



Research review paper

Computational design of enzymes for biotechnological applications

Joan Planas-Iglesias^{a,b}, Sérgio M. Marques^{a,b}, Gaspar P. Pinto^{a,b}, Milos Musil^{a,b,c},
Jan Stourac^{a,b}, Jiri Damborsky^{a,b,*}, David Bednar^{a,**}

^a Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kamenice 5/A13, 625 00 Brno, Czech Republic

^b International Clinical Research Center, St. Anne's University Hospital Brno, Pekarska 53, 656 91 Brno, Czech Republic

^c IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 61266 Brno, Czech Republic

ARTICLE INFO

Keywords:

Biocatalyst
Catalytic efficiency
Computational enzyme design
Enzyme biotechnologies
Protein engineering
Protein dynamics
Software
Solubility
Stability

ABSTRACT

Enzymes are the natural catalysts that execute biochemical reactions upholding life. Their natural effectiveness has been fine-tuned as a result of millions of years of natural evolution. Such catalytic effectiveness has prompted the use of biocatalysts from multiple sources on different applications, including the industrial production of goods (food and beverages, detergents, textile, and pharmaceuticals), environmental protection, and biomedical applications. Natural enzymes often need to be improved by protein engineering to optimize their function in non-native environments. Recent technological advances have greatly facilitated this process by providing the experimental approaches of directed evolution or by enabling computer-assisted applications. Directed evolution mimics the natural selection process in a highly accelerated fashion at the expense of arduous laboratory work and economic resources. Theoretical methods provide predictions and represent an attractive complement to such experiments by waiving their inherent costs. Computational techniques can be used to engineer enzymatic reactivity, substrate specificity and ligand binding, access pathways and ligand transport, and global properties like protein stability, solubility, and flexibility. Theoretical approaches can also identify hotspots on the protein sequence for mutagenesis and predict suitable alternatives for selected positions with expected outcomes. This review covers the latest advances in computational methods for enzyme engineering and presents many successful case studies.

1. Introduction

Proteins play a unique role in triggering, regulating, and executing all sorts of biological processes. The wide range of different functions that proteins can accomplish – from mere structural support to fine-tuning of imbricated signalling pathways – makes of them unique performers in the mosaic of biological processes. Amongst all other types of proteins, enzymes are characterized by their ability to accelerate chemical reactions, allowing the process to happen in a time-scale appropriate for life preservation (Stryer et al., 2002; Mak and Siegel, 2014; Wolfenden and Snider, 2001). Such an improved efficiency makes the use of these biocatalysts a naturally attractive alternative to the bare application of chemical reactions (Hughes and Lewis, 2018). Being natural amazing pieces of bio-machinery, enzymes would become even more valuable if their catalytic properties and stability were enhanced.

This is the purpose of enzyme engineering: (i) improving the performance of enzymes by rendering the chemical step of catalysis more efficient, (ii) accelerating the transport of substrate and products from the enzyme surface to the catalytic site and vice versa, (iii) suppressing substrate and product inhibitions by optimization of active sites and access tunnels, (iv) developing specific or promiscuous enzymes regarding the converted substrate, (v) improving protein production by increasing its expressibility and solubility, and (vi) enhancing protein stability in water or in the presence of co-solvents.

Besides their power as catalysts and their industrial applications, enzymes also have special relevance in different aspects of human health. Protein therapeutics (Faber and Whitehead, 2019) aims to improve the catalytic efficiency and stability of enzymes relevant to medical care. This is the case of TEM β -lactamase (Klesmith et al., 2017), for which an engineering effort provided a better understanding of the

* Corresponding author at: Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kamenice 5/A13, 625 00 Brno, Czech Republic.

** Corresponding author.

E-mail addresses: jiri@chemi.muni.cz (J. Damborsky), 222755@mail.muni.cz (D. Bednar).

<https://doi.org/10.1016/j.biotechadv.2021.107696>

Received 7 October 2020; Received in revised form 12 January 2021; Accepted 13 January 2021

Available online 26 January 2021

0734-9750/© 2021 Elsevier Inc. All rights reserved.

enzyme machinery with the prospects of improving its catalytic properties and developing specific inhibitors. Engineered enzymes have also shown potential to neutralize the effects of drugs during detoxification (Xue et al., 2011; Liu et al., 2013) or in removing pollutants from the environment (Wang et al., 2019). The goal of metabolic engineering is not only to design individual enzymes but cascades of enzymatic reactions constituting complete metabolic pathways. Different yeast and bacteria have been engineered to either be more vulnerable to the human immune system (Workalemahu et al., 2014) or to produce *ex-vivo* humanized biomolecules (Murakami et al., 2015). This technology holds a great promise in the context of personalized medicine.

Enzymes have been progressively tailored by Darwinian selection to optimize specific chemical reactions over evolutionary history, achieving a paramount competence in the efficiency, specificity and selectivity inside living cells. Experimentally, this process can be paralleled and hastened by directed evolution (Arnold, 2018). The huge impact of this revolutionary technique awarded its pioneer Professor Frances Arnold the Chemistry Nobel Prize in 2018 (<https://www.nobelprize.org/prizes/chemistry/2018/summary/>, n.d.). However, engineering enzymes by directed evolution can be costly, time- and labour-demanding. In this context, computational techniques represent an attractive complement to such experiments. This theoretical approach, known as rational or computational protein design, is often inspired on or directly builds from the ideas by the 2013 Nobel laureates Martin Karplus, Michael Levitt, and Arieh Warshel (<https://www.nobelprize.org/prizes/chemistry/2013/summary/>, n.d.). In this article, we will review the state of the art of computational methods for enzyme engineering, highlight the most commonly used tools in the rational design and illustrate some of their successful applications. Pretending to cover a broad range of areas in enzyme engineering and aiming to engage with the wider audience possible, we surmise that this review might appeal to both experts and non-experts in the field, expecting they could find in here information useful to their practical purposes. However, for deeper analyses of some of the methods herein presented we can refer the readers to more specific reviews in the context of machine learning (Mazurenko et al., 2020), or the engineering of thermostability (Xu et al., 2020), access pathways (Kokkonen et al., 2019), or dynamics (Kreß et al., 2018). Finally, it has to be noted that as experiments become increasingly miniaturized enabling high throughput output (Neun et al., 2020; Vasina et al., 2020), these become a stronger complement to calculations, which can be used as a starting point for further experimental engineering by directed evolution.

2. Computational methods for biocatalyst design

2.1. Biochemical reactivity

In-depth knowledge of the reaction of an enzyme is often important for the successful engineering of any of its features, be it catalytic efficiency, specificity, or stability. We will briefly explain the different approaches that may be used to understand the reaction mechanism of a target protein. Four major approaches will be mentioned, all of which have advantages and disadvantages (Siegbahn and Himo, 2011; Friesner and Guallar, 2005; Sousa et al., 2017; Lu et al., 2016; Kamerlin and Warshel, 2011; Kästner, 2011; Hutter, 2012). All these approaches are conceived to decipher an energetic profile for the reaction mechanism, either in the form of a potential energy surface or a free energy surface (Fig. 1). Such an energetic profile represents the energy for all the stationary points and transition states along the reaction coordinate. Enzyme design usually targets the free energy barrier, the energy difference between the enzyme-substrate complex and its transition state. This barrier can be engineered either by lowering the overall energy of the transition state or by designing an enzyme-substrate complex that is energetically and geometrically closer to the transition state (Andrews et al., 2013; Mak and Siegel, 2014; Roston and Cui, 2016). This can create some problems to the substrate binding process which would

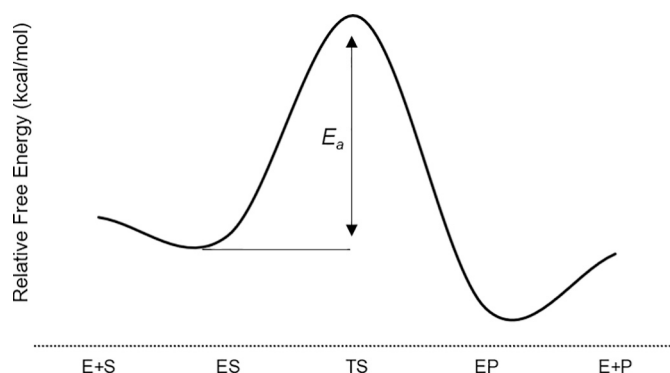


Fig. 1. Reaction coordinate obtained from a quantum-mechanics approach or hybrid quantum mechanics/molecular mechanics. The reaction coordinate is represented on the x-axis, starting with the enzyme-substrate (ES) complex, overcoming the energy barrier of the transition state (TS) and ending with the enzyme-product (EP) complex. The difference between ES and TS is activation energy (E_a) and correlates with k_{cat} .

need to be addressed in an enzyme engineering endeavour as seen in section 2.2. Table 1 summarizes the most common software tools that implement available methods to obtain a potential or free energy surface.

One of the most popular techniques for elucidating the enzymatic reaction mechanisms is the *quantum mechanics-cluster* approach (Himo, 2017). Quantum mechanics normally requires a starting model that represents a limited number of catalytic residues in the active pocket and the substrate of the reaction. The model should also include important residues for the stabilization of the catalytic residues and residues that anchor the substrate in the pocket, generally encompassing no more than 300 atoms. Historically, this size restriction was due to hardware limitations but nowadays, however, it is rather a methodological one. Models over the 300 atoms result in barrier problems such as the appearing of multiple local minima. In contrast, an excessively small model may insert unrealistic constraints into the system, limiting its capability for adjusting to the reaction mechanism changes that are

Table 1

Examples of commonly used software for the calculation of potential or free energy surface profiles. All the tools herein referred are stand-alone applications.

Software	Principle	Availability	URL	Reference
NWCHEM	QM, MM, MD	Free License	http://www.nwchem-sw.org	(Valiev et al., 2010)
Turbomole	QM	Paid License	http://www.turbomole.com/	(Ahlich et al., 1989)
Gaussian	QM and MM	Paid License	https://gaussian.com/	(Hehre et al., n.d.)
Q-Chem	QM and MM	Paid License	http://www.q-chem.com/	(Shao et al., 2015)
Q	EVB	Free for non-commercial use	http://xray.bmc.uu.se/~aqwww/q/	(Bauer et al., 2018)
Molaris	EVB	Paid License	https://laetro.usc.edu/software.html	(Lee et al., 1993)
AMBER	QM/MM; MD	Paid License	http://ambermd.org/	(Salomon-Ferrer et al., 2013)
NAMD	MD	Free for non-commercial use	http://www.ks.uiuc.edu/Research/namd/	(Phillips et al., 2005)
Gromacs	MD	Free License	http://www.gromacs.org/	(Van Der Spoel et al., 2005)
YASARA	MD; Docking	Paid License	http://www.yasara.org/	(Korendovych, 2018)

QM – quantum mechanics; MM – molecular mechanics; MD – molecular dynamics; EVB – electron valence bond; QM/MM – quantum mechanics/molecular mechanics

expected to occur during the catalytic process (Siegbahn and Himo, 2009). Since the construction of the active site model implies truncating the enzyme, the atoms at which the enzyme is cropped should be frozen in space to avoid the introduction of artificial freedom. Freezing the atoms can also prevent unfavourable steric interactions with the remaining modelled parts of the enzyme (Georgieva and Himo, 2010). Implicit solvation (also known as continuum solvation) models can also be used with a cluster approach to help simulate the presence of the enzyme around the active site and/or the solvent itself around the enzyme. A continuous and homogenous medium is introduced as a set of parameters to the system, but no coordinates are added since there is no explicit solvent (Skyner et al., 2015; Cramer and Truhlar, 1999). The parameter used and user-defined in implicit solvation models is the dielectric constant (ϵ), which defines the polarizability of the model being used. Different values of ϵ are used upon the simulation of a water model around the system or of the rest of the protein. When adopting this approach, the *density functional theory* and particularly the B3LYP functional are commonly used because of the trade-off between computational cost and accuracy (Himo, 2017). On the one hand, the principal advantage of the cluster model approach, over hybrid methods, is the relative ease of setting up the system for the calculations, especially with dedicated graphical user interfaces. However, finding the correct transition state and its internal reaction coordinate remains as arduous and lengthy tasks. On the other hand, the model on which QM calculations can be performed is necessarily partial and incomplete, and this is considered to be its main disadvantage.

Hybrid methods can be used as an alternative to quantum mechanics calculations focussing on a small and yet important part of the protein. Such methods consider two separated parts of the protein: while the active site and other important reactive parts are treated with quantum mechanics, the rest of the enzyme is treated using molecular mechanics. The interface of the two layers needs a more complex treatment, since severing the bond and summing the energies from quantum and molecular mechanics is not accurate (Hall et al., 2000). The atomic bonds at the interface of the quantum and molecular mechanics regions can be treated at different levels of approximation (Amara and Field, 2003) by methods such as capping potential (DiLabio et al., 2002), hybrid orbitals (Amara et al., 2000), and link atoms (Lyne et al., 1999). These methods rely on either the additive or the subtractive coupling schemes for the interface treatment. The latter, developed by Keiji Morokuma in the mid-nineties is implemented in Gaussian (Dapprich et al., 1999) under the acronym ONIOM. Initially, the whole system is treated under molecular mechanics conditions, using the parameters from the AMBER and Dreiding force fields to calculate the energy of the intact enzyme system. After obtaining the energy for the intact system, the energy of the cluster model is calculated also using molecular mechanics. This energy will later be subtracted to the energy from the intact system. In the last step, the energy in the cluster model is calculated with a quantum mechanics level of theory. In comparison to quantum mechanics approach, using hybrid methods is more computationally expensive because it models the full protein and a small layer of water molecules in a gas phase or solution.

Another method for modelling enzymatic reactions is the *empirical valence bond* (EVB) developed by the Nobel Laureate Arieh Warshel (Warshel and Weiss, 1981). The method models the wave function of the different atomic bonds as a linear combination of important covalent and ionic resonance forms. To describe the reactive part of the enzyme, each studied system is uniquely described by a set of parameters, which are chosen from existing force fields or built from scratch. This means, that the method combines two valence bond states representing the reactants and the products. To interpolate the energies, EVB makes use of classical force fields and free energy perturbation/umbrella sampling (commonly referred to as FEP/US) to calculate the potential energies of each state (Purg and SCL, 2018). The transition state is then parameterised using one of many possibilities: the same reaction occurring in another solvent with a simpler surrounding, experimental values or

quantum mechanical calculations for that same reaction. The first step in the EVB approach is the calibration of the free energy surface with a non-catalysed reference reaction in solution, and only then the contributions of the enzyme are considered. It is considerably faster than self-consistent field *quantum mechanics/molecular mechanics* hybrid methods. Given its equitable computational cost, empirical valence bond allows for sampling the different configurations of the reaction coordinates in a way that would not be possible with the more precise *ab-initio* methods. Although there is a clear gain in computational resources when using the empirical valence bond approach, its parameterization is challenging when no pre-calculated parameters are available for a given system and every little change in the system means a new parameterisation must be done. Another limitation of EVB is the calculation of multi-step reactions. Although theoretically, it is possible to use EVB for this purpose, it becomes a slower method if several steps need to be parameterised. More detailed descriptions about the usage of the EVB method can be found elsewhere (Warshel and Weiss, 1981; Kamerlin and Warshel, 2011; Luzhkov and Åqvist, 1998).

