inf. Model.* **2007**, *47*, 547−556.

(20) Mortier, W. J.; Ghosh, S. K.; Shankar, S. Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. *J. Am. Chem. Soc.* **1986**, *108*, 4315−4320.

(21) Czodrowski, P.; Dramburg, I.; Sotriffer, C. A.; Klebe, G. Development, Validation, and Application of Adapted PEOE Charges to Estimate pK_a Values of Functional Groups in Protein−Ligand Complexes. *Proteins Struct. Funct. Bioinf.* **2006**, *65*, 424−437.

(22) Tehan, B. G.; Lloyd, E. J.; Wong, M. G.; Pitt, W. R.; Montana, J. G.; Manallack, D. T.; Gancia, E. Estimation of pK_a Using Semiempirical Molecular Orbital Methods. Part 1: Application to Phenols and Carboxylic Acids. *Quant. Struct.-Act. Relat.* **2002**, *21*, 457−472.

(23) NCI Open Database Compounds. National Cancer Institute. http://cactus.nci.nih.gov/ (acccesed August 10, 2010).

(24) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. ReV.* **1993**, *93*, 2567−2581.

(25) Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H. In *Annual Reports in Computational Chemistry*; Wheeler, R., Spellmeyer, D., Eds.; Elsevier, 2008; Vol. *4*; Chapter 12.

(26) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(27) Vainio, M. J.; Johnson, M. S. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462−2474.

(28) Leite, T. B.; Gomes, D.; Miteva, M.; Chomilier, J.; Villoutreix, B.; Tuffery, P. Frog: a FRee Online druG 3D conformation generator. *Nucleic Acids Res.* **2007**, *35*, W568−W572.

(29) O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, 33−47.

(30) Landrum, G. RDKit: Open-Source Cheminformatics. http://www.rdkit.org (accessed January 10, 2014).

(31) Gross, K. C.; Seybold, P. G. Substituent Effects on the Physical Properties and pK_a of Phenol. *Int. J. Quantum Chem.* **2001**, *85*, 569−579.

(32) Habibi-Yangjeh, A.; Danandeh-Jenagharad, M.; Nooshyar, M. Application of Artificial Neural Networks for Predicting the Aqueous Acidity of Various Phenols Using QSAR. *J. Mol. Model.* **2006**, *12*, 338−347.

(33) Howard, P.; Meylan, W. *Physical/Chemical Property Database (PHYSPROP)*; Syracuse Research Corporation, Environmental Science Center: North Syracuse, NY, 1999.

(34) Frisch, M. J. et al. Gaussian 09, Revision E.01. Gaussian, Inc.: Wallingford, CT, 2004.

(35) Gront, D.; Kolinski, A. BioShell − A Package of Tools for Structural Biology Computations. *Bioinformatics* **2006**, *22*, 621−622.

(36) Gront, D.; Kolinski, A. Utility Library for Structural Bioinformatics. *Bioinformatics* **2008**, *24*, 584−585.

(37) Svobodová Vařeková, R.; Koča, J. Optimized and Parallelized Implementation of the Electronegativity Equalization Method and the Atom-Bond Electronegativity Equalization Method. *J. Comput. Chem.* **2006**, *3*, 396−405.

(38) Svobodová Vařeková, R.; Jiroušková, Z.; Vaněk, J.; Suchomel, S.; Koča, J. Electronegativity Equalization Method: Parameterization and Validation for Large Sets of Organic, Organohalogene and Organometal Molecule. *Int. J. Mol. Sci.* **2007**, *8*, 572−582.

(39) Chaves, J.; Barroso, J. M.; Bultinck, P.; Carbo-Dorca, R. Toward an Alternative Hardness Kernel Matrix Structure in the Electronegativity Equalization Method (EEM). *J. Chem. Inf. Model.* **2006**, *46*, 1657−1665.

(40) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, J. P. The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. *J. Phys. Chem. A* **2002**, *106*, 7895−7901.

DOI: 10.1021/ci500758w
J. Chem. Inf. Model. 2015, 55, 1088−1097

(41) Bultinck, P.; Vanholme, R.; Popelier, P. L. A.; De Proft, F.; Geerlings, P. High-speed Calculation of AIM Charges Through the Electronegativity Equalization Method. *J. Phys. Chem. A* **2004**, *108*, 10359−10366.
(42) Skřehota, O.; Svobodová Vařeková, R.; Geidl, S.; Kudera, M.; Sehnal, D.; Ionescu, C.-M.; Koča, J. QSPR Designer − A Program To Design and Evaluate QSPR models. Case Study on pK_a Prediction. *J. Cheminf.* **2011**, *3* (Suppl1), P16.
(43) Lemm, S.; Blankertz, B.; Dickhaus, T.; Müller, K.-R. Introduction to Machine Learning for Brain Imaging. *NeuroImage* **2011**, *56*, 387−399.
(44) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties From Structure. *Chem. Soc. Rev.* **1995**, *24*, 279−287.

DOI: 10.1021/ci500758w
J. Chem. Inf. Model. 2015, 55, 1088−1097

AtomicChargeCalculator: interactive web-based calculation of atomic charges in large biomolecular complexes and drug-like molecules

Ionescu *et al. J Cheminform (2015) 7:50*
DOI 10.1186/s13321-015-0099-x

Journal of Cheminformatics

SOFTWARE Open Access

AtomicChargeCalculator: interactive web-based calculation of atomic charges in large biomolecular complexes and drug-like molecules

Crina-Maria Ionescu[1†], David Sehnal[1,2,3†], Francesco L. Falginella[1], Purbaj Pant[2], Lukáš Pravda[1,2], Tomáš Bouchal[1,2], Radka Svobodová Vařeková[1,2], Stanislav Geidl[1,2] and Jaroslav Koča[1,2*]

Abstract

Background: Partial atomic charges are a well-established concept, useful in understanding and modeling the chemical behavior of molecules, from simple compounds, to large biomolecular complexes with many reactive sites.

Results: This paper introduces AtomicChargeCalculator (ACC), a web-based application for the calculation and analysis of atomic charges which respond to changes in molecular conformation and chemical environment. ACC relies on an empirical method to rapidly compute atomic charges with accuracy comparable to quantum mechanical approaches. Due to its efficient implementation, ACC can handle any type of molecular system, regardless of size and chemical complexity, from drug-like molecules to biomacromolecular complexes with hundreds of thousands of atoms. ACC writes out atomic charges into common molecular structure files, and offers interactive facilities for statistical analysis and comparison of the results, in both tabular and graphical form.

Conclusions: Due to high customizability and speed, easy streamlining and the unified platform for calculation and analysis, ACC caters to all fields of life sciences, from drug design to nanocarriers. ACC is freely available via the Internet at http://ncbr.muni.cz/ACC.

Keywords: Conformationally dependent atomic charges, Biomacromolecules , Drug-like molecules, Paracetamol, Benzoic acids, Protegrin, Proteasome, Allostery, Chemical reactivity

Background

Partial atomic charges are real numbers meant to quantify the uneven distribution of electron density in the molecule, and have been used for decades in theoretical and applied chemistry in order to understand the chemical behavior of molecules. Atomic charges are extensively used in many molecular modeling and chemoinformatics applications. With respect to biomacromolecules, charges can elucidate electrostatic effects critical for long range molecular recognition phenomena, protein folding, dynamics and allostery, directed adduction of substrates and egression of products in enzymes, ligand binding and complex formation for proteins and nucleic acids, etc. [1–3]. With respect to drug-like molecules, atomic charges provide information related to reactivity and can be used in the prediction of various pharmacological, toxicological or environmental properties [4, 5].

Although, in principle, it is possible to estimate atomic charges based on experimental measurements (e.g., [6, 7]), such calculations are impractical. Most commonly, atomic charges are estimated based on theoretical approaches. Quantum mechanical (QM) approaches first solve the Schrödinger equation [8] and calculate the electron density using a combination of theory level and basis set. They then partition the obtained molecular

*Correspondence: jaroslav.koca@ceitec.muni.cz
†Crina-Maria Ionescu and David Sehnal contributed equally
[2] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic
Full list of author information is available at the end of the article

electron density (or a density-derived quantity) into atomic contributions (atomic partial charges) according to various population analyses [9–19]. Empirical approaches to atomic charge calculation (e.g., [20–27]) have been proposed as resource-efficient alternatives to QM approaches, as they do not require the demanding step of solving the Schrödinger equation. In particular, approaches based on the equalization of molecular electronegativity [22, 23, 28–35] are of interest because they are sensitive to both the chemical environment and molecular conformation.

Due to the essential role of atomic charges, many modeling tools currently include atomic charge calculation capabilities (e.g., [36–50]). However, in the case of drug-like molecules, only a few tools can provide QM quality charges which respond to changes in conformation or chemical environment without needing to first obtain the QM electron density or electrostatic potential [47, 48, 50]. Moreover, these tools are not sufficiently general, resource-efficient or interactive. In the case of biomacromolecules, no freely available software tool can readily provide atomic charges of QM quality, despite repeated reports that such quality is necessary [51–54]. We have accepted these challenges and set out to provide a robust and accessible software solution for atomic charge calculation for molecules of all nature and size.

This contribution presents the AtomicChargeCalculator (ACC), a free web application for the calculation and analysis of atomic charges which respond to changes in molecular conformation and chemical environment. The calculation is based on the electronegativity equalization method (EEM [22]), a powerful empirical approach which can provide atomic charges similar to those generated by various QM approaches, but using much lower computational resources. Along with the classical EEM algorithm, ACC implements two additional EEM approximations with increased efficiency, specifically tailored for studying very large molecular systems. A single calculation may take from less than a second (small molecules), to a few minutes (large biomacromolecular complexes). ACC outputs the most common molecular structure formats containing atomic charges. Additionally, it provides facilities for statistical analysis and comparison of the results, in tabular and graphical form. ACC also includes interactive 3D visualization of the molecules based on atomic charges. A command line version is also available.

Implementation

The challenge was to provide a robust web based software solution for atomic charge calculation for molecules of all nature and size. Therefore, we first focused on identifying and optimizing a suitable algorithm for atomic charge calculation, and then on implementing the optimal workflow for setting up an ACC calculation and interpreting the results.

The application was constructed using the client-server architecture (Fig. 1): the charge computation is carried out on the server and implemented in the C# programming language. The JavaScript Object Notation (JSON) is used to transfer data to the client that provides the user interface (UI) implemented using HTML5 and JavaScript. Additionally, the UI uses the WebGL technology to provide a custom built 3D visualization of the computed charges.

Computational details

The Electronegativity Equalization Method (EEM) is the general approach followed by ACC to calculate atomic charges. EEM-based methods have been successfully applied to zeolites and metal-organic frameworks, small organic molecules, polypeptides and proteins [55–61].

EEM is an empirical approach which relies on parameters usually fitted to data from reference QM calculations. The values of atomic charges computed using EEM support chemical reasoning, and generally correlate well with values from reference QM calculations. The accuracy of each set of EEM parameters is documented in the respective literature. On the other hand, classical EEM approaches incorrectly predict superlinear scaling of the polarizability with increasing molecular size, making the models developed on small molecules difficult to transfer to extended systems like biomacromolecules [62, 63]. This artifact can be

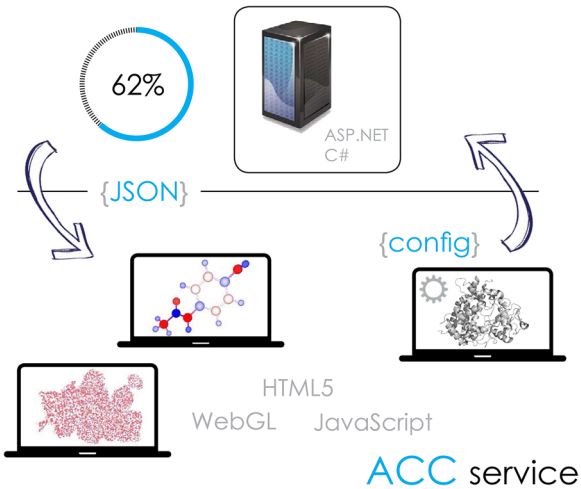


Fig. 1 ACC application architecture. The client allows the user to setup the calculation via the user interface. The settings are sent to the web server in the form of a configuration file. The charge computation takes place on the web server. The results are sent back to the client for visualization and download by the user

tempered by applying charge conservation constraints to small molecular units. Such extensions to EEM have been proposed [24, 64].

Computationally efficient implementations of EEM-based methods are integrated in tools specialized for reactive molecular dynamics simulations [65] and for generating conformers of drug-like molecules [50, 66]. However, ACC is the first to implement EEM in a manner which is not only computationally efficient, but also independent of the subsequent intended application, and specifically designed to allow users with little background in computational sciences to run charge calculations and interactively analyze the results.

ACC can solve the EEM matrix equation (see the Computational details section of the Additional file 1) if the following input is provided: the 3D structure of the molecular system, the total molecular charge, and a set of EEM parameters. Solving the EEM matrix equation requires solving a dense system of equations. The computational complexity of this procedure is $\mathcal{O}(N^3)$. The space complexity, which refers to the memory required to store the EEM matrix, is $\mathcal{O}(N^2)$, where N is the number of atoms. For very large molecules with tens of thousands of atoms, the EEM approach is too demanding on conventional desktop hardware. We thus propose two new approaches for solving the EEM matrix, namely EEM Cutoff and EEM Cover. These approaches work by splitting the EEM matrix into multiple smaller matrices.

Within the EEM Cutoff approach, for each atom in the molecule, ACC generates a fragment made up of all atoms within a cutoff radius R of the original atom. Thus, for a molecule containing N atoms, the EEM Cutoff approach solves N smaller EEM matrices, for a set of N overlapping fragments of the original molecule. EEM Cutoff effectively reduces the time complexity of the calculation to $\mathcal{O}(R^6N + R^2NlogN)$, and the space complexity to $\mathcal{O}(R^4N + NlogN)$. A detailed description of the EEM Cutoff approach is given in the Computational details section of the Additional file 1.

To further enhance the run-time and memory efficiency of calculations in ACC, we propose EEM Cover, an approach for tackling molecules with hundreds of thousands of atoms. EEM Cover also splits the EEM matrix into smaller matrices, but it generates fragments only for a subset of atoms in the molecule. While the asymptotic complexity remains the same, the number of EEM matrices that need to be solved is reduced by at least 50 % compared to EEM Cutoff, while maintaining high accuracy. A detailed description of the EEM Cover approach is given in the Computational details section of the Additional file 1.

Workflow

The ACC workflow is organized into four phases, namely: upload, setup, calculation and results. Each phase is characterized by a set of operations as follows:

1. *Upload molecules* Multiple molecules can be uploaded in the most common file formats (PDB, PDBx/mmCIF, PQR, MOL, MOL2, SDF, or .zip with multiple files of a suitable format). The molecular structures should be complete and properly protonated. There is no limitation regarding the size, number or nature of the chemical entities in a single structure file (proteins, nucleic acids, ligands, water, etc.), as all these are loaded and identified as a single molecule within ACC. The total size of the upload is limited to 50 MB.
2. *Setup* Upon uploading the molecule(s), ACC parses the molecular structure to identify the number and types of atoms in the system, as well as the interatomic distances. Based on this information, ACC tries to prefill the submission form with suitable default settings (see the Default settings section of the Additional file 1). These settings can be adjusted by the user before the calculation is started. Each distinct setup (Fig. 2) will result in a certain number of ACC

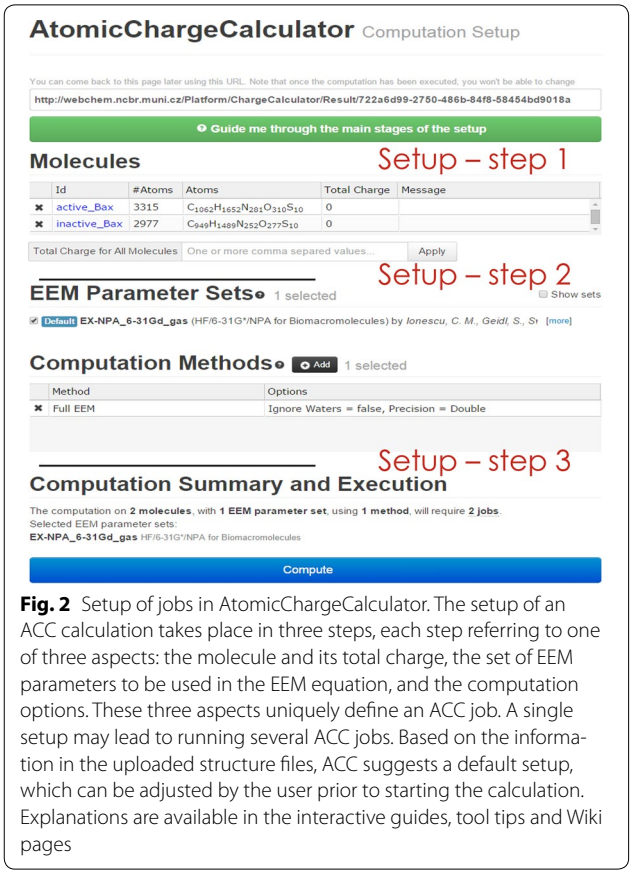


Fig. 2 Setup of jobs in AtomicChargeCalculator. The setup of an ACC calculation takes place in three steps, each step referring to one of three aspects: the molecule and its total charge, the set of EEM parameters to be used in the EEM equation, and the computation options. These three aspects uniquely define an ACC job. A single setup may lead to running several ACC jobs. Based on the information in the uploaded structure files, ACC suggests a default setup, which can be adjusted by the user prior to starting the calculation. Explanations are available in the interactive guides, tool tips and Wiki pages

jobs, each defined by the molecule, total molecular charge, the set of EEM parameters, and the computation options. For the command line version of ACC, the setup workflow is identical to the steps described below, and is scripted into a configuration file.

2.1. *Total molecular charge* The total molecular charge quantifies the amount of charge that will be distributed across the molecule during the EEM calculation. By default, ACC assumes that all uploaded molecules are neutral. The user must provide the correct total molecular charge for each non-neutral molecule uploaded.

2.2. *Set of EEM parameters* EEM employs special parameters for each type of atom (H, C, N, O, halogens, metals, etc., depending on the target molecules). EEM parameters are generally developed based on reference QM calculations. The applicability domain of a given EEM parameter set is generally limited to the target molecules, and closely related to the applicability domain of the particular QM approach used as reference. Performance is further influenced by the procedure used when fitting the EEM parameters to the reference data. Many EEM parameter sets have been published in literature, and are available in ACC as built-in sets [28, 34, 67–70] with full information regarding the parameter development procedure (atom types covered, target molecules, QM reference data, literature reference). By default, ACC tries to select one of these sets based on the atom types present in the uploaded molecules. The user can select a different set of EEM parameters by choosing from the list of available built-in sets, or even uploading customized sets in an XML template. Multiple sets of EEM parameters can be tested in a single ACC run.

2.3 *Computation options* ACC may compute atomic charges based on one of the three available EEM approaches implemented, namely Full EEM, EEM Cutoff, and EEM Cover. Further options refer to the precision (64 or 32-bit representation of numbers), cutoff radius parameter, and including water molecules into the calculation. By default, ACC picks computation options most suitable to the size of the uploaded molecules. These computation options can be adjusted by the user. Up to 10 computation options can be tested in a single ACC run.

3. *Calculation* Once the setup phase is complete, the calculation is launched. A single ACC run may consist of multiple atomic charge calculation jobs. Each job is uniquely defined by the molecule, total molecular charge, set of EEM parameters, and computation options, and produces one set of atomic charges. Each job may use a different amount of time and memory resources, depending on the size of the molecule and the complexity of the computation.

4. *Results* The ACC results are organized into hierarchical reports which are stored on the server for download or inspection for up to a month, at a unique URL visible only to the user. The command line version of ACC produces the same overall and single molecule reports described below, but does not facilitate interactive 3D visualization.

4.1. The *overall report* contains information and downloadable content (molecular structure files containing atomic charges, statistics of the results, information about all jobs) for all molecules. Single molecule reports are also accessible from here.

4.2. The *single molecule report* (Fig. 3), which can be downloaded or examined directly in the browser, consists of a few sections:

4.2.1. *Summary report* containing general information about the input molecule (molecular formula, total charge), calculation setup, a list of all sets of charges produced during the calculation, information about all ACC jobs (duration, warnings, errors) for that molecule.

4.2.2. *Interactive list of values* for all sets of charges produced by all ACC jobs for that molecule. The atomic charges and residue charges are given.

4.2.3. *Statistics within each set of charges*, in both tabular and graphical form. The statistics are available for both atomic and residue charges, and are computed for relevant properties such as chemical element, type of residue, etc. The statistical indicators are the minimum value, maximum value, standard deviation, average, median, etc.

4.2.4. *Pairwise comparison statistics* between sets of charges resulted from different ACC jobs, or uploaded by the user. A graphical representation for each comparison is also provided. The comparison is available for atomic and residue charges. The comparison indicators computed are the squared Pearson's correlation coefficient, Spearman's rank correlation coefficient, RMSD, sum of absolute differences.

4.2.5. *Interactive* 3D visualization of molecules. The 3D model can be built based on atomic positions,

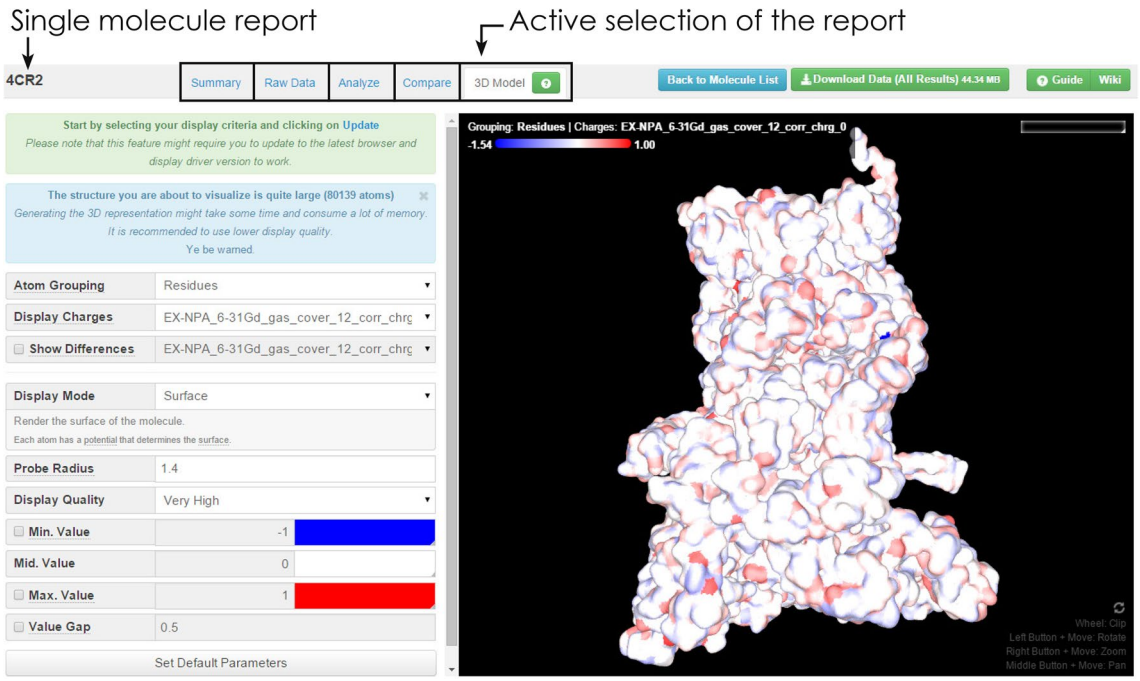


Fig. 3 Single molecule report provided by AtomicChargeCalculator. It can be downloaded or inspected in the browser, and consists of several sections: summary information about the molecule, interactive list of values for all sets of charges, statistics within each set of charges, pairwise comparison statistics between sets of charges, and interactive 3D visualization of molecules

and colored based on atomic charges, or built based on residue positions, and colored based on residue charges. The coloring scheme can also use differences in charges resulted from distinct ACC jobs, or uploaded by the user.

The applicability of ACC is limited by three main aspects: related to the concept of atomic partial charges and its definitions, related to the concept of EEM and its parameters, and related to the 3D structure of the molecule and its total charge. These aspects are discussed in detail in the Limitations section of the Additional file 1.

Full documentation explaining the methodology, functionality and interface, along with interesting examples are provided on the web page. Embedded interactive guides assist first-timers and beginners in setting up their calculations and interpreting the results. A command line version of the application is available as an executable for users who wish to streamline more complex calculations.

Results and discussion

Implementation benchmark

We have evaluated the accuracy and computational efficiency of the EEM Cutoff and EEM Cover approaches in a benchmark. The evaluation was performed against reference calculations which solved the full EEM matrix, with a few exceptions. We give here a brief overview (Fig. 4), whereas the full details can be found in the Benchmark section of the Additional file 1. Both EEM Cutoff and EEM Cover are sufficiently accurate, but EEM Cutoff is slightly more accurate. Using a cutoff radius of 8 Å may lead to deviations of up to 0.015*e*, but on average less than 0.008*e*. Using a cutoff radius of 12 Å may lead to deviations of up to 0.008*e*, but on average less than 0.004*e*. The approaches are time efficient compared to Full EEM only when the molecule contains at least 10,000 atoms, but they are always more memory efficient.

Below we provide a few brief examples of uses for AtomicChargeCalculator in the form of case studies. These case studies are focused on a direct interpretation of the ACC results, and show how important hints about the reactivity of a molecule can be obtained in just a few seconds.

Case study I: atomic charges and chemical reactivity in small drug-like molecules

N-acetyl-p-aminophenol, commonly known as paracetamol, is a widely used analgesic and antipyretic. Its mechanism of action is believed to be the inhibition of the protein cyclooxygenase 2, regulating the production of

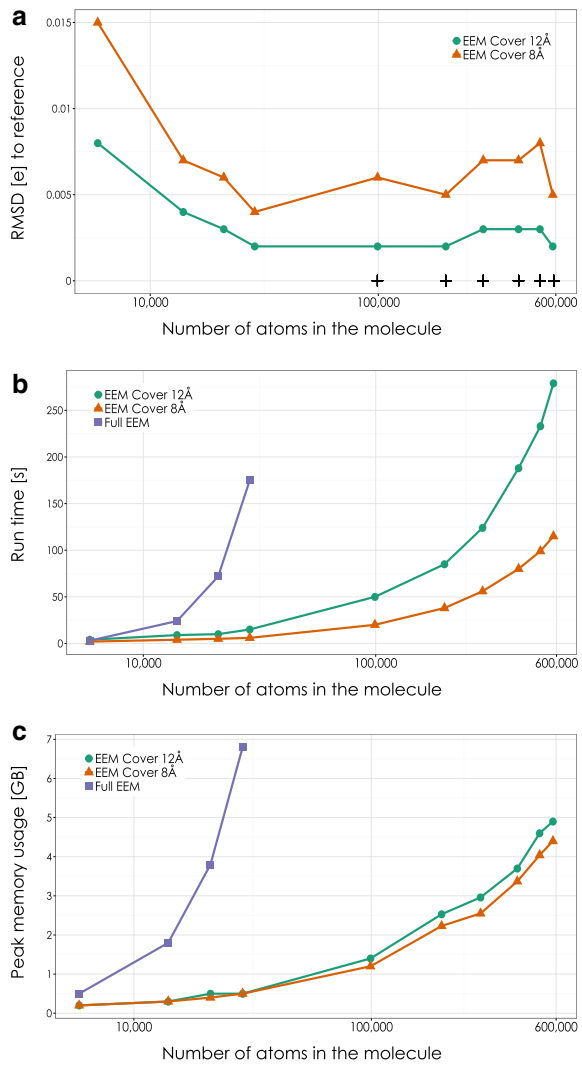


Fig. 4 Benchmark of the EEM Cover approach. Full details can be found in the Benchmark section of the Additional file 1. **a** EEM Cover is sufficiently accurate, and its accuracy increases with the value of the cutoff radius. The root mean square deviation (RMSD) is calculated by comparison against the reference calculation which solves the entire EEM matrix (Full EEM), using 64-bit precision numbers. Due to the limitation on computational resources, for some molecules the reference calculation was *EEM Cutoff, with a cutoff radius of 17 Å. **b** EEM Cover is significantly faster than Full EEM only for molecules with over 10,000 atoms. **c** EEM Cover is always more memory efficient than Full EEM

pro-inflammatory compounds [71]. The metabolic breakdown of paracetamol has been the subject of intense study, since it holds the key to both its therapeutic action and toxicity.

We calculated atomic charges in paracetamol using ACC. The geometry of the paracetamol molecule corresponded to the ideal coordinates [wwPDB CCD: TYL]. The default ACC settings were used. The computation took less than 1s, and the complete results are available on the ACC web page at http://ncbr.muni.cz/ACC/CaseStudy/Paracetamol.

A quick analysis reveals that the phenolic H (position HO4 in Fig. 5) is the most acidic proton (highest positive charge) in the molecule, suggesting a faster and easier dissociation of this O–H bond. Indeed up to 90 % of metabolic degradation happens at position HO4 (glucuronidation, sulphonation) [72]. Additionally, up to 15 % of metabolic degradation involves oxidation at the phenolic (HO4) and amidic positions (HN) [72], the two most positive H in the paracetamol molecule. While paracetamol is a very small molecule with few polar sites, the same principle can be applied in reasoning out highly reactive sites in more complex molecules.

Having found out that the most probable dissociation site on the paracetamol molecule is the phenolic H, we were able then to calculate the acid dissociation constant pK_a, a property which significantly affects the ability of the drug to cross cellular membranes and thus exert its therapeutic effect. For this purpose we used Quantitative Structure-Property Relationship (QSPR) modeling, as atomic partial charges have been shown to be successful QSPR descriptors in pK_a prediction [58, 73, 74], and QSPR models are available in literature for this purpose. Because the dissociating group on paracetamol is phenolic, we chose a QSPR model specifically developed for the prediction of pK_a in phenols [58], and which employed descriptors based on the EEM charges we computed in our interpretation of local reactivity in paracetamol. The necessary descriptors consisted of the partial charges on the phenolic oxygen (q_{O4}) and hydrogen (q_{HO4}), and on the carbon atom binding the phenolic group (q_{C4}). The equation and parameters of the QSPR model are given in Fig. 5.

We thus computed a pK_a value for paracetamol of 9.36, which is close to the experimental value of 9.38 [75]. The computed pK_a suggests that paracetamol is completely unionized at stomach pH, and only 1.1 % ionized at physiological pH, therefore highly efficient at crossing cellular membranes both via oral and intravenous delivery.

While the above described approach was able to provide useful information for paracetamol, it is important to keep in mind that there are limitations to the accuracy of EEM charges. We illustrate such limitations on a series of benzoic acid derivatives. For this purpose, we downloaded the structures of 45 molecules representing benzoic acid derivatives from the NCI Open Database [76]. The data set contained benzoic acids with a wide range of donating and accepting substituents on the phenyl ring in o-, m- and p- positions (Fig. 6a; Additional file 1: Table S3). We chose only compounds for which pK_a values were available in Physprop [77]. Furthermore, we did

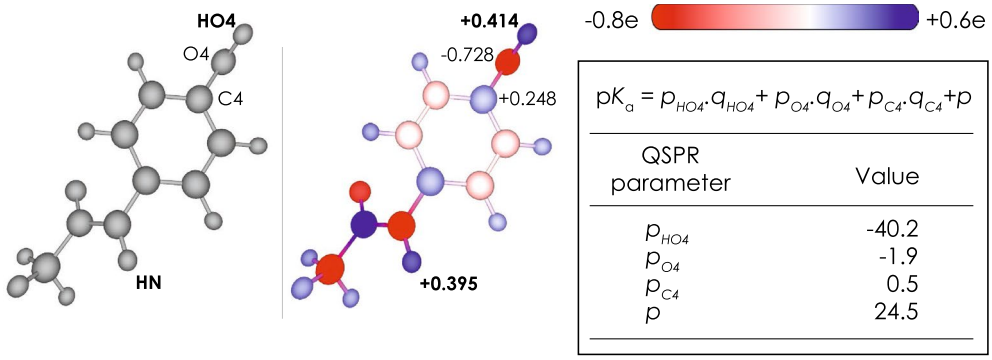


Fig. 5 Atomic charges reveal reactive sites involved in the metabolic degradation of paracetamol. In the picture on the right, the atoms are colored according to their charge. The majority of metabolic degradation of paracetamol (glucuronidation, sulphonation, oxygenation) involves the phenolic (HO4) and amidic (HN) positions, the two most acidic protons in the paracetamol molecule (labels marked in bold) [72]. The relatively higher positive charge on these H atoms marks more active bonds compared to the rest of the molecule. In the equation of the QSPR model, q_{HO4} is the charge on the phenolic H, q_{O4} is the charge on the phenolic O, q_{C4} is the charge on the C binding the phenolic OH, and p_{HO4}, p_{O4}, p_{C4} and p are the parameters of the QSPR model, taken from the QSPR model 3d EEM Ouy2009_elem [58]

not include compounds with halogens because the EEM parameter set used in this particular case study cannot treat halogens.

We first wanted to know if and how the effect of different substituents in different positions on the phenyl ring is visible on the charge of the atoms of the carboxyl group. For this purpose, we used Gaussian [36] to compute reference QM atomic charges from a natural population analysis on the electron density obtained at the B3LYP/6-31G* level of theory. Despite the different nature and position of the substituents, the spread of reference QM charges for the atoms in the carboxyl group is very narrow (Fig. 6b). Specifically, the QM values for O1 are within 0.06*e*, the values for O2 within 0.01*e*, and the values for H within 0.01*e*. On the other hand, the EEM parameter set employed here is expected to reproduce the reference QM values within 0.09*e* [34]. Based on the documented accuracy, we do not expect EEM charges to reflect suitable changes based on the nature or position of the substituents. We used ACC to compute EEM charges for the benzoic acid derivatives, in the same manner as for paracetamol. The computation took less than 2s, and the complete results (structures, QM charges, EEM charges) are available on the ACC web page at http://ncbr.muni.cz/ACC/CaseStudy/BenzoicAcids.

Indeed, we found that EEM charges could not accurately reflect the QM spread for the charges on the carboxyl atoms (Fig. 6b). We then wondered if this accuracy, though unable to reflect suitable changes depending on the nature or position of the substituents, was sufficient to build acceptable QSPR models for pK_a prediction. No suitable QSPR models are available for benzoic acids, but such models for aliphatic carboxylic acids have been reported [58, 78]. The descriptors used by these QSPR models consist of charges on the atoms of the carboxylic group in both the neutral (q_H, q_{O1}, q_{O2}, q_{C1}) and dissociated forms (q_{O1D}, q_{O2D}, q_{C1D}). We thus built QSPR models for benzoic acids based on these descriptors (Fig. 6c). We obtained the structures of the dissociated acids by removing the carboxylic H atoms. We computed EEM charges for the dissociated molecules using the same ACC setup, but setting the total charge for each molecule at −1. We then built QSPR models using multilinear regression. We performed a 5-fold cross-validation of the QSPR models, whereby, in each round, 35 randomly chosen molecules were used to train the model, and the remaining 10 molecules were used to validate the model. The QSPR model parameters and full details of the cross-validation procedure are given in the Additional file 1: Table S4. The models showed adequate predictive capability (Fig. 6c; Additional file 1: Table S4). On average, the mean absolute error during validation was 0.27 pK_a units compared to experiment, suggesting that EEM charges can be used to predict dissociation constants for benzoic acids, despite their inability to reflect local changes caused by different substituents.

Case study II: atomic charges and activity of antimicrobial peptides

Protegrins are a family of antimicrobial peptides active against a wide range of pathogens [79]. Protegrin-1 (PG1, Fig. 7) has been intensely studied for its potential in treating infections caused by antibiotic resistant bacteria [80–82]. PG1 shows activity against several pathogens, but also toxicity against the host. Useful mutations are those which maintain the antimicrobial activity, and at the same time reduce toxicity [83].

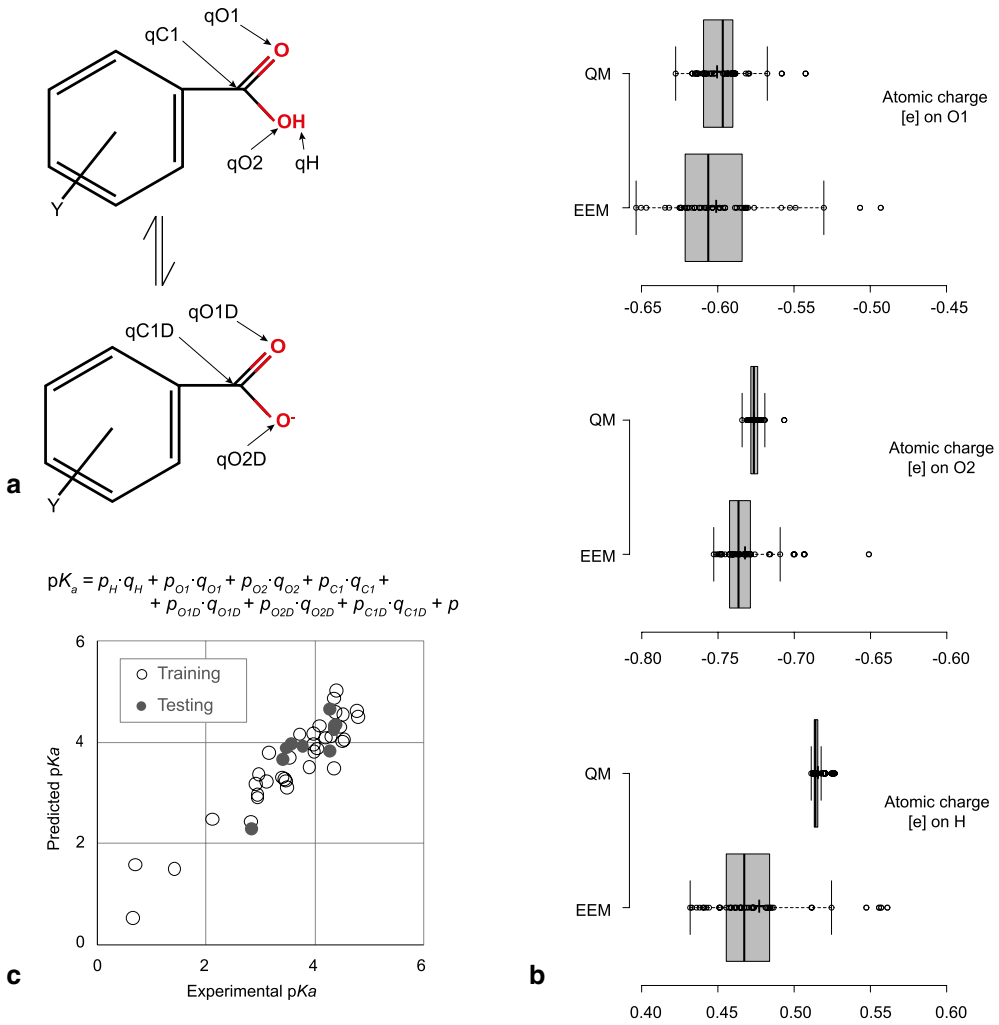


Fig. 6 Limitations of EEM charges illustrated on a series of substituted benzoic acid derivatives. **a** Denotation of relevant atomic charges in the neutral and dissociated molecules. **b** The reference QM charges have a narrow spread despite the different position and wide range of electron donating or withdrawing effects of the subsituents. EEM charges are not accurate enough to reflect the small changes induced by different substituents. **c** The QSPR descriptors are EEM charges of the atoms of the carboxylic group in both the neutral (q_H, q_{O1}, q_{O2}, q_{C1}) and dissociated forms (q_{O1D}, q_{O2D}, q_{C1D}). The symbols p_H, p_{O1}, p_{O2}, p_{C1}, p_{O1D}, p_{O2D}, p_{C1D}, and p are parameters of the QSPR model. The graph displays the correlation between experimental pK_a values, and the values predicted by one of the QSPR models developed in this study. EEM charge descriptors are sufficiently accurate for the prediction of dissociation constants of benzoic acid derivatives

We calculated atomic charges in PG1 using ACC. The geometry of the PG1 molecule corresponded to a low energy NMR model [PDB: 1PG1] [84]. The system contained all H atoms expected at pH 6.5, as they were listed in the NMR model. The total molecular charge was +7, owing to the many ARG residues. The EEM parameter set used was EX-NPA_6-31Gd_gas [70], but it was necessary to add to this set EEM parameters for deuterium (D), because this element was present in the input file. The EEM parameters for D were identical to the parameters for H. The rest of the ACC default settings were kept. The computation took less than 1s, and the complete results are available on the ACC web page at http://ncbr.muni.cz/ACC/CaseStudy/Protegrin.

The calculation produced one set of atomic charges. PG1 contains 18 residues, and rather than analyzing atomic charges, we analyzed the residue charges, which are also reported by ACC (Fig. 7b). PG1 is special because of its high positive charge. It contains 6 ARG residues. However, not all have the same charge. In particular, ARG at positions 4 and 9 have the least positive charge (around +0.5*e*), whereas the rest have much higher positive charge (over +0.8*e*). Keeping in mind that these charges are likely affected by the polarizability

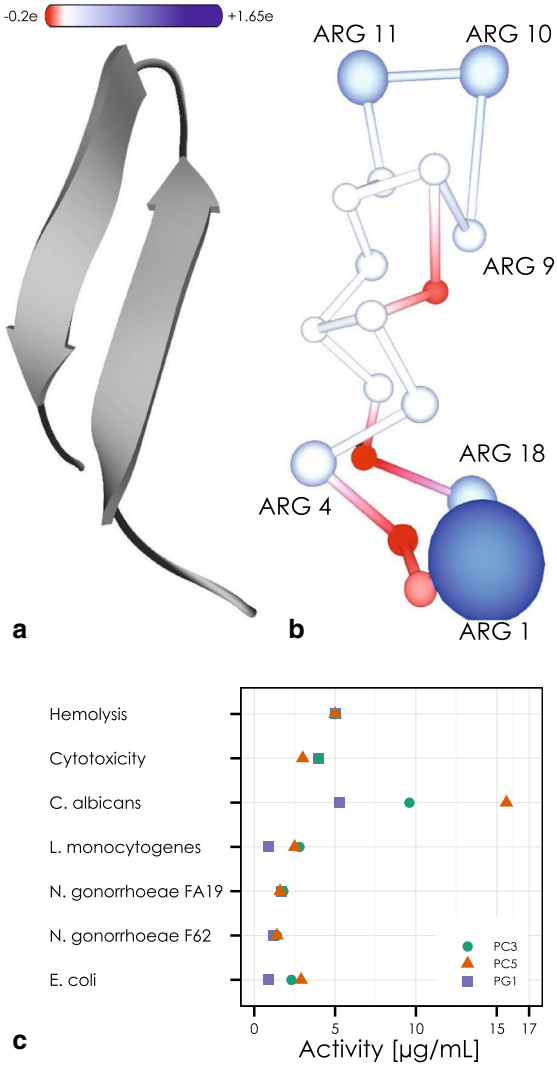


Fig. 7 Residue charges in protegrin-1 (PG1) indicate relevant mutation sites. **a** Cartoon representation of the structure of PG1. **b** Schematic representation of the PG1 residues and their charge. Each residue position is represented by a *sphere*, the coordinates of which correspond to the average coordinates of all atoms in the residue. The *color* of the sphere is given by the residue charge, and the *size* of the sphere is proportional to the absolute charge. Cystein bridges are also displayed. The ARG residues at positions 4 and 9 are significantly less positive than the rest of the ARG residues, indicating that mutations at these positions may be less effective. **c** Antimicrobial activity (against five pathogens) and toxicity (cytotoxicity, hemolysis) of PG1 and two single mutants, namely PC3 (R4G) and PC5 (R10P) [83]. Although both mutants have increased activity, the mutation of ARG 4 in PC3 alters the antimicrobial activity against *C. albicans* significantly less than the mutation of ARG 10 in PC5, as predicted by the residue charges

exaggeration artifact of EEM described in the Computational details section, the results suggest that mutations of ARG into a neutral residue at positions 4 or 9 would have a lower effect than mutations at positions 1, 10, 11 or 18.

A literature search reveals that many mutants and derivatives of PG1 have been studied [83, 85–87]. In particular, Ostberg and Kastnessis [83] have logged the antimicrobial and toxic activity of sixty-two PG1 mutants and PG1-analogue peptides. Out of these, there are two single point mutants where one ARG was mutated into a neutral residue. These mutants are PC3 (R4G) and PC5 (R10P). The study found that, indeed, the mutation of ARG 4 in PC3 alters the antimicrobial activity against *C. albicans* significantly less than the mutation of ARG 10 of PC5 (Fig. 7c). Such biologically relevant insight can be gained by analyzing the residue charges on a single structure of PG1.

Case study III: atomic charges and allostery of large biomacromolecular complexes

The 26S proteasome is a large biomacromolecular complex which facilitates the targeted degradation of intracellular proteins, and thus plays an essential role in keeping protein homeostasis [88]. It consists of a core particle, made up of alpha rings and beta rings, controlled by regulatory particles, which are made up of a number of proteins (Fig. 8a). The proteasome is an intricate molecular machine which requires complex regulation to unfold and deubiquitylate the substrate, and push it through the catalytic machinery located in the beta rings [89]. Necessarily, the proteasome undergoes large conformational changes during its operation. However, due to its size, such changes are very difficult to study. Recent work in the field of cryo-electron microscopy [90] has led to the discovery of intermediate conformers during the initial binding of ubiquitylated substrates. While the conformational changes in the regulatory particle are easily distinguishable (average backbone atom RMSD 10.4 Å), the changes in the core particle are very subtle (average backbone atom RMSD 1.5 Å), due to the fact that all studied conformers refer to the initial phase of substrate binding.

Using ACC, we calculated atomic charges in these intermediate conformers of the 26S proteasome [PDB: 4CR2, 4CR3, 4CR4]. The default ACC settings were used. The computation took 130s, and the complete results are available on the ACC web page at http://ncbr.muni.cz/ACC/CaseStudy/Proteasome.

The calculation produced one set of atomic charges for each conformer of the 26S proteasome. Since the proteasome is very large, we analyzed the residue charges, which are also reported by ACC, and subsequently the charges for the various subunits that make up the proteasome (details regarding the charge analysis on subunits can be found in the section Case study III of the Additional file 1). The first observation was that, during

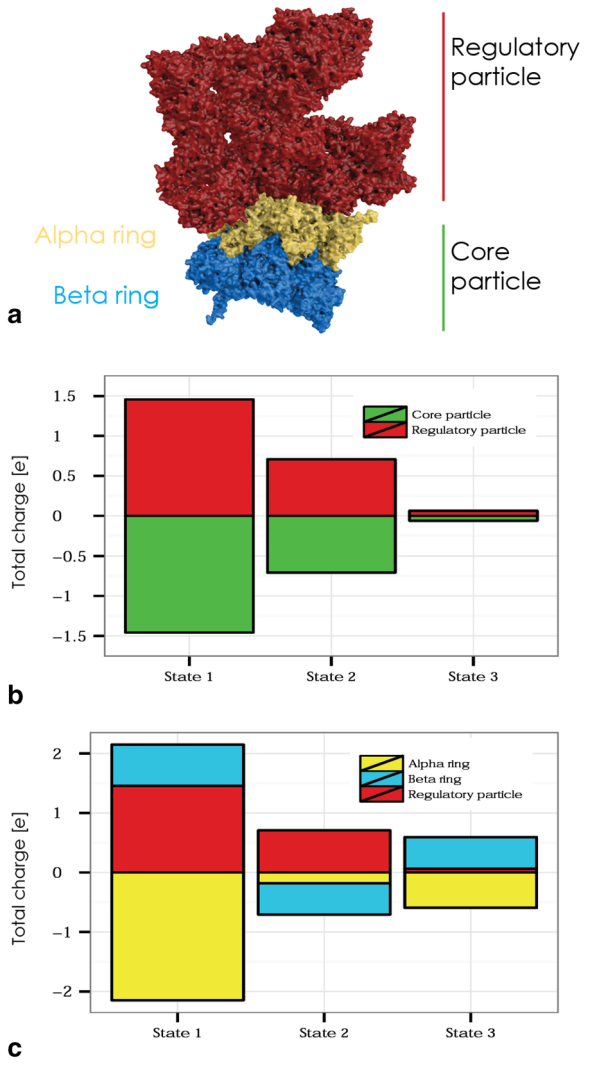


Fig. 8 Using information about total subunit charge to track allosteric regulation in the 26S proteasome. **a** Surface representation of the structure of half of the 26S proteasome in state 1 [PDB: 4CR2]. The regulatory particle (*red*) enables the substrate to unfold and enter the core particle, which is made up of alpha (*yellow*) and beta rings (*blue*). **b** As the proteasome evolves from state 1 [PDB: 4CR2], through state 2 [PDB: 4CR3], to state 3 [PDB: 4CR4], the movement of the regulatory particle is translated into information which is exchanged with the core particle. The difference in total particle charge is a clear indication of regulation, despite the negligible movement observed in the core particle. **c** Information is transferred not just within each particle, but also vertically, between the alpha and the beta ring, and between the alpha ring and the regulatory particle. This suggests that vertical structures such as pairs of alpha-beta subunits may act in a correlated manner

the conformational changes from state 1, to state 2, and then to state 3, a significant amount of electron density is transferred between the core particle and the regulatory particle (Fig. 8b). This suggests that, even though there is no significant movement observed for the core particle, allosteric information is exchanged with the core particle, and this information can be tracked at the electrostatic level. The next observation is that significant information is disseminated not only horizontally (within the alpha ring, or within the beta ring), but also vertically. In the overall transition it appears that the alpha ring loses electron density to the regulatory particle. By checking the intermediate state 2 it is possible to see that there is also transfer between the alpha ring and beta ring (Fig. 8c). This vertical shuttling of electron density within the core particle suggests that the activity of alpha and beta subunits may cross-correlate. Such phenomena have indeed been reported. For example, alpha5 and beta1 may translocate together [91], while knockdown of alpha1 leads to loss of chymotrypsin activity associated with beta5 [92]. Further analysis can even yield the residues involved in the allosteric regulation, as those residues which exhibit a high variation in total charge (e.g., approximately 10 sites on the Rpn-13 regulatory subunit).

It is important to note that the structures used in the EEM calculation were incomplete. Specifically, due to the size of these molecular machines, the resolution of the structures was too low to distinguish H atoms or even parts of residues. No modifications were made to the structures of the proteasome conformers prior to the EEM calculation. Thus, the charge distribution of each conformer is not expected to be physically relevant taken on its own. Moreover, the results are very likely affected by the polarizability exaggeration artifact of EEM, discussed in the Computational details section. Therefore, the analysis here focused on how the amount of charge in functional parts of the proteasome changes with the conformation. This case study shows how a brief calculation using only a crude structural approximation can give insight regarding allosteric regulation in large biomolecular complexes.

Conclusions

We present AtomicChargeCalculator (ACC), a web-based application for the calculation and analysis of atomic charges which respond to changes in molecular conformation and chemical environment. ACC also provides interactive facilities for statistical analysis and comparison of the results. We illustrate how direct analysis of atomic charges can give basic information about chemical reactivity in paracetamol, and how residue charges hold clues about biochemical relevance in the antimicrobial peptide protegrin-1. Additionally, ACC provides molecular structure files containing atomic charges, which can be used in further modelling studies. We illustrate how such data can be used for pK_a calculation using QSPR models. Another advantage of ACC is that it can

handle any type of molecular system, regardless of size and chemical complexity, from drug-like molecules to biomacromolecular complexes with hundreds of thousands of atoms. We show how the direction and intensity of allosteric regulation can be tracked in large biomacromolecular systems like the proteasome even in the absence of high resolution structures. ACC thus caters to all fields of life sciences, from drug design to nano-carriers. AtomicChargeCalculator is freely available online at http://ncbr.muni.cz/ACC.

Availability and requirements

- *Project name* AtomicChargeCalculator
- *Project home page* http://ncbr.muni.cz/ACC
- *Operating system(s)* Web server - platform independent. Command line application—Windows, Linux, Mac OS
- *Programming language* C#
- *Other requirements* For the web-server - modern internet browser with JavaScript enabled, WebGL support for 3D visualization. For the command line application - NET 4.0 for Windows based systems, Mono framework 3.10 or newer (http://www.mono-project.com) for other OS.
- *License* ACC license for the downloadable command line version.
- *Any restrictions to use by non-academics* Free of charge. No login requirement for running or accessing the results in the web server.

Additional file

Additional file 1. Full computational details, benchmark and limitations of ACC, along with supporting information for case studies I and III.

Authors' contributions

CMI, RSV and JK conceived the study, and participated in its design and coordination. DS designed and optimized the algorithms. DS, SG and LP implemented the web-application. CMI, DS, FLF, SG and PP performed extensive testing of the web-application. CMI and FLF prepared the documentation. CMI and TB performed the calculations for the case studies. CMI, TB and LP wrote the manuscript. All authors read and approved the manuscript.

Author details

[1] CEITEC-Central European Institute of Technology, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic. [2] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic. [3] Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic.

Acknowledgements

This work was supported by the Grant Agency of the Czech Republic [13-25401S] and the European Community's Seventh Framework Programme [CZ.1.05/1.1.00/02.0068] from the European Regional Development Fund. Additional support was provided by the project "Employment of Newly Graduated Doctors of Science for Scientific Excellence" [CZ.1.07/2.3.00/30.0009] to CMI, co-financed from the European Social Fund and the state budget of the Czech Republic. The authors thank Mr. Ravi Ramos and Mr. Tomáš Raček for useful discussions.

Competing interests

The authors declare that they have no competing interests.

Received: 20 June 2015 Accepted: 8 October 2015
Published online: 22 October 2015

References

1. Giese B, Graber M, Cordes M (2008) Electron transfer in peptides and proteins. Curr Opin Chem Biol 12(6):755–759. doi:10.1016/j.cbpa.2008.08.026
2. Grodick MA, Muren NB, Barton JK (2015) DNA charge transport within the cell. Biochemistry 54(4):962–973. doi:10.1021/bi501520w
3. Li L, Wang L, Alexov E (2015) On the energy components governing molecular recognition in the framework of continuum approaches. Front Mol Biosci. doi:10.3389/fmolb.2015.00005
4. Zheng G, Xiao M, Lu XH (2005) QSAR study on the Ah receptor-binding affinities of polyhalogenated dibenzo-p-dioxins using net atomic-charge descriptors and a radial basis neural network. Anal Bioanal Chem 383:810–816. doi:10.1007/s00216-005-0085-7
5. Karelson M, Karelson G, Tämm T, Tulp I, Jänes J, Tämm K, Lomaka A, Deniss S, Dobchev D (2009) QSAR study of pharmacological permeabilities. Arkivoc 2009(2):218–238
6. Wood JS (1995) An X-ray determination of the electron distribution in crystals of hexapyridine-N-oxide cobalt(II) perchlorate and the electronic structure of the Co2+ ion. Inorganica Chimica Acta 229(1–2):407–415. doi:10.1016/0020-1693(94)04272-W
7. Belokoneva EL, Gubina YK, Forsyth JB, Brown PJ (2002) The charge-density distribution, its multipole refinement and the antiferromagnetic structure of dioptase, $Cu_6[Si_6O_{18}] \cdot 6H_2O$. Phys Chem Min 29(6):430–438. doi:10.1007/s00269-002-0246-6
8. Schrödinger E (1926) An undulatory theory of the mechanics of atoms and molecules. Phys Rev 28(6):1049
9. Mulliken RS (1935) Electronic structures of molecules XI. Electroaffinity, molecular orbitals and dipole moments. J Chem Phys 3(9):573–585. doi:10.1063/1.1749731
10. Mulliken RS (1962) Criteria for the construction of good self-consistent-field molecular orbital wave functions, and the significance of LCAO-MO population analysis. J Chem Phys 36(12):3428. doi:10.1063/1.1732476
11. Löwdin P-O (1950) On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. J Chem Phys 18(3):365–375. doi:10.1063/1.1747632
12. Reed EA, Weinstock RB, Weinhold F (1985) Natural population analysis. J Chem Phys 83(2):735–746. doi:10.1063/1.449486
13. Bader RFW, Larouche A, Gatti C, Carroll MT, MacDougall PJ, Wiberg KB (1987) Properties of atoms in molecules: dipole moments and transferability of properties. J Chem Phys 87(2):1142–1152. doi:10.1063/1.453294
14. Hirshfeld FL (1977) Bonded-atom fragments for describing molecular charge densities. Theoretica Chimica Acta 44(2):129–138. doi:10.1007/BF00549096
15. Bultinck P, Van Alsenoy C, Ayers PW, Carbó-Dorca R (2007) Critical analysis and extension of the Hirshfeld atoms in molecules. J Cheml Phys. doi:10.1063/1.2715563
16. Breneman CM, Wiberg KB (1990) Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. J Comput Chem 11(3):361–373. doi:10.1002/jcc.540110311
17. Besler BH, Merz KM, Kollman PA (1990) Atomic charges derived from semiempirical methods. J Comput Chem 11:431–439. doi:10.1002/jcc.540110404
18. Kelly CP, Cramer CJ, Truhlar DG (2005) Accurate partial atomic charges for high-energy molecules using class IV charge models with the MIDI! basis set. Theor Chem Acc 113(3):133–151. doi: 10.1007/s00214-004-0624-x
19. Manz TA, Sholl DS (2010) Chemically meaningful atomic charges that reproduce the electrostatic potential in periodic and nonperiodic materials. J Chem Theory Comput 6(8):2455–2468. doi:10.1021/ct100125x
20. Abraham RJ, Griffiths L, Loftus P (1982) Approaches to charge calculations in molecular mechanics. J Comput Chem 3(3):407–416. doi:10.1002/jcc.540030316
21. Shulga DA, Oliferenko AA, Pisarev SA, Palyulin VA, Zefirov NS (2008) Parameterization of empirical schemes of partial atomic charge

calculation for reproducing the molecular electrostatic potential. Doklady Chem 419(1):57–61. doi:10.1007/s10631-008-3004-6
22. Mortier WJ, Ghosh SK, Shankar S (1986) Electronegativity-equalization method for the calculation of atomic charges in molecules. J Am Chem Soc 108(15):4315–4320. doi:10.1021/ja00275a013
23. Rappé AK, Goddard WA III (1991) Charge equilibration for molecular dynamics simulations. J Phys Chem 95:3358–3363. doi:10.1021/j100161a070
24. Nistor RA, Polihronov JG, Müser MH (2006) A generalization of the charge equilibration method for nonmetallic materials. J Chem Phys. doi:10.1063/1.2346671
25. Gasteiger J, Marsili M (1980) Iterative partial equalization of orbital electronegativity-a rapid access to atomic charges. doi:10.1016/0040-4020(80)80168-2
26. Cho K-H, Kang YK, No KT, Scheraga HA (2001) A fast method for calculating geometry-dependent net atomic charges for polypeptides. J Phys Chem A 105(17):3624–3634. doi:10.1021/jp0023213
27. Oliferenko AA, Pisarev SA, Palyulin VA, Zefirov NS (2006) Atomic charges via electronegativity equalization: generalizations and perspectives. doi:10.1016/S0065-3276(06)51004-4
28. Baekelandt B, Mortier W, Lievens J (1991) Probing the reactivity of different sites within a molecule or solid by direct computation of molecular sensitivities via an extension of the electronegativity equalization. J Am Chem Soc 113(18):6730–6734. doi:10.1021/ja00018a003
29. York DM, Yang W (1996) A chemical potential equalization method for molecular simulations. J Chem Phys 104(1):159. doi:10.1063/1.470886
30. Yang Z-Z, Wang C-S (1997) Atom-bond electronegativity equalization method. 1. Calculation of the charge distribution in large molecules. J Phys Chem A 101(35):6315–6321. doi:10.1021/jp9711048
31. Njo SL, Fan J, Van De Graaf B (1998) Extending and simplifying the electronegativity equalization method. J Mol Catal A Chem 134:79–88. doi:10.1016/S1381-1169(98)00024-7
32. Dias LG, Shimizu K, Farah JPS, Chaimovich H (2002) A simple method for the fast calculation of charge redistribution of solutes in an implicit solvent model. Chem Phys 282(2):237–243. doi:10.1016/S0301-0104(02)00717-6
33. Chaves J, Barroso JM, Bultinck P, Carbó-Dorca R (2006) Toward an alternative hardness kernel matrix structure in the Electronegativity Equalization Method (EEM). J Chem Inform Model 46(4):1657–1665. doi:10.1021/ci050505e
34. Ouyang Y, Ye F, Liang Y (2009) A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules. Phys Chem Chem Phys PCCP 11(29):6082–6089. doi:10.1039/b821696g
35. Verstraelen T, Van Speybroeck V, Waroquier M (2009) The electronegativity equalization method and the split charge equilibration applied to organic systems: Parametrization, validation, and comparison. J Chem Phys. doi:10.1063/1.3187034
36. Frisch M, Trucks G, Schlegel H, Scuseria G, Robb M, Cheeseman J, Scalmani G, Barone V, Mennucci B, Petersson G et al (2010) Gaussian 09 (revision a. 02), gaussian, inc., wallingford ct (USA). In: Naturforsch Z (ed) Vol. 10
37. Hutter J, Iannuzzi M, Schiffmann F, Vandevondele J (2014) Cp2k: atomistic simulations of condensed matter systems. Wiley Interdiscip Rev Comput Mol Sci 4(1):15–25. doi:10.1002/wcms.1159
38. Manz TA, Sholl DS (2012) Improved atoms-in-molecule charge partitioning functional for simultaneously reproducing the electrostatic potential and chemical states in periodic and nonperiodic materials. J Chem Theory Comput 8(8):2844–2867
39. Verstraelen T, Vandenbrande S, Chan M, Zadeh FH, González C, Limacher PA, Horton AM (2013). http://theochem.github.com/horton/
40. Keith TA (2013) Aimall (version 13.05. 06). TK Gristmill Software, Overland Park
41. Marenich AV, Jerome SV, Cramer CJ, Truhlar DG (2012) Charge model 5: an extension of hirshfeld population analysis for the accurate description of molecular interactions in gaseous and condensed phases. J Chem Theory Comput 8(2):527–541. doi:10.1021/ct200866d
42. Malde AK, Zuo L, Breeze M, Stroet M, Poger D, Nair PC, Oostenbrink C, Mark AE (2011) An automated force field topology builder (ATB) and repository: Version 1.0. J Chem Theory Comput 7(12):4026–4037. doi:10.1021/ct200196m
43. Medeiros DDJ, Cortopassi WA, Costa França TC, Pimentel AS (2013) ITP adjuster 1.0: A new utility program to adjust charges in the topology files generated by the PRODRG server. J Chem. doi:10.1155/2013/803151
44. Wang J, Wang W, Kollman PA, Case DA (2006) Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Modelling 25(2):247–260. doi:10.1016/j.jmgm.2005.12.005
45. Vanquelef E, Simon S, Marquant G, Garcia E, Klimerak G, Delepine JC, Cieplak P, Dupradeau FY (2011) R.E.D. Server: a web service for deriving RESP and ESP charges and building force field libraries for new molecules and molecular fragments. Nucl Acids Res. doi:10.1093/nar/gkr288
46. Mukherjee G, Patra N, Barua P, Jayaram B (2011) A fast empirical GAFF compatible partial atomic charge assignment scheme for modeling interactions of small molecules with biomolecular targets. J Comput Chem 32(5):893–907. doi:10.1002/jcc.21671
47. Vainio MJ, Johnson MS (2007) Generating conformer ensembles using a multiobjective genetic algorithm. J Chem Inform Modeling 47(6):2462–2474. doi:10.1021/Ci6005646
48. Vařeková RS, Koča J (2006) Software news and update optimized and parallelized implementation of the electronegativity equalization method and the atom-bond electronegativity equalization method. J Comput Chem 27(3):396–405. doi:10.1002/jcc.20344
49. Dolinsky TJ, Nielsen JE, McCammon JA, Baker NA (2004) PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. Nucl Acids Res. doi:10.1093/nar/gkh381
50. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open babel: an open chemical toolbox. J Cheminform. doi:10.1186/1758-2946-3-33
51. Cho AE, Guallar V, Berne BJ, Friesner R (2005) Importance of accurate charges in molecular docking: Quantum Mechanical/Molecular Mechanical (QM/MM) approach. J Comput Chem 26(9):915–931. doi:10.1002/jcc.20222
52. Anisimov VM (2010) Quantum-mechanical molecular dynamics of charge transfer. Kinetics Dynamics. doi:10.1007/978-90-481-3034-4_9
53. Nielsen JE, Gunner MR, García-Moreno EB (2011) The pK a Cooperative: a collaborative effort to advance structure-based calculations of pK a values and electrostatic effects in proteins. doi:10.1002/prot.23194
54. Baker CM (2015) Polarizable force fields for molecular dynamics simulations of biomolecules. Wiley Interdiscip Rev Comput Mol Sci 5(2):241–254. doi:10.1002/wcms.1215
55. Heidler R, Janssens GOA, Mortier WJ, Schoonheydt RA (1996) Charge sensitivity analysis of intrinsic basicity of Faujasite-type zeolites using the electronegativity equalization method (EEM). J Phys Chem 100(50):19728–19734. doi:10.1021/jp9615619
56. Haldoupis E, Nair S, Sholl DS (2012) Finding MOFs for highly selective CO2/N2 adsorption using materials screening based on efficient assignment of atomic point charges. J Am Chem Soc 134(9):4313–4323. doi:10.1021/ja2108239
57. Bultinck P, Langenaeker W, Carbó-Dorca R, Tollenaere JP (2003) Fast calculation of quantum chemical molecular descriptors from the Electronegativity Equalization Method. J Chem Inform Comp Sci 43:422–428. doi:10.1021/ci0255883
58. Svobodová Vařeková R, Geidl S, Ionescu C-M, Skřehota O, Bouchal T, Sehnal D, Abagyan R, Koča J (2013) Predicting p Ka values from EEM atomic charges. J Cheminform 5(1):18. doi:10.1186/1758-2946-5-18
59. Shimizu K, Chaimovich H, Farah JPS, Dias LG, Bostick DL (2004) Calculation of the dipole moment for polypeptides using the generalized born-electronegativity equalization method: results in vacuum and continuum-dielectric solvent. J Phys Chem B 108(13):4171–4177. doi:10.1021/jp037315w
60. Chen S, Yang Z (2010) Molecular dynamics simulations of a β-hairpin fragment of protein G by means of atom-bond electronegativity equalization method fused into molecular mechanics (ABEEM$\delta\pi$/MM). Chin J Chem 28(11):2109–2118. doi:10.1002/cjoc.201090350
61. Ionescu CM, Svobodová Vařeková R, Prehn JHM, Huber HJ, Koča J (2012) Charge profile analysis reveals that activation of pro-apoptotic regulators bax and bak relies on charge transfer mediated allosteric regulation. PLoS Comput Biol. doi:10.1371/journal.pcbi.1002565
62. Chelli R, Procacci P, Righini R, Califano S (1999) Electrical response in chemical potential equalization schemes. J Chem Phys 111(18):8569. doi:10.1063/1.480198

63. Warren Lee G, Davis JE, Patel S (2008) Origin and control of superlinear polarizability scaling in chemical potential equalization methods. J Chem Phy 128(14):144110. doi:10.1063/1.2872603
64. Verstraelen T, Pauwels E, De Proft F, Van Speybroeck V, Geerlings P, Waroquier M (2012) Assessment of atomic charge models for gas-phase computations on polypeptides. J Chem Theory Comput 8(2):661–676. doi:10.1021/ct200512e
65. van Duin ACT, Strachan A, Stewman S, Zhang Q, Xu X, Goddard WA (2003) ReaxFF SiO reactive force field for silicon and silicon oxide systems. J Phys Chem A 107(19):3803–3811. doi:10.1021/jp0276303
66. Puranen JS, Vainio MJ, Johnson MS (2009) Accurate conformation-dependent molecular electrostatic potentials for high-throughput in silico drug discovery. J Comput Chem. doi:10.1002/jcc.21460
67. Bultinck P, Langenaeker W, Lahorte P, De Proft F, Geerlings P, Van Alsenoy C, Tollenaere JP (2002) The electronegativity equalization method II: applicability of different atomic charge schemes. J Phys Chem A 106(34):7895–7901. doi:10.1021/jp020547v
68. Bultinck P, Vanholme R, Popelier PLA, De Proft F, Geerlings P (2004) High-speed calculation of AIM charges through the electronegativity equalization method. J Phys Chem A 108(46):10359–10366. doi:10.1021/jp0469281
69. Varekova RS, Jirouskova Z, Vanek J, Suchomel S, Koca J (2007) Electronegativity equalization method: parameterization and validation for large sets of organic, organohalogene and organometal molecule. Int J Mol Sci 8(7):572–582
70. Ionescu CM, Geidl S, Svobodová Vařeková R, Koča J (2013) Rapid calculation of accurate atomic charges for proteins via the electronegativity equalization method. J Chem Inform Model 53(10):2548–2558. doi:10.1021/ci400448n
71. Graham GG, Davies MJ, Day RO, Mohamudally A, Scott KF (2013) The modern pharmacology of paracetamol: therapeutic actions, mechanism of action, metabolism, toxicity and recent pharmacological findings. Inflammopharmacology 21(3):201–232. doi:10.1007/s10787-013-0172-x
72. Bertolini A, Ferrari A, Ottani A, Guerzoni S, Tacchi R, Leone S (2006) Paracetamol: new vistas of an old drug. CNS Drug Rev. 12(3–4):250–275. doi:10.1111/j.1527-3458.2006.00250.x
73. Svobodová Vařeková R, Geidl S, Ionescu CM, Skřehota O, Kudera M, Sehnal D, Bouchal T, Abagyan R, Huber HJ, Koča J (2011) Predicting pKa values of substituted phenols from atomic charges: comparison of different quantum mechanical methods and charge distribution schemes. J Chem Inform Modeling 51(8):1795–1806. doi:10.1021/ci200133w
74. Ugur I, Marion A, Parant S, Jensen JH, Monard G (2014) Rationalization of the pKa values of alcohols and thiols using atomic charge descriptors and its application to the prediction of aminoacid pKa's. J Chem Inform Modeling. doi:10.1021/ci500079w
75. Dastmalchi S, Rashidi M, Rassi M (1995) Simultaneous determination of the pka and octanol/water partition coefficient (pm) of acetaminophen. J Sch Pharm Med Sci Univ Tehran 4:7–14
76. NCI Open Database Compounds. National Cancer Institute. http://cactus.nci.nih.gov/. Accessed Aug 2015
77. Howard P, Meylan W (1999) Physical/chemical property database (PHYSPROP). Syracuse Research Corporation, Environmental Science Center, North Syracuse
78. Geidl S, Svobodová Vařeková R, Bendová V, Petrusek L, Ionescu CM, Jurka Z, Abagyan R, Koča J (2015) How does the methodology of 3D structure preparation influence the quality of p K a prediction? J Chem Inform Modeling 55(6):1088–1097. doi:10.1021/ci500758w
79. Bellm L, Lehrer RI, Ganz T (2000) Protegrins: new antibiotics of mammalian origin. Exp Opin Investig Drugs 9(8):1731–1742. doi:10.1517/13543784.9.8.1731
80. Steinberg DA, Hurst MA, Fujii CA, Kung AHC, Ho JF, Cheng FC, Loury DJ, Fiddes JC (1997) Protegrin-1: a broad-spectrum, rapidly microbicidal peptide with in vivo activity. Antimicrob Agents Chemotherap 41(8):1738–1742
81. Dong N, Zhu X, Chou S, Shan A, Li W, Jiang J (2014) Antimicrobial potency and selectivity of simplified symmetric-end peptides. Biomaterials 35(27):8028–8039. doi:10.1016/j.biomaterials.2014.06.005
82. Mohanram H, Bhattacharjya S (2014) Cysteine deleted protegrin-1 (CDP-1): anti-bacterial activity, outer-membrane disruption and selectivity. Biochimica et Biophysica Acta (BBA) General Subjects 1840(10):3006–3016. doi:10.1016/j.bbagen.2014.06.018
83. Ostberg N, Kaznessis Y (2005) Protegrin structure-activity relationships: using homology models of synthetic sequences to determine structural characteristics important for activity. Peptides 26(2):197–206. doi:10.1016/j.peptides.2004.09.020
84. Fahrner RL, Dieckmann T, Harwig SSL, Lehrer RI, Eisenberg D, Feigon J (1996) Solution structure of protegrin-1, a broad-spectrum antimicrobial peptide from porcine leukocytes. Chem Biol 3(7):543–550. doi:10.1016/S1074-5521(96)90145-3
85. Bolintineanu DS, Langham AA, Davis HT, Kaznessis YN (2007) Molecular dynamics simulations of three protegrin-type antimicrobial peptides: interplay between charges at the termini, β-sheet structure and amphiphilic interactions. doi:10.1080/08927020701393481
86. Langham AA, Khandelia H, Schuster B, Waring AJ, Lehrer RI, Kaznessis YN (2008) Correlation between simulated physicochemical properties and hemolycity of protegrin-like antimicrobial peptides: Predicting experimental toxicity. Peptides 29(7):1085–1093. doi:10.1016/j.peptides.2008.03.018
87. Lai JR, Huck BR, Weisblum B, Gellman SH (2002) Design of non-cysteine-containing antimicrobial β-hairpins: Structure-activity relationship studies with linear protegrin-1 analogues. Biochemistry 41(42):12835–12842. doi:10.1021/bi026127d
88. Bedford L, Paine S, Sheppard PW, Mayer RJ, Roelofs J (2010) Assembly, structure, and function of the 26S proteasome. doi:10.1016/j.tcb.2010.03.007
89. Gallastegui N, Groll M (2010) The 26S proteasome: assembly and function of a destructive machine. doi:10.1016/j.tibs.2010.05.005
90. Unverdorben P, Beck F, Śledź P, Schweitzer A, Pfeifer G, Plitzko JM, Baumeister W, Förster F (2014) Deep classification of a large cryo-EM dataset defines the conformational landscape of the 26S proteasome. Proc Natl Acad Sci USA 111(15):5544–5549. doi:10.1073/pnas.1403409111
91. O'Hara A, Howarth A, Varro A, Dimaline R (2013) The role of proteasome beta subunits in gastrin-mediated transcription of plasminogen activator inhibitor-2 and regenerating protein1. PLoS One. doi:10.1371/journal.pone.0059913
92. Cron KR, Zhu K, Kushwaha DS, Hsieh G, Merzon D, Rameseder J, Chen CC, D'Andrea AD, Kozono D (2013) Proteasome inhibitors block DNA repair and radiosensitize non-small cell lung cancer. PLoS One. doi:10.1371/journal.pone.0073710

Anatomy of enzyme channels

Pravda et al. BMC Bioinformatics 2014, 15:379
http://www.biomedcentral.com/1471-2105/15/379

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Anatomy of enzyme channels

Lukáš Pravda[1†], Karel Berka[2†], Radka Svobodová Vařeková[1], David Sehnal[1,3], Pavel Banáš[2], Roman A Laskowski[4], Jaroslav Koča[1*] and Michal Otyepka[2*]

Abstract

Background: Enzyme active sites can be connected to the exterior environment by one or more channels passing through the protein. Despite our current knowledge of enzyme structure and function, surprisingly little is known about how often channels are present or about any structural features such channels may have in common.

Results: Here, we analyze the long channels (i.e. >15 Å) leading to the active sites of 4,306 enzyme structures. We find that over 64% of enzymes contain two or more long channels, their typical length being 28 Å. We show that amino acid compositions of the channel significantly differ both to the composition of the active site, surface and interior of the protein.

Conclusions: The majority of enzymes have buried active sites accessible via a network of access channels. This indicates that enzymes tend to have buried active sites, with channels controlling access to, and egress from, them, and that suggests channels may play a key role in helping determine enzyme substrate.

Background

Channels inside biomacromolecular structures (proteins, nucleic acids and their complexes) play many significant biological roles as they enable traffic between the interior spaces and the exterior. In enzymes they allow passage of substrates and products to/from the active site [1-12], in the ribosome they allow nascently synthetized proteins to pass from the proteosynthetic center to the outside [13], and in membrane proteins they provide high specificity of passage in either direction through the membrane [14,15]. Thus channels have attracted the attention of many researchers, who have rationalized their biological roles using a variety of experimental and theoretical methods. The ribosome, for example, prevents nascently synthetized polypeptides getting stuck in its polypeptide egress channel by lining the wall of the channel with a mosaic of alternating negative and positive electrostatic potentials [13,16]. Gramicidin provides polar holes for biomembranes, enabling free diffusion of monovalent ions and water through the membrane [17-19], while transmembrane ion channels maintain their high selectivity by a combination of structural and electrostatic features of the channel-lining amino acids [14,20].

Enzymes are proteins that catalyse reactions changing substrates to products. The enzymatic reactions occur in the enzymes' active sites. Thanks to the many analyses of enzymatic reactions, we now have a better understanding of how active site chemistry works [21-24] and which amino acids are present in the sites [25]. However, relatively little is known about how substrates enter active sites and how the respective products leave them. While some active sites are positioned on the protein's surface, in clefts or pockets, other enzymes have deeply buried active sites, which are connected to the outside by one or more channels. Here we focus on these channels, as the active site access paths play an important role in substrate and product trafficking between active site and outside. It has been shown that mutations in enzymes' active site access channels alter the substrate preferences of haloalkane dehalogenase enzymes and may be utilized in rational design of enzymes [26,27]. The amino acids lining the access channels of cytochrome P450 (CYP) are important for the selectivity of these enzymes [28] while the flexibility of these channels, i.e. their

* Correspondence: jaroslav.koca@ceitec.muni.cz; michal.otyepka@upol.cz
†Equal contributors
[1]National Centre for Biomolecular Research, Faculty of Science and CEITEC, Central European Institute of Technology, Masaryk University Brno, Kamenice 5, Brno-Bohunice 625 00, Czech Republic
[2]Department of Physical Chemistry, Regional Centre of Advanced Technologies and Materials, Faculty of Science, Palacký University Olomouc, tř. 17. listopadu 12, Olomouc 771 46, Czech Republic
Full list of author information is available at the end of the article

opening and closing motions, contributes to the broad substrate specificity of CYP [10,29].

Despite a large effort, and recent progress in the field, an in-depth analysis of enzyme channels is lacking. Here, we use an advanced software tool, MOLE 2.0, developed for analysis of biomacromolecular channels [30], to survey 4,306 enzymes annotated in the Catalytic Site Atlas (CSA). We identify that more than 64% of enzyme structures contain channels at least 15 Å long from the active site. A typical enzyme channel is ~20 Å long and its walls are made preferentially of histidine, arginine, tryptophan and tyrosine residues and, to a lesser extent, by phenylalanine, asparagine, and aspartic acid (Figure 1). These residues can be considered as gate-keepers controlling the entry to and from the active site.

Results

Geometry features

We identified that at least 64.2% of enzymes contain channels at least 15 Å long (Table 1) and 86.8% contain channels at least 5 Å long. However, as the short channels may correspond to paths connecting cleft-like active sites with the exterior, we decided to use only channels longer than 15 Å for our analysis. Enzymes with channels (of ≥15 Å in length) leading to a buried active site contained on average two such channels (Table 2). Some had more than five such channels, the highest number being 68 in 6,7-dimethyl-8-ribityllumazine synthase from *Aquifex aeolicus* (PDBID: 1NQU; Additional file 1: Figure S1). Whereas one might expect larger proteins to contain larger numbers of channels, we found that the number of long channels does not correlate with protein size (Additional file 1: Figure S2). The median channel length is 27.7 Å (Table 2 and Additional file 1: Figure S3), 40% of channels are 15–30 Å long and 10% of enzymes contain channels longer than 50 Å. The longest channel (172 Å long) was found in Penicillium vitale catalase from *Penicillium janthinellum* (PDBID: 2IUF; Additional file 1: Figure S1). It should be noted that although the size of small enzymes (i.e. those containing fewer than 5,000 atoms) does not limit the number of long channels they contain, it does limit the maximum length these channels can have (Additional file 1: Figure S2).

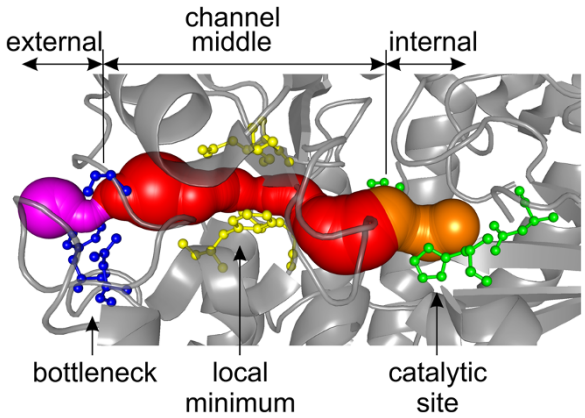


Figure 1 An example of an enzyme channel identified by MOLE 2.0. Internal, middle and external parts of the channel in pyridoxal-5'-phosphate-dependent acyl-CoA transferase (PDBID: 3KKI) are colored orange, red and magenta, respectively. Active site amino acids (present in the internal part of the channel) are shown in green, amino acids in the middle part making the wall of a local minimum (channel narrowing) are in yellow, and amino acids in the external part lining the bottleneck are in blue.

Table 1 Number of enzyme entries in each EC class, and numbers of channels of different lengths

EC	Enzyme class	Number of enzymes	Enzymes with channels of length			
			≥ 5 Å	≥10 Å	≥ 15 Å	≥ 20 Å
EC1	Oxidoreductases	879	781	736	**684**	612
EC2	Transferases	1096	963	863	**747**	639
EC3	Hydrolases	1455	1228	1019	**749**	542
EC4	Lyases	465	401	361	**318**	262
EC5	Isomerases	252	226	204	**165**	140
EC6	Ligases	159	139	118	**98**	88
	Sum	4306	3738	3301	**2761**	2283

Numbers of enzymes in the dataset containing at least one channel of the given length. Bold values indicate 15 Å threshold for channel detection used thorough the study.

Channel occurrence and length varies among the enzymatic classes (Table 2). The highest percentage (77.8%) of proteins with channels longer than 15 Å was identified in oxidoreductases (EC1), while the lowest percentage (51.8%) applied to hydrolases (EC3). The number of channels is slightly elevated in oxidoreductases (EC1), transferases (EC2) and isomerases (EC5). Oxidoreductases (EC1) have median channel length longer by about 2 Å than other enzymes, whereas transferases (EC2) and ligases (EC6) have average channel length shorter, also by about 2 Å (Figure 2). As a result, oxidoreductases stand out of

Table 2 Geometrical channel features for all enzyme classes

EC	Na	P	M[a]	L
EC1	6518	77.8	3	29.8
EC2	4728	68.2	3	26.3
EC3	3454	51.8	2	27.5
EC4	5780	68.4	2	27.9
EC5	5107	65.5	3	26.4
EC6	5289	61.6	2	25.2
All	4823	64.2	2	27.7

[a]Enzymes not containing channels were excluded.
Average number of atoms (Na), percentage of enzymes containing at least one channel from the active site longer than 15 Å (P in %), median number of channels (M), and median length of channels (L in Å) for each enzyme class.

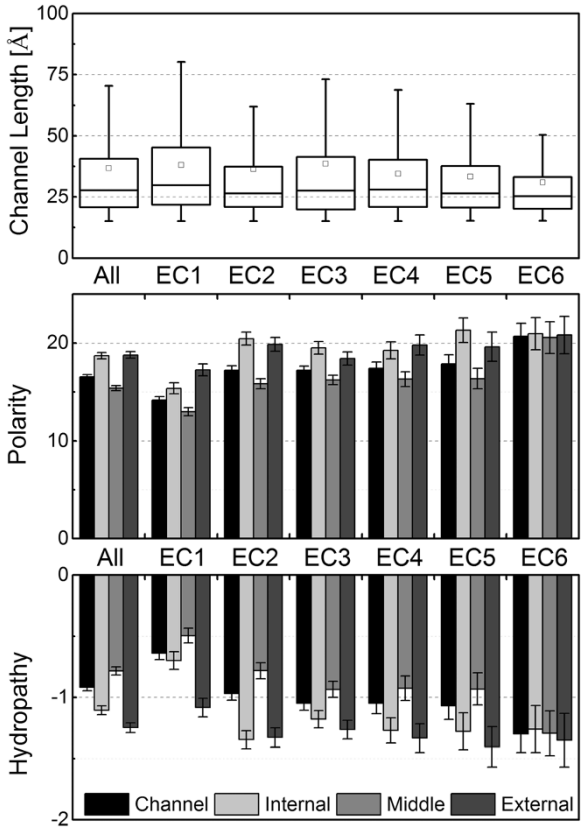


Figure 2 Properties of channels in different enzymatic classes. The topmost panel shows the variability of average channel length in individual enzymatic classes in comparison with the overall average channel length. The middle panel shows the average channel polarity for each enzyme class as well as the polarity of individual parts of the channels. The colors of the bars correspond to different parts of the channels, as shown in the key at the bottom. The bottom panel shows the variability in average hydropathy. Error bars show the standard error of the mean.

the crowd, as they have both higher channel occurrence and longer channels (Additional file 1: Figure S2).