Before the turn of the millennium, the term *theozyme* (Tantillo et al., 1998) was proposed by Tantillo et al. to describe the building of model system where a given reaction would be simulated and studied to learn how to modify the real system to achieve a certain goal. The authors described this concept as “an array of functional groups in a geometry predicted by theory to provide transition-state stabilization” (Tantillo et al., 1998). This concept leads to the development of other methods and tools (Houk and Cheong, 2008; Richter et al., 2011; Zanghellini et al., 2006) that are now widely used. There are two main approaches for the use of theozymes: i) biomimetic enzymes to aid the understanding of enzymes of interest and ii) *de novo* design of enzymes as it will be explained in more detail in the section dedicated to *de novo* design. There are other applications where theozymes might be employed (Zipse et al., 1996; Bach and Canepa, 1997; Ben-Nun and Levine, 1995; Jansen et al., 1997) but those topics are beyond the scope of this review. Whereas at first, a theozyme would focus mostly on some key active residues in a geometrical disposition that was beneficial for catalysis, there is now an effort to show what mutations away from the active site (Osuna, 2020) can lead to changes in the most visited state of an enzyme and thus, change the catalytic cycle. These methods are, however, more akin to molecular dynamics simulations, which are discussed later (section 2.4).

Lastly, there exist hybrid models that additionally exploit *molecular dynamics* simulations, often referred to as quantum mechanics/molecular mechanics/molecular dynamics. Within this category, there are two main approaches: *umbrella sampling* and *metadynamics* (Laio and Parrinello, 2002; Kumar et al., 1992; Lence et al., 2018). When using hybrid methods, the reaction mechanism is still modelled and explored using quantum mechanics methods. After the stationary points of the reaction coordinate are located on the potential energy surface, geometrically optimized molecular dynamics simulations are performed to ensure the sampling of the reaction coordinate from reactants to products. In the case of umbrella sampling, several independent molecular dynamics simulations are conducted that introduce a biasing potential at different points along the reaction coordinate. After sampling the transition from reactants to products, the potentials are combined and unbiased to calculate the free energy surface. Since calculating even a few atoms with quantum mechanics along the reaction coordinate would be computationally very expensive, the quantum mechanics part is frozen and uncoupled from the molecular mechanics part of the protein. The main disadvantage of umbrella sampling is that finding the correct reaction path may be challenging (Mills and Andricioaei, 2008; Yang et al., 2019). Conversely, metadynamics starts with the huge advantage of waving the requirement of calculating an initial potential energy surface. Instead, the method assumes that the whole system can be described by a few collective variables and only their location on the energy surface is calculated. A potential is added biasing the system to sample further towards the products in the reaction coordinate until a

barrier is overcome. When the potentials added become close to a constant, the system has been sampled enough and the free energy surface is the inverse of the potentials added. Although it might seem that the required knowledge of the system is lesser when using metadynamics, the choosing of collective variables is not trivial. This method has been used for complex energy surfaces and rare events successfully (Bussi and Laio, 2020; Valsson et al., 2016).

Some of the methods herein described were used to increase the efficiency of the organophosphate degrading enzyme from *Agrobacterium radiobacter* P230, and to design a highly efficient cocaine-detoxifying enzyme. These case examples are described in section 3.1.

2.2. Ligand binding

The interactions of biomolecules with smaller ligands play a crucial role in the function of most natural proteins. Hence, it is not surprising that the prediction of ligand binding is very important in multiple fields of experimental sciences such as medicinal chemistry, protein engineering, biochemistry and molecular biology. *Molecular docking* is the most commonly used of all the computational methods developed for this purpose. Such popularity arises from the quickness and versatility of the technique for predicting the binding modes, interactions, and affinity of ligands with their biological targets. It can be used in enzyme engineering for quickly assessing whether a new compound is a potential substrate or inhibitor of the studied enzyme or to explain preferences for particular substrates. Docking is often used to provide good starting points for more advanced modelling techniques whose goals are to explore the bound ensembles with molecular dynamics or to evaluate the chemical reactivity with quantum mechanics. All protein-ligand docking programs aim to predict the preferential binding mode by determining the minimum energy conformation of the protein-ligand complexes and to provide a score that quantifies the strength of such binding. The method involves an iterative cycle of three main steps: i) search the space of possible molecular conformations; ii) score those protein-ligand binding poses; iii) compare the scores and proceed towards the energetic minimum. Multiple approaches and algorithms have been developed to optimize each of these steps and to find a good trade-off between accuracy and calculation speed. When the binding site of some ligand is unknown, such as in the case of newly discovered proteins, *blind docking*, in which the entire protein surface is considered for docking the ligand, can be used. This approach requires high exhaustiveness in the search, but it can provide important information about the catalytic site of novel enzymes or allosteric binding sites. However, most commonly the active site is already known and *site-specific docking* can be applied. In this case, the docking search is restricted to a smaller region where the ligand is expected to bind, and thus safe computational resources and time. The receptor search space is typically defined by the centre of the docking box (specified either by the coordinates or by a residue or ligand found in the active site) together with a box size or radius.

An initial challenge is to account for all possible conformers of the ligand, which can be overcome by very different algorithms: (i) systematic search performs an exhaustive rotation and translation of different groups of atoms from the ligand; (ii) molecular mechanics displace the ligand from a random starting conformation to a local minimum; (iii) Monte Carlo simulated annealing applies random changes on the ligand conformations based on a probabilistic function; (iv) genetic algorithms introduce mutations and cross-overs on a random population of conformations; and (v) incremental construction is based on the ligand fragmentation and its incremental reconstruction (Sousa et al., 2006; Guedes et al., 2014). Traditionally, the receptor is considered rigid and only the conformation of the ligand is changed. In a more realistic situation, the receptor can be considered flexible as well, at the cost of considerably increasing the computation time. This is called *flexible docking*. In soft docking, the archetype of flexible docking, a certain degree of protein-ligand overlapping is allowed by decreasing

the *Lennard-Jones repulsion potentials*. Side-chain flexibility docking makes the side chains of one or more residues flexible while keeping the protein's backbone fixed. The binding poses are optimized by molecular relaxation usually on a post-processing refinement step. Ensemble docking is characterized by considering several protein conformations instead of a single one. The method may consider collective degrees of freedom, meaning that the whole protein can be flexible but the high number of degrees of freedom are reduced to the main components of the motion of the protein (Pujadas et al., 2008; Naqvi et al., 2018; Saikia and Bordoloi, 2019).

Scoring functions represent the core of the second step of the molecular docking and are used to estimate the quality of every docked pose by a quantitative score. At the end of the docking cycle, the most likely conformations of the protein-ligand complex and their respective binding affinity are determined based on the scoring function. There exist many scoring functions that may be classified according to their rationale. In force field-based functions, the atom-pair energy terms are calculated using physical principles. Empirical scoring schemes build the final binding score by a linear combination of intuitive interaction terms, to best fit experimental training data sets. Knowledge-based functions rely on atom-pair potential terms obtained from statistical analysis of large datasets of protein-ligand complexes with known three-dimensional structures. Machine-learning approaches are trained on experimental datasets using algorithms of various types and are mostly aimed for post-processing rescoring (Guedes et al., 2018; Li and Fu, 2019). Some of the most popular docking programs are AutoDock (Morris et al., 2009), AutoDock Vina (Trott and Olson, 2009), GOLD (Jones et al., 1997), FlexX (Rarey et al., 1996), DOCK (Kuntz et al., 1982), ICM (Abagyan et al., 1994), QuickVina-W (Hassan et al., 2017), and Glide (Friesner et al., 2004). A comparison between these tools can be found elsewhere (Pujadas et al., 2008; Saikia and Bordoloi, 2019; Pagadala et al., 2017).

Virtual screening is closely related to molecular docking. Its main goal is to computationally evaluate a large library of molecules –typically from hundreds to millions, and predict which ones are likely to bind to the biological target. This procedure generates a smaller set of candidates for biological testing and avoids the experimental screening which can be costly and time-consuming (Westermaier et al., 2015; Rognan, 2017). Virtual screening becomes an ideal tool to predict potential inhibitors, substrates, or novel metabolic pathways. In protein engineering, it can assess which proteins amongst a library of variants can better accommodate and catalyse a given substrate of interest (Zhang et al., 2019) or to identify potential substrates from a library of compounds (Xu et al., 2009; Daniel et al., 2015). Some of the most popular compound libraries are ZINC (Irwin et al., 2012), ChEMBL (Gaulton et al., 2012), PubChem (Kim et al., 2016), BindingDB (Liu et al., 2007), and PDBbind (Liu et al., 2015). Virtual screening can be ligand- or structure-based. The first is used when the biological activity of a number of ligands is experimentally known (Stahura and Bajorath, 2005). The structure-based approach can only be applied when the three-dimensional structure of the biological target is known. The method docks all the ligands into the binding site of the receptor and ranks the resulting poses with scores that estimate their corresponding binding affinities. The results from virtual screening can be narrowed down by extra filters such as predicted solubility and bioavailability (Naqvi et al., 2018; Rognan, 2017; Lionta et al., 2014). A large number of other methods can be used to predict the binding modes with higher levels of accuracy. However, they can be more computationally demanding and require more skills from the users (Cournia et al., 2017; Pietrucci, 2017; Dickson, 2018). A summary of available tools for predicting the binding of ligands to enzymes is provided in Table 2. Molecular docking techniques are widely used in many different enzyme engineering strategies. Examples of their application can be found on the original pieces of research referring to a number of the cases presented in section 3.

Table 2

Examples of commonly used software tools for protein-ligand docking. All the tools herein referred are stand-alone applications.

Software	Principle	Availability	URL	Reference
AutoDock	Rigid and limited flexibility	Free for non-commercial use	http://autodock.scripps.edu/	(Morris et al., 2009)
Autodock Vina	Rigid and limited flexibility	Free for non-commercial use	http://vina.scripps.edu/	(Trott and Olson, 2009)
GOLD	Rigid and flexible docking	Commercial software	https://www.cc-dc.cam.ac.uk/solutions/csd-discovers/compounds/gold/	(Jones et al., 1997; Verdonk et al., 2003)
FlexX	Flexible docking	Commercial software	https://www.biosolveit.de/FlexX/	(Rarey et al., 1996)
DOCK	Rigid and flexible docking	Free for academic use	http://dock.compbio.ucsf.edu/	(Kuntz et al., 1982)
ICM	Flexible docking and virtual screening	Commercial software	http://www.molsoft.com/doc/king.html	(Abagyan et al., 1994)
QuickVina-W	Rigid, blind docking	Free for non-commercial use	http://www.qvina.org/	(Hassan et al., 2017)
Glide	Rigid and flexible docking	Commercial software	https://www.schrodinger.com/Glide/	(Friesner et al., 2004)

2.3. Access pathways and ligand transport

Proteins usually form highly complex three-dimensional structures with many internal grooves, protrusions, clefts, and voids (Calland, 2003). Even though the majority of such empty spaces have no biological role, they form functionally important substructures in enzymes such as active and allosteric sites, tunnels, and channels (Ringe and Petsko, 2008; Nussinov and Tsai, 2013; Gora et al., 2013). Enzymes with a buried active site often require tunnels that allow the transit of substrates, inhibitors, co-factors, products, and solvent molecules from the outside environment to the active site. Tunnels can also connect two distinct active sites (Kingsley and Lill, 2015; Marques et al., 2016; Raushel et al., 2003). Channels are found mainly in transmembrane proteins, allowing for the transport of small molecules, ions, and water solvent through biomembranes (Saier, 2000). Geometrical, physicochemical, and dynamical properties of tunnels and channels significantly influence the function of enzymes, making these structural features interesting targets for engineering (Gora et al., 2013; Dalby, 2007; Prokop et al., 2012). Indeed, key characteristics of enzymes such as substrate specificity, activity, enantioselectivity, and thermostability can be altered by modifying the tunnels and channels (Biedermannova et al., 2012; Kaushik et al., 2017; Brezovsky et al., 2016; Liskova et al., 2015; Bayley and Jayasinghe, 2004; Chaloupkova et al., 2003). Tunnels and channels are usually recognized as continuous empty spaces within the Voronoi diagram of the protein, which is a geometrical approximation to its Van der Waals volume (Okabe et al., 2000). CAVER (Chovancova et al., 2012), MolAxis (Yaffe et al., 2008), BetaCavityWeb (Kim et al., 2015), PoreWalker (Pellegrini-Calace et al., 2009), Mole (Sehnal et al., 2013), and ChExVis (Bin et al., 2015) are amongst the most commonly used tools for channel detection. These methods are easy to use and yield high-quality results (Brezovsky et al., 2013), including the coordinates of the located pathways, their geometrical properties like length and bottleneck radius, and an estimate of their importance.

Another family of methods is aimed to analyse trajectories of ligand transportation. This problem can be either addressed using molecular dynamics simulations or faster approaches such as iterative docking and robotic algorithms (Kokkonen et al., 2019). Methods based on molecular

dynamics usually use different enhanced sampling techniques to simulate the transport process in a reasonable time scale. These include *steered molecular dynamics* (Do et al., 2018), *metadynamics* (Furini and Domene, 1858), umbrella sampling (Zhang and Voth, 2011) and adaptive sampling (Marques et al., 2019), and are succinctly described in section 2.4. *Random accelerated molecular dynamics* is a method specifically developed to study ligand transport that applies a force to the ligand in a random direction and evaluates its position. The method allows more extensive sampling of the conformational space while still being effectively unbiased (Kokh et al., 2018; Lüdemann et al., 1997; Lüdemann et al., 2000), and has been implemented in various molecular dynamics packages, including AMBER (Case et al., 2018), GROMACS (Berendsen et al., 1995) and NAMD (Phillips et al., 2005). Despite the high quality of their results, the methods based on molecular dynamics are time demanding and require expert knowledge (Gelpi et al., 2015).

Robotic algorithms iteratively construct an exploration tree to rapidly find a feasible path between two states: bound and ligand-free. A representative of this family is MoMA-LigPath (Devaurs et al., 2013), implementing a Manhattan-like Rapidly-exploring Random Tree algorithm (LaValle and Kuffner, 2001). The method does not provide any energetic evaluation of the process. SLITHER (Lee et al., 2009) and CaverDock (Filipovic et al., 2019; Vavra et al., 2019) are iterative docking methods. The former implements different docking algorithms such as AutoDock3 (Morris et al., 1998), AutoDock4 (Morris et al., 2009), or MEDock (Chang et al., 2005) in an iterative scheme. The docking step is combined with a puddle-skimming procedure that repeatedly elevates the potential energies of identified global minima to determine the binding modes of the ligand inside the protein. CaverDock divides the tunnel into discrete slices and iteratively docks the ligand to each of them using AutoDock Vina (Trott and Olson, 2009). Restraints can be enforced in the process, both spatial (lower-bound) and rotational (upper bound), where the second guarantees a continuous motion of the ligand during the docking process. The two tasks described in this section, (i) the detection of access pathways and (ii) the analysis of the ligand transport, are typically performed separately, and the most popular tools for addressing these tasks are summarized in Table 3. To overcome this limitation, Caver Web (Stourac et al., 2019) has been recently developed to integrate CAVER and CaverDock in one single interactive user interface, providing the community with an intuitive and easy-to-use web platform to perform both tasks at once (Fig. 2). A successful application of access pathways and ligand transport analysis is presented in the haloalkane dehalogenase LinB example on section 3.5.

2.4. Protein dynamics

All biomolecules are dynamic entities, and the dynamic properties of enzymes are often important determinants of their biological function (Henzler-Wildman and Kern, 2007; Bahar et al., 2010; Petrovic et al., 2018). Therefore, it comes with no surprise that the study of protein dynamics is increasingly considered a standard requirement in the majority of biomolecular research.

Molecular dynamics (MD) is the most common computational method to study protein dynamics (Karplus and McCammon, 2002; Adcock and McCammon, 2006; Kumari et al., 2017). The method emerged around from the works by Levitt and Warshel (Levitt and Warshel, 1975; Warshel, 1976), while Karplus and co-workers announced the first simulation of a biologically relevant macromolecule (McCammon et al., 1977). These three authors were awarded the Nobel Prize in Chemistry in 2013 for their pioneering work on “the development of multi-scale models for complex chemical systems”. In simple terms, molecular dynamics consists of numerically solving the classical equations of motion, based on Newton's law, $F = m \cdot a$. These equations are applied to a number of hard spheres connected by springs, which are used as an approximation to the atoms and the bonds connecting them. The forces involved are described by a set of potential energy functions –the *force field* (Lazaridis

Table 3

Examples of commonly used tools for identification of access pathways and analyses of ligand transport.

Software	Presentation	Principle	Availability	URL	Reference
HTMD/AceMD	Standalone	Molecular dynamics	Paid License	https://www.htmd.org/	(Harvey et al., 2009)
MoMA-LigPath	Webserver, standalone on request	Robotic	Free	https://moma.laas.fr/	(Devaurs et al., 2013)
SLITHER	Webserver	Iterative docking	Free	http://slither.rcas.sinica.edu.tw/	(Lee et al., 2009)
CaverDock	Standalone	Iterative docking	Free for non-commercial use	https://loschmidt.chemi.muni.cz/caverdock/	(Filipovic et al., 2019; Vavra et al., 2019)
Caver Web	Webserver	Hybrid	Free for non-commercial use	https://loschmidt.chemi.muni.cz/caverweb/	(Stourac et al., 2019)
CAVER	Standalone, PyMol plugin	Path search in Voronoi diagram	Free	https://www.caver.cz	(Chovancova et al., 2012)
Mole	Standalone	Path search in Voronoi diagram	Free	https://mole.upol.cz/	(Sehnal et al., 2013)
MolAxis	Webserver, standalone	Path search in Voronoi diagram	Free for non-commercial use	http://bioinfo3d.cs.tau.ac.il/MolAxis/	(Yaffe et al., 2008)
BetaCavityWeb	Webserver	Path search in Voronoi diagram	Free	http://voronoi.hanyang.ac.kr/betacavityweb	(Kim et al., 2015)
PoreWalker	Webserver	Analysis of channel formed by protein main axis	Free	https://www.ebi.ac.uk/thornton-srv/software/PoreWalker/	(Pellegrini-Calace et al., 2009)
ChExVis	Webserver	Path search in Voronoi diagram	Free	http://vgl.serc.iisc.ernet.in/chexvis/	(Bin et al., 2015)

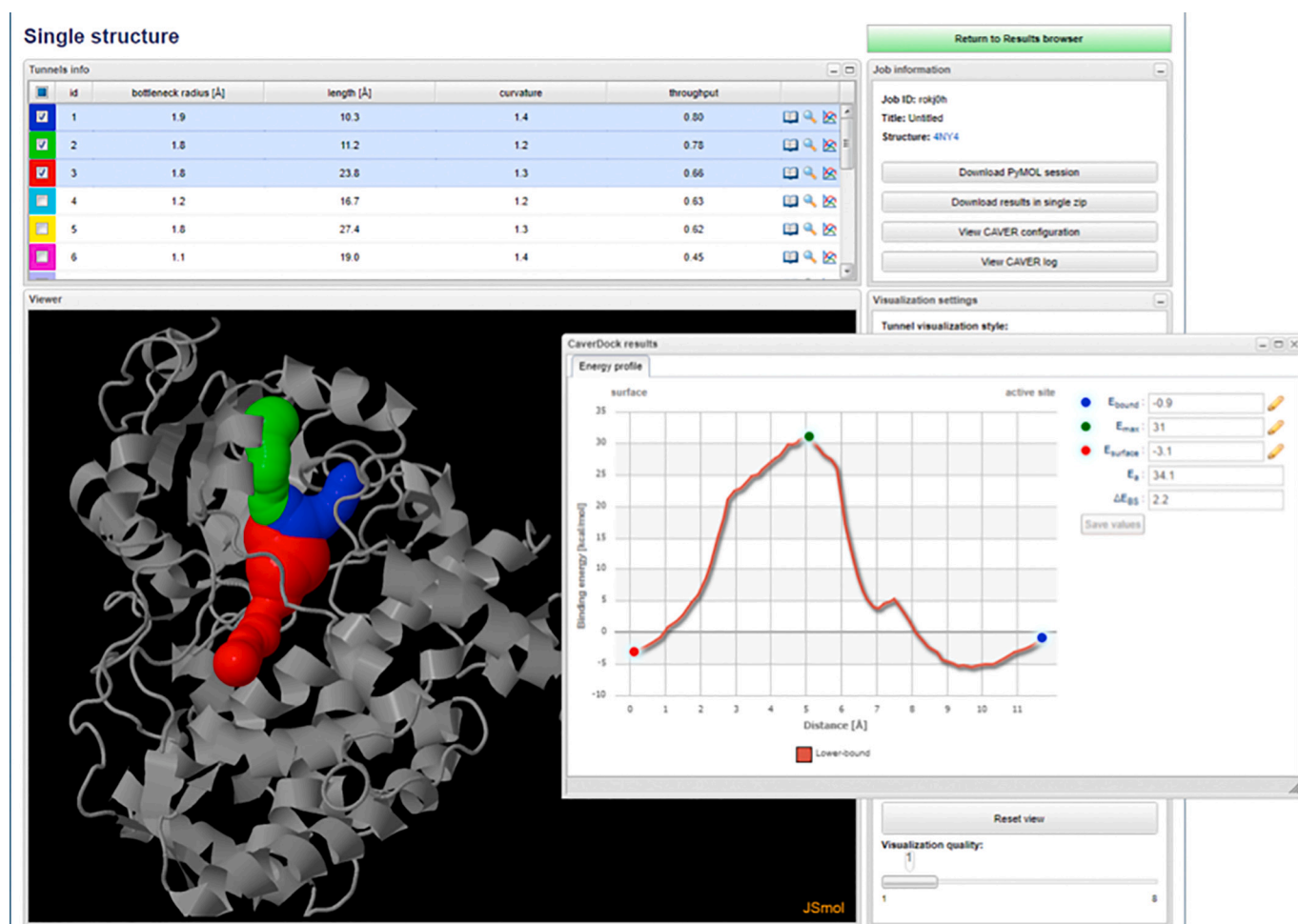


Fig. 2. Caver Web graphical user interface. Screenshot of a graphical user interface of Caver Web showing the results for the CYP3A4 protein (PDB ID 4NY4). After running the analysis, the page shows in the lower-left corner the visualization panel that allows for the interactive inspection of the detected tunnels on the original protein structure using JSmol Viewer. On the upper left corner, all the detected tunnels are listed on the 'Tunnels info' section. The most important geometrical properties of the listed tunnels are provided in this section and also several buttons that allow for zooming in each tunnel on the visualization panel (bottom left) and for opening popup windows with detailed information of the tunnels and their profiles. On the upper right section the 'Job information' panel displays the unique identifier for the job, a custom job title, the name of the structure. On the same panel, different buttons allow the user to download a pre-generated PyMOL session with the results, the raw data in zip format, or to show a popup window with the configuration parameters used during CAVER calculation and its corresponding log file. The displayed pop-up window (titled 'CaverDock results') shows the lower-bound energy profile calculated using CaverDock, coloured points in the graph define the energy of the bound state (blue, E_{bound}), the energy on the surface (red, E_{surface}) and the energy maximum (green, E_{max}). The activation energy (E_a) and the energy difference between the bound and surface states are automatically calculated.

and Karplus, 2000), and are assumed to accurately govern the evolution of the entire system as a function of time (Nick Pace et al., 2014; Vlachakis et al., 2014; Paquet and Viktor, 2015). The results produced can be used to estimate the configurational distribution in equilibrium for each atom of the system, to study a particular transient process such as the binding/unbinding of a ligand, or the opening/closing of a molecular gate. The most popular force fields for the simulation of the dynamics of proteins (Adcock and McCammon, 2006; Paquet and Viktor, 2015) are AMBER, CHARMM, GROMOS and OPLS (Vlachakis et al., 2014; Monticelli and Tieleman, 2013). Nevertheless, classical molecular dynamics is not devoid of limitations: i) feasible time scales are in practical terms narrow, currently circumscribed to milliseconds; ii) force fields are approximate and commonly lack polarization; iii) the solvent model consists of a simplistic approximation to the real water molecules, and iv) the models neither consider cleavage nor formation of bonds that are essential in chemical reactions. These issues have been tackled by different approaches and methods aimed at improving the sampling accuracy namely with the development of hybrid quantum mechanics/molecular mechanics methods (Kumari et al., 2017; Maximova et al., 2016).

To reduce the computational cost of molecular dynamics, parallel calculations in graphics processing unit computing technology have accelerated this process by orders of magnitude (Stone et al., 2007). Also, different methods have been developed to enable the enhanced sampling of events to longer time scales, including metadynamics, replica-exchange MD, umbrella sampling, steered MD, accelerated MD, and adaptive sampling (Bernardi et al., 1950; Spiwok et al., 2015). *Metadynamics* is informally described as “filling the free energy wells with computational sand” (Darve and Pohorille, 2001) because the conformations already visited during the simulation are restricted by a positive Gaussian potential (the computational sand) added to the real energy landscape (the energy wells). The system is thus encouraged to escape the local energy minima and explore other parts of the energy landscape. The advantage of this method is that it does not require any initial estimate of the energy landscape and that the real energy landscape can be fully recovered. However, it requires selecting one or more of the so-called collective variables to guide the conformational sampling, which may be a very difficult task in the case of more complex systems. The choice of the collective variables is critical, for instance, for obtaining a correct estimation of kinetic rates of molecular events due to slow convergence issues, although new approaches are being developed to tackle this problem (Furini and Domene, 1958; Zhang and Voth, 2011; Bernardi et al., 1950; Sultan and Pande, 2017). *Replica-exchange molecular dynamics* consists of running multiple replicas of independent simulations at different temperatures, which are periodically exchanged in a way to ensure the proper canonical ensembles. This facilitates the overcoming of energy barriers and a wider exploration of the conformational space. It has to be noted that this methodology is resource-intensive and that the sampling of energetic barriers may lack accuracy due to difficulties in correctly sampling the canonical ensemble (Swendsen and Wang, 1986; Sugita and Okamoto, 1999).

Umbrella sampling (Zhang and Voth, 2011) divides the event path into small windows and applies a bias (umbrella) potential function to restrain the system within each window. The event path is defined by a variable such as distance, angle, or root-mean-square deviation. The bias is aimed to counteract high energetic barriers and let the ligand pass through them. A part of the system, typically a ligand, is pulled by a constant velocity or force towards a predefined direction in *steered molecular dynamics*. The system is moved step by step along with a predefined metric, e.g., distance, centre of masses, or angle, and is restrained by a harmonic potential. The measured force is calculated over time and then presented as a diagram showing the force distribution at different atomic states of the binding (Do et al., 2018; Chen, 2015; Skovstrup et al., 2012). *Accelerated molecular dynamics* adds a boost of potential energy to the system that is greater for lower values of the real energy. This means that the potential energy landscape becomes

flattened and the energy barriers are decreased, leading to more conformational transitions. An advantage of this method is that the bias does not depend on a user-defined parameter (Hamelberg et al., 2004). *Adaptive sampling molecular dynamics* (Marques et al., 2019; Betz and Dror, 2019) allows the user to specify a metric to be sampled, e.g. distance, contacts, or root-mean-square deviation. A pre-defined number of parallel simulations are run in sequential batches (epochs), iteratively scheduled according to the analysis of previous epochs. Snapshots of previous epochs from conformational states with lower sampling are used preferentially to start the new simulations. This allows the simulation to quickly explore under-sampled regions, overcome energetic barriers and reach the desired conformation without applying extra force. Very often used in combination with adaptive sampling, *Markov state model* (MSM) analysis is increasingly used for assessing the relevant molecular states of a system and the inter-conversion rates between them. This robust method relies on the clustering of simulations according to a metric, or collective variable, that describe the process being studied, and can be used to identify slow events, kinetic rates, equilibria, affinities, and pathways (Chodera and Noé, 2014). New developments have recently implemented unsupervised machine learning-based MSM analysis of MDs, which can more easily select the optimal collective variable for the identification and study of relevant molecular processes (Mardt et al., 2018; Xie et al., 2019).

Available force fields are being constantly improved and updated. For instance, polarizability has been introduced to force fields, either implicitly or explicitly (Monticelli and Tieleman, 2013), bringing higher accuracy to the simulations at the expense of larger computation time. Many *water models* are available for explicitly considering the solvent with different degrees of approximation, such as TIP3P, TIP4P, TIP5P, SPC, and SPC/E (Adcock and McCammon, 2006). TIP3P is one of the most commonly used models despite its simplicity. Since the solvent model can indirectly influence the dynamics of the studied enzyme, the choice of both force field and solvent model is extremely important but it is not trivial. The optimal combination can very much depend on the purpose of the study and on how they have demonstrated to perform in a specific context (Zeng et al., 2016; Zhang et al., 2018; Palazzesi et al., 2016). Finally, quantum mechanics/molecular mechanics methods can be coupled to molecular dynamics (section 2.1) to precisely describe catalytic events occurring in the most important regions of the enzyme, allowing for the study of chemical processes with a great level of detail (Ferrer et al., 2011; Chang et al., 2016).

Monte Carlo simulations can also be used to assess the structural and thermodynamic properties of macromolecules in equilibrium (Adcock and McCammon, 2006; Maximova et al., 2016). Monte Carlo is a stochastic method that sequentially applies random perturbations on the system to sample the configurational space based on the potential energy. At each step, the probability of accepting the newly generate configuration is dictated by the Metropolis criterion (Metropolis et al., 1953). According to this algorithm, the generated frame is evaluated and accepted if the potential energy decreases. Otherwise, if the energy increases, the probability of accepting such a frame is given by the respective Boltzmann factor $e^{-\Delta V/k_B T}$, where ΔV is the potential energy difference between the new and previous conformations, k_B is the Boltzmann constant, and T the temperature (Paquet and Viktor, 2015; Maximova et al., 2016; Earl and Deem, 2008). In a well-converged simulation, the ensemble of accepted conformations describes properly Boltzmann-weighted averages for structure and thermodynamic properties of the system. Such outcome enables the possibility of determining the thermodynamic equilibrium, the different microstates, the probability of a given conformation and its physical properties. Monte Carlo simulations accelerate the sampling of the conformational space since they do not need to explicitly solve Newton's equations of motion (Earl and Deem, 2008). However, the probability of atomic collisions after long movements increases with the system size, leading to reject many conformations due to their high energy. This can result in the inefficient exploration of the conformational space in the case of

large systems such as proteins. Moreover, Monte Carlo lacks the time component and thus the ability to calculate the kinetic properties of the reaction. Finally, these simulations also present the risk of stalling the system in local minima. Many methods have been developed to overcome these limitations, such as the *configurational bias Monte Carlo* (Deem and Bader, 1996), *parallel tempering Monte Carlo* (Geyer and Thompson, 1995), *density of states* (Wang and Landau, 2001), or even *hybrid Monte Carlo/molecular dynamics* (Duane et al., 1987; Izaguirre and Hampton, 2004).