Physico-chemical features

For each channel, we calculated basic physico-chemical features such as hydropathy [31] and polarity [32] using the method introduced in our previous paper [30]. In brief, the method sums up the length-weighted physico-chemical properties of the amino acids lining the channel. As the algorithm is rather approximate, we would like to note that the estimated physico-chemical properties should be interpreted with care. The average channel hydropathy is −0.92 (Table 3) and this value is close to the hydropathy of tyrosine and tryptophan (it is worth noting that the hydropathy of amino acids varies from −4.5 of Arg to 4.5 of Ile). The distribution plots of hydropathy (Additional file 1: Figure S3) also indicate that the values are shifted in most channels to negative values indicating

Table 3 Physicochemical features of channels

EC	Hydropathy	Polarity	N(+)	N(−)	Δ
EC1	−0.64 ± 0.05	14.2 ± 0.4	3	2	1
EC2	−0.97 ± 0.06	17.2 ± 0.4	3	2	1
EC3	−1.05 ± 0.05	17.2 ± 0.4	3	2	0
EC4	−1.05 ± 0.08	17.4 ± 0.6	3	2	1
EC5	−1.07 ± 0.11	17.9 ± 0.9	3	2	1
EC6	−1.30 ± 0.15	20.7 ± 1.3	4	2	1
All	**−0.92 ± 0.03**	**16.5 ± 0.2**	**3**	**2**	**0**
Internal	−1.11 ± 0.04	18.7 ± 0.3	1	1	0
Middle	−0.78 ± 0.03	15.4 ± 0.2	1	0	0
External	−1.25 ± 0.04	18.8 ± 0.3	1	1	0

The average hydropathy and polarity are shown for each enzyme class as well as for the internal, middle and external parts of channels. Also shown are the median number of charged amino acid side chains lining the channels (N(+): positive Arg, Lys and His, N(−): negative Asp and Glu, and Δ: median of overall charge). Bold values indicate features of all channels.

that hydrophilic channels are more preferred over hydrophobic ones.

The average channel polarity is 16.5, which falls between the values of highly polar amino acids (Asp, Glu, Lys, Arg and His having polarities of 49.5 – 52.0) and those of other amino acids (with polarities of 0.0 – 3.5). It indicates that the channels are rather polar as well as hydrophilic. Taking all this information into account we may conclude that the average channel has slightly negative hydropathy and higher polarity. However, highly hydrophobic and nonpolar, as well as highly hydrophilic and polar, channels were also detected. We also analyzed the presence of charged amino acid side chains (Asp, Glu, His, Lys and Arg) in channels walls (Additional file 1: Table S2). On average the channel walls are lined by two negative and two positive side chains resulting in sum neutral channel walls (Table 3).

We also identified channels with significant extreme physico-chemical properties (Additional file 1: Table S3 and Additional file 1: Figure S1). Here we present two examples. A highly hydrophilic channel (hydropathy index −3.8) of length 18.7 Å occurs in 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase (PDB ID: 1N8F) from *E. coli*. The high hydrophilicity of the channel is in accord with its function [33] since this enzyme catalyses a condensation reaction between two highly polar substrates: phosphoenolpyruvate and erythrose-4-phosphate. It is worth noting that transferases (EC2) show the largest variability in hydrophobicity as illustrated by the fact that six times transferases are in the top 10 having highly hydrophilic channels, and four occur in the top 10 with highly hydrophobic channels.

At the other extreme is peroxidase (PDBID: 1LYK) from the fungus *Coprinus cinereus* [34] which has a highly hydrophobic channel (hydropathy index of 3.59)

of length 23.8 Å. This channel enables transport of simple phenols and smaller aromatic dye molecules for their oxidation in lignin decomposition [35] in a process which has been exploited in biotechnology as a dye-transfer inhibitor in a laundry detergent [36].

These examples show that enzymatic classes differ in their average physico-chemical properties: (i) oxidoreductases (EC1) show the most hydrophobic as well as the least polar channels among the enzyme classes, while (ii) ligases (EC6), and to some extent also isomerases (EC5), lyases (EC4) and hydrolases (EC3), show the most hydrophilic as well as the most polar channels (Figure 2, Table 3 and Additional file 1: Table S3).

We also identified that some physicochemical features differ across the three channel layers: internal, middle and external (Table 3). The polarity of middle part of the channel is always lower than polarity of both internal and external parts, respectively. The lower polarity of the middle part of channel is also reflected by its significantly more hydrophobic behaviour. The charged residues occur mainly in external parts of enzyme channels, while the internal and middle part contains more aromatic residues.

Channel-lining residues

We calculated the frequencies of channel-lining amino acids and compared them with frequencies of amino acids in the same enzyme structures. On the basis of this data, the channel propensities of individual amino acids can be defined as a ratio of the frequency of amino acid in the channel walls to the frequency of amino acid anywhere within the protein structure. The resulting channel propensities of the individual amino acids differ significantly. The rather bulky and aromatic amino acids (His, Tyr, Trp, Arg), occur over 1.25 times more frequently in the channel walls than in the whole enzyme. Additionally, other amino acids (Asn, Phe, Asp, Thr, Met, Ser) also show a slightly higher frequency in the channel walls than in the rest of the protein. Conversely, nonpolar aliphatic amino acids (Pro, Gly, Ile, Leu, Ala, and Val) are significantly less localized in channel walls (Figure 3). We also looked at the amino acid composition at each channel's local minimum. Whereas this reflected the composition of the whole channel, the channel bottlenecks contain significantly more cysteine (Cys), histidine (His) and tyrosine (Tyr) residues than usual and much fewer small aliphatic amino acids (Pro, Gly and Ala) (Additional file 1: Figure S4). As histidine (His) and cysteine (Cys) have unique binding properties, it is possible to hypothesize that these binding properties might provide a gate-keeping activity at the channel bottlenecks, whereas small aliphatic residues cannot undergo large changes and as such cannot serve as gate-keepers.

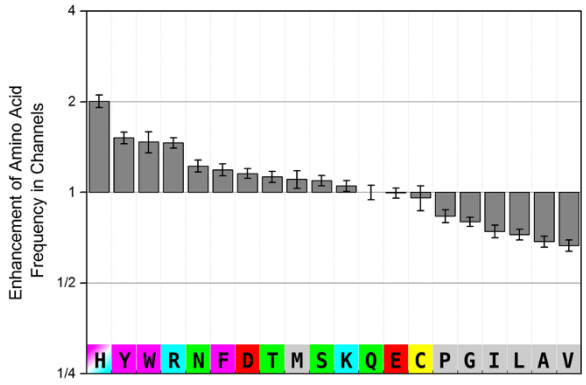


Figure 3 Channel propensity of amino acids. Enhancement of frequencies of individual amino acids in channels indicates which amino acids are more likely to occur in the channel walls than anywhere else in the protein structure. Hydrophobic aliphatic residues are shown in gray, aromatic amino acids in magenta, polar residues in green, negatively charged in red, positively charged in blue and cysteine in yellow.

The frequencies of amino acids in active sites, on the protein surface and inside the protein, or in general channels, are markedly different from both the average protein amino acid composition and the composition of the channels (Figure 4). The active sites contain significantly more amino acids that can be part of a catalytic cycle (His, Asp, Cys, Glu, Arg, Tyr, Lys) enabling proton and electron shuffling and covalent bond reorganization. On the other hand, the frequency of less reactive amino acids (Trp, Thr, Gln, Phe) or amino acids with nonreactive side-chain (Met, Ala, Pro, Ile, Val, Leu) is lower in

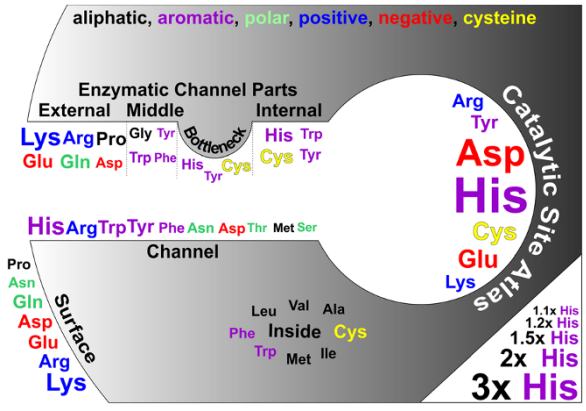


Figure 4 Enhancement of amino acid frequency in different parts of the enzyme structure. Amino acids that are found more often than average in different regions of an enzyme structure. Their labels are scaled to reflect their propensity for each compartment; the key in the bottom right-hand corner indicates how the label size relates to propensity. Hydrophobic aliphatic residues are shown in gray, aromatic amino acids in magenta, polar residues in green, negatively charged in red, positively charged in blue and cysteine in yellow.

the active sites. These results are in perfect agreement with data published by Holliday and coworkers [37]. Protein cores are rich in hydrophobic aliphatic (Ile, Leu, Val, Ala, Met) and aromatic amino acids (Phe, Trp) as well as in cysteines (Cys). These amino acids have a structural function to maintain the formation and stability of protein hydrophobic cores [38] and the formation of disulphide bonds [39]. Conversely, the surface regions contain mainly charged (Lys, Arg, Glu, Asp) and polar residues (Asn, Gln), which facilitate contact with the polar water environment. Also, the surface has a higher than average frequency of prolines (Pro) as these helix-breaker amino acids are common in turns in the protein structures rigidifying the protein fold (Additional file 1: Figure S4).

Channel-lining amino acids are not uniformly distributed along the length of the channel. Internal parts of the channel tend to contain more aromatic residues (His, Trp, Tyr) together with cysteine (Cys), while glutamine and glutamate (Gln and Glu) and aliphatic amino acids (Pro, Ala, Leu) are underrepresented. These trends are similar to the catalytic site propensities, as discussed in the previous paragraph. The frequencies of amino acids in the middle regions of the channels correspond to the frequencies in the entire channel with the exception of glycine (Gly) and aromatic amino acids (Trp, Tyr, Phe), which are present more frequently. We may hypothesize that the higher frequency of glycine (Gly) in the middle channel parts is because it facilitates flexibility, which may be important for substrate/product channeling between the active site and protein surface [29], whereas aromatic amino acids can serve as gate-keepers. External parts of the channel bear more charged residues than any other part (Arg, Lys, Glu, Asp) together with proline (Pro) and glutamine (Gln), whereas other polar residues (Ser, Thr) are surprisingly less common in the external parts even though that both Ser and Thr are evenly distributed within the structure of proteins (Additional file 1: Figure S4). External parts of channels are less occupied by aliphatic (Ile, Ala, Val) and aromatic amino acids (Phe, Tyr) as well as typical catalytic amino acids – histidine (His), aspartate (Asp) and cysteine (Cys). A higher frequency of charged or rather bulky and aromatic amino acids in channels may have functional implications because such amino acids may work as gate-keepers, regulating traffic between active site and surface via conformational changes. It is worth noting that some mutations of such residues have been shown to alter the catalytic efficiency and substrate preferences in haloalkane dehalogenases [26].

All the results above concern channels leading to buried enzyme active sites. For comparison, we also analyzed all channels of length greater than 15 Å connecting a cavity in the protein interior to the outside exterior (Additional file 1: Figure S4). We compared the amino acid compositions of the two types of channels. The active site access channels have higher frequencies of aromatic (Tyr, Trp and Phe), polar amino acids (Asn, Thr and Ser), catalytically active amino acids (Cys and His), and glycine (Gly). On the other hand, they contain fewer aliphatic amino acids (Pro, Leu, Ile and Val), charged and some polar amino acids (Lys, Glu, Arg and Gln). In sum, the active site access channels contain more functional amino acids than generic channels. This finding agrees with the idea that some amino acids bear functional roles in channels, e.g., as gate keepers or to maintain their flexibility.

The analysis of amino acids leading to active sites, divided according to the six enzymatic groups, shows that amino acid channel propensities correspond to overall channel propensities. However, some differences were identified (Additional file 1: Figure S4). As can be expected from their higher hydropathy and lower polarity, channels in oxidoreductases (EC1) have significantly lower frequencies of charged lining amino acids (Arg, Asp, Lys, Glu), but higher frequencies of aliphatic lining amino acids (Met, Pro, Ile, Leu, Ala, Val). Channels in transferases (EC2) contain fewer aromatic (Trp) and more charged (Arg, Asp, Lys) amino acids. Channels in hydrolases (EC3) contain fewer arginine (Arg), threonine (Thr) and aliphatic amino acids (Pro, Ile, Ala, Val), whereas they contain more aromatic (Trp, Tyr) and smaller charged (Lys, Asp, Glu) amino acids. Channels in lyases (EC4) show only lower amounts of aromatic (Trp) and sulphur containing (Met, Cys) amino acids. Channels in isomerases (EC5) contain fewer glycines (Gly). Channels in ligases (EC6) are the most hydrophilic channels, so it is not surprising that their channels contain fewer cysteine (Cys), aromatic (Trp, Tyr) and aliphatic (Pro, Leu) amino acids and more charged (Arg, Lys, Glu) amino acids and glycine (Gly). It should be noted that the differences between the individual enzyme classes should be interpreted with care because of larger statistical error bars, especially in the case of less populated EC5 and EC6 classes.

Discussion

Long channels (>15 Å) are a common feature of enzymes, with over 64% containing at least two such channels. This shows that the majority of enzymes have buried active sites accessible via a network of access channels. Hence there is an apparent tendency for enzymes to bury their active site, i.e., to limit and control direct connection of active sites with the surrounding environment. This may be the result of two evolutionary pressures; i) steric, because active sites have to be structurally well arranged – a buried active site enables full spatial arrangement better than a pocket-like active site can give that half of the space of the latter is open to

surrounding environment and ii) functional, as active site access paths may enable pre-selection of substrates, and may be involved in features co-determining enzyme substrate preferences. In another words, the active site access channels may limit access to the enzyme active sites and function as keyholes, enabling passage only of some classes of substrates.

The amino acid frequencies in the whole protein structures and channel walls differ significantly. Aliphatic amino acids are more involved in the formation of enzyme hydrophobic cores, which are important to maintain a protein fold. In turn, they are less frequently involved in channel wall lining or within the active sites. The aromatic, charged and polar amino acids occur more frequently in the channels walls. In addition, we identified a higher frequency of glycine in the middle parts of channels, which may function here to support channel flexibility enabling the channelling of bulkier substrates to active sites. This finding can be explained by the fact that the polar and charged amino acids line the channels to enable passage of polar substrates/products and water. The enhanced frequency of rather bulky and aromatic amino acids in channel external parts may have functional implications, because such amino acids may work as gatekeepers, regulating traffic between active site and outside.

The functional implications deduced from these global analyses are also supported by the fact that individual enzyme classes differ in their channel features. Typically oxidoreductases have the most hydrophobic, the least polar and longest channels among the enzyme classes, while ligases have the most hydrophilic, the most polar and the shortest channels. This indicates that evolution of enzymatic substrate preferences might also include evolution of active site access channels.

Conclusions

To conclude, we analyzed channels in 4,306 enzyme structures annotated in the Catalytic Site Atlas. We identified that at least 64% of enzyme structures contain on average two channels longer than 15 Å leading to the catalytic site. Consequently, we may anticipate that the same number of enzymes have buried active sites. The longest, and also the most hydrophobic, channels are found in oxidoreductases, while the smallest number of channels can be found in hydrolases and the shortest and also the most hydrophilic channels in ligases. The composition of channel walls differs from the average composition of enzyme structures as well as from the composition of the protein surface. Hydrophobic aliphatic amino acids, which are the most common amino acids present in protein cores, occur in channel walls less frequently, whereas aromatic, charged and polar amino acids occur more frequently in channel walls. All these findings indicate that the active site access channels bear significant biological function as they are involved in co-determining enzyme substrate preferences.

Methods

Dataset

We analyzed 4,306 enzymes which were annotated in the Catalytic Site Atlas (CSA) database release of 4th March 2013 [25,40]. The dataset contained structures determined by X-ray diffraction at a resolution better than 2.5 Å, and had no two structures with a sequence identity higher than 90% (more quality checks can be found in Additional file 1: Table S1). It should be noted that when we used a dataset containing structures with a sequence identity less than 50%, the results did not significantly differ from the results obtained with the dataset containing structures with a sequence identity less than 90%. The enzymes in the dataset were grouped according to their Enzymatic Commission (EC) class (Table 1).

Channel Identification

An active site is a cavity, which walls contain amino acids residues annotated in the CSA. A channel is a pathway inside an active site cavity connecting the deepest apex of the cavity with an exterior. The MOLE 2.0 program [30] was used for channel identification and characterization. Briefly, the MOLE 2.0 algorithm calculates the Delaunay triangulation/Voronoi diagram of the atomic centers, splitting it into several smaller parts and identifying suitable start and end points in the interior and surface, respectively. Dijkstra's algorithm is used to identify tunnels as the shortest paths between the start and end points (see Additional file 1 for further details). This algorithm is used also in the MOLEonline 2.0 web application [41]. The setup of MOLE 2.0 for these calculations was as follows: Probe Radius and Origin Radius 5 Å, Interior Threshold 1.1 Å and default values for Bottleneck Length, Bottleneck Radius, Cutoff Ratio and Surface Cover Radius. The CSA active sites were used as starting points. We used biological assemblies for the enzymes structures, which were obtained from the PDB database as *.pdb1 files. [42] Ten structures of EC 3.6.4 group were removed from the dataset. Hydrogen atoms and ligands not covalently bound to the structure were deleted prior to calculation. In cases where the system contained more than one active site, the site having the most channels was used. In order to study only relevant channels, we analysed only those channels longer than 15 Å. Table 1 shows the numbers of channels of different lengths for each of the six different enzyme classes, which are listed together with their properties in Additional file 2.

In the text we use the following terminology (Figure 1); Lining amino acids are all the amino acids fully encapsulating the detected channel and are divided into three

classes: internal, middle, and external according to their respective positions in the layers perpendicular the channel centerline. Internal or external lining amino acids are those lying within a 5 Å distance of the start or end point, respectively, with middle amino acids constituting the remainder. A bottleneck is where the channel radius is a minimum.

Additional files

Additional file 1: Description of MOLE methodology, and explanation of statistical methods used for channel analysis, Supplementary Figures S1–S4 and Supplementary Tables S1–S3.

Additional file 2: Full set of channels in csv format. The MOLE 2.0 application and enzymatic data set are available for download at http://mole.chemi.muni.cz/enzymes.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
LP carried out channel calculations. KB calculated the statistical analysis of the data and drafted the manuscript. RSV contributed to the design of the study and data interpretation. DS wrote the software used. PB participated in statistical analysis of the data. RAL participated in the design of the study and data interpretation. JK coordinated the study and wrote the manuscript. MO conceived of the study, designed and coordinated the study and wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
This work was supported by the Czech Science Foundation [P208/12/G016 to M.O., P303/12/P019 to K.B.] and the Operational Program Research and Development for Innovations–European Regional Development Fund [CZ.1.05/2.1.00/03.0058 to M.O., K.B., P.B. and CZ.1.05/1.1.00/02.0068 to J.K. and R.S.V.], European Social Fund [CZ.1.07/2.3.00/20.0058 to K.B.]. D.S. acknowledges Brno City Municipality for financial support provided through the program Brno Ph.D. Talent. Access to the MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is gratefully acknowledged.

Author details
[1]National Centre for Biomolecular Research, Faculty of Science and CEITEC, Central European Institute of Technology, Masaryk University Brno, Kamenice 5, Brno-Bohunice 625 00, Czech Republic. [2]Department of Physical Chemistry, Regional Centre of Advanced Technologies and Materials, Faculty of Science, Palacký University Olomouc, tř. 17. listopadu 12, Olomouc 771 46, Czech Republic. [3]Faculty of Informatics, Masaryk University Brno, Botanická 68a, Brno 602 00, Czech Republic. [4]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

Received: 19 July 2014 Accepted: 5 November 2014
Published online: 18 November 2014

References
1. Huang X, Holden HM, Raushel FM: **Channeling of substrates and intermediates in enzyme-catalyzed reactions.** *Annu Rev Biochem* 2001, **70:**149–180.
2. Park J, Czapla L, Amaro RE: **Molecular simulations of aromatase reveal new insights into the mechanism of ligand binding.** *J Chem Inf Model* 2013, **53:**2047–2056.
3. Sgrignani J, Magistrato A: **Influence of the membrane lipophilic environment on the structure and on the substrate access/egress routes of the human aromatase enzyme. A computational study.** *J Chem Inf Model* 2012, **52:**1595–1606.
4. Madrona Y, Hollingsworth SA, Khan B, Poulos TL: **P450cin active site water: implications for substrate binding and solvent accessibility.** *Biochemistry* 2013, **52:**5039–5050.
5. Cui Y-L, Zhang J-L, Zheng Q-C, Niu R-J, Xu Y, Zhang H-X, Sun C-C: **Structural and dynamic basis of human cytochrome P450 7B1: a survey of substrate selectivity and major active site access channels.** *Chemistry* 2013, **19:**549–557.
6. Lee SJ, McCormick MS, Lippard SJ, Cho U-S: **Control of substrate access to the active site in methane monooxygenase.** *Nature* 2013, **494:**380–384.
7. Pryor EE, Horanyi PS, Clark KM, Fedoriw N, Connelly SM, Koszelak-Rosenblum M, Zhu G, Malkowski MG, Wiener MC, Dumont ME: **Structure of the integral membrane protein CAAX protease Ste24p.** *Science* 2013, **339:**1600–1604.
8. Xu S, Mueser TC, Marnett LJ, Funk MO: **Crystal structure of 12-lipoxygenase catalytic-domain-inhibitor complex identifies a substrate-binding channel for catalysis.** *Structure* 2012, **20:**1490–1497.
9. Guskov A, Nordin N, Reynaud A, Engman H, Lundbäck A-K, Jong AJO, Cornvik T, Phua T, Eshaghi S: **Structural insights into the mechanisms of Mg2+ uptake, transport, and gating by CorA.** *Proc Natl Acad Sci U S A* 2012, **109:**18459–18464.
10. Otyepka M, Berka K, Anzenbacher P: **Is there a relationship between the substrate preferences and structural flexibility of cytochromes P450?** *Curr Drug Metab* 2012, **13:**130–142.
11. Rengachari S, Aschauer P, Schittmayer M, Mayer N, Gruber K, Breinbauer R, Birner-Gruenberger R, Dreveny I, Oberer M: **Conformational plasticity and ligand binding of bacterial monoacylglycerol lipase.** *J Biol Chem* 2013, **288:**31093–31104.
12. Salter MD, Blouin GC, Soman J, Singleton EW, Dewilde S, Moens L, Pesce A, Nardini M, Bolognesi M, Olson JS: **Determination of ligand pathways in globins: apolar tunnels versus polar gates.** *J Biol Chem* 2012, **287:**33163–33178.
13. Voss NR, Gerstein M, Steitz TA, Moore PB: **The geometry of the ribosomal polypeptide exit tunnel.** *J Mol Biol* 2006, **360:**893–906.
14. Lemoine D, Jiang R, Taly A, Chataigneau T, Specht A, Grutter T: **Ligand-gated Ion channels: new insights into neurological disorders and ligand recognition.** *Chem Rev* 2012, **112:**6285–6318.
15. Kasianowicz JJ: **Introduction to Ion channels and disease.** *Chem Rev* 2012, **112:**6215–6217.
16. Knight AM, Culviner PH, Kurt-Yilmaz N, Zou T, Ozkan SB, Cavagnero S: **Electrostatic effect of the ribosomal surface on nascent polypeptide dynamics.** *ACS Chem Biol* 2013, **8:**1195–1204.
17. Eisenberg B: **Ionic channels in biological membranes: natural nanotubes.** *Acc Chem Res* 1998, **4842:**117–123.
18. Wallace B: **Gramicidin channels and pores.** *Annu Rev Biophys Biophys Chem* 1990, **19:**127–157.
19. Roux B: **Computational studies of the gramicidin channel.** *Acc Chem Res* 2002, **35:**366–375.
20. Maffeo C, Bhattacharya S, Yoo J, Wells D, Aksimentiev A: **Modeling and simulation of Ion channels.** *Chem Rev* 2012, **112:**6250–6284.
21. Kraut DA, Carroll KS, Herschlag D: **Challenges in enzyme mechanism and energetics.** *Annu Rev Biochem* 2003, **72:**517–571.
22. Warshel A, Sharma PK, Kato M, Xiang Y, Liu H, Olsson MHM: **Electrostatic basis for enzyme catalysis.** *Chem Rev* 2006, **106:**3210–3235.
23. Garcia-Viloca M, Gao J, Karplus M, Truhlar DG: **How enzymes work: analysis by modern rate theory and computer simulations.** *Science* 2004, **303:**186–195.
24. Benkovic S, Hammes-Schiffer S: **A perspective on enzyme catalysis.** *Science* 2003, **301:**1196–1202.
25. Porter CT, Bartlett GJ, Thornton JM: **The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, **32**(Database issue):D129–D133.
26. Pavlova M, Klvana M, Prokop Z, Chaloupkova R, Banas P, Otyepka M, Wade RC, Tsuda M, Nagata Y, Damborsky J: **Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate.** *Nat Chem Biol* 2009, **5:**727–733.
27. Stepankova V, Khabiri M, Brezovsky J, Pavelka A, Sykora J, Amaro M, Minofar B, Prokop Z, Hof M, Ettrich R, Chaloupkova R, Damborsky J: **Expansion of access tunnels and active-site cavities influence activity of haloalkane dehalogenases in organic cosolvents.** *Chembiochem* 2013, **14:**890–897.
28. Skopalík J, Anzenbacher P, Otyepka M: **Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences.** *J Phys Chem B* 2008, **112:**8165–8173.

29. Hendrychová T, Berka K, Navrátilová V, Anzenbacher P, Otyepka M: **Dynamics and hydration of the active sites of mammalian cytochromes P450 probed by molecular dynamics simulations.** *Curr Drug Metab* 2012, **13:**177–189.
30. Sehnal D, Svobodová Vařeková R, Berka K, Pravda L, Navrátilová V, Banáš P, Ionescu C-M, Otyepka M, Koča J: **MOLE 2.0: advanced approach for analysis of biomacromolecular channels.** *J Cheminform* 2013, **5:**39.
31. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157:**105–132.
32. Zimmerman JM, Eliezer N, Simha R: **The characterization of amino acid sequences in proteins by statistical methods.** *J Theor Biol* 1968, **21:**170–201.
33. Webby CJ, Lott JS, Baker HM, Baker EN, Parker EJ: **Crystallization and preliminary X-ray crystallographic analysis of 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase from Mycobacterium tuberculosis.** *Acta Crystallogr Sect F: Struct Biol Cryst Commun* 2005, **61**(Pt 4):403–406.
34. Houborg K, Harris P, Petersen J, Rowland P, Poulsen J-CN, Schneider P, Vind J, Larsen S: **Impact of the physical and chemical environment on the molecular structure of Coprinus cinereus peroxidase.** *Acta Crystallogr Sect D: Biol Crystallogr* 2003, **D59:**989–996.
35. Lundell TK, Mäkelä MR, Hildén K: **Lignin-modifying enzymes in filamentous basidiomycetes–ecological, functional and phylogenetic review.** *J Basic Microbiol* 2010, **50:**5–20.
36. Cherry JR, Lamsa MH, Schneider P, Vind J, Svendsen A, Jones A, Pedersen AH: **Directed evolution of a fungal peroxidase.** *Nat Biotechnol* 1999, **17:**379–384.
37. Holliday GL, Mitchell JBO, Thornton JM: **Understanding the functional roles of amino acid residues in enzyme catalysis.** *J Mol Biol* 2009, **390:**560–577.
38. Dill KA: **Dominant forces in protein folding.** *Biochemistry* 1990, **29:**7133–7155.
39. Wilkinson B, Gilbert HF: **Protein disulfide isomerase.** *Biochim Biophys Acta* 2004, **1699:**35–44.
40. Furnham N, Holliday GL, de Beer TAP, Jacobsen JOB, Pearson WR, Thornton JM: **The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes.** *Nucleic Acids Res* 2014, **42:**D485–D489.
41. Berka K, Hanák O, Sehnal D, Banáš P, Navrátilová V, Jaiswal D, Ionescu C-M, Svobodová Vařeková R, Koča J, Otyepka M: **MOLEonline 2.0: interactive web-based analysis of biomacromolecular channels.** *Nucleic Acids Res* 2012, **40**(Web Server issue):W222–W227.
42. Berman H, Henrick K, Nakamura H, Markley JL: **The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.** *Nucleic Acids Res* 2007, **35**(Database issue):D301–D303.

doi:10.1186/s12859-014-0379-x
Cite this article as: Pravda *et al.*: Anatomy of enzyme channels. *BMC Bioinformatics* 2014 **15**:379.

MotiveValidator: interactive web-based validation of ligand and residue structure in biomolecular complexes

Nucleic Acids Research Advance Access published May 21, 2014
Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku426

MotiveValidator: interactive web-based validation of ligand and residue structure in biomolecular complexes

Radka Svobodová Vařeková[1,2,†], Deepti Jaiswal[1,†], David Sehnal[1,2,3], Crina-Maria Ionescu[1], Stanislav Geidl[1,2], Lukáš Pravda[1,2], Vladimír Horský[3], Michaela Wimmerová[1,2,*] and Jaroslav Koča[1,2,*]

[1]CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno, Czech Republic, [2]National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic and [3]Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic

Received January 30, 2014; Revised May 02, 2014; Accepted May 2, 2014

ABSTRACT

Structure validation has become a major issue in the structural biology community, and an essential step is checking the ligand structure. This paper introduces MotiveValidator, a web-based application for the validation of ligands and residues in PDB or PDBx/mmCIF format files provided by the user. Specifically, MotiveValidator is able to evaluate in a straightforward manner whether the ligand or residue being studied has a correct annotation (3-letter code), i.e. if it has the same topology and stereochemistry as the model ligand or residue with this annotation. If not, MotiveValidator explicitly describes the differences. MotiveValidator offers a user-friendly, interactive and platform-independent environment for validating structures obtained by any type of experiment. The results of the validation are presented in both tabular and graphical form, facilitating their interpretation. MotiveValidator can process thousands of ligands or residues in a single validation run that takes no more than a few minutes. MotiveValidator can be used for testing single structures, or the analysis of large sets of ligands or fragments prepared for binding site analysis, docking or virtual screening. MotiveValidator is freely available via the Internet at http://ncbr.muni.cz/MotiveValidator.

INTRODUCTION

Validation arose as a major issue in the structural biology community when it became apparent that some published structures contained serious errors (1–6). Various tools for the validation of the protein and nucleic acid 3D structures are well established, such as WHAT_CHECK (7), PROCHECK (8), MolProbity (9) and OOPS (10).

An essential step in the validation process is checking the ligand structure. Ligands are chemical compounds which form a complex with a biomacromolecule (e.g. sugar, drug, heme) and play a key role in its function. The ligands are also the main source of errors in structures (11,12). Nonetheless, ligand validation is a very challenging task (13), because of the high diversity and nontriviality of their structure and the general lack of information about correct structures. Therefore, early validation tools focused on selected types of ligands (PDB-care (14) focused on carbohydrates) and their scope only widened later (ValLigURL (15)). Ligand validation features were recently added to existing software (e.g. Mogul (16), Coot (17)). New tools such as PHENIX (18) were developed to include ligand validation functionality. However, the functionality of some available tools (i.e. ValLigURL, Mogul, Coot, PHENIX) is aimed at the validation of selected properties (atom clashes, bond lengths, bond angles, etc.) or is limited to a selected type of molecules (e.g. PDB-care validates only carbohydrates).

This article presents the web-based application MotiveValidator, which offers a user-friendly, interactive and platform-independent environment for the validation of ligands and residues in PDB (http://www.wwpdb.org/docs.html) or PDBx/mmCIF (19) files provided by the user.

*To whom correspondence should be addressed. Tel: +420 54949 4947; Fax: +420 54949 2556; Email: Jaroslav.Koca@ceitec.muni.cz
Correspondence may also be addressed to Michaela Wimmerová. Tel: +420 54949 3805; Fax: +420 54949 2690; Email: michaw@chemi.muni.cz
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Residues refer to any component of a biomacromolecule or a biomacromolecular complex (i.e. amino acids, nucleotides, ligands). Specifically, MotiveValidator is able to evaluate in a straightforward manner whether the ligand or residue under study has a correct annotation (3-letter code), i.e. if it has the same topology and stereochemistry as the model ligand or residue with this annotation. If not, MotiveValidator explicitly describes the differences. Validation is performed against so-called model residues, which can be either correct structures of the residue obtained from the wwPDB Chemical Components Dictionary (20) (accessed via the web interface provided by LigandExpo (21)), or against templates provided by the user. The output provides a report of the validation results, including summary and detailed information in both tabular and graphical form. MotiveValidator can process thousands of ligands or residues in a single validation run that takes no more than a few minutes.

MotiveValidator can be used for testing single structures, or the analysis of large sets of ligands or fragments prepared for binding site analysis, docking or virtual screening. A significant advantage of MotiveValidator is the ability to process structures obtained by any type of experiment and not requiring the user to have any additional knowledge in the field of X-ray crystallography or nuclear magnetic resonance.

DESCRIPTION OF THE TOOL

MotiveValidator incorporates several tools for the detection and extraction of residues (MotiveQuery; D. Sehnal *et al.*, unpublished work), motif superimposition (SiteBinder (22)), chirality verification (OpenBabel (23)), statistical evaluation of results (in-house program) and interactive visualization of 3D structures (ChemDoodle, http://www.chemdoodle.com). All these tools are integrated into a single program which runs on a server and is accessible under any operating system. The built-in 3D molecular visualizer requires an up-to-date web browser with WebGL enabled. In addition to running validations on the server, a command line version of MotiveValidator is also available.

MotiveValidator enables three kinds of validation to be performed, accessible via three modules. Residue Validation is the most general module, meant for any residue, including ligands. Sugar Validation is focused on carbohydrates and Motif/Fragment Validation on biomolecular fragments (motifs). A motif can in principle be any part of a biomacromolecule. Nonetheless, MotiveValidator is focused on the validation of residues, thus here motif generally refers to the residue under study, together with its immediate environment. Validation via any module involves three steps, namely setup, calculation and finally visualization and the analysis of results. We provide here an extensive description of the Residue Validation module and then briefly point out the differences for the other two modules.

Residue validation

Setup. Two kinds of input are required, namely the structure of a biomolecule or biomolecular complex to be validated and a model residue to serve as the reference template for validation (Supplementary Figure S1). The structure to be validated and model residue must be uploaded in PDB format, or can be retrieved in this format from the mirrors of the Protein Data Bank (24) and LigandExpo databases maintained on the MotiveValidator server and updated every week. The structure to be validated can also be uploaded in PDBx/mmCIF format. A single MotiveValidator run can validate multiple residues in multiple structures.

Calculation. After the setup, the validation proceeds in several steps. The sequence of steps performed during validation is as follows (see also Supplementary Figure S2 for a graphical dictionary of the main terms that appear in this section):

(i) In the structure(s) to be validated, find all instances of residues with the same 3-letter code as the model residue.
(ii) Extract the identified residues (i.e. residues to be validated) together with their immediate surroundings (i.e. atoms within one or two bonds of any atom of the residue to be validated), to obtain input motifs for validation.
(iii) For each input motif:
  (a) Superimpose the input motif with the model residue to find the best atom pairing, i.e. the correspondence (mapping) between atoms from the model residue and from the input motif. Mathematically, it is the bijection which matches the most atoms from the input motif to the most atoms from the model residue and provides the lowest RMSD (root mean square deviation) for the structural superimposition. PDB names of atoms are not used in this step. The subset of atoms from the input motif paired with atoms in the model residue forms the validated motif. The atoms in a validated motif are checked for connectivity, to ensure that it is the same as in the model residue. Report any discrepancy between the inter-atomic bonds in the validated motif and in the model residue (section Processing Errors/Warnings).
  (b) Establish the validated motif according to the best atom pairing identified in the previous step. Based on the validated motif, detect and report errors:
    - missing atoms: atom in the model residue with no corresponding atom in the validated structure
    - missing rings: missing atoms originating from cycles (rings)
    - wrong chirality: atom from the validated motif with different chirality than the corresponding atom from the model residue;

    and warnings:
    - substitutions: atom from the validated motif with different chemical symbol than the corresponding atom in the model residue (e.g. O mapped to N)
    - different atom name: atom from the validated motif with different PDB name than the corresponding atom from the model residue (e.g. the C1 atom mapped to the C7 atom)

- foreign atoms: atom from the model residue mapped to atom from outside the validated residue (i.e. from its surroundings).

Note: An occurrence of a warning does not mean that the validated motif is wrong. The warning serves only as information to the user.

Visualization and processing of results. All setup information, along with all input and output structures and files are deposited on the server in a unique directory, translated as a unique URL accessible for visualization and download for at least a month. The MotiveValidator output provides a straightforward report of the validation results, including a summary and detailed information in both tabular and graphical form, along with a 3D structure visualizer for closer inspection of the problematic structures.

The Summary section first provides a description of the validation process and then a validation report for each validated residue (Figure 1). The report contains information about the model residue (annotation, 2D structure) and an overview (table and pie chart) of issues found during validation, namely, the number of residues with missing atoms, missing (incomplete) rings, wrong chirality, correct chirality, substitutions, different atom names and foreign atoms. A list of specific issues and their localization within the residue (i.e. number of residues with particular missing atoms or atoms having wrong chirality) is also given.

The Details section (Figure 2, top) provides detailed information for each validated motif. It is organized into a table with one line per motif, containing basic identification of the motif inside the original input file and a list of all issues identified during validation. Each motif can be examined in the 3D space and a complete validation report is available in graphical form using the individual motif links (Figure 2, bottom).

The additional section Processing Errors/Warnings lists the issues found while processing the input files. Processing warnings are issues that may cause incorrect validation, such as atoms that are too close in the 3D space. Processing errors are major issues preventing the finalization of the validation, such as parts of the residue which are completely disconnected from the rest of the structure, probably due to missing atoms at multiple locations throughout the structure.

Sugar validation

A notable case of ligand validation is the analysis of carbohydrate structures because they have complex topology and many chiral atoms. Carbohydrates are involved in a variety of fundamental biological processes and have significant pharmaceutical and diagnostic potential. Additionally, more than 60% of nontrivial-sized ligands (>10 atoms) from the PDB contain a carbohydrate. For these reasons, MotiveValidator includes the mode Sugar Validation, which was developed specifically for the validation of carbohydrates. Unlike Residue Validation, the Sugar Validation setup stage requires only one input, namely the biomolecule(s) containing residues to be validated. This mode enables the automatic validation of all carbohydrate residues identified in the input structure(s). Specifically, MotiveValidator identifies all motifs containing pyran or furan rings as saccharides and validates them against the corresponding model residues (same 3-letter code) retrieved from the LigandExpo mirror.

Motif/fragment validation

The Motif/Fragment Validation mode uses the model residue and fragments of biomolecules as the input, as opposed to entire biomolecules in the Residue Validation mode. The motifs (fragments) should contain the validated residue and its closest surrounding. The surrounding can include, e.g. atoms within one or two bonds of any atom of the validated residue or more. However, it must stay clear, which residue is the validated one. Therefore, the surrounding can contain just fragments of neighboring residues, but not the whole neighboring residues. It is very useful for the efficient processing of very large amounts of data, such as validating all instances of a residue in the entire PDB. The calculation skips steps (i) and (ii) related to residue detection and extraction, and instead starts directly with the superimposition [step (iii)] of the model residue and validated fragments. The fragments can be prepared manually or automatically. The MotiveValidator website also provides the utility MotifExtractor to enable automatic extraction of the desired motifs (residues and their surroundings) from large datasets of biomolecular structures.

RESULTS AND DISCUSSION

We provide examples of uses for MotiveValidator in the form of case studies for each of the three validation modes.

Residue validation: all proteins containing cholic acid

Cholic acid (CHD) is the best known bile acid and includes four rings and 11 chiral atoms. It contains three 6-member rings A, B and C in chair conformation and a 5-member ring D (Supplementary Figure S3A and B) (25). The PDB contains 299 instances of CHD as ligand in a total of 55 PDB entries (access date: 5.1.2014). We collected all 55 structures and validated all occurrences of CHD using the Residue Validation mode in MotiveValidator. The validation (Figure 1) took 15 s and showed that all 299 CHD instances are complete (no missing atoms). However, the validation revealed that almost 13% of the CHD ligands have incorrect chirality. The problematic molecules can be organized into three groups. The first group contains 18 ligands from nine PDB entries, with incorrect chirality at atoms C3, C8, C9, C12 and C14. The errors are caused by the unnatural boat conformation of rings A, B and C in these particular structures (Supplementary Figure S3C). All these structures come from bovine heart cytochrome c oxidase and were published by the same lab. The second group contains 18 ligands from the same nine PDB entries, with incorrect chirality at atoms C8, C9, C12, C14 and C17. The errors are caused by the unnatural twist-boat conformation of rings A, B and C (Supplementary Figure S3D). The third group contains two ligands from the H240A variant of human ferrochelatase (PDB ID 3AQI), with incorrect chirality at atom C20.

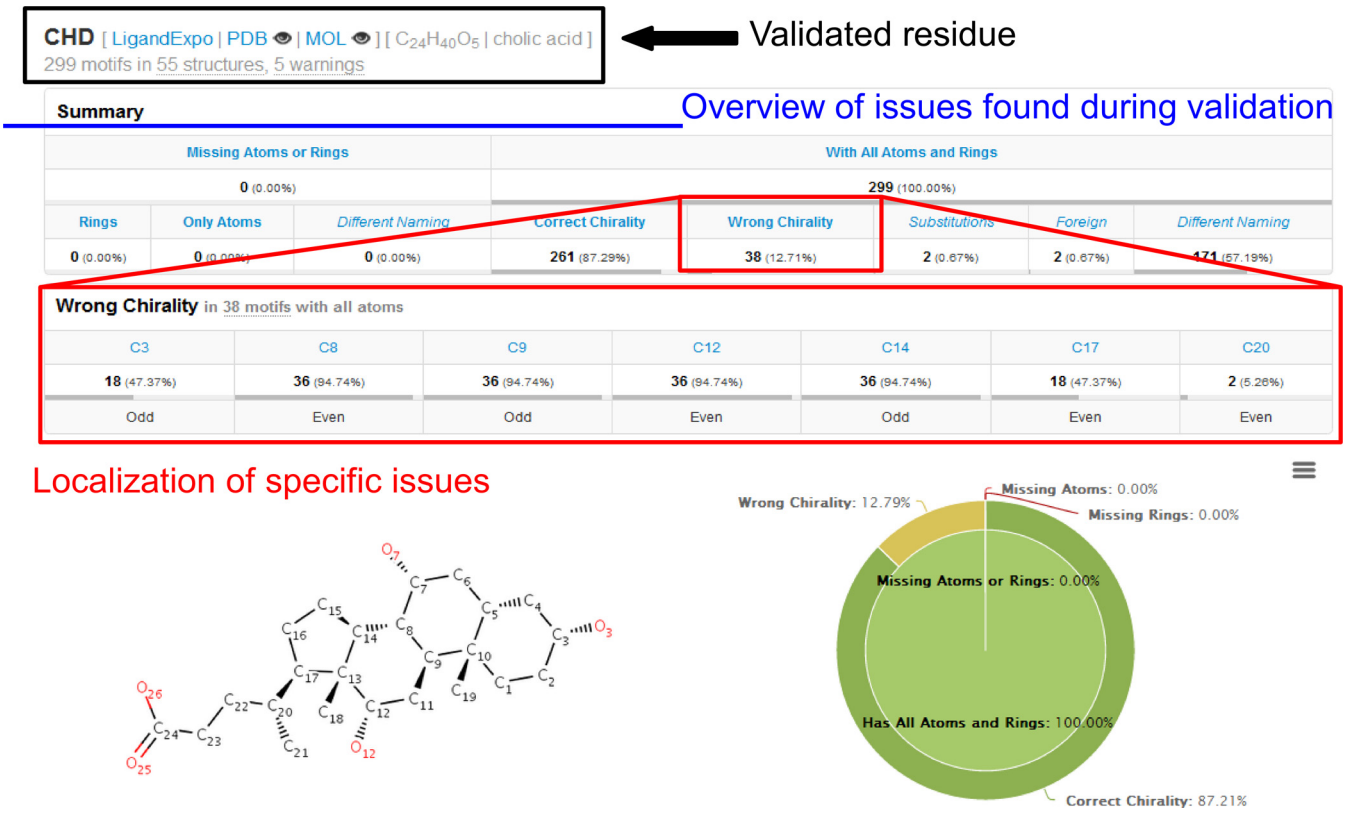


Figure 1. The Summary tab first provides a description of the validation process and a summary of the results in tabular and graphical form. An overview of the issues related to incomplete structure or incorrect chirality is given here, along with other useful notes. The problematic atoms are further highlighted to better localize the problems in the structures.

The complete results are available at the MotiveValidator website as a Sample calculation (http://ncbr.muni.cz/MotiveValidator/ProteinsWithCHD).

Sugar validation: nipah G attachment glycoprotein complexed with ephrin-B3

Nipah virus infection may lead to severe respiratory disease and fatal encephalitis in humans. The Nipah virus relies on the Nipah G attachment glycoprotein for host cell recognition. The crystal structure of the glycoprotein complexed with its receptor ephrin-B3 (PDB ID 3D12, (26)) contains 30 instances of 11 different carbohydrates, each with one ring and five chiral atoms: β-D-glucose (BGC), β-D-mannose (BMA), β-D-gulopyranose (GL0), α-D-glucose (GLC), α-L-galactopyranose (GXL), 2-(acetylamino)-2-deoxy-β-D-gulopyranose (LXB), 2-(acetylamino)-2-deoxy-α-D-idopyranose (LXZ), α-D-mannose (MAN), *N*-acetyl-D-glucosamine (NAG), *N*-acetyl-D-galactosamine (NGA) and 2-(acetylamino)-2-deoxy-α-L-glucopyranose (NGZ). Note that the names of the carbohydrates were obtained from LigandExpo and prefixes alpha- and beta- were replaced with α− and β− (see Supplementary Table TS1 for IUPAC systematic names). We validated all carbohydrate structures in this biomacromolecular complex using the Sugar Validator mode. The validation showed that 13 of these ligands had incorrect chirality (Supplementary Figure S4). In the few cases with GLC or NGA ligands, all five chiral atoms exhibited incorrect chirality. Manual inspection of the structure showed further discrepancies in the ligand part. This is discussed in details in the Supplementary material (Supplementary Figure S5).

The complete results are available at the MotiveValidator website as a Sample calculation (http://ncbr.muni.cz/MotiveValidator/ComplexedGlycoprotein).

Motif/fragment validation: all *N*-acetyl-D-glucosamine residues from PDB

N-acetyl-D-glucosamine (NAG) is the second most frequent hetero-atom chemical component found in the PDB, amounting to 24 357 instances as ligands in a total of 3905 PDB entries (access date: 9.1.2014). NAG includes one pyran ring and five chiral atoms (Supplementary Figure S6A). We extracted all 24 357 NAG instances from the PDB using MotifExtractor. Each file contained one NAG motif, composed of a NAG residue and the atoms in its immediate surroundings (atoms within one or two bonds of the NAG residue). These motifs were validated using the Motif/Fragment Validation mode. The validation (Figure 2) took 195 s and revealed that 94% of NAG instances in the PDB are complete and have correct chirality. In addition, several issues were reported.

First, 16 NAG residues exhibit serious problems: some only contain a few atoms, others have errors in their bond information described by CONNECT keywords (see exam-

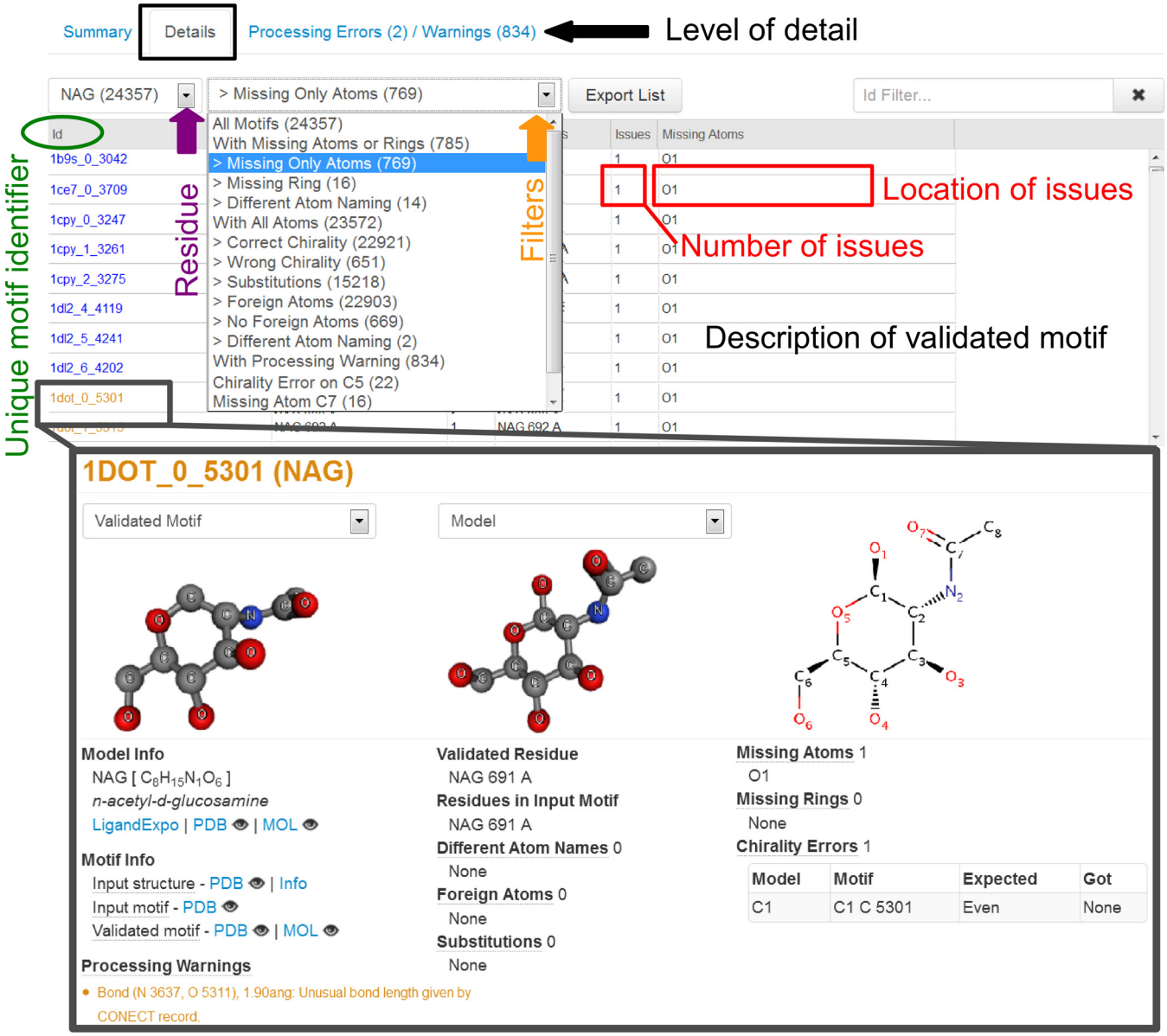


Figure 2. The Details tab enables the issues in selected groups of motifs to be inspected by specifying the residue name and type of issue. All information pertaining to a given motif is provided in a single row. Further, each motif can be examined in the 3D space and a complete validation report is accessible via the individual motif links.

ple in Supplementary Figure S6B). Second, approximately 3.5% of NAG residues are missing at least one atom. In most cases, the O1 atom is missing. Third, 2.7% of NAG residues have wrong chirality, mostly at C1, since that is the main site of covalent connection to other residues, which can cause a change in chirality. Some of the chirality errors are caused by incorrect placement of the ligand inside the electron density map. For example, residue NAG 2 A from the PDB entry 3A4X exhibits incorrect chirality at atom C2 (Supplementary Figure S6C). Using Coot and the corresponding electron density maps downloaded from the EDS server at Uppsala University (27), we found that NAG is not placed correctly in the electron density map, leading to a deformation in the vicinity of C2. New positioning leads to a conformation which fits the experimental 3D electron density map markedly better and which has the correct chirality at position C2 (Supplementary Figure S6D).

Additionally, MotiveValidator found that over 60% of NAG residues in the PDB have a nitrogen substitution at O1, which indicates their participation in *N*-glycosylation. The ability to process and validate also residues with substitutions is an advantage of MotiveValidator.

The complete results are available at the MotiveValidator site as a Sample calculation (http://ncbr.muni.cz/MotiveValidator/MotifsNAG).

Limitations

MotiveValidator is limited in three main ways. First, there is the requirement to ensure that the model residue serv-

ing as the reference during validation is indeed correct. This limitation is overcome by using high-quality reference residues from LigandExpo. Second, the superimposition phase might not identify the optimal matching between the atoms of the model residue and those of the validated residue if their 3D structures are too different. Finally, software and data handling on the server currently limits the maximum size of the input file with structures to be validated (PDB or ZIP file) to 300 MB. We plan to minimize these limitations in the next version of MotiveValidator. For example, we will explore the use of additional metrics to improve the second limitation.

CONCLUSION

In this article we introduced MotiveValidator, a web-based interactive tool for validating ligand and residue structure in biomolecular complexes. The MotiveValidator interface is easy to use and platform-independent, enables interactive analyses with a high degree of automation, e.g. retrieving structures from local mirrors of the PDB and LigandExpo databases, automatic detection and extraction of sugars or selected residues, including their immediate surroundings. Results are presented in a clear graphical and tabular form, facilitating their interpretation and further processing.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

The authors wish to thank Prof. Gerard Kleywegt and Dr Sameer Velankar, both EMBL-EBI, Hinxton, UK, for their useful comments on the manuscript.

FUNDING

This work was funded by the Ministry of Education, Youth and Sports of the Czech Republic [LH13055], the CEITEC - Central European Institute of Technology [CZ.1.05/1.1.00/02.0068] from the European Regional Development Fund, the "Capacities" specific program [286154] and by INBIOR [CZ.1.07/2.3.00/20.0042] from the European Social Fund and the state budget of the Czech Republic. Additional support was provided by the project "Employment of Newly Graduated Doctors of Science for Scientific Excellence" [CZ.1.07/2.3.00/30.0009] co-financed from the European Social Fund and the state budget of the Czech Republic. Funding for open access charge: INBIOR [CZ.1.07/2.3.00/20.0042] from the European Social Fund and the state budget of the Czech Republic.
Conflict of interest statement. None declared.

REFERENCES

1. Kleywegt,G.J. (2000) Validation of protein crystal structures. *Acta Crystallogr. D Biol. Crystallogr.*, **56**, 249–265.
2. Kleywegt,G.J. (2009) On vital aid: the why, what and how of validation. *Acta Crystallogr. D. Biol. Crystallogr.*, **65**, 134–139.
3. Kleywegt,G.J. (2007) Crystallographic refinement of ligand complexes. *Acta Crystallogr. D. Biol. Crystallogr.*, **63**, 94–100.
4. Davis,A.M., St-Gallay,S.A. and Kleywegt,G.J. (2008) Limitations and lessons in the use of X-ray structural information in drug design. *Drug Discov. Today*, **13**, 831–841.
5. Spek,A.L. (2009) Structure validation in chemical crystallography. *Acta Crystallogr. D. Biol. Crystallogr.*, **65**, 148–155.
6. Gutmanas,A., Oldfield,T.J., Patwardhan,A., Sen,S., Velankar,S. and Kleywegt,G.J. (2013) The role of structural bioinformatics resources in the era of integrative structural biology. *Acta Crystallogr. D. Biol. Crystallogr.*, **69**, 710–721.
7. Hooft,R.W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272.
8. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
9. Chen,V.B., Arendall,W.B., Headd,J.J., Keedy,D.A., Immormino,R.M., Kapral,G.J., Murray,L.W., Richardson,J.S. and Richardson,D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 12–21.
10. Kleywegt,G.J. and Jones,T.A. (1996) Efficient rebuilding of protein structures. *Acta Crystallogr. D. Biol. Crystallogr.*, **52**, 829–832.
11. Read,R.J., Adams,P.D., Arendall,W.B., Brunger,A.T., Emsley,P., Joosten,R.P., Kleywegt,G.J., Krissinel,E.B., Lütteke,T., Otwinowski,Z. *et al.* , (2011) A new generation of crystallographic validation tools for the protein data bank. *Structure*, **19**, 1395–1412.
12. Gore,S., Velankar,S. and Kleywegt,G.J. (2012) Implementing an X-ray validation pipeline for the Protein Data Bank. *Acta Crystallogr. D. Biol. Crystallogr.*, **68**, 478–483.
13. Kleywegt,G.J., Henrick,K., Dodson,E.J. and van Aalten,D.M.F. (2003) Pound-wise but penny-foolish: how well do micromolecules fare in macromolecular refinement? *Structure*, **11**, 1051–1059.
14. Lütteke,T. and von der Lieth,C.-W. (2004) pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files. *BMC Bioinformatics*, **5**, 69.
15. Kleywegt,G.J. and Harris,M.R. (2007) ValLigURL: a server for ligand-structure comparison and validation. *Acta Crystallogr. D. Biol. Crystallogr.*, **63**, 935–938.
16. Bruno,I.J., Cole,J.C., Kessler,M., Luo,J., Motherwell,W.D.S., Purkis,L.H., Smith,B.R., Taylor,R., Cooper,R.I., Harris,S.E. *et al.* , Retrieval of crystallographically-derived molecular geometry information. *J. Chem. Inf. Comput. Sci.*, **44**, 2133–2144.
17. Debreczeni,J.É. and Emsley,P. (2012) Handling ligands with Coot. *Acta Crystallogr. D. Biol. Crystallogr.*, **68**, 425–430.
18. Adams,P.D., Afonine,P. V, Bunkóczi,G., Chen,V.B., Davis,I.W., Echols,N., Headd,J.J., Hung,L.-W., Kapral,G.J., Grosse-Kunstleve,R.W. *et al.* , (2010) PHENIX: a comprehensive python-based system for macromolecular structure solution. *Acta Crystallogr. D. Biol. Crystallogr.*, **66**, 213–221.
19. The Protein Data Bank exchange dictionary (PDF), Westbrook,J., Henrick,K., Ulrich,E.L. and Berman,H.M. (2005) *International Tables for Crystallography G. Definition and exchange of crystallographic data*. In: Hall,SR and McMahon,B (eds). Springer, Dordrecht, pp. 295–298.
20. Henrick,K., Feng,Z., Bluhm,W.F., Dimitropoulos,D., Doreleijers,J.F., Dutta,S., Flippen-Anderson,J.L., Ionides,J., Kamada,C., Krissinel,E. *et al.* , (2008) Remediation of the protein data bank archive. *Nucleic Acids Res.*, **36**, D426–D433.
21. Feng,Z., Chen,L., Maddula,H., Akcan,O., Oughtred,R., Berman,H.M. and Westbrook,J. (2004) Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics*, **20**, 2153–2155.
22. Sehnal,D., Vařeková,R.S., Huber,H.J., Geidl,S., Ionescu,C.-M., Wimmerová,M. and Koča,J. (2012) SiteBinder: an improved approach for comparing multiple protein structural motifs. *J. Chem. Inf. Model.*, **52**, 343–359.
23. O'Boyle,N.M., Banck,M., James,C.A., Morley,C., Vandermeersch,T. and Hutchison,G.R. (2011) Open Babel: an open chemical toolbox. *J. Cheminform.*, **3**, 33.
24. Berman,H., Henrick,K., Nakamura,H. and Markley,J.L. (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.*, **35**, D301–D303.
25. Mukhopadhyay,S. and Maitra,U. (2004) Chemistry and biology of bile acids. *Curr. Sci.*, **87**, 1666–1683.

26. Xu,K., Rajashankar,K.R., Chan,Y.-P., Himanen,J.P., Broder,C.C. and Nikolov,D.B. (2008) Host cell recognition by the henipaviruses: crystal structures of the Nipah G attachment glycoprotein and its complex with ephrin-B3. *Proc. Natl. Acad. Sci. U.S.A.*, **105**, 9953–9958.

27. Kleywegt,G.J., Harris,M.R., Zou,J.Y., Taylor,T.C., Wählby,A. and Jones,T.A. (2004) The uppsala electron-density server. *Acta Crystallogr. D Biol. Crystallogr.*, **60**, 2240–2249.

MOLE 2.0: advanced approach for analysis of biomacromolecular channels

Sehnal et al. Journal of Cheminformatics 2013, 5:39
http://www.jcheminf.com/content/5/1/39

Journal of Cheminformatics

SOFTWARE Open Access

MOLE 2.0: advanced approach for analysis of biomacromolecular channels

David Sehnal[1,2], Radka Svobodová Vařeková[1], Karel Berka[3], Lukáš Pravda[1], Veronika Navrátilová[3], Pavel Banáš[3], Crina-Maria Ionescu[1], Michal Otyepka[3*] and Jaroslav Koča[1*]

Abstract

Background: Channels and pores in biomacromolecules (proteins, nucleic acids and their complexes) play significant biological roles, e.g., in molecular recognition and enzyme substrate specificity.

Results: We present an advanced software tool entitled MOLE 2.0, which has been designed to analyze molecular channels and pores. Benchmark tests against other available software tools showed that MOLE 2.0 is by comparison quicker, more robust and more versatile. As a new feature, MOLE 2.0 estimates physicochemical properties of the identified channels, i.e., hydropathy, hydrophobicity, polarity, charge, and mutability. We also assessed the variability in physicochemical properties of eighty X-ray structures of two members of the cytochrome P450 superfamily.

Conclusion: Estimated physicochemical properties of the identified channels in the selected biomacromolecules corresponded well with the known functions of the respective channels. Thus, the predicted physicochemical properties may provide useful information about the potential functions of identified channels. The MOLE 2.0 software is available at http://mole.chemi.muni.cz.

Keywords: Channels, Tunnels, Pores, Protein structures, Cytochrome P450, CAM, BM3

Background

The number of known three-dimensional (3D) structures of biomacromolecules (proteins, nucleic acids and their complexes) has increased rapidly over recent years, enabling relationships between structure and function to be analyzed at an atomic level. The functions of biomacromolecules usually depend on interactions with other biomacromolecules as well as ions and small molecules, such as water, messenger and endogenous compounds, pollutants and drugs, which can occupy "otherwise empty spaces" in biomacromolecular structures [1]. Thus, information about the nature of empty spaces in a biomacromolecule can provide valuable insights into its functions.

Biomacromolecular empty spaces can be classified as pockets, cavities, voids, channels (tunnels) or pores (Figure 1). A *pocket* usually refers to a shallow depression on a biomacromolecular surface, whereas a *cavity* describes a deeper pocket or cleft. If the cavity is encapsulated inside a biomolecule (having no connection to a water accessible surface), it is called a *void*. A *channel* or *tunnel* is a pathway inside a cavity connecting an internal point (typically the deepest apex) with an exterior. A *pore* is considered here as a channel that passes through the biomacromolecule from one point on the surface to another.

The present work focused on pores and channels because they have been shown to play significant roles in many biologically relevant systems. For example, internal pores of ion channels maintain a highly selective ionic balance between the cell interior and exterior, [2-6] photosystem II channels are involved in photosynthesis, [7,8] ribosomal polypeptide exit channels allow nascent peptides to leave the ribosome during translation, [9] and active site access/egress channels enable substrate/product to enter/leave the occluded active sites of various enzymes (e.g., cytochrome P450, [10-15] acetylcholinesterase, [16-18] etc.). Information about the nature of active site access paths can also be utilized in biotechnology applications aimed at designing more effective and selective enzymes [19-21]. Unquestionably,

* Correspondence: michal.otyepka@upol.cz; jaroslav.koca@ceitec.muni.cz
[3]Department of Physical Chemistry, Regional Centre of Advanced Technologies and Materials, Faculty of Science, Palacký University Olomouc, tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
[1]National Centre for Biomolecular Research, Faculty of Science and CEITEC-Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic
Full list of author information is available at the end of the article

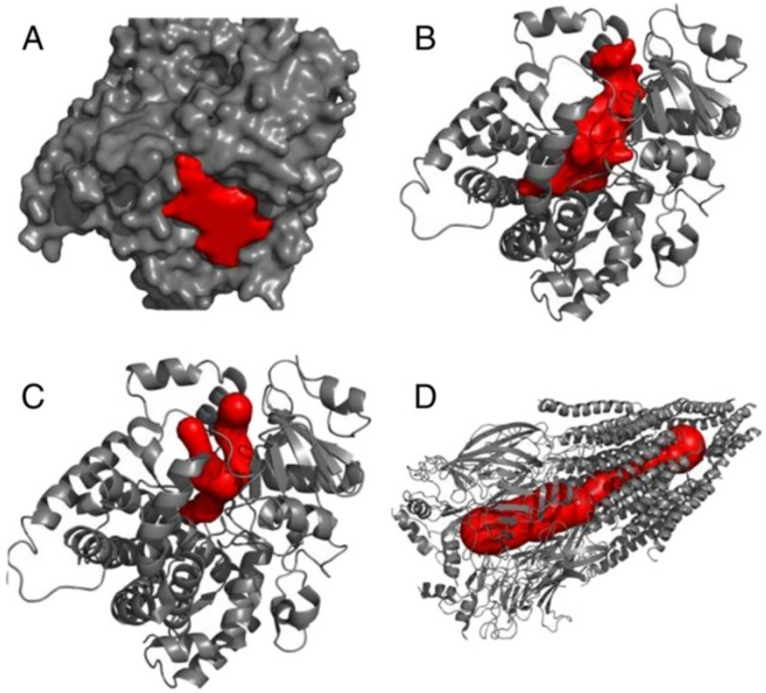


Figure 1 Classification of biomacromolecular "empty spaces": A) pockets, B) cavities, C) channels (or tunnels), and D) pores.

identification and characterization of channels are fundamental to understanding numerous biologically relevant processes and serve as a starting point for rational drug design, protein engineering and biotechnological applications.

Over the last few years, numerous computational approaches have been developed for detection and characterization of empty spaces in biomacromolecules, particularly proteins [22]. The main strategies used in the developed algorithms can be grouped into four classes [23]. The first class comprises grid-based methods, which project biomacromolecular structures onto a 3D grid, process the void grid voxels and connect them into pockets or tunnels. These methods are used in numerous software tools, such as POCKET [24], LIGSITE [25,26], dxTuber [27], HOLLOW [28], 3V [29], CAVER 1.x [30] and CHUNNEL [31]. Sphere-filling methods belong to a second class. These methods carpet biomacromolecules with spheres layer by layer. A cluster of carpeting spheres is considered a pocket. This method is implemented in PASS [32] and SURFNET [33]. The third class involves slice and optimization methods. These methods split a biomacromolecular structure into slices along a start vector defined by the user and then optimization methods are used to determine the largest sphere. These approaches are implemented in the software HOLE [34] and PoreWalker [35]. The fourth class represents methods utilizing Voronoi diagrams, in which the shortest path is searched from a starting point to the biomacromolecular surface. This approach was used in the previous version of MOLE 1.x [19] and it is also utilized in other software tools, e.g., MolAxis [36,37], CAVER 2.0 [38] and CAVER 3.0 [39].

Here, we present an advanced and fully automatic software tool, MOLE 2.0, based on a new, fast and robust algorithm for finding channels and pores. MOLE 2.0 provides an improved approach for channel identification. The algorithm introduces several preprocessing steps that result in increased speed (up to several times faster), accuracy (more relevant channels are identified) and robustness. New capabilities include the computation of pores and better identification of channel start points. It contains extended options for starting point selection and allows improved computation of channel profiles together with estimation of their basic physicochemical properties. The implemented automatic filtering of obtained channels facilitates selection of the relevant channels. MOLE 2.0 offers an innovative user experience, as it can be used effectively even without knowledge of the underlying algorithms whilst at the same time allows the tunnel detection algorithm to be tweaked interactively, such that the results are immediately available for inspection and comparison. MOLE 2.0 also introduces a new, intuitive and user-friendly interface. MOLE 2.0 can be used as a stand-alone application or as a plugin for the widely used software PyMOL [40]. Some functionality is also available in a platform-independent manner via the web-based application MOLE*online* 2.0 [41].

Implementation

MOLE 2.0 algorithm

The algorithm for finding channels implemented in MOLE 2.0 involves seven steps: i) computation of the Delaunay triangulation/Voronoi diagram of the atomic centers, ii) construction of the molecular surface, iii) identification of cavities, iv) identification of possible channel start points, v) identification of possible channel end points, vi) localization of channels, and vii) filtering of the localized channels (Figure 2).

Step i: computing the delaunay triangulation/voronoi diagram

In the first step, the Delaunay triangulation of the atomic centers is computed using an incremental algorithm that utilizes pre-sorted input points according to the Hilbert curve [19,42]. The Voronoi diagram is then constructed as the dual of the Delaunay triangulation. The Voronoi diagram can be represented as a graph with vertices corresponding to the circumcenters of the Delaunay tetrahedrons and edges present if two tetrahedrons share a common side (i.e., share exactly three vertices).

Steps ii and iii: approximating the molecular surface and identifying cavities

The molecular surface is approximated by iterative removal of boundary tetrahedrons from the outermost layers (i.e., tetrahedrons found at the interface between the molecule and the external environment). Boundary tetrahedrons produced by the triangulation are removed in this step if they are sufficiently large to fit a sphere with a given *probe radius* (tetrahedron T fits a sphere S with probe radius r if the center C of sphere S can be placed inside the tetrahedron and the distance to all vertices of T is greater than or equal to the sum of r, with the van der Waals radius of an atom corresponding to the given vertex). Next, tetrahedrons that are too small to fit a sphere with *interior radius* are removed. Remaining tetrahedrons form one or more connected components. We call the components that contain at least one tetrahedron on the molecular surface *cavity diagrams*. It should be noted here that the *cavity diagram* is a purely geometrical concept to help identify regions of space (volume) that can contain tunnels and only very loosely corresponds to the cavities shown in Figure 1B).

Steps iv and v: identifying possible start and end points of channels

The algorithm includes two ways to specify potential channel start and end points:

- *Computed*: Start and end points are defined as the centers of the *deepest* tetrahedrons in each cavity. The depth of the tetrahedron is defined as the number of Voronoi edges from the closest boundary tetrahedron.
- *User-defined*: Specified by a 3D point (that can also be defined as a centroid of several residues). Next, cavities that have at least one tetrahedron with a centroid within the *origin radius* from the user-specified point are found. Finally, for each such cavity, the start point is selected as the circumsphere center of the tetrahedron closest to the original point. Potential channel end points are placed in the centers of certain boundary tetrahedrons in such a way that the distance between two end points is at least the *cover radius*. This is achieved by picking the largest boundary tetrahedron and marking it as an exit, then

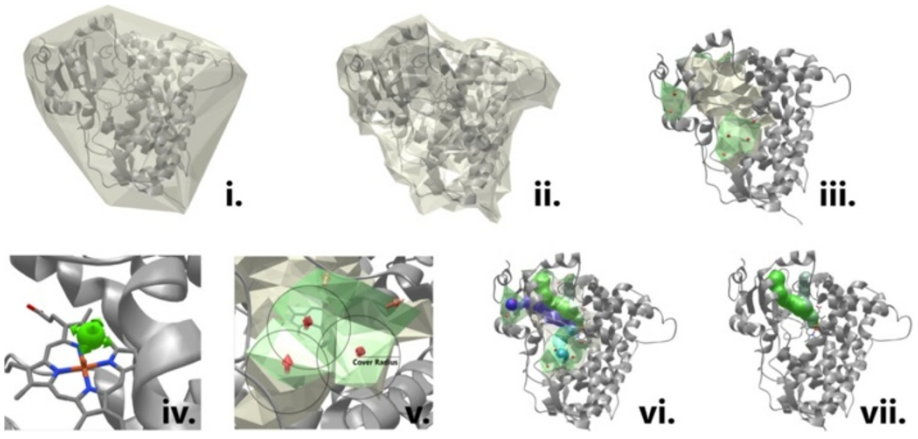


Figure 2 Scheme showing the steps i-vii involved in the channel calculation algorithm (see the text for details). Illustrated for cytochrome P450 3A4 (PDB ID: 1TQN).

removing all boundary tetrahedrons within the *cover radius*. This process is repeated until all non-exit boundary tetrahedrons are removed.

Step vi: computing channels

Once the potential start and end points have been identified, channels are computed as the shortest paths between all pairs of start and end points in the same cavity diagram. To achieve this, Dijkstra's algorithm is used with edge weights given by the following formula:

$$w(e) = \frac{l(e)}{d(e)^2 + \varepsilon}, \quad (1)$$

where *l(e)* is the length of the edge, *d(e)* is the distance of the edge to the closest atom van der Waals sphere and ε is a small number to avoid division by zero [19].

At this stage, each channel is represented by a sequence of tetrahedrons. The next step is to approximate the channel centerline by a natural cubic spline of the circumsphere centers of the tetrahedrons. Additionally, a "radius spline" is computed that determines the centerline distance to the closest atom van der Waals sphere.

Step vii: filtering of channels

The above-described steps usually generate a large number of channels. However, many of these channels are either too narrow (i.e., have a bottleneck with a small radius) to be considered relevant or are duplicated (i.e., too similar to each other). To obtain the most relevant channels, the algorithm contains a filter with two criteria.

The first criterion deals with bottlenecks using parameters that define the maximum *bottleneck length* and minimum *bottleneck radius*. These two parameters ensure that there is enough room for a ligand to pass through each region of the tunnel.

The second criterion is necessary because channels generated using steps (i-vi) of the algorithm often have very similar centerlines that only deviate towards the ends of the channels near the molecular surface. Therefore, for practical purposes, these channels can be considered identical. To remove duplicate channels, a parameter called the *cutoff ratio* is introduced. The centerlines of each pair of tunnels are compared, and if two channels "share" at least the *cutoff ratio* percentage of the centerline, the longer one is removed.

Lining and physicochemical properties of identified channels

The channel lining amino acids residues are the residues that surround the centerline of the channel. The centerline is uniformly divided into layers, and each layer is defined by the residues lining it. A new layer starts whenever there is a change in the list of residues lining the tunnel along its length. The lining of the channel is then described as a sequence of layer lining residues. For each layer, the length (distance between the first and last atom of the layer projected to the tunnel centerline) and radius (bottleneck) are computed. Additionally, the orientation of each residue is determined to check whether the residue faces the tunnel with its backbone or side-chain moiety.

Basic physicochemical properties of protein channels are computed from the set of lining amino acids residues. In MOLE 2.0, the *charge* according to the amino acid side-chain type (Arg, Lys +1*e*; Glu, Asp −1*e*), *hydropathy* [43], *hydrophobicity* [44], *mutability* [45] and *polarity* [46] are computed. The properties are calculated for the unique residues surrounding the channel by averaging tabulated values (Additional file 1: Table S1) for every amino acid residue that has a side chain oriented towards the tunnel. The only exception is charge, which is calculated as the sum of the charges of individual amino acid side chains. For amino acids that have their main chains oriented towards the tunnel, tabulated values for glycine (Gly) are used to compute the hydrophobicity and hydropathy, and the value for asparagine (Asn) is used to evaluate polarity. Amino acids residues that have their main chains lining the channel are not considered when computing mutability. MOLE 2.0 also enables calculation of the weighted physicochemical properties (except the charge) of the channel. The weighted properties are evaluated by applying the above methods separately for each layer and then computing the weighted average, where the weight is given by the length of the layer. We note that the calculated physicochemical properties should be interpreted with care, because the used calculation comes from an assumption that the side chains making the channel wall determine the internal environment of the channel.

Merging channels to pores

The MOLE 2.0 algorithm can compute pores by merging channels. There are three modes for computing pores. The first automatic mode evaluates pores as "channels" between all pairs of end points in a given cavity. In the second mode channels are computed among a set of user-selected end points. Finally, the third mode first computes channels from a user-defined start point and then merges them to form a pore. This mode also imposes a so called "pore criterion" that stipulates that the end points of the pore must be further away than the average length of the channels that formed the pore. In all modes, pores that are too similar are removed using the same criteria as for channels.

Complexity of the algorithm

The worst-case complexity of the algorithm is (N^2 log N), where N is the number of atoms in the molecule. However, in most practical cases, the complexity is $O(M$

log M), where M is the number of vertices in the Voronoi diagram. In the worst case, $M = N^2$. However, as shown by Dwyer *et al.* [47], in most cases $M = O(N)$. Thus, as a result of the use of the incremental algorithm and Hilbert curve ordering, the complexity of calculating the Delaunay triangulation of most molecular structures is $O(N \log N)$. Finally, the complexity of all the remaining steps of the algorithm is at most $O(M \log M)$.

MOLE 2.0 (Figure 3) supports protein files in PDB format. Once the protein is loaded, the GUI provides a full interactive 3D rendering of the protein and the option to tune individual parameters of the channel computation. The GUI displays information about the identified cavities and once channels or pores are computed, a detailed view of them can be displayed that provides information about the channel's profile, lining and physicochemical properties (Figure 4). Information on the channels can be exported in several formats, including XML, CSV, PDB and PyMOL for enhanced visualization.

The command line version of MOLE 2.0 requires the user to specify the input parameters in an XML file. The output can be obtained in XML format as well as a PDB or PyMOL script together with 3D representations of channels that can be loaded to Jmol [48] (http://www.jmol.org). The complete documentation can be found on the web page http://mole.chemi.muni.cz.

Case study: properties of channels of cytochrome P450s BM3 and P450cam

Channels were calculated using MOLE 2.0 with parameters set as follows: minimal bottleneck radius 1.25 Å, probe radius 3 Å, surface cover radius 10 Å and origin radius 5 Å. The heme cofactor was used as the start point in all structures, while all other non-protein ("HETATM") groups were ignored. The PDB database contains a relatively large number of X-ray structures of the two selected cytochrome P450s: 43 structures with 54 chains for P450cam (CAM) and 37 structures with 80 chains for P450BM3 (BM3). All crystal structures were divided into monomers and superimposed using the PyMOL 0.99rc program [40]. The identified channels were sorted into specific families according to the nomenclature of Wade and coworkers [15]: channels were included in a particular family if they had at least one point that trespassed a 4 Å wide cube in space assigned to a specific area for that channel family (i.e., through the B/C loop for channel 2e). Only the shortest channel in each channel family was selected for each protein

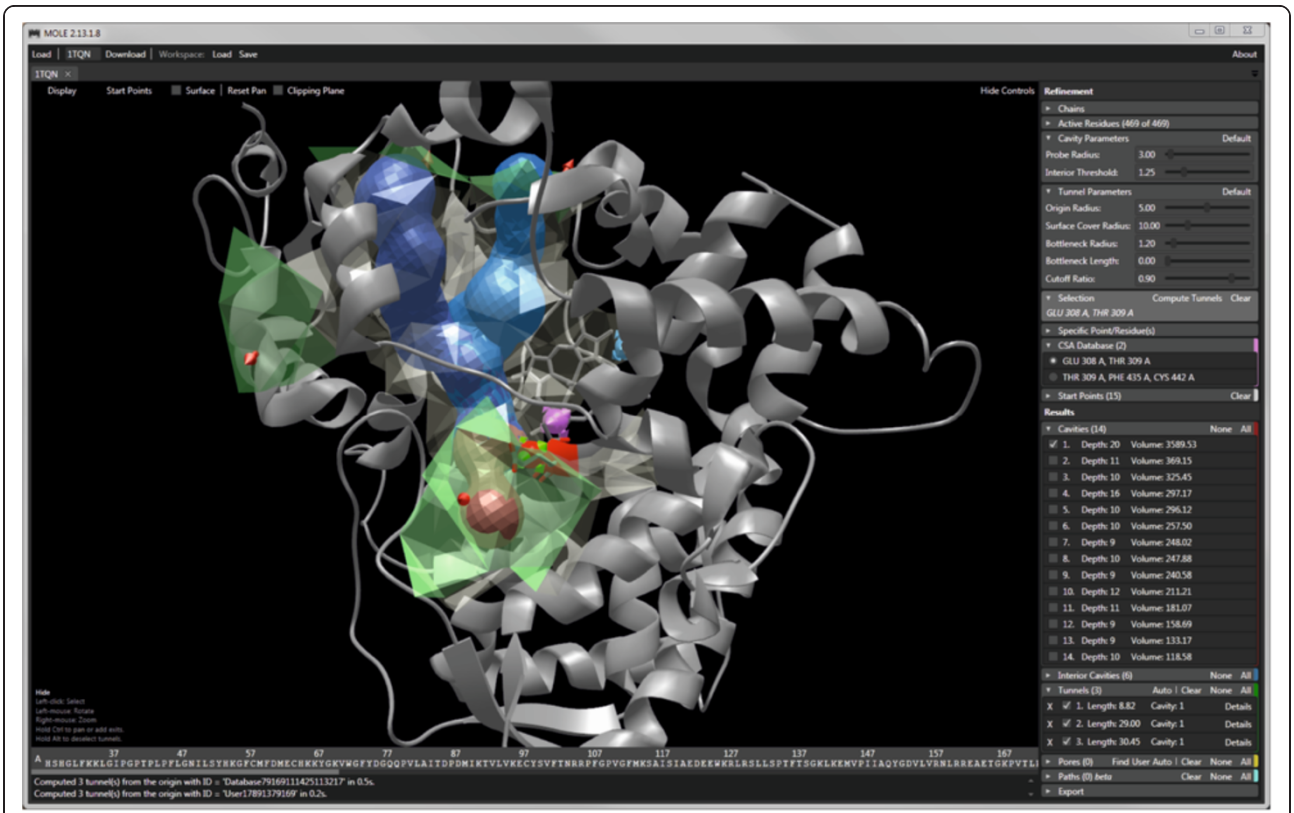


Figure 3 MOLE 2.0 graphical user interface. The left side of the window contains an interactive visualization of the molecule, cavities and computed tunnels. The panel on the right allows the user to tune the computation parameters, select which results are visualized and export them.

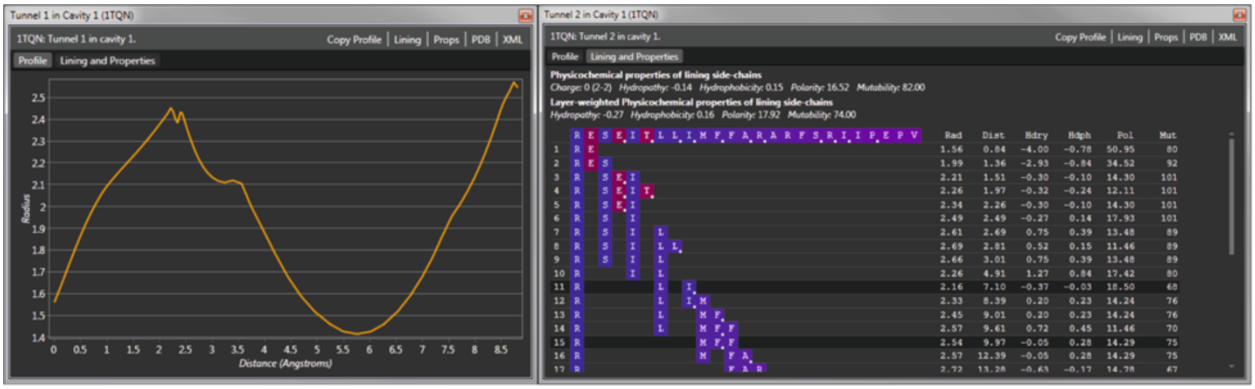


Figure 4 MOLE 2.0 channel details. Channel profile, i.e., plot of radius vs. distance from the start point (left), together with a list of lining amino acid residues and physicochemical properties (right).

structure. Other similar channels were designated as duplicates. The remaining channels were visually checked and meandering channels were also removed. Duplicates were also excluded from the comparison of physicochemical properties.

Results and discussion

Benchmarking study

MOLE 2.0 was compared with four other software tools: MOLE 1.4 [19], MolAxis [36], CAVER 2.0 [38] and CAVER 3.0 [39] (beta version). The main features of the software tools are listed in Table 1. By comparison, MOLE 2.0 provides the richest set of input and output features and has the advantage that both command line and graphical user interfaces are available. The need for a start point is made easier by the fact that MOLE 2.0 enables active sites annotated in the Catalytic Site Atlas (CSA, http://www.ebi.ac.uk/thornton-srv/databases/CSA/) [49] to be used as well as automatic identification of start points in a given structure. Data generated by MOLE 2.0 can be exported to PyMOL [40], which is a popular visualization software, and conveniently, MOLE 2.0 can also be called directly from PyMOL via a plug-in module. In the MOLE 2.0 GUI, a user can select and change the channel end points, which may facilitate the detection of complex channels and pores. The calculation of channels can be customized through nine parameters, whose default values enable automatic identification of channels in many common protein structures. Hence, MOLE 2.0 can be readily used by a new user but provides sufficient flexibility for an advanced user. Besides setup of these parameters, users can adjust the surface of a molecule and filtering of detected channels. It should be noted that MOLE 2.0 is the only software currently available that allows a user to compute cavities and estimate physicochemical properties of identified channels.