Compared to molecular dynamics, *normal mode analysis* on elastic network models represents a cruder approach to enzyme dynamics. Normal mode analysis entails a simplification of the molecular dynamics principles at two different levels. The granularity of the model can be coarser in normal mode analysis. For instance, elastic network models usually represent whole amino acids as beads, whereas molecular dynamics uses individual atoms as simulating units instead. The potential functions that these methods implement – *harmonic potentials and derivatives* – match the complexity of the protein representation and are thus less precise than the mechanistic functions implemented in molecular dynamics. The approach is adequate for predicting collective motions in proteins (Tirion, 1996), usually being only the slower ones (corresponding to smaller eigenvectors) informative, and is commonly implemented using the *Gaussian* or the *anisotropic network models* (Bahar et al., 1997; Atilgan et al., 2001). Similarities and differences between normal mode analysis and molecular dynamics have been previously described (Rueda et al., 2007), leading to the conclusion that normal mode analysis can be regarded as a cheap alternative to molecular dynamics for very large systems. A summary of available tools for the prediction of protein dynamics is provided in Table 4.

The critical importance of protein dynamics to biocatalysis has become increasingly recognized, namely with the recent understanding that the intrinsic dynamic properties have evolved with the enzyme function (Maria-Solano et al., 2018; Campbell et al., 2018; Gardner et al., 2020). For this reason, any robust computational enzyme design exercise can greatly benefit from the assessment of dynamical ensembles and the impact of mutations in such ensembles. Examples of flexibility and dynamical modifications that led to improved enzymes are presented in section 3.5.

2.5. Hotspot identification and smart library design

Directed evolution is an effective tool for improving protein properties with minimal knowledge of the studied system (Romero and Arnold, 2009; Currin et al., 2015; Cheng et al., 2015; Bornscheuer et al., 2019). Repeated rounds of random mutagenesis are used to construct libraries of variants to cover a large mutational landscape. These libraries are screened to select variants with an improved target property. Such screenings are laborious and yet not always available for very large libraries. To cope with these hindrances, rational design was introduced as a reasonably successful strategy to design site-specific mutations that improve the designed property (Bednar et al., 2015; Goldenzweig et al., 2016; Zhou et al., 2019; Sinha and Shukla, 2019). However, a vast knowledge of the studied system and prior experimental information is necessary for correct identification of the suitable mutations. By combining aspects from rational design and directed evolution, semi-rational design takes advantage of both approaches (Sinha and Shukla, 2019; Verma et al., 2012) and leads to the construction of “smart-but-small libraries” containing a smaller number of variants to be screened. Commonly used tools implementing all these strategies are summarised in Table 5.

The appropriate strategy for hot spots identification greatly depends on the targeted property. Stability hot spots can be predicted with B-fitter (Reetz and Carballera, 2007). The method relies on the crystallographic structure of the enzyme with known flexible regions, which can be randomized by experimental mutagenesis. Molecular Dynamics can also be used to identify the most flexible positions (Osuna, 2020;

Table 4

Examples of commonly used tools for the simulation of protein dynamics. All the tools herein referred are stand-alone applications, except when stated otherwise.

Method	Principle	Software	URL	Reference
Metadynamics	Molecular Dynamics	AMBER, GROMACS, NAMD, etc.	http://amb-ermd.org/ http://www.gromacs.org/ http://www.ks.uiuc.edu/Research/namd/	(Laio and Parrinello, 2002)
Replica exchange	Molecular Dynamics	AMBER, GROMACS, NAMD, etc.	http://amb-ermd.org/ http://www.gromacs.org/ http://www.ks.uiuc.edu/Research/namd/	(Sugita and Okamoto, 1999)
Umbrella sampling	Molecular Dynamics	AMBER, GROMACS, NAMD, etc.	http://amb-ermd.org/ http://www.gromacs.org/ http://www.ks.uiuc.edu/Research/namd/	(Torrie and Valleau, 1977)
Simulated annealing	Molecular Dynamics	AMBER, GROMACS, NAMD, etc.	http://amb-ermd.org/ http://www.gromacs.org/ http://www.ks.uiuc.edu/Research/namd/	(Tsallis and Stariolo, 1996)
Accelerated Molecular Dynamics	Molecular Dynamics	AMBER, GROMACS, NAMD, etc.	http://amb-ermd.org/ http://www.gromacs.org/ http://www.ks.uiuc.edu/Research/namd/	(Hamelberg et al., 2004)
Hybrid quantum mechanics/molecular mechanics methods	Quantum Mechanics, Molecular Mechanics, Molecular Dynamics	AMBER, GROMACS, NAMD, etc.	http://amb-ermd.org/ http://www.gromacs.org/ http://www.ks.uiuc.edu/Research/namd/	(Warshel and Levitt, 1976)
Configurational bias Monte Carlo	Monte Carlo	CMBC	Not applicable	(Deem and Bader, 1996)
Parallel tempering	Monte Carlo	AMBER, GROMACS, NAMD, etc.	http://amb-ermd.org/ http://www.gromacs.org/ http://www.ks.uiuc.edu/Research/namd/	(Geyer and Thompson, 1995)
Density of states	Monte Carlo	DOS	https://github.com/DavidAce/WL	(Wang and Landau, 2001)
Hybrid Monte Carlo/molecular	Monte Carlo, Molecular Dynamics	ProtoMol	http://proteinomol.sourceforge.net/	(Duane et al., 1987)

(continued on next page)

Table 4 (continued)

Method	Principle	Software	URL	Reference
dynamics methods				
Anisotropic Network Model	Normal Mode Analysis	Anisotropic Network Model ^a	http://anm.csb.pitt.edu/	(Eyal et al., 2015)
Gaussian Network Model	Normal Mode Analysis	DynOmics ^a	http://enm.pitt.edu/	(Li et al., 2017)
Cartesian pseudo-trajectories	Normal Mode Analysis	FlexServ ^a	http://mmb.pcb.ub.es/FlexServ/	(Camps et al., 2009)

^a These tools are provided as web servers

Table 5

Examples of commonly used strategies for hotspots identification.

Strategy	Residues Targeted	Effect	Targeted Property
B-fitting	Residues with high B-factors	Rigidification	Stability
Proline rule	Proline substitutions in β -turns	Rigidification	Stability
Filling voids	Around cavities and tunnels	The increasing number of interactions	Stability
Tunnel bottleneck	The narrowest part of a tunnel	Modifying tunnel throughput	Ligand transport, activity, stability, selectivity
Molecular complementarity	First shell residues	Changing orientation or affinity of a ligand	Catalysis, selectivity
Druggability	First shell residues	Changing orientation or affinity of a ligand	Catalysis, selectivity
Evolutionary variable position	Non-conserved residues in alignments	Identification of mutable positions	Catalysis, selectivity, stability

Dauber-Osguthorpe et al., 1999; Liao et al., 2019). Another approach to improve stability is based on filling existing voids in the protein (Borgo and Havranek, 2012), which normally correspond to the access pathways to the catalytic site. Not only stability but also other properties (ligand transport, specificity, selectivity, or catalysis) are often influenced by the amino acids that directly interact with the ligand during specific phases of the catalytic cycle. These residues can be identified by molecular dynamics and the enhanced sampling variants (Ebert et al., 2017; Marques et al., 2019; Curado-Carballada et al., 2019), molecular docking (Xie and Hwang, 2015; Carlson et al., 2016) coupled to cavity detection (Le Guilloux et al., 2009; Huang, 2009) or druggability (Volkamer et al., 2012; Yuan et al., 2013) analyses (section 2.2), or as the residues lining protein tunnels (Chovancova et al., 2012; Sehnal et al., 2013; Jurcik et al., 2018) (section 2.3).

Phylogenetic analysis is the most common approach for selecting hot spots. Multiple sequence alignments of homologous proteins are often used to identify functionally important residues in the studied enzyme (Pei, 2008; Ashkenazy et al., 2016). On this context, highly conserved positions correlate with residues presumably relevant for activity, stability, or folding of the enzyme; in contrast, variable positions represent safe mutagenesis hot spots. The identified a pool of evolutionary allowed substitutions can be used to design and construct smart libraries of codons encoding for the desired protein variants (Reetz and Wu, 2008; Jochens and Bornscheuer, 2010). The size of such libraries can be further reduced by using a restricted codon alphabet that exclusively translates into the requested set and frequency of amino acids (Pines et al., 2015), or using specifically designed sets to improve particular

properties (Goldsmith and Tawfik, 2013). Using this strategy, the introduction of potential stop codons should be ideally kept at reasonably low frequencies and the amino acids redundancy minimised (Gaytán et al., 2009; Nov, 2014). As a result, a proper smart library of degenerated codons encoding the desired frequencies and properties for all selected amino acids is obtained.

The whole process of hot spot identification and smart library construction can be tedious. HotSpot Wizard (Sumbalova et al., 2018) simplifies the process by gathering a large number of methods and databases under a single web portal. HotSpot Wizard requires a three-dimensional structure of the studied enzyme as input. If the structure is not available, HotSpot Wizard can prepare a homology model from the protein sequence. In the first place, a set of homologous sequences is obtained and used to build a multiple-sequence alignment to single out conserved and mutable residues. Then, the catalytic residues are identified and structural analyses of pockets and access tunnels are performed to detect any residue in the catalytic pocket or lining transport pathways. Finally, stability hot spots from B-fitter and consensus mutations are identified and statistical analysis of evolutionary correlated positions is performed. Once identified, mutations on the hotspots can then be further assessed using Rosetta (Kellogg et al., 2011) to evaluate the stability of selected substitutions. Also, molecular docking by AutoDock Vina (Trott and Olson, 2009) can be used to identify residues in direct contact with a selected ligand. Taking advantage of these analyses, HotSpot Wizard allows for the selection of identified substitutions for individual hotspots and designs suitable degenerated codons, all within one graphical user interface. The Hotspot Wizard workflow is summarized in Fig. 3. Details on successful examples of hotspot identification can be found in the references herein provided (i. e. 217) or in the haloalkane dehalogenase LinB example in section 3.5.

2.6. Mutagenesis

Design of mutations is one of the main tasks of protein engineering. Fast and accurate *in silico* methods have been developed to tackle this problem, and can be divided into three categories: (i) *machine learning*, (ii) *force-field calculations*, and (iii) *phylogenetic analyses*. These tools are normally aimed to predict the effect of mutations on protein stability or solubility (section 3.4) and the most commonly used tools are summarized in Tables 6–7. Different examples of practical applications of these techniques are presented in section 3.4. Machine learning exploits statistical techniques to grasp hidden patterns in large sets of unstructured data (Bishop, 2006). Machine learning models are trained and validated using independent experimental data and hold the potential to elicit any possible feature. Through this process, the explored features are weighted according to their relevance for the prediction outcome. Machine learning does not require an understanding of the mechanistic principles underlying the enzymatic catalysis, conforming a remarkably useful approach for the evaluation of the effect of introduced mutations on the target enzyme (Magnan et al., 2009; Ozen et al., 2009). However, the reliability of machine learning approaches is strongly dependent on the size, balance, and quality of the experimental data (Kotsiantis et al., 2005). Predictors trained on small sets lead to higher risks of being over-trained, prone to ignore the actual diversity existing in the population the training set represents (Pucci et al., 2018). Machine learning methods are also notably sensible to data unbalance (Pucci et al., 2018; Chawla and Japkowicz, 2004). Nevertheless, some methods such as support vector machines: EASE-MM (Folkman et al., 2016), MuStab (Teng et al., 2010), I-Mutant (Capriotti et al., 2005), and MuPro (Cheng et al., 2006); and random forest classifiers: ProMaya (Wainreb et al., 2011) and PROTS-RF (Li and Fang, 2012); are relatively robust. However, the risk of overtraining is prevalent in all machine learning techniques (Breiman, 2001; Liaw and Wiener, 2002; Boughorbel et al., 2017). To cope with this liability, several machine learning approaches can be combined into a single system to increase the precision of the final predictions. For instance, MAESTRO (Laimer et al., 2015) combines

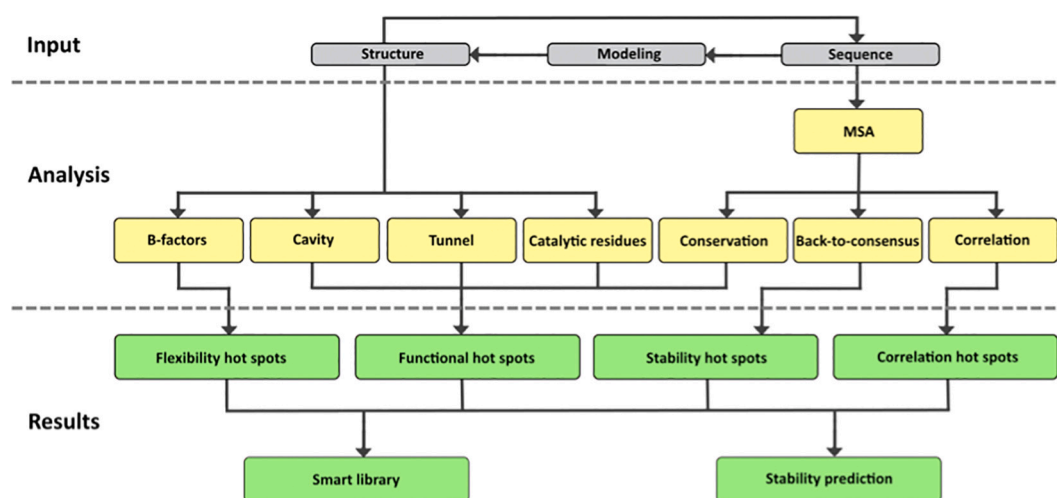


Fig. 3. General workflow of HotSpot Wizard. Input files are highlighted in grey, individual analyses in yellow and results on the output in green (Bendl et al., 2016).

support vector machines, multiple linear regression and statistical potentials.

Force field methods rely on energy models built on top of the known laws of physics. Over the last 20 years, different approximations in the form of *physical*, *empirical*, and *statistical effective energy* functions have been developed (Guerois et al., 2002; Mendes et al., 2002). Physical effective energy functions are closely related to classical molecular mechanics force fields (MacKerell et al., 1998; Oostenbrink et al., 2004) and, despite overestimating hydrophobicity (Musil et al., 2017; Tokuriki et al., 2008; Arabnejad et al., 2017), allow for a fundamental analysis of molecular interactions. Hence, physical effective energy functions represent accurate and versatile methods for predicting the structural changes upon introduction of mutations under non-standard conditions. Representative tools implementing physical effective energy functions are Rosetta (Kellogg et al., 2011), ERIS (Yin et al., 2007) and CC/PBSA (Benedix et al., 2009). Statistical effective energy functions parameterize molecular interactions based on curated data sets of folded protein structures and provide an overall energy function of a system (Dehouck et al., 2006; Liu, 2015). Empirical effective energy functions combine both physical and statistical terms, providing a reasonable compromise between computational cost and accuracy. The distinction between statistical and empirical effective energy functions is vague, and collectively include methods such as PoPMuSiC (Dehouck et al., 2011), FoldX (Schymkowitz et al., 2005) and DMutant (Hoppe and Schomburg, 2005). Despite being conveniently rapid, the applicability of these methods is restricted to the conditions under which the experimental data used for their parameterization were acquired (Kepp, 1854; Christensen and Kepp, 2012).

Phylogenetic analyses aim to extract information from the available evolutionary history of the studied enzyme to infer the effects potentially introduced by mutations. *Consensus design* and *ancestral sequence reconstruction* are the two most widely used phylogenetic approaches in protein engineering. Both methods are based on the identification of conserved positions in multiple-sequence alignments of homologous proteins (described in section 2.5). Consensus design was conceived to identify deviations at conserved positions in the alignment and revert such mutations to the consensus state to agree with the rest of the aligned sequences (Amin et al., 2004; Sullivan et al., 2012). Despite being unavailable as stand-alone tools, consensus design as implemented in EMBOSS (Rice et al., 2000), 3DM (Kuipers et al., 2010), and Vector NTI (Lu and Moriyama, 2004) can be used as a filter during core calculations of hybrid workflows or as a component of predictive tools for hotspot identification such as HotSpot Wizard (Quan et al., 2016). Ancestral sequence reconstruction derives a phylogenetic tree from the

multiple-sequence alignment to infer the sequence of the studied protein in a theoretical ancestor organism. This inference can be either Bayesian (Huelsenbeck et al., 2001): HandAlign (Westesson et al., 2012), MrBayes (Ronquist et al., 2012), and BALi-Phy (Suchard and Redelings, 2006); or obtained by maximum-likelihood (Xu and Yang, 2013; Stamatakis, 2006): RAXML (Stamatakis, 2014), FastML (Ashkenazy et al., 2012), and Ancestors (Diallo et al., 2010). Ancestral sequence reconstruction is a complex multi-step technique that requires in-depth knowledge of the biological system of interest. The aforementioned tools only provide the means to deal with the last steps of the procedure while earlier ones (such a selection of relevant homologs, the construction of a reliable multiple sequence alignment or rooting the evolutionary tree) are left to the user to deal with. FireProt^{ASR} (Musil et al., 2020), is a maximum-likelihood tool that was explicitly designed to address those issues, providing with a fully automated workflow that also allows for the reconstruction of ancestral gaps, a feature not available in most of the existing tools. The method is implemented in a web-server that requires only the sequence of the protein of interest as input and presents the inferred ancestral sequences in an interactive interface. Both consensus design and ancestral sequence reconstruction rely on the assumption of the thermophilic origin of primordial life, expecting the consensus/ancestral sequence to be more stable than the modern protein (Pey et al., 2008).

3. Applications of computational design of biocatalysts

3.1. Design of efficient biocatalysts

The well-established Michaelis-Menten kinetic model describes the catalytic efficiency of an enzyme as the ratio k_{cat}/K_M . In this model, k_{cat} represents the catalytic rate, whereas K_M is the Michaelis constant defining the amount of substrate needed to obtain half of the maximum reaction rate. In computational terms, the catalytic efficiency of enzymes is normally associated with the free energy barrier, a commonly used expression to refer to the energy difference between the transition state and the enzyme-substrate complex (Hammes-Schiffer, 2013). A deep understanding of the catalytic mechanism by means of computational methods must be obtained prior to engineering catalytic efficiency (Fasan et al., 2008; Zhang and Klinman, 2011; Miton et al., 2018; Wang et al., 2018; Korendovych and DeGrado, 2014; Wijma and Janssen, 2013). Russo and collaborators (Alberto et al., 2015) used a cluster model to represent the active site of the organophosphate degrading enzyme from *Agrobacterium radiobacter* P230. The enzyme has two cobalt atoms in the active site and catalyses the hydrolysis of a

Table 6
Examples of commonly used tools for predicting protein stability.

Software	Method	Input	Availability	Reference
EASE-MM	Support vector machine	Sequence	Web	(Folkman et al., 2016)
mCSM	Graph-based	Structure	Web	(Pires et al., 2014)
ELASPIC	Support vector machine, hidden Markov model	Structure	Web	(Witvliet et al., 2016)
I-Mutant	Support vector machine	Sequence or structure	Web	(Capriotti et al., 2005)
MAESTRO	Artificial neural networks, support vector machine, multiple linear regression, statistical potentials	Structure	Stand alone, web	(Laimer et al., 2015)
Prethermut	Support vector machine, random forest classifier	Structure	Stand alone	(Tian et al., 2010a)
ProMaya	Random forest classifier	Structure	Web	(Wainreb et al., 2011)
Iptree-stab	Decision tree	Partial sequence	Web	(Huang et al., 2007)
Rosetta	Physical effective energy function	Structure	Stand alone	(Kellogg et al., 2011)
ERIS	Physical effective energy function	Structure	Stand alone	(Yin et al., 2007)
CC/PBSA	Physical effective energy function	Structure	Stand alone	(Benedix et al., 2009)
CUPSAT	Atom potentials and torsion angles	Structure	Web	(Parthiban et al., 2006)
PopMuSiC	Statistical effective energy function	Structure	Web	(Dehouck et al., 2011)
DMutant	Amino acid potentials and torsion angles	Structure	Stand alone	(Hoppe and Schomburg, 2005)
FoldX	Statistical effective energy function	Structure	Web	(Schymkowitz et al., 2005)
STRUM	Statistical effective energy function	Structure	Stand alone, web	(Quan et al., 2016)
HotSpotWizard	Combined approach	Sequence and Structure	Web	(Bendl et al., 2016)
FastML	Maximum-likelihood	Multiple sequence alignment and	Web	(Ashkenazy et al., 2012)

Table 6 (continued)

Software	Method	Input	Availability	Reference
RAXML	Maximum-likelihood	phylogenetic tree Multiple sequence alignment	Stand alone and web	(Stamatakis, 2014)
Ancestors	Maximum-likelihood	Multiple sequence alignment and phylogenetic tree	Web	(Diallo et al., 2010)
HandAlign	Bayesian inference	Multiple sequence alignment and phylogenetic tree	Web	(Westesson et al., 2012)
PAML	Maximum-likelihood	Multiple sequence alignment and phylogenetic tree	Stand alone	(Yang, 2007)
PhyloBot	Maximum-likelihood	Multiple sequence alignment	Stand alone and web	(Hanson-Smith and Johnson, 2016)
MrBayes	Bayesian inference	Multiple sequence alignment	Stand alone	(Ronquist et al., 2012)
FireProt	Hybrid	Structure	Web	(Musil et al., 2017)
PROSS	Hybrid	Structure	Web	(Goldenzweig et al., 2016)
FRESCO	Hybrid	Structure	Stand alone	(Wijma et al., 2014)

phosphotriester into a phosphodiester, with a subsequent hydrolysis of the phosphodiester into a phosphomonoester. The authors substituted the two cobalt atoms in the active site by three other pairs of metal atoms (Pinto et al., 2017). Mimicking the natural availability of metallic atoms in the system, they observed significant changes in the free energy barriers. The enzyme showed better efficiency when two manganese atoms or a zinc-iron pair was present instead of the two naturally occurring metal atoms (Pinto et al., 2017).

Zahn and collaborators exploited molecular dynamics simulations and quantum mechanics/molecular mechanics to computationally design a highly efficient cocaine-detoxifying enzyme (Zheng et al., 2014). The engineering effort started identifying human butyrylcholinesterase as a promiscuous enzyme, which primarily catalyses the hydrolysis of acetylcholine, but is also able to process cocaine. The authors engineered a new variant E30-6 with a much higher catalytic efficiency for cocaine conversion, even when compared to the catalysis of the primary butyrylcholine substrate. It is important here to dissect how the efficiency became one order of magnitude higher in this enzyme for the hydrolysis of cocaine. Even though the turnover (k_{cat}) number is very similar, by building this new variant the affinity of the non-natural substrate is one order of magnitude higher. Although the chemical step was indeed improved, altering the physical steps by design turned to be the leading factor in the improvement of hydrolysis of cocaine when compared to the enzyme's natural substrate. Their protocol consists of three steps. First, the transition state of the rate-limiting step was modelled and the full protein minimized using molecular dynamics. Since the transition state is not a local or global minimum, the geometric parameters of the modelled transition state were constrained during the simulation. Then, the authors applied the hydrogen bonding energy equation from Autodock (Morris et al., 2009) to create a library of variants. The screened library had 67,000 possible variants, from single-point up to septuple-point mutants. Finally, hybrid quantum mechanics/molecular mechanics calculations were used to obtain the energies and geometries of the three states of the rate-limiting step on the most

Table 7
Examples of commonly used tools for predicting protein solubility.

Software	Method	Input	Availability	Reference
SolPro	Two-layer support vector machine	Sequence	Stand alone	(Magnan et al., 2009)
PROSO II	Logistic regression, Parzen window	Sequence	Web	(Smialowski et al., 2012)
ccSOL	Support vector machine	Sequence	Web	(Agostini et al., 2014)
DeepSol	Convolutional neural network	Sequence	Stand alone	(Khurana et al., 2018)
ESPRESSO	Support vector machine	Sequence, expression system	Web	(Hirose and Noguchi, 2013)
AGGRESKAN	Custom regression	Sequence	Web	(Conchillo-Solé et al., 2007)
TANGO	Custom regression and statistical potentials	Sequence	Stand alone and web	(Fernandez-Escamilla et al., 2004)
WALTZ	PSSM	Sequence	Web	(Maurer-Stroh et al., 2010)
PASTA	Custom regression and statistical potentials	Sequence	Web	(Walsh et al., 2014)
BETASCAN	Pairwise probabilistic analysis	Sequence	Stand alone and web	(Bryan et al., 2009)
FoldAmyloid	Custom regression and statistical potentials	Sequence	Web	(Garbuzynskiy et al., 2010)
OptSolMut	Linear programming	Structure	Stand alone	(Tian et al., 2010b)
CamSol	Custom regression	Sequence or structure	Web	(Sormanni et al., 2015)
AGGRESKAN3D	Custom regression	Structure	Web	(Zambrano et al., 2015)
SolubiS	Statistical and physical potentials	Structure	Stand alone and web	(Van Durme et al., 2016)
SODA	Custom regression	Sequence or structure	Web	(Paladin et al., 2017)
PON-SOL	Random forest classifier	Sequence	Web	(Yang et al., 2016)
SoluProt	Gradient boosting machines	Sequence	Web	(Hon et al., 2021)

promising variants. The most efficient enzyme had a 25-fold better efficiency for cocaine hydrolysis than for acetylcholine hydrolysis (Table 8).

3.2. Design of biocatalysts with novel activity

De novo protein design (Khare and Fleishman, 2013) is regarded as the opportunity to explore the universe of the “protein space” (Woelfson et al., 2015). *De novo* enzyme design aims to create a nature-inspired catalyst tailored to produce a desired compound of interest. Spanning from only a few modifications on the active site of an enzyme to the design of a new catalyst from scratch, there is an ever-growing number of computer-based strategies developed to pursue that goal (Tiwari et al., 2012; Kries et al., 2013; Świderek et al., 2015; Vaissier Welborn and Head-Gordon, 2018). Some software tools and web servers were developed for *de novo* design, such as DEZYMER (Hellings and Richards, 1991), ORBIT (Dahiyat and Mayo, 2008; Dahiyat and Mayo, 1997),

Rosetta (Zanghellini et al., 2006; Richter et al., 2011; Kiss et al., 2013), SABER (Nosrati and Houk, 2012), and EvoDesign (Mitra et al., 2013). The workflow for computationally designing a novel enzymatic activity typical comprises: i) building the active site models based on the chemical requirements, the so-called *theozymes*; ii) matching the model with a protein scaffold, normally searched over a database of known structures; iii) designing the active site pocket; iv) ranking and selecting different possible active sites; v) experimentally constructing and assessing the activity of several variants; and vi) additionally optimizing the active site. The computationally designed biocatalysts often do not perform very well, but they can be highly improved by successive rounds of directed evolution (Woelfson et al., 2015; Tiwari et al., 2012; Świderek et al., 2015; Vaissier Welborn and Head-Gordon, 2018). As happens with many natural enzymes, the *de novo* designed enzymes can have complex multi-step catalytic cycles. This means that after the chemical step has been optimized, one can hit a new rate-limiting step, such as the substrate binding or the product release. When this is the case, the engineering problem must be shifted to a different bottleneck, possibly with a different approach.

Historically, the earliest attempts at *de novo* design date back to 1991 by Hellings and Richards on thioredoxin to graft a copper-binding site into an unrelated protein scaffold (Hellings and Richards, 1991; Hellings et al., 1991). The authors devised the computational tool DEZYMER, that searched for backbone scaffolds and appropriate side chains to construct a ligand-binding site that matched the desired interaction patterns (Hellings and Richards, 1991). Unfortunately, the tool has never been released to the community and some other data presented in papers using DEZYMER were questioned (Check Hayden, 2008). In 2001, Bolon and Mayo in their pioneering work introduced a single histidine residue to the active site to confer esterase activity to the thioredoxin scaffold (Bolon and Mayo, 2001) using the ORBIT program (Dahiyat and Mayo, 2008; Dahiyat and Mayo, 1997). The target reaction was the histidine-mediated nucleophilic hydrolysis of *p*-nitrophenyl acetate into *p*-nitrophenol. Due to its high stability, thioredoxin was chosen once more as a template once devoid of its relevant catalytic activity. The high-energy state of a theoretical acyl-histidine complex was modelled. The hydrophobic solvent-accessible surface area of a substrate was used as a metric to evaluate recognition and to rank the different active-site designs. This protocol resulted in two variants, bearing two F12H+Y70A and three F12A+L17H+Y70A mutations, respectively, and both demonstrated hydrolysis of *p*-nitrophenyl acetate. The three-point mutant named PZD2 showed an enzyme-like behaviour (Table 8), and a 180-fold increase in k_{cat} as compared to the uncatalysed reaction ($k_{cat} = 4.6 \times 10^{-4} \text{ s}^{-1}$). A catalyst with such activity represents a good starting point for directed evolution.

Successful *de novo* design of active enzymes was established in 2008 with three major significant works: the Kemp elimination (Röthlisberger et al., 2008), Retro-Aldase reaction (Jiang et al., 2008), and Diels-Alder reaction (Siegel et al., 2010). Diels-Alder reactions involve the stereoselective cyclo-addition of two substrates implicating the formation of two new carbon-carbon bonds and rarely occur in nature (Gao et al., 2020; Jamieson et al., 2019). Bimolecular bond-forming reactions are very challenging because both substrates must be properly oriented in the active site for the reaction to proceed. The first step was to decide upon the reaction mechanism to use and construct the ideal active site around the reacting molecules to stabilize the transition state. Quantum mechanics calculations were fundamental to two important tasks for engineering the active site. First, to determine the factors influencing the energy gap between the highest and lowest occupied molecular orbitals of the chemical species involved in the transition state. Second, to test different scenarios, such as the proper placing of H-bond donors and acceptors. Rosetta and quantum mechanics calculations were followed by RosettaMatch (Zanghellini et al., 2006) to search for backbone scaffolds that would allow the required geometry of the designed active sites. Each match was then optimized with RosettaDesign (Liu and Kuhlman, 2006) and filtered based on the geometric and energetic

Table 8

Examples of computationally designed biocatalysts with listed methods, tools, level of improvement and introduced mutations.