The performance of all the considered software tools was compared on a set of thirteen diverse biomacromolecules containing several channels or pores: two RNAs, three

Table 1 Basic features of software tools for channel identification

Features		Software				
		MOLE 2.0	MOLE 1.4	MolAxis	CAVER 2.0	CAVER 3.0
Input and output	Command line interface	Yes	Yes	Yes	Yes	Yes
	GUI	Yes	Web	Web	No	No
	Suggested start points from CSA	Yes	Yes	No	No	No
	Automatic suggestion of start points	Yes	No	Yes	No	No
	Possibility to set end point	Yes	No	No	No	No
	PyMOL export	Yes	Yes	No	Yes	Yes
	PyMOL plugin	Yes	Yes	No	Yes	Yes
Settings of calculation	Number of parameters	9	9	11	8	35
	Adjustable surface of a molecule	Yes	No	Limited	No	Yes
	Channel filtering	Yes	No	Limited	No	Yes
	Cavity computation	Yes	No	No	No	No
	Computation of physicochemical properties	Yes	No	No	No	No

membrane proteins, the photosystem II oxygen evolving center and seven representatives of enzymatic groups, which have all been targeted in research studies dealing with molecular channels (Figure 5). This comparison was carried out on the laptop with CPU Intel Core i5-430 M 2.26 GHz and 4GB RAM, running native Windows 7. For MolAxis, the webserver (http://bioinfo3d.cs.tau.ac.il/MolAxis/) was used. The software tools were used to identify channels with a radius of at least 1.25 Å along most of their length. Because some channels may be "partially closed" by an amino acid side chain, we also considered channels with a radius less than 1.25 Å provided this narrowing was not longer than 3 Å. Such channels may still be biologically active because they allow at least adaptive penetration of a water molecule (radius ~1.4 Å) upon dynamical changes. If two channels shared more than 70% of their length, only the shortest one was reported. This feature eliminated very similar (duplicate) channels. Full details of the setup of all the software tools and post-processing of results are provided in the Additional file 1. We used the same start points for all the software tools (in Additional file 1: Table S2).

Both versions of MOLE (2.0 and 1.4) together with MolAxis were able to process the largest molecular system considered in the benchmarking, i.e., the large ribosomal subunit containing almost 100,000 atoms. Consistently, MOLE 2.0 displayed the shortest processing times for both small and large systems. For small systems, MOLE 2.0 gave similar processing times to

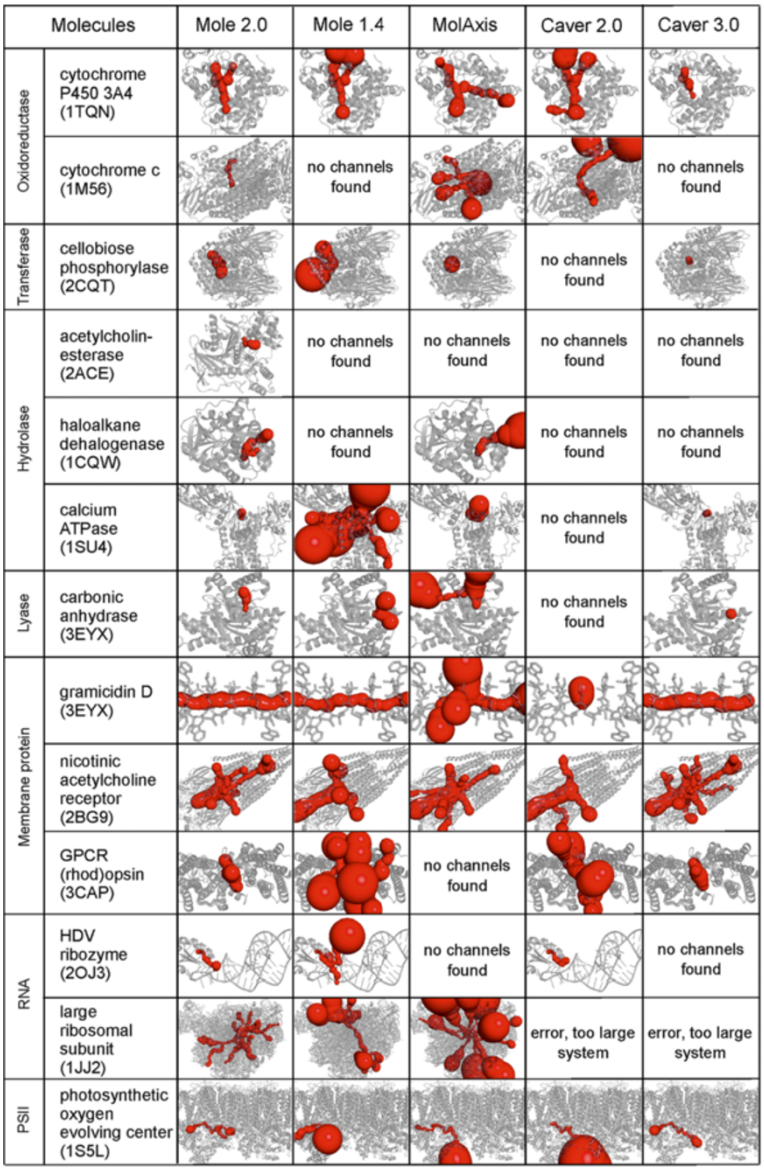


Figure 5 Channels found in the analyzed molecules.

those of MolAxis (one order of magnitude faster than the CAVER tools), whereas for large systems, MOLE 2.0 was one order of magnitude faster than MolAxis and the CAVER tools were not able to calculate the largest system (large ribosomal subunit 1JJ2) (Figure 6 and Additional file 1: Table S3). Such enhancement of processing times may be a considerable advantage if a large number of structures need to be processed (e.g., in analyses of structures from molecular dynamics simulations).

MOLE 2.0 found channels in all the tested molecules, whereas the other software tools did not detect any channels in some cases: MOLE 1.4 and MolAxis in three cases, CAVER 2.0 in six cases and CAVER 3.0 in five cases (Figure 5 and Additional file 1: Table S4). All software tools predicted a rather similar set of channels. The software tools that had end points localized directly on the convex hull (e.g., MOLE 1.4, CAVER 2.0) predicted longer channels with large radii where the probe left the biomacromolecular surface (this behavior could be easily recognized from the "bulky ends" of the identified channels outside the structure). In the case of gramicidin D, which forms a transmembrane pore, MolAxis and CAVER 2.0 predicted a clearly incorrect set of channels, whereas the other tools identified appropriate channels inside the pore. It should be noted that MOLE 2.0 has a new feature of automatic identification of pores in a biomacromolecular structure, which makes it easier to characterize pores and avoids the need for manually merging two (or more) channels into a single pore (a process that cannot be overlooked if one wants to analyze pores with software tools primarily designed for the analysis of channels rather than pores).

For several of the molecules containing biologically important channels/pores with known functionality and properties, we evaluated the physicochemical properties by MOLE 2.0 and related them to the known function of the channel/pore (Figure 7 and Table 2).

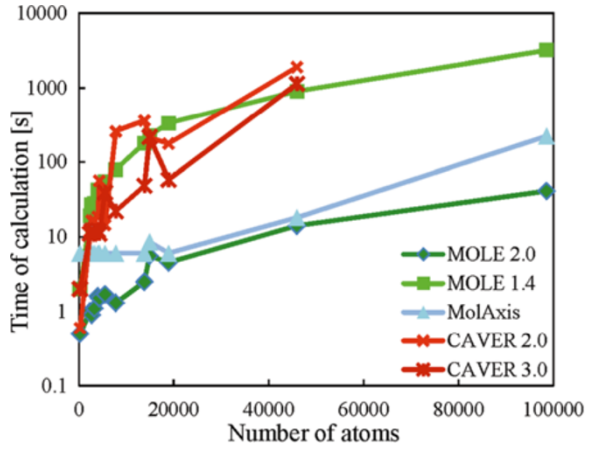


Figure 6 Performance of software tools. Time taken for the channel calculation with respect of the number of atoms in a biomacromolecule (*cf.* Additional file 1: Table S3).

- Gramicidin D (1GRM) is known to form a polar pore in membranes (Figure 7A), [50] which was also reflected in the physicochemical properties identified using MOLE 2.0 as the polar part of the pore surface was predicted to be 100%. However, the predicted polarity of the pore was not high.
- The ribosomal polypeptide (1JJ2) exit channel directs a nascent protein from the proteosynthetic center to the outside of the ribosome [9]. MOLE 2.0 showed that the channel (Figure 7B) is highly polar and lined by amino acids side chains bearing positive charges (7 arginines). In addition, the channel is also lined by 16 RNA backbone phosphate groups. This clearly suggests a fragmental charge along the channels, which is necessary to prevent the nascent peptide from sticking to the channel wall inside the ribosome.
- In the cytochrome c oxidase (1M56), MOLE 2.0 identified two channels with different polarities (Figure 7C), which may be involved in the transfer process required for the proper functioning of this enzyme [51].
- The central pore (Figure 7D) of the nicotinic acetycholine receptor (2BG9) was suggested to be lined by 18 negatively charged amino acids, which explains the experimentally observed selectivity for cation permeation [52].
- The final analyzed channel was present in carbonic anhydrase (3EYX), which can utilize inorganic carbon sources CO_2 and HCO_3^- [53]. MOLE 2.0 predicted that the channel (Figure 7E) is highly polar, in agreement with expectations.

Taken together, the above findings indicate that physicochemical properties may provide useful information about the nature of the channel and its biological function. However, the predicted physicochemical properties may be highly sensitive to the choice of X-ray structure, as discussed later.

Case study: properties of channels in cytochrome P450 BM3 and P450cam

Cytochrome P450s (P450) are heme-containing monoxygenases the active sites of which are deeply buried inside their structures [11,54] and are connected to the exterior by access channels [15]. Hence, channels are considered to play an important role in the metabolism of P450 substrates [12]. Two bacterial cytochrome P450 enzymes - P450cam (CAM, which is also known as CYP101) [55] and P450 BM3 (BM3, which is also known

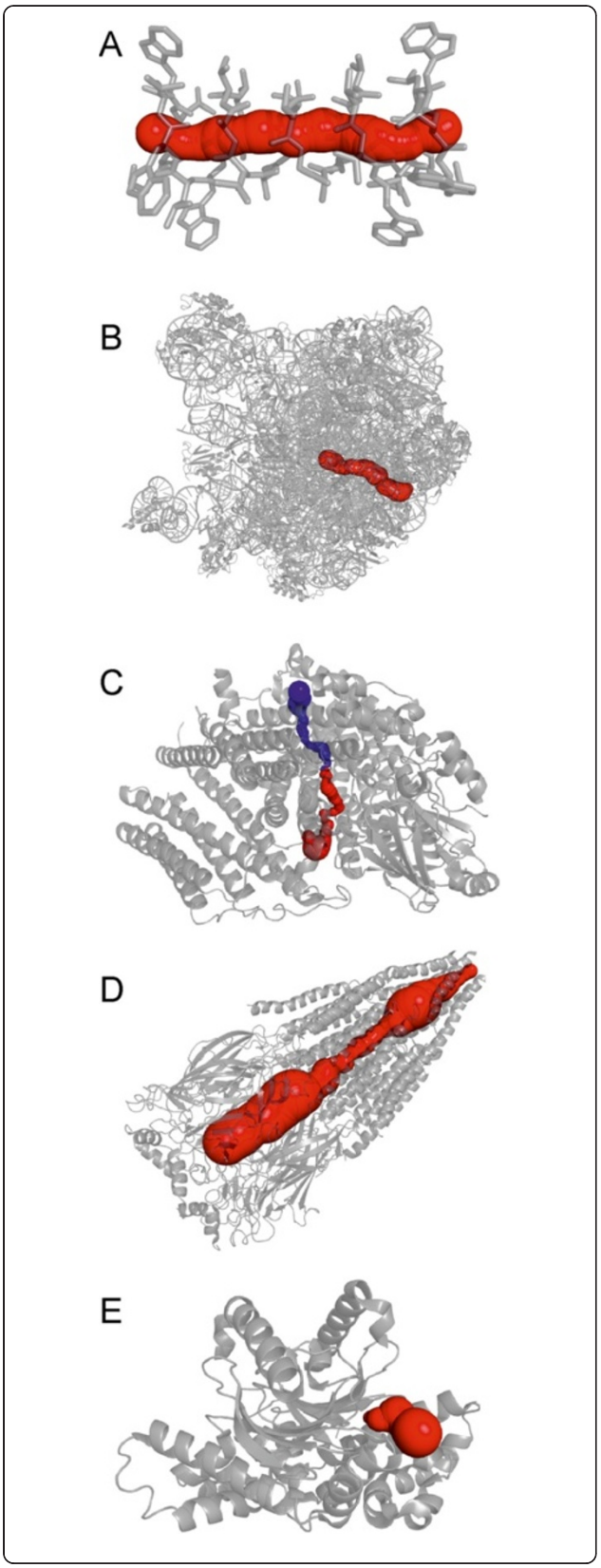


Figure 7 Found channels. A–gramicidin D (1GRM), **B**–large ribosomal subunit (1JJ2), **C**–cytochrome c oxidase (1M56), **D**–nicotinic acetylcholine receptor (2BG9), **E**–carbonic anhydrase (3EYX) by MOLE 2.0. Nonpolar channel in cytochrome c oxidase structure is shown in blue, polar channel is shown in red.

as CYP102) [56]-have been extensively studied by X-ray diffraction in both ligand-free and ligand-bound states; to date, more than 80 structures have been published. Thus, both cytochrome P450s are suitable systems for testing the performance of MOLE 2.0 in predicting the physicochemical properties of channels.

Channel families

More channels were identified in BM3 than in CAM structures. As each independent chain within an asymmetric unit can have different channels [57], it is worthwhile testing all chains within a crystal structure for channel identification. Therefore, we analyzed all 80 chains within the 37 BM3 crystal structures and 54 chains within the 43 CAM crystal structures. It should be noted that CAM can be found in either closed or open states, which differ in the conformation of the F/G loop. Channels were found (using the setup described in the Methods section) only in the open CAM structures (i.e., only in 5 crystal structures: 1K2O, 1PHA, 1QMQ, 1RE9 and 1RF9).

CYP structures contain several different types of active site access channels, which have been classified according to their position in relation to conserved secondary structures in the cytochrome P450 fold by Wade and coworkers [15]. There are two specifically named channels, which are considered to enable the exchange of water molecules between the active site and the enzyme exterior, i.e., the water channel neighboring the B-helix, which is the only channel leading to the CYP proximal side [12], and the solvent channel between the β4 sheet, F and I helices. Other channels are labeled by numerals and only those that are present either in CAM or BM3 structures are noted here. Channels close to the B/C and F/G loops belong to the 2× family–channel 2a is located close to the β1 sheet, F/G and B/B' loops and it has been suggested to be the main access channel of CAM [58,59]; channel 2f neighbors channel 2a and the solvent channel and it is located between the β5 sheet and F/G loop; channel 2b also neighbors channel 2a and is located between the B/C loop, β1 and β3 sheets; channel 2c neighbors channel 2a and is located close to the B/C loop, G and I helices; channel 2ac connects channels 2a and 2c and is located between the B/C and F/G loops; channel 2d is located between the N-terminus and A helix (Figure 8).

Table 2 Physicochemical properties of the studied biologically important channels/pores

PDB	Length (Å)	Hydropathy	Hydrophobicity	Polarity	Charge	Mutability	Polar length	Nonpolar length
1GRM	25.2	−0.4	−0.8	3.38	0 (0–0)	-	100%	0%
1JJ2	79.8	−1.7	−0.6	20.8	4 (6–2)[c]	68	92%	8%
1M56[a]	36.2	3.0	1.0	0.6	0	83	4%	96%
1M56[b]	41.9	1.3	0.8	12.3	0	84	48%	52%
2BG9	143.7	−1.1	−0.2	22.3	−8 (10–18)	85	81%	19%
3EYX	11.5	0.1	0.1	17.0	1 (2–1)	73	100%	0%

[a] the nonpolar channel in Figure 7C (blue), [b] the polar channel in Figure 7C (red), [c] MOLE 2.0 counts the charge on amino acids only, whereas the ribosome channel is also lined by 16 phosphates.

Variability of results

We identified 209 channels along with 73 duplicates within the 80 BM3 chains. Such a large number of channels allowed us to analyze the variability in geometrical or physicochemical properties of the identified channels between individual X-ray structures of a specific protein. The variability was evaluated as the standard deviation calculated for each channel type (W, S, 2a, 2b, 2c, 2ac, 2d, 2f). Then, the total standard deviation of a given property was calculated as a channel-number weighted average of the channels' individual standard deviations. We also calculated the relative variability as the total standard deviation divided by the channel-number weighted mean value of a given property.

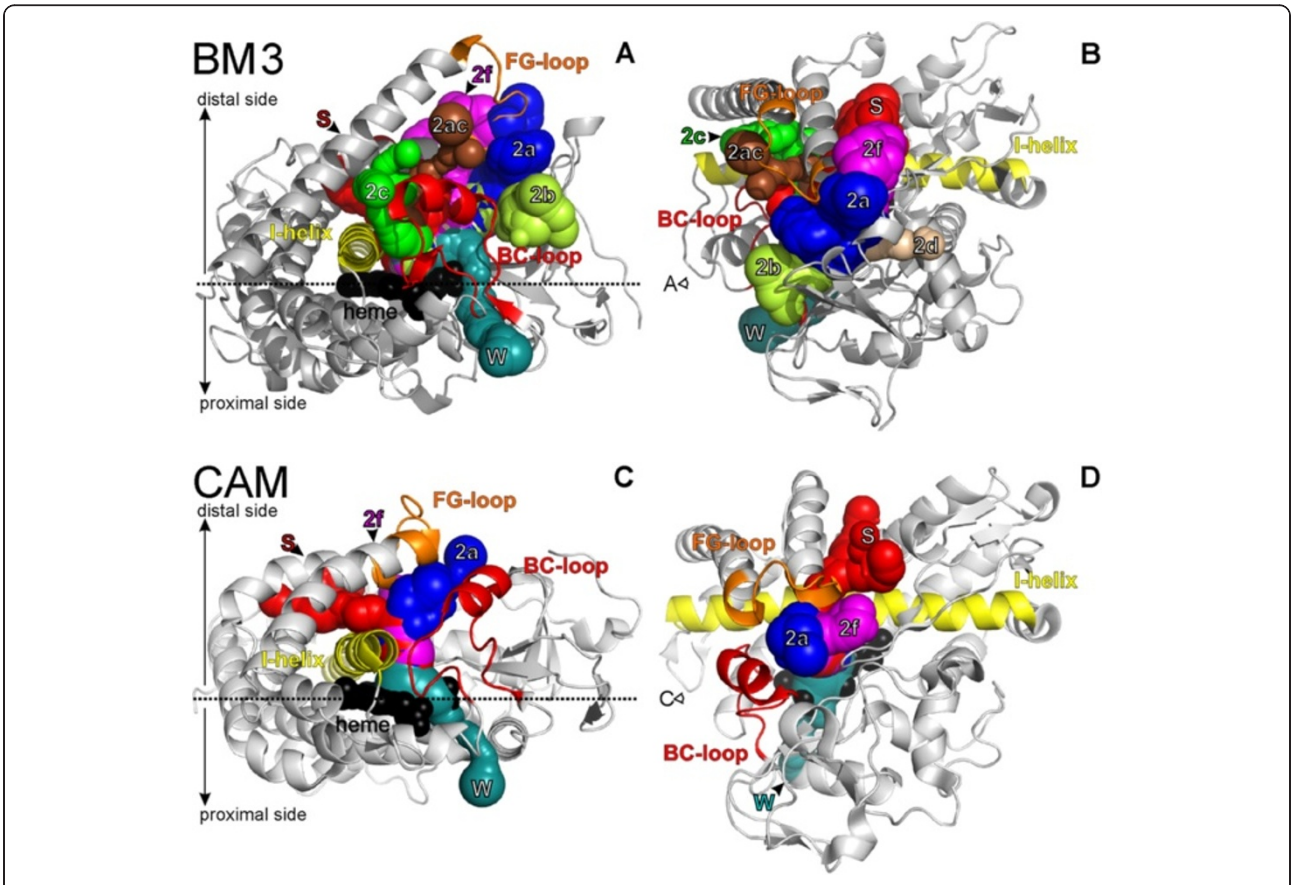


Figure 8 Cytochrome P450 access and egress channels calculated by MOLE 2.0. Channels were imposed on a cartoon representation of structures of cytochrome P450 BM3 (in views **A** and **B**; PDB structure 1BU7 was used) and cytochrome P450 CAM (in views **C** and **D**; PDB 1RE9 was used). Important secondary structures are colored as follows: the I helix is yellow, the F/G-loop is orange, the B/C-loop is red and the heme cofactor is shown as black balls. The images on the left **(A** and **C)** show views from the side in a plane horizontal to the plane of the heme; the images on the right **(B** and **D)** show views from above the distal side. Arrows indicate the viewpoints of the respective images. Channels are shown as connected spheres colored as follows: on the proximal side, channel W is colored in cyan; on the distal side channel S is shown in red; 2a–blue; 2ac–brown; 2b–light green; 2c–green; 2d–pink; 2f–magenta.

- The channel length variation was usually between 10% and 20% of the average channel length, i.e., around 5 Å in the case of BM3.
- The bottleneck radius showed a deviation of about ± 0.23 Å (less than 15%).
- The variability in the distance of bottlenecks from the start point was rather large, i.e., up to 8 Å (53%). This is not surprising because the position of bottlenecks is sensitive to the actual structure of the channel (and conformation of the lining amino acids side chains), i.e., it depends on the choice of X-ray structure [14]. The large variability in the position of bottlenecks has been also identified in molecular dynamics simulations [60]. Based on the large variability of this parameter, we do not recommend that this parameter is viewed as a robust feature of any channel found in only one crystal structure.
- The charge along a channel exhibited a deviation in the order of 0.6 *e* (about 21%).
- The hydropathy index of amino acids ranges between hydrophilic (−4.5) and hydrophobic (4.5). The variation of this value was in the order of 0.5 (less than 9%).
- The hydrophobicity index is a similar measure to the hydropathy index but has a smaller range of values between hydrophilic (−1.14) and hydrophobic amino acids (1.81). It exhibited a lower variation than the hydropathy index of about 0.14. However, its relative error was similar (less than 9%). It also seemed to be more consistent between systems as values for the same types of channels did not differ much between both proteins.
- Polarity values range from 0 for nonpolar amino acids through values of about 2 for polar amino acids towards values around 50 for charged amino acids. Polarity can therefore easily distinguish between polar channels and channels lined with charged amino acids. For instance, the solvent channel in BM3 was predicted to have a similar charge to that of channel 2f (−0.7 vs. -0.4). However, the solvent channel showed a significantly higher polarity index (9.4 vs. 2.0 for channel 2f). This indicates that the solvent channel is lined with more highly charged residues that cancel each other out, whereas channel 2f is mostly lined with nonpolar and polar residues. The variation of the polarity was in the order of ± 2.5. The relative error was about 47%. However, this value should be interpreted with care owing to the low polarity of the analyzed channels (the channel number weighted mean value was only 6.4 out of a possible range of 0–50).
- Mutability values range from the lowest mutability of 44 for Cys to a value of 177 for the most easily interchangeable Ser. The variation of mutability was in the order of ±3 and the relative error was the lowest of all the indices mentioned (less than 4%).

The results showed that the geometrical properties and physicochemical properties of the found channels typically varied by less than 20% except for the distance of bottlenecks from the starting point.

Properties of CAM and BM3 channels

From a geometrical perspective, the most open channels were usually found within the open CAM structures, particularly 2a channels, which have a bottleneck radius larger than 2.6 Å. Channels belonging to the 2× family (mainly channels 2a, 2f, and in the case of BM3, channel 2b) were predicted to have bottleneck radii large enough to allow substrates/products to pass (> 2 Å) in both the CAM and BM3 structures, i.e., comparable or even larger than the solvent channel bottleneck radius (> 1.4 Å, radius of water molecule). The most closed channel was the water channel. However this does not necessarily mean that small molecules cannot pass through it as it might partially open to allow molecules to enter due to bottleneck fluctuations, as shown previously for the 2b channel within the structure of mammalian cytochrome P450 2A6 [14]. It is also worth noting that the solvent channel was predicted to be ~7 Å longer in CAM than in BM3, whereas other channels were typically longer in BM3. In contrast, the most open channels 2a and 2f in CAM were ~12 Å shorter than in BM3. However, this was partly because we used a probe radius of 3 Å to construct the overall shape of the protein, and therefore we only detected channels below this radius.

The water and solvent channels were clearly the most hydrophilic. The hydrophilicity also appeared to correlate with the polarity of the channels because the water and solvent channels were also predicted to be the most polar channels. The higher polarity index indicates that polar and charged amino acid residues line the solvent and water channels. On the other hand, the mutability index did not differ significantly between the individual channels. The mutability was also relatively high, which may indicate that the channels are lined with amino acids that can be relatively easily interchanged. This finding is in accord with the relatively low sequence homology between individual members of CYP family [60].

Ranking the channels according to their average hydrophobicity supported the hypothesis that the water and solvent channels are involved in water transfer into the active site [61], as the water channel was the most hydrophilic channel in both the CAM and BM3 structures, followed by the solvent channel (according to the hydropathy and hydrophobicity indices). BM3 was also predicted to contain the rather polar channel 2b. The

more hydrophobic channels 2f and 2a were present in both the CAM and BM3 structures. Channels 2ac and 2d were more hydrophobic still. Finally, the most hydrophobic channel was channel 2c. However, the last three channels were found rather infrequently, i.e., only present in some BM3 structures (Additional file 1: Tables S5 and S6).

Conclusions

We present the advanced software tool MOLE 2.0 designed to analyze molecular channels and pores. We benchmarked MOLE 2.0 against similar software tools and showed that by comparison it is faster and capable of analyzing large and complex systems containing up to hundreds of thousands of atoms. As a new feature, MOLE 2.0 estimates physicochemical properties of the identified channels. We compared the estimated physicochemical properties with the known functions of selected biomacromolecular channels and concluded that the properties correlated with the functions. We also assessed the variability of physicochemical properties by analyzing a large number of X-ray structures of two members of the cytochrome P450 superfamily. We propose that the physicochemical properties may provide useful clues about the potential functions of identified channels. The software is available free of charge at http://mole.chemi.muni.cz.

Availability and requirements

Project name: MOLE 2.0
Project home page: http://mole.chemi.muni.cz
Operating systems: Mac OS, Linux, Windows
Programming language: C#
Other requirements: NET 4.0 for Windows based systems, Mono framework 2.10. or newer (http://www.mono-project.com) for other OS.
License: MOLE 2.0 license
Restrictions: free of charge

Additional file

Additional file 1: Table S1. Physicochemical properties of amino acids residues, setup of all software tools used for the benchmarking study. **Table S2.** Channel starting points used in the benchmarking study. **Table S3.** Duration of channel calculations for all biomacromolecules used in the benchmarking study. **Table S4.** Numbers of channels found in the analyzed molecules in the benchmarking study. **Table S5.** Comparison of geometrical and physicochemical properties of channels detected in CAM structures. **Table S6.** Comparison of geometrical and physicochemical properties of channels detected in BM3 structures.

Abbreviations

BM3 Cytochrome P450 BM3; CAM Cytochrome P450cam; all amino acids are represented by their respective three-letter abbreviations.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

All authors contributed extensively to the work presented in this paper. DS wrote MOLE 2.0 application. All authors read and approved the final manuscript.

Acknowledgement

This work was supported by the Czech Science Foundation (GD301/09/H004 to CMI, P208/12/G016 to MO, P303/12/P019 to KB) and the Operational Program Research and Development for Innovations–European Regional Development Fund (CZ.1.05/2.1.00/03.0058 to MO, KB, PB and CZ.1.05/1.1.00/02.0068 to JK and RSV), European Social Fund (CZ.1.07/2.3.00/20.0017 to KB, PB) and a student project of Palacký University (PrF_2013_028 to VN). C.M.I. and D.S. thank Brno City Municipality for financial support provided through the program Brno Ph.D. Talent. Access to the MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is gratefully acknowledged.

Author details

[1]National Centre for Biomolecular Research, Faculty of Science and CEITEC-Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic. [2]Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic. [3]Department of Physical Chemistry, Regional Centre of Advanced Technologies and Materials, Faculty of Science, Palacký University Olomouc, tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic.

Received: 4 April 2013 Accepted: 13 August 2013
Published: 16 August 2013

References

1. Matthews BW, Liu L: **A review about nothing: are apolar cavities in proteins really empty?** *Protein Sci* 2009, **18:**494–502.
2. Walz T, Smith BL, Agre P, Engel A: **The three-dimensional structure of human erythrocyte aquaporin CHIP.** *EMBO J* 1994, **13:**2985–2993.
3. Jiang Y, Lee A, Chen J, Cadene M, Chait BT, MacKinnon R: **Crystal structure and mechanism of a calcium-gated potassium channel.** *Nature* 2002, **417:**515–522.
4. Doyle DA, Morais Cabral J, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R: **The structure of the potassium channel: molecular basis of K + conduction and selectivity.** *Science* 1998, **280:**69–77.
5. Alexander SPH, Mathie A, Peters JA: **Guide to Receptors and Channels (GRAC), 5th edition.** *Br J Pharmacol* 2011, **164**(Suppl):S1–S324.
6. MacKinnon R: **Potassium channels and the atomic basis of selective ion conduction (Nobel Lecture).** *Angewandte Chemie (International ed. in English)* 2004, **43:**4265–4277.
7. Murray JW, Barber J: **Structural characteristics of channels and pathways in photosystem II including the identification of an oxygen channel.** *J Struct Biol* 2007, **159:**228–237.
8. Guskov A, Kern J, Gabdulkhakov A, Broser M, Zouni A, Saenger W: **Cyanobacterial photosystem II at 2.9-A resolution and the role of quinones, lipids, channels and chloride.** *Nat Struct Mol Biol* 2009, **16:**334–342.
9. Voss NR, Gerstein M, Steitz TA, Moore PB: **The geometry of the ribosomal polypeptide exit tunnel.** *J Mol Biol* 2006, **360:**893–906.
10. Wade RC, Winn PJ, Schlichting I, Sudarko: **A survey of active site access channels in cytochromes P450.** *J Inorg Biochem* 2004, **98:**1175–1182.
11. Otyepka M, Skopalík J, Anzenbacherová E, Anzenbacher P: **What common structural features and variations of mammalian P450s are known to date?** *Biochim Biophys Acta* 2007, **1770:**376–389.
12. Otyepka M, Berka K, Anzenbacher P: **Is there a relationship between the substrate preferences and structural flexibility of cytochromes P450?** *Curr Drug Metab* 2012, **13:**130–142.
13. Berka K, Hendrychová T, Anzenbacher P, Otyepka M: **Membrane position of ibuprofen agrees with suggested access path entrance to cytochrome P450 2C9 active site.** *J Phys Chem A* 2011, **115:**11248–11255.
14. Hendrychova T, Berka K, Navratilova V, Anzenbacher P, Otyepka M: **Dynamics and hydration of the active sites of mammalian cytochromes P450 probed by molecular dynamics simulations.** *Curr Drug Metab* 2012, **13:**177–189.
15. Cojocaru V, Winn PJ, Wade RC: **The ins and outs of cytochrome P450s.** *Biochim Biophys Acta* 2007, **1770:**390–401.

16. Gilson MK, Straatsma TP, McCammon JA, Ripoll DR, Faerman CH, Axelsen PH, Silman I, Sussman JL: **Open "back door" in a molecular dynamics simulation of acetylcholinesterase.** *Science* 1994, **263:**1276–1278.
17. Wiesner J, Kriz Z, Kuca K, Jun D, Koca J: **Acetylcholinesterases–the structural similarities and differences.** *J Enzyme Inhib Med Chem* 2007, **22:**417–424.
18. Sanson B, Colletier J-P, Xu Y, Lang PT, Jiang H, Silman I, Sussman JL, Weik M: **Backdoor opening mechanism in acetylcholinesterase based on X-ray crystallography and molecular dynamics simulations.** *Protein Sci* 2011, **20:**1114–1118.
19. Petrek M, Kosinová P, Koca J, Otyepka M: **MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels.** *Structure* 2007, **15:**1357–1363.
20. Pavlova M, Klvana M, Prokop Z, Chaloupkova R, Banas P, Otyepka M, Wade RC, Tsuda M, Nagata Y, Damborsky J: **Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate.** *Nat Chem Biol* 2009, **5:**727–733.
21. Biedermannová L, Prokop Z, Gora A, Chovancová E, Kovács M, Damborsky J, Wade RC: **A single mutation in a tunnel to the active site changes the mechanism and kinetics of product release in haloalkane dehalogenase LinB.** *J Biol Chem* 2012, **287:**29062–29074.
22. Brezovsky J, Chovancova E, Gora A, Pavelka A, Biedermannova L, Damborsky J: **Software tools for identification, visualization and analysis of protein tunnels and channels.** *Biotechnol Adv* 2012, **31:**38–49.
23. Lee P-H, Helms V: **Identifying continuous pores in protein structures with PROPORES by computational repositioning of gating residues.** *Proteins* 2011, **80:**421–432.
24. Levitt DG, Banaszak LJ: **POCKET: a computer graphies method for identifying and displaying protein cavities and their surrounding amino acids.** *J Mol Graph* 1992, **10:**229–234.
25. Hendlich M, Rippmann F, Barnickel G: **LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins.** *J Mol Graphics Model* 1997, **15:**359–363.
26. Huang B, Schroeder M: **LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation.** *BMC Struct Biol* 2006, **6:**19.
27. Raunest M, Kandt C: **dxTuber: detecting protein cavities, tunnels and clefts based on protein and solvent dynamics.** *J Mol Graph Model* 2011, **29:**895–905.
28. Ho BK, Gruswitz F: **HOLLOW: generating accurate representations of channel and interior surfaces in molecular structures.** *BMC Struct Biol* 2008, **8:**49.
29. Voss NR, Gerstein M: **3V: cavity, channel and cleft volume calculator and extractor.** *Nucleic Acids Res* 2010, **38:**W555–W562.
30. Petrek M, Otyepka M, Banás P, Kosinová P, Koca J, Damborský J: **CAVER: a new tool to explore routes from protein clefts, pockets and cavities.** *BMC Bioinformatics* 2006, **7:**316.
31. Coleman RG, Sharp KA: **Finding and characterizing tunnels in macromolecules with application to ion channels and pores.** *Biophys J* 2009, **96:**632–645.
32. Brady GP, Stouten PFW, Brady GP Jr: **Fast prediction and visualization of protein binding pockets with PASS.** *J Comput Aided Mol Des* 2000, **14:**383–401.
33. Laskowski RA: **SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions.** *J Mol Graph* 1995, **13:**323–330.
34. Smart OS, Neduvelil JG, Wang X, Wallace BAA, Sansom MSP: **HOLE: a program for the analysis of the pore dimensions of ion channel structural models.** *J Mol Graph* 1996, **14:**354–360.
35. Pellegrini-Calace M, Maiwald T, Thornton JM: **PoreWalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure.** *PLoS Comput Biol* 2009, **5:**e1000440.
36. Yaffe E, Fishelovitch D, Wolfson HJ, Halperin D, Nussinov R: **MolAxis: efficient and accurate identification of channels in macromolecules.** *Proteins* 2008, **73:**72–86.
37. Yaffe E, Fishelovitch D, Wolfson HJ, Halperin D, Nussinov R: **MolAxis: a server for identification of channels in macromolecules.** *Nucleic Acids Res* 2008, **36:**W210–W215.
38. Medek P, Benes P, Sochor J: **Multicriteria tunnel computation. CGIM '08.** In *Proceedings of the Tenth IASTED International Conference on Computer Graphics and Imaging; Innsbruck, Austria.* 2008:57–61.
39. Chovancova E, Pavelka A, Benes P, Strnad O, Brezovsky J, Kozlikova B, Gora A, Sustr V, Klvana M, Medek P, Biedermannova L, Sochor J, Damborsky J: **CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures.** *PLoS Comput Biol* 2012, **8:**e1002708.
40. Schrödinger L: *The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC.* 2010 (see PyMOL page: http://www.pymol.org/citing).
41. Berka K, Hanák O, Sehnal D, Banáš P, Navrátilová V, Jaiswal D, Ionescu C-M, Svobodová Vareková R, Koca J, Otyepka M: **MOLEonline 2.0: interactive web-based analysis of biomacromolecular channels.** *Nucleic Acids Res* 2012, **40:**W222–W227.
42. Liu Y, Snoeyink J: **A comparison of five implementations of 3D Delaunay tessellation in combinatorial and computational geometry.** *Combinatorial Computational Geometry* 2005, **52:**439–458.
43. Kyte J, Doolittle RF: **A simple method for displaying the hydropathic character of a protein.** *J Mol Biol* 1982, **157:**105–132.
44. Cid H, Bunster M, Canales M, Gazitúa F: **Hydrophobicity and structural classes in proteins.** *Protein Eng Design Selection* 1992, **5:**373–375.
45. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Bioinformatics* 1992, **8:**275–282.
46. Zimmerman JM, Eliezer N, Simha R: **The characterization of amino acid sequences in proteins by statistical methods.** *J Theor Biol* 1968, **21:**170–201.
47. Dwyer RA: **Higher-dimensional voronoi diagrams in linear expected time.** *Discrete Comput Geom* 1991, **6:**343–367.
48. Herráez A: **Biomolecules in the computer: Jmol to the rescue.** *Bioch Mol Biol Educ* 2006, **34:**255–261.
49. Porter CT, Bartlett GJ, Thornton JM: **The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data.** *Nucleic Acids Res* 2004, **32:**D129–D133.
50. Andersen OS, Koeppe RE, Roux B: **Gramicidin channels.** *IEEE Trans Nanobioscience* 2005, **4:**10–20.
51. Brzezinski P, Gennis RB: **Cytochrome c oxidase: exciting progress and remaining mysteries.** *J Bioenerg Biomembr* 2008, **40:**521–531.
52. Unwin N: **Refined structure of the nicotinic acetylcholine receptor at 4A resolution.** *J Mol Biol* 2005, **346:**967–989.
53. Teng Y-B, Jiang Y-L, He Y-X, He W-W, Lian F-M, Chen Y, Zhou C-Z: **Structural insights into the substrate tunnel of Saccharomyces cerevisiae carbonic anhydrase Nce103.** *BMC Struct Biol* 2009, **9:**67.
54. Pochapsky TC, Kazanis S, Dang M: **Conformational plasticity and structure/function relationships in cytochromes P450.** *Antioxid Redox Signal* 2010, **13:**1273–1296.
55. Poulos TL, Finzel BC, Howard AJ: **High-resolution crystal structure of cytochrome P450cam.** *J Mol Biol* 1987, **195:**687–700.
56. Ravichandran K, Boddupalli S, Hasermann C, Peterson J, Deisenhofer J: **Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal P450's.** *Science* 1993, **261:**731–736.
57. DeVore NM, Scott EE: **Nicotine and 4-(methylnitrosamino)-1-(3-pyridyl)-1-butanone binding and access channel in human cytochrome P450 2A6 and 2A13 enzymes.** *J Biol Chem* 2012, **287:**26576–26585.
58. Ludemann SK, Lounnas V, Wade RC: **How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms.** *J Mol Biol* 2000, **303:**797–811.
59. Lüdemann SK, Lounnas V, Wade RC: **How do substrates enter and products exit the buried active site of cytochrome P450cam? 2. Steered molecular dynamics and adiabatic mapping of substrate pathways.** *J Mol Biol* 2000, **303:**813–830.
60. Nebert DW, Nelson DR, Coon MJ, Estabrook RW, Feyereisen R, Fujii-Kuriyama Y, Gonzalez FJ, Guengerich FP, Gunsalus IC, Johnson EF: **The P450 superfamily: update on new sequences, gene mapping, and recommended nomenclature.** *DNA Cell Biol* 1991, **10:**1–14.
61. Skopalík J, Anzenbacher P, Otyepka M: **Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences.** *J Phys Chem B* 2008, **112:**8165–8173.

doi:10.1186/1758-2946-5-39
Cite this article as: Sehnal *et al.*: MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *Journal of Cheminformatics* 2013 **5**:39.

Predicting pK_{a} values from EEM atomic charges

Journal of Cheminformatics

RESEARCH ARTICLE Open Access

Predicting pK_a values from EEM atomic charges

Radka Svobodová Vařeková[1], Stanislav Geidl[1], Crina-Maria Ionescu[1], Ondřej Skřehota[1], Tomáš Bouchal[1], David Sehnal[1], Ruben Abagyan[2] and Jaroslav Koča[1*]

Abstract

The acid dissociation constant pK_a is a very important molecular property, and there is a strong interest in the development of reliable and fast methods for pK_a prediction. We have evaluated the pK_a prediction capabilities of QSPR models based on empirical atomic charges calculated by the Electronegativity Equalization Method (EEM). Specifically, we collected 18 EEM parameter sets created for 8 different quantum mechanical (QM) charge calculation schemes. Afterwards, we prepared a training set of 74 substituted phenols. Additionally, for each molecule we generated its dissociated form by removing the phenolic hydrogen. For all the molecules in the training set, we then calculated EEM charges using the 18 parameter sets, and the QM charges using the 8 above mentioned charge calculation schemes. For each type of QM and EEM charges, we created one QSPR model employing charges from the non-dissociated molecules (three descriptor QSPR models), and one QSPR model based on charges from both dissociated and non-dissociated molecules (QSPR models with five descriptors). Afterwards, we calculated the quality criteria and evaluated all the QSPR models obtained. We found that QSPR models employing the EEM charges proved as a good approach for the prediction of pK_a (63% of these models had $R^2 > 0.9$, while the best had $R^2 = 0.924$). As expected, QM QSPR models provided more accurate pK_a predictions than the EEM QSPR models but the differences were not significant. Furthermore, a big advantage of the EEM QSPR models is that their descriptors (i.e., EEM atomic charges) can be calculated markedly faster than the QM charge descriptors. Moreover, we found that the EEM QSPR models are not so strongly influenced by the selection of the charge calculation approach as the QM QSPR models. The robustness of the EEM QSPR models was subsequently confirmed by cross-validation. The applicability of EEM QSPR models for other chemical classes was illustrated by a case study focused on carboxylic acids. In summary, EEM QSPR models constitute a fast and accurate pK_a prediction approach that can be used in virtual screening.

Keywords: Dissociation constant, Quantitative structure-property relationship, QSPR, Partial atomic charges, Electronegativity equalization method, EEM, Quantum mechanics, QM

Background

The acid dissociation constant pK_a is an important molecular property, and its values are of interest in pharmaceutical, chemical, biological and environmental research. The pK_a values have found application in many areas, such as the evaluation and optimization of candidate drug molecules [1-3], ADME profiling [4,5], pharmacokinetics [6], understanding of protein-ligand interactions [7,8], etc.. Moreover, the key physicochemical properties like lipophilicity, solubility, and permeability are all pK_a dependent. For these reasons, pK_a values are important for virtual screening. Therefore, both the research community and pharmaceutical companies are interested in the development of reliable and above all fast methods for pK_a prediction.

Several approaches for pK_a prediction have been developed [8-11], namely LFER (Linear Free Energy Relationships) methods [12,13], database methods, decision tree methods [14], ab initio quantum mechanical calculations [15,16], ANN (artificial neural networks) methods [17] or QSPR (quantitative structure-property relationship) modelling [18-20]. However, pK_a values remain one of the most challenging physicochemical properties to predict.

*Correspondence: jkoca@chemi.muni.cz
[1] National Centre for Biomolecular Research, Faculty of Science and CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic
Full list of author information is available at the end of the article

A promising approach for pK_a prediction is to use QSPR models which employ partial atomic charges as descriptors [21-24].

The partial atomic charges cannot be determined experimentally and they are also not quantum mechanical observables. For this reason, the rules for determining partial atomic charges depend on their application (e.g. molecular mechanics energy, pK_a etc.), and many different methods have been developed for their calculation. The charge calculation methods can be divided into two main groups, namely quantum mechanical (QM) approaches and empirical approaches.

The quantum mechanical approaches first calculate a molecular wave function by a combination of some theory level (e.g., HF, B3LYP, MP2) and basis set (e.g., STO-3G, 6–31G*), and then partition this wave function among the atoms (i.e., the assignment of a specific part of the molecular electron density to each atom). This partitioning can be done using an orbital-based population analysis, such as MPA (Mulliken population analysis) [25,26], Löwdin population analysis [27] or NPA (natural population analysis) [28]. Other partitioning approaches are based on a wavefunction-dependent physical observable. Such approaches are, for example, AIM (atoms in molecules) [29], Hirshfeld population analysis [30] and electrostatic potential fitting methods like CHELPG [31] or MK (Merz-Singh-Kollman) [32]. Another partitioning method is the mapping of QM atomic charges to reproduce charge-dependent observables (e.g., CM1, CM2, CM3 and CM4) [33].

Empirical approaches determine partial atomic charges without calculating a quantum mechanical wave function for the given molecule. Therefore they are markedly faster than QM approaches. One of the first empirical approaches developed, CHARGE [34], performs a breakdown of the charge transmission by polar atoms into one-bond, two-bond, and three-bond additive contributions. Most of the other empirical approaches have been derived on the basis of the electronegativity equalization principle. One group of these empirical approaches invoke the Laplacian matrix formalism, and result in a redistribution of electronegativity. Such methods are PEOE (partial equalization of orbital electronegativity) [35], GDAC (geometry-dependent atomic charge) [36], KCM (Kirchhoff charge model) [37], DENR (dynamic electronegativity relaxation) [38] or TSEF (topologically symmetric energy function) [38]. The second group of approaches use full equalization of orbital electronegativity, and such approaches are, for example, EEM (electronegativity equalization method) [39], QEq (charge equilibration) [40] or SQE (split charge equilibration) [41]. The empirical atomic charge calculation approaches can also be divided into 'topological' and 'geometrical'. Topological charges are calculated using the 2D structure of the molecule, and they are conformationally independent (i.e., CHARGE, PEOE, KCM, DENR, and TSEF). Geometrical charges are computed from the 3D structure of the molecule and they consider the influence of conformation (i.e., GDAC, EEM, Qeq, and SQE).

The prediction of pK_a using QSPR models which employ QM atomic charges was described in several studies [21-24], which have analyzed the precision of this approach and compared the quality of QSPR models based on different QM charge calculation schemes. All these studies show that QM charges are successful descriptors for pK_a prediction, as the QSPR models based on QM atomic charges are able to calculate pK_a with high accuracy. The weak point of QM charges is that their calculation is very slow, as the computational complexity is at least $\theta(E^4)$, where E is the number of electrons in the molecule. Therefore, pK_a prediction by QSPR models based on QM charges cannot be applied in virtual screening, as it is not feasible to compute QM atomic charges for hundreds of thousands of compounds in a reasonable time. This issue can be avoided if empirical charges are used instead of QM charges. A few studies were published, which give QSPR models for predicting pK_a using topological empirical charges as descriptors (specifically PEOE charges) [22,42,43]. But these models provided relatively weak predictions.

The geometrical charges seem to be more promissing descriptors, because they are able to take into consideration the influence of the molecule's conformation on the atomic charges. The conformation of the atoms surrounding the dissociating hydrogens strongly influences the dissociation process, and also the atomic charges.

The EEM method is a geometrical empirical charge calculation approach which can be useful for pK_a prediction by QSPR. This approach calculates charges using the following equation system:

$$\begin{pmatrix} B_1 & \frac{\kappa}{R_{1,2}} & \dots & \frac{\kappa}{R_{1,N}} & -1 \\ \frac{\kappa}{R_{2,1}} & B_2 & \dots & \frac{\kappa}{R_{2,N}} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\kappa}{R_{N,1}} & \frac{\kappa}{R_{N,2}} & \dots & B_N & -1 \\ 1 & 1 & \dots & 1 & 0 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ \overline{\chi} \end{pmatrix} = \begin{pmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_N \\ Q \end{pmatrix} \quad (1)$$

where q_i is the charge of atom i; $R_{i,j}$ is the distance between atoms i and j; Q is the total charge of the molecule; N is the number of atoms in the molecule; $\overline{\chi}$ is the molecular electronegativity, and A_i, B_i and κ are empirical parameters. These parameters are obtained by a parameterization process, which uses QM atomic charges to calculate a set of parameters for which EEM best reproduces these QM charges. EEM is very popular, and despite the fact that it was developed more than twenty years ago, new

parameterizations [39,44-50] and modifications [47,51,52] of EEM are still under development. Its accuracy is comparable to the QM charge calculation approach for which it was parameterized. Additionally, EEM is very fast, as its computational complexity is $\theta(N^3)$, where N is the number of atoms in the molecule.

Therefore, in the present study, we focus on pK_a prediction using QSPR models which employ EEM charges. Specifically, we created and evaluated QSPR models based on EEM charges computed using 18 EEM parameter sets. We also compared these QSPR models with corresponding QSPR models which employ QM charges computed by the same charge calculation schemes used for EEM parameterization.

Methods

EEM parameter sets

In our study, we used all EEM parameters published till now. Specifically, we found 18 different EEM parameters sets, published in 8 different articles [39,44-50]. The parameters cover two QM theory levels (HF and B3LYP), two basis sets (STO-3G and 6–31G*) and six population analyses (MPA, NPA, Hirshfeld, MK, CHELPG, AIM). Unfortunately, only some combinations of QM theory levels, basis sets and population analyses are available. On the other hand, more parameter sets were published for some combinations (i.e., 6 parameter sets for HF/STO-3G/MPA). All the parameter sets include parameters for C, O, N and H. Some sets include also parameters for S, P, halogens and metals. Most of the sets do not include parameters for C and N bonded by triple bond. Summary information about all these parameter sets is given in Table 1.

EEM charge calculation

The EEM charges were calculated by the program EEM SOLVER [53] using each of the 18 EEM parameter sets.

QM charge calculation

We calculated QM atomic charges for all the combinations of QM theory level, basis set and population analysis for which we have EEM parameters (see Table 1). Specifically, atomic charges were calculated for these eight QM approaches: HF/STO-3G/MPA, HF/6–31G*/MK, and B3LYP/6–31G* with MPA, NPA, Hirshfeld, MK, CHELPG and AIM). The QM charge calculations were carried out using Gaussian09 [54]. In the case of AIM population analysis, the output from Gaussian09 was further processed by the software package AIMAll [55].

Data set for phenols

There are two main ways to create a QSPR model for a feature to be predicted. The first is to create as general a model as possible, with the risk that the accuracy of such a model may not be high. The second approach is to develop more models, each of them being dedicated to a certain class of compounds. Here we took the second approach, following a similar methodology as in previous studies [21-24]. Specifically, we focus on substituted phenols, because they are the most common test set molecules employed in the evaluation of novel pK_a prediction approaches [21-24,56-58]. Our data set contains the 3D structures of 74 distinct phenol molecules. This data set is of high structural diversity and it covers molecules with pK_a values from 0.38 to 11.1. The molecules were obtained from the NCI Open Database Compounds [59] and their 3D structures were generated by CORINA 2.6 [60], without any further geometry optimization. Our data set is a subset of the phenol data set used in our previous work related to pK_a prediction from QM atomic charges [24]. The subset is made up of phenols which contain only C, O, N and H, and none of the molecules contain triple bonds. This limitation is necessary, because the EEM parameters of all 18 studied EEM parameter sets are available only for such molecules (see Table 1). For each phenol molecule from our data set, we also prepared the structure of the dissociated form, where the hydrogen is missing from the phenolic OH group. This dissociated molecule was created by removing the hydrogen from the original structure without subsequent geometry optimization. The list of the molecules, including their names, NCS numbers, CAS numbers and experimental pK_a values, can be found in the (Additional file 1: Table S1a). The SDF files with the 3D structures of molecules and their dissociated forms are also in the (Additional file 2: Molecules).

Data set for carboxylic acids

An aspect which is very important for the applicability of the pK_a prediction approach is its transferability to other chemical classes. In this work, we provide a case study showing the performance of the approach on carboxylic acids, which are also very common testing molecules for pK_a prediction approaches [19-21,43]. The data set contains 71 distinct molecules of carboxylic acids and their dissociated forms. The 3D structures of these molecules were obtained in the same way as for the phenols. The list of the molecules, including their names, NCS numbers, CAS numbers and experimental pK_a values can be found in the (Additional file 3: Table S1b). The SDF files with the 3D structures of the molecules and their dissociated forms are also included in the (Additional file 2: Molecules).

pK_a values

The experimental pK_a values were taken from the Physprop database [61].

Table 1 Summary information about the EEM parameter sets used in the present study

QM theory level + basis set	PA	EEM parameter set name	Published by	Year of publication	Elements included
HF/STO-3G	MPA	Svob2007_cbeg2	Svobodova et al. [44]	2007	C, O, N, H, S
		Svob2007_cmet2	Svobodova et al. [44]	2007	C, O, N, H, S, Fe, Zn
		Svob2007_chal2	Svobodova et al. [44]	2007	C, O, N, H, S, Br, Cl, F, I
		Svob2007_hm2	Svobodova et al. [44]	2007	C, O, N, H, S, F, Cl, Br, I, Fe, Zn
		Baek1991	Baekelandt et al. [45]	1991	C, O, N, H, P, Al, Si
		Mort1986	Mortier et al. [39]	1986	C, O, N, H
HF/6–31G*	MK	Jir2008_hf	Jirouskova et al. [46]	2008	C, O, N, H, S, F, Cl, Br, I, Zn
B3LYP/6–31G*	MPA	Chaves2006	Chaves et al. [47]	2006	C, O, N, H, F
		Bult2002_mul	Bultinck et al. [48]	2002	C, O, N, H, F
	NPA	Ouy2009	Ouyang et al. [49]	2009	C, O, N, H, F
		Ouy2009_elem	Ouyang et al. [49]	2009	C, O, N, H, F
		Ouy2009_elemF	Ouyang et al. [49]	2009	C, O, N, H, F
		Bult2002_npa	Bultinck et al. [48]	2002	C, O, N, H, F
	Hir.	Bult2002_hir	Bultinck et al. [48]	2002	C, O, N, H, F
	MK	Jir2008_mk	Jirouskova et al. [46]	2008	C, O, N, H, S, F, Cl, Br, I, Zn
		Bult2002_mk	Bultinck et al. [48]	2002	C, O, N, H, F
	CHELPG	Bult2002_che	Bultinck et al. [48]	2002	C, O, N, H, F
	AIM	Bult2004_aim	Bultinck et al. [50]	2004	C, O, N, H, F

Descriptors and QSPR models for phenols

Our descriptors were atomic charges. We analyzed two types of QSPR models, namely QSPR models with three descriptors (3d QSPR models) and QSPR models with five descriptors (5d QSPR models).

The 3d QSPR models used those descriptors which proved to be the most relevant for pK_a prediction in our previous study [24]. Therefore these descriptors were the atomic charge of the hydrogen atom from the phenolic OH group (q_H), the charge on the oxygen atom from the phenolic OH group (q_O), and the charge on the carbon atom binding the phenolic OH group (q_{C1}). These descriptors were used to establish the QSPR models by the general equation:

$$\mathrm{p}K_a = p_H \cdot q_H + p_O \cdot q_O + p_{C1} \cdot q_{C1} + p \quad (2)$$

where p_H, p_O, p_{C1} and p are parameters of the QSPR model (i.e., constants derived by multiple linear regression). The 5d QSPR models employ the above mentioned descriptors q_H, q_O and q_{C1} and additionally also the charge on the phenoxide O^- from the dissociated molecule (q_{OD}), and the charge on the carbon atom binding this oxygen (q_{C1D}). Using the charges from the dissociated molecules for pK_a prediction was inspired by the work of Dixon et al. [19]. The equation of the 5d QSPR models is therefore:

$$\mathrm{p}K_a = p'_H \cdot q_H + p'_O \cdot q_O + p'_{C1} \cdot q_{C1} + p'_{OD} \cdot q_{OD} + p'_{C1D} \cdot q_{C1D} + p' \quad (3)$$

where p'_H, p'_O, p'_{C1}, p'_{OD}, p'_{C1D} and p' are parameters of the QSPR model.

Descriptors and QSPR models for carboxylic acids

The descriptors were again atomic charges and, similarly as for phenols, two types of QSPR models were developed and evaluated. Specifically, QSPR models with four descriptors (4d QSPR models) and QSPR models with seven descriptors (7d QSPR models). The 4d QSPR models used similar descriptors as the 3d models for phenols - the atomic charge of the hydrogen atom from the COOH group (q_H), the charge on the hydrogen bound oxygen atom from the COOH group (q_O), and the charge on the carbon atom binding the COOH group (q_{C1}). Additionally, also the charge of the second carboxyl oxygen (q_{O2}) is included. These 4d QSPR models are represented by the equation:

$$\mathrm{p}K_a = p_H \cdot q_H + p_O \cdot q_O + p_{O2} \cdot q_{O2} + p_{C1} \cdot q_{C1} + p \quad (4)$$

where p_H, p_O, p_{O2}, p_{C1} and p are parameters of the QSPR model. The 7d QSPR models employ also charges from the dissociated forms, namely the charge on the carboxyl oxygens (q_{OD}, q_{O2D}) and the charge on the carboxylic carbon atom (q_{C1D}). The equation of the 7d QSPR models is therefore:

$$\mathrm{p}K_a = p'_H \cdot q_H + p'_O \cdot q_O + p'_{O2} \cdot q_{O2} + p'_{C1} \cdot q_{C1} + p'_{OD} \cdot q_{OD} + p'_{O2D} \cdot q_{O2D} + p'_{C1D} \cdot q_{C1D} + p' \quad (5)$$

where p'_H, p'_O, p'_{O2}, p'_{C1}, p'_{OD}, p'_{O2D}, p'_{C1D} and p' are parameters of the QSPR model.

QSPR model parameterization

The parameterization of the QSPR models was done by multiple linear regression (MLR) using the software tool QSPR Designer [62].

Results and discussion

QM and EEM QSPR models for phenols

We prepared one 3d QSPR model and one 5d QSPR model using atomic charges calculated by each of the above mentioned 18 EEM parameter sets. These models are denoted 3d or 5d EEM QSPR models. Additionally, we created one 3d and one 5d QSPR model using atomic charges calculated by each of the corresponding 8 QM charge calculation approaches (denoted as 3d or 5d QM QSPR models). The data set of 74 phenol molecules was used for the parameterization of the QSPR models, and the obtained models were validated for all molecules in the data set.

The parameterization of the 3d EEM QSPR models showed that several molecules in the data set perform as outliers. For this reason, we created also EEM QSPR models without outliers (i.e., EEM QSPR models for which parameterization was done using a data set that excluded the previously observed outliers). These models are denoted 3d EEM QSPR WO models. We classified as outliers 10% of the molecules from our data set, which had the highest Cook's square distance. Therefore the 3d EEM QSPR WO models were parameterized using 67 molecules, and their validation was also done on the data set excluding outliers.

The quality of the QSPR models, i.e. the correlation between experimental pK_a and the pK_a calculated by each model, was evaluated using the squared Pearson correlation coefficient (R^2), root mean square error (RMSE), and average absolute pK_a error ($\overline{\Delta}$), while the statistical criteria were the standard deviation of the estimation (s) and Fisher's statistics of the regression (F).

Table 2 contains the quality criteria (R^2, RMSE, $\overline{\Delta}$) and statistical criteria (s and F) for all the QSPR models analyzed. All these models are statistically significant at $p = 0.01$. Since our data sets contained 74 and 67 molecules, respectively, the appropriate F value to consider was that for 60 samples. Thus, the 3d QSPR models are statistically significant (at $p = 0.01$) when $F > 4.126$ and the 5d QSPR models when $F > 3.339$. Figure 1 summarizes the R^2 of all QSPR models for ease of visual comparison, and Tables 3 and 4 provide a comparison of the models from specific points of view. The parameters of the QSPR models are summarized in the (Additional file 4: Table S2) and all charge descriptors and pK_a values are contained in the (Additional file 5: Table S6). The most relevant graphs of correlation between experimental and calculated pK_a are visualized in Figure 2.

Prediction of pK_a using EEM charges

The key question we wanted to answer in this paper is whether EEM QSPR models are applicable for pK_a prediction. For this purpose we selected a set of phenol molecules and generated QSPR models which used EEM atomic charges as descriptors. We then evaluated the accuracy of these models by comparing the predicted pK_a values with the experimental ones. The results (see Tables 2 and 3, Figure 1) clearly show that QSPR models based on EEM charges are indeed able to predict the pK_a of phenols with very good accuracy. Namely, 63% of the EEM QSPR models evaluated in this study were able to predict pK_a with $R^2 > 0.9$. The average R^2 for all 54 EEM QSPR models considered was 0.9, while the best EEM QSPR model reached $R^2 = 0.924$. Our findings thus suggest that EEM atomic charges may prove as efficient QSPR descriptors for pK_a prediction. The only drawback of EEM is that EEM parameters are currently not available for some types of atoms. Nevertheless, EEM parameterization is still a topic of research, therefore more general parameter sets are being developed.

Improvement of EEM QSPR models by removing outliers

The quality of 3d EEM QSPR models can be markedly increased by removing the outliers. In this case, the models have average $R^2 = 0.911$ and 83% of them have $R^2 > 0.9$. The disadvantage of these models is that they are not able to cover the complete data set (i.e., 10% of molecules must be excluded as outliers).

On the other hand, the outliers are similar for all EEM QSPR models. For example, while 16 molecules from our data set are outliers for at least one parameter set, 10 out of these 16 molecules are outliers for five or more parameter sets. From the chemical point of view, most of the outliers contain one or more nitro groups. This may be related to reported lower accuracy of EEM for these groups [48]. In general one limitation of the 3d EEM QSPR models is that they are very sensitive to the quality of EEM charges. Therefore, if the EEM charges are inaccurate for certain compounds or class of compounds, the 3d QSPR models based on these EEM charges will have lower performance for these compounds or class of compounds. In addition, a lower experimental accuracy of these pK_a values may also be a reason for low performance in some cases. A table containing information about outlier molecules is given in the (Additional file 6: Table S3).

Improvement of EEM QSPR models by adding descriptors

Our first EEM QSPR models contained three descriptors (3d), namely atomic charges originating from the non-dissociated molecule. Nonetheless, in our study we found

Table 2 Quality criteria and statistical criteria for all the QSPR models analyzed in the present study and focused on phenols

QM theory level + basis set	PA	EEM parameter set name	QSPR model	R^2	RMSE	$\overline{\Delta}$	s	F
HF/STO-3G	MPA	-	3d QM	0.9515	0.490	0.388	0.504	458
		-	5d QM	0.9657	0.412	0.310	0.430	358
		Svob2007_cbeg2	3d EEM	0.8671	0.812	0.571	0.835	152
			3d EEM WO	0.9239	0.482	0.382	0.497	255
			5d EEM	0.9179	0.638	0.481	0.666	152
		Svob2007_cmet2	3d EEM	0.8663	0.814	0.577	0.837	151
			3d EEM WO	0.9239	0.482	0.386	0.497	255
			5d EEM	0.9189	0.634	0.476	0.661	154
		Svob2007_chal2	3d EEM	0.8737	0.792	0.554	0.814	161
			3d EEM WO	0.9127	0.483	0.387	0.498	220
			5d EEM	0.9203	0.629	0.473	0.656	157
		Svob2007_hm2	3d EEM	0.8671	0.812	0.578	0.835	152
			3d EEM WO	0.9241	0.481	0.387	0.496	256
			5d EEM	0.9179	0.638	0.478	0.666	152
		Baek1991	3d EEM	0.9099	0.669	0.531	0.688	236
			3d EEM WO	0.9166	0.531	0.423	0.548	231
			5d EEM	0.9195	0.632	0.493	0.659	155
		Mort1986	3d EEM	0.8860	0.752	0.577	0.773	181
			3d EEM WO	0.9151	0.520	0.405	0.536	226
			5d EEM	0.9142	0.652	0.524	0.680	145
HF/6–31G*	MK	-	3d QM	0.8405	0.890	0.727	0.915	123
		-	5d QM	0.8865	0.750	0.641	0.782	106
		Jir2008_hf	3d EEM	0.8612	0.830	0.582	0.853	145
			3d EEM WO	0.9182	0.500	0.394	0.516	236
			5d EEM	0.9154	0.648	0.488	0.676	147
B3LYP/6–31G*	MPA	-	3d QM	0.9671	0.404	0.317	0.415	686
		-	5d QM	0.9724	0.370	0.274	0.386	479
		Chaves2006	3d EEM	0.891	0.735	0.570	0.756	191
			3d EEM WO	0.9198	0.505	0.398	0.521	241
			5d EEM	0.9192	0.633	0.489	0.660	155
		Bult2002_mul	3d EEM	0.8876	0.747	0.589	0.768	184
			3d EEM WO	0.9151	0.520	0.416	0.536	226
			5d EEM	0.9158	0.646	0.504	0.674	148
B3LYP/6–31G*	NPA	-	3d QM	0.9590	0.451	0.349	0.464	546
		-	5d QM	0.9680	0.399	0.295	0.416	411
		Ouy2009	3d EEM	0.8731	0.793	0.541	0.815	161
			3d EEM WO	0.9043	0.505	0.379	0.521	198
			5d EEM	0.9094	0.670	0.503	0.699	137
		Ouy2009_elem	3d EEM	0.8727	0.795	0.546	0.817	160
			3d EEM WO	0.9113	0.487	0.382	0.502	216
			5d EEM	0.9132	0.656	0.495	0.684	143
		Ouy2009_elemF	3d EEM	0.8848	0.756	0.519	0.777	179
			3d EEM WO	0.9012	0.512	0.386	0.528	192
			5d EEM	0.8866	0.750	0.520	0.782	106

Table 2 Quality criteria and statistical criteria for all the QSPR models analyzed in the present study and focused on phenols *(continued)*

	Bult2002_npa	3d EEM	0.9044	0.689	0.532	0.708	221
		3d EEM WO	0.9098	0.523	0.405	0.539	212
		5d EEM	0.9180	0.638	0.488	0.666	152
Hir.	-	3d QM	0.9042	0.689	0.503	0.708	220
	-	5d QM	0.9477	0.509	0.356	0.531	246
	Bult2002_hir	3d EEM	0.8415	0.887	0.636	0.912	124
		3d EEM WO	0.8838	0.579	0.414	0.597	160
		5d EEM	0.9050	0.687	0.522	0.717	130
MK	-	3d QM	0.8447	0.878	0.705	0.903	127
	-	5d QM	0.8960	0.718	0.594	0.749	117
	Jir2008_dft	3d EEM	0.8696	0.804	0.555	0.827	156
		3d EEM WO	0.9224	0.487	0.371	0.502	250
		5d EEM	0.9148	0.650	0.489	0.678	146
	Bult2002_mk	3d EEM	0.8639	0.822	0.610	0.845	148
		3d EEM WO	0.9053	0.519	0.384	0.535	201
		5d EEM	0.9131	0.657	0.508	0.685	143
Chel.	-	3d QM	0.8528	0.854	0.712	0.878	135
	-	5d QM	0.9087	0.673	0.552	0.702	135
	Bult2002_che	3d EEM	0.8695	0.805	0.597	0.828	155
		3d EEM WO	0.8863	0.588	0.436	0.606	164
		5d EEM	0.9057	0.684	0.540	0.714	131
AIM	-	3d QM	0.9609	0.440	0.332	0.452	573
	-	5d QM	0.9677	0.400	0.285	0.417	407
	Bult2004_aim	3d EEM	0.8646	0.819	0.619	0.842	149
		3d EEM WO	0.8972	0.590	0.438	0.608	183
		5d EEM	0.9017	0.698	0.571	0.728	125

that using two additional charge descriptors from the dissociated molecule can markedly improve the predictive power of the EEM QSPR models. Tables 2 and 3, Figure 1 show that these new 5d EEM QSPR models provide better pK_a prediction than their corresponding 3d EEM QSPR models. Specifically, adding the descriptors derived from the dissociated molecules increased the average R^2 value for the EEM QSPR models from 0.876 to 0.913.

Comparison of EEM QSPR models and QM QSPR models

Another important question is how accurate the EEM QSPR models are in comparison with QM QSPR models. Table 2 and Figure 1 show that QM QSPR models provide, in most cases, more precise predictions. This is confirmed also by the average R^2 values from Table 3. This is not surprising, since EEM is an empirical method which just mimics the QM approach for which it was parameterized. An interesting fact is that the differences in accuracy between QM QSPR models and EEM QSPR models are not substantial. For example, 5d EEM QSPR models have average $R^2 = 0.913$, while 5d QM QSPR models have average $R^2 = 0.951$. We also note that adding more descriptors to a QM QSPR model brings less improvement than adding more descriptors to an EEM QSPR model.

Influence of theory level and basis set

EEM parameters are available only for a relatively small number of theory levels (HF and B3LYP) and basis sets (STO-3G and 6–31G*). Therefore we can not perform such a deep analysis of theory level and basis set influence on pK_a prediction capability from EEM atomic charges, as was done for QM QSPR models by Gross et al. [22] or Svobodova et al. [24]. We can only compare the models employing HF/STO-3G and B3LYP/6–31G* charges, as these are the only combinations for which EEM parameters are available for the same population analysis (MPA). Therefore we can study only the influence of the combination of theory level / basis set, and not the isolated influence of the theory level or basis set. Our analysis revealed that B3LYP/6–31G* charges provide slightly more accurate QM QSPR models than HF/STO-3G charges (see

QM theory level + basis set	PA	EEM parameter set name	R^2 of QSPR model				
			3d EEM	3d EEM WO	5d EEM	3d QM	5d QM
HF/STO-3G	MPA	Svob2007_cbeg2	0.8671	0.9239	0.9179	0.9515	0.9657
		Svob2007_cmet2	0.8663	0.9239	0.9189		
		Svob2007_chal2	0.8737	0.9127	0.9203		
		Svob2007_hm2	0.8671	0.9241	0.9179		
		Baek1991	0.9099	0.9166	0.9195		
		Mort1986	0.8860	0.9151	0.9142		
HF/6-31G*	MK	Jir2008_hf	0.8696	0.9182	0.9154	0.8405	0.8865
B3LYP/6-31G*	MPA	Chaves 2006	0.8910	0.9198	0.9192	0.9671	0.9724
		Bult2002_mul	0.8876	0.9151	0.9158		
	NPA	Ouy2009	0.8731	0.9043	0.9094	0.9590	0.9680
		Ouy2009_elem	0.8727	0.9113	0.9132		
		Ouy2009_elemF	0.8848	0.9012	0.8866		
		Bult2002_npa	0.9044	0.9098	0.9180		
	Hir.	Bult2002_hir	0.8415	0.8838	0.9050	0.9042	0.9477
	MK	Jir2008_mk	0.8696	0.9224	0.9148	0.8447	0.8960
		Bult2002_mk	0.8639	0.9053	0.9131		
	Chel.	Bult2002_che	0.8695	0.8863	0.9057	0.8528	0.9087
	AIM	Bult2004_aim	0.8646	0.8972	0.9017	0.9609	0.9677

Legend	excellent	very good	good	satisfactory	acceptable	weak
R^2	0.95 – 0.97	0.92 – 0.95	0.91 – 0.92	0.9 – 0.91	0.85 – 0.9	0.8 – 0.85

Figure 1 R^2 for the correlation between calculated and experimental pK_a.

Table 3 Average R^2 between experimental and predicted pK_a for all QSPR models of a certain type and percentages of QSPR models whose R^2 values are in a certain interval

QSPR model		3d EEM	3d EEM WO	5d EEM	3d QM	5d QM
Average R^2		0.876	0.911	0.913	0.929	0.951
Interval of R^2	$R^2 > 0.9$	11%	83%	94%	78%	83%
	$0.9 \geq R^2 > 0.85$	83%	17%	6%	6%	17%
	$0.85 \geq R^2 > 0.8$	6%	0%	0%	17%	0%
QSPR model		**EEM based models**			**QM based models**	
Average R^2		0.900			0.940	
Interval of R^2	$R^2 > 0.9$	63%			81%	
	$0.9 \geq R^2 > 0.85$	35%			13%	
	$0.85 \geq R^2 > 0.8$	2%			6%	

Table 4 Average R^2 between experimental and predicted pK_a for all QSPR models using atomic charges calculated by a specific combination of theory level and basis set, or by a specific population analysis

QSPR model		3d EEM	3d EEM WO	5d EEM	3d QM	5d QM
Theory level and basis set *	HF/STO-3G	0.878	0.919	0.918	0.952	0.966
	B3LYP/6–31G*	0.889	0.917	0.918	0.967	0.972
Population analysis **	MPA	0.889	0.917	0.918	0.967	0.972
	NPA	0.884	0.907	0.907	0.959	0.968
	Hirshfeld	0.842	0.884	0.905	0.904	0.948
	MK	0.867	0.914	0.914	0.845	0.896
	CHELPG	0.870	0.886	0.906	0.853	0.909
	AIM	0.865	0.897	0.902	0.961	0.968

*Only QSPR models employing MPA were included in this analysis.
**Only QSPR models using B3LYP/6–31G* were included in this analysis.

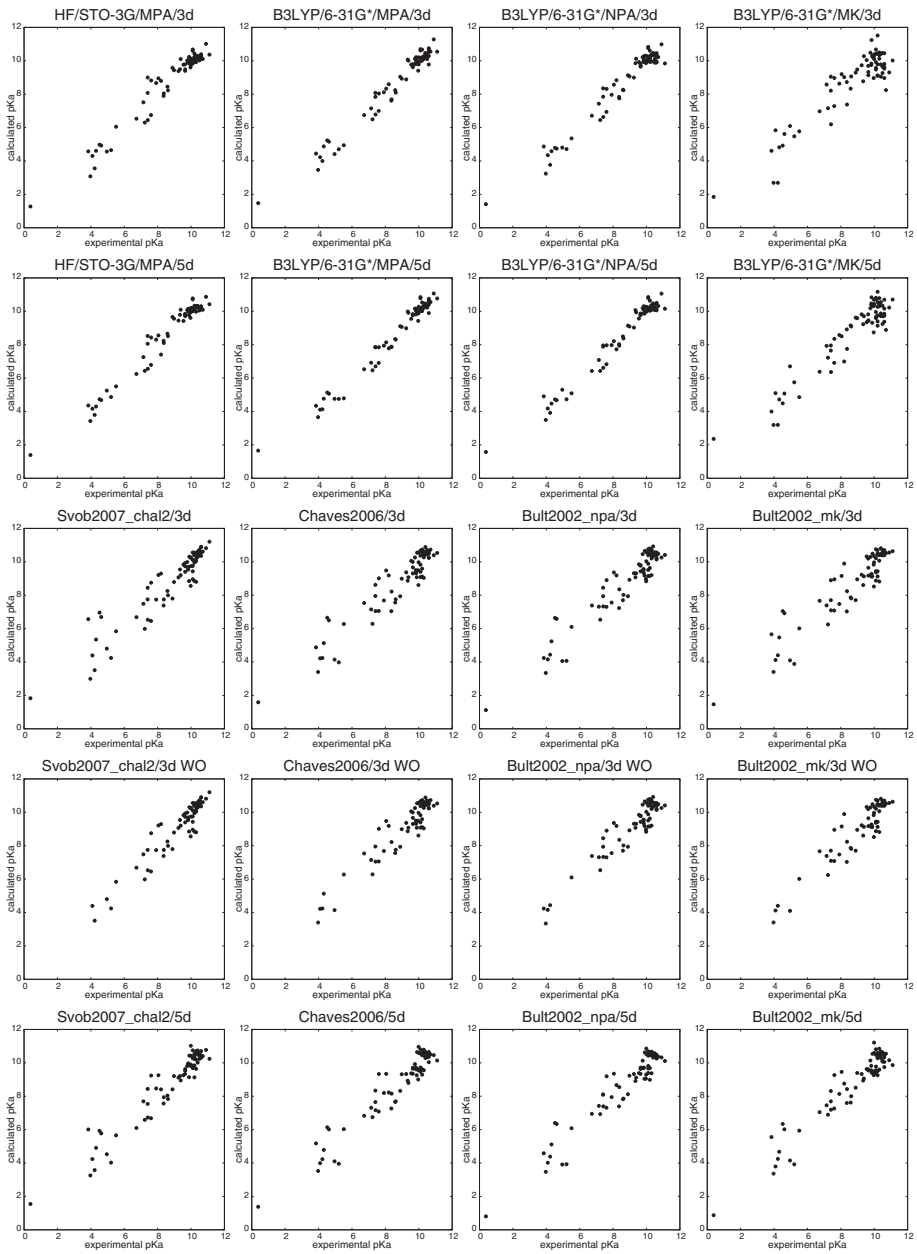


Figure 2 Correlation graphs. Graphs showing the correlation between experimental and calculated pK_a for selected QSPR models.

Table 4). This is in agreement with our previous findings [24], and it can be explained by the fact that 6–31G* is a more robust basis set than STO-3G. However, the difference is not marked in the case of EEM QSPR models.

Influence of population analysis

Eleven EEM parameter sets were published for B3LYP/6–31G* with six different population analyses (see Table 1). Therefore it is straightforward to analyze the influence of the population analysis on the predictive power of the resulting QSPR models (see Table 4). We found that MPA and NPA provide the best QM models, while MK and CHELPG (PAs based on fitting the atomic charges to the molecular electrostatic potential) provide weak QM models. Our results are in agreement with previous studies [22,24]. QM QSPR models based on AIM predict pK_a with accuracy comparable to MPA and NPA. In the case of EEM QSPR models, we did indeed find that MPA provided the best models, but most of the other population analyses gave comparable results. This confirms previous observations that the EEM approach is not able to faithfully mimic MK charges [63]. On the other hand,

Table 5 Comparison between the performance of the QSPR models developed here, and previously developed models

Method	Theory level	PA	Basis set	Descriptors	R^2	*s*	*F*	Number of molecules	Source
QM	B3LYP	NPA	6–311G**	q_{OH}	0.789	1.300	48	15	Kreye and Seybold [23][a]
	B3LYP	NPA	6–311G**	q_O	0.731	1.500	38	15	Kreye and Seybold [23][a]
	B3LYP	NPA	6–31+G*	q_{OH}	0.880	0.970	95	15	Kreye and Seybold [23][b]
	B3LYP	NPA	6–31+G*	q_O	0.865	1.000	38	15	Kreye and Seybold [23][b]
	B3LYP	NPA	6–311G(d,p)	q_{O-}	0.911	0.252	173	19	Gross and Seybold [22]
	B3LYP	NPA	6–311G(d,p)	q_H	0.887	0.283	134	19	Gross and Seybold [22]
	B3LYP	NPA	6–31G*	q_H, q_O, q_{C1}	0.961	0.440	986	124	Svobodova and Geidl [24]
	B3LYP	NPA	6–311G	q_H, q_O, q_{C1}	0.962	0.435	1013	124	Svobodova and Geidl [24]
	B3LYP	NPA	6–31G*	q_H, q_O, q_{C1}	0.959	0.464	545	74	This work
	B3LYP	NPA	6–31G*	q_H, q_O, q_{C1}, q_{OD}, q_{C1D}	0.968	0.410	705	74	This work
EEM	B3LYP	NPA	6–31G*	q_H, q_O, q_{C1}, q_{OD}, q_{C1D}	0.918	0.656	261	74	This work[c]
QM	B3LYP	MPA	6–311G(d,p)	q_H	0.913	0.248	179	19	Gross and Seybold [22]
	B3LYP	MPA	6–311G(d,p)	q_{O-}	0.894	0.274	144	19	Gross and Seybold [22]
	B3LYP	MPA	6–311G	q_H, q_O, q_{C1}	0.938	0.556	605	124	Svobodova and Geidl [24]
	B3LYP	MPA	6–31G*	q_H, q_O, q_{C1}	0.959	0.450	936	124	Svobodova and Geidl [24]
	B3LYP	MPA	6–31G*	q_H, q_O, q_{C1}	0.967	0.415	685	74	This work
	B3LYP	MPA	6–31G*	q_H, q_O, q_{C1}, q_{OD}, q_{C1D}	0.972	0.380	822	74	This work
EEM	B3LYP	MPA	6–31G*	q_H, q_O, q_{C1}, q_{OD}, q_{C1D}	0.919	0.651	265	74	This work[d]
QM	B3LYP	MK	6–311G(d,p)	q_H	0.344	0.682	9	19	Gross and Seybold [22]
	B3LYP	MK	6–311G(d,p)	q_{O-}	0.692	0.467	38	19	Gross and Seybold [22]
	B3LYP	MK	6–311G	q_H, q_O, q_{C1}	0.822	0.941	185	124	Svobodova and Geidl [24]
	B3LYP	MK	6–31G*	q_H, q_O, q_{C1}	0.808	0.978	168	124	Svobodova and Geidl [24]
	B3LYP	MK	6–31G*	q_H, q_O, q_{C1}	0.845	0.902	126	74	This work
	B3LYP	MK	6–31G*	q_H, q_O, q_{C1} q_{OD}, q_{C1D}	0.896	0.739	201	74	This work
EEM	B3LYP	MK	6–31G*	q_H, q_O, q_{C1} q_{OD}, q_{C1D}	0.915	0.669	250	74	This work[e]

[a]With solvent model SM5.4.
[b]With solvent model SM8.
[c]EEM parameter set Bult2002 npa.
[d]EEM parameter set Chaves2006.
[e]EEM parameter set Jir2008 mk.

this drawback of EEM allowed the EEM QSPR models employing MK charges to predict pK_a more accurately than the corresponding QM QSPR models.

Influence of the EEM parameter set

Two or more EEM parameter sets are available in literature for four combinations of theory level, basis set and population analysis (see Table 1). We found that the quality of EEM QSPR models employing the same types of charges slightly varies when using EEM parameters coming from different studies (see Table 2 and Figure 1). Even EEM parameters from the same study, but obtained by different approaches, lead to QSPR models of slightly different quality. In any case, these differences are minimal.

Comparison with previous work

QM QSPR models for pK_a prediction in phenols, similar to those presented in this paper (i.e., employing similar

Table 6 Comparison of the quality criteria and statistical criteria for the training set, test set and complete set for some selected charge calculation approaches

5d EEM QSPR model employing Svob2007_chal2 EEM parameters:

Complete set:

R^2	RMSE	*s*	*F*	Number of molecules
0.920	0.629	0.647	269	74

Cross-validation:

Cross-validation step	Training set					Test set				
	R^2	RMSE	*s*	*F*	Number of molecules	R^2	RMSE	*s*	*F*	Number of molecules
1	0.9283	0.5211	0.5498	137	59	0.9202	1.0754	1.3884	21	15
2	0.9210	0.6538	0.6899	124	59	0.9029	0.5394	0.6963	17	15
3	0.9191	0.6442	0.6796	120	59	0.9275	0.5823	0.7517	23	15
4	0.9207	0.6244	0.6588	123	59	0.9271	0.6878	0.8880	23	15
5	0.9274	0.6302	0.6643	138	60	0.9008	0.6678	0.8834	15	14

5d EEM QSPR model employing Ouy2009_elemF EEM parameters:

Complete set:

R^2	RMSE	*s*	*F*	Number of molecules
0.8866	0.7501	0.7825	106	74

Cross-validation:

Cross-validation step	Training set					Test set				
	R^2	RMSE	*s*	*F*	Number of molecules	R^2	RMSE	*s*	*F*	Number of molecules
1	0.8936	0.6349	0.6698	89	59	0.8704	1.2857	1.6598	12	15
2	0.8953	0.7526	0.7940	91	59	0.8018	0.7802	1.0072	7	15
3	0.8908	0.7481	0.7893	86	59	0.8647	0.7983	1.0306	12	15
4	0.8821	0.7614	0.8033	79	59	0.9154	0.7481	0.9658	19	15
5	0.8956	0.7557	0.7966	93	60	0.8089	0.8396	1.1107	7	14

charges) were previously published by Gross and Seybold [22], Kreye and Seybold [23] and Svobodova and Geidl [24]. Table 5 shows a comparison between these models and the models developed in this study. Our work is the first which presents QSPR models for pK_a prediction based on EEM charges. Therefore, we can not provide a comparison between EEM QSPR models, but we can compare against QSPR models based on QM charges only. It is seen therein that our 3d QM QSPR models show markedly higher R^2 and *F* values than the models published by Gross and Seybold and Kreye and Seybold (even if some of these models employ higher basis sets) and comparable R^2 and *F* values as models published by Svobodova and Geidl. Moreover, our 5d QM QSPR models outperform the models from Svobodova and Geidl. Our best EEM QSPR models (i.e., 5d EEM QSPR models) provide even better results than QM QSPR models from Gross and Seybold and Kreye and Seybold. These EEM QSPR models are not as accurate as the QM QSPR models published by Svobodova and Geidl or those developed in this work, but the loss of accuracy is not too high (R^2 values are still > 0.91).

Cross-validation

Our results show that 5d EEM QSPR models provide a fast and accurate approach for pK_a prediction. Nonetheless, the robustness of these models should be proved. Therefore, all the 5d EEM QSPR models (i.e., 18 models) were tested by cross-validation. For comparison, also the cross-validation of all 5d QM QSPR models (i.e., 8 models) was done. The *k*-fold cross-validation procedure was used [64,65], where $k = 5$. Specifically, the set of phenol molecules was divided into five parts (each contained 20% of the molecules). The division was done randomly, and included stratification by pK_a value. Afterwards, five cross validation steps were performed. In the first step, the first part was selected as a test set, and the remaining four parts were taken together as the training set. The test and training sets for the other steps were prepared in a similar manner, by subsequently considering

QM theory level + basis set	PA	EEM parameter set name	R^2 of QSPR model 7d EEM	7d QM
HF/STO-3G	MPA	Svob2007_cbeg2	0.8831	0.9327
		Svob2007_cmet2	0.8810	
		Svob2007_chal2	0.8822	
		Svob2007_hm2	0.8793	
		Baek1991	0.9211	
		Mort1986	0.9176	
B3LYP/6-31G*	MPA	Chaves2006	0.9238	0.9059
		Bult2002_mul	0.9248	
	NPA	Ouy2009	0.8825	0.9169
		Ouy2009_elem	0.8777	
		Ouy2009_elemF	0.8478	
		Bult2002_npa	0.9094	

Legend	very good	good	satisfactory	acceptable	weak
R^2	0.92 – 0.95	0.91 – 0.92	0.9 – 0.91	0.85 – 0.9	0.8 – 0.85

Figure 3 Correlation between calculated and experimental pK_a for carboxylic acids.

one part as a test set, while the remaining parts served as a training set. For each step, the QSPR model was parameterized on the training set. Afterwards, the pK_a values of the respective test molecules were calculated via this model, and compared with experimental pK_a values. The results are summarized in the (Additional file 7: Table S4), while the cross-validation results for the best and the worst performing 5d EEM QSPR models are shown in Table 6. The cross-validation showed that the models are stable and the values of R^2 and RMSE are similar for the test set, the training set and the complete set. The robustness of EEM QSPR models and QM QSPR models is comparable.