Enzyme	Method	Tools	Improvement	Mutations	Reference
Butyryl cholinesterase	Molecular dynamics, quantum mechanics/ molecular mechanics	AMBER, Gaussian	catalysis (k_{cat}) improved 25-fold	6 mutations	(Zheng et al., 2014)
Protozyme esterase	<i>De novo</i>	ORBIT	catalysis (k_{cat}/k_{uncat}) improved 180-fold	3 mutations	(Bolon and Mayo, 2001)
Diels-Alderase	Molecular dynamics, quantum mechanics, <i>de novo</i>	RosettaMatch, RosettaDesign	catalysis (k_{cat}/k_{uncat}) improved 89-fold	13+6 mutations	(Siegel et al., 2010)
Guanine deaminase	Loop re-modelling	MiniRosetta fold tree	activity improved 1000000-fold towards ammelide	7 residues replaced	(Murphy et al., 2009)
TesA thioesterase	Rigid body docking and energy minimisation	Iterative protein redesign and optimization	production of C8 fatty acids improved 10-fold	8 mutations	(Grisewood et al., 2017)
Limonene epoxide hydrolase	Docking and molecular dynamics simulations	Rosetta Design	enantiomeric preference 93%	7 mutations	(Wijma et al., 2015)
α -Galactosidase	Force-field, machine learning	Solubis, TANGO, FoldX	solubility increased by 40 %	3 mutations	(Ganesan et al., 2016)
Adenosine kinase	Ancestral sequence reconstruction	BALiPhy	thermostability (T_m) increased by 35 °C	66 mutations	(Nguyen et al., 2017)
Haloalkane dehalogenase DhaA	Energetic assessment and phylogenetic analysis	FireProt	thermostability (T_m) increased by 24.6 °C	11 mutations	(Bednar et al., 2015)
Transketolase	Molecular dynamics simulations; analysis of coordinated motions	Gromacs	thermostability (T_m) increased by 3 °C half-life ($t_{1/2}$) increased 10-fold at 55 °C	2 mutations	(Yu and Dalby, 2018a)
Haloalkane dehalogenase LinB	Molecular dynamics simulations and directed evolution	AMBER, AceMD	catalysis (k_{cat}) improved 5-fold	4 mutations	(Kokkonen et al., 2018)
Nitrating cytochrome P450	Molecular dynamics simulations	AMBER	binding affinity increased 15-fold; regioselectivity switch	1 mutation	(Dodani et al., 2016)

complementarity with the transition state, which resulted in 84 designs selected for experimental validation. To improve these two enzymes, the residues in direct contact with the transition state were mutagenized. Several of these new designs proved successful, being the catalytic efficiency increased up to 100-fold compared to the original variant (Table 8). These computationally designed enzymes were more efficient than the uncatalyzed reaction and had also high stereoselectivity towards the desired enantiomers (section 3.3). Later works have also created new artificial Diels-Alderase enzymes based on metalloproteins (Reetz, 2012; Bos et al., 2012; Deuss et al., 2013).

3.3. Design of specific and enantioselective biocatalysts

The capability to bind one or more ligands and catalyse their conversion represents one of the defining properties of an enzyme and is referred to as substrate specificity (Eaton et al., 1995). Different enzymes present a wide range of specificity levels ranging from highly specific enzymes that only catalyse one particular substrate, through enzymes acting on a broad range of analogous molecules, to promiscuous ones catalysing even unrelated reactions (Chakraborty et al., 2012). Designing substrate specificity towards new small molecules has a great potential for biomedical, chemical and pharmaceutical industries (Ollikainen et al., 2015). Directed evolution has been usually employed to modify enzymes that already possessed at least a minimal activity towards a target substrate (Brustad and Arnold, 2011). Different strategies have been developed to predict the effect of enzyme redesign on substrate specificity. Most of them are focused on engineering the binding specificity of protein-protein complexes (Joachimik et al., 2006; Kapp et al., 2012; Sammond et al., 2007), simply considering them as fully rigid systems. However, predicting the affinity and the accurate orientation of small molecules requires a high-resolution sampling of the conformational flexibility of both the ligand and the binding site residues (Blomberg et al., 2013). Most of the studies modifying substrate specificity focus on the engineering of the active site whereas other parts of the enzyme are typically neglected. Protein access tunnels can be engineered (section 2.3) to discriminate certain ligands while keeping the catalytic site intact (Kokkonen et al., 2019).

The Rosetta package gradually has become the gold standard for engineering substrate specificity and has been used to solve multiple

engineering challenges. Its predictive power was analysed in a reverse-engineering experiment using RosettaLigand, showing its ability to recover mutations to alanine on polar (33%) and non-polar (62%) residues (Davis and Baker, 2009). The package has been used as preliminary screening to predict the effect of mutations on the activity of the enzyme (Aldeghi et al., 2018), to increase the catalytic activity towards a specific substrate (Tinberg et al., 2013), or to concurrently consider multiple design goals by working on an ensemble of structures (Löffler et al., 2017; Leaver-Fay et al., 2011). A paradigmatic example of specificity engineering using Rosetta involved redesigning the guanine deaminase enzyme to catalyse ammelide deamination using a loop remodelling protocol (Murphy et al., 2009). First, a hypothetical transition state structure for ammelide was placed into the active site of the guanine deaminase enzyme. Anchoring residues were then freely placed to maximize the interaction with the new ligand, and the backbone conformations capable of holding these anchor residues were identified. Subsequently, the protein structure was optimized to stabilize the novel backbone, and thus the conformational requirements to bind the desired ligand were predicted. Finally, unconstrained *de novo* loop remodelling (Rohl et al., 2004) was applied to confirm that the new sequence had propensity to fold into the predicted conformation. After applying this protocol, the selectivity towards ammelide was shifted around 6 orders of magnitude, and the activity improved by 100-fold (Table 8). Conversely, the catalytic efficiency was still several orders of magnitude lower than that of the wild type enzyme towards guanine.

The outcomes from basic Rosetta design suite can be improved by a 5% using PocketOptimizer (Malisi et al., 2012), even disregarding backbone flexibility (Richter et al., 2011). Other strategies have been developed to account for backbone displacement and have been successfully used to redesign enzymatic activity, such as the coupled motions method (Chakraborty et al., 2012) or K^* (Chen et al., 2009). Among these, an iterative protein redesign and optimization strategy (Pantazes et al., 2015) was profitably applied to modify the substrate specificity of TesA thioesterase, enabling the enzyme to produce different chain-length fatty acids (Grisewood et al., 2017). Each iteration in the strategy is evaluated by a Metropolis criterion and consists of i) local backbone perturbation, ii) suggestion of mutations by a mixed-integer linear program, iii) rigid-body docking to reorient ligand, iv) energy minimization of the complex, and v) evaluation of energy and

constraint geometry. To shift TesA products from the natural long-chain fatty acids medium-chain (C6-C12) ones, four rounds of the iterative protein redesign and optimization were performed to select variants with reduced binding to C14 and increased affinity for C12 or C8. All the 54 designed variants, were constructed and experimentally tested, and the best two showed a 2- and 10-fold improvement towards the production of C12 or C8 fatty acids, respectively (Table 8). This proved to be a significant enhancement in comparison to the activity of other 61 variants of the same protein produced by random mutagenesis.

Design of enantioselectivity is similar to other substrate specificity engineering, except that while designing a good affinity towards one enantiomer, the binding of the undesired enantiomer needs to be suppressed (Korendovych, 2018). Wijma and co-workers developed the catalytic selectivity by computational design, a strategy to produce small mutant libraries of enantioselective enzymes (Wijma et al., 2015). The method was successfully applied to design variants of the limonene epoxide hydrolase for the production of enantio-enriched (*S,S*)- and (*R,R*)-diols. The desired reactive conformations of the ligand were docked into the enzyme and RosettaDesign (Richter et al., 2011) was then used to generate a library of 1876 mutants. Such variants were designed to favour the binding of the target enantiomer and to concurrently introduce steric hindrances to the binding of the second enantiomer. The ability of each variant to specifically bind the target enantiomer in its reactive orientation was evaluated using multiple independent molecular dynamics simulations. This high-throughput virtual characterization led to the selection of only 37 mutants that were further characterized experimentally. The best variants for both (*S,S*)- and (*R,R*)- substrates showed a significant enantiomeric preference of 93% and 80%, respectively (Table 8). The enantiomeric excess of the final variants was comparable to that of other designs identified by directed evolution, which required much larger experimental effort (Zheng and Reetz, 2010).

3.4. Design of stable and soluble biocatalysts

The design approaches described above aim to enhance the catalytic properties of enzymes and yet, the redesign of their sequence may easily result in a reduced solubility or stability (Tokuriki et al., 2008; Dellus-Gur et al., 2013; Johansson et al., 2016). These negative effects can be reverted by introducing stabilizing or solubilizing mutations by protein engineering (Sormanni et al., 2015; Ganesan et al., 2016; Bloom et al., 2006). Most of the available tools for solubility engineering rely on machine learning. The models use collections of data (primary sequences, sequence profiles or compilations of mutations) associated to the solubility measurements. The methods based on plain protein sequences are: SOLpro (Magnan et al., 2009), PROSO II (Smialowski et al., 2012), ccSOL omics (Agostini et al., 2014), DeepSol (Khurana et al., 2018), ESPRESSO (Hirose and Noguchi, 2013), or the recently developed SoluProt (Hon et al., 2021). The majority of such methods exploit support vector machines or random forests to construct their predictors. Approaches relying on sequence profiles assign to each residue a solubility score that contextually describes its relative contribution to the solubility of the protein. This family of methods includes: AGGRESCAN (Conchillo-Sol   et al., 2007), TANGO (Fernandez-Escamilla et al., 2004), WALTZ (Maurer-Stroh et al., 2010), PASTA (Walsh et al., 2014), BETASCAN (Bryan et al., 2009), and FoldAmyloid (Garbuzynskiy et al., 2010). On top of these approaches, methods for predicting the solubility outcome of individual mutations usually require of the protein tertiary structure; this category includes: OptSolMut (Tian et al., 2010b), CamSol (Sormanni et al., 2015), AGGRESCAN3D (Zambrano et al., 2015), Solubis (Van Durme et al., 2016), SODA (Paladin et al., 2017), and PON-Sol (Yang et al., 2016). Solubility predictors have still a wide space to improve their accuracy, considering that the best tools report around 58 % accuracy on independent data sets. Nevertheless, the combination of several tools – Solubis, TANGO, and FoldX (Ganesan et al., 2016) – lead to a designed mutant of the α -galactosidase showing a 40 % increase in

solubility (Table 8).

A plethora of methods aimed to predict the effects of individual mutations on protein stability has emerged in recent years. These tools mainly exploit force-field calculations, machine learning or phylogenetic analyses. Force-field methods rely on the simulation of known laws of physics. The widely employed Rosetta ddg.monomer module (Kellogg et al., 2011) has been used to increase the thermal stability of the cutinase enzyme by 5.7 °C (Shirke et al., 2016). A different approach, based on phylogenetic analyses, relies on the knowledge of the evolutionary history of the studied enzyme and unlike the force-field approaches do not require a three-dimensional structure. One phylogenetic method denominated: *ancestral sequence reconstruction* allowed for a notable stabilization of the adenosine kinase. Kern and collaborators compiled a set of homologous sequences from the NCBI database, estimated their phylogeny using Bayesian inference (Suchard and Redelings, 2006), and designed 8 evolutionary ancestors with melting temperatures ranging from 64 °C to 89 °C. The best variant encompassed 66 different mutations (Nguyen et al., 2017) and achieved a stability improvement of 35 °C (Table 8).

Hybrid methods combine results from different approaches, increasing the robustness and reliability of their predictions. This enables the design of multiple-point mutants minimising the risk of suggesting mutations with antagonistic effects as exemplified in the stabilised halohydrin dehalogenase (Arabnejad et al., 2017) and acetylcholinesterase (Goldenzweig et al., 2016). The FireProt (Musil et al., 2017) method and web application combine energy- and evolution-based approaches in a single pipeline, where the sequence-derived information is used as a primary filter for time-demanding force-field calculations. Its workflow is presented in Fig. 4. The pipeline was used to enhance the stability of the haloalkane dehalogenase DhaA without compromising its catalytic efficiency (Bednar et al., 2015). After constructing a multiple sequence alignment of homologs, sequence positions were removed from further analysis if identified either as conserved or mutually correlated. Saturation mutagenesis on the remaining positions was performed using FoldX and 22 potentially stabilizing mutations were identified by Rosetta, as part of the energy-based branch of the pipeline. On the evolution-based side of the workflow, all mutations suggested by back-to-consensus analysis (Lehmann et al., 2002) were evaluated with FoldX, identifying another 13 potentially stabilizing mutations (Table 8). Various subsets of those mutations were characterized experimentally and the final set of 11 mutations (8 energy- and 3 evolution-based) improved the thermal stability of DhaA by 24.6 °C (Bednar et al., 2015).

3.5. Design of biocatalysts with tailored flexibility

Protein flexibility, from side-chain level to backbone dynamics, is a common theme in many enzyme engineering strategies. Yet, engineering protein flexibility itself is one of the hardest challenges to engage. First, because the output of this task is not only the modification of specific characteristics but introducing very precise changes in the motion of the engineered region aiming for specific amplitudes and frequencies of such motions. Second, because long-range interactions that enhance or constrict such dynamic behaviour are yet not completely understood (Yu and Dalby, 2018b; Chen et al., 2018). Engineering flexibility is important because internal motions of the enzyme often influence the ligand binding, the catalytic steps of the reaction, as well as the product release. For instance, the importance of dynamics on monooxygenases was recently reviewed (F  rst et al., 2019). The two structural features that are particularly susceptible to flexibility engineering are tunnels and loops. Surprisingly, the approaches undertaken to engage these features are opposite. Rational-driven design is prevalent in the engineering of tunnels whereas mainly experimental-guided efforts have been successful in loop design (Kre   et al., 2018). Nevertheless, different protocols have been devised to systematize this task either in the complementarity-determining regions of antibodies (Baran

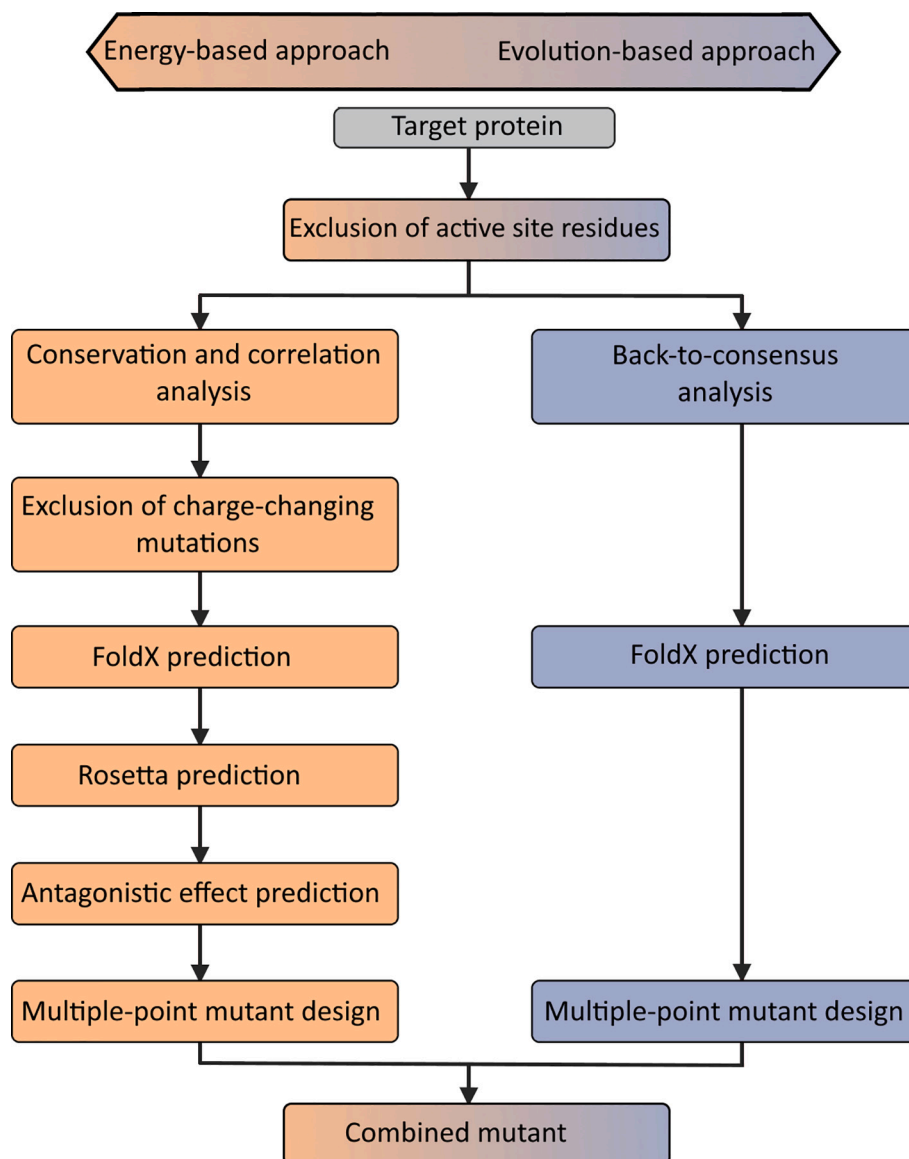


Fig. 4. General workflow of FireProt. Orange denotes strategies leading to a mutation pool designed by energetic methods, while blue indicates the steps required to select mutations by a back-to-consensus approach (Musil et al., 2017).

et al., 2017; Bonet et al., 2018) or insoluble enzymes (Murphy et al., 2009). The main methods and software tools for tailoring the functional aspects associated with tunnels have been reviewed (Gora et al., 2013; Kokkonen et al., 2019) and are covered in section 2.3. Loop design has also been recently discussed (Kundert and Kortemme, 2019).