Case study for carboxylic acids

We have shown that QSPR models based on EEM atomic charges can be used for predicting pK_a in phenols. In order to evaluate the general applicability of this approach for pK_a prediction, we tested the performance of such models for carboxylic acids. This case study is done for the charge schemes found to provide the best QM and EEM QSPR models in the case of phenols. Specifically, QM charges calculated by HF/STO-3G/MPA, B3LYP/6–31G*/MPA and B3LYP/6–31G*/NPA, and EEM charges calculated using the corresponding EEM parameters. Because 5d QSPR models provide the most accurate prediction for phenols, the case study is focused on their analogue for carboxylic acids, i.e., 7d QSPR models. Squared Pearson correlation coefficients of the analysed QSPR models are summarized in Figure 3, and all the quality and statistical criteria can be found in (Additional file 8: Table S5). The results show that 7d EEM QSPR models are able to predict the pK_a of carboxylic acids with very good accuracy. Namely, 5 out of 12 analysed 7d EEM QSPR models were able to predict pK_a with $R^2 > 0.9$, while the best EEM QSPR model reached $R^2 = 0.925$. Therefore, we concluded that EEM QSPR models are indeed applicable also for carboxylic acids. Again QM QSPR models perform better than EEM QSPR models, but the differences are not substantial.

Conclusions

We found that the QSPR models employing EEM charges can be a suitable approach for pK_a prediction. From our 54 EEM QSPR models focused on phenols, 63% show a correlation of $R^2 > 0.9$ between the experimental and predicted pK_a. The most successful type of these EEM QSPR models employed 5 descriptors, namely the atomic charge of the hydrogen atom from the phenolic OH group, the charge on the oxygen atom from the phenolic OH group, the charge on the carbon atom binding the phenolic OH group, the charge on the oxygen from the phenoxide O^- from the dissociated molecule, and the charge on the carbon atom binding this oxygen. Specifically, 94% of these models have $R^2 > 0.9$, and the best one has $R^2 = 0.920$. In general, including charge descriptors from dissociated molecules, which was introduced in our work, always increases the quality of a QSPR model. The only drawback of EEM QSPR models is that the EEM parameters are currently not available for all types of atoms. Therefore the EEM parameter sets need to be expanded to larger sets of molecules and further improved.

As expected, the QM QSPR models provided more accurate pK_a predictions than the EEM QSPR models. Nevertheless, these differences are not substantial. Furthermore, a big advantage of EEM QSPR models is that one can calculate the EEM charges markedly faster than the QM charges. Moreover, the EEM QSPR models are not so strongly influenced by the charge calculation approach as the QM QSPR models are. Specifically, the QM QSPR models which use atomic charges obtained from calculations with higher basis set perform better, while the EEM QSPR models do not show such marked differences. Similarly, the quality of QM QSPR models depends a lot on population analysis, but EEM QSPR models are not influenced so much. Namely, QM QSPR models which use atomic charges calculated from MPA, NPA and Hirshfeld PA performed very well, while MK provides only weak models. In the case of EEM QSPR models, MPA performs also the best, but all other PAs (including MK) provide accurate results as well. The source of the EEM parameters also did not affect the quality of the EEM QSPR models significantly.

The robustness of EEM QSPR models was successfully confirmed by cross-validation. Specifically, the accuracy of pK_a prediction for the test, training and complete set were comparable. The applicability of EEM QSPR models for other chemical classes was tested in a case study focused on carboxylic acids. This case study showed that EEM QSPR models are indeed applicable for pK_a prediction also for carboxylic acids. Namely, 5 from 12 of these models were able to predict pK_a with $R^2 > 0.9$, while the best EEM QSPR model reached $R^2 = 0.925$.

Therefore, EEM QSPR models constitute a very promising approach for the prediction of pK_a. Their main advantages are that they are accurate, and can predict pK_a values very quickly, since the atomic charge descriptors used in the QSPR model can be obtained much faster by EEM than by QM. Additionally, the quality of EEM QSPR models is less dependent on the type of atomic charges used (theory level, basis set, population analysis) than in the case of QM QSPR models. Accordingly, EEM QSPR models constitute a pK_a prediction approach which is very suitable for virtual screening.

Additional files

Additional file 1: Table S1a. The list of the phenol molecules, including their names, NCS numbers, CAS numbers and experimental pK_a values.

Additional file 2: Molecules. The SDF files with the structures of the molecules and also their dissociated forms.

Additional file 3: Table S1b. The list of the carboxylic acid molecules, including their names, NCS numbers, CAS numbers and experimental pK_a values.

Additional file 4: Table S2. The parameters of all the QSPR models for phenols.

Additional file 5: Table S6. The table containing charge descriptors for all charge calculation approaches and predicted pK_a values for all QSPR models (for phenols).

Additional file 6: Table S3. The information about outlier molecules for phenols.

Additional file 7: Table S4. The table of cross-validation results for phenols.

Additional file 8: Table S5. The quality and statistical criteria of QSPR models for carboxylic acids.

Abbreviations

3d: 3 descriptors; 4d: 4 descriptors; 5d: 5 descriptors; 7d: 7 descriptors; AIM: Atoms in Molecules; ANN: Artificial Neural Networks; B3LYP: Becke, three-parameter, Lee-Yang-Parr; DENR: Dynamic Electronegativity Relaxation; EEM: Electronegativity Equalization Method; GDAC: Geometry-Dependent Atomic Charge; HF: Hartree-Fock; KCM: Kirchhoff Charge Model; LFER: Linear Free Energy Relationships; MK: Merz-Singh-Kollman; MLR: Multiple Linear Regression; MP2: Møller-Plesset Perturbation Theory; MPA: Mulliken Population Analysis; NPA: Natural Population Analysis; PA: Population Analysis; PEOE: Partial Equalization of Orbital Electronegativity; QEq: Charge Equilibration; QM: Quantum Mechanical; QSPR: Quantitative Structure-Property Relationship; RMSE: Root Mean Square Error; SQE: Split Charge Equilibration; TSEF: Topologically Symmetric Energy Function; WO: Without Outliers.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

The concept of the study originated from JK and was reviewed and extended by RA, while the design was put together by RSV and SG and reviewed by JK and RA. SG and CMI collected and prepared the input data. SG, OS, DS and TB performed the acquisition and post-processing of data. The data were analyzed and interpreted by RSV, SG, CMI and JK. The manuscript was written by RSV and SG in cooperation with JK and CMI, and reviewed by all authors.

Authors' information

Radka Svobodová Vařeková and Stanislav Geidl wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Acknowledgements

This work was supported by the Ministry of Education of the Czech Republic (LH13055), by the European Community's Seventh Framework Programme (CZ.1.05/1.1.00/02.0068 to J.K. and R.S.V.) from the European Regional Development Fund and by the EU Seventh Framework Programme under the "Capacities" specific programme (Contract No. 286154 to J.K.). C.M.I. and D.S. would like to thank Brno City Municipality for the financial support provided to them through the program Brno Ph.D. Talent. This work was also supported in part by NIH grants R01 GM071872, U01 GM094612, and U54 GM094618 to R.A.. The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly appreciated. Also the access to the CERIT-SC computing facilities provided under the programme Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, reg. no. CZ. 1.05/3.2.00/08.0144 is acknowledged.

Author details

[1] National Centre for Biomolecular Research, Faculty of Science and CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic. [2] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, 9500 Gilman Drive, MC 0657, San Diego, USA.

Received: 16 November 2012 Accepted: 27 March 2013
Published: 10 April 2013

References

1. Ishihama Y, Nakamura M, Miwa T, Kajima T, Asakawa N: **A rapid method for pK_a determination of drugs using pressure-assisted capillary electrophoresis with photodiode array detection in drug discovery.** *J Pharm Sci* 2002, **91**(4):933–942.
2. Babić S, Horvat A J, Pavlović D M, Kaštelan-Macan M: **Determination of pK_a values of active pharmaceutical ingredients.** *TrAC* 2007, **26**(11):1043–1061.
3. Manallack D: **The pK$_a$ distribution of drugs: application to drug discovery.** *Perspect Med Chem* 2007, **1:**25–38.
4. Wan H, Ulander J: **High-throughput pK_a screening and prediction amenable for ADME profiling.** *Expert Opin Drug Metabx Toxicol* 2006, **2:**139–155.
5. Cruciani G, Milletti F, Storchi L, Sforna G, Goracci L: **In silico pK_a prediction and ADME profiling.** *Chem Biodivers* 2009, **6**(11): 1812–1821.
6. Comer J, Tam K: *Pharmacokinetic Optimization in Drug Research: Biological, Physicochemical, and Computational Strategies*. Switzerland: Wiley-VCH. Verlag Helvetica Chimica Acta, Postfach, CH-8042 Zürich; 2001.
7. Klebe G: **Recent developments in structure-based drug design.** *J Mol Med* 2000, **78:**269–281.
8. Lee AC, Crippen GM: **Predicting pK_a.** *J Chem Inf Model* 2009, **49:**2013–2033.
9. Rupp M, Körner R, Tetko IV: **Predicting the pK$_a$ of small molecules.** *Comb Chem High Throughput Screen* 2010, **14**(5):307–327.
10. Fraczkiewicz R: *In Silico Prediction of Ionization, Volume 5*. Oxford: Elsevier; 2006.
11. Ho J, Coote M: **A universal approach for continuum solvent pK_a calculations: Are we there yet?** *Theor Chim Acta* 2010, **125**(1–2): 3–21.
12. Clark J, Perrin DD: **Prediction of the strengths of organic bases.** *Q ReV Chem Soc* 1964, **18:**295–320.
13. Perrin DD, Dempsey B, Serjeant EP: *pK_a Prediction for Organic Acids and Bases*. New York: Chapman and Hall; 1981.
14. Blower PE, Cross KP: **Decision tree methods in pharmaceutical research.** *Curr Top Med Chem* 2006, **6:**31–39.
15. Liptak MD, Gross KC, Seybold PG, Feldgus S, Shields G: **Absolute pK_a determinations for substituted phenols.** *J Am Chem Soc* 2002, **124:**6421–6427.
16. Toth AM, Liptak MD, Phillips DL, Shields GC: **Accurate relative pK_a calculations for carboxylic acids using complete basis set and Gaussian-n models combined with continuum solvation methods.** *J Chem Phys* 2001, **114:**4595–4606.
17. Hagan MT, Demuth HB, Beale M: *In Neural, Network Design*. Boston: PWS, MA; 1996.
18. Jelfs S, Ertl P, Selzer P: **Estimation of pK_a for druglike compounds using semiempirical and information-based descriptors.** *J Chem Inf Model* 2007, **47:**450–459.
19. Dixon SL, Jurs PC: **Estimation of pK_a for organic oxyacids using calculated atomic charges.** *J Comput Chem* 1993, **14:**1460–1467.
20. Zhang J, Kleinöder T, Gasteiger J: **Prediction of pK_a values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors.** *J Chem Inf Model* 2006, **46:**2256–2266.
21. Citra MJ: **Estimating the pK_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods.** *Chemosphere* 1999, **1:**191–206.
22. Gross KC, Seybold PG, Hadad CM: **Comparison of different atomic charge schemes for predicting pK_a variations in substituted anilines and phenols.** *Int J Quantum Chem* 2002, **90:**445–458.
23. Kreye WC, Seybold PG: **Correlations between quantum chemical indices and the pK_as of a diverse set of organic phenols.** *Int J Quantum Chem* 2009, **109:**3679–3684.
24. Svobodová Vařeková R, Geidl S, Ionescu CM, Skřehota O, Kudera M, Sehnal D, Bouchal T, Abagyan R, Huber HJ, Koča J: **Predicting pK_a values of substituted phenols from atomic charges: Comparison of different quantum mechanical methods and charge distribution schemes.** *J Chem Inf Model* 2011, **51**(8):1795–1806.
25. Mulliken RS: **Electronic structures of molecules XI. Electroaffinity, molecular orbitals and dipole moments.** *J Chem Phys* 1935, **3**(9):573–585.
26. Mulliken RS: **Criteria for construction of good self-consistent-field molecular orbital wave functions, and significance of LCAO-MO population analysis.** *J Chem Phys* 1962, **36**(12):3428–3439.
27. Lowdin PO: **On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals.** *J Chem Phys* 1950, **18**(3):365–375.
28. Reed AE, Weinstock RB, Weinhold F: **Natural-population analysis.** *J Chem Phys* 1985, **83**(2):735–746.
29. Bader RFW, Larouche A, Gatti C, Carroll MT, Macdougall PJ, Wiberg KB: **Properties of atoms in molecules - dipole-moments and transferability of properties.** *J Chem Phys* 1987, **87**(2):1142–1152.
30. Hirshfeld FL: **Bonded-atom fragments for describing molecular charge-densities.** *Theor Chim Acta* 1977, **44**(2):129–138.
31. Breneman CM, Wiberg KB: **Determining atom-centered monopoles from molecular electrostatic potentials - the need for high sampling density in formamide conformational-analysis.** *J Comput Chem* 1990, **11**(3):361–373.
32. Besler BH, Merz KM, Kollman PA: **Atomic charges derived from semiempirical methods.** *J Comput Chem* 1990, **11**(4):431–439.
33. Kelly CP, Cramer CJ, Truhlar DG: **Accurate partial atomic charges for high-energy molecules using class IV charge models with the MIDI! basis set.** *Theor Chem Acc* 2005, **113**(3):133–151.
34. Abraham RJ, Griffiths L, Loftus P: **Approaches to charge calculations in molecular mechanics.** *J Comput Chem* 1982, **3**(3):407–416.
35. Gasteiger J, Marsili M: **Iterative partial equalization of orbital electronegativity - a rapid access to atomic charges.** *Tetrahedron* 1980, **36**(22):3219–3228.
36. Cho KH, Kang YK, No KT, Scheraga HA: **A fast method for calculating geometry-dependent net atomic charges for polypeptides.** *J Phys Chem B* 2001, **105**(17):3624–3634.
37. Oliferenko AA, Pisarev SA, Palyulin VA, Zefirov NS: **Atomic charges via electronegativity equalization: Generalizations and perspectives.** *Adv Quantum Chem* 2006, **51:**139–156.
38. Shulga DA, Oliferenko AA, Pisarev SA, Palyulin VA, Zefirov NS: **Parameterization of empirical schemes of partial atomic charge calculation for reproducing the molecular electrostatic potential.** *Dokl Chem* 2008, **419:**57–61.
39. Mortier WJ, Ghosh SK, Shankar S: **Electronegativity equalization method for the calculation of atomic charges in molecules.** *J Am Chem Soc* 1986, **108:**4315–4320.
40. Rappe AK, Goddard WA: **Charge equilibration for molecular-dynamics simulations.** *J Phys Chem* 1991, **95**(8):3358–3363.
41. Nistor RA, Polihronov JG, Muser MH, Mosey NJ: **A generalization of the charge equilibration method for nonmetallic materials.** *J Chem Phys* 2006, **125**(9):094108–094118.
42. Czodrowski P, Dramburg I, Sotriffer CA, Klebe G: **Development, validation, and application of adapted PEOE charges to estimate pK_a values of functional groups in protein–ligand complexes.** *Proteins Struct Funct Bioinf* 2006, **65:**424–437.
43. Gieleciak R, Polanski J: **Modeling robust QSAR. 2. Iterative variable elimination schemes for CoMSA: Application for modeling benzoic acid pK_a values.** *J Chem Inf Model* 2007, **47:**547–556.
44. Svobodová Vařeková R, Jiroušková Z, Vaněk J, Suchomel S, Koča J: **Electronegativity equalization method: Parameterization and validation for large sets of organic, organohalogene and organometal molecule.** *Int J Mol Sci* 2007, **8:**572–582.
45. Baekelandt BG, Mortier WJ, Lievens JL, Schoonheydt RA: **Probing the reactivity of different sites within a molecule or solid by direct computation of molecular sensitivities via an extension of the electronegativity equalization method.** *J Am Chem Soc* 1991, **113**(18):6730–6734.
46. Jiroušková Z, Svobodová Vařeková R, Vaněk J, Koča J: **Electronegativity equalization method: Parameterization and validation for organic molecules using the Merz–Kollman–Singh charge distribution scheme.** *J Comput Chem* 2009, **30:**1174–1178.
47. Chaves J, Barroso JM, Bultinck P, Carbo-Dorca R: **Toward an alternative hardness kernel matrix structure in the Electronegativity Equalization Method (EEM).** *J Chem Inf Model* 2006, **46**(4):1657–1665.

48. Bultinck P, Langenaeker W, Lahorte P, De Proft, F, Geerlings P, Waroquier M, Tollenaere J: **The electronegativity equalization method I: Parametrization and validation for atomic charge calculations.** *J Phys Chem A* 2002, **106**(34):7887–7894.
49. Ouyang Y, Ye F, Liang Y: **A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules.** *Phys Chem* 2009, **11:**6082–6089.
50. Bultinck P, Vanholme R, Popelier PLA, De Proft, F, Geerlings P: **High-speed calculation of AIM charges through the electronegativity equalization method.** *J Phys Chem A* 2004, **108**(46):10359–10366.
51. Yang ZZ, Wang CS: **Atom-bond electronegativity equalization method. 1. Calculation of the charge distribution in large molecules.** *J Phys Chem A* 1997, **101:**6315–6321.
52. Menegon G, Loos M, Chaimovich H: **Parameterization of the electronegativity equalization method based on the charge model 1.** *J Phys Chem A* 2002, **106:**9078–9084.
53. Svobodová Vařeková R, Koča J: **Optimized and parallelized implementation of the electronegativity equalization method and the atom-bond electronegativity equalization method.** *J Comput Chem* 2006, **3:**396–405.
54. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JAJr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, et al.: *Gaussian 09, Revision E.01*. Wallingford: Gaussian, Inc.; 2004.
55. Keith TA: *AIMAll, Version 11.12.19*. USA: TK Gristmill Software, Overland Park KS; 2011. [aim.tkgristmill.com].
56. Habibi-Yangjeh A, Danandeh-Jenagharad M, Nooshyar M: **Application of artificial neural networks for predicting the aqueous acidity of various phenols using QSAR.** *J Mol Model* 2006, **12:**338–347.
57. Hanai T, Koizumi K, Kinoshita T, Arora R, Ahmed F: **Prediction of pK_a values of phenolic and nitrogen-containing compounds by computational chemical analysis compared to those measured by liquid chromatography.** *J Chromatogr A* 1997, **762:**55–61.
58. Tehan BG, Lloyd EJ, Wong MG, Pitt WR, Montana JG, Manallack DT, Gancia E: **Estimation of pK_a Using semiempirical molecular orbital methods. Part 1: Application to phenols and carboxylic acids.** *Quant Struct-Act Relat* 2002, **21:**457–472.
59. **NCI Open Database Compounds.** Retrieved from [http://cactus.nci.nih.gov/] on August 10, 2010.
60. Sadowski J, Gasteiger J: **From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders.** *Chem ReV* 1993, **93:**2567–2581.
61. Howard P, Meylan W: *Physical/Chemical Property Database (PHYSPROP)*. North Syracuse NY: Syracuse Research Corporation, Environmental Science Center; 1999.
62. Skřehota O, Svobodová Vařeková R, Geidl S, Kudera M, Sehnal D, Ionescu CM, Koča J: **QSPR designer – a program to design and evaluate QSPR models. Case study on pK_a prediction.** *J Cheminf* 2011, **3**(Suppl 1):P16.
63. Bultinck P, Langenaeker W, Lahorte P, De Proft, F, Geerlings P, Van Alsenoy, C, Tollenaere JP: **The electronegativity equalization method II: Applicability of different atomic charge schemes.** *J Phys Chem A* 2002, **106**(34):7895–7901.
64. Lemm S, Blankertz B, Dickhaus T, Müller KR: **Introduction to machine learning for brain imaging.** *NeuroImage* 2011, **56**(2):387–399.
65. Organisation for Economic Co-operation and Development: *Guidance Document on the Validation of (Quantitative)Structure-Activity Relationships [(Q)SAR] Models*. Paris: OECD; 2007. [http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en] (accessed April 6,2013).

doi:10.1186/1758-2946-5-18
Cite this article as: Svobodová Vařeková *et al.*: Predicting pK_a values from EEM atomic charges. *Journal of Cheminformatics* 2013 **5**:18.

Rapid calculation of accurate atomic charges for proteins via the Electronegativity Equalization Method

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

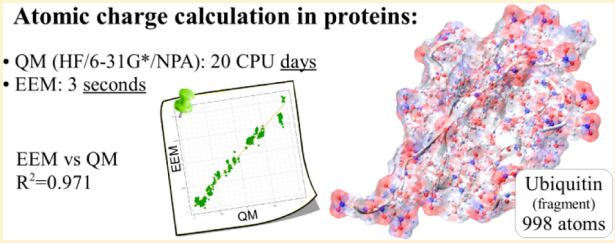
Rapid Calculation of Accurate Atomic Charges for Proteins via the Electronegativity Equalization Method

Crina-Maria Ionescu, Stanislav Geidl, Radka Svobodová Vařeková,* and Jaroslav Koča*

CEITEC—Central European Institute of Technology, and National Centre for Biomolecular Research, Faculty of Science, Masaryk University Brno, Kamenice 5, 625 00, Brno-Bohunice, Czech Republic

Supporting Information

ABSTRACT: We focused on the parametrization and evaluation of empirical models for fast and accurate calculation of conformationally dependent atomic charges in proteins. The models were based on the electronegativity equalization method (EEM), and the parametrization procedure was tailored to proteins. We used large protein fragments as reference structures and fitted the EEM model parameters using atomic charges computed by three population analyses (Mulliken, Natural, iterative Hirshfeld), at the Hartree−Fock level with two basis sets (6-31G*, 6-31G**) and in two environments (gas phase, implicit solvation). We parametrized and successfully validated 24 EEM models. When tested on insulin and ubiquitin, all models reproduced quantum mechanics level charges well and were consistent with respect to population analysis and basis set. Specifically, the models showed on average a correlation of 0.961, RMSD 0.097 e, and average absolute error per atom 0.072 e. The EEM models can be used with the freely available EEM implementation EEM_SOLVER.



INTRODUCTION

The concept of atomic point charges is well established in theoretical chemistry. Atomic point charges have played an important role in understanding and modeling chemical behavior by allowing the extraction and quantification of information stored in the molecular electron distribution of chemical compounds. Thus, atomic point charges have been used to estimate reactivity indices, dissociation constants, partition coefficients, the electrostatic contribution in molecular dynamics or docking studies, etc.[1−4] Virtual screening, the process by which potential drug candidates or targets are identified by screening huge libraries of compounds, employs atomic point charges, for instance, as parameters in the docking stage. Absorbtion, distribution, metabolism, and excretion profiling studies, involved in drug discovery programs, employ atomic point charges as descriptors useful in predicting various pharmacological properties. It is therefore desirable to have knowledge of the values of atomic charges.

While atomic charges are very intuitive, they are not physical observables. Nevertheless, atomic partial charges can be derived from physical observables, the values of which are most commonly obtained by quantum mechanical (QM) calculations. A remarkable fact is that a unique definition of atomic partial charges has yet to be accepted. As such, a score of methods have been developed to estimate the values of atomic charges, which is another indication that atomic charges are of great interest. Roughly, these methods can be classified as QM-based or empirical.

QM-based approaches first compute the wave function and subsequently the molecular electron density. In further, various population analyses can be performed in order to partition the molecular electron density into its atomic contributions. Some partitioning approaches are based on the wave function itself;[5−8] others on a wave function-dependent physical observable.[9−13] QM atomic charges can also be mapped to reproduce charge-dependent observables.[14] Along with physical soundness, the advantage of QM based approaches for charge calculation is their general applicability.

Empirical approaches, on the other hand, do not need the molecular wave function and instead employ a wide array of principles and intuitive derivations, coupled with cost-effective algorithms. Empirical approaches for atomic charge calculation can be divided into those which depend only on the 2D structure of the molecule,[15−18] and those which depend on the 3D structure of the molecule.[19−24] The majority of empirical approaches are based on partial or complete equalization of the electronegativity. Atomic charge calculation via empirical approaches is an extremely cost-effective alternative to QM-based methods, though appropriate parametrizations need to be performed before any empirical method can be used with confidence. Unlike QM based approaches, the applicability of empirical methods is many times limited to certain kinds of systems, either because of the intrinsic approximations that are made or by the parametrization procedure.

Although these atomic charge definitions differ in both the principles and algorithms they employ, many such definitions have already proven suitable for the prediction of relevant

Received: May 1, 2013
Published: August 22, 2013

   dx.doi.org/10.1021/ci400448n | J. Chem. Inf. Model. 2013, 53, 2548−2558

chemical phenomena. Following a statistical analysis of 25 charge definition schemes, Meister and Schwarz[25] concluded that the same physical principles govern all the schemes under study, and therefore, all charge definition schemes are relevant, despite the differences in scale.

In the present study, we focus on proteins, due to their capital importance on all aspects of cellular life. Charge transfer was found to be significant in protein folding and many biomolecular interactions,[4,26−28] and functionally linked to protein structural dynamics.[29] Cho et al.[4] have shown that using conformationally dependent, QM-level atomic charges is critical to identifying the correct binding partners in docking studies on various proteins. Later, Cho and Rinaldo[30] found that, especially in the case of metalloproteins, it is necessary to compute QM charges for as many atoms as possible close to the metal ion in order to allow the correct prediction of ligand binding modes. Wallin et al.[31] emphasized the need to generalize and automatize the charge assignment procedure and include this procedure in the work flow of ligand preparation for binding free energy calculations.

In the light of the above-mentioned findings, it is clear that there is a need for a fast, accurate, and accessible procedure to calculate atomic charges in proteins. Nonetheless, an affordable, accurate, and highly available solution to handle large biomolecular systems at the QM level has not yet been found. Clearly, direct QM calculations are not yet a feasible option for charge calculation in proteins, and one must rely on some empirical method. In the present study, we focus on the electronegativity equalization method (EEM) as a method for rapid and accurate estimation of atomic partial charges.[20,32−45] EEM is fast, has a straightforward implementation, and is generally applicable and sensitive to a molecule's 3D structure. EEM has been successfully applied to zeolites, small organic molecules, and polypeptides.[46−50] A recent EEM investigation proved useful in mapping the allosteric activation mechanism of two apoptotic proteins.[44]

Before an empirical model can be used with confidence, appropriate parameters must be determined, i.e., the model must be parametrized. EEM models are parametrized by fitting the model parameters to a set of reference data usually made up of atomic charges obtained by various QM-based approaches. A QM-based charge calculation approach is characterized by the setup of the wave function calculation (theory level, basis set, environment), as well as by the population analysis (PA) used to partition the molecular electron density. We will refer to the sum of these characteristics as the "type" of charge produced by the QM approach. The maximum accuracy and potential application of any EEM model is given by the type of charge used during its parametrization. Previous works have parametrized EEM mainly for inorganic or small organic molecules, and especially for drug-like molecules,[20,33−45] but very few with specific focus on biomolecules. Some EEM models have been successfully tested on peptides of various size,[20,41] but these models cover only two types of charges and contain no parameters for sulfur, an element commonly found in proteins. Some EEM models were successfully tested on large protein fragments,[44] but these models only cover one type of charge.

Thus, the goal of the present study was to provide wider coverage and evaluation of the EEM approach for proteins. In this study, we have developed and validated 24 EEM models specifically tailored to cover 12 types of charges in proteins. Our reference data came from QM calculations at the HF level with two basis sets (6-31G*, 6-31G**), in the gas-phase and implicit solvent. Three atomic charge definitions were considered (Mulliken, Natural, iterative Hirshfeld). The accuracy of the proposed models was evaluated on insulin and ubiquitin.

THEORY AND METHODS

Electronegativity Equalization Method. The electronegativity equalization method (EEM)[20] enables the determination of atomic charges that are sensitive to the molecule's topology and three-dimensional structure. EEM is based on the electronegativity equalization principle,[51] which has received theoretical grounding within the density functional theory,[52,53] and which states that the electronegativity of all atoms is equalized throughout a molecule:

$$X_1 = X_2 = \ldots X_i = \ldots = \bar{X} \tag{1}$$

Within EEM, the electronegativity X_i of each atom i in a molecule can be approximated as a linear function of several terms:

$$X_1 = (X_i^0 + \Delta X_i) + 2(\eta_i^0 + \Delta\eta_i)q_i + k\sum_{i\neq j}\frac{q_j}{r_{ij}} \tag{2}$$

The first term is the electronegativity of the isolated atom X_i^0, empirically corrected for the presence of the molecular environment (ΔX_i). The second term is the product between the charge of the atom q_i and the hardness of the isolated atom η_i^0, empirically corrected for the presence of the molecular environment ($_\Delta\eta_i$). The last term $k\sum_{i\neq j}(q_j/r_{ij})$ accounts for the electrostatic interaction with every other charged atom j in the molecule. k is an adjusting factor first introduced by Yang and Shen.[54] Setting $A_i = X_i^0 + \Delta X_i$ and $B_i = 2(\eta_i^0 + \Delta\eta_i)$, the molecular electronegativity can be formally expressed as

$$\bar{X} = A_i + B_i q_i + k\sum_{i\neq j}\frac{q_j}{r_{ij}} \tag{3}$$

Additionally, the total molecular charge Q is the sum of all partial atomic charges q_i:

$$\sum q_i = Q \tag{4}$$

Taken all together, eqs 1, 3, and 4 can be expressed as a system of equations from which the partial atomic charges q_i and the molecular electronegativity $\bar{X}$ can be calculated, provided that the rest of the terms (Q, r_{ij}, k, A_i, B_i) are known:

$$\begin{pmatrix} B_1 & \frac{k}{r_{1,2}} & \ldots & \frac{k}{r_{1,N}} & -1 \\ \frac{k}{r_{2,3}} & B_2 & \ldots & \frac{k}{r_{2,N}} & -1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \frac{k}{r_{N,1}} & \frac{k}{r_{N,2}} & \ldots & B_N & -1 \\ 1 & 1 & \ldots & 1 & 0 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ \ldots \\ q_N \\ \bar{X} \end{pmatrix} = \begin{pmatrix} -A_1 \\ -A_2 \\ \ldots \\ A_N \\ Q \end{pmatrix} \tag{5}$$

This formalism estimates atomic charges via a set of coupled linear equations which can be efficiently solved by a Gaussian elimination procedure.[55] The computational complexity is $\theta(N^3)$, where N is the number of atoms in the molecule.

dx.doi.org/10.1021/ci400448n | J. Chem. Inf. Model. 2013, 53, 2548−2558

Table 1. Atomic Composition of the Data Sets Used in This Study

	Nr molecules	Nr atoms	H	C	N	O	S	Ca
training set	41	40142	19879	11912	3188	4954	148	61
insulin	1	802	397	248	63	88	6	
ubiquitin	1	998	505	306	88	99		

EEM Model Parameterization Theory. Given the three-dimensional structure of the molecule and its total charge, partial charges for all atoms in the molecule can be calculated only if the values of the EEM parameters (k, $\langle A_i, B_i \rangle$) are known. Otherwise, the EEM model needs to be parametrized by fitting against reference data (training data set). We employed QM atomic charges to fit the EEM parameters given by the electronegativity and hardness contributions $A_i = X_i^0 + \Delta X_i$ and $B_i = 2(\eta_i^0 + \Delta\eta_i)$, respectively, as well as the value of the parameter k present in this formalism. The EEM parametrization methodology employed here was described by Svobodová Vařeková et al.,[37] and its principles are given below.

For the purpose of EEM model parametrization, eq 3 can be rearranged as a linear equation in A_i and B_i for each atom i in the system:

$$A_i + B_i q_i = \bar{X} - k \sum_{i \neq j} \frac{q_j}{r_{ij}} \tag{6}$$

For a given value of the parameter k, sets of eqs 6 can be grouped together according to the kind of atom they refer to. The values of atomic charges q_i and interatomic distances r_{ij} are taken from the training data set, and the value of the molecular electronegativity $\bar{X}$ can be calculated as an average of the isolated atom electronegativities X_i^0. Under these circumstances, each group of linear equations becomes an overdetermined system of equations, enabling the determination of a set of parameters $(A, B)_m$ for each atom type m by least-squares minimization. The classification of atoms into types can be done according to various criteria, such as chemical element, hybridization, binding partners, etc. Once a set of parameters $(A, B)_m$ has been obtained for all M atom types, for all given values of k, it is possible to determine the optimal EEM parameter set $k[(A, B)_1, (A, B)_2, ..., (A, B)_M]$, in further denoted simply as $k[(A, B)_m]_1^M$, as the parameter set which produces the best EEM model in the internal validation step (see below).

EEM Model Validation Theory. The accuracy of an EEM model in reproducing the reference data (here, QM atomic charges) can be evaluated by internal and external validation. In the internal validation step, the EEM model is evaluated for its ability to reproduce the reference data which was used during its parametrization. In other words, EEM is first validated on the training data set. Next, in the external validation step, the EEM model is evaluated for its ability to reproduce reference data which was not used during its parametrization. In other words, EEM is also validated on test molecules. The correlation between the reference QM charges and predicted EEM charges can be assessed by various indicators.

The first indicator used in the present study was the average correlation coefficient (squared Pearson's correlation coefficient), computed for each molecule, and averaged over all molecules in the data set:

$$R_{\text{avg}} = \frac{\sum_{I=1}^{N} \left(\frac{\sum_{i=1}^{n_I} (q_i^{\text{QM}} - \overline{q_I^{\text{QM}}})(q_i^{\text{EEM}} - \overline{q_I^{\text{EEM}}})}{(n_I - 1)\sigma_I^{\text{QM}}\sigma_I^{\text{EEM}}} \right)}{N} \tag{7}$$

where the index $i = 1...n_I$ described all atoms in molecule I, $\overline{q_I^{\text{QM}}}$ and $\overline{q_I^{\text{EEM}}}$ represented average atomic charges in molecule I, σ_I^{QM} and σ_I^{EEM} were standard deviations of the atomic charges in molecule I, n_I was the number of atoms in molecule I, and N was the number of molecules in a given set.

The second indicator was the root-mean-square deviation, computed for each molecule, and averaged over all molecules in the data set:

$$\text{RMSD}_{\text{avg}} = \frac{\sum_{I=1}^{N} \sqrt{\frac{\sum_{i=1}^{n_I} (q_i^{\text{QM}} - q_i^{\text{EEM}})^2}{n_I}}}{N} \tag{8}$$

The third indicator was the average absolute difference, computed for each molecule and averaged over all molecules in the data set:

$$D_{\text{avg}} = \frac{\sum_{I=1}^{N} \sqrt{\frac{\sum_{i=1}^{n_I} (q_i^{\text{QM}} - q_i^{\text{EEM}})^2}{n_I}}}{N} \tag{9}$$

Reference Structures—Training Data Set. The reference molecules used in the model parametrization and internal validation steps were 41 fragments of proteins from the Protein Data Bank (PDB), whose structures had been determined by X-ray crystallography or solution state NMR experiments. Protein fragments, rather than small molecules, were chosen as reference structures since they reflect the complex nature of proteins as long, non-neutral molecular chains with complex 3D assembly.

The 41 reference structures consisted of amino acid chains, water molecules, and calcium ions, and were obtained from the 3D structure of their parent protein using the program Triton.[56] Hydrogen atoms were added for all crystal structures to satisfy the missing valences. The protonation states of the amino acid residues were +1 for Arg, Lys, and His and charged amino ends of the polypeptide chains, −1 for Glu, Asp, and charged carboxyl ends of the polypeptidic chains, 0 for the rest. Only the first structural model was used in the case of NMR structures.

Reference Structures—Test Molecules. Two small proteins were used in the external validation step, namely insulin (PDB ID 3E7Y[57]) and ubiquitin (PDB ID 1UBQ[58]). The nature of the calculations imposed some limitations to the size of the systems used for testing. Thus, only one of the two insulin monomers was kept, containing chains C and D, along with a few water molecules cocrystallized with the protein. None of the Cl and Zn ions found cocrystallized with insulin were kept, as no EEM parameters were obtained for these ions. In the case of ubiquitin, the first 14 residues had to be removed in order to reduce the system to a manageable size.

dx.doi.org/10.1021/ci400448n | J. Chem. Inf. Model. 2013, 53, 2548−2558

Table 2. Overview of the 12 Types of QM Calculations Performed, and 24 EEM Models Obtained in This Study

EEM model characteristics						
	QM scheme			training data sets		
atom type classification	PA	basis set	environment	Nr molecules	Nr atoms	EEM model
E	MPA	6-31G*	gas phase	41	40142	E-MPA/6-31G*/gas
			PCM	41	40142	E-MPA/6-31G*/PCM
		6-31G**	gas phase	41	40142	E-MPA/6-31G**/gas
			PCM	41	40142	E-MPA/6-31G**/PCM
	NPA	6-31G*	gas phase	41	40142	E-NPA/6-31G*/gas
			PCM	40	39061	E-NPA/6-31G*/PCM
		6-31G**	gas phase	41	40142	E-NPA/6-31G**/gas
			PCM	40	39061	E-NPA/6-31G**/PCM
	HiI	6-31G*	gas phase	41	40142	E-HiI/6-31G*/gas
			PCM	38	36875	E-HiI/6-31G*/PCM
		6-31G**	gas phase	41	40142	E-HiI/6-31G**/gas
			PCM	40	39061	E-HiI/6-31G**/PCM
EX	MPA	6-31G*	gas phase	41	40142	EX-MPA/6-31G*/gas
			PCM	41	40142	EX-MPA/6-31G*/PCM
		6-31G**	gas phase	41	40142	EX-MPA/6-31G**/gas
			PCM	41	40142	EX-MPA/6-31G**/PCM
	NPA	6-31G*	gas phase	41	40142	EX-NPA/6-31G*/gas
			PCM	40	39061	EX-NPA/6-31G*/PCM
		6-31G**	gas phase	41	40142	EX-NPA/6-31G**/gas
			PCM	40	39061	EX-NPA/6-31G**/PCM
	HiI	6-31G*	gas phase	41	40142	EX-HiI/6-31G*/gas
			PCM	38	36875	EX-HiI/6-31G*/PCM
		6-31G**	gas phase	41	40142	EX-HiI/6-31G**/gas
			PCM	40	39061	EX-HiI/6-31G**/PCM

An overview of the composition of all molecules used for EEM model parametrization and validation is given in Table 1. The 3D structures of these fragments are available as Supporting Information in PDB format.

Reference QM Atomic Charges. For each reference structure, QM atomic charges were obtained from the Mulliken Population Analysis (MPA),[5,6] Natural Population Analysis (NPA),[8] and iterative Hirshfeld analysis (HiI)[11] performed at the Hartree−Fock (HF) theory level with 6-31G* and 6-31G** basis sets in the gas phase and polarizable continuum model (PCM).[59] These three population analyses were chosen because they are known to work well with EEM.[34,50] Moreover, these population analyses can be performed by currently available software tools for very large protein fragments that we used as training and test molecules.

A total of 12 QM reference data sets were thus obtained: MPA/6-31G*/gas, MPA/6-31G*/PCM, MPA/6-31G**/gas, MPA/6-31G**/PCM, NPA/6-31G*/gas, NPA/6-31G*/PCM, NPA/6-31G**/gas, NPA/6-31G**/PCM, HiI/6-31G*/gas, HiI/6-31G*/PCM, HiI/6-31G**/gas, HiI/6-31G**/PCM. All QM single point energy calculations and MPA were performed using Gaussian 09,[60] while NPA was performed using the NBO program,[61] and HiI was performed using the HiPart program.[62] The values of all QM atomic charges obtained in this study are included as Supporting Information in csv format.

EEM Models. Two classifications of atoms into atom types were employed here. One classification was based on chemical elements (denoted "E"), and the other on chemical elements and maximum bond multiplicity for each atom (denoted "EX", so that, for example, "O1" indicates sp^3 hybridized oxygen and "O2", sp^2 hybridized oxygen). Both atom classification approaches were applied to all 12 QM reference data sets.

Within a given atom classification, the parameter k was sampled as discrete values on several intervals. For each discrete value of k, the set of parameters $(A, B)_m$ were obtained for all atom types in the given classification. For each atom type m, the parameters $(A, B)_m$ were determined by least-squares minimization, as the values of all other variables were known: the interatomic distances were calculated from the 3D atomic coordinates of the reference structures, the atomic charges were calculated by the 12 QM schemes described above, and the value of the average electronegativity for each molecule was calculated as the harmonic average of the electronegativities of the constituent atoms[63] $\bar{X} = n(\sum_{i=1}^{n}(1/X_i^0))^{-1}$, where n is the number of atoms in the molecule, and the values of X_i^0 correspond to Pauling electronegativities.[64,65] Thus, within each of the two atom type classifications, for each discrete value of k, an EEM model described by the parameter set $k[(A, B)_m]_1^M$ was obtained. Internal validation was then performed for all such EEM models. Namely, each model was used to predict the EEM charges of the reference molecules in the training data set. The atomic charges predicted by the EEM model were compared against the reference QM atomic charges used during the model parametrization. The EEM model which gave the highest R_{avg} between the EEM predicted charges and the QM reference charges for the whole training set was designated as the final EEM model within the given atom type classification.

Finally, 12 EEM models were obtained for the first atom classification E, based on chemical elements: E-MPA/6-31G*/gas, E-MPA/6-31G*/PCM, E-MPA/6-31G**/gas, E-MPA/6-31G**/PCM, E-NPA/6-31G*/gas, E-NPA/6-31G*/PCM, E-NPA/6-31G**/gas, E-NPA/6-31G**/PCM, E-HiI/6-31G*/gas, E-HiI/6-31G*/PCM, E-HiI/6-31G**/gas, E-HiI/

dx.doi.org/10.1021/ci400448n | J. Chem. Inf. Model. 2013, 53, 2548−2558

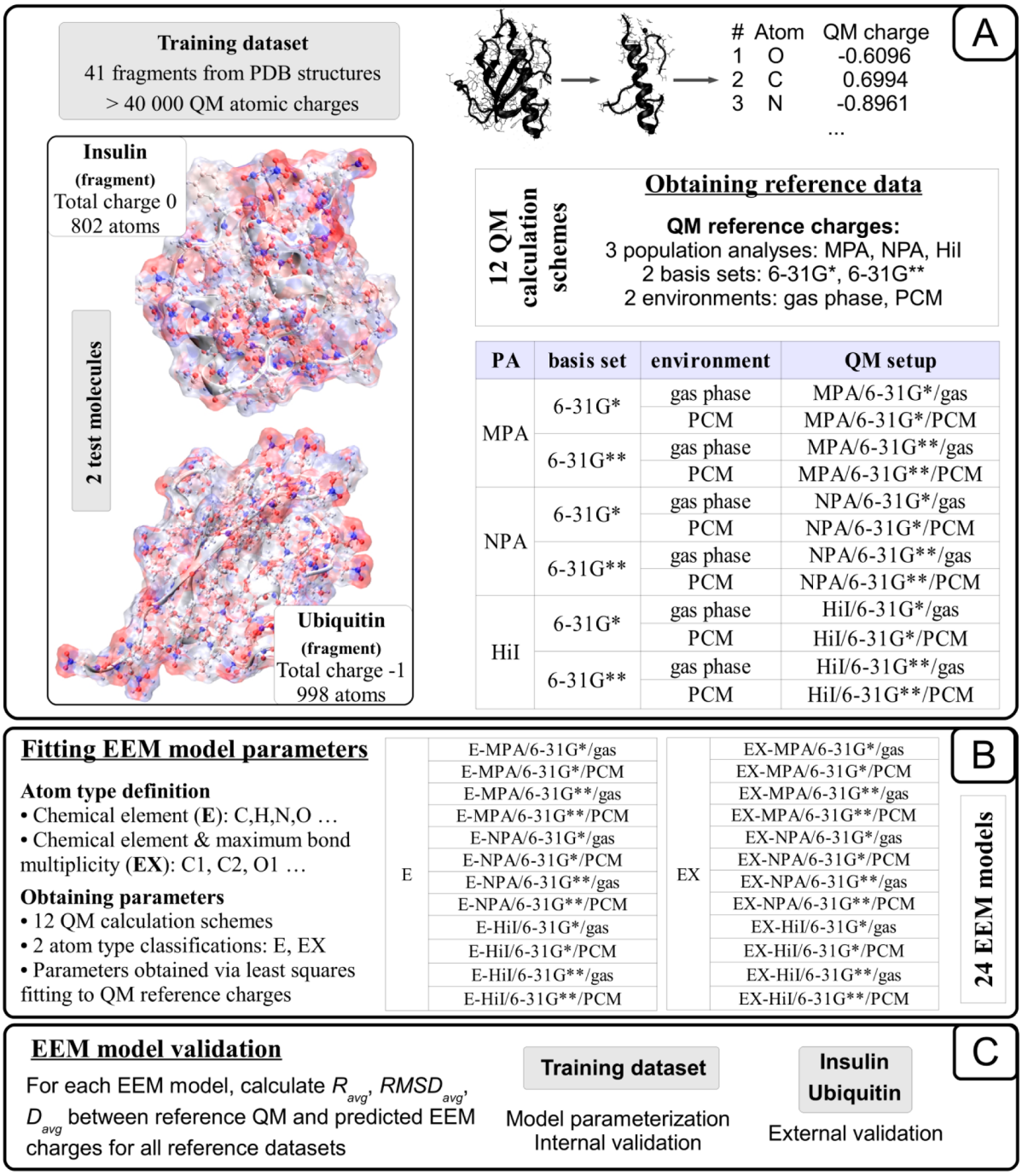


Figure 1. Flowchart of parametrization of EEM models for calculating partial atomic charges in proteins. (A) The reference data used in this study consisted of QM atomic charges for one training data set of protein fragments, and two test molecules (insulin and ubiquitin). Due to the system size limitations imposed by the QM calculation, only one insulin monomer was used, and the first 14 residues were removed from the ubiquitin structure. In total, 12 different QM atomic charge calculation schemes were used. (B) For each of the 12 QM charge schemes, two atom type classification approaches were employed for EEM model parametrization. Parameters for 24 EEM models were fitted onto the QM charges of the training data set. (C) Each EEM model was subjected to internal and external validation by comparing the EEM charges against reference QM charges of the training and test molecules, respectively. Performance was evaluated by 3 indicators, namely the average correlation coefficient (R_{avg}), the average root-mean-square deviation ($RMSD_{avg}$), and the average absolute difference (D_{avg}).

6-31G**/PCM. Additionally, 12 EEM models were obtained for the second atom classification EX, based on chemical elements and maximum bond multiplicity: EX-MPA/6-31G*/gas, EX-MPA/6-31G*/PCM, EX-MPA/6-31G**/gas, EX-MPA/6-31G**/PCM, EX-NPA/6-31G*/gas, EX-NPA/6-31G*/PCM, EX-NPA/6-31G**/gas, EX-NPA/6-31G**/PCM, EX-HiI/6-31G*/gas, EX-HiI/6-31G*/PCM, EX-HiI/6-31G**/gas, EX-HiI/6-31G**/PCM. All 24 EEM models

dx.doi.org/10.1021/ci400448n | J. Chem. Inf. Model. 2013, 53, 2548−2558

Table 3. Validation of the 24 EEM Models Obtained in This Study[a]

EEM model	internal validation (training data set)			external validation (insulin)			external validation (ubiquitin)		
	R_{avg}	$RMSD_{avg}$ [e]	D_{avg} [e]	R_{avg}	$RMSD_{avg}$ [e]	D_{avg} [e]	R_{avg}	$RMSD_{avg}$ [e]	D_{avg} [e]
E-MPA/6-31G*/gas	0.975	0.079	0.062	0.972	0.080	0.064	0.969	0.083	0.065
E-MPA/6-31G*/PCM	0.976	0.080	0.065	0.973	0.081	0.066	0.972	0.081	0.066
E-MPA/6-31G**/gas	0.977	0.067	0.053	0.976	0.065	0.052	0.973	0.069	0.054
E-MPA/6-31G**/PCM	0.977	0.068	0.053	0.976	0.066	0.052	0.973	0.071	0.054
E-NPA/6-31G*/gas	0.975	0.083	0.064	0.971	0.086	0.066	0.971	0.086	0.066
E-NPA/6-31G*/PCM	0.975	0.092	0.067	0.971	0.091	0.065	0.973	0.087	0.063
E-NPA/6-31G**/gas	0.976	0.084	0.065	0.971	0.086	0.067	0.971	0.086	0.066
E-NPA/6-31G**/PCM	0.975	0.093	0.068	0.971	0.092	0.066	0.973	0.088	0.063
E-HiI/6-31G*/gas	0.965	0.125	0.092	0.965	0.114	0.083	0.969	0.113	0.080
E-HiI/6-31G*/PCM	0.948	0.123	0.087	0.948	0.116	0.081	0.958	0.108	0.075
E-HiI/6-31G**/gas	0.967	0.121	0.088	0.965	0.111	0.080	0.970	0.109	0.077
E-HiI/6-31G**/PCM	0.962	0.128	0.094	0.961	0.119	0.087	0.965	0.116	0.084
EX-MPA/6-31G*/gas	0.976	0.076	0.056	0.970	0.082	0.059	0.974	0.076	0.058
EX-MPA/6-31G*/PCM	0.979	0.073	0.056	0.976	0.075	0.057	0.975	0.076	0.058
EX-MPA/6-31G**/gas	0.975	0.070	0.053	0.969	0.073	0.056	0.971	0.072	0.055
EX-MPA/6-31G**/PCM	0.961	0.090	0.069	0.958	0.087	0.067	0.959	0.088	0.065
EX-NPA/6-31G*/gas	0.979	0.077	0.059	0.974	0.081	0.062	0.975	0.080	0.061
EX-NPA/6-31G*/PCM	0.978	0.079	0.063	0.975	0.080	0.063	0.974	0.081	0.063
EX-NPA/6-31G**/gas	0.979	0.077	0.062	0.975	0.081	0.062	0.975	0.080	0.061
EX-NPA/6-31G**/PCM	0.980	0.076	0.062	0.977	0.078	0.062	0.976	0.079	0.062
EX-HiI/6-31G*/gas	0.919	0.148	0.109	0.913	0.146	0.107	0.934	0.130	0.095
EX-HiI/6-31G*/PCM	0.910	0.166	0.118	0.911	0.156	0.110	0.930	0.142	0.099
EX-HiI/6-31G**/gas	0.910	0.156	0.118	0.903	0.154	0.117	0.927	0.137	0.103
EX-HiI/6-31G**/PCM	0.917	0.161	0.117	0.912	0.156	0.112	0.931	0.141	0.101

[a]Internal validation denotes validation for the molecules in the training data set, while external validation denotes validation on the test molecules. Statistical descriptors comprising the average correlation coefficient (R_{avg}), the average root mean square deviation ($RMSD_{avg}$), and the average absolute difference (D_{avg}) are given. All quantities are given in elementary charges (1 e $\sim$ 1.602×10^{-19} coulombs).

were further subjected to external validation, whereby they were evaluated for their ability to predict the atomic charges of the test molecules insulin and ubiquitin.

An overview of all QM calculations performed, and all EEM models obtained is given in Table 2. An overview of the entire EEM model parametrization procedure employed in this study is given in Figure 1. The parametrization and validation of the EEM models were done using an in-house program. The visualization of structural models in various representations was performed using VMD.[66] The values of all EEM atomic charges obtained in this study are included as Supporting Information in csv format. The parameters of all 24 EEM models developed in this study are given in Table S1 of the Supporting Information.

RESULTS AND DISCUSSION

Performance of the EEM Models. In this study, we obtained 24 EEM models (Figure 1, Table 2). Each EEM model was validated with respect to R_{avg}, $RMSD_{avg}$, and D_{avg} between reference QM charges and predicted EEM charges. The results of the internal and external validation procedures are given in Table 3 and Figure 2 (see also Supporting Information Table S2).

Overall, there is good agreement between QM and EEM charges, suggesting that these EEM models can be used reliably for rapid atomic charge calculation in proteins. With respect to internal validation (i.e., validation against the training data sets), all 24 models have R_{avg} over 0.91, while 19 models have R_{avg} over 0.95, and 15 models have R_{avg} over 0.97. All models have $RMSD_{avg}$ less than 0.17 e and D_{avg} less than 0.12 e, while 16 models have $RMSD_{avg}$ less than 0.10 e and D_{avg} less than 0.07 e. With respect to external validation on insulin and ubiquitin, all EEM models perform comparably as in the internal validation step, suggesting high transferability. Specifically, all 24 models have R_{avg} over 0.9, while 19 models have R_{avg} over 0.95, and 14 models have R_{avg} over 0.97. All models have $RMSD_{avg}$ less than 0.16 e and D_{avg} less than 0.12 e, while 16 models have $RMSD_{avg}$ less than 0.10 e and D_{avg} less than 0.07 e. The values obtained for the majority of models are comparable to previous EEM parametrizations published in the literature.

There are diverging trends in literature regarding the level of detail that should be included in the atom type classification used for EEM model parametrization. One direction supports the use of general atom types,[33] while the other direction supports the use of very detailed atom types.[40,41] Our results in Table 3 (see also Supporting Information Table S2) suggested that the finer grained atom classification EX only modestly improved the accuracy compared to the classification E based on chemical elements alone, and not for all models. In any case, both atom classifications can provide satisfactory results for proteins, but we rather support the use of general atom types to avoid the extra step and possible errors associated with assigning hybridizations. Thus, in further we discuss only the results for the models based on general atom types E, while the complete results can be found in Supporting Information Table S2.

Accuracy of the EEM Models. A more detailed view of the accuracy that can be expected for the EEM models developed in this study is given in Figure 3. Approximately 50% of the predicted values are expected to contain an error of less than 0.05 e, while approximately 30% of the predicted values are expected to contain an error between 0.05 and 0.1 e. Finally, approximately 5% of the values are expected to contain an error larger than 0.2 e.

dx.doi.org/10.1021/ci400448n | J. Chem. Inf. Model. 2013, 53, 2548−2558

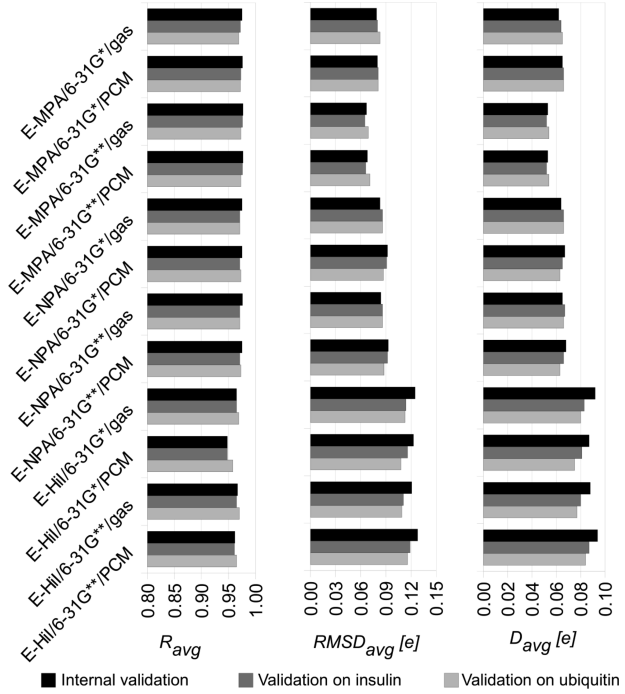


Figure 2. Validation of EEM models by comparing the predicted EEM atomic charges against the reference QM atomic charges. Internal validation denotes validation for the molecules in the training data set. Statistical descriptors comprising the average correlation coefficient (R_{avg}), the average root-mean-square deviation ($RMSD_{avg}$), and the average absolute difference (D_{avg}) are given. All quantities are given in elementary charges (1 e $\sim 1.602 \times 10^{-19}$ coulombs). The names of the EEM models encode the QM scheme and atom classification used in their parametrization. Only the models based on general atom types (E) are given here, while the complete results can be found in Table 3 (see also Supporting Information Table S2). Good agreement between QM and EEM charges was found for all data sets, as R_{avg} is close to 1, and $RMSD_{avg}$ and D_{avg} are minimal.

With respect to the practical implications of employing EEM models for calculating atomic partial charges, it is worth discussing the expected accuracy of the EEM models in reproducing not only QM atomic charges, but also various quantities derived from these charges. Since it is not possible to give a general evaluation of the expected accuracy of the EEM models in all possible practical applications, we focus here on two common uses of atomic charges, namely electrostatic potential (ESP) and docking calculations.

Figure 4 (see also Supporting Information Table S3) shows that the ESP computed from EEM charges deviates from the ESP computed from QM charges on average by 0.0071 au for insulin and 0.0058 au for ubiquitin, respectively; whereas, the average correlation coefficient is 0.9 for insulin and 0.87 for ubiquitin.

To evaluate the accuracy of EEM charges in a typical docking calculation, we performed the blind docking of glycerol onto the ubiquitin fragment previously used in our study as a test molecule for the validation of EEM models. Glycerol was chosen as a potential ligand because it has been found to stabilize the native state of ubiquitin.[68] The ubiquitin fragment was used as a receptor in favor of the native ubiquitin dimer because of the size restrictions imposed by the reference QM calculations. Figure 5 (see also Supporting Information Table S4) shows that the docking results obtained using QM charges are well reproduced using the EEM charges given by the corresponding EEM model. The EEM binding pose differs by 0.07 kcal/mol and an RMSD of 0.131 Å from the binding pose given using QM charges. By comparison, using Gasteiger−Marsili or AMBER ff94[69] charges on ubiquitin produces different binding poses. Specifically, the binding pose given by Gasteiger−Marsili charges differs from that given by QM charges by 0.99 kcal/mol and an RMSD of 3.244 Å. The binding pose given by AMBER ff94 charges differs by 1.57 kcal/mol and an RMSD of 3.235 Å.

Sensitivity of the EEM Models. We have evaluated the performance and predicted the accuracy of the EEM models developed in this study. It is also necessary to translate the meaning of this accuracy with respect to the ability of the EEM models to differentiate between types of charges, since many times different modeling applications rely on different types of charges. For example, in the particular case of pK_a prediction via

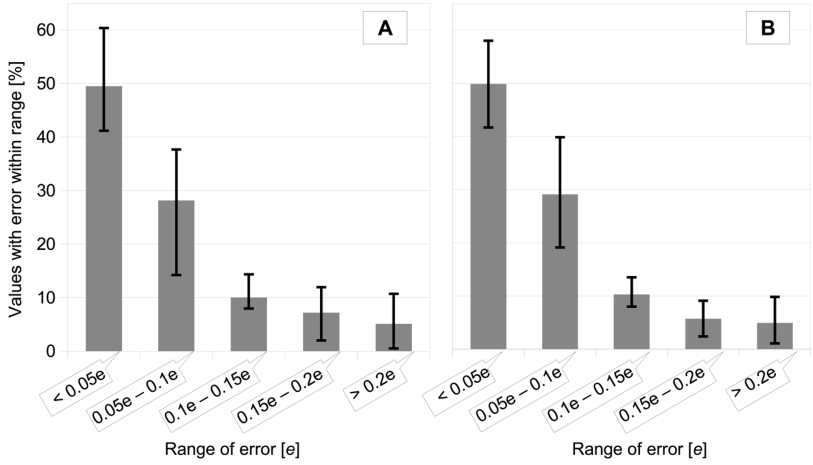


Figure 3. Expected accuracy of the EEM models developed in this study, measured as the percentage of atoms for which the error of the EEM vs QM prediction lies in a certain interval. Only the models based on general atom types E are included in this analysis. The error is computed as the absolute difference between the predicted EEM value and reference QM value of the atomic charge. The gray blocks represent the error distribution averaged over all 12 EEM models obtained in this study, while the black lines indicate the minimum and maximum values for these distributions. (A) Values obtained for insulin. (B) Values obtained for ubiquitin.

dx.doi.org/10.1021/ci400448n | J. Chem. Inf. Model. 2013, 53, 2548−2558

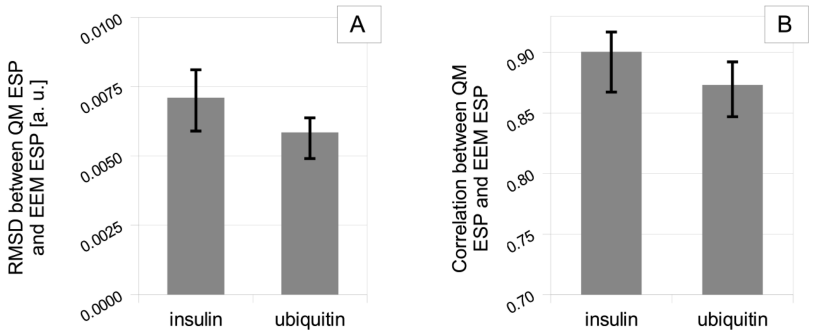


Figure 4. Comparison between the QM and EEM based ESP for insulin and ubiquitin. Only the models based on general atom types E are included in this analysis. The gray blocks represent averages over all 12 EEM models, while the black lines indicate the minimum and maximum values. The ESP was calculated using the biomolecular electrostatics software APBS.[67] (A) Deviation between QM and EEM based ESP measured as RMSD. (B) Correlation between QM and EEM based ESP measured as Pearson's squared correlation coefficient.

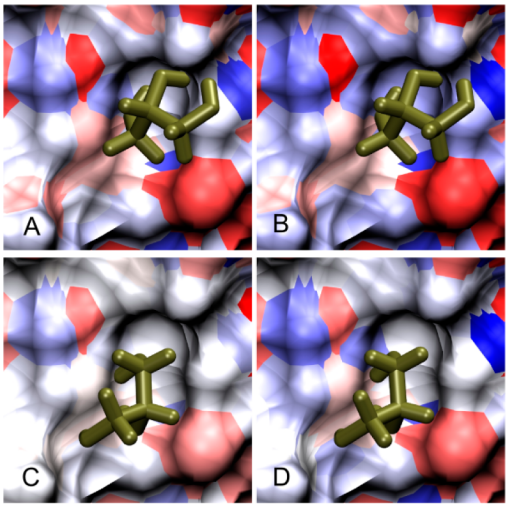


Figure 5. Blind docking of glycerol (ochre, licorice representation) onto the ubiquitin fragment (surface representation colored according to atomic charges, where red signifies more negative charge, while blue signifies more positive charge). The fixed receptor was placed in a 100 × 90 × 90 sized grid, with 0.35 Å spacing between the grid points. The ligand's initial conformation was taken from the coordinates of the ideal glycerol model available in Ligand Expo[70] and contained five rotatable bonds. All calculations were set up using Triton[56] and performed with AutoDock4.2.[71] Each docking run employed 200 iterations of the Genetic Algorithm, with up to 1 000 000 energy evaluations for each iteration (the rest of the parameters were kept in their default values). The calculations differ in the type of atomic charges used for the receptor, while the ligand always carried Gasteiger−Marsili charges. In each case, the binding pose which gives the best binding energy in five independent docking runs is chosen for comparison. (A) Using MPA/6-31G*/gas QM charges—estimated binding energy −9.64 kcal/mol. (B) Using E-MPA/6-31G*/gas EEM charges—estimated binding energy −9.71 kcal/mol, RMSD 0.132 Å compared to the QM pose. (C) Using Gasteiger−Marsili charges—estimated binding energy −8.65 kcal/mol, RMSD 3.244 Å compared to the QM pose; (D) Using AMBER ff94 charges—estimated binding energy −8.07 kcal/mol, RMSD 3.235 Å compared to the QM pose.

QSPR modeling, it was shown that QSPR models which use QM MPA charges as descriptors perform better than QSPR models which use QM charges derived from electrostatic potentials (ESP charges).[72,73] On the other hand, QSPR models which use EEM ESP charges perform comparably well as the QSPR models which use EEM MPA charges,[50] suggesting that the error of the EEM vs QM prediction could very well overwhelm the distinction between different atomic charge definitions.

Therefore, one last aspect to be studied is the sensitivity of the EEM models produced in this study with respect to population analysis, basis set, and environment. For this purpose, we inspected the correlation between the QM charges produced by each of the 12 QM schemes, and the EEM charges produced by all 24 EEM models (see Supporting Information Table S2). Table 4 contains the conclusions for the 12 EEM models based on general atom types E, for insulin and ubiquitin.

It is clear from Table 4 (see also Supporting Information Table S2) that the models obtained in this study easily distinguish between population analyses. In all cases, EEM models parametrized onto MPA charges provide the best prediction for QM MPA data. Similarly, EEM models parametrized onto NPA charges always provide the best prediction for QM NPA data, and EEM models parametrized onto HiI charges always provide the best prediction for QM HiI data. Further, in the case of MPA charges, the models easily distinguish between basis sets, which is expected, since MPA charges are known to be basis set dependent. As such, EEM models parametrized onto MPA/6-31G* charges provide the best prediction for QM MPA/6-31G* data. The equivalent is true for 6-31G**. In the case of NPA and HiI charges, the models do not distinguish between basis sets, which correlates with the fact that NPA and HiI charges have low basis set dependence.[74] As such, EEM models parametrized onto NPA/6-31G* charges provide the best prediction for both QM NPA/6-31G* and NPA/6-31G** data. Similarly, EEM models parametrized onto HiI/6-31G* charges provide the best prediction for both QM HiI/6-31G* and HiI/6-31G** data. On the other hand, the EEM models are not able to distinguish between the gas phase and implicit solvation conditions. In many cases, EEM models parametrized onto gas phase data provide the best prediction for QM PCM data, and vice versa.

Availability of Implementation. The models developed in the present study are fully compatible with any implementation of the EEM formalism given by eq 3, allowing the computation of accurate, conformationally dependent atomic charges in proteins with hundreds of residues in only a few minutes. In particular, these models can be used with EEM_SOLVER, our previously published EEM implementation, which has already been implemented in the Parallel Virtual Machine (PVM) environment and scales very favorably for large systems.[55]

dx.doi.org/10.1021/ci400448n | J. Chem. Inf. Model. 2013, 53, 2548−2558

Table 4. Verification of EEM Model Consistency with QM Calculation Scheme with Respect to Population Analysis, Basis Set, and Environment[a]

R_{avg}	best E-EEM model	
EEM model	insulin	ubiquitin
MPA/6-31G*/gas	MPA/6-31G*/gas	MPA/6-31G*/PCM
MPA/6-31G*/PCM	MPA/6-31G*/PCM	MPA/6-31G*/PCM
MPA/6-31G**/gas	MPA/6-31G**/gas	MPA/6-31G**/gas
MPA/6-31G**/PCM	MPA/6-31G**/gas	MPA/6-31G**/gas
NPA/6-31G*/gas	NPA/6-31G*/PCM	NPA/6-31G*/PCM
NPA/6-31G*/PCM	NPA/6-31G*/gas	NPA/6-31G*/PCM
NPA/6-31G**/gas	NPA/6-31G*/gas	NPA/6-31G**/PCM
NPA/6-31G**/PCM	NPA/6-31G*/gas	NPA/6-31G*/PCM
HiI/6-31G*/gas	HiI/6-31G**/gas	HiI/6-31G**/gas
HiI/6-31G*/PCM	HiI/6-31G**/gas	HiI/6-31G**/gas
HiI/6-31G**/gas	HiI/6-31G**/gas	HiI/6-31G**/gas
HiI/6-31G**/PCM	HiI/6-31G**/gas	HiI/6-31G**/gas
D_{avg}	best E-EEM model	
EEM model	insulin	ubiquitin
MPA/6-31G*/gas	MPA/6-31G*/gas	MPA/6-31G*/gas
MPA/6-31G*/PCM	MPA/6-31G*/gas	MPA/6-31G*/gas
MPA/6-31G**/gas	MPA/6-31G**/gas	MPA/6-31G**/gas
MPA/6-31G**/PCM	MPA/6-31G**/gas	MPA/6-31G**/gas
NPA/6-31G*/gas	NPA/6-31G**/PCM	NPA/6-31G**/PCM
NPA/6-31G*/PCM	NPA/6-31G*/gas	NPA/6-31G**/PCM
NPA/6-31G**/gas	NPA/6-31G*/gas	NPA/6-31G**/PCM
NPA/6-31G**/PCM	NPA/6-31G*/gas	NPA/6-31G**/PCM
HiI/6-31G*/gas	HiI/6-31G*/gas	HiI/6-31G*/PCM
HiI/6-31G*/PCM	HiI/6-31G*/PCM	HiI/6-31G*/PCM
HiI/6-31G**/gas	HiI/6-31G*/PCM	HiI/6-31G*/PCM
HiI/6-31G**/PCM	HiI/6-31G*/PCM	HiI/6-31G*/PCM
$RMSD_{avg}$	best E-EEM model	
EEM model	insulin	ubiquitin
MPA/6-31G*/gas	MPA/6-31G*/gas	MPA/6-31G*/PCM
MPA/6-31G*/PCM	MPA/6-31G*/gas	MPA/6-31G*/PCM
MPA/6-31G**/gas	MPA/6-31G**/gas	MPA/6-31G**/gas
MPA/6-31G**/PCM	MPA/6-31G**/gas	MPA/6-31G**/gas
NPA/6-31G*/gas	NPA/6-31G*/gas	NPA/6-31G**/PCM
NPA/6-31G*/PCM	NPA/6-31G*/gas	NPA/6-31G**/PCM
NPA/6-31G**/gas	NPA/6-31G*/gas	NPA/6-31G*/gas
NPA/6-31G**/PCM	NPA/6-31G*/gas	NPA/6-31G*/gas
HiI/6-31G*/gas	HiI/6-31G**/gas	HiI/6-31G*/PCM
HiI/6-31G*/PCM	HiI/6-31G*/PCM	HiI/6-31G*/PCM
HiI/6-31G**/gas	HiI/6-31G**/gas	HiI/6-31G*/PCM
HiI/6-31G**/PCM	HiI/6-31G*/PCM	HiI/6-31G*/PCM

[a]Only the analysis for models based on E atom type classification is given here, while the complete results can be found in Supporting Information Table S2.

CONCLUSION

The goal of the study was to extend the coverage of the Electronegativity Equalization Method (EEM) for atomic charge calculation in proteins. For the purpose of EEM model parametrization, fragments of experimentally determined protein structures were used as reference systems. Reference QM atomic charges given by three population analyses were calculated at the HF level, with two basis sets and in two environments. A total of 24 EEM models were parametrized and evaluated in the present study.

Upon validation on insulin and ubiquitin, the models exhibited high correlation and low deviations between the QM reference charges and EEM predicted charges. Very good accuracy is expected for 80% of the predictions, while poor accuracy is expected for 5% of the predictions. Last, the models were found to be consistent with respect to basis set and population analysis.

Overall, the results of the evaluation suggest that the EEM models obtained in this study can be used reliably for the rapid calculation of conformationally dependent atomic charges in proteins and protein complexes. The models can be used with our previously published EEM implementation EEM_SOLVER, which allows for an efficient estimation of atomic charges with application in biomolecular modeling investigations.

ASSOCIATED CONTENT

Supporting Information

Structures of all reference molecules (.pdb format), values of all QM and EEM atomic charges (.csv format), values of parameters for all EEM models (Table S1, .pdf format), results of the validation of all EEM models against all QM schemes (Table S2, .pdf format), reproduction of ESP for insulin and ubiquitin (Table S3, .pdf format), and comparison of docking runs using different types of charges for the receptor (Table S4, .pdf format). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: jkoca@ceitec.muni.cz.

*E-mail: svobodova@chemi.muni.cz.

Author Contributions

The study was conceived by J.K., while the procedure was designed by C.-M.I. and R.S.V. The calculations were performed by C.-M.I. and S.G., and the data was analyzed by all authors. The manuscript was drafted by C.-M.I. and critically revised by all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

C.-M.I. is very grateful to Tomáš Bouchal for the graphic design ideas and to Dr. Sushil Kumar Mishra for the useful discussions. This work was funded by the Ministry of Education, Youth and Sports of the Czech Republic (contract number LH13055) and the European Community's Seventh Framework Programme (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund and from the "Capacities" specific program (Contract No. 286154). The access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the program "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005) is highly appreciated. C.-M.I. would like to thank Brno City Municipality for the financial support provided to her through the program Brno Ph.D. Talent.

ABBREVIATIONS

E, atom type classification approach based on chemical elements; EEM, Electronegativity Equalization Method; ESP, electrostatic potential; EX, atom type classification approach based on chemical elements and maximum bond multiplicity; HiI, iterative Hirshfeld population analysis; MPA, Mulliken Population Analysis; NMR, nuclear magnetic resonance; NPA, natural

population analysis; PA, population analysis; PCM, polarizable continuum model; PDB, Protein Data Bank; PVM, parallel virtual machine; QM, quantum mechanics/mechanical; QSPR, quantitative structure−property relationships; RMSD, root-mean-square deviation

REFERENCES

(1) Wang, R.; Gao, Y.; Lai, L. Calculating partition coefficient by atom-additive method. *Persp. Drug Discov. Design* **2000**, *19*, 47−66.

(2) Czodrowski, P.; Dramburg, I.; Sotriffer, C. A.; Klebe, G. Development, validation, and application of adapted PEOE charges to estimate pKa values of functional groups in protein-ligand complexes. *Proteins Struct. Funct. Bioinf.* **2006**, *65*, 424−437.

(3) Cherkasov, A.; Shi, Z.; Li, Y.; Jones, S. J. M.; Fallahi, M.; Hammond, G. L. 'Inductive' charges on atoms in proteins: comparative docking with the extended steroid benchmark set and discovery of a novel SHBG ligand. *J. Chem. Inf. Model.* **2005**, *45*, 1842−1853.

(4) Cho, A. E.; Guallar, V.; Berne, B. J.; Friesner, R. Importance of accurate charges in molecular docking: quantum mechanical/molecular mechanical (QM/MM) approach. *J. Comput. Chem.* **2005**, *26*, 915−931.

(5) Mulliken, R. S. Electronic Structures of Molecules XI. Electro-affinity, Molecular Orbitals and Dipole Moments. *J. Chem. Phys.* **1935**, *3*, 573−585.

(6) Mulliken, R. S. Criteria for the Construction of Good Self-Consistent-Field Molecular Orbital Wave Functions, and the Significance of LCAO-MO Population Analysis. *J. Chem. Phys.* **1962**, *36*, 3428−3439.

(7) Löwdin, P.-O. On the Non-Orthogonality Problem Connected with the Use of Atomic Wave Functions in the Theory of Molecules and Crystals. *J. Chem. Phys.* **1950**, *18*, 365−375.

(8) Reed, A. E.; Weinstock, R. B.; Weinhold, F. Natural population analysis. *J. Chem. Phys.* **1985**, *83*, 735−746.

(9) Bader, R. F. W.; Larouche, A.; Gatti, C.; Carroll, M. T.; Macdougall, P. J.; Wiberg, K. B. Properties of atoms in molecules: Dipole moments and transferability of properties. *J. Chem. Phys.* **1987**, *87*, 1142−1152.

(10) Hirshfeld, F. L. Bonded-atom fragments for describing molecular charge densities. *Theor. Chim. Acta* **1977**, *44*, 129−138.

(11) Bultinck, P.; Van Alsenoy, C.; Ayers, P. W.; Carbó-Dorca, R. Critical analysis and extension of the Hirshfeld atoms in molecules. *J. Chem. Phys.* **2007**, *126*, 144111.

(12) Breneman, C. M.; Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *J. Comput. Chem.* **1990**, *11*, 361−373.

(13) Besler, B. H.; Merz, K. M.; Kollman, P. A. Atomic Charges Derived from Semiempirical Methods. *J. Comput. Chem.* **1990**, *11*, 431−439.

(14) Kelly, C. P.; Cramer, C. J.; Truhlar, D. G. Accurate partial atomic charges for high-energy molecules using class IV charge models with the MIDI! basis set. *Theor. Chem. Acc.* **2005**, *113*, 133−151.

(15) Abraham, R. J.; Griffiths, L.; Loftus, P. Approaches to charge calculations in molecular mechanics. *J. Comput. Chem.* **1982**, *3*, 407−416.

(16) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(17) Oliferenko, A. A.; Pisarev, S. A.; Palyulin, V. A.; Zefirov, N. S. Atomic Charges via Electronegativity Equalization: Generalizations and Perspectives. *Adv. Quantum Chem.* **2006**, *51*, 139−156.

(18) Shulga, D. A.; Oliferenko, A. A.; Pisarev, S. A.; Palyulin, V. A.; Zefirov, N. S. Parameterization of empirical schemes of partial atomic charge calculation for reproducing the molecular electrostatic potential. *Dokl. Chem.* **2008**, *419*, 57−61.

(19) Cho, K.-H.; Kang, Y. K.; No, K. T.; Scheraga, H. A. A Fast Method for Calculating Geometry-Dependent Net Atomic Charges for Polypeptides. *J. Phys. Chem. B* **2001**, *105*, 3624−3634.

(20) Mortier, W. J.; Ghosh, S. K.; Shankar, S. Electronegativity-equalization method for the calculation of atomic charges in molecules. *J. Am. Chem. Soc.* **1986**, *108*, 4315−4320.

(21) Rappé, A. K.; Goddard, W. A. Charge Equilibration for Molecular Dynamics Simulations. *J. Phys. Chem.* **1991**, *95*, 3358−3363.

(22) Nistor, R. A.; Polihronov, J. G.; Muser, M. H.; Mosey, N. J. A generalization of the charge equilibration method for nonmetallic materials. *J. Chem. Phys.* **2006**, *125*, 094108−094118.

(23) Yang, Z.-Z.; Wang, C.-S. Atom−Bond Electronegativity Equalization Method. 1. Calculation of the Charge Distribution in Large Molecules. *J. Phys. Chem. A* **1997**, *101*, 6315−6321.

(24) Chaves, J.; Barroso, J. M.; Bultinck, P.; Carbó-Dorca, R. Toward an alternative hardness kernel matrix structure in the Electronegativity Equalization Method (EEM). *J. Chem. Inf. Model.* **2006**, *46*, 1657−1665.

(25) Meister, J.; Schwarz, W. H. E. Principal Components of Ionicity. *J. Phys. Chem.* **1994**, *98*, 8245−8252.

(26) Van der Vaart, A.; Bursulaya, B. D.; Brooks, C. L.; Merz, K. M. Are many body effects important in protein folding? *J. Phys. Chem. B* **2000**, *104*, 9554−9563.

(27) Bucher, D.; Raugei, D.; Guidoni, L.; Dal Peraro, M.; Rothlisberger, U.; Carloni, P.; Klein, M. L. Polarization effects and charge transfer in the KcsA potassium channel. *Biophys. Chem.* **2006**, *124*, 292−301.

(28) Ufimtsev, I. S.; Luehr, N.; Martinez, T. J. Charge Transfer and Polarization in Solvated Proteins from Ab Initio Molecular Dynamics. *J. Phys. Chem. Lett.* **2011**, *2*, 1789−1793.

(29) Anisimov, V. M.; Cavasotto, C. N. In *Challenges and Advances in Computational Chemistry and Physics*; Paneth, P., Dybala-Defratyka, A., Eds.; Springer: Netherlands, 2010; Vol. *12*, Chapter 9, pp 247−266.

(30) Cho, A. E.; Rinaldo, D. Extension of QM/MM docking and its applications to metalloproteins. *J. Comput. Chem.* **2009**, *30*, 2609−2616.

(31) Wallin, G.; Nervall, M.; Carlsson, J.; Åqvist, J. Charges for Large Scale Binding Free Energy Calculations with the Linear Interaction Energy Method. *J. Chem. Theory. Comput.* **2009**, *5*, 380−395.

(32) Yang, Z.; Shen, E.; Wang, L. A scheme for calculating atomic charge distribution in large molecules based on density functional theory and electronegativity equalization. *J. Mol. Str.: Theochem* **1994**, *312*, 167−173.

(33) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Waroquier, M.; Tollenaere, J. P. The electronegativity equalization method I: Parametrization and validation for atomic charge calculations. *J. Phys. Chem. A* **2002**, *106*, 7887−7894.

(34) Bultinck, P.; Langenaeker, W.; Lahorte, P.; De Proft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, J. P. The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. *J. Phys. Chem. A* **2002**, *106*, 7895−7901.

(35) Bultinck, P.; Vanholme, R.; Popelier, P. L. A.; De Proft, F.; Geerlings, P. High-Speed Calculation of AIM Charges through the Electronegativity Equalization Method. *J. Phys. Chem. A* **2004**, *108*, 10359−10366.

(36) Berente, I.; Czinki, E.; Naray-Szabo, G. A Combined Electronegativity Equalization and Electrostatic Potential Fit Method for the Determination of Atomic Point Charges. *J. Comput. Chem.* **2007**, *28*, 1936−1942.

(37) Svobodová Vařeková, R.; Jiroušková, Z.; Vaněk, J.; Suchomel, Š; Koča, J. Electronegativity Equalization Method: Parameterization and Validation for Large Sets of Organic, Organohalogene and Organometal Molecule. *Int. J. Mol. Sci.* **2007**, *8*, 572−582.

(38) Jiroušková, Z.; Svobodová Vařeková, R.; Vaněk, J.; Koča, J. Electronegativity Equalization Method: Parameterization and Validation for Organic Molecules using the Merz-Kollman-Singh Charge Distribution Scheme. *J. Comput. Chem.* **2009**, *30*, 1174−1178.

(39) Verstraelen, T.; Van Speybroeck, V.; Waroquier, M. The electronegativity equalization method and the split charge equilibration applied to organic systems: Parametrization, validation, and comparison. *J. Chem. Phys.* **2009**, *131*, 044127−19.

(40) Puranen, J. S.; Vainio, M. J.; Johnson, M. S. Accurate Conformation-Dependent Molecular Electrostatic Potentials for High-

Throughput *In Silico* Drug Discovery. *J. Comput. Chem.* **2010**, *31*, 1722−1732.
(41) Ouyang, Y.; Ye, F.; Liang, Y. A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules. *Phys. Chem. Chem. Phys.* **2009**, *11*, 6082−6089.
(42) Verstraelen, T.; Bultinck, P.; Van Speybroeck, V.; Ayers, P. W.; Van Neck, D.; Waroquier, M. The Significance of Parameters in Charge Equilibration Models. *J. Chem. Theory Comput.* **2011**, *7*, 1750−1764.
(43) Verstraelen, T.; Sukhomlinov, S. V.; Van Speybroeck, V.; Waroquier, M.; Smirnov, K. S. Computation of Charge Distribution and Electrostatic Potential in Silicates with the Use of Chemical Potential Equalization Models. *J. Phys. Chem. C* **2012**, *116*, 490−504.
(44) Ionescu, C.-M.; Svobodová Vařeková, R.; Prehn, J. H. M.; Huber, H. J.; Koča, J. Charge Profile Analysis Reveals That Activation of Pro-apoptotic Regulators Bax and Bak Relies on Charge Transfer Mediated Allosteric Regulation. *PloS Comput. Biol.* **2012**, *8*, e1002565−11.
(45) Cedillo, A.; Van Neck, D.; Bultinck, P. Self-consistent methods constrained to a fixed number of particles in a given fragment and its relation to the electronegativity equalization method. *Theor. Chem. Acc.* **2012**, *131*, 1227−1233.
(46) Heidler, R.; Janssens, G. O. A.; Mortier, W. J.; Schoonheydt, R. A. Charge Sensitivity Analysis of Intrinsic Basicity of Faujasite-Type Zeolites Using the Electronegativity Equalization Method (EEM). *J. Phys. Chem.* **1996**, *100*, 19728−19734.
(47) Bultinck, P.; Langenaeker, W.; Carbó-Dorca, R.; Tollenaere, J. P. Fast Calculation of Quantum Chemical Molecular Descriptors from the Electronegativity Equalization Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 422−428.
(48) Shimizu, K.; Chaimovich, H.; Farah, J. P. S.; Dias, L. G.; Bostick, D. L. Calculation of the Dipole Moment for Polypeptides Using the Generalized Born-Electronegativity Equalization Method: Results in Vacuum and Continuum-Dielectric Solvent. *J. Phys. Chem. B* **2004**, *108*, 4171−4177.
(49) Chen, S.; Yang, Z. Molecular Dynamics Simulations of a β-Hairpin Fragment of Protein G by Means of Atom-Bond Electronegativity Equalization Method Fused into Molecular Mechanics (ABEEM$\sigma\pi$/MM). *Chin. J. Chem.* **2010**, *28*, 2109−2118.
(50) Svobodová Vařeková, R.; Geidl, S.; Ionescu, C.-M.; Skřehota, O.; Bouchal, T.; Sehnal, D.; Abagyan, R.; Koča, J. Predicting pKa values from EEM atomic charges. *J. Cheminf.* **2013**, *5*, 18.
(51) Sanderson, R. T. An Interpretation of Bond Lengths and a Classification of Bonds. *Science* **1951**, *114*, 670−672.
(52) Parr, R. G.; Donnelly, R. A.; Levy, M.; Palke, W. E. Electronegativity- the density functional viewpoint. *J. Chem. Phys.* **1978**, *68*, 3801−3807.
(53) Parr, R. G.; Pearson, R. G. Absolute hardness: companion parameter to absolute electronegativity. *J. Am. Chem. Soc.* **1983**, *105*, 7512−7516.
(54) Yang, Z.; Shen, E. Molecular electronegativity in density functional theory (I). *Ser. B Sci. China* **1995**, *38*, 521−528.
(55) Svobodová Vařeková, R.; Koča, J. Optimized and Parallelized Implementation of the Electronegativity Equalization Method and the Atom-Bond Electronegativity Equalization Method. *J. Comput. Chem.* **2005**, *27*, 396−405.
(56) Prokop, M.; Adam, J.; Križ, Z.; Wimmerová, M.; Koča, J. TRITON: a graphical tool for ligand-binding protein engineering. *Bioinformatics* **2008**, *24*, 1955−1956.
(57) Timofeev, V. I.; Baidus, A. N.; Kislitsyn, Y. A.; Kuranova, I. P. 2009, DOI: 10.2210/pdb3e7y/pdb. http://www.rcsb.org/pdb/explore.do?structureId=3E7Y (accessed Oct 31, 2012).
(58) Vijay-Kumar, S.; Bugg, C. E.; Cook, W. J. Structure of ubiquitin refined at 1.8 A resolution. *J. Mol. Biol.* **1987**, *194*, 531−544.
(59) Miertuš, S.; Scrocco, E.; Tomasi, J. Electrostatic interaction of a solute with a continuum. A direct utilizaion of ab initio molecular potentials for the prevision of solvent effects. *Chem. Phys.* **1981**, *55*, 117−29.
(60) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, Jr., J. A.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision C.01; Gaussian, Inc.: Wallingford, CT, 2009.
(61) Glendening, E. D.; Reed, A. E.; Carpenter, J. E.; Weinhold, F. *NBO Version 3.1*; Gaussian, Inc.: Wallingford, CT, 2009.
(62) Verstraelen T. HiPart Program. http://molmod.ugent.be/software (accessed May 23, 2013).
(63) Wilson, M. S.; Ichikawa, S. Comparison between the Geometric and Harmonic Mean Electronegativity Equilibration Techniques. *J. Phys. Chem.* **1989**, *93*, 3087−3089.
(64) Pauling, L. The Nature of the Chemical Bond. IV. The Energy of Single Bonds and the Relative Electronegativity of Atoms. *J. Am. Chem. Soc.* **1932**, *54*, 3570−3582.
(65) Allred, A. L. Electronegativity values from thermochemical data. *J. Inorg. Nucl. Chem.* **1961**, *17*, 215−221.
(66) Humphrey, W.; Dalke, A.; Schulten, K. VMD - Visual Molecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(67) Baker, N. A.; Sept, D.; Joseph, S.; Holst, M. J.; McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10037−10041.
(68) Page, R. C.; Pruneda, J. N.; Amick, J.; Klevit, R. E.; Misra, S. Structural insights into the conformation and oligomerization of E2∼ubiquitin conjugates. *Biochemistry* **2012**, *51*, 4175−4187.
(69) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179−5197.
(70) Feng, Z.; Chen, L.; Maddula, H.; Akcan, O.; Oughtred, R.; Berman, H. M.; Westbrook, J. Ligand Depot: a data warehouse for ligands bound to macromolecules. *Bioinformatics* **2004**, *13*, 2153−2155.
(71) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. Autodock4 and AutoDockTools4: automated docking with selective receptor flexiblity. *J. Comput. Chem.* **2009**, *16*, 2785−91.
(72) Gross, K. C.; Seybold, P. G.; Hadad, C. M. Comparison of different atomic charge schemes for predicting pKa variations in substituted anilines and phenols. *Int. J. Quantum Chem.* **2002**, *90*, 445−458.
(73) Svobodová Vařeková, R.; Geidl, S.; Ionescu, C.-M.; Skřehota, O.; Kudera, M.; Sehnal, D.; Bouchal, T.; Abagyan, R.; Huber, H. J.; Koča, J. Predicting pK_a values of substituted phenols from atomic charges. *J. Chem. Inf. Model.* **2011**, *51*, 1795−1806.
(74) Bultinck, P.; Ayers, P. W.; Fias, S.; Tiels, K.; Van Alsenoy, C. Uniqueness and basis set dependence of iterative Hirshfeld charges. *Chem. Phys. Lett.* **2007**, *444*, 205−208.

Charge profile analysis reveals that activation of proapoptotic regulators Bax and Bak relies on charge transfer mediated allosteric regulation

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Charge Profile Analysis Reveals That Activation of Pro-apoptotic Regulators Bax and Bak Relies on Charge Transfer Mediated Allosteric Regulation

Crina-Maria Ionescu[1,2], Radka Svobodová Vařeková[1,2], Jochen H. M. Prehn[3,4], Heinrich J. Huber[3,4], Jaroslav Koča[1,2]*

1 CEITEC - Central European Institute of Technology, Masaryk University, Brno, Czech Republic, **2** National Centre for Biomolecular Research, Masaryk University, Brno, Czech Republic, **3** Centre for Systems Medicine, Royal College of Surgeons in Ireland, Dublin, Ireland, **4** Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, Dublin, Ireland

Abstract

The pro-apoptotic proteins Bax and Bak are essential for executing programmed cell death (apoptosis), yet the mechanism of their activation is not properly understood at the structural level. For the first time in cell death research, we calculated intra-protein charge transfer in order to study the structural alterations and their functional consequences during Bax activation. Using an electronegativity equalization model, we investigated the changes in the Bax charge profile upon activation by a functional peptide of its natural activator protein, Bim. We found that charge reorganizations upon activator binding mediate the exposure of the functional sites of Bax, rendering Bax active. The affinity of the Bax C-domain for its binding groove is decreased due to the Arg94-mediated abrogation of the Ser184-Asp98 interaction. We further identified a network of charge reorganizations that confirms previous speculations of allosteric sensing, whereby the activation information is conveyed from the activation site, through the hydrophobic core of Bax, to the well-distanced functional sites of Bax. The network was mediated by a hub of three residues on helix 5 of the hydrophobic core of Bax. Sequence and structural alignment revealed that this hub was conserved in the Bak amino acid sequence, and in the 3D structure of folded Bak. Our results suggest that allostery mediated by charge transfer is responsible for the activation of both Bax and Bak, and that this might be a prototypical mechanism for a fast activation of proteins during signal transduction. Our method can be applied to any protein or protein complex in order to map the progress of allosteric changes through the proteins' structure.

Citation: Ionescu C-M, Svobodová Vařeková R, Prehn JHM, Huber HJ, Koča J (2012) Charge Profile Analysis Reveals That Activation of Pro-apoptotic Regulators Bax and Bak Relies on Charge Transfer Mediated Allosteric Regulation. PLoS Comput Biol 8(6): e1002565. doi:10.1371/journal.pcbi.1002565

Editor: James M. Briggs, University of Houston, United States of America

Received March 8, 2012; **Accepted** May 4, 2012; **Published** June 14, 2012

Funding: This work was funded by the Czech Science Foundation (GD301/09/H004), and the European Community's Seventh Framework Programme (CZ.1.05/1.1.00/02.0068) from the European Regional Development Fund and from the "Capacities" specific programme (Contract No. 286154). The access to the MetaCentrum computing facilities was provided under the programme "Projects of Large Infrastructure for Research, Development, and Innovations" LM2010005 funded by the Ministry of Education, Youth, and Sports of the Czech Republic. CMI would like to thank Brno City Municipality for the financial support provided to her through the program Brno Ph.D. Talent. HJH and JHMP wish to acknowledge funding by Science Foundation Ireland grant PI 08/IN.1/B1949. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jkoca@ceitec.muni.cz

Introduction

Mitochondrial outer membrane permeabilization (MOMP) is a hallmark of programmed cell death (apoptosis). Following MOMP, apoptotic proteins from the mitochondrial inter-membrane space are released, causing the activation of cell death proteases which cleave the cell's cytoskeleton and genetic material. MOMP is executed by the Bcl-2 family proteins Bak and Bax that, upon activation during apoptosis, oligomerize and form pores in the mitochondrial membrane [1–4].

Bak and Bax oligomerisation is controlled by the interplay of further Bcl-2 proteins [5–8]. While pro-survival Bcl-2 proteins bind to and deactivate Bak and Bax [9], other apoptotic Bcl-2 proteins de-repress this inhibition, leaving Bak and Bax free to oligomerize [10]. Nevertheless, a separate step, whereby a subclass of apoptotic Bcl-2 proteins such as Bim and Bid directly activate Bak and Bax, was proposed to be required for oligomerization [11–13].

The activation steps required for Bax oligomerization were extensively investigated [14–18]. These steps were found to comprise Bax translocation from the cytosol to the mitochondrial membrane, and changes of Bax conformation. Conformational changes of Bax include exposure of its C-domain, insertion of this C-domain into the membrane, and exposure of the Bax BH3 domain, one of four homology domains of Bcl-2 proteins (Figure 1).

In inactive Bax, the C-domain is tightly bound inside a hydrophobic pocket which we henceforth denote as the 'BH groove'. This tight binding was suggested to increase the solubility of Bax and to keep Bax in the cytosol in the absence of stress [16]. Gavathiotis et al. [18] synthesized a helix mimicking the BH3 domain of the activator Bim (Bim-stabilized α-helix of Bcl-2 domains, Bim-SAHB). They subsequently performed NMR spectroscopy to study the interaction of Bax with the Bim-SAHB activator. They found that, in the absence of Bim-SAHB, the Bax

Author Summary

Apoptosis is a physiological form of cell death that is fundamental for development, growth and homeostasis in multi-cellular organisms. Deviations in the apoptosis machinery are known to be involved in cancer, neurodegenerative disorders, and autoimmune diseases. The proteins Bax and Bak are essential for executing apoptosis, yet the mechanism of their activation is not properly understood at the structural level. To understand this mechanism, we investigated how the electronic density is reorganized (i.e., how charge is transferred) inside the Bax molecule when Bax binds a functional peptide of its natural activator protein. We identified the specific interactions responsible for the exposure of the functional sites of Bax, rendering Bax active. Furthermore, we found a network of charge transfer that conveys activation information from the Bax activation site, through the hydrophobic core of Bax, to the well-distanced functional sites of Bax. This network consists of three residues inside the hydrophobic core of Bax, which are present also in the hydrophobic core of Bak, suggesting that these residues are functionally important and thus potential drug targets. We provide a straightforward and accessible methodology to identify the key residues involved in the fast activation of proteins during signal transduction.

activation site was blocked by a largely unstructured loop (loop 1–2), which opens upon incubation with Bim-SAHB. Using Bax mutants with reduced loop 1–2 mobility, Gavathiotis et al. later demonstrated that the opening of this loop was a prerequisite for Bax activation [19]. Interestingly, the suggested Bax activation site and the Bax C-domain are separated by over 25 Å. Since the binding of Bim-SAHB to Bax is weak and transient, and neither significant disturbances in the helical packing, nor covalent modifications have been observed in Bax upon activation, the mechanism of how C-domain exposure occurs following this activation remains elusive [20,21].

Charge transfer was found to be significant in many biomolecular interactions [22–24], and functionally linked to protein structural dynamics [25]. In this paper, we therefore investigate the role of charge transfer during Bax activation by employing an electronegativity equalization model for the calculation of atomic charges. Following our investigation, we propose that a charge transfer network is intimately connected to the way that the activation information travels across Bax, and that a similar network is plausible in Bak.

Results

Calibration of an EEM Model for Calculating Partial Atomic Charges in Proteins

The Electronegativity Equalization Method (EEM) [26] is a fast technique for estimating partial atomic charges, and has been successfully applied to zeolites, small organic molecules and polypeptides [27–31]. To use EEM for studying charge transfer during Bax activation, EEM model parameters need to be calibrated to charges of reference molecules.

For this purpose we followed our previously published EEM model calibration procedure [32], with a few modifications that

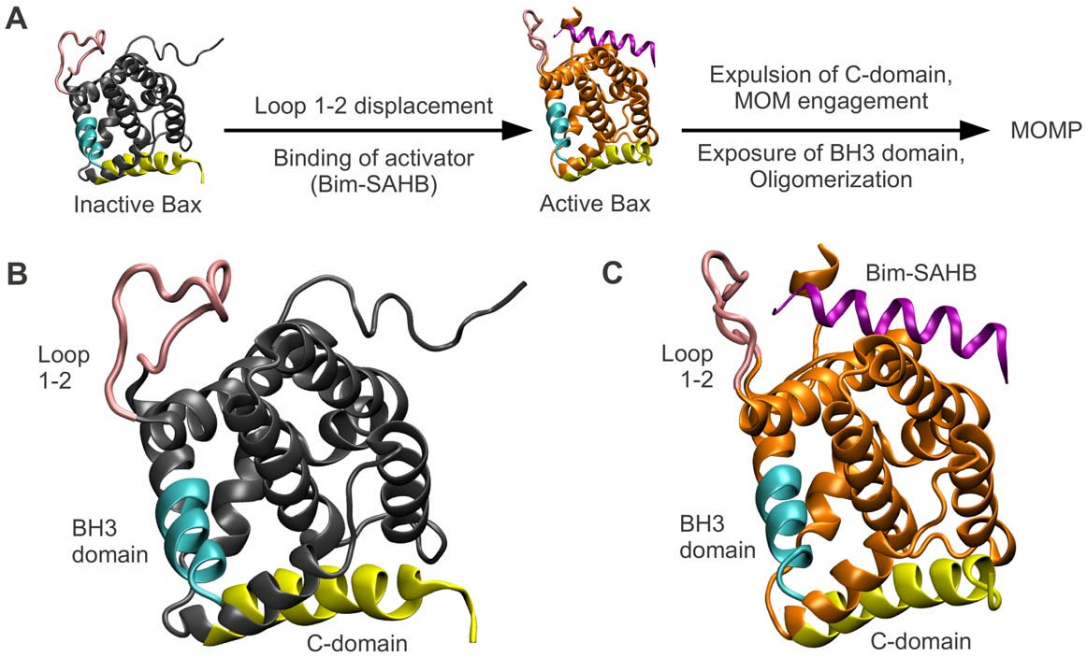


Figure 1. Bax undergoes several conformational changes enabling it to form pores in the mitochondrial outer membrane. (A) Bax activation leads to mitochondrial outer membrane permeabilization (MOMP). Inactive Bax is a cytosolic monomer. Activator-induced opening of loop 1–2 allows the activator to bind. Subsequently, the C-domain of the now active Bax vacates the BH groove and inserts into the mitochondrial membrane. Additionally, the Bax BH3 domain gets exposed. Bax oligomerization ensues via the BH groove and the BH3 domain, eventually leading to the formation of pores which permeabilize the membrane. (B) Location of the BH3 domain (cyan), the C-domain (yellow), loop 1–2 (pink) in inactive Bax (the rest of the protein in gray). (C) Location of above domains (same color coding) in active Bax (the rest of the protein in orange). The activator peptide Bim-SAHB is shown in purple.
doi:10.1371/journal.pcbi.1002565.g001

address the complex nature of proteins. The reference data consisted of atomic charges for molecules of two disjoint reference sets RS1 and RS2, and these charges were calculated using the quantum mechanics (QM) scheme detailed in the Methods section. Since calculation of QM charges for large molecules such as proteins would require too high computational costs, previous EEM models available in literature were calibrated to reference charges from small molecules [32–38]. Moreover, to make EEM as generally applicable as possible, these calibrations used mostly inorganic or drug-like compounds, which do not reflect the complex nature of proteins as long, non-neutral molecular chains with complex 3D assembly. Therefore, to retain properties that are characteristic for proteins and allow a fast calculation of reference QM charges at the same time, large fragments of experimentally determined protein structures were used as reference sets in the present study (see the Methods section for details). We next determined values for the EEM model parameters by fitting them to the reference QM charges using a least squares algorithm. Prior to fitting, we classified atoms according to two schemes. One scheme was based on chemical elements only (denoted 'E'), and the other on chemical elements and maximum bond order for each atom (denoted 'EX', so that, for example, 'O1' indicates simple bonded, and 'O2' double bonded oxygen). Fitting the model parameters for each of the two atom classification schemes and each of the two reference sets of atomic charges, we obtained four parameter sets, denoted RS1-E, RS1-EX, RS2-E, RS2-EX. Finally, we validated our EEM model by assessing the accuracy of the model in reproducing the original QM charges from reference sets RS1 and RS2, and from five additional test molecules T1-T5. Results were evaluated by the average correlation coefficient R_{avg} (squared Pearson's correlation coefficient), the root mean square deviation $RMSD_{avg}$, and the average absolute difference D_{avg}. An overview of the EEM model calibration procedure is given in Figure 2, and the complete details can be found in the Methods section.

Overall, the results in Figure 3 suggested that the finer grained atom classification scheme 'EX' only modestly improved the accuracy compared to the scheme 'E' based on chemical elements alone. The good agreement between QM and EEM charges for all data sets suggested that both atom classification schemes can provide satisfactory calibration results. Moreover, our model was able to compute EEM atomic charges in less than 1 second for any of the reference or test molecules using our previously published implementation [39].

Bax Is Activated by Arg94-mediated Abrogation of the Ser184-Asp98 Interaction, Decreasing the Affinity of the Bax C-domain for Its Binding Groove

Having developed an EEM based method for rapid calculation of atomic partial charges, we investigated whether atomic charge distribution prior and subsequent to Bax activation would reveal any clues about the mechanisms of the activation. To this end, we obtained the 3D structure of inactive Bax (Figure 1B), and of active Bax in complex with the activator peptide Bim-SAHB (Figure 1C) from the Protein Data Bank (PDB IDs 1F16 [16] and 2K7W [18] respectively). We then computed EEM atomic charges using parameter set RS2-E (Figure 2B) for both structures, and assessed the absolute charge transfer per residue (total difference in charge per amino acid residue, ΔQ_{res}), and the intra-residue charge density reorganization (root mean square deviation in charge per residue, $RMSD_{res}$). The mathematical derivation of these descriptors can be found in the Methods section, and their values for all Bax residues are available in Table S1.

Experimental evidence suggests that, in inactive Bax, the C-terminal helix is bound tightly to its hydrophobic pocket ('BH-groove'). During activation, this binding gets destabilized, causing the C-domain to subsequently vacate the BH-groove and insert into the mitochondrial outer membrane. Early mutagenesis studies revealed a critical interaction between residues Ser184 and Asp98 at the C-domain-BH-groove interface, whose abrogation is sufficient to immediately activate Bax [16,40]. We therefore focused on the changes in charge density distribution in the vicinity of this interaction. While our calculations did not show any change in the charge profile of Ser184, they indicated that any interaction that might have taken place between Asp98 and Ser184 in the inactive structure has been replaced by an Asp98-Arg94 salt bridge in the active structure (Figure 4). Upon activation, Arg94 becomes more positive (see also Table S1), which is suggested to lead to the recruitment of Asp98, the abrogation of the Asp98-Ser184 interaction, and ultimately the destabilization of the C-domain. This demonstrates that the binding of Bim-SAHB to Bax can activate Bax by destabilizing the interaction between the Bax C-domain and its binding groove.

A Network of Charge Transfer Extends from the Bax Activation Site, through Its Hydrophobic Core, to the C-domain Binding Groove

It remains puzzling how the BH groove is influenced by the binding of Bim-SAHB to Bax, given that this interaction occurs on the opposite side of the Bax molecule, at a distance of 25 Å from the BH groove.

Interestingly, the residues which showed a transfer of charge one standard deviation higher than average (Table S1) provided a clue as to how the activation information proceeds through the protein. Foremost, significant changes in the net residue charges were found at the Bax activation site, the BH3-domain (required for oligomerization) and the C-domain (required for membrane insertion). Since these are all functional sites of Bax, these changes were not unexpected. For example, George et al [41] found that a triple alanine mutant at residues 63–65 (on the BH3 domain of Bax) ablated Bax oligomerisation and apoptotic activity, which correlates perfectly with the high charge transfer we found on residues 64 and 65 upon Bax activation (Table S1).

However, in addition to the expected changes, our method surprisingly identified significant charge transfer also on the central helix, inside the hydrophobic core of Bax (residues Trp107, Arg109 and Lys119 on helix 5). The presence of significant charges and charge transfer in a predominantly hydrophobic environment suggests that helix 5 acts as a hub which collects and distributes charge density (Figure 5). We further calculated the intra-residue redistributions of charge density upon activator (Bim-SAHB) binding. Significant such redistributions were observed at the Bax activation site, BH groove and loop 1–2. Since these are the functional regions of Bax, these calculations provide further support for the notion of a charge transfer network that conveys information across the entire Bax molecule (Figure S1, Table S1).

Hints of such an interaction transfer phenomenon were found by Gavathiotis et al. [19]. They titrated Bax with increasing amounts of Bim-SAHB and observed small, but reproducible dose-responsive changes in NMR resonance behavior for the backbone N atoms of residues on the Bax C-domain, as well as on helix 5 inside the hydrophobic core of Bax. They concluded that the binding of the activator induces reverberations in the core of the Bax protein, which serve to mobilize the C-domain (allosteric sensing). Our charge analysis explains these reverberations by a network of charge transfer through the entire Bax molecule.

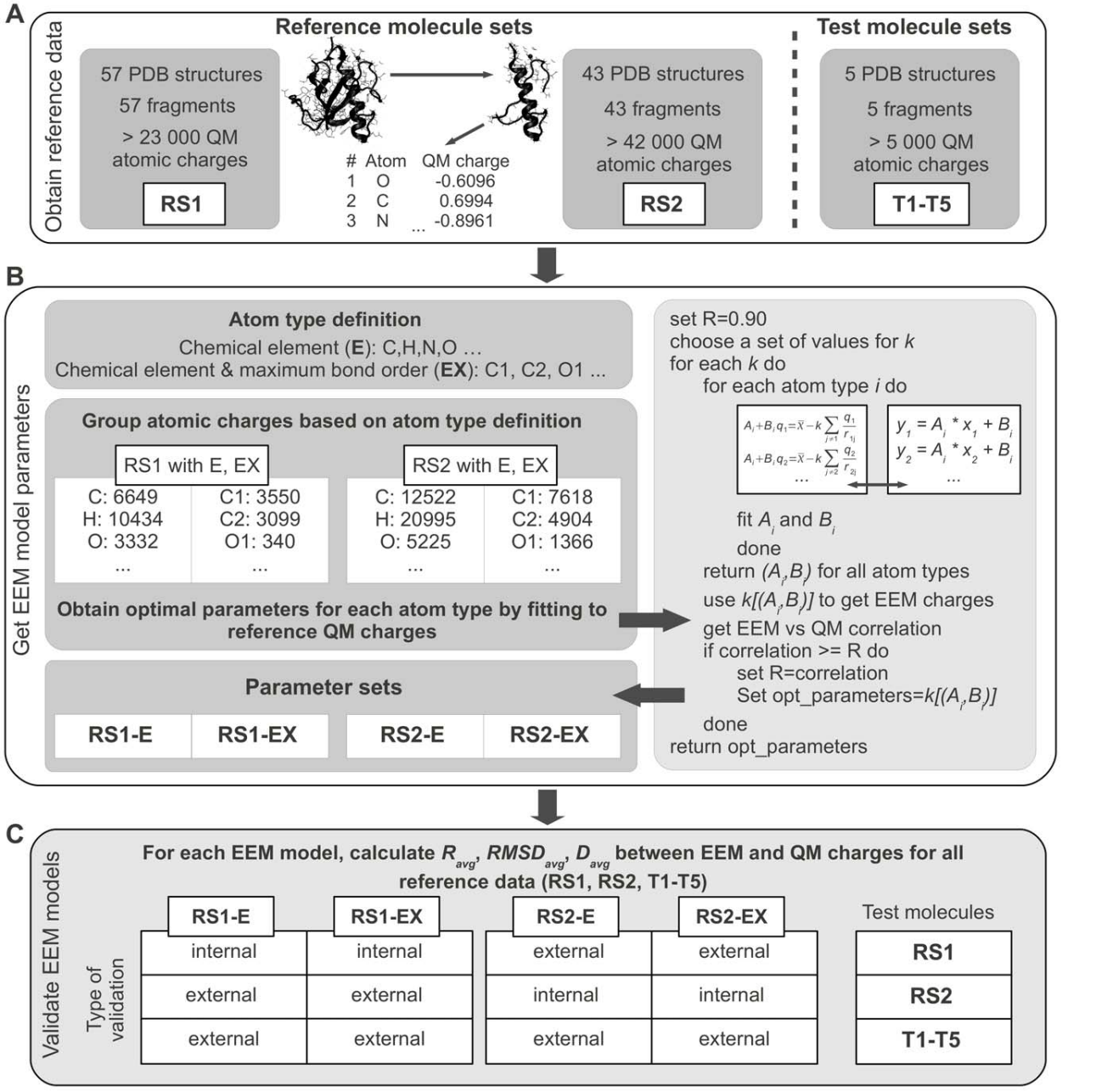


Figure 2. Flowchart of calibration of an EEM model for calculating partial atomic charges in proteins. (A) The reference data used in this study consisted of QM atomic charges for protein fragments in two reference sets (RS1, RS2) and one test set (T1–T5). (B) Two atom type definitions were used. The atomic electronegativity equations were grouped together based on the atom type. The EEM model parameters for each atom type were then obtained by least squares fitting to reference QM charges. (C) Each EEM model was subjected to internal and external validation by comparing the EEM charges with reference QM charges for all available data sets (RS1, RS2, T1–T5).
doi:10.1371/journal.pcbi.1002565.g002

The Residues Essential for the Charge Transfer Network in Bax Are Conserved in Bak

Unlike Bax, Bak is present at the outer mitochondrial membrane in absence of apoptotic stimuli. Evidence suggests that the inactive form of Bak gets recruited to the mitochondrial outer membrane and forms complexes with the membrane protein VDAC2. Upon apoptotic stimuli, pro-apoptotic Bcl-2 proteins such as Bid transiently bind to Bak. This binding breaks down the VDAC2/Bak complex and exposes the BH3 domain of Bak, which is essential for Bak oligomerization [42–46]. As the activation information may be conveyed by a similar charge transfer network to induce abrogation of VDAC2/Bak binding, we wondered whether a comparable transfer hub may exist also in Bak. Since residues that are essential for functionality are most often conserved in proteins with similar functions, we therefore first performed the sequence alignment of Bax and Bak. While the sequence identity between the two proteins was rather low (ClustalW2 score 19%, see Figure S2), we found that the residues

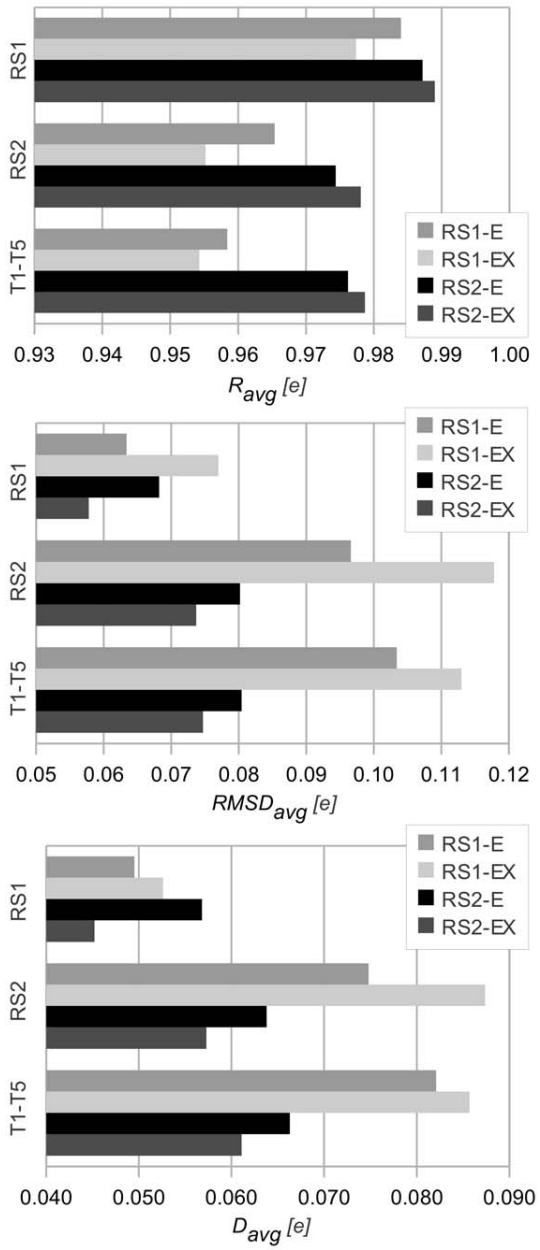


Figure 3. Validation of EEM models by comparing EEM atomic charges against QM atomic charges. Statistical descriptors comprising the average correlation coefficient (R_{avg}), the average root mean square deviation ($RMSD_{avg}$) and the average absolute difference (D_{avg}) are given. These descriptors quantify the agreement between EEM model charges and QM charges for molecules belonging to the reference sets RS1 and RS2, and for five further test molecules T1–T5. All quantities are given in elementary charges (1 $e \sim 1.602 \times 10^{-19}$ coulombs). The names of the parameter sets encode the reference set and atom classification scheme based on which they were developed (RS1-E, RS1-EX, RS2-E, RS2-EX). Good agreement between QM and EEM charges was found for all data sets, as R_{avg} is close to 1, and $RMSD_{avg}$ and D_{avg} are minimal. Calibrations that used the coarse atom type classification 'E' gave a similarly good agreement as those where the more detailed classification scheme 'EX' was used.
doi:10.1371/journal.pcbi.1002565.g003

involved in the charge transfer network in Bax were conserved in Bak. These homologous residues were Trp125, Arg127 and Arg137 (Figure 6A). We subsequently compared the 3D structures of Bax and Bak (PDB ID 2IMT [47]), and found that above Bak residues were organized in a very similar manner to their Bax homologues (Figures 5B and 6B). These findings suggest that the mechanism of charge transfer via the hydrophobic core of Bax is also plausible for Bak, and that similar residues may also play an essential role during Bak activation.

Discussion

Allosteric proteins are characterized by a regulatory site that is distinct and often well distanced from the protein's active site. Regulation of the protein's activity which occurs via this distinct site is termed allosteric regulation. Recent reports indicate that allosteric regulation is particularly important during cell signaling processes, where it has been shown to stabilize receptor proteins, or to be responsible for the rapid, stress induced release of dormant signaling proteins bound to the cytoskeleton [48,49]. An interesting structure-function analysis of Bax performed by George et al. [41] concluded that monomeric Bax may be held in an inactive conformation by multiple helices in the absence of stress, and that Bax may be activated through perturbation at multiple sites. Nevertheless, later Gavathiotis et al. identified a unique and well defined activation site on Bax [18], and subsequently demonstrated that binding of an activator BH3 peptide induces reverberations in the core of the Bax protein, a phenomenon they named allosteric sensing [19]. The present study found that this allosteric regulation is mediated by a charge transfer network, which conveys the activation information from the Bax activation site to the functional regions of Bax without compromising the structure of the BH groove (essential for pro-apoptotic activity). As charge transfer is significantly faster than domain rearrangements, the charge transfer mediated alosteric regulation in Bax also allows for a swift control of the apoptotic fate [50].

In addition to suggesting that charge transfer mediated allosteric regulation is responsible for Bax activation by pro-apoptotic Bcl-2 proteins, our charge profile analysis also indicated several residues that actively mediate this charge interaction, providing an opportunity for further in-depth mutagenesis studies or even pharmacological intervention.

We first confirmed that the abrogation of the Asp98-Ser184 interaction, which has been reported to be responsible for the mobilization of the C-domain from the BH groove [16,40], indeed occurs during Bax activation. We propose that Arg94 plays an essential role in this abrogation, as it can sequester Asp98 and prevents the formation of the stabilizing Asp98-Ser184 interaction in active Bax. Indeed, previous mutational studies [41] showed that a triple alanine Bax mutant at residues 92 to 94 is biologically inactive, supporting our finding that residue Arg94 plays a role in Bax activation.

Furthermore, we found that helix 5 acts as a central hub for the charge transfer network in Bax. We identified three residues, Trp107, Arg109 and Lys119, that may act as the main mediators of this charge transfer. Helix 5 has been found to react to Bim-SAHB binding in NMR experiments [19]. It was then found that the Bim-SAHB-induced opening of the Bax loop 1–2 is essential for Bax activity, and that this opening induces reverberations in the protein's hydrophobic core. A deeper look at the NMR data from their supplement (Figure S1D from [19]) reveals that activator binding induces pronounced chemical shifts in the Bax backbone N atoms in the area of residue Trp107 even when the mobility of loop 1–2 is restricted by chemical tethering. In Figure

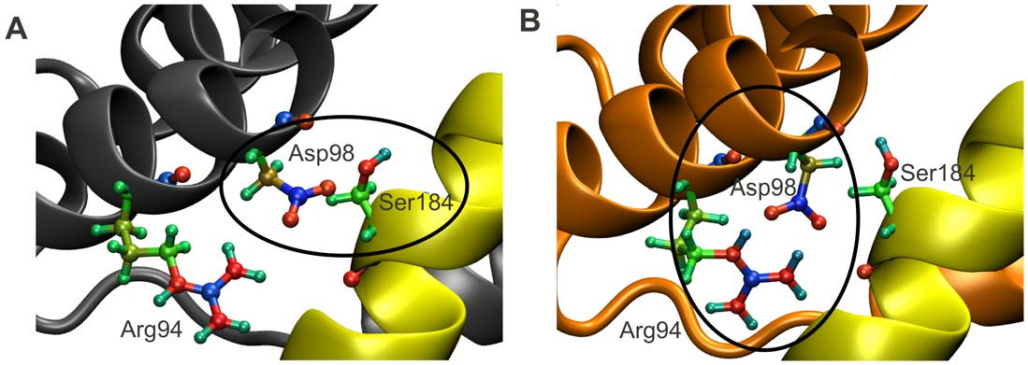


Figure 4. In active Bax, Arg94 recruits Asp98, destabilizing the C-domain inside the BH groove. Upon activation, Arg94 becomes more positive, leading to the recruitment of Asp98, abrogation of the Asp98-Ser184 interaction, and ultimately destabilization of the C-domain inside the BH groove [16,40]. The color coding from Figure 1 is maintained. Additionally, the atoms in residues Arg94, Asp98 and Ser184 are displayed explicitly. Colors are coded according to their EEM charges, where the color scale ranges from red, through green, to blue, as values of atomic charges go from negative to positive. The EEM charges were computed using parameter set RS2-E (see Figures 2 and 3). (A) In inactive Bax, Asp98 is engaged in an interaction with Ser184, which keeps the C-domain in its binding pocket. (B) In active Bax, the now more positively charged Arg94 (see also Table S1) has sequestered Asp98, which no longer contributes to the stabilization of the Bax C-domain in its BH groove.
doi:10.1371/journal.pcbi.1002565.g004

S1B from the same publication [19], we observe that opening this loop similarly affects the backbone N atoms in the vicinity of Lys119. While the authors [19] did not explicitly focus on these residues, our charge calculations make it reasonable to assume that they are indeed important for allosteric Bax activation. Moreover, another study [41] found that triple alanine Bax mutants at residues 109–111 or 118–120 showed decreased biological activity in the presence of the activator tBid. Therefore, influencing the activity of Trp107, Arg109 or Lys119 may readily influence the biological activity of Bax. Because of their positioning, residues Trp107 and Arg109 are easily accessible and therefore excellent drug targets.

The results of our investigation agree well with the mutational study of George et al [41], in that helix 5 is a central mediator of Bax activity. Both studies further agree that Arg94 is essential for Bax oligomerisation, and that residues Arg109 or Lys119 may influence the biological activity of Bax. In addition, George et al. suggested that the block of four central residues (113–116) is mandatory for Bax activity. Comparatively, the amount of charge transferred by these 4 residues upon Bax activation was only

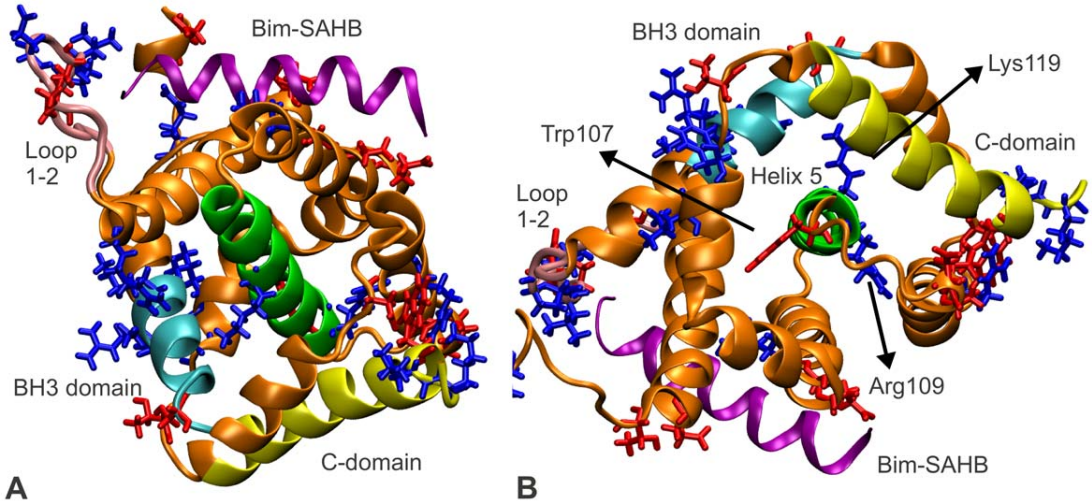


Figure 5. Proposed charge transfer network in Bax, indicated by net changes in residue charges. The information of the Bim-SAHB induced activation of Bax is transmitted from the Bax activation site via a charge transfer network through the core of the Bax protein, up to the Bax C- and BH3-domains. Inside the hydrophobic core of Bax, the central helix, helix 5, acts as a hub which collects and distributes charge density, mainly through residues Trp107, Arg109 and Lys119. The color coding from Figure 1 is maintained. Additionally, helix 5 is highlighted in green. The Bax residues which transfer an amount of charge of one standard deviation higher than average (Table S1) are explicitly displayed and color coded according to whether they become more positive (blue) or negative (red) upon activation. (A) The residues which transfer a significant amount of charge were found at the Bax activation site, on the loop 1–2, inside the BH groove holding the Bax C-domain, and at one end of the C-domain (see also Figure S1). Additionally, several such residues were found on helix 5, the central helix in Bax, and on the Bax BH3 domain, suggesting that the interaction at the Bax activation site is transmitted via a network of charges from the activation site, through the protein core, to the C- and BH3-domains. (B) Top view of helix 5 is given. The organization of residues Trp107, Arg109 and Lys119 inside the hydrophobic core of Bax suggests that helix 5 acts as a charge transfer hub, which integrates and distributes charge density.
doi:10.1371/journal.pcbi.1002565.g005

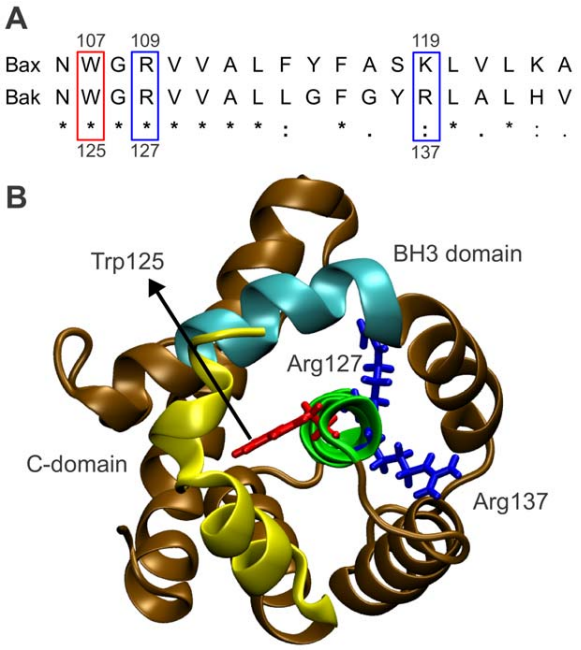


Figure 6. Proposed Charge Transfer Network in Bak. (A) Sequence alignment of central helices of Bak and Bax. An asterisk indicates a single, fully conserved residue. A colon indicates conservation between groups of strongly similar biochemical properties. A period indicates conservation between groups of weakly similar biochemical properties. The residues involved in the charge transfer network in Bax are conserved in Bak as Trp125, Arg127 and Arg137. (B) Bak structure (ochre) is displayed according to Figure 5B, with the same top view of the central helix, and the same color coding for C-domain (yellow), BH3-domain (cyan), central helix (green) and hub residues (red and blue). Residues Trp125, Arg127 and Arg137 are organized in a similar manner to their Bax homologues, suggesting that they may also play an essential role during Bak activation.
doi:10.1371/journal.pcbi.1002565.g006

slightly above average (Table S1). Nevertheless, we note that this block of residues resides at the centre of the Bax molecule and is very bulky, and therefore these residues are very likely essential for maintaining the helical fold of Bax. Therefore, the observation of George et al. that a quadruple mutation in the centre of the Bax molecule impairs biological activity can be easily explained by change of stability of the helical fold, rather than by a disruption of the Bax activation mechanism.

Finally, by performing sequence and structural alignment of Bax and Bak, we identified that Bak residues Trp125, Arg127, and Arg137 potentially constitute a similar hub of charge transfer inside the Bak protein. The membrane protein VDAC2 was reported to recruit Bak to the mitochondrial membrane in the absence of apoptotic stimuli. Upon apoptotic stimuli, this VDAC2/Bak binding can be abrogated by Bcl-2 proteins that transiently bind to Bak [42–46]. Cheng et al. [42] showed that Bak mutations at residues Leu78 (within Bak's BH3 domain), Trp125, Gly126 and Arg127 (on Bak's central helix) impair complex formation with VDAC2, and thus concluded that VDAC2 binding to Bak must occur in the proximity of the above mentioned residues. Having identified those residues that may constitute a transfer charge hub during Bak activation, we propose a charge transfer mediated allosteric activation mechanism of Bak that is similar to that of Bax. We propose that transient activator binding in the vicinity of Arg137 is transmitted through allosteric sensing to the VDAC2 binding site, which is in the neighborhood of Trp125 and Arg127. Subsequently, Bak disengages VDAC2, exposes its BH3 domain and oligomerizes. Since the relevance of residue Arg137 has not been assessed to date, its investigation may further our understanding of the details of Bak activation and its de-repression of VDAC2.

Understanding the structural changes of Bax and Bak during apoptosis provides important insights into the mechanisms of Bax and Bak activation, which steps and key players are involved, how aberrant protein folding or mutations may influence the protein's function, and how their activation may be influenced by inhibitors or synthetic drugs. While X-ray crystallography and NMR spectroscopy provide excellent experimental techniques to obtain 3D-structures, further theoretical data analysis tools are needed to obtain better mechanistic and functional insights into the structural aspects of protein activation. We report here the first successful application of the Electronegativity Equalization Method to study protein activation during programmed cell death, which enabled us to detect and track the allosteric effects responsible for Bax activation by BH3-only proteins. We have thus shown how knowledge of atomic charges can yield insight into biological phenomena even without further simulations or intricate computations. Moreover, the methodology we developed is directly applicable to other molecular systems, and thus of interest in biomedical and pharmacological research.

Methods

EEM Formalism

The Electronegativity Equalization Method (EEM) [26] was employed here to estimate the charge transfer upon Bax activation using structural data obtained from the Protein Data Bank (PDB). EEM enables the determination of connectivity- and geometry-dependent atomic charges. Various formalisms are available in literature [32–38,51–53]. In the present implementation, we focused on the original work by Mortier et al. [26] with a minor modification as suggested by Yang and Wang [54] and described below. Besides enabling a fast and versatile calibration, this formalism estimates atomic charges via a set of coupled linear equations which can be efficiently solved by a Gaussian elimination procedure (see below).

EEM is based on the Electronegativity Equalization Principle [55], which was proven within the Density Functional Theory [56] and which states that electronegativity $\bar{\chi}$ is equalized throughout a molecule ($\bar{\chi}=\chi_i$). The electronegativity χ_i of each atom i in this molecule can be approximated as a linear function of several terms:

$$\chi_i = \left(\chi_i^0 + \Delta\chi_i\right) + 2\left(\eta_i^0 + \Delta\eta_i\right)q_i + k\sum_{i\neq j}\frac{q_j}{r_{ij}}.$$

The first term is the effective electronegativity (i.e., the electronegativity χ_i^0 of the neutral atom, corrected for the presence of the molecular environment $\Delta\chi_i$). The second term is the charge of the atom q_i multiplied by its effective hardness (i.e., the hardness η_i^0 of the neutral atom corrected for the presence of the molecular environment $\Delta\eta_i$). Hardness was defined by Parr and Pearson [57] as the second derivative of the energy with respect to the charge, and can be thought of as the resistance to change in charge. The last term $k\sum_{i\neq j}\frac{q_j}{r_{ij}}$ accounts for the electrostatic interaction with

every other charged atom j in the molecule. k is an adjusting factor first introduced by Yang and Wang [54].

Setting $A_i=\chi_i^0+\Delta\chi_i$ and $B_i=2\left(\eta_i^0+\Delta\eta_i\right)$, the molecular electronegativity can be written as:

$$\overline{\chi}=A_i+B_iq_i+k\sum_{i\neq j}\frac{q_j}{r_{ij}}.$$

Considering the total molecular charge Q to be the sum of all partial atomic charges q_i ($Q=\sum q_i$), a system of equations results, from which the partial atomic charges q_i and the molecular electronegativity $\overline{\chi}$ can be calculated:

$$\begin{pmatrix} B_1 & \frac{k}{R_{1,2}} & \cdots & \frac{k}{R_{1,N}} & -1 \\ \frac{k}{R_{2,3}} & B_2 & \cdots & \frac{k}{R_{2,N}} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{k}{R_{N,1}} & \frac{k}{R_{N,2}} & \cdots & B_N & -1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix}\cdot\begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ \overline{\chi} \end{pmatrix}=\begin{pmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_N \\ Q \end{pmatrix}.$$

Calibration of the EEM Model

The corrections for electronegativity $\Delta\chi_i$ and hardness $\Delta\eta_i$ cannot be measured [26]. Therefore, the effective electronegativity and hardness contributions given by $A_i=\chi_i^0+\Delta\chi_i$ and $B_i=2\left(\eta_i^0+\Delta\eta_i\right)$ respectively were calibrated in this study. The additional parameter k was also determined, as it allows for a computationally cheap and straightforward sampling of the *(A,B)* parameter space, as previously demonstrated by Svobodová Vařeková et al. [32].

The EEM equation of the molecular electronegativity can be rearranged as a linear equation in A and B for each atom in the system:

$$A_i+B_iq_i=\overline{\chi}-k\sum_{i\neq j}\frac{q_j}{r_{ij}}.$$

The above linear equations can be grouped together according to the type of atom they refer to, as each parameter will be valid only for a particular type of atom. The classification of atoms into types can be done according to various criteria. As schemes in literature use different levels of granularity [e.g., 32,33,35], two schemes of atom classification were tested in the present study (see also the Results section).

For each atom type, the parameters A and B can be determined by least squares minimization, provided that the values of all the other variables in the equation are known. Here the interatomic distances were calculated from the 3D atomic coordinates. The reference atomic charges were calculated by the QM scheme described below. For each molecule, the value of the global electronegativity was approximated as the harmonic average of the electronegativities of its constituent atoms [58]:

$$\overline{\chi}=n\left(\sum_{i=1}^{n}\frac{1}{\chi_i^0}\right)^{-1},$$

where n is the number of atoms in the molecule, and the values of χ_i^0 correspond to Pauling electronegativities [59,60]. The extra parameter k present in this particular formalism was sampled on several intervals. For each discrete value of k, the least squares minimization was performed in order to obtain the $(A,B)_x$ parameters, where x goes over all atom types considered. Upon internal validation, the result was the set of parameters $[k,(A,B)_x]$ which gives the best R_{avg} (see below) between the reference QM values and the predicted EEM values for atomic charges. A scheme of the calibration step is given in Figure 2B, while a detailed description of this procedure can be found in the work of Svobodová Vařeková et al. [32].

Reference Data Used for EEM Model Calibration

Obtaining appropriate reference data is essential for the accuracy and applicability of a predictive model. The reference data used in this study consisted of atomic charges for molecules of two disjoint reference sets RS1 and RS2, and these charges were calculated using quantum mechanics (QM) (see below). These reference molecules were fragments extracted from calcium containing proteins which were obtained from PDB and whose structures had been determined by X-ray crystallography or solution state NMR experiments. Each of the fragments consisted of amino acid chains, calcium ions and water molecules, and was obtained from the 3D structure of its parent protein using the program Triton [61]. The fragments were curated manually to ensure that, while they are sufficiently small for QM calculations, they remained biochemically meaningful. For each fragment, reference QM atomic charges were obtained from a Mulliken population analysis performed at the HF/6-31G* theory level using the program Gaussian 03 [62].

An overview of the composition of all fragments used for EEM model calibration is given in Table 1, while the 3D structures of these fragments are available online in PDB format at www.ncbr.muni.cz/~ionescu/Supporting_Data_Sets.zip. Reference sets RS1 and RS2 were used for model calibration, and internal and external validation (see below). Five additional test molecules T1-T5 were used for external validation. A brief summary regarding the reference data is given in Figure 2A.

Validation of the EEM Model

The accuracy of the EEM models in reproducing the original QM charges from reference sets RS1 and RS2, and from five

Table 1. Summary of the atomic composition of all the protein fragments used for EEM model calibration.

System	RS1	RS2	T1	T2	T3	T4	T5
Atoms	23259	42295	1167	1125	1065	1040	1075
Fragments	57	43	1	1	1	1	1
C	6649	12522	350	342	305	298	260
H	10434	20995	562	505	498	503	558
N	2730	3339	117	129	123	104	75
O	3332	5225	136	144	136	133	176
S	30	149	0	3	2	0	6
Ca	84	65	2	2	1	2	0

Reference sets RS1 and RS2 were used for model calibration, and internal and external validation. Five additional test molecules T1–T5 were used for external validation.

doi:10.1371/journal.pcbi.1002565.t001

additional test molecules T1–T5 was evaluated by internal and external validation. In the internal validation step, the charges predicted by the EEM model with parameter sets RS1-E and RS1-EX (RS2-E and RS2-EX respectively) were compared against QM charges from the associated reference set RS1 (RS2 respectively). In the external validation step, EEM and QM charges were compared for five test molecules T1–T5 which were not contained in the original reference sets RS1 and RS2, but were obtained in a similar manner. Since the reference sets RS1 and RS2 were disjoint, two further external validations were performed. Therefore, EEM charges obtained by using parameter sets RS1-E and RS1-EX (RS2-E and RS2-EX respectively) were compared against QM charges from the non associated reference set RS2 (RS1 respectively). A schematic representation of the EEM model validation step is given in Figure 2C.

The correlation between the sets of QM and EEM charges was assessed by three indicators. The first indicator was the average correlation coefficient (squared Pearson's correlation coefficient), computed for each molecule, and averaged over all molecules in a set:

$$R_{avg} = \frac{\sum_{I=1}^{N} \left(\frac{\sum_{i=1}^{n_I} \left(q_i^{QM} - \overline{q_I^{QM}}\right)\left(q_i^{EEM} - \overline{q_I^{EEM}}\right)}{(n_I - 1)\sigma_I^{QM}\sigma_I^{EEM}} \right)^2}{N},$$

where the index $i = 1 \ldots N$ described all atoms in molecule I, $\overline{q_I^{QM}}$ and $\overline{q_I^{EEM}}$ represented average atomic charges in molecule I, σ_I^{QM} and σ_I^{EEM} were standard deviations of the atomic charges in molecule I, n_I was the number of atoms in molecule I, and N was the number of molecules in a given set.

The second indicator was the root mean square deviation, computed for each molecule, and averaged over all molecules in a set:

$$RMSD_{avg} = \frac{\sum_{I=1}^{N} \sqrt{\frac{\sum_{i=1}^{n_I} \left(q_i^{QM} - q_i^{EEM}\right)^2}{n_I}}}{N}.$$

The third indicator was the average absolute difference, computed for each molecule and averaged over all molecules in a set:

$$D_{avg} = \frac{\sum_{I=1}^{N} \frac{\sum_{i=1}^{n_I} |q_i^{QM} - q_i^{EEM}|}{n_I}}{N}.$$

Evaluating Differences in Charges upon Bax Activation

The EEM charge calculations for both Bax structures were done using the program EEM_SOLVER [39] which implemented the above mentioned EEM formalism and employed the parameter set RS2-E developed in the present study.

Two indicators were employed in order to quantify the changes in the charge profile of Bax upon activation. The first indicator was the total difference in charge per amino acid residue:

$$\Delta Q_{res} = \frac{\sum_{i=1}^{n_{res}} \left(q_i^{active} - q_i^{inactive}\right)}{n_{res}},$$

where q_i^{active} denoted atomic charges in the active Bax, $q_i^{inactive}$ denoted atomic charges in the inactive Bax, and n_{res} was the number of atoms in the residue. ΔQ_{res} assessed the amount of charge that had been transferred to or from each residue. The second indicator was the root mean square deviation in atomic charge per residue:

$$RMSD_{res} = \sqrt{\frac{\sum_{i=1}^{n_{res}} \left(q_i^{active} - q_i^{inactive}\right)^2}{n_{res}}}.$$

$RMSD_{res}$ assessed the intra-residue charge density redistributions.

Sequence and Structural Alignment between Bax and Bak

The Bax/Bak sequence alignment was done for the UniProtKB/Swiss-Prot entries Q07812 (BAX_HUMAN) and Q16611 (BAK_HUMAN), and was performed using ClustalW2 with default parameters on the EBI server [63]. The structural models were visualized using VMD [64].

Supporting Information

Figure S1 Activator binding induces significant reorganization of intra-residue charge density in the functional regions of Bax. Upon activator (Bim-SAHB) binding to Bax, significant reorganization of the intra-residue charge density is observed in the functional regions of Bax, suggesting that the activation is conveyed across the entire Bax molecule. The color coding from Figure 4 is maintained, with the C-domain in yellow, BH3 domain in cyan, central helix in green, the rest of active Bax in orange, and Bim-SAHB in purple. Additionally, the amino acid residues which suffer significant redistributions of their charge density are displayed explicitly ($RMSD_{res}$ one standard deviation higher than average; see Table S1). These residues can be found at the Bax activation site, on loop 1–2, inside the BH groove holding the Bax C-domain, and at the two ends of the C-domain itself. (A) Side view is given. (B) Top view of helix 5 is given.
(PDF)

Figure S2 Sequence alignment between Bax and Bak reveals rather low sequence identity (ClustalW2 Score 19%). An asterisk indicates a single, fully conserved residue. A colon indicates conservation between groups of strongly similar biochemical properties. A period indicates conservation between groups of weakly similar biochemical properties. The sequence alignment for the central helices and the structural alignment are given in Figure 6.
(PDF)

Table S1 Changes in the charge profile of all Bax residues upon Bax activation. Net charge transfer was computed as the total difference in charge per residue (Q_{res}). The intra-residue charge density redistributions were evaluated as the root mean square deviation in charge per residue ($RMSD_{res}$). Both descriptors were computed using EEM atomic charges, and their mathematical derivation can be found in the Methods section. All quantities are given in elementary charges (1 e has approximately 1.602×10^{-19} coulombs). The cell background colors mark the

various domains of the Bax molecule in agreement with Figure 1 (BH3-domain in cyan, C-domain in yellow, loop 1–2 in pink, and helix 5 in green). The residues which exhibited a net charge transfer of more than one standard deviation over the average are marked in bold, and the color of the font indicates whether the respective residues became more positive (red) or more negative (blue) upon activation. These residues are also displayed explicitly in Figure 5.
(PDF)

Acknowledgments

The authors would like to thank Ms. Niamh M. Connolly and Mr. Sushil Kumar Mishra for thorough revision of the original manuscript.

Author Contributions

Conceived and designed the experiments: RSV HJH JK. Performed the experiments: CMI. Analyzed the data: CMI RSV JHMP HJH JK. Contributed reagents/materials/analysis tools: RSV HJH. Wrote the paper: CMI RSV JHMP HJH JK.

References

1. Kroemer G, Galluzzi L, Brenner C (2007) Mitochondrial Membrane Permeabilization in Cell Death. Physiol Rev 87: 99–163.
2. Wei MC, Cheng EH-Y, Zong W-X, Lindsten T, Panoutsakopoulou V, et al. (2001) Proapoptotic BAX and BAK: A Requisite Gateway to Mitochondrial Dysfunction and Death. Science 292: 727–730.
3. Kuwana T, Newmeyer DD (2003) Bcl-2 family proteins and the role of mitochondria in apoptosis. Curr Opin Cell Biol 15: 691–699.
4. Tait SW, Green DR (2010) Mitochondria and cell death: outer membrane permeabilization and beyond. Nat Rev Mol Cell Biol 11: 621–632.
5. Letai A, Bassik MC, Walensky LD, Sorcinelli MD, Weiler S, et al. (2002) Distinct BH3 domains either sensitize or activate mitochondrial apoptosis, serving as prototype cancer therapeutics. Cancer Cell 2: 183–192.
6. Marsden VS, Strasser A (2003) Control of apoptosis in the immune system: Bcl-2, BH3-only proteins and more. Annu Rev Immunol 21: 71–105.
7. Leber B, Lin J, Andrews DW (2007) Embedded Together: The Life and Death Consequences of Interaction of the Bcl-2 Family with Membranes. Apoptosis 12: 897–911.
8. Chipuk JE, Green DR (2008) How do BCL-2 proteins induce mitochondrial outer membrane permeabilization? Trends Cell Biol 18: 157–164.
9. Oltvai ZN, Milliman CL, Korsmeyer SJ (1993) Bcl-2 heterodimerizes in vivo with a conserved homolog, Bax, that accelerates programmed cell death. Cell 74: 609–619.
10. Willis SN, Fletcher JI, Kaufmann T, van Delft MF, Chen L, et al. (2007) Apoptosis initiated when BH3 ligands engage multiple Bcl-2 homologs, not Bax or Bak. Science 315: 856–859.
11. Eskes R, Desagher S, Antonsson B, Martinou JC (2000) Bid induces the oligomerization and insertion of Bax into the outer mitochondrial membrane. Mol Cell Biol 20: 929–935.
12. Kuwana T, Bouchier-Hayes L, Chipuk JE, Bonzon C, Sullivan BA, et al. (2005) BH3 domains of BH3-only proteins differentially regulate Bax-mediated mitochondrial membrane permeabilization both directly and indirectly. Mol. Cell 17, 525–535.
13. Walensky LD, Pitter K, Morash J, Oh KJ, Barbuto S, et al. (2006) A stapled BID BH3 helix directly binds and activates BAX. Mol Cell 24: 199–210.
14. Wolter KG, Hsu Y-T, Smith CL, Nechushtan A, Xi XG, Youle RJ (1997) Movement of Bax from the cytosol to mitochondria during apoptosis. J Cell Biol 139: 1281–1292.
15. Hsu Y-T, Wolter KG, Youle RJ. (1997) Cytosol-to-membrane redistribution of Bax and Bcl-xL during apoptosis. Proc Natl Acad Sci U S A 94: 3668–3672.
16. Suzuki M, Youle RJ, Tjandra N (2000) Structure of Bax: coregulation of dimer formation and intracellular localization. Cell 103: 645–654.
17. Lovell JF, Billen LP, Bindner S, Shamas-Din A, Fradin C, et al. (2008) Membrane binding by tBid initiates an ordered series of events culminating in membrane permeabilization by Bax. Cell 135: 1074–1084.
18. Gavathiotis E, Suzuki M, Davis ML, Pitter K, Bird GH, et al. (2008) BAX activation is initiated at a novel interaction site. Nature 455: 1076–1081.
19. Gavathiotis E, Reyna DE, Davis ML, Bird GH, Walensky LD (2010) BH3-triggered structural reorganization drives the activation of proapoptotic BAX. Mol Cell 40: 481–492.
20. Czabotar PE, Colman PM, Huang DCS (2009) Bax activation by Bim? Cell Death Differ 16: 1187–1191.
21. Westphal D, Dewson G, Czabotar PE, Kluck RM (2011). Molecular biology of Bax and Bak activation and action. Biochim Biophys Acta 1813: 521–531.
22. Van der Vaart A, Bursulaya BD, Brooks CL, III, Merz KM, Jr (2000) Are Many-Body Effects Important in Protein Folding? J Phys Chem B 104: 9554–9563.
23. Cho AE, Guallar V, Berne BJ, Friesner R (2005) Importance of Accurate Charges in Molecular Docking: Quantum Mechanical/Molecular Mechanical (QM/MM) Approach J Comput Chem 26: 915–931.
24. Bucher D, Raugei D, Guidoni L, Dal Peraro M, Rothlisberger U, et al. (2006) Polarization effects and charge transfer in the KcsA potassium channel. Biophys Chem 124: 292–301.
25. Anisimov VM, Cavasotto CN (2010) Quantum-Mechanical Molecular Dynamics of Charge Transfer. In: Paneth P, Dybala-Defratyka A. Kinetics and Dynamics: From Nano- to Bio-scale. Springer. pp. 247–266.
26. Mortier WJ, Ghosh SK, Shankar S (1986) Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. J Am Chem Soc 108: 4315–4320.
27. Rappe AK, Goddard WA, III (1991) Charge Equilibration for Molecular Dynamics Simulations. J Phys Chem 95: 3358–3363.
28. Smirnov KS (2001) Computer Modeling Study of Interaction of Acetonitrile with Hydroxyl Groups of HY Zeolite. J Phys Chem B 105: 7405–7413.
29. Cong Y, Yang Z-Z, Wang C-S, Liu X-C, Bao X-H (2002) Investigation of the regio- and stereoselectivity of Dies-Alder reactions by newly developed ABEEM $\sigma\pi$ model on the basis of local HSAB principle and maximum hardness principle. Chem Phys Lett 357: 59–64.
30. Shimizu K, Chaimovich H, Farah JPS, Dias LG, Bostick DL (2004) Calculation of the dipole moment for polypeptides using the generalized born-electronegativity equalization method: Results in vacuum and continuum-dielectric solvent. J Phys Chem B 108: 4171–4177.
31. Wallin G, Nervall M, Carlsson J, Aqvist J (2009) Charges for Large Scale Binding Free Energy Calculations with the Linear Interaction Energy Method. J Chem Theory Comput 5: 380–395.
32. Svobodova RV, Jiroušková Z, Vaněk J, Suchomel Š, Koča J (2007) Electronegativity Equalization Method: Parameterization and Validation for Large Sets of Organic, Organohalogene and Organometal Molecules. Int J Mol Sci 8: 572–582.
33. Bultinck P, Langenaeker W, Lahorte P, De Proft F, Geerlings P, et al. (2002) The Electronegativity Equalization Method I: Parameterization and Validation for Atomic Charge Calculations. J Phys Chem 106: 7887–7894.
34. Berente I, Czinki E, Naray-Szabo G (2007) A Combined Electronegativity Equalization and Electrostatic Potential Fit Method for the Determination of Atomic Point Charges. J Comput Chem 28: 1936–1942.
35. Kang YK, Scheraga HA (2008) An Efficient Method for Calculating Atomic Charges of Peptides and Proteins from Electronic Populations. J Phys Chem B 112: 5470–5478.
36. Verstraelen T, Van Speybroeck V, Waroquier M (2009) The electronegativity equalization method and the split charge equilibration applied to organic systems: Parameterization, validation, and comparison. J Chem Phys 131: 044127–19
37. Purannen JS, Vainio MJ, Johnson MS (2009) Accurate conformation-dependent molecular electrostatic potentials for high-throughput in silico drug discovery. J Comp Chem 31: 1722–1732.
38. Ouyang Y, Ye F, Liang Y (2009) A modified electronegativity equalization method for fast and accurate calculation of atomic charges in large biological molecules. Phys Chem Chem Phys 11: 6082–6089.
39. Svobodová Vařeková R, Koča J. (2005) Optimized and Parallelized Implementation of the Electronegativity Equalization Method and the Atom-Bond Electronegativity Equalization Method. J Comput Chem 27: 396–405.
40. Nechushtan A, Smith CL, Hsu Y-T, Youle RJ (1999) Conformation of the Bax C-terminus regulates subcellular location and cell death. EMBO J 18: 2330–2341.
41. George NM, Evans JJD, Luo X (2007) A three-helix homo-oligomerization domain containing BH3 and BH1 is responsible for the apoptotic activity of Bax. Genes Dev 21:1937–1948.
42. Cheng EH-Y, Sheiko TV, Fisher JK, Craigen WJ, Korsmeyer SJ (2003) VDAC2 Inhibits BAK Activation and Mitochondrial Apoptosis. Science 301: 513–517.
43. Roy SS, Ehrlich AM, Craigen WJ, Hajnoczky G (2009) VDAC2 is required for truncated BID-induced mitochondrial apoptosis by recruiting BAK to the mitochondria. EMBO Rep 10: 1341–1347.
44. Dewson G, Kratina T, Sim HW, Puthalakath H, Adams JM, et al. (2008) To Trigger Apoptosis, Bak Exposes Its BH3 Domain and Homodimerizes via BH3:Groove Interactions. Mol Cell 30: 369–380.
45. Lazarou M, Stojanovski D, Frazier AE, Kotevski A, Dewson G, et al. (2010) Inhibition of Bak activation by VDAC2 is dependent on the Bak transmembrane anchor. J Biol Chem 285: 36876–36883.
46. Dai H, Smith A, Meng XW, Schneider PA, Pang YP, et al. (2011) Transient binding of an activator BH3 domain to the Bak BH3-binding groove initiates Bak oligomerization. J Cell Biol 194: 39–48.
47. Moldoveanu T, Liu Q, Tocilj A, Watson M, Shore G, et al. (2006) The X-ray structure of a BAK homodimer reveals an inhibitory zinc binding site. Mol Cell 24: 677–688.
48. Tsai CJ, del Sol A, Nussinov R (2008) Allostery: absence of a change in shape does not imply that allostery is not at play. J Mol Biol 378: 1–11.
49. Bu Z, Callaway DJ (2011) Proteins move! Proteins dynamics and long-range allostery in cell signaling. Adv Protein Chem Struct Biol 83: 163–221.

50. Düssmann H, Rehm M, Concannon CG, Anguissola S, Würstle M, et al. (2010) Single-cell quantification of Bax activation and mathematical modelling suggest pore formation on minimal mitochondrial Bax accumulation. Cell Death Differ 17: 278–290.
51. Cong Y, Yang Z-Z (2000) General atom-bond electronegativity equalization method and its application in prediction of charge distributions in polypeptides. Chem Phys Lett 316: 324–329.
52. Menegon G, Shimizu K, Farah JPS, Dias LG, Chaimovich H (2002) Parameterization of the electronegativity equalization method based on the charge model 1. Phys Chem Chem Phys 4: 5933–5936.
53. Chaves J, Barroso JM, Bultinck P, Carbó-Dorca R (2005) Toward an Alternative Hardness Kernel Matrix Structure in the Electronegativity Equalization Method (EEM). J Chem Inf Model 46: 1657–1665.
54. Yang Z-Z, Wang C-S (1997) Atom-Bond Electronegativity Equalization Method. 1. Calculation of the Charge Distribution in Large Molecules. J Phys Chem A 101: 6315–6321.
55. Sanderson RT (1951) An Interpretation of Bond Lengths and a Classification of Bonds. Science 114: 670–672.
56. Parr RG, Donnelly RA, Levy M, Palke WE (1978) Electronegativity: the density functional viewpoint. J Chem Phys 68: 3801–3807.
57. Parr RG, Pearson RG (1983) Absolute Hardness: Companion Parameter to Absolute Electronegativity. J Am Chem Soc 105: 7512–7516.
58. Wilson MS, Ichikawa S (1989) Comparison between the Geometric and Harmonic Mean Electronegativity Equilibration Techniques. J Phys Chem 93: 3087–3089.
59. Pauling L (1932) The Nature of the Chemical Bond. IV. The Energy of Single Bonds and the Relative Electronegativity of Atoms. J Am Chem Soc 54: 3570–3582.
60. Allred AL (1961) Electronegativity values from thermochemical data. J Inorg Nucl Chem 17: 215–221.
61. Prokop M, Adam J, Kriz Z, Wimmerova M, Koca J (2008) TRITON: a graphical tool for ligand-binding protein engineering. Bioinformatics 24: 1955–1956.
62. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, et al. (2004) Gaussian 03 (Gaussian, Inc, Wallingford, CT), Revision C.02.
63. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) ClustalW and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
64. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. J Mol Graphics 14: 33–38: 27–38

MOLEonline 2.0: interactive web-based analysis of biomacromolecular channels

W222–W227 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 2 May 2012*
doi:10.1093/nar/gks363

MOLE*online* 2.0: interactive web-based analysis of biomacromolecular channels

Karel Berka[1], Ondřej Hanák[1], David Sehnal[2,3], Pavel Banáš[1], Veronika Navrátilová[1], Deepti Jaiswal[2], Crina-Maria Ionescu[2], Radka Svobodová Vařeková[2], Jaroslav Koča[2,*] and Michal Otyepka[1,*]

[1]Regional Centre of Advanced Technologies and Materials, Department of Physical Chemistry, Faculty of Science, Palacky University Olomouc, tr. 17. listopadu 12, 771 46 Olomouc, [2]Central European Institute of Technology and National Centre for Biomolecular Research, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic and [3]Faculty of Informatics, Masaryk University Brno, Botanická 68a, 602 00 Brno, Czech Republic

Received January 30, 2012; Revised April 6, 2012; Accepted April 10, 2012

ABSTRACT

Biomolecular channels play important roles in many biological systems, e.g. enzymes, ribosomes and ion channels. This article introduces a web-based interactive MOLE*online* 2.0 application for the analysis of access/egress paths to interior molecular voids. MOLE*online* 2.0 enables platform-independent, easy-to-use and interactive analyses of (bio)macromolecular channels, tunnels and pores. Results are presented in a clear manner, making their interpretation easy. For each channel, MOLE*online* displays a 3D graphical representation of the channel, its profile accompanied by a list of lining residues and also its basic physicochemical properties. The users can tune advanced parameters when performing a channel search to direct the search according to their needs. The MOLE*online* 2.0 application is freely available via the Internet at http://ncbr.muni.cz/mole or http://mole.upol.cz.

INTRODUCTION

Tunnels or channels, pores, cavities and voids are structural features of many biomolecular systems possessing significant biological functions. The following are just a few of the numerous examples where channels play an important biological function; highly selective ion channels (1–6), channels and pathways in photosystem II (7,8), ribosomal polypeptide exit channel (9), substrate-determining active site access channels of Cytochrome P450 (10–15) and haloalkane dehalogenases, where mutagenesis of substrate access channels alters enzyme activity (16,17). As an empty interior space is a key feature of this type of biomolecule, a considerable amount of attention has been paid to analyzing its properties (18–20). Many algorithms and software tools have been developed to identify these structures in (bio)macromolecules, including grid (16,21–24), space filling (25) and slice methods (26,27), and Voronoi diagrams (18,28–30).

CAVER (16), MOLE (28), MolAxis (29,30) and PROPORES (31) are all dedicated software tools for analyzing molecular channels. CAVER 1.0 (16) involves grid nodes evaluated by a cost function based on the square of reciprocal distance to the closest atom, and then employs the Dijskstra's algorithm (32) to select the shortest and most geometrically convenient pathway from an internal to external point. In 2005, CAVER 1.0 represented a considerable advance in the automatic detection of channels. However, its algorithm suffered from several limitations, which have since been overcame in the later issued software named MOLE (28). The core of the MOLE 1.0 algorithm again utilizes the Dijkstra's path search algorithm, which is applied to a Voronoi mesh (33,34). A later published software, MolAxis (29,30), uses an algorithm similar to MOLE. Another recent tool, PROPORES (31), searches for channels in a similar fashion to CAVER, but it also rotates side chains along the channel so that they adopt sterically allowed positions in order to enlarge possible bottlenecks.

This article presents the web-based MOLE*online* application (ver. 2.0), which offers a user-friendly, interactive and platform-independent environment for the setup, manipulation, analysis and printing of channel search results. Besides structural features, MOLE*online* also allows analysis of the basic physicochemical properties of (bio)macromolecular channels, tunnels and pores.

*To whom correspondence should be addressed. Tel: +420 585634756; Fax: +420 585634761; Email: Michal.Otyepka@upol.cz
Correspondence may also be addressed. to Jaroslav Koča. Tel: +420 549492685; Fax: +420 549491060; Email: Jaroslav.Koca@ceitec.muni.cz

DESCRIPTION OF THE TOOL

The procedure in using the MOLE*online* 2.0 application involves three steps: (i) setup; (ii) calculation; and (iii) results visualization and manipulation.

Setup

The structure to be analyzed can be either taken from the Protein data bank (PDB) server (35) or uploaded in the PDB format. Once the structure is uploaded, it is visualized by the Jmol Java plugin (36). In addition, the sequence corresponding to the structure can be explored in an interactive window, enabling selection of the starting residues (Figure 1). MOLE*online* enables the user to define the starting point based on the center of mass of selected residues, either by selection from the sequence or manually by selection of *x*, *y* and *z* coordinates. In the case of known and annotated enzyme structures, MOLE*online* allows the use of information on the active site residues from the catalytic site atlas (CSA) database (37) and use of biological unit instead of asymmetric one. The last possibility is to use 'Automatic starting points'. These points are the deepest points in the protein's cavities and using them can provide primary information on the layout of channels inside a protein.

Calculation

After setup, the calculation of channels is executed by the MOLE 2.0 software (D. Sehnal *et al.*, unpublished data) running on a server. All setup and structure information are deposited on the server in a unique directory (which is translated as a unique URL for a web browser). After the MOLE 2.0 calculation, further analyses of the channel results are carried out, providing comprehensive and easily interpretable information about the channels (see below).

The channel computation in the MOLE 2.0 software is performed in several steps as follows:

(1) the Voronoi diagram is computed;
(2) the Voronoi diagram is refined and split into several smaller parts called cavity diagrams, representing all the empty space in the molecule;
(3) starting and ending points are identified in each of the cavity diagrams; and
(4) Dijkstra's shortest path algorithm is used to find the channels between the pairs of starting and ending points.

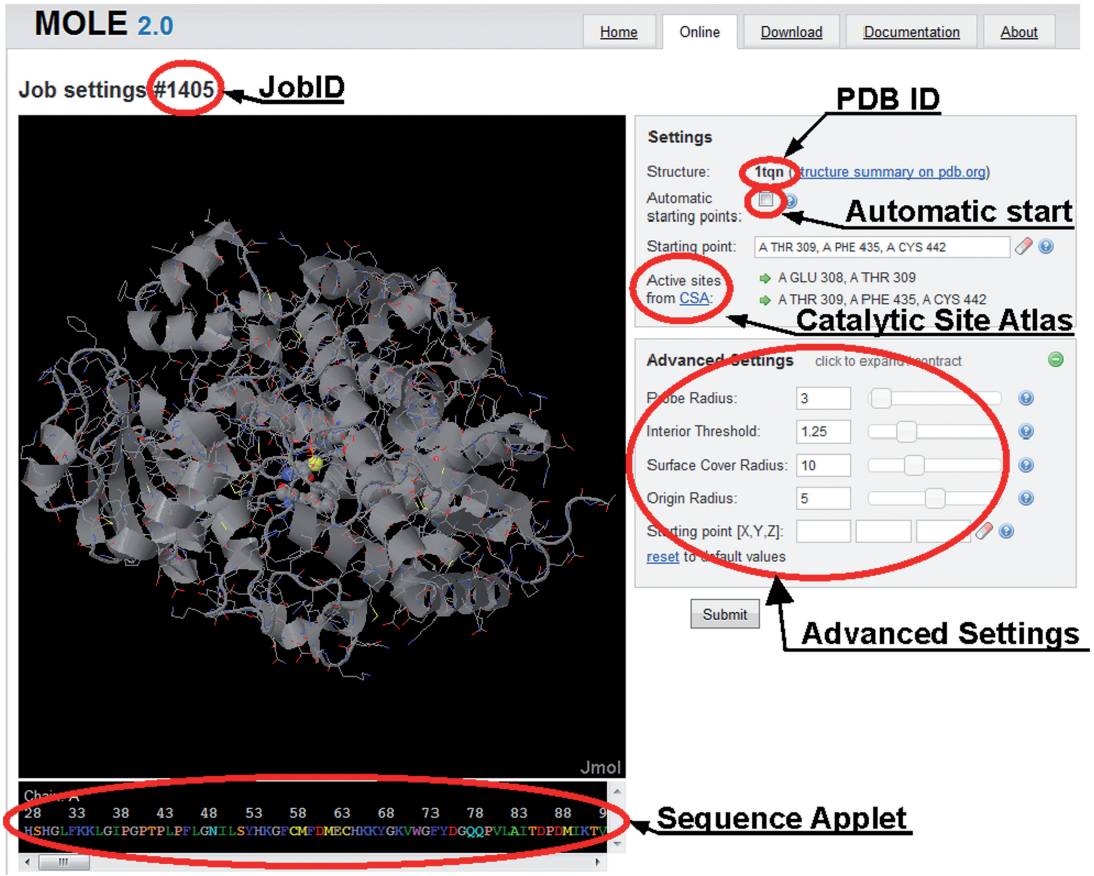


Figure 1. MOLE*online* 2.0 setup webpage for channel calculation. Each job is assigned a job ID to allow easy access to the results. Setup starts with the selection of a PDB file (here 1TQN) either from the PDB database or uploaded as a user file. The tunnel starting point can be selected automatically (inside cavities detected by MOLE 2.0 algorithm) or manually, by using CSA (37), via selection through the interactive sequence applet on the bottom of the page or by specifying of *x*, *y*, *z* coordinates in advanced settings. Advanced settings also enable the adjustment of parameters determining the tunnel searching algorithm. All parameters are set in Ångströms (for details see the text).

The Voronoi diagram divides a metric space according to the distances between discrete sets of specified objects. In our case, the objects are atomic centers with van der Waals (vdW) radii assigned according to the parm99 force field (38). Molecular surface is calculated as a probe accessible surface with a defined *probe radius* (default 3 Å). A vertex of the Voronoi diagram is removed if a sphere with *interior threshold* (default 1.25 Å) radius cannot pass through any of the tetrahedron sides. The Voronoi diagram is split into several smaller cavity diagrams that are analyzed for suitable channel start and end points between vertices of the cavity. Starting points are initially estimated by considering a centroid from all the corresponding atomic centers of the residues selected by the user. Starting points are then selected within a specified *origin radius* (default 3 Å) as the closest vertex for each cavity. End points are selected for each cavity diagram as the tetrahedra on the boundary vertices. Channel exits can only be assigned to those tetrahedra that are separated by a distance equivalent to the *surface cover radius* (default 10 Å). Finally, when the set of start and end points has been identified for each cavity, the Dijkstra's shortest path algorithm is used to find the channels between all pairs of start and end points. The edge weight function used in the algorithm takes into account the distance to the surface of the closest vdW sphere and the edge length. The channel centerline is represented as a 3D natural spline. Depending on the density of computed exits, the algorithm may find duplicate channels. Therefore, in the final post-processing step, if two channels are nearly identical, the longer one is removed. A detailed explanation of the algorithm (also as a scheme) and parameters of the calculation can be found on the MOLE*online* webpage (e.g. http://mole.upol.cz/documentation/). MOLE 2.0 outperforms the original MOLE (28) algorithm in many aspects. For instance, it is quicker due to the division of the internal space within the macromolecule to separate subcavities. There is no need to determine the number of channels prior calculation. MOLE 2.0 enables automatic selection of the starting points and calculation of some basic physicochemical properties of the channel-lining residues.

Results visualization and analysis

Profiles of the channels found are presented in three ways: (i) plots of channel radii against length (visualized using gnuplot—http://www.gnuplot.info—as PNG images); (ii) an interactive table summarizing the set of lining residues and physicochemical properties; and (iii) the channel isosurface along its centerline, which is visualized in the Jmol plugin (36).

MOLE*online* 2.0 also allows calculation of basic physicochemical properties along the unique channel-lining amino acids side chains (these properties are not calculated for nucleobases). Charge, hydropathy and hydrophobicity indices (39,40), polarity (41) and mutability (42) can be estimated.

Charge is calculated as the sum of charges on the side chains (at pH ~7) lining the channel.

Hydropathy (39) is calculated as an average of the hydropathy index of lining side chains, where the most hydrophilic is Arg (−4.5) and the most hydrophobic is Ile (+4.5).

Hydrophobicity (40) is calculated as an average of normalized hydrophobicity scales, where the most hydrophilic residue is Glu (−1.14) and the most hydrophobic residue is Ile (1.81).

Polarity (41) is calculated as an average of amino acid polarity. Polarity values range from zero for non-polar amino acids (Ala and Gly), through values of around 1.5 for polar residues (e.g. Ser 1.67), and finally, to two digits values for charged residues (Glu 49.90, Arg 52.00).

Mutability (42) is calculated as an average of relative mutability index. Relative mutability is high for mutatable amino acids, e.g. small polar amino acids (Ser 117, Thr 107, Asn 104) or small aliphatic amino acids (Ala 100, Val 98, Ile 103). On the other hand, the mutability is low for amino acids that play important structural roles, such as aromatic amino acids (Trp 25, Phe 51, Tyr 50) or special amino acids (Cys 44, Pro 58, Gly 50).

Such an approach gives only an approximate value of mutability, whereas sequence specific analyses can be performed using the multiple sequence alignment tools in other programs, e.g. ConSurf (43) and Hotspot Wizard (44). It is worth noting that the estimated physicochemical properties should be interpreted with care, as the calculation is based on an assumption that the side chains of the lining amino acids significantly determine the environment within the identified channel. The calculation might be sensitive to exact position of the starting and ending points.

Users of MOLE*online* can download all results as a report. Channel centerline positions with radii of maximally inscribed balls values can also be downloaded in two formats for further analysis and storage: (i) as a generic PDB file or (ii) as a python script for visualization in PyMol (http://www.pymol.org).

RESULTS AND DISCUSSION

Examples of usage

Microsomal Cytochrome P450 (CYP) enzymes are important for the metabolism of many endogenous compounds and xenobiotics (45,46). CYPs share a buried active site (47), which is connected to the outside environment by various access/egress channels. (15) These channels are responsible for substrate passage to and product release from the active site, and they are considered to be involved in substrate preferences of CYP, which has been shown to vary considerably among CYP enzymes (12,13). Figure 2 shows all the channels connecting the active site of a CYP enzyme [calculation started from Glu 308 and Thr 309 according to the CSA (37)] of CYP3A4 (PDB: 1TQN) with the exterior. The top ranked channel found by MOLE*online* (white in Figure 2) is the solvent channel (15). The solvent channel is 17 Å long and its bottle-neck is 1.41 Å wide. The solvent channel is also rather hydrophilic as its hydropathy equals −1.9. By comparison, the hydropathy values of channels 2e, 2a and 2f are −0.2, −0.3 and 0.4, respectively, which suggests that these channels are less hydrophilic. The same trend can be

seen in the hydrophobicity index, which again suggests that the solvent channel is also more hydrophilic (−0.68) than channels 2e, 2a and 2f, with values of 0.1, 0.08 and 0.1, respectively. These findings are consistent with previous data, which have identified the solvent channel as the main channel responsible for active site solvation (48) and hydrophilic product release (13,49), while channels of 2x family are considered to be involved in hydrophobic substrate binding (13,50).

The ribosomal exit tunnel (RET) allows nascent peptide chains synthesized at a peptidyl transferase center to exit the ribosome (9). Analysis of ribosomal channels represents a challenge for software tools like MOLE*online*, due to the considerable size and complexity of ribosomes [approximately 100 000 heavy atoms (28)]. Figure 3 shows the RET of a large ribosomal unit from *Haloarcula marismortui* (PDB: 1JJ2 containing 90 650 atoms). In order to achieve optimal results, the channel search parameters had to be adjusted. Since the RET is large enough for passage of nascent peptide with a channel bottleneck radius of ∼3 Å, the probe radius has to be greater (6 Å) to capture the channel. The interior threshold also has to be increased to avoid additional small channels in the structure (2.4 Å). In addition, the surface cover radius should be enlarged to avoid redundant channels appearing (20 Å). Two residues of the peptidyl transferase center were chosen as the start of the RET (Chain 0: U 2620, A 2486). The calculation takes ∼35 s on the server (CPU Intel i5 760 2.8 GHz, 4 GB RAM), while the total time, including transfer of data onto client web pages, takes ∼1–2 min. The length of the ribosomal exit channel is ∼100 Å with three bottlenecks of minimum radii ∼4.5 Å (Figure 3). The RET is highly hydrophilic, polar and mostly lined by negatively charged residues (11 nt from 23 S rRNA have their negatively charged main chains oriented toward the channel and two Glu residues have side chains facing the channel); the negative charge is to some extent compensated by six Arg residues. The distributed charge of the ribosomal polypeptide exit channel is important to prevent the nascent peptides from becoming 'stuck' inside the ribosome.