After successfully using saturated mutagenesis for engineering *Escherichia coli* transketolase to accept aromatic substrates, Yu and Dalby faced a bewildering dilemma. The designed active site showed a greatly increased flexibility that compromised the enzyme stability, while further mutations on the active site were discouraged due to the risk of compromising the newly gained activity. Previous knowledge existed on stabilizing mutations for the wild type enzyme in the flexible loop regions, and the authors hypothesized that those regions could have coupled motions to those in the over-flexible engineered active site. They devised the use of dynamics cross-correlation maps to analyse coordinated motions (Lehmann et al., 2002), and confirmed that the previously introduced mutations correlated positively with the active site (Table 8). Hence, new variants of the engineered enzyme carrying one or two of the aforementioned mutations were designed, of which the most effective one improved the thermal stability of the enzyme by 3 °C

and its half-life at 55 °C by ten-fold (Yu and Dalby, 2018a).

Brezovsky and co-workers (Brezovsky et al., 2016) introduced a novel auxiliary tunnel to the haloalkane dehalogenase LinB using a combination of rational design and directed evolution. The residues lining the bottleneck of a very narrow tunnel were mutagenized and screened for activity towards 1,2-dibromoethane (Table 8). The mutant with a newly open tunnel is the most catalytically proficient haloalkane dehalogenase reported to date, with suppressed substrate inhibition. A follow-up study revealed that opening of this auxiliary tunnel resulted in the molecular gating ($k_{\text{on}} = 1822 \text{ s}^{-1}$, $k_{\text{off}} = 60 \text{ s}^{-1}$), which accelerated the rate-limiting product release in the open conformation while keeping the two chemical steps sufficiently fast when its conformation was closed (Kokkonen et al., 2018).

Arnold and co-workers exploited molecular dynamics simulations to study the conformational flexibility of two loops known to have a coordinated activity in a member of the cytochrome P450 superfamily that can nitrate L-tryptophan in its fourth carbon. Given the high disorder of one of such loops (F/G), it needed to be reconstructed from scratch using a hierarchical approach to all-atom protein prediction (Jacobson et al., 2004). Adaptive sampling simulations on the model followed by

reconstruction of Markov-state clusters revealed two different populations of the F/G loop. These states correspond with the closed and open conformations of the enzyme. It was expected that a partial stabilization of the closed state would result in an increased affinity for the substrate. A tyrosine residue located in the F/G loop stabilizes the substrate in the closed conformation and shows a disparate orientation in the open one (Dodani et al., 2016). Interestingly, in an isolated intermediate state, a tyrosine-tyrosine interaction between the two loops was substituted by a much weaker histidine-tyrosine interaction. Mutating that histidine to either phenylalanine or tyrosine resulted in an up to a 15-fold increase of the binding affinity for the L-tryptophan substrate, and surprisingly also in a shift of the nitrated position from C4 to C5. A single mutation in the target loop not only modified its dynamic behaviour but also the enzyme regioselectivity (Table 8).

4. Conclusions and perspectives

The development of methods and tools for computational enzyme design goes hand-to-hand with advances in structural methods, which provide structural information at the atomic resolution required for obtaining reliable computational predictions. Structural genomics projects have systematically mapped protein domains into the structure space. Advances in protein crystallography and nuclear magnetic resonance spectroscopy have been instrumental in these endeavours. Rapid improvements in cryo-electron microscopy provide structural information for membrane-bound enzymes, which were previously difficult to study using other structural techniques. Bioinformatics and phylogenetic approaches greatly benefit from next-generation sequencing technologies, resulting in many new genomes sequenced and in a geometric growth of genomic databases. Novel gene mining tools and automated high-throughput laboratory techniques will likely be developed to take advantage of ever-increasing structural and sequence diversity available in the publically accessible databases.

We envisage that individual computational and bioinformatics tools will be more often integrated into complex and automated workflows. These workflows will simplify the consecutive steps of the bioinformatics analyses and will eliminate potential errors introduced during repetitive actions. Workflows will also allow the simultaneous optimization of multiple properties by integrating a wide range of tools. For instance, we expect workflows to facilitate the introduction of stabilizing mutations to the studied enzyme right after the engineering of its activity or specificity, minimizing in this way the number of designs that would not fold properly. Intuitive web interfaces will expand to non-experts the accessibility to sophisticated computational methods, making the execution of calculations and the analysis of output data easier. Nevertheless, a proper interpretation of the collected results will still be crucial, but this can be assisted by advanced visualization techniques.

Computationally demanding tasks will benefit from the development of novel algorithms, parallelization and improvements in the hardware suitable for biomolecular simulations. Modelling enzymatic reactions using high-level hybrid quantum mechanics/molecular mechanics methods, or investigating the binding and release of substrates and products when these require large-scale conformational changes are amongst such challenging tasks. The use of graphical processing units already has allowed increasing the effective times of feasible simulations by an order of magnitude, and also to replicate such calculations to improve their sampling and to obtain better statistics. Gaining access to quantum computing will be a game-changer in this context, enabling the testing of countless designs or simulating the folding of structures containing thousands of atoms. However, quantum computing is still in its early stage of development and very likely will not be available for solving practical problems of computational design in a near future.

Catalysts to accelerate specific chemical reactions have been *de novo* designed for numerous systems, representing one of the greatest developments ever achieved in computational enzyme design. However, enzymes designed from scratch do not match natural ones in their

catalytic efficiency. This may be due to the requirement for atomic arrangements on the catalytic residues at sub-angstrom precision, which are stabilized by complex networks of hydrogen bonds. Such fine arrangements are beyond the prediction ability of currently available computational tools. Particularly challenging is the design of biocatalysts for multi-step reactions, which require the stabilization of several transition states within a single active site. Another difficult case is that of designing an enzyme to catalyse large and hydrophobic substrates, which normally require large conformational changes on the protein during their binding and unbinding steps. Yet another limitation of *de novo* design is that the currently available methods cannot precisely model the protein flexibility in terms of frequency and amplitude at specific and important regions. Algorithms and software tools for analysing the passage of ligands through protein tunnels and channels are maturing and can be potentially coupled to the design of transition states. Notably, multiple rounds of directed evolution can refine *de novo* designs to obtain native-like catalytic efficiencies.

We have witnessed impressive progress in the design of stable enzymes. Stabilizing mutations can be predicted accurately using available force fields. Mutations from energetic calculations should be ideally complemented by phylogenetic analyses, which can efficiently capture epistatic effects. Hybrid methods based on force-field calculations and phylogeny are accessible as web services and facilitate the design of stable multiple-point mutants. Conversely, the computational design of protein solubility is still in its infancy. Two approaches based on phylogenetic analyses – back-to-consensus and ancestral sequence reconstruction – often lead to robust, stable and expressible proteins. Mechanism-based models may be difficult to develop due to the high complexity of the factors and processes involved in protein expression, folding and aggregation at molecular and cellular levels. Machine learning holds greater promise, but require consistent and well-balanced data sets for independent training and validation. Such datasets can be collected in the future thanks to newly developed high-throughput methods, like fluorescence-activated cell sorting, microfluidics, fluorescence resonance energy transfer, deep sequencing, and deep mutational scanning.

The rational design of enzyme enantioselectivity is extremely challenging since it requires accurate calculations of the binding affinity and the catalytic rates for both *R*- and *S*-enantiomers. The rate-limiting phase of enantioselective catalysis can be either the chemical step or the binding and release of substrates and products, which often require conformational changes in the protein structure. Current limitations on the estimation of the transition states for such a rate-limiting step, especially in the latter case, render the precise calculation of enantioselectivity unfeasible. A possible solution to this problem is to identify hot spots and design small-but-smart libraries, ideally taking advantage of phylogenetic information for the prediction of safe mutations. Several approaches for the interactive introduction of mutations improving the enzyme enantioselectivity in a step-wise manner have been developed and experimentally validated achieving impressive results. It is to be seen whether machine learning approaches will provide some valuable information for a more reliable selection of hot spots and/or narrowing the alphabet of introduced amino acid residues. Yet, another level of complexity is expected for enzymes employing large entropic components to discriminate between *R*- and *S*-enantiomers. Proper treatment of protein dynamics and solvation will be needed for the estimation of binding affinities and reactivities, which is far from trivial. Looking backwards, what has been achieved in designing enzymes during the past two decades, and looking forward to newly developed software tools and emerging experimental techniques, there is much to be expected from computational enzyme design.

Acknowledgements

We thank all five referees for the time and effort devoted to careful reading of our manuscript and for making many excellent comments. We

also thank the Czech Ministry of Education (CZ.02.1.01/0.0/0.0/16.026/0008451), to European Union (814418), and Technology Agency of Czech Republic (TN01000013) for financial support. Joan Planas-Iglesias is grateful to the project Postdoc@MUNI (CZ.02.2.69/0.0/16_027/0008360). Milos Musil has received funding from Brno University of Technology (FIT-S-20-6293).

Conflict of interest declaration

All authors are requested to disclose any actual or potential conflict of interest including any financial, personal or other relationships with other people or organizations within five years of beginning the submitted work that could inappropriately influence, or be perceived to influence, their work. See also <https://www.elsevier.com/conflictsofinterest>.