Limitations

The presented application has four main limitations. The first limitation stems from the initial concept that the channels are extrapolated as sets of maximally inscribed

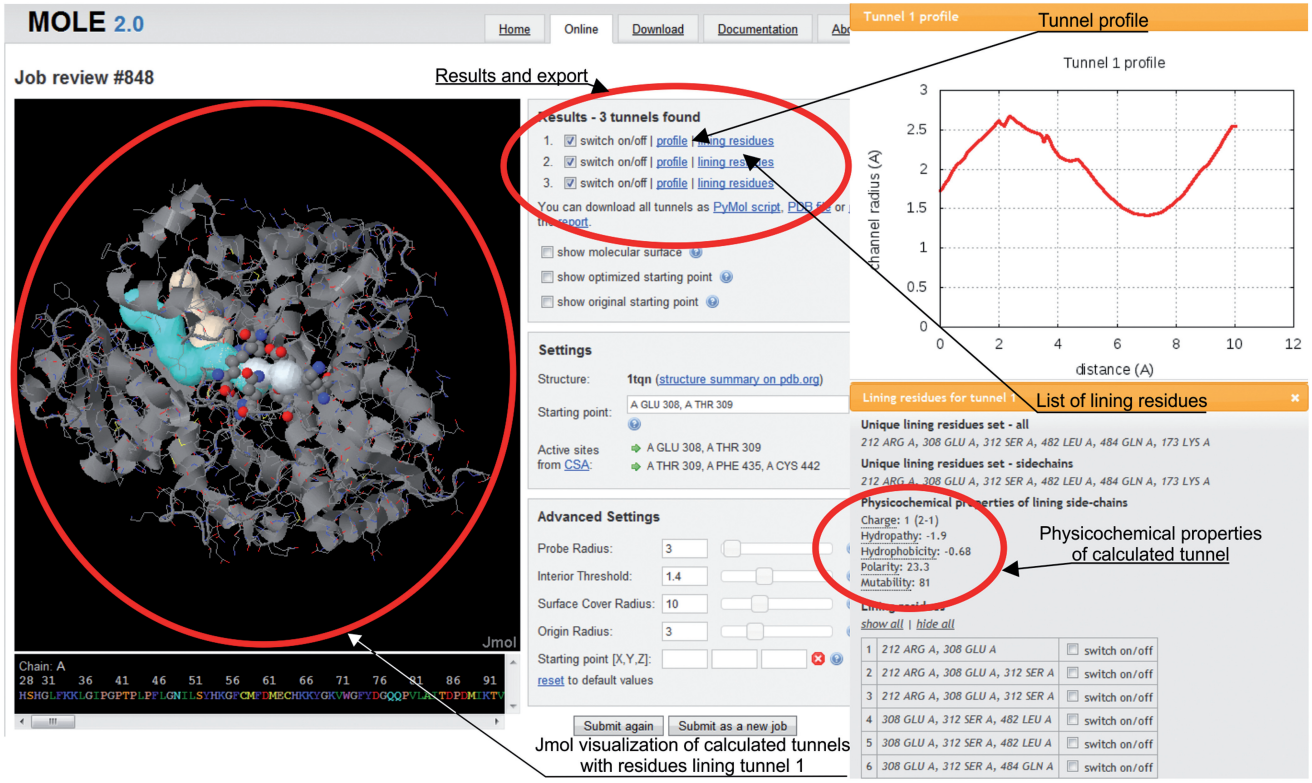


Figure 2. Results of channel analysis of Cytochrome P450 3A4 (CYP3A4) using the setup shown in Figure 1. Four channels found from user-specified starting point are shown, whereas the automatic detection also found additional 17 tunnels which are not shown for clarity. The profile of the tunnel #1 along the centerline and list of lining residues are shown in the external windows (right-hand side). A list of all the unique lining residues and the corresponding side chains alone is displayed along with physicochemical properties of the respective channel. Lining residues can also be visualized along the channel centerline, with the channel represented by maximally inscribed spheres in the Jmol window. It is also possible to show molecular surface and all detected cavities and their volumes. In addition, starting points can be shown as small cubes for original user-defined starting point (in magenta), for optimized position of such starting point (in green) and for all automatically detected starting points (in yellow). Information about tunnel profiles and lining residues can be further exported in form of report, PDB file or python file for visualization in Pymol.

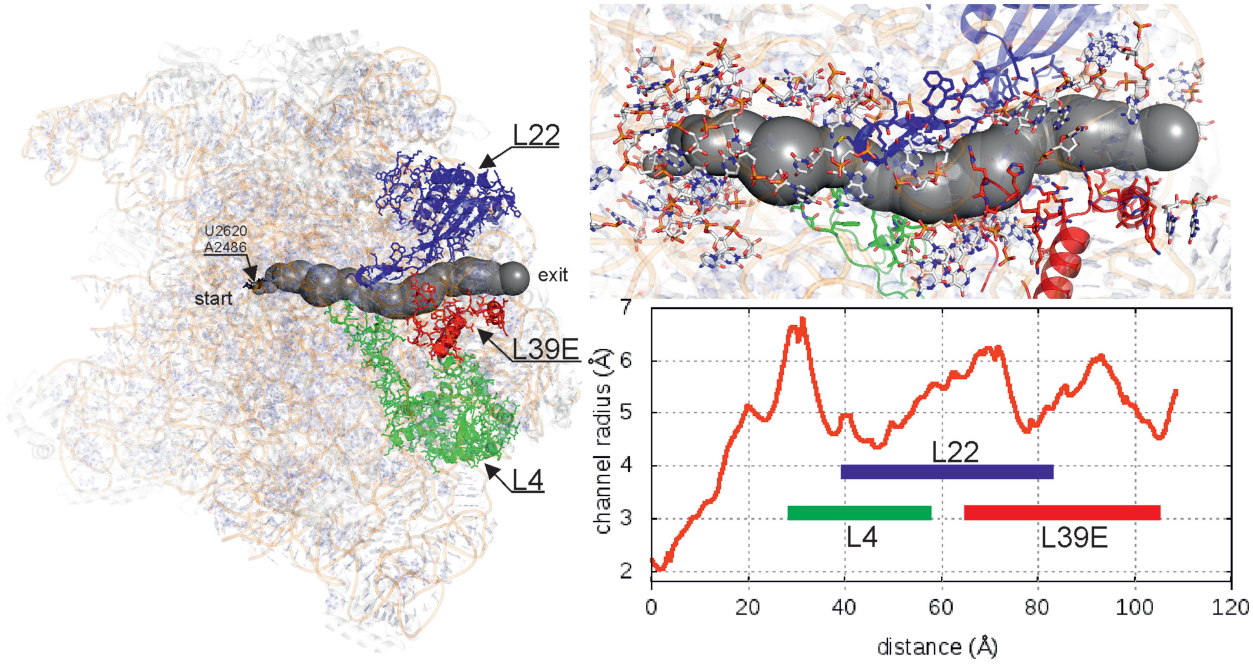


Figure 3. Visualization of ribosomal exit tunnel (RET) of a large ribosomal unit from *Haloarcula marismortui* (PDB: 1JJ2). The figure was prepared in Pymol using an exported python file containing positions of all the channels identified by MOLE*online*. Only the RET is shown and ribosomal proteins L4 (green), L22 (blue), L39E (red) lining the tunnel are highlighted. The channel profile shows the positions of three bottlenecks.

balls along the channel centerline. Such an extrapolation does not allow complex channels with bulges to be mapped accurately. The second limitation arises because the channel-finding algorithm is applied to an atom-centered Voronoi mesh. In principle, the additively weighted Voronoi graph or power diagram offers some benefits in terms of precision, but the gain in precision is small compare to the uncertainties associated with the chosen structures (e.g. X-ray structures with finite resolution, which is generally higher than 0.8 Å), treatment of hydrogen atoms and atomic radii set. The analysis of transmembrane pores is also limited (or not so convenient) because the transmembrane pores have to merged from pore segments identified as tunnels by MOLE*online* 2.0. The final limitation relates to the software and data handling on the server, which limits the maximal size of the studied system to around 100 000 atoms (8 MB).

CONCLUSIONS

In this article, we described MOLE*online* 2.0 (http://ncbr.muni.cz/mole or http://mole.upol.cz), a new web-based interactive tool for the analysis of molecular channels and pores. The MOLE*online* interface enables platform-independent, easy-to-use and interactive analyses and offers the prospect of high automation, e.g. by downloading structures from the PDB database and employing automatic active site identification based on the CSA. The results of the channel search using MOLE*online* are presented in a clear visual or data form, making their interpretation and further manipulation easy.

FUNDING

Czech Science foundation [GD301/09/H004, 303/09/1001, P303/12/P019, P208/12/G016]; Ministry of Education of the Czech Republic (ME 08008); Palacky University [Student Project PrF_2012_028]; the European Community's Seventh Framework Program from the Operational Program Research and Development for Innovations—European Regional Development Fund [CZ.1.05/2.1.00/03.0058, CZ.1.05/1.1.00/02.0068], 'Capacities' specific program [Contract No. 286154]; and European Social Fund [CZ.1.07/2.3.00/20.0017]. D.S. and C.M.I. also acknowledges financial funding through a PhD. Talent program by Brno City Municipality. Funding for the open access charge: Czech Science foundation [P208/12/G016].

Conflict of interest statement. None declared.

REFERENCES

1. Walz,T., Smith,B.L., Agre,P. and Engel,A. (1994) The 3-dimensional structure of human erythrocyte aquaporin chip. *EMBO J.*, **13**, 2985–2993.
2. Engel,A., Fijiyoshi,Y. and Agre,P. (2000) The importance of aquaporin water channel protein structures. *EMBO J.*, **19**, 800–806.
3. Jiang,Y.X., Lee,A., Chen,J.Y., Cadene,M., Chait,B.T. and MacKinnon,R. (2002) Crystal structure and mechanism of a calcium-gated potassium channel. *Nature*, **417**, 515–522.
4. Doyle,D.A., Cabral,J.M., Pfuetzner,R.A., Kuo,A.L., Gulbis,J.M., Cohen,S.L., Chait,B.T. and MacKinnon,R. (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science*, **280**, 69–77.
5. Gouaux,E. and MacKinnon,R. (2005) Principles of selective ion transport in channels and pumps. *Science*, **310**, 1461–1465.

6. MacKinnon,R. (2003) Potassium channels. *FEBS Lett.*, **555**, 62–65.
7. Murray,J.W. and Barber,J. (2007) Structural characteristics of channels and pathways in photosystem II including the identification of an oxygen channel. *J. Struct. Biol.*, **159**, 228–237.
8. Guskov,A., Kern,J., Gabdulkhakov,A., Broser,M., Zouni,A. and Saenger,W. (2009) Cyanobacterial photosystem II at 2.9-A resolution and the role of quinones, lipids, channels and chloride. *Nat. Struct. Mol. Biol.*, **16**, 334–342.
9. Voss,N.R., Gerstein,M., Steitz,T.A. and Moore,P.B. (2006) The geometry of the ribosomal polypeptide exit tunnel. *J. Mol. Biol.*, **360**, 893–906.
10. Wade,R.C., Winn,P.J., Schlichting,E. and Sudarko. (2004) A survey of active site access channels in cytochromes P450. *J. Inorg. Biochem.*, **98**, 1175–1182.
11. Otyepka,M., Skopalik,J., Anzenbacherova,E. and Anzenbacher,P. (2007) What common structural features and variations of mammalian P450s are known to date? *Biochim. Biophys. Acta*, **1770**, 376–389.
12. Otyepka,M., Berka,K. and Anzenbacher,P. (2012) Is there a relationship between the substrate preferences and structural flexibility of cytochromes P450? *Curr. Drug Metab.*, **13**, 130–142.
13. Berka,K., Hendrychova,T., Anzenbacher,P. and Otyepka,M. (2011) Membrane position of ibuprofen agrees with suggested access path entrance to cytochrome P450 2C9 active site. *J. Phys. Chem. A*, **115**, 11248–11255.
14. Hendrychova,T., Berka,K., Navratilova,V., Anzenbacher,P. and Otyepka,M. (2012) Dynamics and hydration of the active sites of mammalian cytochromes P450 probed by molecular dynamics simulations. *Curr. Drug Metab.*, **13**, 177–189.
15. Cojocaru,V., Winn,P.J. and Wade,R.C. (2007) The ins and outs of cytochrome P450s. *Biochim. Biophys. Acta*, **1770**, 390–401.
16. Petrek,M., Otyepka,M., Banas,P., Kosinova,P., Koca,J. and Damborsky,J. (2006) CAVER: a new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinformatics*, **7**, 316.
17. Pavlova,M., Klvana,M., Prokop,Z., Chaloupkova,R., Banas,P., Otyepka,M., Wade,R.C., Tsuda,M., Nagata,Y. and Damborsky,J. (2009) Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat. Chem. Biol.*, **5**, 727–733.
18. Liang,J., Edelsbrunner,H. and Woodward,C. (1998) Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci.*, **7**, 1884–1897.
19. Damborsky,J., Petrek,M., Banas,P. and Otyepka,M. (2007) Identification of tunnels in proteins, nucleic acids, inorganic materials and molecular ensembles. *Biotechnol. J.*, **2**, 62–67.
20. Perot,S., Sperandio,O., Miteva,M.A., Camproux,A.C. and Villoutreix,B.O. (2010) Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov. Today*, **15**, 656–667.
21. Coleman,R.G. and Sharp,K.A. (2009) Finding and characterizing tunnels in macromolecules with application to ion channels and pores. *Biophys. J.*, **96**, 632–645.
22. Ho,B.K. and Gruswitz,F. (2008) HOLLOW: Generating accurate representations of channel and interior surfaces in molecular structures. *BMC Struct. Biol.*, **8**, 49.
23. Voss,N.R. and Gerstein,M. (2010) 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res.*, **38**, W555–W562.
24. Raunest,M. and Kandt,C. (2011) dxTuber: Detecting protein cavities, tunnels and clefts based on protein and solvent dynamics. *J. Mol. Graphics Model.*, **29**, 895–905.
25. Laskowski,R.A. (1995) Surfnet - a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graphics*, **13**, 323–330.
26. Smart,O.S., Neduvelil,J.G., Wang,X., Wallace,B.A. and Sansom,M.S.P. (1996) HOLE: A program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graphics Model.*, **14**, 354–360.
27. Pellegrini-Calace,M., Maiwald,T. and Thornton,J.M. (2009) PoreWalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS Comp. Biol.*, **5**, e1000440.
28. Petrek,M., Kosinova,P., Koca,J. and Otyepka,M. (2007) MOLE: a Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure*, **15**, 1357–1363.
29. Yaffe,E., Fishelovitch,D., Wolfson,H.J., Halperin,D. and Nussinov,R. (2008) MolAxis: efficient and accurate identification of channels in macromolecules. *Proteins*, **73**, 72–86.
30. Yaffe,E., Fishelovitch,D., Wolfson,H.J., Halperin,D. and Nussinov,R. (2008) MolAxis: a server for identification of channels in macromolecules. *Nucleic Acids Res.*, **36**, W210–W215.
31. Lee,P.H. and Helms,V. (2012) Identifying continuous pores in protein structures with PROPORES by computational repositioning of gating residues. *Proteins*, **80**, 421–432.
32. Dijkstra,E.W. (1959) A note on two problems in connexion with graphs. *Numerische Mathematik*, **1**, 269–271.
33. Richards,F.M. (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. *J. Mol. Biol.*, **82**, 1–14.
34. Poupon,A. (2004) Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Curr. Opin. Struct. Biol.*, **14**, 233–241.
35. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
36. Herraez,A. (2006) Biomolecules in the computer: Jmol to the rescue. *Biochem. Mol. Biol. Educ.*, **34**, 255–261.
37. Porter,C.T., Bartlett,G.J. and Thornton,J.M. (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Res.*, **32**, D129–D133.
38. Cornell,W.D., Cieplak,P., Bayly,C.I., Gould,I.R., Merz,K.M., Ferguson,D.M., Spellmeyer,D.C., Fox,T., Caldwell,J.W. and Kollman,P.A. (1995) A 2nd generation force-field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.*, **117**, 5179–5197.
39. Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
40. Cid,H., Bunster,M., Canales,M. and Gazitua,F. (1992) Hydrophobicity and structural classes in proteins. *Protein Eng.*, **5**, 373–375.
41. Zimmerman,J.M., Eliezer,N. and Simha,R. (1968) Characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.*, **21**, 170–201.
42. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.
43. Ashkenazy,H., Erez,E., Martz,E., Pupko,T. and Ben-Tal,N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–W533.
44. Pavelka,A., Chovancova,E. and Damborsky,J. (2009) HotSpot wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Res.*, **37**, W376–W383.
45. Anzenbacher,P. and Anzenbacherova,E. (2001) Cytochromes P450 and metabolism of xenobiotics. *Cell. Mol. Life Sci.*, **58**, 737–747.
46. Guengerich,F.P. (2005) Human cytochrome P450 enzymes. In: Ortiz de Montellano,P.R. (ed.), *Cytochrome P450: Structure, Mechanism, and Biochemistry* 3rd edn. Kluwer Academic/Plenum Publishers, New York, pp. 377–530.
47. Anzenbacher,P., Anzenbacherová,E., Lange,R., Skopalík,J. and Otyepka,M. (2008) Active sites of cytochromes P450: what are they like? *Acta Chim. Slov.*, **55**, 63–66.
48. Skopalik,J., Anzenbacher,P. and Otyepka,M. (2008) Flexibility of human cytochromes P450: molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences. *J. Phys. Chem. B*, **112**, 8165–8173.
49. Fishelovitch,D., Shaik,S., Wolfson,H.J. and Nussinov,R. (2009) Theoretical characterization of substrate access/exit channels in the human cytochrome P450 3A4 enzyme: involvement of phenylalanine residues in the gating mechanism. *J. Phys. Chem. B*, **113**, 13018–13025.
50. Conner,K.P., Woods,C.M. and Atkins,W.M. (2011) Interactions of cytochrome P450s with their ligands. *Arch. Biochem. Biophys.*, **507**, 56–65.

SiteBinder: an improved approach for comparing multiple protein structural motifs

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

SiteBinder: An Improved Approach for Comparing Multiple Protein Structural Motifs

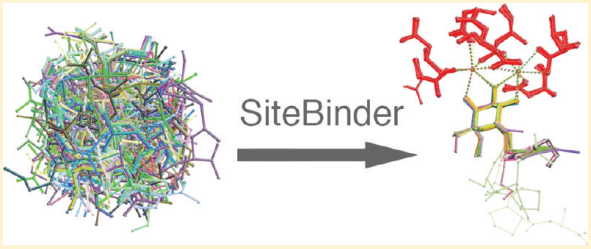
David Sehnal,[†] Radka Svobodová Vařeková,[†,*] Heinrich J. Huber,[‡] Stanislav Geidl,[†] Crina-Maria Ionescu,[†] Michaela Wimmerová,[†] and Jaroslav Koča[†,*]

[†]National Centre for Biomolecular Research, Faculty of Science and CEITEC - Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00 Brno-Bohunice, Czech Republic

[‡]Centre of Systems Medicine, Department of Physiology and Medical Physics, Royal College of Surgeons in Ireland, 123 St Stephens Green, Dublin 2, Ireland

Supporting Information

ABSTRACT: There is a paramount need to develop new techniques and tools that will extract as much information as possible from the ever growing repository of protein 3D structures. We report here on the development of a software tool for the multiple superimposition of large sets of protein structural motifs. Our superimposition methodology performs a systematic search for the atom pairing that provides the best fit. During this search, the RMSD values for all chemically relevant pairings are calculated by quaternion algebra. The number of evaluated pairings is markedly decreased by using PDB annotations for atoms. This approach guarantees that the best fit will be found and can be applied even when sequence similarity is low or does not exist at all. We have implemented this methodology in the Web application SiteBinder, which is able to process up to thousands of protein structural motifs in a very short time, and which provides an intuitive and user-friendly interface. Our benchmarking analysis has shown the robustness, efficiency, and versatility of our methodology and its implementation by the successful superimposition of 1000 experimentally determined structures for each of 32 eukaryotic linear motifs. We also demonstrate the applicability of SiteBinder using three case studies. We first compared the structures of 61 PA-IIL sugar binding sites containing nine different sugars, and we found that the sugar binding sites of PA-IIL and its mutants have a conserved structure despite their binding different sugars. We then superimposed over 300 zinc finger central motifs and revealed that the molecular structure in the vicinity of the Zn atom is highly conserved. Finally, we superimposed 12 BH3 domains from pro-apoptotic proteins. Our findings come to support the hypothesis that there is a structural basis for the functional segregation of BH3-only proteins into activators and enablers.



INTRODUCTION

Nowadays, a large amount of information about the 3D structure of proteins is available, and more and more structures are being solved every year because of advances in experimental techniques and their increased availability. This amount of data provides the opportunity to compare large sets of protein structural motifs like binding sites, secondary structure elements, cavities, and tunnels. Such analyses can help identify the main characteristics of important protein motifs. The obtained characteristics can subsequently be used as patterns in drug discovery,[1,2] to understand the relationship between a protein's structure and its function and even predict its function,[3−5] to classify proteins,[6,7] to identify evolutionary relationships between proteins,[8−10] etc. Collecting large sets of protein structural motifs is a fairly simple task. This task can be accomplished by employing available software tools or in-house scripts that retrieve data from structural databases on the basis of primary or secondary protein structure queries. The more sophisticated challenge is to perform the comparison of these large sets of protein structural motifs, as this requires specifically adapted algorithms and software tools. Such a comparison is a particular topic because, on the one hand, these motifs are small compounds, but on the other hand, the motifs are parts of proteins. To our knowledge, no software tool available to date can process hundreds of protein structural motifs at one time and allow for a straightforward comparison within these large sets of structures. Therefore, our goal was to develop and implement a new methodology for comparing large sets of protein structural motifs in an efficient, flexible, and intuitive manner.

The comparison of 3D structures is a complex topic that can be divided into several subtopics. We distinguish between methods that compare compounds with identical (or very similar) 2D structure, as opposed to methods dealing with compounds for which the 2D structure differs significantly. The term "2D structure", as it is introduced in chemoinformatics,[11] refers to the topology of the molecule, meaning the nature and connectivity of the atoms contained in the molecule. We also

Special Issue: 2011 Noordwijkerhout Cheminformatics

Received: September 19, 2011
Published: February 1, 2012

differentiate between the methods on the basis of the type of molecules they process—organic molecules, proteins, or protein motifs.

Organic molecules with different 2D structures can be compared by two principal methods, namely, the rigid body approach and the flexible body approach.[12,13] Rigid body methods[12,14] keep the structure of both molecules fixed and try to find an alignment by maximizing some kind of volume overlap (i.e., van der Waals overlap, electron density overlap, electrostatic potential overlap, etc.). The overlap optimization methods range from simplex optimization, gradient optimization, and Fourier space methods to Monte Carlo optimization. Flexible (or semiflexible) body methods[13,15] change the structure of one molecule during the comparison, thus simulating the process of how the molecule adapts its shape when undergoing a chemical reaction.

The comparison of proteins with different 2D structures can be classified as global or local.[16] The algorithms and available software for both of these approaches were reviewed by Gherhardini et al.[17] Global comparison approaches use various algorithms, such as dynamic programming,[18] double dynamic programming,[19] branch and bound approach,[20,21] subgraph isomorphism,[22,23] or extension of seed matches.[24] Global comparison is used to classify protein structures and to identify evolutionary links between distant homologues. Nevertheless, the function of a protein usually depends more on the identity and location of a few residues comprising the active site than on the overall structure. In order to directly analyze and compare the residues involved in protein function, local (as opposed to global) structural comparison methods have been developed. These methods focus on detecting a similar 3D arrangement of a small set of residues, possibly in the context of completely different protein structures. Local structure comparison approaches are mainly based on algorithms that employ geometric hashing,[25,26] subgraph isomorphism,[27,28] recursive search connected with the branch and bound algorithm,[29,30] and graph-based heuristics.[31] To identify local similarities within two entire protein structures such algorithms can be applied without any a priori assumption or by using a predefined structural template to screen a structure. The structural templates can be user defined.[32] A special case of local structure comparison is searching for a structural motif in a protein by comparing the motif with a relevant part of the protein. These approaches are reviewed in a recent paper.[33]

The development of comparison methods for protein structural motifs with different 2D structures has become an important topic of research within the past few years.[34,35] These comparisons are, among others, necessary for the functional annotation of proteins.[36,37] General purpose software tools able to compare all types of molecules with different 2D structure (i.e., organic molecules, proteins, and protein motifs) are also available (e.g., Bauer et al.[38]).

Superimposition or superposition[16] is the comparison of molecules with identical (or very similar) 2D structures. Superimposition can be applied to study different conformers of one molecule, and these conformers can be obtained from experiment, from molecular dynamics simulations, or from different databases of 3D structures. Likewise, superimposition is often useful to study substructures that were obtained by the analysis and comparison of the 2D structure of molecules or the primary structure of proteins. Superimposition approaches are similar for organic molecules, proteins, and protein motifs.

In brief, superimposition consists of several interdependent stages.[39] First, it is necessary to find the correspondence between the atoms coming from different structures. We will refer to this first step, as well as to its results, as atom pairing or simply pairing. Using an atom pairing is necessary so that the structures can be processed as sequences of points in the 3D space. In the second step, the sets of paired 3D points are fitted together as tightly as possible by a geometrical transformation. We will refer to this step as optimal fitting because its final result gives the coordinates of the superimposed structures. The last phase of the superimposition is to evaluate the quality of the fit. This is done by computing the root-mean-square deviation (RMSD) between the sets of 3D coordinates belonging to the structures that have been superimposed. We further discuss the currently available methodology for performing the steps of atom pairing and optimal fitting.

From the mathematical point of view, pairings are bijections, which are functions where every element from the first set is assigned to exactly one element from the second set. For structures with n atoms, $n!$ such bijections can be constructed, and therefore, $n!$ pairings may exist. It is desirable to find the best pairing, meaning the pairing that will eventually lead to the lowest RMSD between the superimposed structures. Finding the best pairing requires testing all constructed pairings and is therefore very time demanding. Nevertheless, an incorrect atom pairing can lead to a poor superimposition. There are several heuristics and algorithms to solve this problem, such as implicit pairing,[40,41] employing sequence alignment,[42,43] systematic approach,[44] or subgraph matching.[45,46] We briefly describe these below.

Implicit pairing associates atoms with the same index or position (i.e., pairing the i-th atom of the first molecule to the i-th atom of the second molecule). Pairing atoms by this algorithm is extremely fast. An additional advantage is that the subsequent fitting will only be performed once because only one pairing is produced. However, implicit pairing is suitable only when the atoms in both molecules are indexed or ordered identically, as in the case of conformers resulted from molecular dynamics simulations. Many state of the art programs that offer the superimposition of organic molecules (e.g., Chimera,[41] VMD,[47] Gromacs,[40] gOpenMol,[48] Pymol[49]) use implicit pairing.

Employing sequence alignment provides an improvement on the implicit pairing approach. First, the sequence alignment is performed by a selected algorithm (e.g., Needleman and Wunsch alignment,[50] ICM ZEGA alignment,[51] etc.). Afterward, the atoms from the aligned residues are paired using an implicit pairing. This approach is applicable only for the superimposition of proteins or protein sequences with a reasonable degree of sequence similarity. Several drug design packages (e.g., MOE,[42] Discovery Studio,[52] ICM,[43] etc.) implement this approach.

The systematic approach finds all possible pairings and is therefore very robust. However, because the fitting will have to be performed for a large number of pairings, this method is time consuming and therefore useful mainly for small molecules. It can be sped up by backtracking,[53] a procedure that is able to discard possible solutions as soon as they appear unfeasible. Further decrease in computational complexity can be achieved by pairing only atoms that have corresponding chemical element symbols and/or come from comparable chemical neighborhoods.

Subgraph matching, which was originally developed for processing molecules with different 2D structure, can also be used for finding a relevant pairing (reviewed by Raymond et al.[46]).

This approach identifies the largest possible atom sets that can be superimposed.

When an atom pairing has been found, the sequences of paired 3D points can be fitted by performing a geometrical transformation (composition of a translation and a rotation in the 3D space). Finding the transformation that will lead to the optimal fit is a fairly cumbersome task. An iterative solution to this problem was published by McLachlan et al.,[54] while a closed form solution that utilizes rotation matrices was published by Kabsch et al.[55] This rotation matrix approach was later reformulated using quaternion algebra. Many authors over the past 20 years have "rediscovered" the application of quaternions in the superimposition of 3D points (i.e., Horn,[56] Diamond,[57] and Kearsley[58]). However, within the community of computational chemists and biologists, quaternions were introduced by Coutsias et al.[39] and are still a topic of research.[59] All the closed form solutions have linear space and time complexity in the number of atoms. These solutions work by translating both structures to their common origin and then using singular value decomposition in the case of rotation matrices or eigenvectors in the case of quaternions.

Superimposition can be performed for two or more structures at once, depending on the nature of the investigation. Superimposing two molecules or motifs is a very useful task if the purpose is the in-depth structural comparison and characterization of the two compounds under investigation. Many software tools offering the superimposition of two compounds are available,[40,41,47−49] all using implicit pairing. Nevertheless, one often needs to compare the structures of tens or hundreds of compounds at a time in order to find structural trends or peculiarities. In this case, it is necessary to perform a multiple superimposition, which is a fairly more complex procedure than the superimposition of only two structures at a time. The quality of a multiple superimposition procedure can be measured using the generalized RMSD,[60] which is the average RMSD between all pairs of structures. Another possibility is to compute the RMSD between each structure and the calculated average structure and then average these RMSD values over all structures.[61] A naive approach to this problem is to pick one of the structures and superimpose all structures to this chosen one. A quadratic complexity algorithm to this problem was published by Konagurthu et al.[60] and is used, for example, in Pymol.[49] A more advanced approach is to superimpose all pairs of structures, order the pairs by the quality of the superimposition, and then superimpose the structures to an iteratively computed average.[6] An improved approach to this problem, with nearly linear complexity (in the number of structures), was published by Eidhammer et al.[16] and later generalized by Wang et al.[61] This method is based on iteratively superimposing each structure onto the average model of the structures superimposed in the previous step until a stable configuration is reached.

In this work, we focus on the comparison of large sets of protein structural motifs. Such large sets are generally collected in an automated fashion by querying the primary or secondary structure of proteins and will thus consist mainly of motifs with similar 2D structure. The possibility to perform the multiple superimposition of a large number of protein motifs with similar 2D structure would open the door to innovative thinking. One could find meaningful structural trends or peculiarities that could identify evolutionarily related proteins or could explain and even predict function and activity related features of known or engineered proteins.

To our knowledge, no implementation of such a methodology is available to date, even though many state of the art software packages offer the possibility to superimpose protein structures to various extents. Thus, our goal was to fill in this gap and to develop and implement a methodology for superimposing large sets of protein structural motifs in an efficient, flexible, and intuitive manner, so as to fuel inquisitiveness and creativity in the investigation of protein structure and function. A challenging aspect of protein structural motif superimposition is that the motifs need not refer only to linear protein subsequences but may also consist of the 3D surroundings of residues or sequences, binding sites of metals or sugars, or any other selected parts of protein 3D structure. This means that some of the superimposed motifs may not have any sequence similarity. Our methodology guarantees the best superimposition even in such cases.

METHODS

When performing the superimposition of two protein structural motifs, one faces two challenges. One challenge is to find the best pairing of chemically corresponding atoms from the first and second motif. This pairing establishes which atoms from the first motif should be fitted to which atoms from the second motif in the optimal fitting phase. The other challenge is to calculate the geometrical transformation that optimally fits the structures of the two protein motifs together.

In our methodology, we address the first issue by a systematic approach employing heuristics tailored to proteins (described in detail below) and the second by using a state of the art quaternion algebra approach.[39] A detailed description of how we employ this approach is provided in the Supporting Information. The main mathematical object employed in our methodology is a molecular graph,[62,63] which was adapted for protein structural motifs. The formalized mathematical description of our methodology is available in the Supporting Information.

Pairing. Using the most appropriate atom pairing is a prerequisite for a successful superimposition, and failure to identify the best pairing leads to poor results, as is shown in Figure 1. For superimposing protein structural motifs, we cannot use implicit pairing (i.e., the i-th atom from one motif with the i-th atom from the other motif) because the order of the atoms or amino acid residues in the PDB file of one motif might differ from the order in the PDB file of the other motif. Figure 1 demonstrates that even for the superimposition of two PHE residues there can be a significant difference between the superimposition calculated using implicit pairing and the superimposition calculated using the best possible pairing. Employing sequence alignment is also not applicable because some of the superimposed motifs may not have any sequence similarity. Subgraph matching (i.e., searching for the largest identical subgraph contained in both motifs) is also not suitable because protein motifs can consist of several identical residues and can be very symmetrical, and thus, many relevant subgraphs can be found. We therefore decided to use a systematic approach, which tests all possible pairings.

The disadvantage of the systematic approach is its complexity. When superimposing two motifs with n atoms, there are $n!$ possible pairings (e.g., about 3×10^{40} pairings for 30 atoms). It is thus desirable to reduce the number of tested pairings as much as possible. An initial decrease in the number of pairings can be achieved by looking only at those pairings that are chemically meaningful, such that two atoms will be paired only if

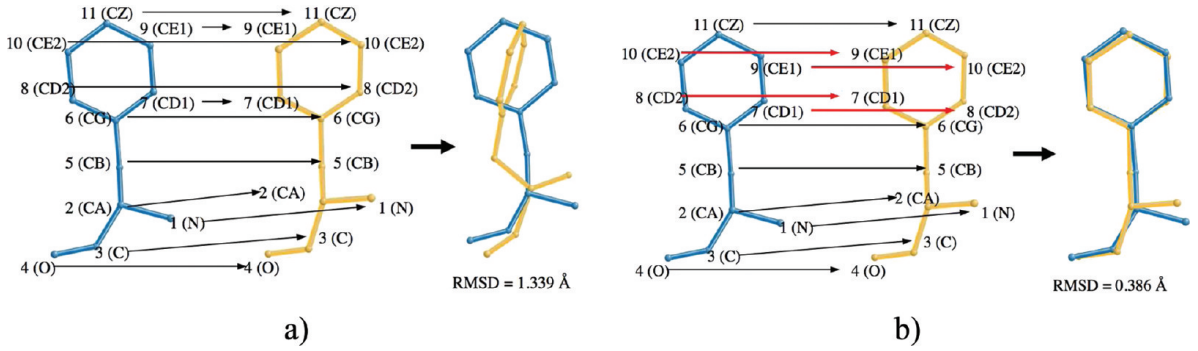


Figure 1. (a) Implicit pairing between residues PHE 83 (blue) and PHE 91 (orange) from the PDB entry 2wh6 and the superimposition calculated by the program VMD, which uses this pairing. (b) The best possible pairing between PHE 83 (blue) and PHE 81 (orange) from 2wh6 and the superimposition calculated by our program SiteBinder, which is able to find this pairing. The differences compared to the implicit pairing are depicted by red arrows. In both (a) and (b), atoms are denoted by their number in the residue, while their PDB name is in brackets.

they are of the same chemical element. A further decrease in the number of tested pairings can be achieved by using the information available in the PDB files. In the PDB file format, each atom is assigned to a residue. Each residue is given a name and a residue identifier (number), which specifies the residue's location in the amino acid sequence. All this information is useful in deciding which atom pairings are worth testing. One can use residue identifiers to make sure that atoms belonging to a single residue in the first motif will be paired only to atoms belonging to a single residue in the second motif and not to atoms belonging to separate residues. Finally, a very effective reduction in the number of possible atom pairings can be achieved if one considers residue names, as this ensures that only atoms belonging to residues with the same names will be paired.

Grouping. We described the basic ideas how to reduce the number of tested atom pairings. To implement these ideas, we need to group the atoms in both motifs into sets and subsets according to the above-mentioned properties. The set containing the atoms from a motif divided into these sets and subsets is denoted as grouping. The groupings help to markedly reduce the number of tested pairings because only the pairings which respect the groupings will be considered. This means that if some atoms from the first grouping are together in a set or subset they can be paired only with atoms from the second grouping that are also together in a relevant set or subset. Conversely, if the respective atoms are not in the same set or subset, they cannot be paired with atoms that are together in a set or subset.

We denote two groupings as compatible if there is at least one pairing (i.e., bijection) that can be created between their atoms. Only compatible groupings can be used in the process of superimposition. We introduce three different types of groupings—residue name, residue identifier, and element symbol grouping.

Residue name grouping assigns atoms to sets according to the name and identifier of the residue they come from. These sets of atoms are further divided into subsets according to their chemical element symbols. For the protein motif in Figure 2, the residue name grouping is $\{\{\{1,3\}^{N}, \{2,4,5\}^{C}\}^{HIS1}, \{\{6,8\}^{N}, \{7,9,10\}^{C}\}^{HIS2}, \{\{11,12\}^{O}, \{13\}^{C}\}^{ASP3}, \{\{14,15\}^{O}, \{16\}^{C}\}^{GLU4}, \{\{17\}^{Zn}\}^{Zn5}\}$. For clarity, the sets and subsets are denoted by the relevant residue name, residue identifier, and element symbol; a similar denotation will be used in further examples of grouping.

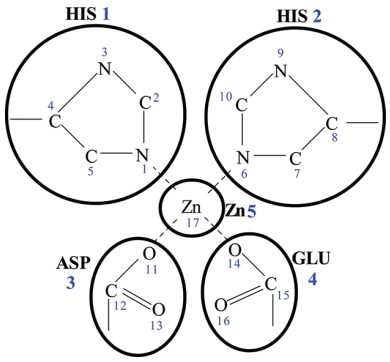


Figure 2. Example of a protein motif.

We use the residue name and identifier jointly for establishing the sets because of two reasons. First, if one uses just the residue names, the atoms from identically named residues will not be separated. For the motif in Figure 2, the grouping would then be $\{\{\{1,3,6,8\}^{N}, \{2,4,5,7,9,10\}^{C}\}^{HIS}, \{\{11,12\}^{O}, \{13\}^{C}\}^{ASP}, \{\{14,15\}^{O}, \{16\}^{C}\}^{GLU}, \{\{17\}^{Zn}\}^{Zn}\}$. Second, if one uses only residue identifiers, the information about the residue name is lost, and it is hard to distinguish for example between the atoms from ASP and GLU in the motif from Figure 2.

Two residue name groupings are compatible if for each set in one grouping there is a set in the other grouping that contains the same number of atoms with the same chemical element symbol that originate from residues with the same name. Thus, using this grouping type is limited (e.g., there are compatible residue name groupings for the dipeptides ALA-GLY and ALA-GLY, but there are no compatible residue name groupings for ALA-GLY and ALA-UNK). On the other hand, the residue name grouping is the most effective grouping type as it reduces the number of tested pairings to a minimum.

Residue identifier grouping assigns atoms to sets according to the identifier of the residue from which they originated. These sets are further divided into subsets according to chemical element symbols. For the protein motif in Figure 2, the residue identifier grouping is $\{\{\{1,3\}^{N}, \{2,4,5\}^{C}\}^{1}, \{\{6,8\}^{N}, \{7,9,10\}^{C}\}^{2}, \{\{11,12\}^{O}, \{13\}^{C}\}^{3}, \{\{14,15\}^{O}, \{16\}^{C}\}^{4}, \{\{17\}^{Zn}\}^{5}\}$. Two residue identifier groupings are compatible if for each set in one grouping there is a set in the other grouping that contains the

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

same number of atoms with the same chemical element symbols. Using this grouping type is also limited (e.g., there can be compatible residue identifier groupings for two dipeptides ALA-GLY and ALA-UNK, but there are no compatible residue identifier groupings for a dipeptide ALA-GLY and a residue UNK). The residue identifier grouping is slightly less effective than the residue name grouping in reducing the number of tested pairings.

Element symbol grouping assigns atoms to sets according to their chemical element symbols. For consistency, we further divide these sets into subsets, but these are also based on chemical element symbols. For the protein motif in Figure 2, the element symbol grouping is $\{\{\{1,3,6,9\}^{N}\}^{N}, \{\{2,4,5,7,8,10,12,14\}^{C}\}^{C}, \{\{11,13,14,16\}^{O}\}^{O}, \{\{17\}^{Zn}\}^{Zn}\}$. Two element symbol groupings are compatible if for each set in one grouping there is a set in the other grouping that contains the same number of atoms that have the same chemical element symbol. This grouping type is very general and can be used in all cases where the superimposed motifs have the same molecular formula. On the other hand, the element symbol grouping has the lowest effectiveness in reducing the number of tested pairings.

Generating Atom Pairings. Before generating all relevant atom pairings that will be tested, it is desirable to find the most effective grouping type that can be used. We first prepare residue name groupings for both motifs and test if these groupings are compatible. If the residue name groupings are compatible, we can employ them. Otherwise, we prepare residue identifier groupings for both motifs and test their compatibility. If the residue identifier groupings are compatible, we can employ them. Otherwise, we prepare element symbol groupings for both motifs. If the element symbol groupings are compatible, we employ these groupings. If no compatible grouping can be found, the motifs cannot be superimposed, and the user needs to change the selection of atoms in at least one of the motifs.

Once we have found compatible groupings for our motifs, we create all possible pairings (i.e., bijections), which respect the groupings (as described above).

Complete Algorithm for Superimposing Two Protein Motifs. To summarize the description given above, we provide a pseudocode of the algorithm for superimposing two motifs.

- Step 1: Prepare the residue name groupings for both motifs. If they are compatible, go to Step 5.
- Step 2: Prepare the residue identifier groupings for both motifs. If they are compatible, go to Step 5.
- Step 3: Prepare the element symbol groupings for both motifs. If they are compatible, go to Step 5.
- Step 4: There is no compatible grouping. Modify the atom selection in at least one motif and go to Step 1.
- Step 5: Use the groupings resulted in the last performed step and generate all possible atom pairings, which respect the groupings.
- Step 6: For each generated pairing do the following: Use quaternion algebra and calculate the transformation that optimally fits one motif to the other. Fit the motifs together using this transformation and calculate the RMSD value.
- Step 7: Find the pairing (among all the generated pairings) that leads to the smallest RMSD.
- Step 8: Superimpose the motifs using the pairing found in Step 7. Return the new coordinates of the motifs (i.e., return the superimposed motifs) and the RMSD value.

Multiple Superimposition of Protein Motifs. Our goal is to provide the most effective solution for this problem that would fit a whole set of protein motifs together as tightly as possible. For this purpose, selecting one of the motifs and superimposing all the others to this one is not a feasible solution as it would only provide an indication of how the rest of the motifs differ from the selected one. Therefore, we designed a multiple superimposition approach that uses the method published by Wang et al.,[61] adapted it to protein motifs and combined it with our algorithm for the superimposition of two motifs. This approach minimizes the RMSD of the whole set of motifs:

$$\mathrm{RMSD}(M) = \sqrt{\binom{m}{2}^{-1} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \mathrm{RMSD}(M_i, M_j)^2} \quad (1)$$

where M is the set of motifs, and m is the number of motifs in this set.

The multiple superimposition approach works in two steps. First, each motif is superimposed to the first one. This simple superimposition of two motifs is done as described in the pseudocode above, and its purpose is to establish an initial pairing of the atoms and calculate an initial RMSD value. We use this atom pairing and calculate an average motif (M_{avg}) as the arithmetic average of the x, y, and z coordinates of the corresponding atoms. Next, all the motifs in the set are superimposed to the average motif. The new coordinates of all these superimposed motifs are stored, together with the new atom pairing. From these new coordinates, we calculate a new RMSD value (denoted RMSD′). We then calculate the normalized difference (δ) between the original and new RMSD

$$\delta = \frac{\mathrm{RMSD} - \mathrm{RMSD}'}{\mathrm{RMSD}} \quad (2)$$

If $\delta \leq \varepsilon$, where ε is a constant set to 0.005, the process is complete, and the new coordinates are returned. If not, we replace the original coordinates by the new ones, the original pairing by the new one, set the value of the RMSD to RMSD′, and repeat the process. For clarity, we provide also the pseudocode of this approach:

- Step 1: Perform the superimposition of each motif to the first one in order to obtain an initial pairing and calculate an initial value for the RMSD.
- Step 2: Calculate the average motif M_{avg} using the pairing.
- Step 3: Superimpose all motifs to M_{avg} and store the new coordinates and new pairing. Calculate RMSD′ and δ.
- Step 4: If $\delta \leq \varepsilon$, go to Step 6.
- Step 5: Replace the original coordinates of the motifs by the new ones, the original pairing by the new one, set RMSD = RMSD′, and go to Step 2.
- Step 6: The process is complete. Return the new coordinates and RMSD′.

Advantages and Limitations of the Methodology. A great advantage of our methodology is that the accuracy of the superimposition does not depend on the sequence similarity of the superimposed motifs, as all the relevant pairings are tested. This guarantees that the methodology will find the best superimposition (i.e., the superimposition providing minimal RMSD), even when the input motifs do not have any sequence similarity. An example of employing our methodology for the superimposition of motifs that have low sequence similarity (Figure S0 a) and that do not have any sequence similarity

(Figure S0 b) is given in the Supporting Information. On the other hand, the degree of sequence similarity may affect the speed of our approach. Generally, the higher sequence similarity, the fewer pairings need to be tested, and thus, the faster the best pairing will be found. Another advantage of our methodology is that it can very effectively employ information from the PDB files and use this information to decrease the number of tested pairings (i.e., by using groupings). A further significant advantage is that the multiple superimposition does not depend on the order of superimposed motifs. Last but not least, the methodology is able to process any residues in the PDB files, including ligands.

Implementation. We implemented the above-described methodology and developed the Web application SiteBinder, which provides an effective, intuitive, and user-friendly IT solution for the superimposition of multiple protein structural motifs. SiteBinder is implemented in C# using the Microsoft Silverlight platform. Currently, the application can be run in any common Internet browser under Windows and Mac. Full Linux support will be available as soon as the new version of the Moonlight framework plugin (Linux adaptation of Microsoft Silverlight) will be released. The user interface of SiteBinder (depicted in Figure 3) consists of three basic elements: the rendering view, the input panel, and the results panel. The input panel includes the list of motifs and the selection tree.

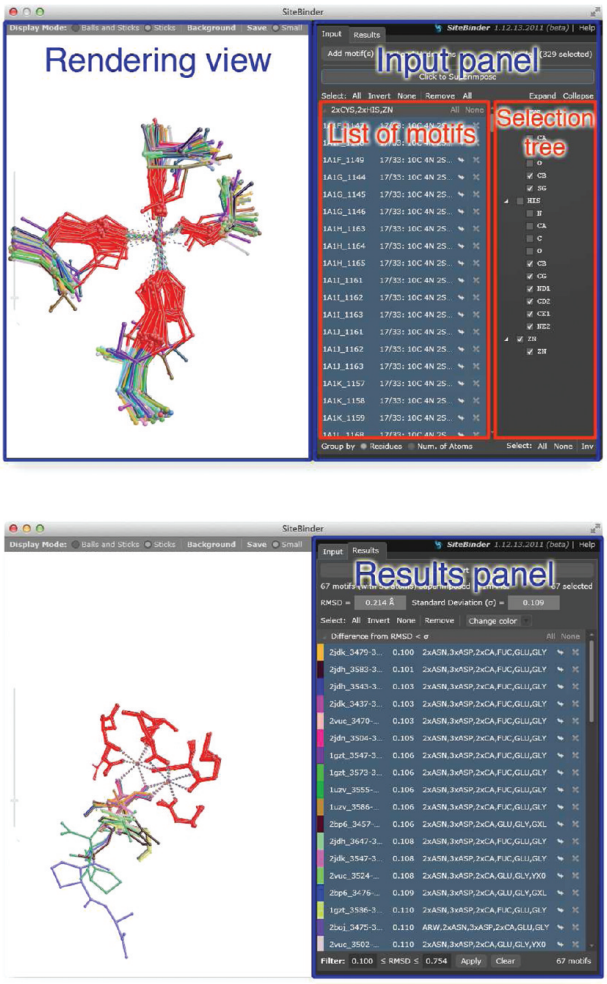


Figure 3. User interface of SiteBinder.

- The rendering view allows the user to view, rotate, and zoom the motifs; change the visualization mode (balls and sticks or sticks); or change the background. Here, the user can also select individual atoms by clicking on them.
- The list of motifs is part of the input panel and shows the loaded motifs grouped by the residues they contain. The user can add or remove motifs from this list and select the particular motifs that will be superimposed at one time.
- The selection tree is also part of the input panel and allows the user to select specific atoms or residues for superimposition.
- The results panel shows the RMSD value of the set of RMSD superimposed motifs. It also provides a list of all superimposed motifs and for each motif its RMSD compared to the average motif ($RMSD_M$). This list of superimposed motifs is sorted according to $RMSD_M$. In addition, the motifs are grouped on the basis of the difference (D_M) between RMSD and $RMSD_M$. There are four groups: $D_M < \sigma$, $\sigma \leq D_M < 2\sigma$, $2\sigma \leq D_M < 3\sigma$, and finally $D_M \geq 3\sigma$, where σ is the standard deviation of the set of D_M values. The RMSD data can be exported into a CSV table and the atomic coordinates into a PDB file. The exported atomic coordinates reflect the superimposition. The structure of the average motif structure can also be written out.

The SiteBinder is a powerful tool but still has some technical limitations. It can superimpose any motifs as long as the atom selections are compatible, meaning that the same number of atoms of the same chemical element need to be selected in each motif. SiteBinder can process at most 7000 to 10000 motifs at a time depending on the computer memory available. For optimum performance, each residue in a superimposed motif should not contain more than 12 atoms of the same element. The reason is that we employ a systematic approach to search for a relevant pairing, which can become significantly slower if each motif contains more than 12 atoms of the same element.

RESULTS AND DISCUSSION

Benchmarking Study—Comparison of Eukaryotic Linear Motifs. Linear motifs (LMs) are short elements embedded within larger protein sequence segments. They operate as regulatory sites and can be found in a wide range of proteins.[64] ELM, the Eukaryotic Linear Motif database,[65,64] is a bioinformatics resource for investigating candidate linear motifs in eukaryotic proteins. ELM currently contains 174 motifs, represented by regular expressions, which describe the occurrence of amino acids in the motif. For example, the regular expression "RF[^P][IV]" indicates that the motif should contain arginine followed by phenylalanine, then any aminoacid except for proline, afterward isoleucine or valine, and finally another amino acid.

This large and heterogeneous resource provides us with a rich area for analysis of protein motifs using SiteBinder. In our investigation, we asked two questions. First, is SiteBinder robust and fast enough to process large sets of low homology linear motifs? Second, do some linear motifs retain conservation at the level of their 3D structure?

In order to address these questions, we first prepared a data set. For each of the 174 linear motifs in ELM (access date: 1.12.2011), we found all its instances in the Protein Data Bank (access date: 1.12.2011). These instances correspond to the ELM regular expressions and may or may not perform the biological function assigned to them in the ELM database.

Table 1. Summary Information about ELM Data Set and Results of Performance and Conservation Study Performed with SiteBinder[a]

information about the motif			performance study			conservation study		
			1000 motifs			1000 motifs	motifs with RMSD < σ	
name	regular expression	no. of res.	no. of compatible atoms in a motif	time (s)	RMSD (Å)	$RMSD_B$ (Å)	no. of motifs	$RMSD_\sigma$ (Å)
LIG_AP2alpha_2	DP[FW]	3	24	10	1.936	0.833	820	0.657
LIG_RGD	RGD	3	23	59	2.603	1.077	883	0.998
LIG_MAPK_2	F.FP	4	33	60	2.693	1.443	833	1.063
LIG_HCF-1_HBM_1	[DE]H.Y	4	32	84	3.238	1.584	816	1.448
LIG_WW_1	PP.Y	4	30	31	2.987	1.601	859	1.519
LIG_EH_1	.NPF.	5	34	50	2.689	1.705	767	1.259
TRG_Cilium_RVxP_2	RV.P.	5	33	80	2.801	1.777	801	1.363
LIG_SPAK-OSR1_1	RF[^P][IV].	5	37	77	3.108	1.962	802	1.428
LIG_TRFH_1	[FY].L.P	5	34	55	3.029	1.869	839	1.525
LIG_APCC_KENbox_2	.KEN.	5	34	79	3.044	1.83	849	1.535
LIG_AP2alpha_1	F.D.F	5	38	79	3.245	1.995	807	1.657
LIG_BIR_III_2	DA.P.	5	28	51	2.641	1.865	853	1.68
LIG_WW_3	.PPR.	5	33	101	2.901	1.962	887	1.714
LIG_BIR_III_4	DA.G.	5	42	30	2.615	1.961	882	1.744
CLV_PCSK_FUR_1	R.[RK]R.	5	37	103	3.65	2.021	819	1.774
LIG_SH3_5	P..DY	5	35	20	3.188	1.963	844	1.856
LIG_EVH1_2	PP..F	5	33	41	3.027	2.096	835	1.944
LIG_PTAP_UEV_1	.P[TS]AP.	6	32	51	2.684	2.121	788	1.895
CLV_PCSK_PC7_1	[R]...[KR]R.	6	41	91	3.884	2.392	791	1.986
LIG_SH3_2	P..P.[KR]	6	33	39	3.033	2.267	883	2.113
LIG_14−3−3_1	R.[^P]([ST])[^P]P	6	35	77	3.257	2.311	861	2.131
LIG_TRAF2_2	P.Q..D	6	36	51	3.337	2.44	854	2.317
LIG_NRBOX	[^P]L[^P][^P]LL[^P]	7	40	73	2.069	1.678	884	0.657
LIG_PP2B_1	.P[^P]I[^P][IV][^P]	7	38	82	2.937	2.45	842	1.924
LIG_SH3_1	[RKY]..P..P	7	36	100	3.079	2.586	870	2.384
LIG_USP7_2	P.E[^P].S[^P]	7	38	42	3.3	2.847	828	2.592
LIG_BRCT_BRCA1_2	.(S)..F.K	7	42	71	3.803	2.928	811	2.607
LIG_RRM_PRI_1	.[ILVM]LG..P.	8	40	110	3.555	3.011	818	2.75
LIG_SH3_4	KP..[QK]...	8	43	92	4.015	3.159	866	2.935
LIG_MDM2	F...W..[LIV]	8	50	211	4.262	3.177	853	2.949
MOD_TYR_ITSM	..T.(Y)..[IV]	8	46	70	3.976	3.31	885	3.134
MOD_PKB_1	R.R..([ST])[^P]..	9	51	181	4.615	3.573	858	3.26

[a]Motifs are sorted first according to their number of residues and then according to $RMSD_\sigma$. Motifs with conserved 3D structure are marked in bold. A brief explanation of the special characters used in the regular expressions can be found on the ELM Help Page.[66]

The files containing the instances of motifs were named *pdbid_index.pdb*, where *pdbid* is the PDB ID of the parent protein, and *index* is the PDB file atom index of the first atom in the motif. Information about the number of instances of each motif and the number of proteins containing at least one instance of each motif is provided in the Supporting Information (Table S1). The program we used for identifying and retrieving ELMs from PDB is also provided in the Supporting Information (program_1). From these 174 linear motifs, we selected 32 as a relevant sample for our benchmarking study. The following criteria were used for this selection. The motif should be frequent enough but not too general (number of instances between 1000 and 30000). The motif should contain at least two identical amino acid residues, and one amino acid residue position defined by a selection of at most four possibilities. In this way, we ensure that it is meaningful to evaluate the structural conservation of the motif. After applying these criteria, we selected the minimum number of motifs so that each of the 20 amino acid residues appears as a firm part of some motif at least once. This procedure provided us with a strong data set for our benchmarking study. The names, regular expressions, and number of residues for the ELMs used in this study are summarized in Table 1.

We then focused on the first question and tested the performance of SiteBinder. For each linear motif in our data set, we selected 1000 instances. Specifically, we went through all P instances of the motif in PDB (sorted alphabetically according to their file names) and took each $(P/1000)^{th}$ instance (e.g., each second instance if the motif appeared 2000 times in PDB). Subsequently, in order to simplify the process of superimposition in SiteBinder, we used a unifying renaming convention for each motif. For instance, the residues in motif "RF[^P][IV]" were renamed as "ARG-PHE-RE1-IL_-RE2". The renaming program (program_2), the unifying residue names for each motif (Table S2), as well as the 1000 renamed instances of each motif are given in the Supporting Information. For each motif, we loaded the 1000 renamed instances into SiteBinder, selected all compatible atoms, and performed the superimposition. By "compatible atoms", we denote all heavy atoms shared by all instances of a particular motif. Table 1 shows the number of atoms used, the duration, and RMSD for each motif. The SiteBinder

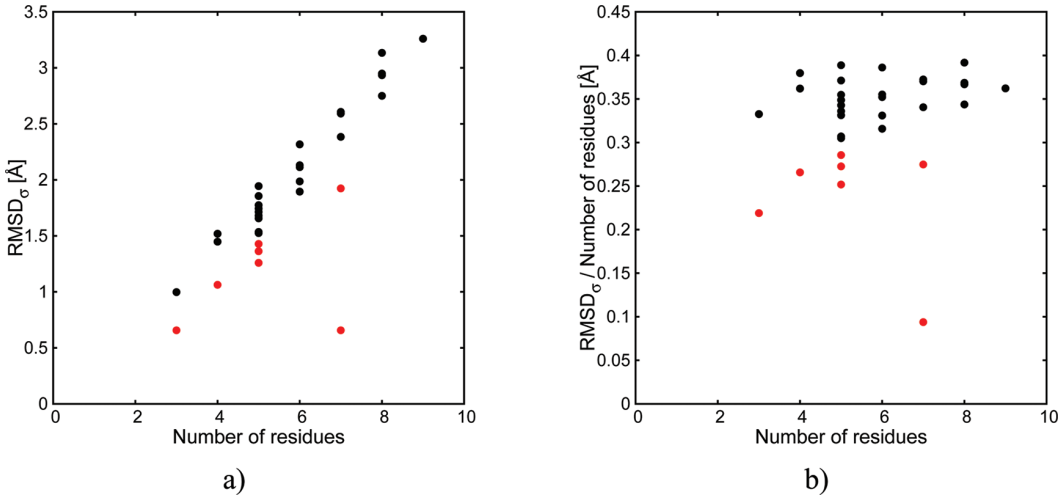


Figure 4. (a) Dependency of RMSD$_\sigma$ on the number of residues in the motif. (b) Dependency of normalized RMSD$_\sigma$ (RMSD$_\sigma$/number of residues) on the number of residues. Motifs with conserved 3D structure are marked red.

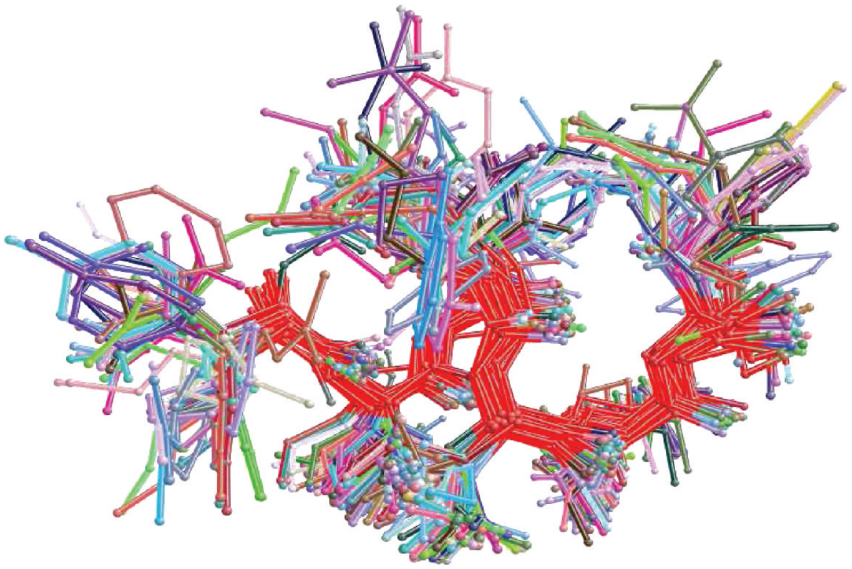


Figure 5. Superimposition of LIG_NRBOX motif instances for which RMSD < σ (only the first 80 instances are shown).

successfully performed the superimposition in all cases, regardless of the number (3 to 9), size (23 to 51 compatible atoms), nature (all 20 amino acids), or degree of conservation of the residues. These results demonstrate the robustness of SiteBinder. The performance test also highlights an exclusive feature of our multiple superimposition methodology, which is that optimal atom pairing can be achieved, and the superimposition can be performed regardless of the degree of amino acid sequence similarity.

We then addressed the second question and investigated whether some linear motifs have a particularly conserved 3D structure. For this stage of the benchmarking, which we denote the "conservation study", we used the same 32 motifs, each with 1000 (renamed) instances, but this time we used only the backbone atoms for the superimposition and obtained RMSD$_B$. To further refine our findings, we performed an additional superimposition for each motif, using only those instances with RMSD$_B$ < σ and thereby obtained RMSD$_\sigma$. The results of the conservation study are also given in Table 1.

The RMSD$_\sigma$ values provide the most relevant information for evaluating the 3D structure conservation of each motif. The RMSD$_\sigma$ grows with the growing number of residues in the motif (Figure 4a), and the dependency is mainly linear. However, seven motifs do not respect this linear trend (marked in red in Figure 4 and in bold in Table 1) and therefore seem much more structurally conserved than the other linear motifs. To clearly identify these motifs, we computed the normalized RMSD$_\sigma$ value by dividing RMSD$_\sigma$ by the number of residues in each motif. We can now clearly visualize the degree of structural conservation. The same seven motifs easily stand out in this analysis, as they have the lowest values of normalized RMSD$_\sigma$ (Figure 4b). The motif LIG_NRBOX seems to be the most structurally conserved by far (Figure 5). Several studies (e.g., Leers et al.,[67] Johansson et al.,[68] Phillips et al.[69]) come to substantiate our finding that LIG_NRBOX is highly conserved. Thus, our analysis was able to easily point out several eukaryotic linear motifs that are conserved at the structural level regardless of the degree of sequence similarity between their

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

parent proteins, and the results of this study are in agreement with published experimental results.

Case Study I—Comparison of Sugar Binding Sites in *Pseudomonas aeruginosa* Lectin II. *Pseudomonas aeruginosa* (PA) is an opportunistic pathogen that can infect almost every human tissue when immunity barriers are lowered.[70] Chronic lung colonization by the bacterium is the major cause of morbidity and mortality in cystic fibrosis patients.[71] *P. aeruginosa* produces the lectin PA-IIL (Pseudomonas lectin II, LecB), which is one of the virulent factors of the pathogen. Each monomer of this lectin contains a sugar binding site that aids the pathogen in host recognition. Knowledge of its structure can lead to better design of new antibacterial−adhesion drugs that minimize the risk of infection. The binding site contains two close calcium cations that mediate the binding of the sugar. These cations are coordinated by seven amino acids, namely, three aspartic acids, two asparagines, one glutamic acid, and one glycine from the adjacent monomer (Figure 6). The sugar is

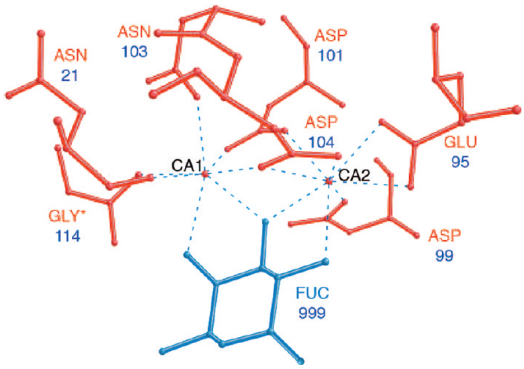


Figure 6. Amino acids coordinating calcium ions in the PA-IIL binding site with α-L-fucose. The depicted binding site originates from the monomer A of the structure with PDB ID 1uzv.

further stabilized by hydrogen bonds with other neighboring amino acids as shown by Mitchell et al.[70]

PA-IIL strongly prefers fucose, but it can also bind other saccharides, albeit with lower affinity. An interesting question that helps to understand the behavior and activity of PA-IIL is whether the structure of this binding site changes when different sugars are bound. We employed SiteBinder to address this question. First, we identified all samples of PA-IIL and its mutants present in the Protein Data Bank (access date: 3.8.2011). We then processed these samples by a program (Supporting Information, program_3) to find and extract the sugar residue, the pair of calcium atoms, and the surrounding seven amino acid residues, as described above and depicted in Figure 6. By this procedure, we obtained 18 structures of PA-IIL and its mutants, which gave us a total of 67 sugar binding sites. Most of these complexes are unique combinations of sugars and the PA-IIL protein or its mutants. There are just three exceptions, i.e., three PDB structures (1gzt, 1oxc, and 1uzv) containing wild-type PA-IIL complexed with α-L-fucose ligands. From these three closely related structures, we kept only the structure with the best resolution (i.e., 1uzv with a resolution of 1 Å) and removed the other two structures. However, we provide a comparison of these three structures in the Supporting Information (Figure S1). It documents the influence of the source organism (1gzt was purified from *P. aeruginosa*, 1ixc and 1uzv were purified from *E. coli*) and the resolution.

We thus obtained a set of 16 PA-IIL structures containing 61 sugar binding sites. These protein structures appear as protein−sugar complexes with nine different sugars. The sugar varies from monosaccharides (i.e., α-L-fucose, α-D-mannose, or α-L-galactose), via their simple derivatives (i.e., methyl-β-D-arabinoside, methyl-α-D-mannoside), to complex synthetic ligands (i.e., 2G0 or LZ0). Basic information about the PA-IIL PDB entries used in this case study can be found in the Supporting Information (Table S3).

In the next step, we used SiteBinder to superimpose the binding sites that bind the same saccharide. The most representative results of this comparison are shown in Figure 7, while the complete set of results can be found in the Supporting Information (Figure S2). These results demonstrate that the binding sites for the same sugar have a very similar structure in different PDB entries (RMSD <0.14 Å), and this feature does not depend on the size of the ligand Figure 7a compared to Figure 7b). The only exception is the binding site of α-methyl-fucoside (RMSD $\leq$ 0.478 Å).

For obtaining a broader overview and in the search for an explanation for the higher RMSD in the case of MFU binding sites, we again employed SiteBinder and superimposed all 61 sugar binding sites. The results of the superimposition are depicted in Figure 8a), and the $RMSD_M$ values are summarized in the Supporting Information (Table S4). This comparison shows that, despite the binding sites originating from different PA-IIL samples (wild types or mutants) and binding different sugars, their structure is very similar (RMSD 0.214 Å). This general comparison also explains the higher RMSD for the binding site of α-methyl-fucoside. The reason is that two of the four binding sites in a mutant of PA-IIL (PDB ID 2jdp) differ from the remaining 59 binding sites (i.e., they have the $RMSD_M$ > 0.7 Å, while the other motifs have the $RMSD_M$ < 0.2 Å). The main difference in these binding sites is that glycine is oriented outward and does not support the binding of the calcium ion (Figure 8b). Nevertheless, this exception does not change the main conclusion, which is that the sugar binding site in PA-IIL is highly conserved.

Our findings that the structure of the sugar binding site in PA-IIL is very similar for nine different sugars could be in direct correlation with the fact that PA is able to infect so many kinds of tissues. In addition, the high level of conservation of this binding site raises the question whether this motif can be used also by other organisms, and because the motif has such a well-defined 3D structure, it can be easily identified. Thus, we used our program_3 to search the complete Protein Data Bank for this motif (access date: 3.8.2011). We searched for two close calcium atoms surrounded by exactly five oxygens from ASP, two oxygens from ASN, two oxygens from GLU, and one oxygen from GLY. From each of the structures found, we obtained the binding site by extracting the sugar residue, the calcium atoms, and the seven surrounding amino acids, as depicted in Figure 6. This way, we collected the 11 sugar binding sites described in Table 2.

These binding sites originate from the proteins *Chromobacterium violaceum* lectin II (CV-IIL) and *Burkholderia cenocepacia* lectin A (BclA). Table 2 shows that the sugar binding sites in these bacteria are very similar as in PA-IIL (RMSD < 0.65 Å). This is in agreement with the fact that the biological activity of BclA[71,72] and CV-IIL[73] is very similar to that of PA-IIL. Moreover, the characteristic propeller assembly of their

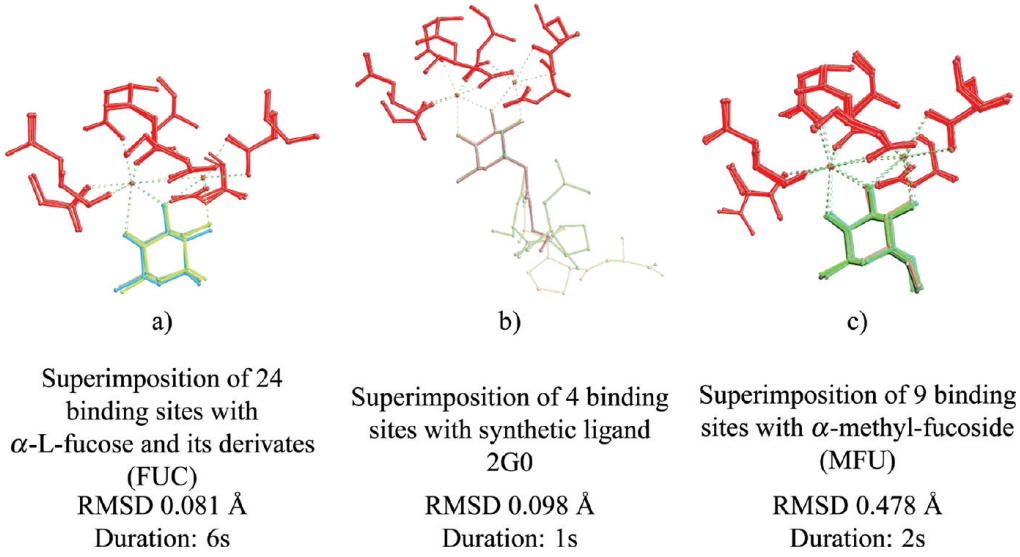


Figure 7. Representative results of the superimposition of PA-IIL binding sites that bind the same sugar-based ligand. Only the atoms in red were used for the superimposition.

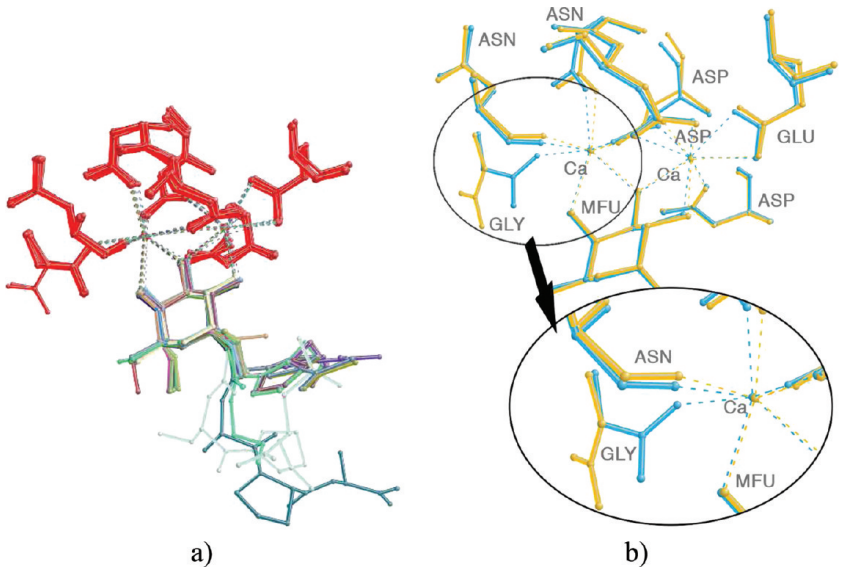


Figure 8. (a) Superimposition of all 61 sugar binding sites, RMSD 0.222 Å, duration 16 s. (b) Comparison of the sugar binding site in the wild type of PA-IIL (PDB ID 2jdm, monomer D, in blue) and in its mutant (PDB ID 2jdp, monomer D, in orange), RMSD 0.754 Å. Part of the calcium binding site in detail.

beta strands classifies these proteins as one family. Our superimposition analysis was able to immediately direct us to identify related family members without any prior knowledge of this fact (i.e., only a calculated model of the sugar binding site in PA-IIL and its mutants was used). One could envision that such analyses could be used to identify related proteins that have been misplaced in different families based on their dominant fold.

Case Study II—Comparison of Zn Binding Sites in Cys_2His_2 Zinc Fingers. Cys_2His_2 zinc fingers are one of the most common structural motifs in eukaryotes.[74,75] Each finger recognizes three to four base pairs of DNA, and several fingers can be linked in tandem to recognize a broad spectrum of DNA sequences with high specificity.[76] There is evidence that some Cys_2His_2 zinc fingers bind RNA and that others may participate in protein−protein interactions, but it appears that their predominant role is in protein−DNA recognition.[74] Individual fingers contain approximately 30 amino acids, and the hallmark of the motif is the presence of two cysteines and two histidines that serve as zinc ligands. The simplest definition of such zinc finger motifs is based on the spacing of the zinc ligands in the amino acid residue sequence. This spacing has the pattern X_2−CYS−X_{2-4}−CYS−X_{12}−HIS−X_{3-5}−HIS,[77] where X represents any amino acid residue. The abundance of this motif, its biological importance, and its simple but apposite description make it an attractive target for research.

We used SiteBinder to determine whether the center of the zinc finger motif (i.e., two CYS, two HIS, and a Zn atom) has a conserved geometry. In order to do this, we went through a few different stages. First, we used a simple program (Supporting

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

Table 2. Sugar Binding Sites with a Very Similar Structure as the PA-IIL Sugar Binding Site

protein name	PDB ID	organism	sugar	monomer	RMSD to the average motif[a] (Å)
CV-IIL	2boi	*Chromobacterium violaceum*	MFU	A	0.136
CV-IIL	2boi	*Chromobacterium violaceum*	MFU	B	0.118
CV-IIL	2bv4	*Chromobacterium violaceum*	MMA	A	0.155
CV-IIL	2bv4	*Chromobacterium violaceum*	MMA	B	0.189
BclA	2vnv	*Burkholderia cenocepacia*	MMA	A	0.621
BclA	2vnv	*Burkholderia cenocepacia*	MMA	B	0.553
BclA	2vnv	*Burkholderia cenocepacia*	MMA	C	0.633
BclA	2vnv	*Burkholderia cenocepacia*	MMA	D	0.567
BclA[b]	2wr9	*Burkholderia cenocepacia*	MAN	A	0.576
BclA	2wr9	*Burkholderia cenocepacia*	MAN	C	0.543
BclA	2wr9	*Burkholderia cenocepacia*	MAN	D	0.521

[a]The average motif was calculated by SiteBinder from all PA-IIL sugar binding sites except those from the mutant 2jdp. [b]The binding site from monomer B of BclA was not included in this study because no sugar was found in the crystal structure at this site.

Information, program_4) and collected all motifs from the Protein Data Bank (access date: 3.8.2011) that fulfill the description of zinc fingers (i.e., Zn coordinated by two CYS and two HIS that are part of a pattern of the type X_2−CYS−X_{2-4}−CYS−X_{12}−HIS−X_{3-5}−HIS). If a protein structure was obtained by NMR, only the motifs from the first model contained in the PDB file were used in our study. We found 329 zinc fingers from 205 different Protein Data Bank entries. For each hit, we extracted the zinc atom and the two HIS and two CYS surrounding this atom. By this procedure, we obtained the zinc finger central motifs and subsequently used these motifs as inputs for SiteBinder. We performed four superimpositions for our set of zinc finger central motifs. These procedures differed in the number of atoms selected for superimposition (displayed in red in Figure 9).

The first superimposition was done using only nine atoms from each motif (zinc, the nitrogens from the imidazole cycle of each HIS, and the sulfur and beta carbon of each CYS), the second superimposition with 15 atoms from each motif (to the previous selection, we added the rest of the imidazole ring atoms of each HIS and the alpha carbon of each CYS), the third with 19 atoms (to the previous selection, we added the beta carbon of each HIS and the carboxylic carbon of each CYS), and the fourth superimposition used all atoms. The superimposed motifs are depicted in Figure 9, which contains also the RMSD values and durations of the superimposition. The values of $RMSD_M$ for each individual motif in all four superimpositions are given in the Supporting Information (Table S5). The RMSD values for the first three superimpositions are similar (between 0.5 and 0.6 Å), which demonstrates that the part of the motif which closely surrounds Zn has a stable structure. Figure 9 demonstrates that the conformation of more distant parts of CYS and HIS may differ.

We further note that, despite the fact that we compared 329 motifs with 9−33 atoms, the superimposition took about 2 min even for the most complex case.

Then we focused on a special group of zinc finger Cys_2His_2 motifs, namely, those known to bind RNA. Superimposing them reveals a very interesting feature. The motifs coming from the PDB entry 1zu1 are markedly different than those in the other RNA binding proteins we investigated (Table 3). This is likely explained by the fact that one of the two HIS residues is facing the binding site with the opposite face of the imidazole ring (Figure 10). What is even more interesting is the biological consequence of this change. Unlike the other zinc finger motifs we discuss here, the motifs contained in 1zu1 have evolved to bind double stranded RNA.[78] The structural peculiarity that we identified by our superimposition analysis without any prior knowledge of RNA binding preference was confirmed by Moller et al.[78] The fact that this structural peculiarity is immediately connected to a functional peculiarity reinforces the structure−function paradigm. This reasoning could be generally applied in order to identify other proteins containing the same functional motif but with slightly different functionality and possibly different behavior toward the same drug molecules.

Case Study III—Comparison of BH3 Domains in Apoptotic Proteins. Apoptosis is a form of cell death that helps to maintain tissue homeostasis and removes malignant cells upon internal and external cellular stress in a biochemically controlled fashion. Apoptosis is downregulated (decreased) in cancer and excessive in neurodegenerative diseases or stroke. The decision whether an initial cellular signal, like a receptor induced stimulus, is tolerated or leads to cell death is controlled by a carefully balanced biochemical cascade of pro-survival or pro-apoptotic proteins of the BCL-2 family.[79,80] The proteins from a pro-apoptotic subgroup of the BCL-2 family, the BH3-only proteins, integrate specific stress signals such as genotoxic stress,[81] serum-deprivation stress,[82] or stress due to the accumulation of unfolded proteins[83] into downstream apoptotic signals. These proteins are called "BH3-only" because they share only the third (of a total of four) BCL-2 homology (BH) domains with the rest of the BCL-2 family. The proteins Bax and Bak, from another pro-apoptotic subgroup, induce the formation of pores into the mitochondrial outer membrane. This phenomenon is a decisive step in apoptosis execution. On the other hand, pro-survival BCL-2 family proteins bind to Bak and Bax, as well as to BH3-only proteins, to prevent unwanted apoptosis. The interaction between pro-survival and pro-apoptotic BCL-2 proteins is mediated by the BH3 domain.[84] The BH3 domains of BH3-only proteins consist of an amphipathic α helix and contain 9−16 amino acids.[85]

A controversy has arisen regarding the role of BH3-only proteins. Originally, it was thought that stress-induced up-regulation (increase) of BH3-only proteins is sufficient to release Bax and Bak from their complexes with pro-survival proteins and thus lead to pore formation. Nonetheless, increasing evidence indicates that an additional step is necessary, namely, the direct activation of Bax and Bak.[86] If such a step is required, two distinct groups of BH3-only proteins are predicted. One group is denoted as "enablers" and comprises the proteins Noxa, Bad, Bmf, Hrk, and Bik. These proteins presumably only bind to pro-survival proteins and thereby release Bax and Bak. The second group of BH3-only proteins, denoted as "activators" are believed to activate Bax and Bak in an explicit activation step. The proteins Bid, Bim, and Puma are examples of activators.[87]

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

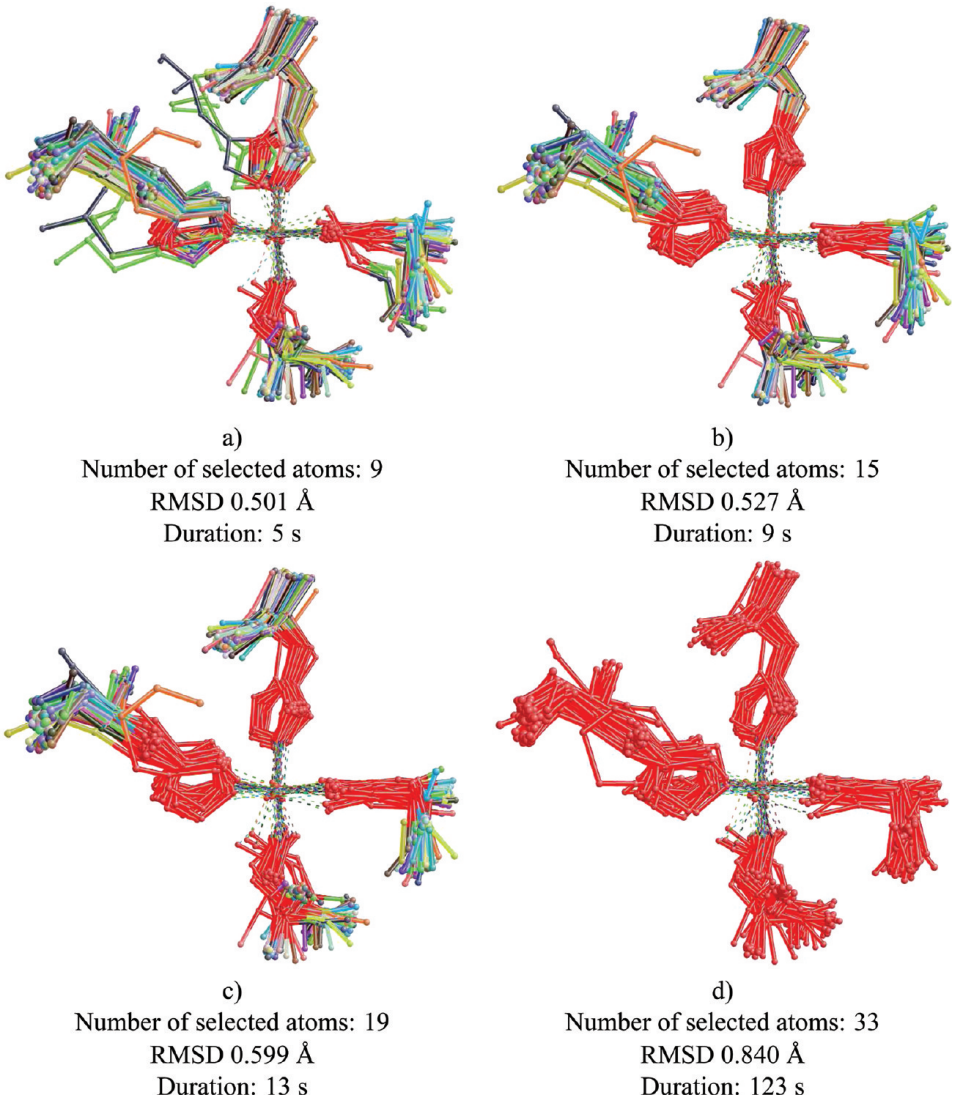


Figure 9. Superimposition of 329 zinc finger central motifs. From (a) to (d), the number of atoms used in the superimposition procedure (displayed in red) increases step by step. For ease of visual interpretation, only the first 80 motifs are displayed.

Table 3. Results of the Superimposition of Zinc Finger Central Motifs of RNA Binding Proteins

protein PDB ID	index of Zn atom	RMSD from the average model (Å)
1un6	4524	0.814
2hgh	3176	0.829
2j7j	717	0.830
1un6	4530	0.880
2hgh	3166	0.881
1un6	4523	0.898
2j7j	718	0.929
1un6	4534	0.956
2ab7	511	0.981
2ab3	494	0.984
2yu5	640	1.108
1zu1	1951	1.821
1zu1	1952	1.834

We used SiteBinder to compare the 3D structures of the BH3 domains of different BH3-only proteins with the goal to investigate whether there is a structural basis of this segregation in activators and enablers. Specifically, we focused on the proteins for which the primary structure of the BH3 domain was described and aligned by Chipuk et al.[79] We obtained the structures of these proteins from the Protein Data Bank, except for the proteins Hrk and Bik, whose structures are not available in this database. The Noxa A protein (PDB ID 2rod) was omitted because its PDB structure was determined by NMR, while the structures of all other BH3-only proteins considered here were determined by X-ray crystallography. The protein names, their PDB identifiers, the BH3-only pro-survival complex from which the structure was derived, and the amino acid sequences are given in Table 4.

As shown in Table 4, the structures of the BH3-only proteins we are using were obtained from larger complexes, in which they are bound to various pro-survival BCL-2 proteins. This complex binding may affect the structural features of the BH3-only proteins. To estimate the influence of these other proteins, we built a reference data set comprising only complexes of the BH3-only protein Bim with all relevant pro-survival BCL-2 proteins (Table 5).

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

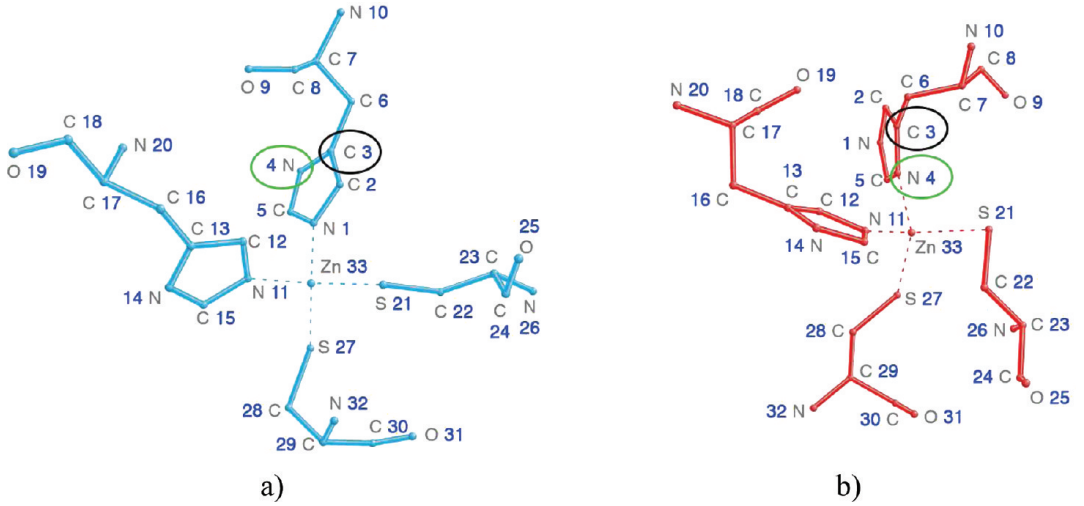


Figure 10. (a) Example of a common structure of the zinc finger central motif in RNA binding proteins (PDB ID 1un6, zinc ion with index 4524). (b) Example of a rare structure of the zinc finger central motif (PDB ID 1zu1, zinc ion with index 1951).

Table 4. Names, PDB Identifiers, and BH3 Domain Amino Acid Sequences of the Activators, and Enablers Used for the Superimposition with SiteBinder[a]

group	PDB ID	BH3-only protein	complexed with	amino acid sequence in BH3 domain[b]
activators	2voi	**Bid**	A1	ILE ALA **ARG** HIS **LEU ALA** GLN **ILE GLY ASP** GLU MET ASP
	2vm6	**Bim**	A1	ILE ALA **GLN** GLU **LEU ARG** ARG **ILE GLY ASP** GLU PHE ASN
	2vof	**Puma**	A1	ILE GLY **ALA** GLN **LEU ARG** ARG **ILE ALA ASP** ASP LEU ASN
enablers	2bzw	**Bad**	BCL-XL	TYR GLY **ARG** GLU **LEU ARG** ARG **MET SER ASP** GLU PHE GLU
	2vog	**Bmf**	A1	ILE ALA **ARG** LYS **LEU GLN** CYS **ILE ALA ASP** GLN PHE HIS
	2nla	**Noxa B**	MCL-1	GLU CYS **ALA** GLN **LEU ARG** ARG **ILE GLY ASP** LYS VAL ASN
SiteBinder denotation				A01 A02 A03 A04 A05 A06 A07 A08 A09 A10 A11 A12 A13

[a]We also mention the pro-survival proteins present in the complexes obtained from PDB. The last row shows our unifying denotation used in the SiteBinder input files. [b]The amino acids that have a degree of conservation higher than 50% for all BH3-only proteins are in bold. Information about the degree of conservation was obtained from the work of Chipuk et al.[79]

Table 5. Summary Information about the Bim Molecules Superimposed Using SiteBinder[a]

PDB ID	2vm6	3fdl	2wh6	2nl9	2pqk	3kj0	3kj1
complexed with	A1	BCL-XL	BHRF1	MCL-1	MCL-1	MCL-1	MCL-1

[a]Entries 3kj0 and 3kj1 contain Bim mutants.

We next extracted the BH3 domains characterized by the amino acid sequence described in Table 4 from the PDB files mentioned in Tables 4 and 5. The amino acid residues that had to be superimposed have different names. In order to simplify their processing by SiteBinder, we introduced a simple unifying denotation for amino acid residue names. Specifically, we renamed the BH3 domain amino acids in the SiteBinder input files according to their position in the sequence (Table 4). This solution was implemented with minimal effort and was feasible because the sequences had already been aligned. The original and modified SiteBinder input files are available in the Supporting Information.

For each of our three groups of motifs (activators, enablers, and Bim samples), we did the following:

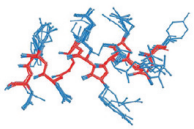
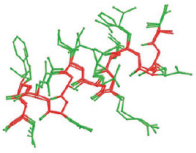
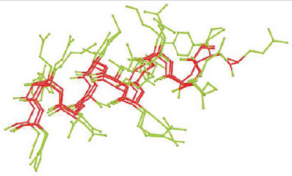
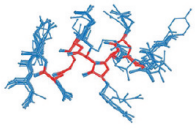
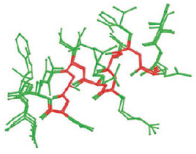
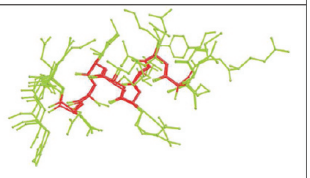
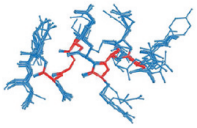
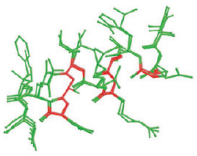
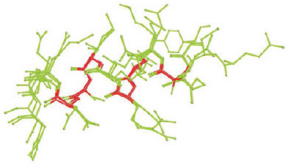
- Superimpose the entire motifs (amino acids A01−A13)
- Superimpose the inner parts of the motifs (amino acids A03−A10)
- Superimpose the conserved parts of the motifs (amino acids A03, A05, A06, A08−A10)

All motifs in a group were employed in the superimposition. Only backbone atoms were used (in red in Table 6), and thus, the RMSD reflects only the backbone geometry conservation.

The results of this superimposition are summarized in Table 6 and in the Supporting Information (Table S6). Each multiple superimposition procedure took less than 5 s. The results in Table 6 indicate that activators have a very conserved BH3 domain (RMSD < 0.25 Å, even when considering the entire motifs). On the contrary, the structure of the BH3 domain in the enablers group shows significant dissimilarity within the group, as well as to the activators group (RMSD > 0.5 Å, even for the inner or most conserved part of the motifs). In addition, comparing Bim motifs extracted from various pro-survival complexes showed smaller RMSD differences than for the activators group in general. This confirms that the pro-survival proteins did not cause significant structural changes upon complex formation. Overall, our results support the hypothesis that activators and enablers may be two functional subgroups of BH3-only proteins. Moreover, they suggest that all activators act in a similar manner to induce cell death. In contrast, the structural

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343−359

Table 6. Superimposition of the BH3 Domains from Several Data Sets (Activators, Enablers, Bim Samples)[a]

	BIM samples	Activators	Enablers
Whole motifs Amino acids: A01–A13 Number of atoms: 39	 RMSD = 0.211 Å	 RMSD = 0.246 Å	 RMSD = 1.438 Å
Middle part of the motifs Amino acids: A03–A10 Number of atoms: 24	 RMSD = 0.147 Å	 RMSD = 0.189 Å	 RMSD = 0.502 Å
Similar parts of the motifs Amino acids: A03, A05, A06, A08–A10 Number of atoms: 18	 RMSD = 0.150 Å	 RMSD = 0.194 Å	 RMSD = 0.534 Å

[a]Only the backbone atoms (in red) were used for the superimposition, and thus, the RMSD values reflect the backbone structural conservation.

heterogeneity of the BH3 domains of different enablers advocate for a specific binding to pro-survival proteins. As different stresses and cells specifically express distinct enablers, this provides a flexible, cell and stress specific, gate-keeping mechanism for enabling or preventing the activation step by the activators.

CONCLUSION

In our work, we focused on the superimposition of very large sets of protein structural motifs. We found the most appropriate state of the art superimposition algorithms available in literature, improved and compiled them, and developed a methodology that is fully tailored to the multiple superimposition of protein structural motifs. This methodology employs the systematic approach for finding the equivalence between atoms and decreases its complexity by using heuristics that consider several types of atom grouping. Fitting the motifs is solved by quaternion algebra. The described superimposition methodology guarantees that the best fit will be found and can be applied even when sequence similarity is low or does not exist at all. Multiple motifs are processed by iteratively superimposing all the structures to an average model until a stable configuration is reached. We have implemented this methodology and have created the Web application SiteBinder. This application is able to process up to thousands of protein structural motifs in a very short time (from a few seconds to a few minutes). Moreover, it provides an intuitive and user-friendly graphical interface, which allows the user to visualize the motifs, select specific atoms or residues for superimposition, export the coordinates of the superimposed structures, as well as the RMSD values, etc.

We have performed a benchmarking analysis by superimposing 1000 experimentally determined structures for each of 32 eukaryotic linear motifs. This analysis shows that our methodology and its implementation are robust, efficient, and versatile. It also demonstrates that SiteBinder can be used for studying general trends in large data sets of low homology protein structural motifs. The applicability of SiteBinder was demonstrated using three case studies that dealt with the comparison of large sets of biochemically important motifs. In the first case study, we compared the structural motifs of 61 PA-IIL sugar binding sites containing nine different sugars. The comparison showed that, despite the binding sites originating from different PA-IIL samples (wild types or mutants) and binding different sugars, their structure is very similar (RMSD 0.222 Å). This finding correlates with the ability of this pathogen to infect many kinds of host cells. In addition, we were able to identify the related proteins CV-IIL and BclA simply by studying the binding site motifs in PA-IIL. This is an example of how a superimposition analysis done with SiteBinder can help in identifying functionally related proteins. The second case study was focused on the analysis of Cys_2His_2 zinc finger structures contained in the Protein Data Bank (more than 300 motifs). We performed four different superimpositions of these motifs, successively increasing the number of superimposed atoms. The results demonstrated that the part of the motifs that closely surrounds Zn has a stable structure (RMSD values are between 0.5 and 0.6 Å). Moreover, we found that a small difference in the structure of RNA binding motifs could be responsible for binding double stranded RNA. In the last case study, we attempted to superimpose 12 BH3 domains from several pro-apoptotic proteins. The results indicated that the activators have a very conserved BH3 domain (RMSD < 0.25 Å, even for the entire motifs). On the contrary, the structure of the BH3 domain in enablers differs across this group of proteins and also differs significantly from the activator group (RMSD > 0.5 Å, even for the most conserved part of the motifs). These results are in agreement with the hypothesis that two functional subgroups of BH3-only proteins, activators and enablers, are present during apoptosis. The three case studies demonstrate the versatility of SiteBinder and show how our

dx.doi.org/10.1021/ci200444d | J. Chem. Inf. Model. 2012, 52, 343–359

software can be used to gain insight into the relationship between protein structure and function. The software is available to the community at http://ncbr.muni.cz/SiteBinder.

■ ASSOCIATED CONTENT

Supporting Information

Superimposition of motifs having low or no sequence similarity (Figure S0), detailed description of the quaternion algebra approach, formalized mathematical description of our superimposition methodology, program to extract the ELMs from PDB (program_1), program for renaming the residues in the ELM PDB files (program_2), summary information about ELMs and their occurrence in PDB (Table S1), unifying residue names for each ELM (Table S2), PDB files of the protein motifs used in benchmarking study, program to extract sugar binding sites (program_3) and zinc fingers (program_4) from PDB, PDB files of the protein motifs used in all case studies, results of the superimposition of PA-IIL sugar binding sites from proteins 1gzt, 1oxc, and 1uzv (Figure S1), results of the superimposition of PA-IIL binding sites which bind the same sugar based ligand (Figure S2), basic information about the PA-IIL PDB entries used in case study I (Table S3), $RMSD_M$ values of the superimposed motifs from case study I (Table S4), case study II (Table S5), and case study III (Table S6). This material is available free of charge via the Internet at http://pubs.acs.org/.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: svobodova@chemi.muni.cz (R.S.V.); jaroslav.koca@ceitec.muni.cz (J.K.).

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic (ME08008 to M.W.), the Czech Science foundation (GD301/09/H004 to C.M.I.), the Science Foundation Ireland Grant 08/IN.1/B1949 to H.J.H. and by the European Community's Seventh Framework Programme (CZ.1.05/1.1.00/02.0068 to J.K. and R.S.V.) from the European Regional Development Fund. C.M.I. and D.S. thank Brno City Municipality for the financial support provided to them through the program Brno Ph.D. Talent. The access to MetaCentrum supercomputing facilities provided under the research intent MSM6383917201 is highly appreciated.

■ REFERENCES

(1) Nayal, M.; Honig, B. On the nature of cavities on protein surfaces: Application to the identification of drug-binding sites. *Proteins* **2006**, *63*, 892−906.

(2) Xie, L.; Xie, L.; Bourne, P. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics* **2009**, *25*, i305−i312.

(3) Thornton, J. M.; Todd, A. E.; Milburn, D.; Borkakoti, N.; Orengo, C. A. O. A. From structure to function: Approaches and limitations. *Nat. Struc. Biol.* **2000**, *7*, 991−994.

(4) Kinoshita, K.; Nakamura, H. Protein informatics towards function identification. *Curr. Opin. Struct. Biol.* **2003**, *13*, 396−400.

(5) Watson, J. D.; Laskowski, R. A.; Thornton, J. M. Predicting protein function from sequence and structural data. *Curr. Opin. Struct. Biol.* **2005**, *15*, 275−284.

(6) Eidhammer, I.; Jonassen, I.; Taylor, W. R. Structure comparison and structure patterns. *J. Comput. Biol.* **2000**, *7*, 685−716.

(7) Chang, Y. S.; Gelfand, T. I.; Kister, A. E.; Gelfand, I. M. New classication of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage. *Proteins* **2007**, *68*, 915−921.

(8) Via, A.; Ferre, F.; Brannetti, B.; Valencia, A.; Helmer-Citterich, M. Three-dimensional view of the surface motif associated with the P-loop structure: *cis* and *trans* cases of convergent evolution. *J. Mol. Biol.* **2000**, *303*, 455−465.

(9) Ausiello, G.; Peluso, D.; Via, A.; Helmer-Citterich, M. Local comparison of protein structures highlights cases of convergent evolution in analogous functional sites. *BMC Bioinformatics* **2007**, *8*, S24.

(10) Gherardini, P. F.; Wass, M. N.; Helmer-Citterich, M.; Sternberg, M. J. E. Convergent evolution of enzyme active sites is not a rare phenomenon. *J. Mol. Biol.* **2007**, *372*, 817−845.

(11) Gasteiger, J.; Engel, T. *Chemoinformatics: A Textbook*; Wiley-VCH: Weinheim, Germany, 2003.

(12) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.

(13) Lemmen, C.; Langauer, T.; Klebe, G. FLEXS: a method for fast exible ligand superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.

(14) Cosgrove, D. A.; Bayada, D. M.; Johnson, A. P. A novel method of aligning molecules by local surface shape similarity. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 573−591.

(15) Baum, D. Multiple semi-flexible 3D superposition of drug-sized molecules. *CompLife* **2005**, *3695*, 198−207.

(16) Eidhammer, I.; Jonassen, I.; Taylor, W. R. *Protein Bioinformatics: An Algorithmic Approach to Sequence and Structure Analysis*; Wiley: Chichester, England, 2004.

(17) Gherardini, P. F.; Helmer-Citterich, M. Structure-based function prediction: Approaches and applications. *Briefings Funct. Genomics Proteomics* **2008**, *7*, 291−302.

(18) Shapiro, J.; Brutlag, D. FoldMiner: Structural motif discovery using an improved superposition algorithm. *Protein Sci.* **2004**, *13*, 278−294.

(19) Taylor, W. R.; Orengo, C. A. Protein structure alignment. *J. Mol. Biol.* **1989**, *208*, 1−22.

(20) Holm, L.; Park, J. DaliLite workbench for protein structure comparison. *Bioinformatics* **2000**, *16*, 566−567.

(21) Michalopoulos, I.; Torrance, G. M.; Gilbert, D. R.; Westhead, D. R. TOPS: An enhanced database of protein structural topology. *Nucleic Acids Res.* **2004**, *32*, D251−D254.

(22) Harrison, A.; Pearl, F.; Sillitoe, I.; Slidel, T.; Mott, R.; Thornton, J.; Orengo, C. Recognizing the fold of a protein structure. *Bioinformatics* **2003**, *19*, 1748−1759.

(23) Madej, T.; Gibrat, J. F.; Bryant, S. H. Threading a database of protein cores. *Proteins* **1995**, *23*, 356−369.

(24) Shindyalov, I. N.; Bourne, P. E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **1998**, *11*, 739−747.

(25) Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. Recognition of functional sites in protein structures. *J. Mol. Biol.* **2004**, *339*, 607−633.

(26) Chang, D. T.; Chen, C.; Chung, W.; Oyang, Y.; Juan, H.; Huang, H. ProteMiner-SSM: A web server for efficient analysis of similar protein tertiary substructures. *Nucleic Acids Res.* **2004**, *32*, W76−W82.

(27) Spriggs, R. V.; Artymiuk, P. J.; Willett, P. Searching for patterns of amino acids in 3D protein structures. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 412−421.

(28) Kinoshita, K.; H., N. Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. *Protein Sci.* **2003**, *12*, 1589−1595.

(29) Ausiello, G.; Via, A.; M., H.-C. Query3d: A new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinf.* **2005**, *6*, S5.

(30) Barker, J. A.; Thornton, J. M. An algorithm for constraint-based structural template matching: Application to 3D templates with statistical analysis. *Bioinformatics* **2003**, *19*, 1644−1649.

(31) Jambon, M.; Imberty, A.; Deleage, G.; Geourjon, C. A new bioinformatic approach to detect common 3D sites in protein structures. *Proteins* **2003**, *52*, 137−145.
(32) Gherardini, P. F.; Ausiello, G.; Helmer-Citterich, M.; Hofmann, A. A local structural comparison program that allows for user-defined structure representations. *PloS One* **2010**, *5*, e11988.
(33) Moll, M.; Bryant, D. H.; Kavraki, L. E. The LabelHash algorithm for substructure matching. *BMC Bioinformatics* **2010**, *11*, 555.
(34) Ferre, F.; Ausiello, G.; Zanzoni, A.; Helmer-Citterich, M.; Helmer-Citterich, M. Functional annotation by identification of local surface similarities: A novel tool for structural genomics. *BMC Bioinformatics* **2005**, *6*, 194.
(35) Powers, R.; Copeland, J. C.; Germer, K.; Mercier, K. A.; Ramanathan, V.; Revesz, P. Comparison of protein active site structures for functional annotation of proteins and drug design. *Proteins* **2006**, *65*, 124−135.
(36) Kellenberger, E.; Schalon, C.; Rognan, D. How to measure the similarity between protein ligand-binding sites? *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 209−220.
(37) Schalon, C.; Surgand, J. S.; Kellenberger, E.; Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **2008**, *71*, 1755−1778.
(38) Bauer, R. A.; Bourne, P. E.; Formella, A.; Frommel, C.; Gille, C.; Goede, A.; Guerler, A.; Guerler, A.; Hoope, A.; Knapp, E. W.; Poschel, T. Others, superimpose: A 3D structural superposition server. *Nucleic Acids Res.* **2008**, *36*, W47.
(39) Coutsias, E. A.; Seok, C.; Dill, K. A. Using quaternions to calculate RMSD. *J. Comput. Chem.* **2004**, *25*, 1849−1857.
(40) Hess, B.; Kutzner, C.; Spoel, D.; Lindahl, E. Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(41) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera: A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(42) *MOE (The Molecular Operating Enviroment)*, version 2005.06; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2009.
(43) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM: A new method for protein modeling and design. Application to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.
(44) Zu-Kang, F.; Sippl, M. J. Optimum superimposition of protein structures: Ambiguities and implications. *Fold. Des.* **1996**, *1*, 123−132.
(45) Raymond, J. W.; Willett, P. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.
(46) Raymond, J. W.; Gardiner, E. J.; Willett, P. Rascal: Calculation of graph similarity using maximum common edge subgraphs. *Comput. J.* **2002**, *45*, 631.
(47) Humphrey, W.; Dalke, A.; Schulten, K. VMD: VisualMolecular Dynamics. *J. Mol. Graphics* **1996**, *14*, 33−38.
(48) Laaksonen, L. *gOpenmol*, version 2.0; CSC — IT Center for Science Ltd.: Espoo, Findland, 2001.
(49) *The PyMOL Molecular Graphics System*, version 1.3r1; Schrödinger, LLC: New York, 2010.
(50) Needleman, S.; Wunsch, C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**, *48*, 443−453.
(51) Abagyan, R. A.; Batalov, S. Do aligned sequences share the same fold? *J. Mol. Biol.* **1997**, *273*, 355−368.
(52) *Discovery Studio*, version 2.5; Accelrys Software, Inc.: San Diego, CA, 2009.
(53) Chen, B. Y.; Fofanov, V. Y.; Kimmel, D. M.; Lichtarge, O.; Kavraki, L. E. Algorithms for structural comparison and statistical analysis of 3d protein motifs. *Pac. Symp. Biocomput.* **2005**, 334−345.
(54) McLachlan, A. D. A mathematical procedure for superimposing atomic coordinates of proteins. *Acta Crystallogr.* **1972**, *28*, 656−657.
(55) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.* **1976**, *32*, 922−923.
(56) Horn, B. K. P. Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Am. A* **1987**, *4*, 629−642.
(57) Diamond, R. A note on the rotational superposition problem. *Acta Crystallogr.* **1988**, *44*, 211−216.
(58) Kearsley, S. K. On the orthogonal transformation used for structural comparisons. *Acta Crystallogr.* **1989**, *45*, 208−210.
(59) Karney, C. F. F. Quaternions in molecular modeling. *J. Mol. Graphics Modell.* **2007**, *25*, 1849−1857.
(60) Konagurthu, A. S.; Whisstock, J. C.; Stuckey, P. J.; Lesk, A. M. MUSTANG: A multiple structural alignment algorith. *Proteins* **2006**, *64*, 559−574.
(61) Wang, X.; Snoeyink, J. Defining and computing optimum RMSD for gapped and weighted multiple-structure alignment. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2008**, *5*, 525−533.
(62) Koca, J. A mathematical model of the logical structure of chemistry. A bridge between theoretical and experimental chemistry and a general tool for computer-assisted molecular design I. An abstract model. *Theor. Chim. Acta* **1991**, *80*, 29−50.
(63) Koca, J. A mathematical model of realistic constitutional chemistry. A synthon approach. II: The model and organic synthesis. *J. Math. Chem.* **1989**, *3*, 73−89.
(64) Gould, C. M.; et al. ELM: The status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.* **2010**, *38*, D167−D180.
(65) Puntervoll, P.; et al. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* **2003**, *31*, 3625−3630.
(66) ELM Help Page. Functional Sites in Proteins, The Eukaryotic Linear Motif. http://elm.eu.org/infos/help.html#regular_expressions (accessed December 1, 2011).
(67) Leers, J.; Treuter, E.; Gustafsson, J. Mechanistic principles in NR box-dependent interaction between nuclear hormone receptors and the coactivator TIF2. *Mol. Cell Biol.* **1998**, *18*, 6001−6013.
(68) Johansson, L.; Bavner, A.; Thomsen, J.; Farnegardh, M.; Gustafsson, J.; Treuter, E. The orphan nuclear receptor SHP utilizes conserved LXXLL-related motifs for interactions with ligand-activated estrogen receptors. *Mol. Cell Biol.* **2000**, *20*, 1124−1133.
(69) Phillips, K. J.; Rosenbaum, D. M.; Liu, D. R. Binding and stability determinants of the PPAR gamma nuclear receptor-coactivator interface as revealed by shotgun alanine scanning and in vivo selection. *J. Am. Chem. Soc.* **2006**, *128*, 11298−11306.
(70) Mitchell, E.; Houles, C.; Sudakevitz, D.; Wimmerova, M.; Gautier, C.; Perez, S.; Wu, A. M.; Gilboa-Garber, N.; Imberty, A. Structural basis foroligosaccharide-mediated adhesion of *Pseudomonas aeruginosa* in the lungs of cystic fibrosis patients. *Nat. Struct. Biol.* **2002**, *9*, 918−921.
(71) Govan, J. R. W.; Deretic, V. Microbial pathogenesis in cystic fibrosis: Mucoid *Pseudomonas aeruginosa* and *Burkholderia cepacia*. *Microbiol. Rev.* **1996**, *60*, 539−574.
(72) Lameignere, E.; Malinovska, L.; Slavikova, M.; Duchaud, E.; Mitchell, E. P.; Varrot, A.; Sedo, O.; Imberty, A.; Wimmerova, M. Structural basis for mannose recognition by a lectin from opportunistic bacteria *Burkholderia cenocepacia*. *Biochem. J.* **2008**, *411*, 307−318.
(73) Pokorna, M.; Cioci, G.; Perret, S.; Rebuffet, E.; Kostlanova, N.; Adam, J.; Gilboa-Garber, N.; Mitchell, E. P.; Imberty, A.; Wimmerova, M. Unusual entropy-driven affinity of *Chromobacterium violaceum* lectin CV-IIL toward fucose and mannose. *Biochemistry* **2006**, *45*, 7501−7510.
(74) Pabo, C. O.; Peisach, E.; Grant, R. A. Design and selection of novel Cys2His2 zinc finger proteins. *Annu. Rev. Biochem.* **2001**, *70*, 313−340.
(75) Krishna, S. S.; Majumdar, I.; Grishin, N. V. Structural classification of zinc fingers: Survey and summary. *Nucleic Acids Res.* **2003**, *31*, 532−550.
(76) Choo, Y.; Sanchez-Garcia, I.; Klug, A. In vivo repression by a site-specific DNA-binding protein designed against an oncogenic sequence. *Nature* **1994**, *372*, 642−645.
(77) Brown, R. S.; Sander, C.; Argos, P. The primary structure of transcription factor TFIIIA has 12 consecutive repeats. *FEBS Lett.* **1985**, *186*, 271−274.

(78) Moller, H.; Martinez-Yamout, M.; Dyson, H.; Wright, P. Solution structure of the N-terminal zinc fingers of the *Xenopus laevis* double-stranded RNA-binding protein ZFa. *J. Mol. Biol.* **2005**, *351*, 718−730.
(79) Chipuk, J. E.; Moldoveanu, T.; Llambi, F.; Parsons, M. J.; Green, D. R. The Bcl-2 family reunion. *Mol. Cell* **2010**, *37*, 299−310.
(80) Huber, H. J.; Duessmann, H.; Wenus, J.; Kilbride, S. M.; Prehn, J. H. Mathematical modelling of the mitochondrial apoptosis pathway. *Biochim. Biophys. Acta* **2011**, *1814*, 608−615.
(81) Herr, I.; Debatin, K. M. Cellular stress response and apoptosis in cancer therapy. *Blood* **2001**, *89*, 2603−2614.
(82) Bruns, C. J.; Harbison, M. T.; Davis, D. W.; Portera, C. A.; Tsan, R.; Hicklin, D. J.; Radinsky, R. Epidermal growth factor receptor blockade with C225 plus gemcitabine results in regression of human pancreatic carcinoma growing orthotopically in nude mice by antiangiogenic mechanisms. *Clin. Cancer Res.* **2000**, *6*, 1936−1948.
(83) Ron, D.; Walter, P. Signal integration in the endoplasmic reticulum unfolded protein response. *Nat. Rev. Mol. Cell Biol.* **2007**, *8*, 519−529.
(84) Han, J.; Flemington, C.; Houghton, A. B.; Gu, Z.; Zambetti, G. P.; Lutz, R. J.; Zhu, L.; Chittenden, T. Expression of bbc3, a pro-apoptotic BH3-only gene, is regulated by diverse cell death and survival signals. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 11318−11323.
(85) Shibue, T.; Taniguchi, T. BH3-only proteins: Integrated control point of apoptosis. *Int. J. Cancer* **2006**, *119*, 2036−2043.
(86) Leber, B.; Lin, J.; Andrews, D. W. Embedded together: the life and death consequences of interaction of the Bcl-2 family with membranes. *Apoptosis* **2007**, *12*, 897−911.
(87) Green, D. R. *Means to an End: Apoptosis and Other Cell Death Mechanisms*; Cold Spring Harbor Laboratory Press: New York, 2011.

Predicting pK_{a} Values of substituted phenols from atomic charges: comparison of different quantum mechanical methods and charge distribution schemes

JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Predicting pK_a Values of Substituted Phenols from Atomic Charges: Comparison of Different Quantum Mechanical Methods and Charge Distribution Schemes

Radka Svobodová Vařeková,[†] Stanislav Geidl,[†] Crina-Maria Ionescu,[†] Ondřej Skřehota,[†] Michal Kudera,[†] David Sehnal,[†] Tomáš Bouchal,[†] Ruben Abagyan,[‡] Heinrich J. Huber,[§] and Jaroslav Koča*[,†]

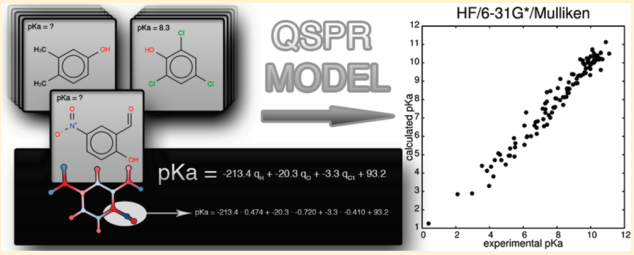
[†]National Centre for Biomolecular Research, Faculty of Science and CEITEC — Central European Institute of Technology, Masaryk University Brno, Kamenice 5, 625 00, Brno-Bohunice, Czech Republic

[‡]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, 9500 Gilman Drive, MC 0657, San Diego, California, United States

[§]Systems Biology Group, Royal College of Surgeons in Ireland, 123 St Stephens Green, Dublin 2, Ireland

Supporting Information

ABSTRACT: The acid dissociation (ionization) constant pK_a is one of the fundamental properties of organic molecules. We have evaluated different computational strategies and models to predict the pK_a values of substituted phenols using partial atomic charges. Partial atomic charges for 124 phenol molecules were calculated using 83 approaches containing seven theory levels (MP2, HF, B3LYP, BLYP, BP86, AM1, and PM3), three basis sets (6-31G*, 6-311G, STO-3G), and five population analyses (MPA, NPA, Hirshfeld, MK, and Löwdin). The correlations between pK_a and various atomic charge descriptors were examined, and the best descriptors were selected for preparing the quantitative structure—property relationship (QSPR) models. One QSPR model was created for each of the 83 approaches to charge calculation, and then the accuracy of all these models was analyzed and compared. The pK_as predicted by most of the models correlate strongly with experimental pK_a values. For example, more than 25% of the models have correlation coefficients (R^2) greater than 0.95 and root-mean-square errors smaller than 0.49. All seven examined theory levels are applicable for pK_a prediction from charges. The best results were obtained for the MP2 and HF level of theory. The most suitable basis set was found to be 6-31G*. The 6-311G basis set provided slightly weaker correlations, and unexpectedly also, the STO-3G basis set is applicable for the QSPR modeling of pK_a. The Mulliken, natural, and Löwdin population analyses provide accurate models for all tested theory levels and basis sets. The results provided by the Hirshfeld population analysis were also acceptable, but the QSPR models based on MK charges show only weak correlations.



INTRODUCTION

The acid dissociation (ionization) constant pK_a is one of the fundamental properties of organic molecules determining the degree of dissociation at a given pH. Dissociation constants are of interest in chemical, biological, environmental, and pharmaceutical research because the important physicochemical properties, like lipophilicity, solubility, and permeability, are all pK_a dependent. The values of these constants are, e.g., essential for absorption, distribution, metabolism, elimination (ADME) profiling.[1] Ionization constants also provide an insight into interactions of drugs containing ionizable groups with a receptor. In drug formulation, pK_a is important for the choice of an appropriate excipient and counterion. Furthermore, pK_a is often used as a descriptor for quantitative structure—activity relationship (QSAR) models. For these reasons, there is a strong interest in the development of reliable methods for pK_a prediction.

Numerous pK_a prediction methods based on different approaches were developed. The linear free energy relationships (LFER) method,[2,3] applying the Hammett and Taft equations, is one of the first approaches used for pK_a prediction. LFER models are still used and have been implemented in popular software packages, such as ACD/pK_a,[4] EPIK,[5] and SPARC.[6] Database methods use similarity metrics[7] to assign the pK_a value of the molecule of interest to the pK_a value obtained from the most similar molecule found in dedicated databases. Likewise, the decision tree method uses similarity metrics and builds a tree which provides a decision path for processing a compound. Ab initio quantum mechanical (QM) methods have often been found to be the most accurate,[8] such as the Jaguar pK_a prediction module,[9] which performs geometry

Received: March 18, 2011

   dx.doi.org/10.1021/ci200133w | J. Chem. Inf. Model. XXXX, XXX, 000–000

optimization at the density functional theory (DFT) B3LYP/6-31G* level, or the approach of Shields et al.,[10] which used the CPCM[11] continuum solvation model in Gaussian 98.[12] On the other hand, the applicability of these approaches is limited due to their computational complexity. A popular way to benefit from quantum mechanical calculations while keeping lower computational costs is to use QM descriptors which have a strong correlation with pK_a. Such descriptors include, e.g., polarizability,[13] free energies [phenoxide highest occupied molecular orbital (HOMO) energy,[14] relative proton transfer energy,[14] minimum surface local ionization energy],[15] partial atomic charges,[16,17] group philicity,[18] molecular electrostatic potential,[19] etc. Information-based descriptors (i.e., molecular tree structured fingerprints or 2D substructure flags,[20] topological sphere descriptors,[21] steric descriptors,[21] etc.) are also applicable. One of the most common techniques that uses descriptors in pK_a prediction is QSAR or quantitative structure − property relationships (QSPR) in combination with partial least-squares (PLS) or multiple linear regression (MLR), while other methods include artificial neural networks (ANN).[8] Unfortunately, pK_a values remain one of the most challenging physicochemical properties to predict.

Using partial atomic charges to estimate the relative acidity or reactivity of organic compounds has a long history in organic chemistry. Specifically, the partial atomic charges concept allows the prediction of relative acidity or reactivity by estimating the extent of charge delocalization based on molecular structure information.

Therefore, the correlation between pK_a and relevant atomic charges calculated by different ab initio or semiempirical approaches has been analyzed. For example, Gross et al.[22] studied which population analyses provide a good correlation at the B3LYP/6-311G** level of theory for substituted phenols and anilines, and Kreye et al.[23] compared three different levels of theory for substituted phenols (RM1 with and without the SM8 solvent model and B3LYP/6-311G** and B3LYP/6-31+G* with the SM8 solvent model). Partial atomic charges are also often and successfully used as part of the descriptors set in QSAR/QSPR models. Dixon et al. calculated pK_a from σ and π partial charges,[17] Citra[16] used partial charges and bond order, Xing et al.[24] charges and polarizabilities, Soriano et al.[25] charges and frontier orbital energy, and Yangjeh[13] combined charges, polarizability, molecular weight, hydrogen-bond accepting capability, and partial-charge weighted topological electronic descriptors. The above studies demonstrate that charges are very powerful descriptors for pK_a modeling and show linear dependency between charges and pK_a. Charge utilization has been limited by the high computational cost of their quantum mechanical calculation.

Nowadays, computers are powerful enough to make QM charges accessible in a much shorter time. Moreover, empirical charge calculation approaches, like equalization methods,[26] are able to mimic QM methods with high accuracy, and such empirical approaches are even markedly faster than QM methods themselves. These facts create a good reason to develop accurate pK_a prediction models that are based on QM charges, because they can subsequently be used in techniques like virtual screening.

In the present study we report on the evaluation of pK_a prediction QSPR models based on 83 different charge calculation approaches. Specifically, we applied 83 combinations of theory levels (MP2, HF, B3LYP, BLYP, BP86, AM1, and PM3), basis sets (6-31G*, 6-311G, STO-3G) and population analyses (MPA, NPA, Hirshfeld, MK, and Löwdin). Then, we compared the correlations between experimental pK_a values and various atomic charge descriptors and used the best descriptors for designing the QSPR models. We created a model for each of the 83 approaches of charge calculation and subsequently analyzed the ability of these models to predict pK_a. The analysis was performed on phenol molecules, a class of compounds frequently used for the evaluation of pK_a prediction models.[16,22,23]

There are basically two possible ways to create a QSPR model of a feature to be predicted. The first is to create as general a model as possible, with the risk that the accuracy of such a model may not be high. The second approach is to develop more models, each of them being dedicated to a certain class of compounds. In our work, we follow the second approach and start with phenols.

METHODS

Data Sets. Our data set contains the 3D structures of 124 distinct phenol molecules. The list of the molecules, including their experimental pK_a values, can be found in the Supporting Information. This data set is of high structural diversity, meaning it contains a wide range of electron-withdrawing and electron-donating substituents, covering a pK_a range of about 10 log units. The molecules were obtained from the NCI Open Database Compounds.[27] This database consists of organic molecules tested against cancer, and it includes their two-dimensional (2D) structures and also their 3D structures predicted by CORINA 2.6.[28] The main reason we used the CORINA approach is speed and compatibility with some other studies. The key point is speed. Our final goal with the approach is to use it when searching large databases for virtual screening purposes. CORINA provides an approximation of the global minimum conformation very quickly. Moreover, it is quite a common software for the preparation of 3D structures used in the validation of pK_a prediction models.[20,21,29]

pK_a Values. The experimental pK_a values were taken from the Physprop database.[30] The structures of phenol molecules from the NCI Open Database and their Physprop pK_a values were paired using the CAS registry numbers, which are unique identifiers in both databases.

Atomic Charge Calculation. All atomic charge calculations were carried out using Gaussian03 from Gaussian Inc.[31] The merely inputs for charge calculations were the 3D coordinates generated by CORINA, i.e., without any further geometry optimization (in a similar way as Ertl et al.).[20] The reason why QM optimization was skipped is again the speed of the approach. An optimization procedure even based on a QM method would bring the problem to a different level of computational complexity and related cost, which would not allow it to be used for our intended purposes, i.e., searching large databases.

Five ab initio levels of theory were examined. The first two were the standard Hartree−Fock (HF) method and the second-order Møller−Plesset (MP2) perturbation theory, which includes more sophisticated approximations of the Hamiltonian compared to HF. A computational cost of HF and MP2 is $\theta(N^4)$ and $\theta(N^5)$, respectively, where N is the number of basis functions. The other three were the DFT methods with BLYP, BP86, and B3LYP functionals. BLYP is a representative of gradient corrected functionals and is denoted according to its authors (Becke, Lee, Yang and Parr). BP86 (Becke Perdew 1986) is

dx.doi.org/10.1021/ci200133w |J. Chem. Inf. Model. XXXX, XXX, 000–000

similar to BLYP but uses an older correlation functional (Perdew-86). B3LYP (Becke, three-parameter, Lee−Yang−Parr) is a hybrid functional constructed as a linear combination of the HF and BLYP functionals. A computational cost of all these DFT methods is $\theta(N^3)$. The basis sets STO-3G, 6-31G*, and 6-311G were used for each level of theory, therefore 15 combinations of theory levels and basis sets were studied (HF/STO-3G, HF/6-31G*, HF/6-311G, ..., BP86/STO-3G, BP86/6-31G*, and BP86/6-311G). Five types of charges were calculated for each of these 15 pairs of theory levels and basis sets—charges derived from: natural population analysis (NPA), Mulliken charges (MPA), Löwdin charges, Hirshfeld charges, and Merz−Singh−Kollman charges fit to the electrostatic potential (MK). Moreover, the application of two semiempirical methods Austin model 1 (AM1) and parameterization method 3 (PM3) with four PA (Mulliken, Löwdin, Hirshfeld, and MK) was analyzed. Both AM1 and PM3 exhibit computational cost of $\theta(N^3)$. Consequently, this publication examines 83 approaches for charge calculation and analyzes their relevance for pK_a calculation.

Descriptors. The selection of appropriate descriptors that are significantly related to the property of interest is very important for predictive QSPR models. The descriptors can be chosen using domain knowledge about the examined property, or the mathematical methods for the selection of descriptors can be applied. In our work, we have utilized both approaches. We have focused on atomic charges and their ability to estimate the pK_a of phenols. Therefore, atomic charges and their sums and differences have been employed as the descriptors. First, according to traditional knowledge about atomic charge influence on pK_a in phenols, we selected the atomic charge of the hydrogen atom from the phenolic OH group (q_H) and the atomic charges of the atoms close to this hydrogen as descriptors. These atoms and their denotations are shown in Figure 1. We also verified in all our molecules that this hydrogen is the most positively charged hydrogen, and therefore this hydrogen will dissociate first. Further descriptors are therefore the charge on the oxygen atom (q_O), the charge on the C1 carbon atom (q_{C1}), and the charge on the C4 carbon atom (q_{C4}). Because it is not possible to distinguish between the charges on C2 and C6, the sum of these charges was used as a descriptor (q_{C2+C6}) and the same for C3 and C5 (q_{C3+C5}). We further evaluated as descriptors also the sums and the differences of these atomic charges—the difference between the O and H charge (q_{O-H}), the sum of charges on C1, C2, and C6 ($q_{C1+C2+C6}$), the sum of charges on C3, C4, and C5 ($q_{C3+C4+C5}$), and the sum of charges on all carbons in the phenolic group (q_{phe}). After this we evaluated the correlation between these 10 descriptors and the experimental pK_a values using the squared Pearson correlation coefficient (R^2) and the Student's statistic of the regression (t) in order to find descriptors significantly correlating with pK_a. These descriptors were used to establish the QSPR models.

QSPR Models: Parametrization and Quality Evaluation. The general equation for our QSPR models is

$$pK_a = param_1 \cdot descr_1 + param_2 \cdot descr_2 + ... + param_n \cdot descr_n + param_{n+1} \quad (1)$$

where $descr_1$, $descr_2$, ..., $descr_n$ are the descriptors mentioned above; $param_1$, $param_2$, ..., $param_{n+1}$ are parameters of the QSPR model (i.e., constants derived by multiple linear regression), and n is the number of descriptors in the QSPR model. The parametrization of the QSPR models was done by MLR. We prepared one model for each procedure of charge calculation;

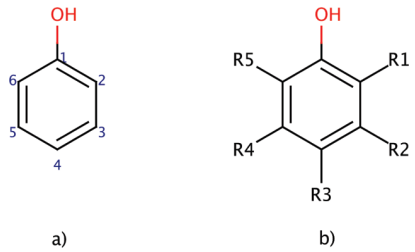


Figure 1. (a) The atom enumeration in phenols. (b) Markush structure of molecules from the data set, where R1, R2, ... R5 = $-CH_3$, $-CH{=}O$, $-C_6H_5$, $-O-CH_2-CH_3$, $-CH(CH_3)_2$, $-O-CH_3$, $-C{=}O-CH_3$, $-C{=}O-NH_2$, $-CH_2-OH$, ... $-Cl$, $-Br$, $-F$, and $-NO_2$.

therefore 83 different QSPR models were generated. The quality of the QSPR models, i.e., the correlation between experimental pK_a and the pK_a calculated by the model, was evaluated using the squared Pearson correlation coefficient (R^2), root-mean-square error (RMSE), average absolute pK_a error ($\overline{\Delta}$), standard deviation of the estimation (s), and Fisher's statistics of the regression (F). The robustness of the models was tested by cross-validation. Details about this procedure and its results are described in the following text.

RESULTS AND DISCUSSION

Evaluation of Descriptors. As the first step of our study, we investigated the pK_a predicting capabilities of all 10 suggested descriptors: q_H, q_O, q_{C1}, q_{C2+C6}, q_{C3+C5}, q_{C4}, ..., q_{phe}. Consequently, we calculated the atomic charges of all 124 phenol molecules from the data set via 83 combinations of theory levels (HF, MP2, B3LYP, BLYP, BP86, AM1, and PM3), population analyses (natural, Mulliken, Löwdin, Hirshfeld, and MK) and basis sets (STO-3G, 6-311G, and 6-31G*). And afterward we calculated the squared Pearson coefficients (R^2) and Student's t-value (t) for the correlations between each descriptor and experimental pK_a values for all 83 procedures of charge calculation.

The Hirshfeld PA demonstrates an untypical correlation between descriptors and pK_a for the basis sets STO-3G and 6-311G with all levels of theory, where the set contains eight strong outliers, all bromophenol molecules in the data set, and there is no reasonable correlation (Figure 2a). When the outliers were removed, the correlations became similar to those for Mulliken, natural, or Löwdin population analyses (Figure 2b). When the polarization basis set 6-31G* is used or when the semiempirical methods are applied, the charges obtained via the Hirshfeld PA do not contain the outliers. Therefore, we removed the bromophenols from the data set and recalculated the correlation coefficients and Student's t-values for the Hirshfeld PA and the basis sets STO-3G and 6-311G using this reduced set of 116 molecules.

The values of R^2 and t for all charge calculation procedures and all descriptors are summarized in the Supporting Information (Table S1), and a set of selected values of R^2 are visualized in Figures 3−5. These results show that q_H and q_O have a high correlation with experimental pK_a, i.e., $R^2 > 0.8$ for most charge calculation approaches. Almost all these correlation coefficients are statistically significant at $p = 0.05$. It is worth mentioning that, for the sets with 124 or 116 molecules, the R^2 is statistically significant (at $p = 0.05$) when $t > 1.66$. Also q_{C1} exhibits a good correlation, i.e., $R^2 > 0.5$ for some approaches

dx.doi.org/10.1021/ci200133w |J. Chem. Inf. Model. XXXX, XXX, 000–000

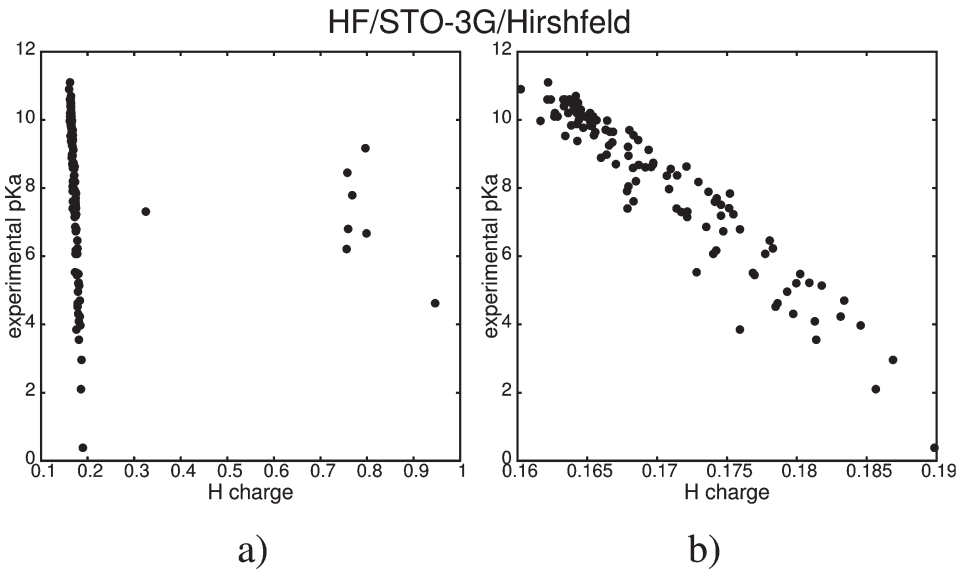


Figure 2. Correlations between q_H and pK_a for HF, STO-3G, and Hirshfeld PA. Graph (a) with and (b) without outliers.

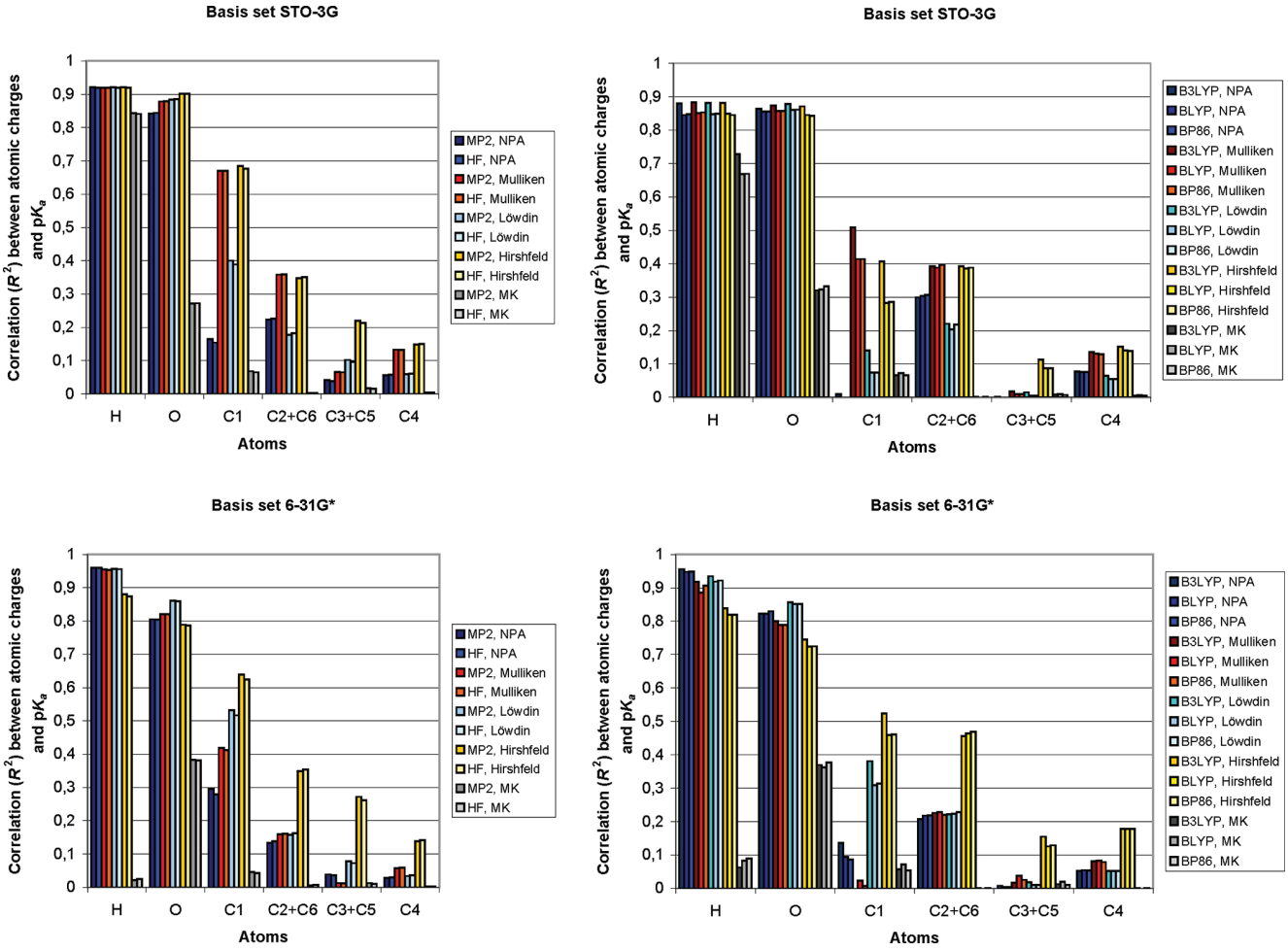


Figure 3. Correlations between descriptors and experimental pK_a.

and $t > 1.66$ for most approaches. The q_{O-H} shows a good correlation ($R^2 > 0.5$ and $t > 1.66$ for many approaches) too, but this descriptor is only a combination of q_O and q_H, and both q_O and q_H are better descriptors than q_{O-H}. Therefore, it

dx.doi.org/10.1021/ci200133w |J. Chem. Inf. Model. XXXX, XXX, 000–000

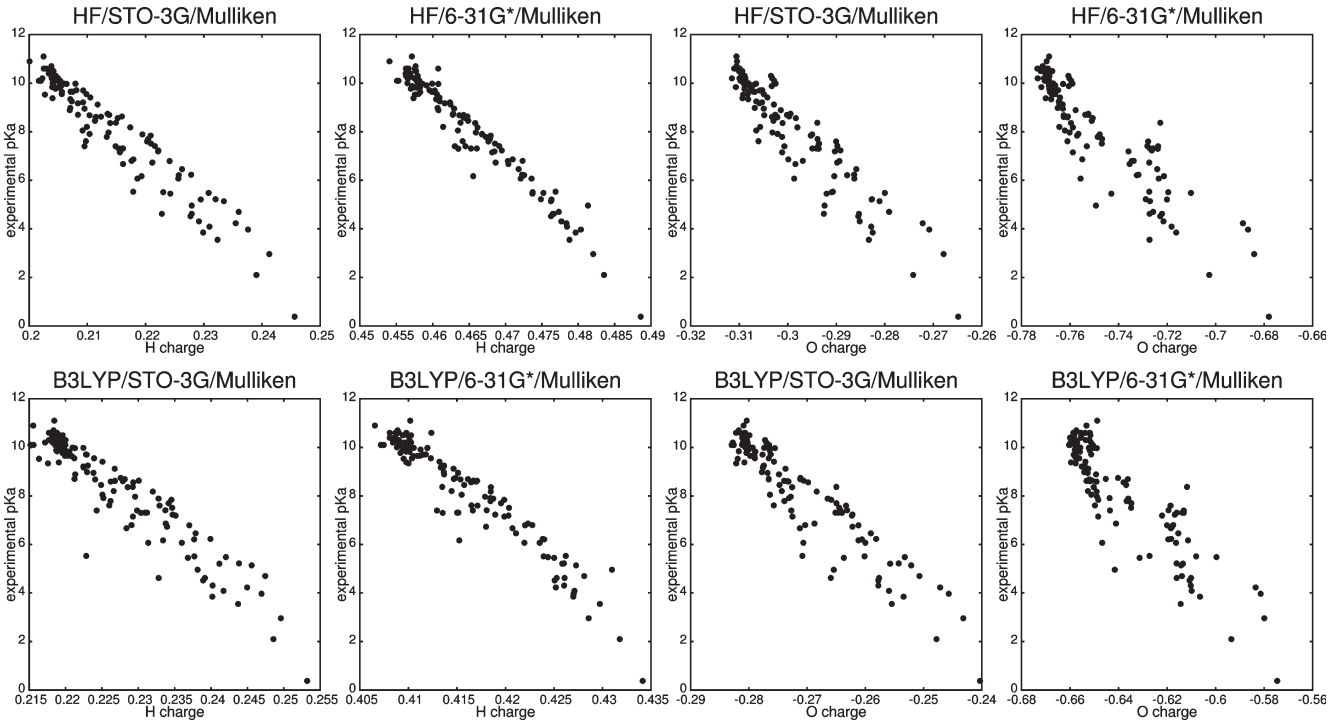


Figure 4. Correlations between q_H, q_O, and experimental pK_a for Mulliken PA and some selected basis sets and theory levels.

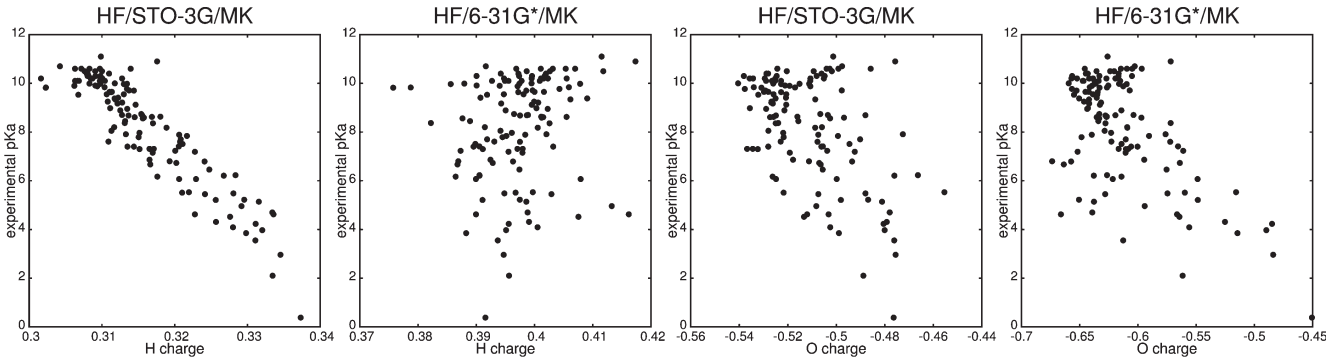


Figure 5. Correlations between q_H, q_O, and experimental pK_a for MK PA and some selected basis sets and theory levels.

makes no sense to introduce q_{O-H} into the models. Further descriptors show only a weak correlation (<0.5). For these reasons, the descriptors q_H, q_O, and q_{C1} were selected to create the QSPR models.

Figures 3−5 and Supporting Information (Table S1) also help us to recognize basic trends in the relevance of the charge calculation approaches for pK_a prediction. All seven theory levels seem applicable for pK_a prediction. Specifically, most of the charge calculation approaches utilizing these theory levels provide q_H or q_O correlating with pK_a with $R^2 > 0.8$ and statistically significant (at $p = 0.05$). In addition, all three basis sets and four out of five examined population analyses seem applicable for the same considerations. However, the MK PA demonstrates only weak correlation with pK_a and most approaches using MK provide q_H or q_O correlating with pK_a with $R^2 < 0.5$. Examples of correlation graphs for the Mulliken and MK population analyses are shown in Figures 4 and 5.

The trends for Mulliken, natural, Löwdin, and MK PA are in agreement with the study of Gross et al.,[22] who examined the correlation between q_O, q_H, and pK_a for different population analyses on the B3LYP/6-311G level of theory on a set of 19 substituted phenols. An interesting discovery of these analyses was also the fact that the correlation between pK_a and the atomic charge descriptors decreases linearly with the distance from the hydrogen atom of the phenolic OH group.

Parameterization and Validation of QSPR Models. The descriptors q_H, q_O, and q_{C1} were selected as inputs for the QSPR models, therefore the models have used the following equation for pK_a calculation:

$$pK_a = param_H \cdot q_H + param_O \cdot q_O + param_{C1} \cdot q_{C1} + constant \quad (2)$$

where $param_H$, $param_O$, $param_{C1}$, and constant are the parameters of the model. The parametrization of the QSPR models

dx.doi.org/10.1021/ci200133w |J. Chem. Inf. Model. XXXX, XXX, 000–000

Table 1. Quality Criteria and Statistical Criteria for All the QSPR Models[a]

model number	theory level	population analysis	basis set	R^2	RMSE	$\overline{\Delta}$	s	F	number of molecules
1	MP2	Mulliken	6-31G*	0.966	0.403	0.315	0.410	1136	124
2	HF	Mulliken	6-31G*	0.966	0.403	0.315	0.410	1136	124
3	MP2	Löwdin	6-31G*	0.966	0.405	0.315	0.412	1136	124
4	HF	Löwdin	6-31G*	0.966	0.406	0.315	0.413	1136	124
5	MP2	NPA	6-31G*	0.964	0.419	0.325	0.426	1071	124
6	B3LYP	Löwdin	6-31G*	0.963	0.421	0.325	0.428	1041	124
7	HF	NPA	6-31G*	0.963	0.421	0.329	0.428	1041	124
8	B3LYP	NPA	6-311G	0.962	0.428	0.336	0.435	1013	124
9	MP2	NPA	6-311G	0.961	0.432	0.344	0.439	986	124
10	B3LYP	NPA	6-31G*	0.961	0.433	0.332	0.440	986	124
11	HF	NPA	6-311G	0.961	0.434	0.346	0.441	986	124
12	BLYP	NPA	6-311G	0.96	0.437	0.341	0.444	960	124
13	BP86	NPA	6-311G	0.959	0.443	0.342	0.450	936	124
14	B3LYP	Mulliken	6-31G*	0.959	0.443	0.355	0.450	936	124
15	BLYP	NPA	6-31G*	0.959	0.444	0.34	0.451	936	124
16	BLYP	Löwdin	6-31G*	0.959	0.444	0.34	0.451	936	124
17	BP86	Löwdin	6-31G*	0.959	0.445	0.341	0.452	936	124
18	BP86	NPA	6-31G*	0.959	0.447	0.342	0.454	936	124
19	BP86	Mulliken	6-31G*	0.954	0.471	0.374	0.479	830	124
20	BLYP	Mulliken	6-31G*	0.953	0.477	0.377	0.485	811	124
21	MP2	Löwdin	6-311G	0.951	0.486	0.375	0.494	776	124
22	HF	Löwdin	6-311G	0.95	0.491	0.38	0.499	760	124
23	HF	Mulliken	6-311G	0.945	0.513	0.399	0.521	687	124
24	MP2	Mulliken	6-311G	0.945	0.514	0.401	0.522	687	124
25	BP86	Mulliken	6-311G	0.939	0.541	0.429	0.550	616	124
26	B3LYP	Mulliken	6-311G	0.938	0.547	0.433	0.556	605	124
27	B3LYP	Löwdin	6-311G	0.937	0.551	0.417	0.560	595	124
28	BLYP	Mulliken	6-311G	0.932	0.573	0.452	0.582	548	124
29	BP86	Löwdin	6-311G	0.931	0.577	0.443	0.587	540	124
30	MP2	Hirshfeld	STO-3G	0.929	0.594	0.467	0.605	488	116
31	HF	Hirshfeld	STO-3G	0.928	0.597	0.47	0.608	481	116
32	BLYP	Löwdin	6-311G	0.926	0.596	0.451	0.606	501	124
33	AM1	Mulliken	–	0.924	0.605	0.452	0.615	486	124
34	AM1	Löwdin	–	0.924	0.605	0.452	0.615	486	124
35	MP2	Mulliken	STO-3G	0.922	0.615	0.502	0.625	473	124
36	MP2	Löwdin	STO-3G	0.921	0.617	0.505	0.627	466	124
37	MP2	NPA	STO-3G	0.921	0.618	0.501	0.628	466	124
38	HF	Mulliken	STO-3G	0.92	0.619	0.508	0.629	460	124
39	HF	Löwdin	STO-3G	0.92	0.621	0.51	0.631	460	124
40	HF	NPA	STO-3G	0.92	0.622	0.506	0.632	460	124
41	AM1	Hirshfeld	–	0.917	0.631	0.499	0.641	442	124
42	MP2	Hirshfeld	6-31G*	0.912	0.652	0.529	0.663	415	124
43	MP2	Hirshfeld	6-311G	0.91	0.667	0.534	0.679	377	116
44	HF	Hirshfeld	6-31G*	0.908	0.665	0.538	0.676	395	124
45	HF	Hirshfeld	6-311G	0.907	0.678	0.541	0.690	364	116
46	B3LYP	Mulliken	STO-3G	0.904	0.68	0.558	0.691	377	124
47	B3LYP	Hirshfeld	STO-3G	0.902	0.698	0.536	0.710	344	116
48	B3LYP	Hirshfeld	6-31G*	0.897	0.705	0.546	0.717	348	124
49	BP86	Mulliken	STO-3G	0.896	0.707	0.575	0.719	345	124
50	BLYP	Mulliken	STO-3G	0.896	0.709	0.581	0.721	345	124
51	B3LYP	Löwdin	STO-3G	0.895	0.71	0.565	0.722	341	124
52	PM3	Hirshfeld	–	0.895	0.711	0.561	0.723	341	124
53	B3LYP	NPA	STO-3G	0.894	0.715	0.567	0.727	337	124
54	BP86	Hirshfeld	6-31G*	0.89	0.729	0.553	0.741	324	124

dx.doi.org/10.1021/ci200133w |J. Chem. Inf. Model. XXXX, XXX, 000–000

Table 1. Continued

model number	theory level	population analysis	basis set	R^2	RMSE	$\overline{\Delta}$	s	F	number of molecules
55	BLYP	Hirshfeld	6-31G*	0.886	0.741	0.567	0.753	311	124
56	BLYP	Hirshfeld	STO-3G	0.886	0.75	0.571	0.763	290	116
57	BP86	Hirshfeld	STO-3G	0.882	0.763	0.578	0.777	279	116
58	B3LYP	Hirshfeld	6-311G	0.882	0.764	0.599	0.778	279	116
59	PM3	Mulliken	–	0.88	0.76	0.581	0.773	293	124
60	PM3	Löwdin	–	0.88	0.76	0.581	0.773	293	124
61	BLYP	Löwdin	STO-3G	0.879	0.764	0.599	0.777	291	124
62	BP86	Löwdin	STO-3G	0.878	0.766	0.597	0.779	288	124
63	BLYP	NPA	STO-3G	0.877	0.769	0.604	0.782	285	124
64	BP86	NPA	STO-3G	0.876	0.772	0.601	0.785	283	124
65	BP86	Hirshfeld	6-311G	0.874	0.789	0.613	0.803	259	116
66	MP2	MK	STO-3G	0.869	0.795	0.634	0.808	265	124
67	BLYP	Hirshfeld	6-311G	0.868	0.807	0.627	0.821	245	116
68	HF	MK	STO-3G	0.867	0.8	0.641	0.813	261	124
69	BLYP	MK	6-311G	0.826	0.917	0.714	0.932	190	124
70	BP86	MK	6-311G	0.825	0.919	0.714	0.934	189	124
71	B3LYP	MK	6-311G	0.822	0.926	0.721	0.941	185	124
72	B3LYP	MK	STO-3G	0.817	0.939	0.749	0.955	179	124
73	BP86	MK	6-31G*	0.813	0.949	0.716	0.965	174	124
74	BLYP	MK	6-31G*	0.813	0.95	0.72	0.966	174	124
75	MP2	MK	6-311G	0.812	0.951	0.746	0.967	173	124
76	HF	MK	6-311G	0.811	0.954	0.747	0.970	172	124
77	B3LYP	MK	6-31G*	0.808	0.962	0.728	0.978	168	124
78	MP2	MK	6-31G*	0.788	1.011	0.773	1.028	149	124
79	HF	MK	6-31G*	0.788	1.012	0.773	1.029	149	124
80	BP86	MK	STO-3G	0.787	1.014	0.8	1.031	148	124
81	BLYP	MK	STO-3G	0.786	1.016	0.799	1.033	147	124
82	AM1	MK	–	0.447	1.633	1.247	1.660	32	124
83	PM3	MK	–	0.445	1.636	1.249	1.663	32	124

[a] The models are sorted according their R^2 (descending) and afterwards according their RMSE (ascending) and $\overline{\Delta}$ (ascending).

was performed for all 83 charge calculation approaches via MLR. The complete data set of 124 phenol molecules was used for the parametrization, and the obtained models were validated for all molecules in the data set. The only exceptions were the charge calculation procedures containing the Hirshfeld PA and basis sets 6-311G and STO-3G. In these cases, only 116 phenols were used for the parametrization and evaluation of the models, because 8 molecules from the original set were strong outliers. Table 1 contains the quality criteria (R^2, RMSE, and $\overline{\Delta}$) and the statistical criteria (s and F) for all the models. The models are sorted according to their quality. The parameters of the models are summarized in the Supporting Information (Table S2). The most relevant graphs of correlation between experimental and calculated pK_a are visualized in Figure 6. Tables 2−4 provide a clue for the comparison of QSPR models. Table 2 summarizes the R^2 of all models, Table 3 contains the average values of R^2 for all QSPR models which use a specific theory level, basis set, or PA, and Table 4 summarizes the quality of the charge calculation approaches which use a specific theory level, basis set or PA, and whose R^2 are in a certain interval.

The results provided in Tables 1−4 and Figure 6 lead us to the following conclusions regarding the relevance of the charge calculation method to the ability of the QSPR model to predict pK_a for phenolic compounds.

Comparison of All Models. All the presented models are statistically significant at $p = 0.01$. For the sets of 124 or 116 molecules, the models with three descriptors are statistically significant (at $p = 0.01$) when $F > 3.949$. The best models are MP2/6-31G*/Mulliken and HF/6-31G*/Mulliken ($R^2 = 0.966$, RMSE = 0.403, $\overline{\Delta} = 0.315$, $s = 0.410$, and $F = 1136$). More than 25% of the analyzed models (22 out of 83) have excellent quality and statistical criteria ($R^2 \geq 0.95$, RMSE ≤ 0.491, $\overline{\Delta} \leq 0.38$, $s \leq 0.5$, and $F \geq 760$), and more than 50% (47 out of 83) have very good statistical criteria ($R^2 > 0.9$, RMSE ≤ 0.698, $\overline{\Delta} \leq 0.54$, $s \leq 0.71$, and $F \geq 344$). About 80% of the models are able to predict pK_a with acceptable quality ($R^2 > 0.85$, RMSE ≤ 0.8, $\overline{\Delta} \leq 0.641$, $s \leq 0.813$, and $F \geq 261$). Only less than 20% of the models show a week correlation.

Influence of Theory Level. Ab initio theory levels: All five examined ab initio theory levels (MP2, HF, B3LYP, BLYP, and BP86) are applicable to pK_a prediction using charges. The best QSPR models are provided by MP2 and HF (models 1 and 2). Surprisingly, the differences between MP2 and HF were very small (illustrated by Tables 2−4). The pK_a values calculated from DFT charges have a slightly weaker correlation with experimental pK_a compared to MP2 and HF. The best performing DFT functional has been B3LYP, the models created by BLYP and BP86 have been less accurate.

dx.doi.org/10.1021/ci200133w |J. Chem. Inf. Model. XXXX, XXX, 000–000

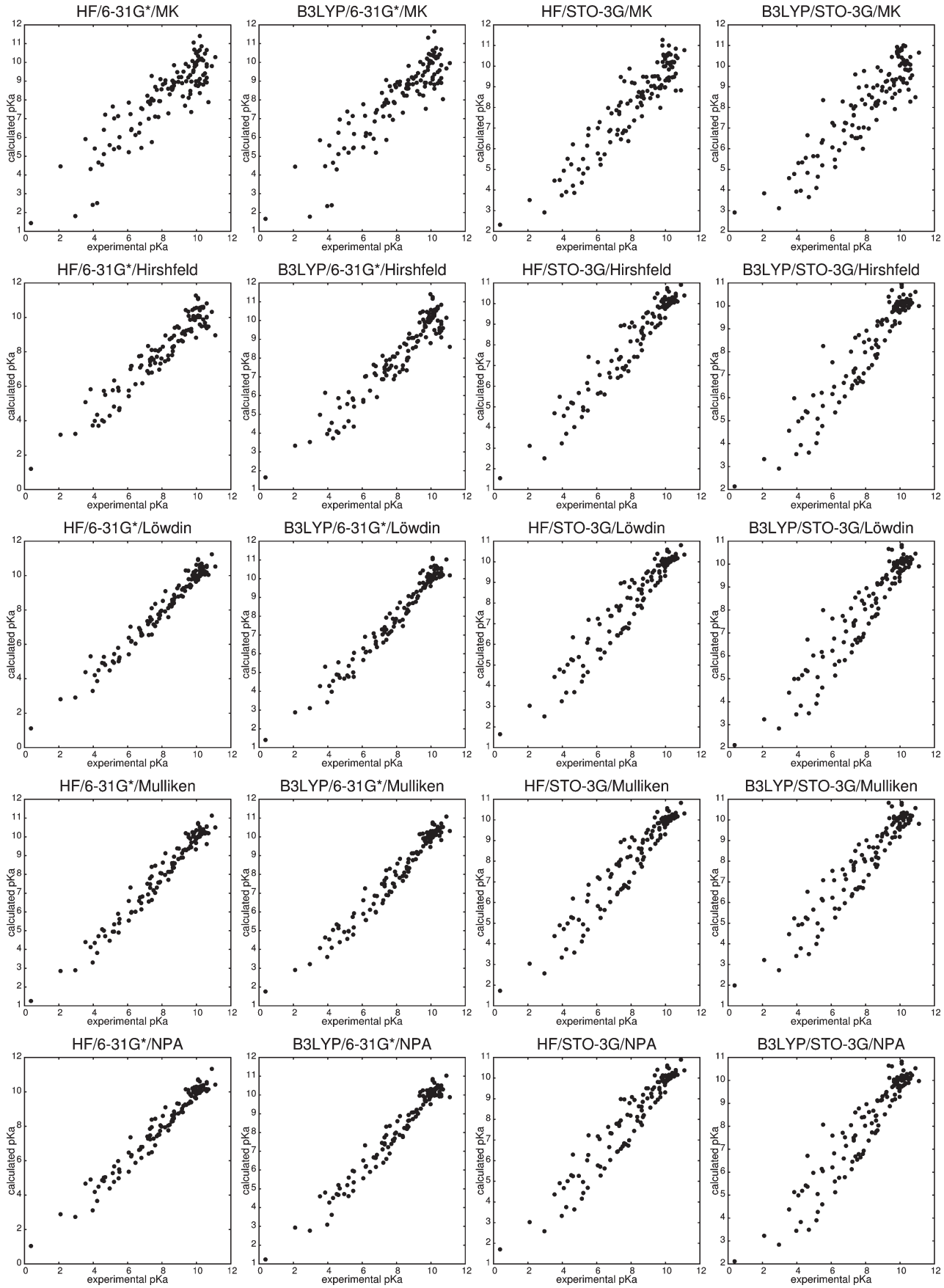


Figure 6. Graphs showing the correlation between experimental and calculated p*Ka* for some selected charge calculation procedures.

dx.doi.org/10.1021/ci200133w |*J. Chem. Inf. Model.* XXXX, XXX, 000–000

Table 2. Squared Pearson Coefficients between Calculated and Experimental pK_a

R^2 for the basis set 6-31G*

Theory level	Population analysis				
	MK	Hir.	Löw.	MPA	NPA
BLYP	0.813	0.886	0.959	0.953	0.959
BP86	0.813	0.890	0.959	0.954	0.959
B3LYP	0.808	0.897	0.963	0.959	0.961
HF	0.788	0.908	0.966	0.966	0.963
MP2	0.788	0.912	0.966	0.966	0.964

R^2 for the basis set 6-311G

Theory level	Population analysis				
	MK	Hir.	Löw.	MPA	NPA
BLYP	0.826	0.868	0.926	0.932	0.960
BP86	0.825	0.874	0.931	0.939	0.959
B3LYP	0.822	0.882	0.937	0.938	0.962
HF	0.811	0.907	0.950	0.945	0.961
MP2	0.812	0.910	0.951	0.945	0.961

R^2 for the basis set STO-3G

Theory level	Population analysis				
	MK	Hir.	Löw.	MPA	NPA
BLYP	0.786	0.886	0.879	0.896	0.877
BP86	0.787	0.882	0.878	0.896	0.876
B3LYP	0.817	0.902	0.895	0.904	0.894
HF	0.867	0.928	0.920	0.920	0.920
MP2	0.869	0.929	0.921	0.922	0.921

Legend

	R^2	$RMSE$	$\bar{\Delta}$
excellent	0.950–0.970	0.40–0.50	0.32–0.38
very good	0.920–0.950	0.50–0.63	0.38–0.51
good	0.900–0.920	0.63–0.70	0.51–0.54
acceptable	0.850–0.900	0.70–0.80	0.54–0.64
weak	0.800–0.850	0.80–0.97	0.71–0.73

R^2 for semiempirical methods

Theory level	Population analysis				
	MK	Hir.	Löw.	MPA	NPA
AM1	0.447	0.917	0.924	0.924	–
PM3	0.445	0.895	0.880	0.880	–

Table 3. Average Squared Pearson Coefficients for All Charge Calculation Approaches Which Use a Specific Theory Level, PA, or Basis Set

theory level	BLYP	BP86	B3LYP	HF	MP2	AM1	PM3
average R^2	0.894	0.895	0.903	0.915	0.916	0.803	0.775

population analysis	MK	Hirshfeld	Löwdin	Mulliken	NPA
average R^2	0.772	0.898	0.930	0.932	0.940

basis set	6-31G*	6-311G	STO-3G
average R^2	0.917	0.909	0.887

Semiempirical theory levels: The semiempirical approaches tested here provide weaker correlation than ab initio methods but are still applicable for pK_a prediction. The models with AM1 and Mulliken, Hirshfeld, or Löwdin PA show good correlation ($R^2 \geq 0.917$). The models using PM3 and Mulliken, Hirshfeld, or Löwdin PA also demonstrate acceptable correlation ($R^2 \geq 0.88$). The combination of semiempirical approaches with MK PA gives the worst models in this study (models 82 and 83).

Influence of Basis Set. The charges most appropriate for QSPR modeling of pK_a are provided by the 6-31G* basis set, and the accuracy of these models is very high. For example, the model with HF/6-31G*/Mulliken charges shows $R^2 = 0.966$. The results for the 6-311G basis set are slightly weaker. Unexpectedly, also the charges obtained using the STO-3G basis set are suitable for QSPR modeling, and the quality of these models is acceptable. For example, the model employing MP2/STO-3G/Hirshfeld charges exhibits $R^2 = 0.929$.

Influence of Population Analysis. Mulliken, natural, and Löwdin PAs with all levels of theory and basis sets provide the charges that are appropriate for pK_a prediction. The Hirshfeld PA with the STO-3G basis set provides results similar to the Mulliken, natural, or Löwdin PA with the same basis set.

Table 4. Percentage of Charge Calculation Approaches Which Use a Specific Theory Level, PA, or Basis Set and Whose Squared Pearson Coefficients Are in a Certain Interval[a]

theory level	interval	BLYP	BP86	B3LYP	HF	MP2	AM1	PM3
	$R^2 \geq 0.95$	27	27	29	33	33	0	0
	$0.95 > R^2 \geq 0.9$	13	13	29	47	47	75	0
	$0.9 > R^2 \geq 0.85$	40	40	21	7	7	0	75
	$R^2 < 0.85$	20	20	21	13	13	25	25

population analysis	interval	MK	Hirshfeld	Löwdin	Mulliken	NPA
	$R^2 \geq 0.95$	0	0	41	29	67
	$0.95 > R^2 \geq 0.9$	0	47	35	53	13
	$0.9 > R^2 \geq 0.85$	12	53	24	18	20
	$R^2 < 0.85$	88	0	0	0	0

basis set	interval	6-31G*	6-311G	STO-3G
	$R^2 \geq 0.95$	60	28	0
	$0.95 > R^2 \geq 0.9$	12	40	40
	$0.9 > R^2 \geq 0.85$	8	12	48
	$R^2 < 0.85$	20	20	12

[a] The percentages are calculated from total number of approaches with the defined theory level, basis set, or PA.

Nevertheless, the Hirshfeld PA with the 6-31G* or 6-311G basis sets lead to less accurate models than the above-mentioned population analyses employing these basis sets. Moreover, the occurrence of strong outliers complicates the application of the Hirshfeld PA. The charges calculated by the MK PA show only weak correlation with pK_a, and the QSPR models based on these charges have low accuracy, i.e., the best of such QSPR models employs HF/STO-3G/MK charges and shows $R^2 = 0.867$.

dx.doi.org/10.1021/ci200133w |J. Chem. Inf. Model. XXXX, XXX, 000–000

Table 5. Comparison of the Presented QSPR Models with Previous Work

theory level	PA	basis set	descriptors	R^2	s	F	number of molecules	source
B3LYP	NPA	6-311G**	q_{O-H}	0.789	1.300	48	15	Kreye and Seybold,[23, a]
B3LYP	NPA	6-311G**	q_O	0.731	1.500	38	15	Kreye and Seybold,[23, a]
B3LYP	NPA	6-31+G*	q_{O-H}	0.880	0.970	95	15	Kreye and Seybold,[23, b]
B3LYP	NPA	6-31+G*	q_O	0.865	1.000	38	15	Kreye and Seybold,[23, b]
B3LYP	NPA	6-311G(d,p)	q_{O^-}	0.911	0.252	173	19	Gross and Seybold[14]
B3LYP	NPA	6-311G(d,p)	q_H	0.887	0.283	134	19	Gross and Seybold[14]
B3LYP	NPA	6-31G*	q_H, q_O, q_{C1}	0.961	0.440	986	124	this work, model 10
B3LYP	NPA	6-311G	q_H, q_O, q_{C1}	0.962	0.435	1013	124	this work, model 8
B3LYP	MPA	6-311G(d,p)	q_H	0.913	0.248	179	19	Gross and Seybold[14]
B3LYP	MPA	6-311G(d,p)	q_{O^-}	0.894	0.274	144	19	Gross and Seybold[14]
B3LYP	MPA	6-311G	q_H, q_O, q_{C1}	0.938	0.556	605	124	this work, model 26
B3LYP	MPA	6-31G*	q_H, q_O, q_{C1}	0.959	0.450	936	124	this work, model 14
B3LYP	MK	6-311G(d,p)	q_H	0.344	0.682	9	19	Gross and Seybold[14]
B3LYP	MK	6-311G(d,p)	q_{O^-}	0.692	0.467	38	19	Gross and Seybold[14]
B3LYP	MK	6-311G	q_H, q_O, q_{C1}	0.822	0.941	185	124	this work, model 71
B3LYP	MK	6-31G*	q_H, q_O, q_{C1}	0.808	0.978	168	124	this work, model 77

[a] With solvent model SM5.4. [b] With solvent model SM8.

Table 6. Comparison of R^2 and RMSE for Test, Training, and Complete Sets for Model 2 (employing HF, Mulliken, 6-31G*) Charge Calculation Approaches

complete set				
R^2	RMSE	s	F	number of molecules
0.966	0.403	0.410	1136	124

cross validation										
	training set					test set				
cross-validation step	R^2	RMSE	s	F	number of molecules	R^2	RMSE	s	F	number of molecules
1	0.965	0.405	0.413	873	99	0.973	0.405	0.442	252	25
2	0.970	0.382	0.390	1024	99	0.930	0.489	0.534	93	25
3	0.964	0.415	0.424	848	99	0.977	0.357	0.390	297	25
4	0.967	0.394	0.402	928	99	0.966	0.444	0.484	199	25
5	0.968	0.403	0.411	968	100	0.957	0.442	0.484	148	24

Table 7. Comparison of R^2 and RMSE for Test, Training, and Complete Sets for Model 14 (employing B3LYP, Mulliken, 6-31G*) Charge Calculation Approaches

complete set				
R^2	RMSE	s	F	number of molecules
0.959	0.443	0.450	936	124

cross validation										
	training set					test set				
cross-validation step	R^2	RMSE	s	F	number of molecules	R^2	RMSE	s	F	number of molecules
1	0.958	0.441	0.450	722	99	0.963	0.452	0.493	182	25
2	0.962	0.434	0.443	802	99	0.925	0.509	0.555	86	25
3	0.955	0.462	0.472	672	99	0.975	0.358	0.391	273	25
4	0.961	0.425	0.434	780	99	0.956	0.516	0.563	152	25
5	0.962	0.435	0.444	810	100	0.950	0.506	0.554	127	24

dx.doi.org/10.1021/ci200133w |J. Chem. Inf. Model. XXXX, XXX, 000–000

Comparison with Previous Work. QSPR models similar to those presented in this paper were previously published by Gross and Seybold[14] and also by Kreye and Seybold.[23] Table 5 shows a comparison of these models with our models. It is seen therein that our models show markedly higher R^2 and F values, even for simpler basis sets. The reason may be that they employ more descriptors and were parametrized within a larger training set.

Cross-Validation. The robustness of the models was tested by cross-validation. The set of phenol molecules was divided into five parts (each contained 20% of the molecules). Afterward, five cross-validation steps were performed. In the first step, the first part was selected as a test set, and the remaining four parts were taken together as the training set. The test and training sets for the other steps were prepared in a similar manner by subsequently considering one part as a test set and the remaining parts served as a training set. For each step, the QSPR model was parametrized on the training set. Afterward, the pK_a values of the respective test molecules were calculated via this model and compared with experimental pK_a values. The cross-validation was performed for all 83 analyzed charge calculation approaches. The results are summarized in the Supporting Information (Table S3), and a part of these results is shown in Tables 6 and 7. The cross-validation showed that the models are stable, and the values of R^2 and RMSE are similar for the test, training, and complete sets.

CONCLUSION

The quantum chemical partial atomic charges have been shown to provide very good QSPR models for the estimation of pK_a. More than 25% of the analyzed models (22 out of 83) have excellent quality and statistical criteria (e.g., $R^2 \geq 0.95$), and more than 50% (47 out of 83) have very good statistical criteria (e.g., $R^2 > 0.9$). The descriptors used in the models we developed are the atomic charges of the hydrogen and oxygen from the phenolic OH group and the charge of the carbon binding to the OH group. Other atomic charges show only a weak correlation with pK_a. All seven examined theory levels (MP2, HF, B3LYP, BLYP, BP86, AM1, and PM3) are applicable to predicting pK_a from charges. The best results have been obtained for MP2 and HF. Utilizing DFT also provides good correlation. Semiempirical methods have generated weaker but acceptable models. The most suitable basis set was 6-31G*, while 6-311G provided slightly weaker correlations, and unexpectedly also the STO-3G basis set proved applicable to the QSPR modeling of pK_a. The Mulliken, natural, and Löwdin population analyses provided accurate models for all tested theory levels and basis sets. The Hirshfeld PA has been also useful, but the QSPR models based on MK charges showed only weak correlations. It is thus clear from our study that it is possible to predict pK_a values with very good accuracy using only partial atomic charges, and even unsophisticated theory levels and basis sets can provide good descriptors.

ASSOCIATED CONTENT

Supporting Information. List of phenol molecules employed in this study, including their experimental pK_a, the table of R^2 values for all charge calculation procedures and all descriptors (Table S1), the table with parameters of all QSPR models (Table S2) and the table of cross-validation results (Table S3). This material is available free of charge via the Internet at http://pubs.acs.org/.

AUTHOR INFORMATION

Corresponding Author
*E-mail: jkoca@chemi.muni.cz.

REFERENCES

(1) Wan, H.; Ulander, J. High-throughput pK_a screening and prediction amenable for ADME profiling. *Expert Opin. Drug Metab. Toxicol.* **2006**, *2*, 139–155.
(2) Clark, J.; Perrin, D. D. Prediction of the strengths of organic bases. *Q. Rev., Chem. Soc.* **1964**, *18*, 295–320.
(3) Perrin, D. D.; Dempsey, B.; Serjeant, E. P. *pKa Prediction for Organic Acids and Bases*; Chapman and Hall: New York, 1981.
(4) *ACD/pK_a*; Advanced Chemistry Development, Inc.: Toronto, Ontario, Canada, 2009.
(5) Shelley, J. C.; Cholleti, A.; Frye, L. L.; Greenwood, J. R.; Timlin, M. R.; Uchimaya, M.; Epik, M. A software program for pK_a prediction and protonation state generation for druglike molecules. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 681–691.
(6) Hilal, S. H.; Karickhoff, S. W. A rigorous test for SPARC's chemical reactiity models: Estimation of more than 4300 ionization pK_as. *Quant. Struct.-Act. Relat.* **1995**, *14*, 348–355.
(7) Sayle, R. *Physiological ionization and pK_a prediction*; Metaphorics LLC: Santa Fe, NM, 2000; http://www.daylight.com/meetings/emug00/Sayle/pkapredict.html. Accessed January 24, 2011.
(8) Lee, A. C.; Crippen, G. M. Predicting pK_a. *J. Chem. Inf. Model.* **2009**, *49*, 2013–2033.
(9) *Jaguar*, version 4.2; Schrödinger, Inc.: New York, 2010.
(10) Liptak, M. D.; Gross, K. C.; Seybold, P. G.; Feldgus, S.; Shields, G. Absolute pK_a determinations for substituted phenols. *J. Am. Chem. Soc.* **2002**, *124*, 6421–6427.
(11) Barone, V.; Cossi, M. Quantum calculaton of molecular energies and energy gradients in solution by a conductor solvent model. *J. Phys. Chem. A* **1988**, *102*, 1995–2001.
(12) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A. *Gaussian 98*, revision A.6; Gaussian, Inc.: Pittsburgh, PA, 1998.
(13) Habibi-Yangjeh, A. Application of artificial neural networks for predicting the aqueous acidity of various phenols using QSAR. *J. Mol. Model.* **2006**, *12*, 338–347.
(14) Gross, K. C.; Seybold, P. G. Substituent effects on the physical properties and pK_a of phenol. *Int. J. Quantum Chem.* **2001**, *85*, 569–579.
(15) Brinck, T.; Murray, J. S.; Politzer, P. Molecular surface electrostatic potentials and local ionization energies of group V - VII hydrides and their anions: Relationships for aqueous and gas-phase acidities. *Int. J. Quantum Chem.* **1993**, *48*, 73–88.
(16) Citra, M. J. Estimating the pK_a of phenols, carboxylic acids and alcohols from semi-empirical quantum chemical methods. *Chemosphere* **1999**, *1*, 191–206.
(17) Dixon, S. L.; Jurs, P. C. Estimation of pK_a for organic oxyacids using calculated atomic charges. *J. Comput. Chem.* **1993**, *14*, 1460–1467.
(18) Parthasarathi, R.; Padmanabhan, J.; Elango, M.; Chitra, K.; Subramanian, V.; Chattaraj, P. K. pK_a prediction using group philicity. *J. Phys. Chem. A* **2006**, *110*, 6540–6544.

(19) Liu, S.; Pedersen, L. G. Estimation of molecular acidity via electrostatic potential at the nucleus and valence natural atomic orbitals. *J. Phys. Chem. A* **2009**, *113*, 3648–3655.
(20) Jelfs, S.; Ertl, P.; Selzer, P. Estimation of pK_a for druglike compounds using semiempirical and information-based descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 450–459.
(21) Zhang, J.; Kleinöder, T.; Gasteiger, J. Prediction of pK_a values for aliphatic carboxylic acids and alcohols with empirical atomic charge descriptors. *J. Chem. Inf. Model.* **2006**, *46*, 2256–2266.
(22) Gross, K. C.; Seybold, P. G.; Hadad, C. M. Comparison of different atomic charge schemes for predicting pK_a variations in substituted anilines and phenols. *Int. J. Quantum Chem.* **2002**, *90*, 445–458.
(23) Kreye, W. C.; Seybold, P. G. Correlations between quantum chemical indices and the pK_as of a diverse set of organic phenols. *Int. J. Quantum Chem.* **2009**, *109*, 3679–3684.
(24) Xing, L.; Glen, R. C. Novel methods for the prediction of logP, pK_a, and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796–805.
(25) Soriano, E.; Cerdán, S.; Ballesteros, P. Computational determination of pK_a values. A comparison of different theoretical approaches and a novel procedure. *J. Mol. Struct.: THEOCHEM* **2004**, *684*, 121–128.
(26) Svobodová Vařeková, R.; Koča, J. Optimized and parallelized implementation of the electronegativity equalization method and the atom-bond electronegativity equalization method. *J. Comput. Chem.* **2006**, *3*, 396–405.
(27) *NCI Open Database Compounds*; National Cancer Institute, National Institutes of Health: Bethesda, MD; http://cactus.nci.nih.gov/. Accessed August 10, 2010.
(28) Sadowski, J.; Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567–2581.
(29) Gieleciak, R.; Polanski, J. Modeling Robust QSAR. 2. Iterative Variable Elimination Schemes for CoMSA: Application for Modeling Benzoic Acid pK_a Values. *J. Chem. Inf. Model.* **2007**, *47*, 547–556.
(30) Howard, P.; Meylan, W. *Physical/chemical property database (PHYSPROP)*. Syracuse Research Corporation, Environmental Science Center: North Syracuse, NY, 1999.
(31) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J. A., Jr.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. *Gaussian 03*, revision C.02; Gaussian, Inc.: Wallingford, CT, 2004.

Electronegativity Equalization Method: parameterization and validation for large sets of organic, organohalogene and organometal molecule

Int. J. Mol. Sci. **2007**, *8*, 572-582

International Journal of
Molecular Sciences
ISSN 1422-0067

www.mdpi.org/ijms/

Full Research Paper

Electronegativity Equalization Method: Parameterization and Validation for Large Sets of Organic, Organohalogene and Organometal Molecule

Radka Svobodová Vařeková [1], Zuzana Jiroušková [1], Jakub Vaněk [1], Šimon Suchomel [1] and Jaroslav Koča [1,*]

1 National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic; Tel.: +420 549 494 947, Fax: +420 549 492 556, E-mail: jkoca@chemi.muni.cz

* Author to whom correspondence should be addressed; E-mail: jkoca@chemi.muni.cz

Received: 10 May 2007; in Revised Form: 6 June 2007 / Accepted: 14 June 2007 / Published: 3 July 2007

Abstract: The Electronegativity Equalization Method (EEM) is a fast approach for charge calculation. A challenging part of the EEM is the parameterization, which is performed using *ab initio* charges obtained for a set of molecules. The goal of our work was to perform the EEM parameterization for selected sets of organic, organohalogen and organometal molecules. We have performed the most robust parameterization published so far. The EEM parameterization was based on 12 training sets selected from a database of predicted 3D structures (NCI DIS) and from a database of crystallographic structures (CSD). Each set contained from 2000 to 6000 molecules. We have shown that the number of molecules in the training set is very important for quality of the parameters. We have improved EEM parameters (STO-3G MPA charges) for elements that were already parameterized, specifically: C, O, N, H, S, F and Cl. The new parameters provide more accurate charges than those published previously. We have also developed new parameters for elements that were not parameterized yet, specifically for Br, I, Fe and Zn. We have also performed crossover validation of all obtained parameters using all training sets that included relevant elements and confirmed that calculated parameters provide accurate charges.

Keywords: Charge distribution, Electronegativity Equalization Method, Parameterization, Organohalogenes, Organometals.

---

1. Introduction

Electronegativity Equalization Method (EEM) [1,2,3] is a fast approach for charge calculation. The basic idea is based on the density functional theory (DFT) [4,5]. First, Parr et al. applied the DFT and formulated a new definition and explanation of electronegativity [6,7]. Later on, Mortier et al. applied Parr's definition of electronegativity and Sanderson's Electronegativity Equalization Principle (EEP) [8,9,10] and created the EEM.

This method is able to calculate atomic charges markedly faster than common *ab initio* approaches. The *ab initio* charge calculations exhibit time complexity of $O(B^4)$, where B is greater or equal to the number of valence electrons. The EEM approach shows a time complexity of $\theta(N^3)$, where N is the number of atoms. Nevertheless, accuracy of the EEM corresponds to the *ab initio* methods.

A challenging part of the EEM is the parameterization that is performed using *ab initio* charges obtained for a set of molecules. The parameterization is very time-consuming with time complexity of $O(S.B^4)$, where S is a number of molecules in the set. The most common parameterization of the EEM is a parameterization for the HF method with the STO-3G basis set, where the charges are calculated by Mulliken population analysis (MPA) [11,12]. Principally, it is also possible to parameterize the EEM for other basis sets (i.e., 6-31G*) and methods for charge calculation (i.e., CHELPG, MK, NPA, ESP, Hirshfeld method) [13,14]. First attempts to calculate EEM parameters were published in eighties [1,2]. These publications contained only parameters for C, H, N and O, which were developed using training sets of about one hundred molecules. Further parameterizations were performed during the nineties and contained parameters for new elements (S, Si, P, F, Cl) and more complex bases [15,16,17]. The EEM parameterization still remains attractive to chemists' attention [18,19,20].

The goal of this work is to perform the EEM parameterization based on large sets of organic, organohalogen and organometal molecules (containing Zn and Fe) selected from databases NCI DIS [21] and CSD [22], and to validate the quality of calculated parameters on reference sets of molecules selected from these databases. The parameterization was performed for STO-3G MPA charges.

2. Theoretical basis

2.1. EEM

Using DFT, the effective (charge-dependent) electronegativity of the atom i in a molecule can be calculated by eq. (1) [1,2 3]:

$$\chi_i = A_i + B_i.q_i + \kappa. \sum_{j=1(j\neq i)}^{N} \frac{q_j}{R_{i,j}} \qquad \textbf{(1)}$$

where N is the number of atoms in the molecule, q_i and q_j are the charges distributed on the atoms i and j, respectively, $R_{i,j}$ is the distance between atoms i and j, and κ is the adjusting factor. The coefficients A_i and B_i are defined by eqs**.** (2):

$$\begin{aligned} A_i &= \chi_i^* = \chi_i^0 + \Delta\chi_i \\ B_i &= 2\eta_i^* = 2\left(\eta_i^0 + \Delta\eta_i\right) \end{aligned} \qquad \textbf{(2)}$$

where χ_i^0 is the electronegativity of an isolated neutral atom *i*, η_i^0 is the hardness, and $\Delta\chi_i^0$ and $\Delta\eta_i$ describe the molecular environment. The coefficients A_i, B_i and κ are empirical parameters, which must be obtained via EEM parameterization. Such a parameterization is a topic of this work.

According to Sanderson's Electronegativity Equalization Principle [8, 9, 10], the effective electronegativity of each atom in the molecule is equal to the molecular electronegativity $\overline{\chi}$:

$$\chi_1 = \chi_2 = ... = \chi_N = \overline{\chi} \quad \textbf{(3)}$$

The total charge *Q* of the molecule is equal to the sum of all the atomic charges:

$$\sum_{i=1}^{N} q_i = Q \quad \textbf{(4)}$$

The atomic charges are described using the equation system (5), which contains *N*+1 equations with *N*+1 unknowns: q_1, q_2, … , q_N and $\overline{\chi}$. This system was derived from equations (1), (3) and (4) [1]:

$$\begin{pmatrix} B_1 & \frac{\kappa}{R_{1,2}} & \cdots & \frac{\kappa}{R_{1,N}} & -1 \\ \frac{\kappa}{R_{2,1}} & B_2 & \cdots & \frac{\kappa}{R_{2,N}} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\kappa}{R_{N,1}} & \frac{\kappa}{R_{N,2}} & \cdots & B_N & -1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ \overline{\chi} \end{pmatrix} = \begin{pmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_N \\ Q \end{pmatrix} \quad \textbf{(5)}$$

The matrix of the equation system (5) is called EEM matrix.

2.2. EEM Parameterization

Empirical parameters A_i, B_i and κ (described by eqs. (1) and (2)) can be calculated in the following way [17]:

From eq. (1) and (3), eq. (6) can be derived:

$$\overline{\chi} = A_i + B_i q_i + \kappa \sum_{j=1(j\neq i)}^{N} \frac{q_j}{R_{i,j}} \quad \textbf{(6)}$$

Eq. (6) can be rewritten as:

$$A_i + B_i q_i = \overline{\chi} - \kappa \sum_{j=1(j\neq i)}^{N} \frac{q_j}{R_{i,j}} \quad \textbf{(7)}$$

Meaning that eq. (7) is in the form:

$$A_i + B_i x_i = y_i \quad \textbf{(8)}$$

where:

$$x_i = q_i, \qquad y_i = \overline{\chi} - \kappa \sum_{j=1(j\neq i)}^{N} \frac{q_j}{R_{i,j}}$$

Then, empirical parameters can be obtained using eq. (8) in the following way:

1. Selection of a set of molecules used for the EEM parameterization.
2. *Ab initio* calculation of atomic charges q_i for all atoms within all selected molecules.
3. Calculation of the molecular electronegativity $\overline{\chi}$ as a harmonic average of atomic electronegativities χ_i^0 (for isolated atoms *i*):

$$\overline{\chi} = N\left(\sum_{i=1}^{N} \frac{1}{\chi_i^0}\right)^{-1} \quad \textbf{(9)}$$

4. Selection of κ values for which the parameterization will be performed.

5. For each of the above selected κ:
    - Calculation of x_i and y_i values for all atoms in all molecules using eq. (8).
    - Separation of x_i and y_i couples into subsets according to the chemical symbol and hybridization of the atom i (for example C in sp^3, C in sp^2 etc.).
    - Calculation of parameters A_i and B_i for each of these subsets using the least square minimization.
6. Finding the optimal κ value.

3. Methods

In this work, two databases were used. The first one was the NCI DIS 3D database [21], created as a part of DTP NCI (Developmental Therapeutics Program of National Cancer Institute). This database contains organic molecules tested against cancer, specifically their topologies and also geometries, predicted by the program CHEM-X [23] and stored in SDF format [24]. The second database used was CSD (Cambridge Structural Database) [22], administered by CCDC (Cambridge Crystallographic Data Centre). Geometries of molecules are stored also in SDF format. However, in this case information is obtained experimentally using the X-ray and/or neutron diffraction. Both these databases are sufficiently large, containing more than two hundred thousand molecules.

Table 1: Sets of molecules that were used as training and testing sets for the EEM parameterization.

Database	Denotation of the set	Number of molecules	Atoms included	Position of the set in the database
NCI DIS (predicted data)	n_{beg}	2000	C, O, N, H, S	beginning (ID between 1 and 3162)
	n_{mid}	2000	C, O, N, H, S	middle (ID between 300 000 and 314 026)
	n_{end}	2000	C, O, N, H, S	end (ID between 705 000 and 712 703)
	n_{all}	6000	C, O, N, H, S	n_{beg}, n_{mid} and n_{end}
	n_{hal}	4000	C, O, N, H, S, Br, Cl, F, I	beginning (ID between 106498 and 114688)
CSD (crystallo-graphic data)	c_{beg}	2000	C, O, N, H, S	beginning (ID starting by A and B)
	c_{mid}	2000	C, O, N, H, S	middle (ID starting by J, K and L)
	c_{end}	2000	C, O, N, H, S	end (ID starting by W and Y)
	c_{all}	6000	C, O, N, H, S	c_{beg}, c_{mid} and c_{end}
	c_{hal}	4000	C, O, N, H, S, F, Cl, Br, I	beginning (ID starting by A, B and C)
	c_{met}	2000	C, O, N, H, S, Fe, Zn	beginning (ID starting by A and B)
	$c_{h,m}$	6000	C, O, N, H, S, F, Cl, Br, I, Fe, Zn	c_{hal} and c_{met}

ID is a unique identification of a molecule in a database. In the NSC DIS database, ID is a number between 1 and about 720 000. Database CSD uses alphabetically sorted string IDs that contains six upper case characters.

From these two databases, several training sets of molecules were selected (see Table 1). Our goal was to generate training sets, which cover most of bonding situations and also conformational variability in real molecules. For that reason, we have chosen large sets containing randomly selected molecules. As molecules are unsorted in NCI DIS and CSD databases, the simplest random selection is to take a continuous part of the database. We selected three training sets, containing elements C, H, O, N and S from each database. To obtain the most versatile training sets, we selected first training set from the beginning, second from the middle and the third from the end of the databases. We have also used unions of these sets. For organohalogenes and organometals, we did not need so many training sets as the process of parameterization was already debugged on above mentioned six training sets and their unions. Therefore we used only one training set of organohalogenes from each database. Just one training set (from CSD database) was used for organometals, because the NCI DIS database does not contain enough organometal molecules. These organometal and organohalogene molecules were selected from the beginning of the databases.

For the parameterization, *ab initio* charges were calculated using the HF method with the STO-3G basis set for all molecules in all sets. The charge calculation was performed by Gaussian 98 program [25]. After that, the EEM parameterization was performed using calculated *ab initio* charges for all training sets.

Table 2: Quality of parameters that were obtained by EEM parameterization using all training sets.

R_{mol}^{avg}		**Training set**												
		Lit.	n_{beg}	n_{mid}	n_{end}	n_{all}	n_{hal}	c_{beg}	c_{mid}	c_{end}	c_{all}	c_{hal}	c_{met}	$c_{h,m}$
Tested set	n_{beg}	0.966	0.955	**0.962**	0.930	0.959	0.961	0.950	0.924	0.938	0.945	0.944	0.928	0.938
	n_{mid}	0.957	0.939	**0.951**	0.910	0.947	**0.951**	0.941	0.902	0.932	0.936	0.930	0.918	0.929
	n_{end}	0.960	0.944	**0.958**	0.922	0.944	0.956	0.956	0.894	0.942	0.945	0.932	0.929	0.942
	n_{all}	0.961	0.946	**0.957**	0.921	0.953	0.956	0.949	0.907	0.937	0.942	0.935	0.925	0.936
	n_{hal}	-	-	-	-	-	**0.928**	-	-	-	-	0.919	-	0.887
	c_{beg}	0.945	0.918	0.928	0.870	0.928	0.930	**0.946**	0.917	0.934	0.941	0.936	0.916	0.937
	c_{mid}	0.934	0.912	0.922	0.867	0.921	0.920	**0.932**	0.902	0.921	0.928	0.922	0.898	0.921
	c_{end}	0.936	0.913	0.925	0.870	0.922	0.922	**0.936**	0.902	0.923	0.930	0.927	0.903	0.927
	c_{all}	0.939	0.914	0.925	0.869	0.924	0.924	**0.938**	0.907	0.926	0.933	0.928	0.906	0.928
	c_{hal}	-	-	-	-	-	0.903	-	-	-	-	**0.910**	-	0.885
	c_{met}	-	-	-	-	-	-	-	-	-	-	-	**0.887**	0.879
	$c_{h,m}$	-	-	-	-	-	-	-	-	-	-	-	-	**0.885**

R_{mol}^{avg} describes quality of parameters obtained via EEM parameterization using the training set. This value is between 0 and 1. The closer it is to 1, the more accurate charges are provided employing the EEM method using the parameters. The R_{mol}^{avg} is an average of R_{mol} values for all molecules in the training set. R_{mol} is the R-squared value of the linear regression line, which was inserted into a set of points [q_i(*ab initio*), q_i(EEM)], where q_i(*ab initio*) and q_i(EEM) are *ab initio* and EEM charges (calculated using the parameters) of atom *i*, respectively. Lit. means parameters obtained from literature [17]. For our parameters, the best R_{mol}^{avg} for each tested set is bolded.

Parameters, calculated for all training sets presented in Table 1, and also parameters obtained from the literature were validated for all training sets that contained suitable atoms. Validation of parameters for a selected training set was done in such a way that *ab initio* charges and the EEM charges calculated for each molecule from the training set using the developed parameters were compared via the least square method. In other words, the linear regression line was fitted to a set of points [q_i(*ab initio*), q_i(EEM)], where q_i(*ab initio*) and q_i(EEM) are *ab initio* and EEM charges of the atom *i*, respectively. Correlation between *ab initio* and EEM charges was described by the R-squared value [26] of this line. This R-squared value is between 0 and 1. The closer it is to 1, the better the correlation is. The R-squared value (R_{mol}) was calculated for each molecule in the training set. An average value of (R_{mol}^{avg}) was calculated from all R_{mol} values in each set to express the quality of parameters for the set.

Table 3: Information about numbers of molecules and atoms in newly created training sets c_{beg2}, c_{hal2}, c_{met2} and $c_{h,m2}$. For more details see the text.

Element	**Bond order**	**Number of molecules and atoms in training set**							
		c_{beg2}		c_{hal2}		c_{met2}		$c_{h,m2}$	
		molecules	atoms	molecules	atoms	molecules	atoms	molecules	atoms
H	1	530	11187	810	13214	1112	25894	3082	60873
C	1	498	4113	729	5128	1070	10918	2847	24359
N	1	325	605	378	689	641	1353	1598	3195
O	1	400	1162	536	1258	830	2636	2185	6030
S	1	58	116	87	160	168	416	358	756
C	2	518	5871	843	10058	1078	12756	3086	37612
N	2	172	350	289	561	374	825	1062	2163
O	2	401	907	546	991	786	1943	2123	4449
Cl	1	-	-	455	1158	-	-	929	2319
Br	1	-	-	211	324	-	-	477	735
F	1	-	-	188	805	-	-	411	1745
I	1	-	-	57	95	-	-	134	202
Zn	1	-	-	-	-	103	178	155	268
Fe	1	-	-	-	-	186	317	203	335
Total		544	24311	870	34441	1154	57236	3258	145041

4. Results

Table 4: EEM parameters *A*, *B* and κ (see eqs. (1) and (2)) obtained via parameterization using training sets c_{beg2}, c_{hal2}, c_{met2} and $c_{h,m2}$.

		EEM parameters created using training sets							
		c_{beg2}		c_{hal2}		c_{met2}		$c_{h,m2}$	
		κ 0.44		κ 0.66		κ 0.42		κ 0.55	
Element	Bond order	*A*	*B*	*A*	*B*	*A*	*B*	*A*	*B*
H	1	2.396	0.959	2.404	1.461	2.386	0.937	2.394	1.212
C	1	2.459	0.611	2.503	0.899	2.452	0.593	2.476	0.772
N	1	2.597	0.790	2.653	1.017	2.550	0.663	2.597	0.835
O	1	2.625	0.858	2.713	1.211	2.624	0.847	2.676	1.077
S	1	2.407	0.491	2.465	0.705	2.424	0.400	2.440	0.665
C	2	2.464	0.565	2.516	0.850	2.462	0.527	2.495	0.704
N	2	2.554	0.611	2.633	0.869	2.547	0.639	2.600	0.790
O	2	2.580	0.691	2.757	1.348	2.567	0.622	2.622	0.850
Cl	1	-	-	2.791	2.365	-	-	2.759	2.092
Br	1	-	-	2.496	1.345	-	-	2.494	1.315
F	1	-	-	2.789	1.494	-	-	3.032	2.985
I	1	-	-	2.421	2.309	-	-	2.454	1.387
Zn	1	-	-	-	-	2.378	0.259	2.422	0.301
Fe	1	-	-	-	-	2.557	0.061	2.575	0.087

For each training set of molecules in Table 1, the parameters were found. As it was described in the Methods section, calculated parameters were validated for all training sets that contained suitable atoms and also compared with the parameters from literature [17]. As the literature does not show the κ value (see eq. (1)), we had to find the κ value via our methodology. The best fit for κ was found to equal 1.25. The results of this parameter quality validation expressed by R_{mol}^{avg} are summarized in Table 2. This table shows that the quality of parameters varies for different training sets. Moreover, the quality of parameters from literature is generally slightly better than the quality of our parameters. Therefore, our effort was to further improve our methodology and parameters. The main idea of this improvement was based on results, obtained for training sets n_{all} and its subsets n_{beg}, n_{mid} and n_{end} and also for training set c_{all} with subsets c_{beg}, c_{mid} and c_{end}. It is seen from Table 2 that R_{mol}^{avg} (n_{all}) is better than the average value from R_{mol}^{avg} (n_{beg}), R_{mol}^{avg} (n_{mid}) and R_{mol}^{avg} (n_{end}), but the best results are obtained for the set n_{mid}. Analogically, in the training set c_{all}, the subset c_{beg} provides the best parameters. Randomly sorted molecules that create the training sets imply the good accuracy of parameters from subsets n_{mid} and c_{beg}. Therefore, the quality of parameters can be increased by selection of an appropriate subset of the input training set. We have tested two methods of appropriate subset selection:

1. Select only molecules, which have R_{mol} greater than a defined limit (for example 0.8).

2. Sort molecules from the training set T randomly and create a sequence of them (1, 2, ..., $|T|$), where $|T|$ is a cardinality of T. Calculate parameters for all subsets ST_i, where ST_i is obtained from T by removing the subset DST_i. The subset DST_i is composed of elements $T_{(i-1).K+1}$, $T_{(i-1).K+2}$, …, $T_{i.K}$; where K can be, for example, 100. Now create the selection in the following way: From the input training set, sorted into the above described sequence, delete every subset DST_i, for which $R_{mol}^{avg}(T) < R_{mol}^{avg}(ST_i)$.

By comparison, the second approach was found to be more successful. It is interesting that sets selected via the first method provide worse quality of parameters than the input training sets themselves (results not shown here).

Using method 2, we have performed selections based on sets c_{beg}, c_{hal}, c_{met} and $c_{h,m}$ and created sets c_{beg2}, c_{hal2}, c_{met2} and $c_{h,m2}$ (see Table 3).

Table 5: Comparison of the quality of parameters obtained using original sets and their selected subsets.

R_{mol}^{avg}		**Training set**								
		Lit.	c_{beg}	c_{beg2}	c_{hal}	c_{hal2}	c_{met}	c_{met2}	$c_{h,m}$	$c_{h,m2}$
Tested set	n_{beg}	0.966	0.950	***0.968***	0.944	0.958	0.928	0.959	0.938	0.950
	n_{mid}	0.957	0.941	***0.962***	0.930	0.952	0.918	0.951	0.929	0.943
	n_{end}	0.960	0.956	***0.970***	0.932	0.953	0.929	0.956	0.942	0.949
	n_{all}	0.961	0.949	***0.967***	0.935	0.954	0.925	0.955	0.936	0.947
	n_{hal}	-	-	-	*0.919*	**0.940**	-	-	*0.887*	*0.927*
	c_{beg}	0.945	*0.946*	***0.960***	0.936	*0.954*	0.916	*0.954*	0.937	*0.947*
	c_{mid}	0.934	0.932	***0.948***	0.922	*0.943*	0.898	*0.941*	0.921	*0.934*
	c_{end}	0.936	*0.936*	***0.951***	0.927	*0.945*	0.903	*0.944*	0.927	*0.937*
	c_{all}	0.939	0.938	***0.953***	0.928	*0.947*	0.906	*0.946*	0.928	*0.939*
	c_{hal}	-	-	-	*0.910*	**0.934**	-	-	*0.885*	*0.921*
	c_{met}	-	-	-	-	-	*0.887*	**0.927**	*0.879*	*0.917*
	$c_{h,m}$	-	-	-	-	-	-	-	*0.885*	**0.919**

For more details about R_{mol}^{avg} see Table 2. R_{mol}^{avg} values of our parameters that are better than the literature parameters (denoted as Lit., taken from reference [17]) are in italics. The best R_{mol}^{avg} value (our parameters) for each tested set is bolded.

We have chosen the CSD database as this database contains high quality experimental data. The set c_{beg} was selected as it exhibits R_{mol}^{avg} higher than c_{mid}, c_{end} and c_{all}. The parameters were calculated for selected subsets c_{beg2}, c_{hal2}, c_{met2} and $c_{h,m2}$ (see Table 4). Then the parameters were validated for all training sets containing corresponding atoms (see Table 5 and graphs in supplementary materials).

It is seen that the selected subsets c_{beg2}, c_{hal2}, c_{met2} and $c_{h,m2}$ provide markedly better parameters than the input sets c_{beg}, c_{hal}, c_{met} and $c_{h,m}$ themselves. In all cases we have found parameters that are better than the literature parameters.

The parameters c_{beg2} are of better quality than parameters obtained from literature [17] for both used databases. The parameters c_{hal2}, c_{met2} and $c_{h,m2}$ are of a worse quality than published parameters for the NCI DIS database, but are better for the experimental database CSD. Moreover, these parameters contain new data for halogens or Fe and Zn.

Generally, we can conclude, that it is possible to calculate parameters using both the predicted and experimental databases. However, parameters that are based on experimental structures exhibit better charge calculation results. It can be caused by the fact that the theoretical structures from NCI DIS database may include some less realistic geometries compared to the experimental structures from CSD database. These parameters are more useful as they are portable and can be used for an arbitrary molecule that contains atoms for which the parameters were developed. Our results also show that it is useful to work with large training sets and select the best subset that provides the highest quality parameters. It is also reasonable to test several training sets.

We did a large validation of our parameters. For demonstration, tables with detailed results of EEM charge calculation method with our parameters for several different organohalogene and organometal molecules are attached in supplementary material. Also coordinates and charges on single atoms are available there.

5. Conclusions

In this work, we have improved the published EEM parameters to calculate the STO-3G MPA charges for C, O, N, H, S, F and Cl. The new parameters provide more accurate charges than those published previously [17]. We have developed parameters for elements not yet parameterized, specifically for Br, I, Fe and Zn.

The EEM parameterization we have performed has been based on 12 training sets, which are also the largest published training sets used for the EEM parameterization ranging from 2000 to 6000 molecules. We have shown that the number of molecules in the training set is very important for the quality of the parameters.

We have performed crossover validation of all obtained parameters using all training sets that include relevant elements. To the best of our knowledge, we have performed the most accurate testing of EEM parameters quality published so far.

This is the first work to compare EEM parameters calculated using two principally different training sets, one being a database of theoretically predicted 3D structures (NCI DIS) and the second being a database of crystallographic structures (CSD). Our results show that it is possible to use both databases, but parameters from the CSD database training sets give more accurate charges. Moreover, the parameters obtained from the NCI DIS database training sets are not very suitable to calculate charges for molecules from the CSD database.

These improved and newly developed parameters can be used for charge calculation using the program EEM SOLVER [27], which we have developed and which is freely available via the internet on http://ncbr.chemi.muni.cz/~n19n/eem_abeem.

Acknowledgements

We would like to thank the Supercomputing Centre in Brno for providing us with access to its computer facilities. This research has been supported in part by Ministry of Education of the Czech Republic (contracts LC06030 (JK) and MSM0021622413 (RSV)). The financial support is gratefully acknowledged.

References and Notes

1. Mortier, W.J.; Van Genechten, K.; Gasteiger, J. Electronegativity Equalization: Application and Parametrization. *J. Am. Chem. Soc.* **1985**, *107*, 829-835.
2. Mortier, W.J.; Ghosh, S.K.; Shankar, S. Electronegativity Equalization Method for the Calculation of Atomic Charges in Molecules. *J. Am. Chem. Soc.* **1986**, *108*, 4315-4320.
3. Van Genechten, K.A.; Mortier, W.J.; Greelings, P. Intrinsic Framework Electronegativity: A Novel Concept in Solid State Chemistry. *J. Chem. Phys.* **1987**, *86*, 5063-5071.
4. Parr, R.G.; Yang, W. *Density Functional Theory of Atoms and Molecules*; Oxford University Press: New York, 1989.
5. Bartolotti, L.J.; Flurchick, K. An Introduction to Density Functional Theory. *Rev. Comp. Chem.* **1996**, *7*, 187-216.
6. Parr, R.G.; Donnelly, R.A.; Levy, R.A.; Palke, W.E.J Electronegativity: The Density Functional Viewpoint. *Chem. Phys.* **1978**, *68*, 3801-3807.
7. Donnelly, R.A.; Parr, R.G. Elementary Properties of an Energy Functional of the First-order Reduced Density Matrix. *J. Chem. Phys.* **1978**, *69*, 4431-4439.
8. Sanderson, R.T. *Chemical Bond and Bond Energies*; Academic Press: New York, 1976.
9. Sanderson, R.T. *Polar Covalence*; Academic Press: New York, 1983.
10. Parr, R.G.; Pearson, R.G. Absolute Hardness: Companion Parameter to Absolute Electronegativity. *J. Am. Chem. Soc.* **1983**, *105*, 7512-7522.
11. Grant, G.H.; Richards, W.G. *Computational Chemistry*; Oxford University Press: USA, 1995.
12. Bachrach, S.M. Population Analysis and Electron Densities from Quantum Mechanics. *Rev. Comp. Chem.* **1994**, *5*, 171-227.
13. Bultinck, P.; Langenaeker, W.; Lahorte, P.; DeProft, F.; Geerlings, P.; Waroquier, M.; Tollenaere, J.P. The Electronegativity Equalization Method I: Parametrization and Validation for Atomic Charge Calculations. *J. Phys. Chem. A* **2002**, *106*, 7887-7894.
14. Bultinck, P.; Langenaeker, W.; Lahorte, P.; DeProft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, J.P. The Electronegativity Equalization Method II: Applicability of Different Atomic Charge Schemes. *J. Phys. Chem. A* **2002**, *106*, 7895-7901.
15. Yang, Z.-Z.; Shen, E.-Z.; Wang, L.-H. A scheme for Calculating Atomic Charge Distribution in Large Molecules Based on Density Functional Theory and Electronegativity Equalization. *J. Mol. Struct. (Theochem)* **1994**, *312*, 167-173.
16. Yang, Z.-Z.; Shen, E.-Z. Molecular Electornegativity in Density Functional Theory (I). *Sci. China B* **1995**, *38*, 521-528.

17. Yang, Z.-Z.; Shen E.-Z. Molecular Electronegativity in Density Functional Theory (II). *Sci. China B* **1996**, *39*, 20-28.
18. Menegon, G.; Shimizu, K.; Farah, J.P.S.; Dias, L.G.; Chaimovich, H. Parameterization of the Electronegativity Equalization Method Based on the Charge Model 1. *Phys. Chem. Chem. Phys.* **2002**, *4*, 5933-5936.
19. Zhang, Q.; Cagin, T.; van Duin, A.; Goddard III, W.A. Adhesion and Nonwetting-Wetting Transition in the *Al/* α*- Al_2O_3* Interface. *Phys. Rev. B* **2004**, *69*, 1-11.
20. Bollmann, L.; Hillhouse, H.W.; Delgass W.N. Calibration of an Electronegativity Equalization Method Based on Dft Results to Evaluate the Partial Charges, Global Softness, and Local Softness of Zeolite Structures. *Comp. Mol. Sci. Eng. Forum* **2005**, *21*, 597d.
21. Milne, G.W.A.; Nicklaus, M.C.; Driscoll, J.S.; Wang, S.; Zaharevitz. D. National Cancer Institute Drug Information System 3D Database. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219-1224.
22. Allen, F.H.; Kennard, O. Cambridge Structural Database. *Chem. Des. Autom. News* **1993**, *8*, 31-37.
23. Eaton, P.E., Millikan, R.; Engel, P. *The CHEMX Reference Manual*; Chemical Design Ltd: Oxford, 1990.
24. Dalby, A.; Nourse, J.G.; Hounshell, W.D.; Gushurst, A.K.I.; Grier, D.L.; Leland, B.A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244-255.
25. Frisch, M.J.; Trucks, G.W.; Schlegel, H.B.; Scuseria, G.E.; Robb, M.A.; Cheeseman, J.R.; Zakrzewski, V.G.; Montgomery Jr., J.A.; Stratmann, R.E.; Burant, J.C.; Dapprich, S.; Millam, J.M.; Daniels, A.D.; Kudin, K.N.; Strain, M.C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G.A.; Ayala, P.Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J.J.; Malick, D.K.; Rabuck, A.D.; Raghavachari, K.; Foresman, J.B.; Cioslowski, J.; Ortiz, J.V.; Baboul, A.G.; Stefanov, B.B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R.L.; Fox, D.J.; Keith, T.; Al-Laham, M.A.; Peng, C.Y.; Nanayakkara, A.; Challacombe, M.; Gill, P.M.W.; Johnson, B.; Chen, W.; Wong, M.W.; Andres, J.L.; Gonzalez, C.; Head-Gordon, M.; Replogle, E.S.; Pople, J.A. *Gaussian 98*; Gaussian Inc.: Pittsburgh, 2001.
26. Nemhauser, G.L.; Rinnooy, A.H.G. *Optimization*; North-Holland: Amsterdam, 1989.
27. Svobodová Vařeková, R.; Koča, J. Optimized and Parallelized Implementation of the Electronegativity Equalization Method and the Atom-Bond Electronegativity Equalization Method. *J. Comput. Chem.* **2005**, *27*, 396-405.

Optimized and parallelized implementation of the Electronegativity Equalization Method and the Atom-Bond Electronegativity Equalization Method

Software News and Update

Optimized and Parallelized Implementation of the Electronegativity Equalization Method and the Atom-Bond Electronegativity Equalization Method

R. SVOBODOVÁ VAŘEKOVÁ, J. KOČA
National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kotlářská 2, 611 37 Brno, Czech Republic

Received 15 May 2005; Accepted 19 September 2005
DOI 10.1002/jcc.20344
Published online in Wiley InterScience (www.interscience.wiley.com).

Abstract: The most common way to calculate charge distribution in a molecule is *ab initio* quantum mechanics (QM). Some faster alternatives to QM have also been developed, the so-called "equalization methods" EEM and ABEEM, which are based on DFT. We have implemented and optimized the EEM and ABEEM methods and created the EEM SOLVER and ABEEM SOLVER programs. It has been found that the most time-consuming part of equalization methods is the reduction of the matrix belonging to the equation system generated by the method. Therefore, for both methods this part was replaced by the parallel algorithm WIRS and implemented within the PVM environment. The parallelized versions of the programs EEM SOLVER and ABEEM SOLVER showed promising results, especially on a single computer with several processors (compact PVM). The implemented programs are available through the Web page http://ncbr.chemi.muni.cz/∼n19n/eem_abeem.

 J Comput Chem 27: 396–405, 2006

Key words: charge distribution; electronegativity equalization method; atom-bond electronegativity equalization method; optimization; parallelization; parallel virtual machine

Introduction

The most common approach for calculation of charge distribution in a molecule is quantum mechanics (especially *ab initio* methods). A disadvantage of quantum mechanics methods is that they are very time-consuming. Their time complexity is $O(B^4)$, where B is greater or equal to the number of valence electrons in the molecule.

Therefore, the Electronegativity Equalization Method (EEM),[1–3] an alternative method of charge calculation, has been developed. This method is based on density functional theory (DFT).[4,5] First, Parr et al.[6,7] applied DFT and formulated a new definition and explanation of electronegativity. Later on, Mortier et al.[8–10] applied Parr's definition of electronegativity and Sanderson's Electronegativity Equalization Principle (EEP) and created the EEM.[1–3] This method is able to calculate atomic charges markedly faster than *ab initio* approaches, as the EEM has a time complexity of $\theta(N^3)$, where N is the number of atoms in the molecule.[3] The accuracy of the EEM corresponds to the *ab initio* method for which the EEM was parametrized. The most common parametrization of the EEM is a parametrization for the SCF-HF method with the STO-3G basis set, where the charges are calculated by Mulliken population analysis.[11,12] Principally, it is also possible to parametrize the EEM for other basis sets (i.e., 6-31G*) and methods for charge calculation (i.e., CHELPG, MK, NPA, ESP, Hirshfeld method).[13,14] The EEM was mostly used to calculate charges on relatively small molecules (composed of less than a hundred atoms),[15,16] although applications on larger systems, for example, biopolymers, are also known.[17] This method was also applied within another context (e.g., calculation of Fukui function[18] and energy,[19] reactivity studies,[20,21] applications within molecular simulations,[22–26] etc.).

Several improvements of the method have been published (e.g.,[27–30]). The best known extension of the EEM is the Atom-Bond Electronegativity Equalization Method (ABEEM),[19,28,31] which also considers charges localized on bonds. This method is more time-consuming than the EEM as it has a time complexity of $\theta((N + M)^3)$, where N is a number of atoms and M a number of bonds.[31] However the ABEEM provides a more exact model of charge distribution than the EEM, which is the reason why it is more frequently used for larger molecules.[28] Similar to EEM, ABEEM also has several times been used in molecular simulations.[32–36]

Correspondence to: J. Koča; e-mail: jkoca@chemi.muni.cz

Contract/grant sponsor: Ministry of Education of the Czech Republic; contract/grant number: MSM0021622413

Today, two serial implementations of EEM are publicly available, GULP[37] (free of charge for academic users) and Vcharge[38] (commercial software). To the best of our knowledge, there is no implementation of ABEEM publicly available.

The effectiveness of equalization methods enables their application to calculations of conformationaly dependent charges in molecular mechanics or, eventually, even in molecular dynamics simulations.[22–25] This article is focused on an implementation, optimization, and parallelization of the methods using the Parallel Virtual Machine (PVM).[39,40]

Theoretical Basis

EEM

Using DFT, the effective (charge-dependent) electronegativity of the atom i in a molecule can be calculated by eq. (1):[1–3]

$$\chi_i = A_i + B_i \cdot q_i + \kappa \sum_{j=1(j\neq i)}^{N} \frac{q_j}{R_{i,j}} \tag{1}$$

where N is the number of atoms in the molecule, q_i and q_j are the charges distributed on the atoms i and j, respectively, $R_{i,j}$ is the distance between atoms i and j, and κ is the adjustive factor. The coefficients A_i and B_i are defined by eq. (2):

$$A_i = \chi_i^* = \chi_i^0 + \Delta\chi_i \qquad B_i = 2\eta_i^* = 2\left(\eta_i^0 + \Delta\eta_i\right) \tag{2}$$

where χ_i^0 is the electronegativity of an isolated neutral atom i, η_i^0 is the hardness, and $\Delta\chi_i$ and $\Delta\eta_i$ describe the molecular environment. The coefficients A_i, B_i and κ [used further in an equation system (5)] are calculated using a calibration.[1,15]

According to Sanderson's Electronegativity Equalization Principle,[8–10] the effective electronegativity of each atom in the molecule is equal to the molecular electronegativity $\overline{\chi}$:

$$\chi_1 = \chi_2 = \cdots = \chi_N = \overline{\chi}. \tag{3}$$

The total charge Q of the molecule is equal to the sum of all the atomic charges:

$$\sum_{i=1}^{N} q_i = Q. \tag{4}$$

The atomic charges are described using the equation system (5), which contains $N+1$ equations with $N+1$ unknowns: $q_1, q_2, \ldots, q_N$ and $\overline{\chi}$. This system was derived from equations (1), (3), and (4):[1]

$$\begin{pmatrix} B_1 & \frac{\kappa}{R_{1,2}} & \cdots & \frac{\kappa}{R_{1,N}} & -1 \\ \frac{\kappa}{R_{2,1}} & B_2 & \cdots & \frac{\kappa}{R_{2,N}} & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\kappa}{R_{N,1}} & \frac{\kappa}{R_{N,2}} & \cdots & B_N & -1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} q_1 \\ q_2 \\ \vdots \\ q_N \\ \overline{\chi} \end{pmatrix} = \begin{pmatrix} -A_1 \\ -A_2 \\ \vdots \\ -A_N \\ Q \end{pmatrix}. \tag{5}$$

The matrix of the equation system (5) is called an EEM matrix.

ABEEM

The ABEEM method[19,28,31] is similar to the EEM method with an extension on bonds. In other words, the charges are located not only on the atoms but also on the bonds. The charge of a bond is situated in the bond center, which is chosen by a suitable apportionment of the bond's length.[28] The ABEEM method uses adapted charges on atoms (Q_i) and bonds (Q_I), which can be calculated using eq. (6):

$$q_i = Q_i + \sum_{I=1\ (i\in I)}^{M} \frac{1}{2} \cdot Q_I \tag{6}$$

where q_i is the charge of the atom i, atoms are denoted by $i, j, \ldots$ and bonds by $I, J, \ldots$. The notation $i \in I$ means that the atom i is one of the bound atoms on the bond I, analogically $i \notin I$ [see eqs. (7) and (8)].

As formulated by Yang et al.,[28] the effective electronegativities of the atom $i(\chi_i)$ and the bond $I(\chi_I)$ can be described by eqs. (7) and (8):

$$\chi_i = A_i + B_i \cdot Q_i + C_i \sum_{I=1\ (i\in I)}^{M} Q_I + \kappa \left(\sum_{j=1\ (j\neq i)}^{N} \frac{Q_j}{R_{i,j}} + \sum_{J=1\ (i\notin J)}^{M} \frac{Q_J}{R_{i,J}} \right) \tag{7}$$

$$\chi_I = A_I + B_I \cdot Q_I + \sum_{i=1\ (i\in I)}^{N} C_{I,i} \cdot Q_i + \kappa \left(\sum_{J=1\ (J\neq I)}^{M} \frac{Q_J}{R_{I,J}} + \sum_{j=1\ (j\notin I)}^{N} \frac{Q_j}{R_{j,I}} \right) \tag{8}$$

where N is the number of atoms and M the number of bonds in the molecule; Q_i and Q_j are, respectively, charges on atoms i and j; Q_I and Q_J are, respectively, charges on bonds I and J; $R_{i,j}$, $R_{i,J}$, $R_{j,I}$, $R_{I,J}$ are distances between appropriate atoms or bonds; A_i and B_i are the same as in the EEM; A_I and B_I are defined by the equations: $A_I = \chi_I^*$ and $B_I = 2\eta_I^*$ [see eq. (2)], and C_i, $C_{I,i}$, and κ are parameters calculated during the calibration.[31]

For the ABEEM method, Sanderson's electronegativity equalization principle is formulated by eq. (9):

$$\chi_1 = \chi_2 = \cdots = \chi_N = \chi_{(1)} = \chi_{(2)} = \cdots = \chi_{(M)} = \overline{\chi} \tag{9}$$

where (1), ..., (M) are indexes of bonds.

The total charge of a molecule Q is equal to the sum of the charges on all the atoms and all the bonds in the molecule [eq. (10)]:

$$\sum_{i=1}^{N} Q_i + \sum_{I=1}^{M} Q_I = Q. \tag{10}$$

Table 1. Molecules Used for EEM SOLVER and ABEEM SOLVER Program Testing.

Molecule	Number of atoms	Number of bonds	Molecule	Number of atoms	Number of bonds
Formaldehyde	4	3	Cyclopentane	15	15
Methane	5	4	Cyclohexane	18	18
Ethene	6	5	Alanine dipeptide	23	22
Ethanol	9	8	Gly-Ala-Gly (folded)	27	25
Maleinic anhydrid	9	9	Gly-Ala-Gly (linear)	27	25
cis-2-Butene	12	11	*cis*-Retinal	49	47
trans-2-Butene	12	11	Tyr-Gly-Phe-Met	68	62

The ABEEM method calculates the charges of atoms and bonds using the equation system (11), which was derived from eqs. (7)–(10):

$$\begin{pmatrix} B_1 & \cdots & \frac{\kappa}{R_{1,N}} & E_{1,(1)} & \cdots & E_{1,(M)} & -1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\kappa}{R_{N,1}} & \cdots & B_N & E_{N,(1)} & \cdots & E_{N,(M)} & -1 \\ E_{(1),1} & \cdots & E_{(1),N} & B_{(1)} & \cdots & \frac{\kappa}{R_{(1),(M)}} & -1 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ E_{(M),1} & \cdots & E_{(M),N} & \frac{\kappa}{R_{(M),(1)}} & \cdots & B_{(M)} & -1 \\ 1 & \cdots & 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} Q_1 \\ \vdots \\ Q_N \\ Q_{(1)} \\ \vdots \\ Q_{(M)} \\ \overline{\chi} \end{pmatrix} = \begin{pmatrix} -A_1 \\ \vdots \\ -A_N \\ -A_{(1)} \\ \vdots \\ -A_{(M)} \\ Q \end{pmatrix}. \tag{11}$$

This equation system contains $N + M + 1$ unknowns (q_1, $q_2, \ldots, q_N$; $q_{(1)}, q_{(2)}, \ldots, q_{(M)}$; $\overline{\chi}$) and $N + M + 1$ equations. The terms $E_{i,I}$ are defined as follows: if $i \in I$ then $E_{i,I} = C_i$ and $E_{I,i} = C_{I,i}$, else $E_{I,i} = E_{i,I} = \kappa/(R_{i,I})$. In most cases, $C_{i-j,i}$ is equal to $C_{i-j,j}$ and both are denoted by C_{i-j}. Only the bonds C—H, N—H, and O—H are exceptions within all sets of molecules for which the parametrization was performed,[28–31] because $C_{X—H,H}$ are not equal to $C_{X—H,X}$. For these bonds, the terms $C_{X—H,H}$ and $C_{X—H,X}$ are labeled by $C_{X—H}$ and $D_{X—H}$, respectively.

Implementation

The EEM and ABEEM methods were implemented using the following algorithm:

1. Calculate distances between atoms and (in ABEEM) between atoms and bonds and between bonds
2. Create the EEM or ABEEM matrix
3. Solve the equation system described by the EEM or ABEEM matrix
4. Distribute bond charges to bound atoms using eq. (6) (only for ABEEM method)

The above algorithms were implemented in C language and the EEM SOLVER and ABEEM SOLVER programs were written. Both programs use two input files: a PDB file with topology and coordinates, and a file with parameters.

Testing the Programs

The parameters used for testing the EEM and ABEEM method were taken from refs. 2 and 31. Charges were calculated for 14 molecules (see Table 1), each in energy minimum configuration. The molecules belonged to different classes and were of different sizes.

The results obtained by EEM SOLVER and ABEEM SOLVER were compared with atomic charges calculated by the software package Gaussian[41] using the *ab initio* HF method with a STO-3G basis set. The comparison is visualized by a linear regression line and the quality of the correlation between *ab initio* and EEM or ABEEM charges is described by the *R*-squared value[42] of this line. The results are collected in Table 2. They correspond well with published data confirming that the methods were implemented correctly.

Figure 1 shows a comparison for an alanine dipeptide. This molecule was selected because it is often used as a reference system for the demonstration of equalization methods (e.g., refs. 2 and 17).

Table 2. Comparison of *Ab Initio* and EEM or ABEEM Charges Using *R*-Squared Values.

	R-squared value			*R*-squared value	
Molecule	EEM	ABEEM	Molecule	EEM	ABEEM
Formaldehyde	0,9932	0,9961	Cyclopentane	0,9913	0,9947
Methane	1,0000	1,0000	Cyclohexane	0,9878	0,9918
Ethene	1,0000	1,0000	Alanine dipeptide	0,9819	0,9840
Ethanol	0,9890	0,9939	Gly-Ala-Gly (folded)	0,9722	0,9804
Maleinic anhydrid	0,9931	0,9946	Gly-Ala-Gly (linear)	0,9779	0,9836
cis-2-Butene	0,9948	0,9972	*cis*-Retinal	0,9639	0,9736
trans-2-Butene	0,9955	0,9974	Tyr-Gly-Phe-Met	0,9701	0,9753

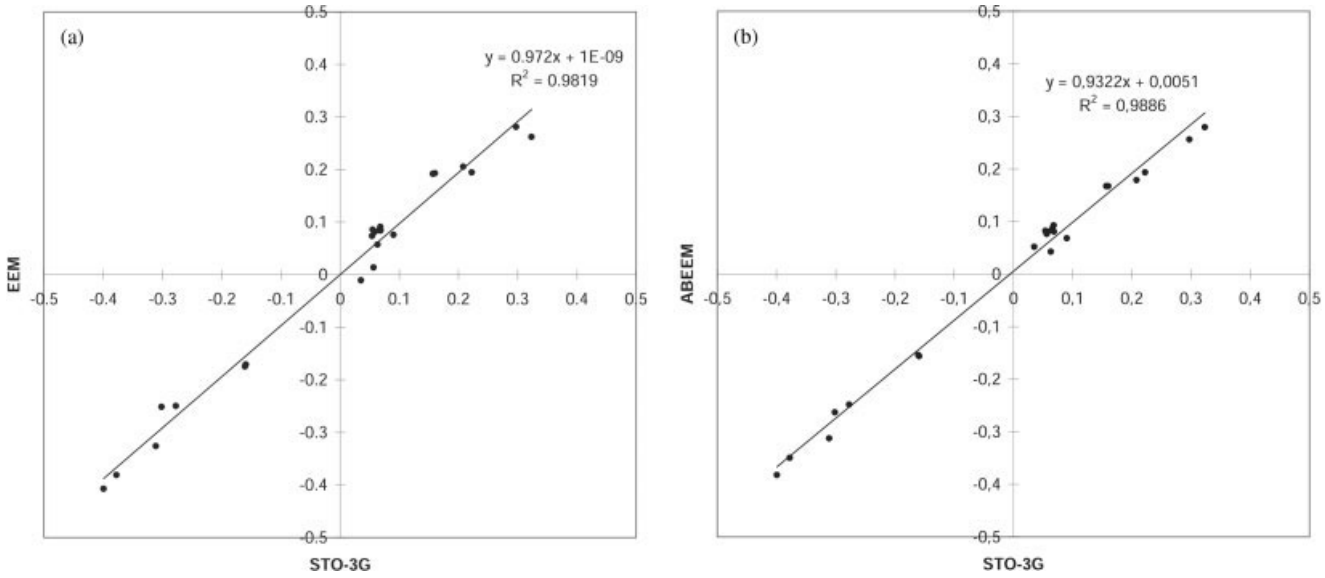


Figure 1. Comparison of the EEM and ABEEM with *ab initio* charges for alanine dipeptide. Each point in the graph represents a single atom.

Using EEM and ABEEM for Charge Calculation on Several Conformations of the Same Molecule

Before using the EEM or ABEEM methods in molecular mechanics calculations or molecular dynamics simulations, it is necessary to verify whether these methods are adequately accurate and can express differences between the charges of corresponding atoms in two different conformers of one molecule. Two conformations of each of three molecules were used to verify whether the EEM and ABEEM methods are sufficiently geometry sensitive: *cis*-2-butene and *trans*-2-butene, *cis*-retinal and *trans*-retinal, folded and linear Gly-Ala-Gly (Fig. 2). *Ab initio* (STO-3G), EEM, and ABEEM charges were calculated for both conformers of each molecule. Then, it was determined how large (or better to say, small) the charge differences are that the EEM and ABEEM methods can detect.

For illustration, the results for two conformers of Gly-Ala-Gly are shown (Table 3, Fig. 3). It is seen that both the EEM and ABEEM methods are sensitive enough to detect even very small charge differences. It is also seen that charge differences between corresponding atoms in different conformers are relatively significant even if the tripeptide Gly-Ala-Gly is not a particularly polar molecule.

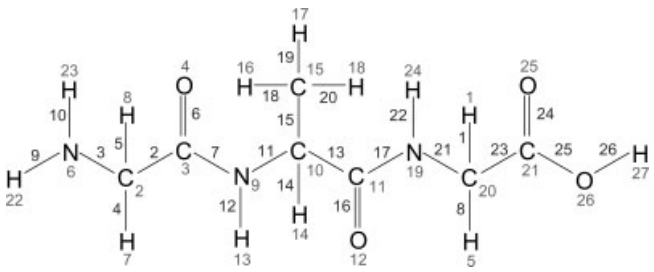


Figure 2. Atom numbering for Gly-Ala-Gly.

Optimization

The programs EEM SOLVER and ABEEM SOLVER can be divided into several parts (modules) described in Table 4.

As we expect to use the programs several (many) times on the same topology file but with different geometries, it would not be helpful to repeatedly read the topology of the molecule and the parameters for the EEM or the ABEEM in each calculation. A more effective solution is to use a partial evaluation. First, therefore, a header file for the input molecule is created. This file contains the topology of the molecule and the parameters for the EEM or the ABEEM. Then, only the coordinates of atoms in the calculated system are read in from the input. Moreover, we may predefine a maximum number of atoms (and bonds) during compilation of the programs. In this way we can replace dynamically allocated fields

Table 3. STO-3G, EEM, and ABEEM Atom Charges in Folded and Linear Conformations of Gly-Ala-Gly.

	Linear conformer			Folded conformer		
Atom	STO-3G	EEM	ABEEM	STO-3G	EEM	ABEEM
H13	0.23	0.22	0.22	0.21	0.21	0.20
H18	0.05	0.06	0.06	0.07	0.09	0.08
C21	0.32	0.29	0.33	0.30	0.28	0.30
H7	0.06	0.07	0.07	0.09	0.08	0.09
H23	0.17	0.19	0.15	0.19	0.22	0.19
H8	0.05	0.07	0.05	0.09	0.08	0.08
O12	−0.30	−0.24	−0.29	−0.25	−0.22	−0.25
O4	−0.31	−0.24	−0.28	−0.26	−0.21	−0.24
C11	0.30	0.30	0.30	0.23	0.26	0.21
C3	0.30	0.32	0.30	0.22	0.23	0.24

Only atoms with a charge difference of greater or equal to 0.02 are shown.

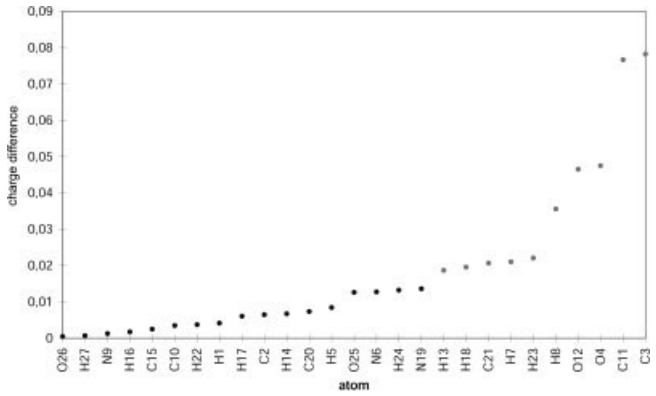


Figure 3. The absolute values of differences between *ab initio* (STO-3G) charges of corresponding atoms in folded and linear Gly-Ala-Gly peptide.

with static ones. This enables the compiler to perform a markedly better level of optimization. In both programs, we have also applied more common optimizations,[43] like the moving of redundant variables and data structures (e.g., fields of distances) into a proper part of the code, simplification of the code in cycles (i.e., the substitution of functions by its bodies, the optimization of dependencies between variables), etc. Some commonly used algorithms were substituted by more effective methods. For example, Gaussian elimination was replaced by the Cholesky method in EEM SOLVER (this substitution was possible only in EEM SOLVER, because the EEM matrix is symmetrical and the Cholesky method can only be used for this type of matrix).

The original and optimized versions of the programs EEM SOLVER and ABEEM SOLVER were executed on five different computers (see Table 5) and tested on 10 molecules of sizes of between 23 and 252 atoms for EEM SOLVER and between 23 and 1608 atoms for ABEEM SOLVER, respectively.

Besides the molecules shown in Table 1, we have also tested the programs by calculating charges on fragments of cyclin dependent kinase 2 (CDK2), a protein composed of 283 aminoacids with 4942 atoms, to obtain an idea of how the programs behave when charge calculations on larger systems are required. The CDK2 fragments were created in such a way that they always included the N-terminal part of the protein and were terminally blocked by the OH group

Table 4. Basic Modules of the EEM SOLVER and ABEEM SOLVER Programs.

Denotation	Description	Complexity[a]
Read	Reading parameters and information about a molecule.	$\theta(P^2)$
Prep	Preparation of the EEM or ABEEM matrix.	$\theta(P^2)$
Calc	Solving equation system, described using the EEM or ABEEM matrix.[b]	$\theta(P^3)$

[a]Time complexity expressed for the worst case ($P = N$ and $P = N + M$ for EEM and ABEEM method, respectively). Here, N and M stands for the number of atoms and number of bonds, respectively.
[b]Bond charges are also distributed over bound atoms using eq. (6)-only for the ABEEM method.

Table 5. Computers Used for EEM SOLVER and ABEEM SOLVER Optimization.

(a) Computers with a single processor

Denotation	Hardware	Operating system
Intel-PIII-700	Dual Pentium III, 700 MHz	Linux
Intel-PIII-1000-a	Dual Pentium III, 1 GHz	Linux
Intel-PIII-1000-b	Dual Pentium III, 1 GHz	Linux

(b) Multiprocessor computers

Denotation	Hardware	Operating system	Number of processors
SGI-P12	MIPS R10000, 200 MHz	IRIX 6.4	12
SGI-P40	MIPS R10000, 196 MHz	IRIX 6.4	40

on the other terminal. The initial geometry was taken as an equilibrated structure from our MD simulations reported elsewhere.[44] The total running time and the periods spent in the Read, Prep, and Calc modules were measured. The results calculated on SGI-P12 are shown in Table 6. Very similar data was also obtained for other computers included in the testing.

The data in Table 6(a) clearly shows that the optimized version of the EEM SOLVER is much more effective compared to the original one. For small molecules (less than 50 atoms) the optimized version was about three to five times faster while for larger molecules (more than 150 atoms) it was about 20 times faster. The most remarkable speed increase was observed in the Calc part, where the Gaussian elimination procedure was replaced by the Cholesky method. The optimization also brings a reduction of time for the ABEEM SOLVER (see Table 6), but the increase in speed was not as remarkable as it was for the EEM SOLVER. The main reason is that it was not possible to replace the Gaussian elimination method in this case. For all studied molecules, the optimized version is about four to eight times faster than the original one.

Parallelization

We have written parallel versions for both EEM SOLVER and ABEEM SOLVER.

As mentioned above, the most time consuming part of EEM and ABEEM SOLVER is the Calc module. It is clear that its high time complexity is caused by the first step of the Gaussian elimination, which is converting the EEM (or ABEEM) matrix to upper triangular form (reduction of a matrix). The algorithm WIRS (Wrapped Interleaved Row Storage)[45] was used to parallelize this step. WIRS works with one master process and allows for an arbitrary number of slave processes. The input of the algorithm is a matrix with K rows and K columns ($K = N + 1$ for EEM and $K = N + M + 1$ for ABEEM, where N is the number of atoms and M the number of bonds in the molecule). The algorithm works as follows. The master process executes all slave processes. The lines of the input matrix are equally distributed between processes. Each process performs reduction on the lines assigned to it. During this calculation, data

Table 6. Comparison of Running Times for Selected Molecules on Computer SGI-P12.

(a) EEM SOLVER

	Not optimized				After optimization			
Number of atoms	Read (ms)	Prep (ms)	Calc (ms)	Total (ms)	Read (ms)	Prep (ms)	Calc (ms)	Total (ms)
23	1.20	0.38	0.74	2.32	0.69	0.06	0.18	0.92
68	3.42	3.62	17.98	25.02	1.88	0.69	1.52	4.09
162	8.10	20.12	215.95	244.17	3.92	2.57	13.82	20.31
252	17.65	54.81	855.61	928.07	7.75	6.72	40.19	54.65

(b) ABEEM SOLVER

	Not optimized				After optimization			
Number of atoms	Read (ms)	Prep (ms)	Calc (ms)	Total (ms)	Read (ms)	Prep (ms)	Calc (ms)	Total (ms)
23	1.56	1.03	4.62	7.21	0.74	0.20	0.72	1.65
162	11.14	72.03	1615.53	1692.70	5.12	19.26	179.72	204.10
536	30.08	962.46	73447.47	74440.00	14.30	302.99	19922.71	20240.00
1608	69.25	9092.24	1998878.51	2008040.00	30.53	4060.65	460288.82	464380.00

sent from other processes is also used. A detailed description of the algorithm is as follows.

The master process:
Executes slave processes on all remaining computers (processors) in the PVM.
Sends selected lines of the input matrix to a slave process, which will own them and elaborate them.*
Sends the first line of the matrix to all slave processes.
Executes CALCULATION.

Each slave process:
Receives selected lines of the input matrix.
Executes CALCULATION.

CALCULATION:

```
FOR (i = 1, 2, ..., K) {
    IF (i == 1) {
        IF (the process is a slave process) {
            The process receives the line i.
        }
    }
    ELSE IF (the process did not send a line in i − 1 step) {
        The process receives the line i.
    }
    FOR (all lines, owned by the process) {
        IF (k > i, where k is the index of the line) {
            Recalculates line k (using line i) according to
                these equations:
                l = a_ki / a_ii
                a_kj = a_kj − l.a_ij (j = i, i + 1, ..., K)
                where a_uv is the element on row u and
                column v of the matrix
        }
    }
    IF (the process owns line i + 1) {
        The process sends this line to other processes
    }
}
```

∗ The process j owns each kth line of the input matrix, for which $(k-1) \mod p = j$, where p is the number of processes. The master process has the number $j = 0$ and slave processes are numbered $1, 2, \ldots, p$.

It is clear from the above description that WIRS is a fine-grained algorithm. This means that the task (matrix reduction) is divided into many very simple subtasks (recalculation of single lines) during parallelization. All effective parallel algorithms for matrix reduction are fine-grained as the largest part of the matrix that can be used in a parallel algorithm as an independent unit is a single line. Fine-grained algorithms are very sensitive to slow computers in the PVM cluster as they contain a number of synchronization points where all processes must wait for the slowest one.

The WIRS algorithm was implemented in PVM[39,40] (Parallel Virtual Machine). We have chosen this platform because of its robustness, efficiency, portability, and ability to connect a heterogeneous collection of Unix and/or Windows computers.

The parallel version of the program ABEEM SOLVER was tested on three types of PVM: compact, homogeneous, and heterogeneous. The parallel version of EEM SOLVER was tested only on a compact PVM as it requires very high communication speed to be effective. A compact PVM consists of only one computer, usually designed as a cluster of processors. A homogeneous PVM includes several computers with the same architecture and software. A heterogeneous PVM is composed of computers with different architecture and software.

Table 7. The EEM SOLVER (a) and ABEEM SOLVER (b) Computational Times as Obtained for Different Sizes of Calculated Molecules and Different Number of Processors (on the SGI-P40 Machine).

(a) EEM SOLVER

		Number of atoms in CDK2 fragments									
		334	433	534	693	1005	1305	1608	1911	2503	3002
Serial	t(s)	0.35	0.70	1.25	2.57	8.19	21.30	48.24	84.67	195.88	355.24
Parallel	p_{ef}	2	2	2	4	4	5	4	7	6	6
	$t(p_{ef})$ (s)	0.47	0.86	1.55	1.99	4.97	10.02	16.69	28.96	53.25	90.74

(b) ABEEM SOLVER

		Number of atoms in CDK2 fragments									
		253	306	334	433	534	601	693	1005	1305	1608
Serial	t(s)	1.42	2.57	3.27	7.04	13.50	21.46	37.49	130.48	300.41	580.3
Parallel	p_{id}	—	—	—	—	—	2	4	8	9	12
	$t(p_{id})$ (s)	—	—	—	—	—	10.53	9.30	16.80	31.61	53.59
	p_{ef}	4	4	5	6	6	7	8	9	11	12
	$t(p_{ef})$ (s)	0.85	1.23	1.48	2.38	4.10	5.21	6.79	16.22	29.62	53.59

The value of p_{id} is defined as the greatest number of processors for which the scaling is still close to ideal (deviation is lower than 1%). The p_{ef} is defined as a borderline after which adding one more processor increases the absolute time of calculation.

Compact PVM

For testing, we used fragments of CDK2 with the following numbers of atoms: 253, 306, 334, 433, 534, 601, 693, 1005, 1305, 1608, 1911, 2503, and 3002. The fragments were created as described above. All processes were executed subsequently on SGI-P12 and SGI-P40 multiprocessor computers. On each of these two machines, both programs were executed gradually with each number of processes until all processors on the computer were used. A relation between the number of processes, p, which are used by the parallel program, and the running time of the program, $t(p)$, is in the case of ideal scaling described by eq. (12).

$$t(p) = t(1)/p \tag{12}$$

where $t(1)$ is the running time on a single processor.

The results for the SGI-P40 computer are collected in Table 7 and Figure 4. The results on the SGI-P12 computer exhibit the same trends.

The distribution of scaling as seen in Table 7 can be explained in the following way. The serial versions of both programs exhibit a time complexity of $\theta(K^3)$, where K is the size of the matrix, and it performs only calculations. The parallel programs perform two types of tasks: calculations and communication. For matrix reduction, the time complexity of calculation is $\theta(K^3/p)$ and the time complexity of communication is $\theta(K^2 \cdot p)$, see ref. 45. When the number of processes is small ($p \leq p_{id}$) and the molecule is sufficiently large, the time spent on communication is negligible compared to calculation time and the scaling of the parallel version is close to ideal. An increasing number of processes causes the calculation time per process to be shorter, but, on the other hand, the program spends more time in communication. If the number of processes is greater than p_{ef}, the increase in communication time is larger than the calculation time saving and, therefore, the overall efficiency decreases. It causes a pathological situation when the absolute time of calculation increases even if more processors are used. The exact explanation is that even if each process is doing less calculations (linear descend), it must do too many communication tasks (quadratic growth), that the total running time is longer than it was before adding the process.

As we assumed, the scaling of the program EEM SOLVER is not as good as for ABEEM SOLVER. However, for large systems also parallel version of EEM SOLVER becomes efficient.

Homogeneous PVM

A homogeneous PVM implementation of the program ABEEM SOLVER was tested on CDK2 fragments with the following number of atoms: 334, 433, 534, 601, and 693. Processes were executed on clusters of computers, as described in Table 8.

The program ABEEM SOLVER was executed gradually with the following number of processes: $1, 2, 3, \ldots, P_C$; where P_C is the number of computers in the cluster. Each process was executed on a different computer. The running time was measured for each number of processes.

The results obtained on the Intel-PIII-700-GE cluster are collected in Table 9 and Figure 5. It is seen that the parallel version of the program ABEEM SOLVER is significantly less effective for a homogeneous PVM than for a compact PVM. The reason is that the speed of communication is substantially higher for a compact PVM than for a homogeneous PVM. The results also demonstrate similar trends for the remaining clusters used for testing (not shown here).

It is seen that using the parallel version in a homogeneous PVM is not effective for small molecules while it is more successful for larger molecules. But also in this case we need to use a relatively

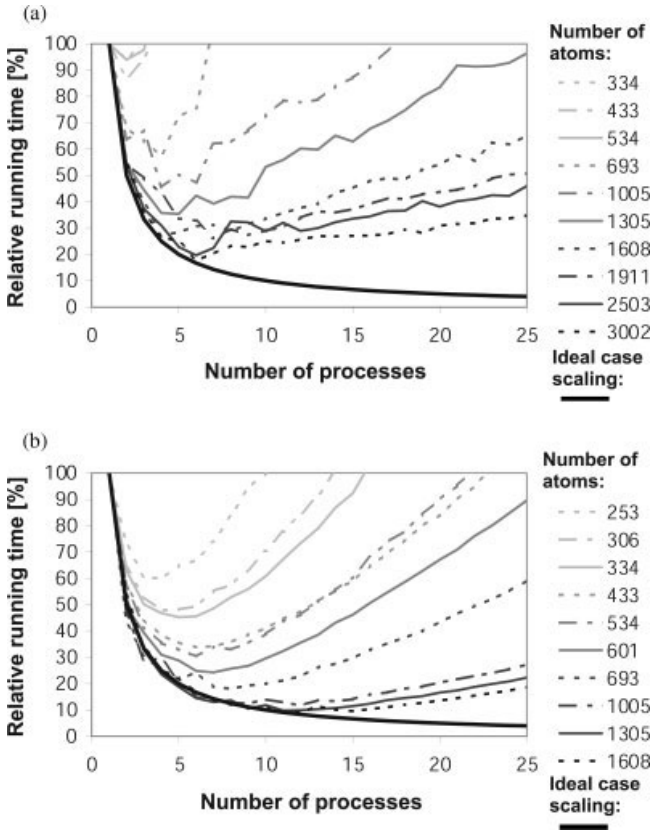


Figure 4. The EEM SOLVER (a) and ABEEM SOLVER (b) scaling on the SGI-P40 machine (see Table 5). The relative time was calculated as a_p/a_1, where a_p and a_1 is the absolute time for p processes and for 1 process, respectively. Adding more processes than 25 did not bring any improvement in scaling.

large number of processes to obtain a considerable decrease in running time. For example, the running time is only twice as short for 10 processes and a molecule with 601 atoms. Unfortunately, we were not able to run the parallel version on a substantially larger system (>1000 atoms) as we did not have a reasonable network infrastructure available. However, trends for larger molecules are visible from our tests. We can summarize that using the parallel version of the program in a homogeneous PVM can only be efficient for large molecules and with a fast network connection.

Table 9. The ABEEM SOLVER Computational Time as Obtained for Different Sizes of Calculated Molecules and Different Numbers of Processes (on the Intel-PIII-700-GE Cluster).

		Number of atoms in CDK2 fragments				
		334	433	534	601	693
Serial	t (s)	4.09	8.77	14.71	20.14	30.22
Parallel	$t(5)$ (s)	18.98	24.53	30.38	34.20	25.80
	$t(10)$ (s)	8.26	10.87	10.74	11.32	21.27
	$t(15)$ (s)	14.23	16.89	14.06	14.21	13.67

In the case of homogeneous PVMs the parameters p_{id} and p_{ef} (see Table 7) have no meaning as running time is very distant from the ideal scaling. Also, the oscillation of curves makes p_{ef} definition impossible.

Heterogeneous PVM

We used the same testing molecules as for the homogeneous PVM, and the program ABEEM SOLVER was also executed in the same way. Processes were run on Intel-PIII-700-GE and Intel-PIII-1000-b-FE clusters (see Table 8) and also on the SGI-P12 computer (see Table 5). We created three heterogeneous PVM's: Intel-PIII-1000-b-FE and Intel-PIII-700-GE, SGI-P12 and Intel-PIII-700-GE, SGI-P12 and Intel-PIII-1000-b-FE. For the first one, the running time

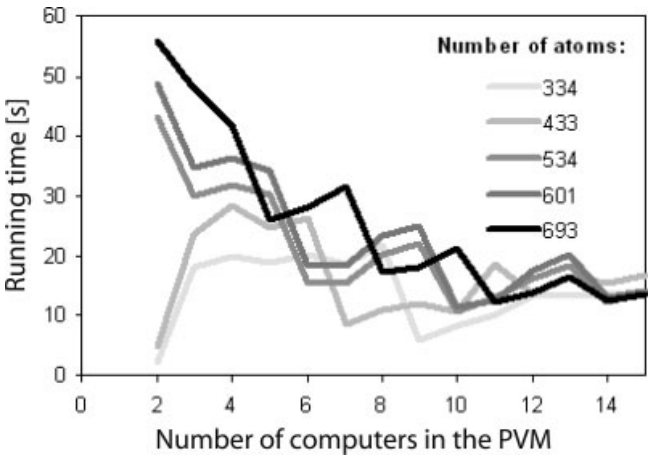


Figure 5. The ABEEM SOLVER scaling on a homogeneous PVM (an Intel-PIII-700-GE cluster, for cluster abbreviation see Table 7).

Table 8. Clusters of Computers, Used to Build Up a Homogeneous PVM.

		Network connection		Number of computers
Denotation	Computers	Name	Transfer speed	
Intel-PIII-1000-b-FE	Intel-PIII-1000-b	Fast Ethernet	100 Mb/s	16
Intel-PIII-700-GE	Intel-PIII-700	Giga Ethernet	1 Gb/s	16
Intel-PIII-1000-a-FE	Intel-PIII-1000-a	Fast Ethernet	100 Mb/s	16
Intel-PIII-1000-a-MN	Intel-PIII-1000-a	Myrinet	1.2 Gb/s	16

All computers in each cluster are identical, details are described in Table 5.

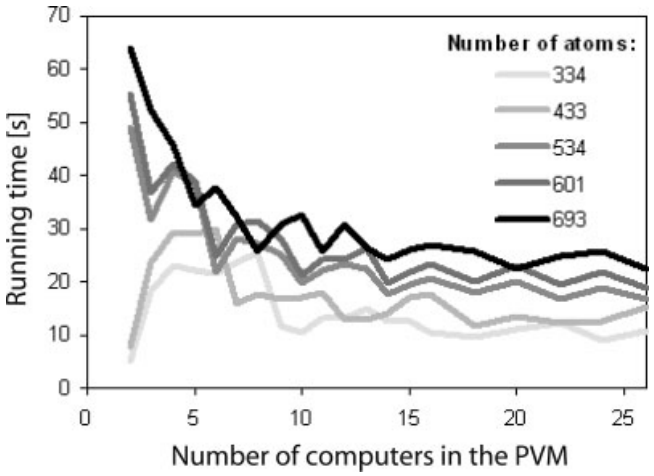


Figure 6. The ABEEM SOLVER scaling on a heterogeneous PVM (Intel-PIII-700-GE and Intel-PIII-1000-b-FE clusters, for cluster abbreviation see Table 7). The serial version was executed on the Intel-PIII-1000-b-FE. The calculation was organized in such a way that the first process was executed on the Intel-PIII-1000-b-FE cluster, the second on the Intel-PIII-700-GE, the third on the Intel-PIII-1000-b-FE, the fourth on the Intel-PIII-700-GE etc.

is shown in Figure 6 and Table 10. It is seen that the running time is limited by the slowest connection between two computers in a heterogeneous PVM.

Other heterogeneous PVM's tested demonstrate similar trends. Therefore, it is even more remarkable than in the previous case that running the parallel version on a heterogeneous PVM only has meaning for large systems.

Conclusion

The most common method of calculating partial atomic charges is quantum mechanics, which may, in many cases, be very time consuming. Electronegativity equalization methods, developed recently, are alternative approaches. These relatively fast (time complexity is of class $\theta(N^3)$ for molecule with N atoms) methods are based on DFT and their accuracy corresponds to *ab initio* quantum mechanical approaches. The goal of this work was to implement two electronegativity equalization methods, EEM and ABEEM. The EEM is a simple and very fast method, and it is mostly used for relatively small molecules. The ABEEM method is more general as it also includes the influence of bonds.

The above methods were implemented and the programs EEM SOLVER and ABEEM SOLVER were written and optimized. The optimized version of the program EEM SOLVER was found to be three to five times faster compared to the original one for small molecules (less than 50 atoms) and about 20 times faster for larger molecules (more than 150 atoms). After optimization, the program EEM SOLVER calculates atomic charges for a molecule with 23 atoms in 0.92 ms and for a molecule with 252 atoms in 54.65 ms on SGI-P12. The optimized version of ABEEM SOLVER was about four to eight times faster compared to the original one and calculates atomic charges for a molecule with 23 atoms in 1.65 ms and for a molecule with 536 atoms in 20.24 s on the same processor.

We have found out that the programs spent the majority of time performing reductions of the EEM or ABEEM matrix. This task has a time complexity of $\theta(N^3)$ while the remaining modules of the programs exhibit a time complexity of no more than $\theta(N^2)$. We have, therefore, implemented the algorithm WIRS (Wrapped Interleaved Rows Storage), which can sufficiently parallelize matrix reduction and used this algorithm for the reduction of EEM and ABEEM matrixes. This parallel algorithm was implemented on a PVM (Parallel Virtual Machine) and tested on a compact PVM, homogenous PVM and heterogeneous PVM. Parallelization has shown promising results for the compact PVM. For example, ABEEM charge calculation for a molecule with 1608 atoms took 53.59 s on 12 processors of SGI-P40 with a scaling of 11. The scaling of EEM SOLVER was not as good as for ABEEM SOLVER. However, for large systems also parallel version of EEM SOLVER becomes efficient. It has also been shown that the scaling of the parallel program ABEEM SOLVER is much worse in both homogenous and heterogeneous PVMs. In this case, it should only be used to calculate charges on large systems exceeding 1000 atoms. The above observations are not surprising, because the problem leads to an algorithm that belongs to fine-grain parallelism, which requires a high level of communication between processes.

Table 10. Comparison of Running Times for Serial and Parallel Versions of the Program ABEEM SOLVER in the Intel-PIII-1000-b-FE and Intel-PIII-700-GE Clusters.

		Numbers of atoms in CDK2 fragments				
		334	433	534	601	693
Serial	t (s)	4.04	8.73	16.45	23.41	36.00
Parallel	$t(8)$ (s)	25.34	17.50	27.02	31.34	25.71
	$t(16)$ (s)	10.38	17.59	20.77	23.17	26.82
	$t(26)$ (s)	10.64	15.20	16.85	18.85	22.58

The serial version was executed on the Intel-PIII-1000-b-FE. The calculation was organized in such a way that the first process was executed on the Intel-PIII-1000-b-FE, the second on the Intel-PIII-700-GE, the third on the Intel-PIII-1000-b-FE, the fourth on the Intel-PIII-700-GE etc.

Availability

The programs are available through the Web page: http://ncbr.chemi.muni.cz/~n19n/eem_abeem

Acknowledgments

One of the authors (J.K.) would like to thank Prof. Chan-Guo Zhan (Division of Pharmaceutical Sciences, University of Kentucky, Lexington, KY) for many fruitful discussions about the role of charges in molecular mechanics and quantum chemistry. We would like to thank the Supercomputing Centre in Brno for providing us with access to its computer facilities. Our thanks are also addressed to R. Turland (UK) for language corrections.

References

1. Mortier, W. J.; Van Genechten, K.; Gasteiger, J. J Am Chem Soc 1985, 107, 829.
2. Mortier, W. J.; Ghosh, S. K.; Shankar, S. J Am Chem Soc 1986, 108, 4315.
3. Van Genechten, K. A.; Mortier, W. J.; Greelings, P. J Chem Phys 1987, 86, 5063.
4. Parr, R. G.; Yang, W. Density Functional Theory of Atoms and Molecules; Oxford University Press: New York, 1989.
5. Bartolotti, L. J.; Flurchick, K. Rev Comp Chem 1996, 7, 187.
6. Parr, R. G.; Donnelly, R. A.; Levy, R. A.; Palke, W. E. J Chem Phys 1978, 68, 3801.
7. Donnelly, R. A.; Parr, R. G. J Chem Phys 1978, 69, 4431.
8. Sanderson, R. T. Chemical Bond and Bond Energies; Academic Press: New York, 1976.
9. Sanderson, R. T. Polar Covalence; Academic Press: New York, 1983.
10. Parr, R. G.; Pearson, R. G. J Am Chem Soc 1983, 105, 7512.
11. Grant, G. H.; Richards, W. G. Computational Chemistry; Oxford University Press: New York, 1995.
12. Bachrach, S. M. Rev Comp Chem 1994, 5, 171.
13. Bultinck, P.; Langenaeker, W.; Lahorte, P.; DeProft, F.; Geerlings, P.; Waroquier, M.; Tollenaere, J. P. J Phys Chem A 2002, 106, 7887.
14. Bultinck, P.; Langenaeker, W.; Lahorte, P.; DeProft, F.; Geerlings, P.; Van Alsenoy, C.; Tollenaere, J. P. J Phys Chem A 2002, 106, 7895.
15. Yang, Z.-Z.; Shen, E.-Z.; Wang, L.-H. J Mol Struct (Theochem) 1994, 312, 167.
16. Heidler, R.; Janssens, G. O. A.; Mortier, W. J.; Schoonheydt, R. A. Microporous Materials 1997, 12, 1.
17. Baeten, A.; Geerlings, P. J Mol Struct 1999, 465, 203.
18. Bultinck, P.; Carbo–Dorca, R.; Langenaeker, W. Phys Chem A 2002, 106, 7887.
19. Wang, Ch.-S.; Li, S.-M.; Yang, Z.-Z. J Mol Struct (Theochem) 1998, 430, 191.
20. Zhang, Y.-L.; Yang, Z.-Z. J Mol Struct (Theochem) 2000, 496, 139.
21. Cong, Y.; Yang, Z.-Z.; Wang, Ch.-S.; Liu, X.-Ch.; Bao, X.-H. Chem Phys Let 2002, 357, 59.
22. Smirnov, K. S.; van de Graaf, B. J Chem Soc Faraday Trans 1996, 430, 2469.
23. Smirnov, K. S.; van de Graaf, B. J Chem Soc Faraday Trans 1996, 92, 2475.
24. van Duin, A. C. T.; Dasgupta, S.; Lorant, F.; Goddard, W. A. J Phys Chem A 2001, 105, 9396.
25. Fan, J. F.; Xia, Q. Y.; Gong, X. D.; Xiao, H. M. Chem Res Chin Univ 2002, 18, 321.
26. Bultinck, P.; Vanholme, R; Popelier, P. L. A.; DeProft, F.; Greelings, P. J Phys Chem A 2004, 108, 10359.
27. Njo, S. L.; Fan, J.; van de Graaf, B. J Mol Catal 1998, 134, 79.
28. Yang, Z.-Z.; Wang, Ch.-S. J Phys Chem 1997, 101, 6315.
29. Cong, Y.; Yang, Z.-Z. Chem Phys Let 2000, 316, 324.
30. Yang, Z.-Z.; Cong, Y.; Wang, C. S. Chem J Chin Univ 2000, 20, 1781.
31. Wang, Ch.-S.; Yang, Z.-Z. J Chem Phys 1999, 110, 6189.
32. Yang, Z.-Z.; Li, X. J Phys Chem A 2005, 109, 3517.
33. Zhang, Q.; Yang, Z.-Z. Chem Phys Let 2005, 403, 242.
34. Wu, Y.; Yang, Z.-Z. J Phys Chem A 2004, 108, 7563.
35. Yang, Z.-Z.; Wu, Y.; Zhao, D.-X. J Chem Phys 2004, 120, 2541.
36. Yang, Z.-Z.; Wang, C. S. J Theo & Comp Chem 2003, 2, 273.
37. Gale, J. D. JCS Faraday Trans 1997, 93, 629.
38. Gilson, M. K.; Gilson, H. S. R.; Potter, M. J. J Chem Inf Comput Sci 2003, 43, 1982.
39. Geist, A.; Beguelin, A.; Dongarra, J.; Weicheng, J.; Mancheck, R.; Sunderam, V. PVM 3 User's Guide and Reference Manual; Oak Ridge National Laboratory, Oak Ridge, Knoxville, TN, 1994.
40. Geist, A.; Beguelin, A.; Dongarra, J.; Jiang, W.; Manchek, R.; Sunderam, V. PVM 3 User's Guide; MIT Press, Cambridge, MA, 1994. http://www.netlib.org/pvm3/book/pvm-book.html
41. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Salvador, P.; Dannenberg, J. J.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Baboul, A. G.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Andres, J. L.; Gonzalez, C.; Head–Gordon, M.; Replogle, E. S.; Pople, J. A. Gaussian 98; Gaussian, Inc., Pittsburgh, PA, 2001.
42. Nemhauser, G. L.; Rinnooy, A. H. G. Optimization; North-Holland: Amsterdam, 1989.
43. Dowd, K. High Performance Computing; O'Reilly & Associates, Inc., Cambridge, 1993.
44. Kříž, Z.; Otyepka, M.; Bártová, I.; Koča, J. Proteins: Struct Funct Bioinform 2004, 55, 258.
45. Golub, G.; Ortega, J. M. Scientific Computing: An Introduction to Parallel Computing; Academic Press: Boston, 1993.