References

- Abagyan, R., Totrov, M., Kuznetsov, D., 1994. ICM-A new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* 15, 488–506.
- Adcock, S.A., McCammon, J.A., 2006. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.* 106, 1589–1615.
- Agostini, F., Cirillo, D., Livi, C.M., Delli Ponti, R., Tartaglia, G.G., 2014. ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. *Bioinformatics* 30, 2975–2977.
- Ahlrichs, R., Bär, M., Häser, M., Horn, H., Kölmel, C., 1989. Electronic structure calculations on workstation computers: the program system turbomole. *Chem. Phys. Lett.* 162, 165–169.
- Alberto, M.E., Pinto, G., Russo, N., Toscano, M., 2015. Triesterase and promiscuous diesterase activities of a Di-Co II -containing organophosphate degrading enzyme reaction mechanisms. *Chem - A Eur J.* 21, 3736–3745.
- Aldeghi, M., Gapsys, V., de Groot, B.L., 2018. Accurate estimation of ligand binding affinity changes upon protein mutation. *ACS Cent Sci* 4, 1708–1718. A.
- Amara, P., Field, M.J., 2003. Evaluation of an ab initio quantum mechanical/molecular mechanical hybrid-potential link-atom method. *Theor. Chem. Accounts Theory, Comput. Model (Theoretica Chim Acta)* 109, 43–52.
- Amara, P., Field, M.J., Alhambra, C., Gao, J., 2000. The generalized hybrid orbital method for combined quantum mechanical/molecular mechanical calculations: formulation and tests of the analytical derivatives. *Theor. Chem. Accounts Theory, Comput. Model (Theoretica Chim Acta)* 104, 336–343.
- Amin, N., Liu, A.D., Ramer, S., Ahle, W., Meijer, D., Metin, M., et al., 2004. Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Eng Des Sel.* 17, 787–793.
- Andrews, L.D., Fenn, T.D., Herschlag, D., 2013. Ground state destabilization by anionic nucleophiles contributes to the activity of phosphoryl transfer enzymes. *PLoS Biol.* 11, e1001599.
- Arabnejad, H., Dal Lago, M., Jekel, P.A., Floor, R.J., Thunnissen, A.-M.W.H., Terwisscha van Scheltinga, A.C., et al., 2017. A robust cosolvent-compatible haloalcohol dehalogenase by computational library design. *Protein Eng Des Sel.* 30, 173–187.
- Arnold, F.H., 2018. Directed evolution: bringing new chemistry to life. *Angew. Chem. Int. Ed.* 57, 4143–4148.
- Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O., et al., 2012. FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Res.* 40, W580–W584.
- Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., et al., 2016. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res.* 44, W344–W350.
- Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., Bahar, I., 2001. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.* 80, 505–515.
- Bach, R.D., Canepa, C., 1997. Theoretical model for pyruvoyl-dependant enzymatic decarboxylation of α -amino acids. *J. Am. Chem. Soc.* 119, 11725–11733.
- Bahar, I., Atilgan, A.R., Erman, B., 1997. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.* 2, 173–181.
- Bahar, I., Lezon, T.R., Yang, L.-W., Eyal, E., 2010. Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.* 39, 23–42.
- Baran, D., Pszolla, M.G., Lapidoth, G.D., Norn, C., Dym, O., Unger, T., et al., 2017. Principles for computational design of binding antibodies. *Proc. Natl. Acad. Sci. U. S. A.* 114, 10900–10905.
- Bauer, P., Barrozo, A., Purg, M., Amrein, B.A., Esguerra, M., Wilson, P.B., et al., 2018. Q6: A comprehensive toolkit for empirical valence bond and related free energy calculations. *SoftwareX* 7, 388–395.
- Bayley, H., Jayasinghe, L., 2004. Functional engineered channels and pores (Review). *Mol. Membr. Biol.* 21, 209–220.
- Bednar, D., Beerens, K., Sebestova, E., Bendl, J., Khare, S., Chaloupkova, R., et al., 2015. FireProt: energy- and evolution-based computational design of thermostable multiple-point mutants. *PLoS Comput. Biol.* 11, e1004556.
- Bendl, J., Stourac, J., Sebestova, E., Vavra, O., Musil, M., Brezovsky, J., et al., 2016. HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Res.* 44, W479–W487.
- Benedix, A., Becker, C.M., de Groot, B.L., Caflich, A., Böckmann, R.A., 2009. Predicting free energy changes using structural ensembles. *Nat. Methods* 6, 3–4.
- Ben-Nun, M., Levine, R.D., 1995. Kinetics and dynamics of reactions in liquids. *Int. Rev. Phys. Chem.* 14, 215–270.
- Berendsen, H.J.C., van der Spoel, D., van Drunen, R., 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* 91, 43–56.
- Bernardi, R.C., Melo, M.C.R., Schulten, K., 1850. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *Biochim. Biophys. Acta, Gen. Subj.* 2015, 872–877.
- Betz, R.M., Dror, R.O., 2019. How effectively can adaptive sampling methods capture spontaneous ligand binding? *J. Chem. Theory Comput.* 15, 2053–2063.
- Biedermannova, L., Prokop, Z., Gora, A., Chovancova, E., Kovacs, M., Damborsky, J., et al., 2012. 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.* 287, 29062–29074.
- Bin, Masood T., Sandhya, S., Chandra, N., Natarajan, V., 2015. CHEXVIS: a tool for molecular channel extraction and visualization. *BMC Bioinformatics* 16, 119.
- Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag, Berlin, Heidelberg.
- Blomberg, R., Kries, H., Pinkas, D.M., Mittl, P.R.E., Grütter, M.G., Privett, H.K., et al., 2013. Precision is essential for efficient catalysis in an evolved Kemp eliminase. *Nature* 503, 418–421.
- Bloom, J.D., Labthavikul, S.T., Otey, C.R., Arnold, F.H., 2006. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U. S. A.* 103, 5869–5874.
- Bolon, D.N., Mayo, S.L., 2001. Enzyme-like proteins by computational design. *Proc. Natl. Acad. Sci.* 98, 14274–14279.
- Bonet, J., Wehrle, S., Schriever, K., Yang, C., Billet, A., Sesterhenn, F., et al., 2018. Rosetta FunFoldS - a general framework for the computational design of functional proteins. *PLoS Comput. Biol.* 14, e1006623.
- Borgo, B., Havranek, J.J., 2012. Automated selection of stabilizing mutations in designed and natural proteins. *Proc. Natl. Acad. Sci.* 109, 1494–1499.
- Bornscheuer, U.T., Hauer, B., Jaeger, K.E., Schwaneberg, U., 2019. Directed Evolution empowered redesign of natural proteins for the sustainable production of chemicals and pharmaceuticals. *Angew. Chem. Int. Ed.* 58, 36–40.
- Bos, J., Fusetti, F., Driessen, A.J.M., Roelfs, G., 2012. Enantioselective artificial metalloenzymes by creation of a novel active site at the protein dimer interface. *Angew. Chem. Int. Ed.* 51, 7472–7475.
- Boughorbel, S., Jarray, F., El-Anbari, M., 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* 12, e0177678. A.
- Breiman, L., 2001. Random Forests. *Mach. Learn.* 41, 5–32.
- Brezovsky, J., Chovancova, E., Gora, A., Pavelka, A., Biedermannova, L., Damborsky, J., 2013. Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnol. Adv.* 31, 38–49.
- Brezovsky, J., Babkova, P., Degtjarik, O., Fortova, A., Gora, A., Iermak, I., et al., 2016. Engineering a de novo transport tunnel. *ACS Catal.* 6, 7597–7610.
- Brustad, E.M., Arnold, F.H., 2011. Optimizing non-natural protein function with directed evolution. *Curr. Opin. Chem. Biol.* 15, 201–210.
- Bryan, A.W., Menke, M., Cowen, L.J., Lindquist, S.L., Berger, B., 2009. BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. *PLoS Comput. Biol.* 5, e1000333.
- Bussi, G., Laio, A., 2020. Using metadynamics to explore complex free-energy landscapes. *Nat. Rev. Phys.* 2, 200–212.
- Calland, P.-Y., 2003. On the structural complexity of a protein. *Protein Eng Des Sel.* 16, 79–86.
- Campbell, E.C., Correy, G.J., Mabbitt, P.D., Buckle, A.M., Tokuriki, N., Jackson, C.J., 2018. Laboratory evolution of protein conformational dynamics. *Curr. Opin. Struct. Biol.* 50, 49–57.
- Camps, J., Carrillo, O., Emperador, A., Orellana, L., Hospital, A., Rueda, M., et al., 2009. FlexServ: an integrated tool for the analysis of protein flexibility. *Bioinformatics* 25, 1709–1710.
- Capriotti, E., Fariselli, P., Casadio, R., 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33, W306–W310.
- Carlson, H.A., Smith, R.D., Damm-Ganamet, K.L., Stuckey, J.A., Ahmed, A., Convery, M. A., et al., 2016. CSAR 2014: a benchmark exercise using unpublished data from pharma. *J. Chem. Inf. Model.* 56, 1063–1077.
- Case, D.A., Ben-Shalom, I.Y., Brozell, S.R., Cerutti, D.S., Cheatham III, T.E., Cruzeiro, V. W.D., et al., 2018. AMBER 2018. University of California, San Francisco.
- Chakraborty, S., Åsgerisson, B., Rao, B.J., 2012. A measure of the broad substrate specificity of enzymes based on ‘duplicate’ catalytic residues. *PLoS One* 7, e49313.
- Chaloupkova, R., Sykorova, J., Prokop, Z., Jesenska, A., Monincova, M., Pavlova, M., et al., 2003. Modification of activity and specificity of haloalkane dehalogenase from *Sphingomonas paucimobilis* UT26 by engineering of its entrance tunnel. *J. Biol. Chem.* 278, 52622–52628.
- Chang, D.T.-H., Oyang, Y.-J., Lin, J.-H., 2005. MEDock: a web server for efficient prediction of ligand binding sites based on a novel optimization algorithm. *Nucleic Acids Res.* 33, W233–W238.
- Chang, C., Huang, Y., Mueller, L., You, W., 2016. Investigation of structural dynamics of enzymes and protonation states of substrates using computational tools. *Catalysts* 6, 82.
- Chawla, N.V., Japkowicz, N., 2004. Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newsl.* 1–6.
- Check Hayden, E., 2008. Chemistry: designer debacle. *Nature* 453, 275–278.

- Chen, L.Y., 2015. Hybrid steered molecular dynamics approach to computing absolute binding free energy of ligand–protein complexes: a Brute force approach that is fast and accurate. *J. Chem. Theory Comput.* 11, 1928–1938.
- Chen, C.-Y., Georgiev, I., Anderson, A.C., Donald, B.R., 2009. Computational structure-based redesign of enzyme activity. *Proc. Natl. Acad. Sci.* 106, 3764–3769.
- Chen, J., Liu, X., Chen, J., 2018. Atomistic peptide folding simulations reveal interplay of entropy and long-range interactions in folding cooperativity. *Sci. Rep.* 8, 13668.
- Cheng, J., Randall, A., Baldi, P., 2006. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*. 62, 1125–1132.
- Cheng, F., Zhu, L., Schwaneberg, U., 2015. Directed evolution 2.0: improving and deciphering enzyme properties. *Chem. Commun.* 51, 9760–9772.
- Chodera, J.D., Noé, F., 2014. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* 25, 135–144.
- Chovanova, E., Pavelka, A., Benes, P., Strnad, O., Brezovsky, J., Kozlikova, B., et al., 2012. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.* 8, e1002708.
- Christensen, N.J., Kepp, K.P., 2012. Accurate stabilities of laccase mutants predicted with a modified FoldX protocol. *J. Chem. Inf. Model.* 52, 3028–3042.
- Conchillo-Solé, O., de Groot, N.S., Avilés, F.X., Vendrell, J., Daura, X., Ventura, S., 2007. AGGRESAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinformatics*. 8, 65.
- Cournia, Z., Allen, B., Sherman, W., 2017. Relative binding free energy calculations in drug discovery: recent advances and practical considerations. *J. Chem. Inf. Model.* 57, 2911–2937.
- Cramer, C.J., Truhlar, D.G., 1999. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. *Chem. Rev.* 99, 2161–2200.
- Curado-Carballada, C., Feixas, F., Iglesias-Fernández, J., Osuna, S., 2019. Hidden conformations in *Aspergillus niger* monoamine oxidase are key for catalytic efficiency. *Angew. Chem. Int. Ed. Engl.* 58, 3097–3101.
- Curran, A., Swainston, N., Day, P.J., Kell, D.B., 2015. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chem. Soc. Rev.* 44, 1172–1239.
- Dahiyat, B.I., Mayo, S.L., 1997. De Novo Protein Design: Fully Automated Sequence Selection. *Science*. 278, 82–87.
- Dahiyat, B.I., Mayo, S.L., 2008. Protein design automation. *Protein Sci.* 5, 895–903.
- Dalby, P.A., 2007. Engineering enzymes for biocatalysis. *Recent Pat. Biotechnol.* 1, 1–9.
- Daniel, L., Buryška, T., Prokop, Z., Damborsky, J., Brezovsky, J., 2015. Mechanism-based discovery of novel substrates of haloalkane dehalogenases using in silico screening. *J. Chem. Inf. Model.* 55, 54–62.
- Dapprich, S., Komáromi, I., Byun, K.S., Morokuma, K., Frisch, M.J., 1999. A new ONIOM implementation in Gaussian98. Part I. The calculation of energies, gradients, vibrational frequencies and electric field derivatives. *J. Mol. Struct. THEOCHEM* 461–462, 1–21.
- Darve, E., Pohorille, A., 2001. Calculating free energies using average force. *J. Chem. Phys.* 115, 9169–9183.
- Dauber-Osguthorpe, P., Osguthorpe, D.J., Stern, P.S., Moulton, J., 1999. Low Frequency Motion in Proteins. *J. Comput. Phys.* 151, 169–189.
- Davis, I.W., Baker, D., 2009. RosettaLigand docking with full ligand and receptor flexibility. *J. Mol. Biol.* 385, 381–392.
- Deem, M.W., Bader, J.S., 1996. A configurational bias Monte Carlo method for linear and cyclic peptides. *Mol. Phys.* 87, 1245–1260. A.
- Dehouck, Y., Gilis, D., Rooman, M., 2006. A new generation of statistical potentials for proteins. *Biophys. J.* 90, 4010–4017.
- Dehouck, Y., Kwasigroch, J.M., Gilis, D., Rooman, M., 2011. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*. 12, 151.
- Dellus-Gur, E., Toth-Petroczy, A., Elias, M., Tawfik, D.S., 2013. What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. *J. Mol. Biol.* 425, 2609–2621.
- Deuss, P.J., Popa, G., Slawin, A.M.Z., Laan, W., Kamer, P.C.J., 2013. Artificial copper enzymes for asymmetric diels-alder reactions. *ChemCatChem*. 5, 1184–1191.
- Devaurs, D., Bouard, L., Vaisset, M., Zanon, C., Al-Blawi, I., Iehl, R., et al., 2013. MoMA-LigPath: a web server to simulate protein–ligand unbinding. *Nucleic Acids Res.* 41, W297–W302.
- Diallo, A.B., Makarenkov, V., Blanchette, M., 2010. Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*. 26, 130–131.
- Dickson, A., 2018. Mapping the ligand binding landscape. *Biophys. J.* 115, 1707–1719.
- DiLabio, G.A., Hurley, M.M., Christiansen, P.A., 2002 Jun. Simple one-electron quantum capping potentials for use in hybrid QM/MM studies of biological molecules. *J. Chem. Phys.* 116, 9578–9584.
- Do, P.-C., Lee, E.H., Le, L., 2018. Steered molecular dynamics simulation in rational drug design. *J. Chem. Inf. Model.* 58, 1473–q482.
- Dodani, S.C., Kiss, G., Cahn, J.K.B., Su, Y., Pande, V.S., Arnold, F.H., 2016. Discovery of a regioselectivity switch in nitrating P450s guided by molecular dynamics simulations and Markov models. *Nat. Chem.* 8, 419–425.
- Duane, S., Kennedy, A.D., Pendleton, B.J., Roweth, D., 1987. Hybrid Monte Carlo. *Phys. Lett B*. 195, 216–222.
- Earl, D.J., Deem, M.W., 2008. Monte Carlo Simulations. In: *Molecular Modeling of Proteins*, pp. 25–36.
- Eaton, B.E., Gold, L., Zichi, D.A., 1995. Let's get specific: the relationship between specificity and affinity. *Chem. Biol.* 2, 633–638.
- Ebert, M.C.C.J.C., Espinola, J.G., Guillaume Lamoureux, G., Pelletier, J.N., 2017. Substrate-specific screening for mutational hotspots using biased molecular dynamics simulations. *ACS Catal.* 7, 6786–6797.
- Eyal, E., Lum, G., Bahar, I., 2015. The anisotropic network model web server at 2015 (ANM 2.0). *Bioinformatics*. 31, 1487–1489.
- Faber, M.S., Whitehead, T.A., 2019. Data-driven engineering of protein therapeutics. *Curr. Opin. Biotechnol.* 60, 104–110.
- Fasan, R., Mehareenna, Y.T., Snow, C.D., Poulos, T.L., Arnold, F.H., 2008. Evolutionary history of a specialized P450 propane monooxygenase. *J. Mol. Biol.* 383, 1069–1080.
- Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., Serrano, L., 2004. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* 22, 1302–1306.
- Ferrer, S., Ruiz-Pernía, J., Martí, S., Moliner, V., Tuñón, I., Bertrán, J., et al., 2011. Hybrid schemes Based on quantum mechanics/molecular mechanics simulations. In: *Advances in Protein Chemistry and Structural Biology*, pp. 81–142.
- Filipovic, J., Vavra, O., Plhak, J., Bednar, D., Marques, S.M., Brezovsky, J., et al., 2019. CaverDock: a novel method for the fast analysis of ligand transport. *IEEE/ACM Trans Comput Biol Bioinforma.* 1, 1.
- Folkman, L., Stantic, B., Sattar, A., Zhou, Y., 2016. EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.* 428, 1394–1405.
- Friesner, R.A., Gualler, V., 2005. Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis. *Annu. Rev. Phys. Chem.* 56, 389–427.
- Friesner, R.A., Banks, J.L., Murphy, R.B., Halgren, T.A., Klicic, J.J., Mainz, D.T., et al., 2004. Glide: a new approach for rapid, accurate docking and scoring. *J. Med. Chem.* 47, 1739–1749.
- Furini, S., Domene, C., 1858. Computational studies of transport in ion channels using metadynamics. *Biochim. Biophys. Acta Biomembr.* 2016, 1733–1740.
- Fürst, M.J., Fiorentini, F., Fraaije, M.W., 2019. Beyond active site residues: overall structural dynamics control catalysis in flavin-containing and heme-containing monooxygenases. *Curr. Opin. Struct. Biol.* 59, 29–37.
- Ganesan, A., Siekierska, A., Beerten, J., Brams, M., Van Durme, J., De Baets, G., et al., 2016. Structural hot spots for the solubility of globular proteins. *Nat. Commun.* 7, 10816.
- Gao, L., Su, C., Du, X., Wang, R., Chen, S., Zhou, Y., et al., 2020. FAD-dependent enzyme-catalysed intermolecular [4+2] cycloaddition in natural product biosynthesis. *Nat. Chem.* 12, 620–628.
- Garbuzynskiy, S.O., Lobanov, M.Y., Galzitskaya, O.V., 2010. FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. *Bioinformatics*. 26, 326–332.
- Gardner, J.M., Biler, M., Risso, V.A., Sanchez-Ruiz, J.M., Kamerlin, S.C.L., 2020. Manipulating conformational dynamics to repurpose ancient proteins for modern catalytic functions. *ACS Catal.* 10, 4863–4870.
- Gaulton, A., Bellis, L.J., Bento, A.P., Chambers, J., Davies, M., Hersey, A., et al., 2012. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* 40, D1100–D1107.
- Gaytán, P., Contreras-Zambrano, C., Ortiz-Alvarado, M., Morales-Pablos, A., Yáñez, J., 2009. TrimerDimer: an oligonucleotide-based saturation mutagenesis approach that removes redundant and stop codons. *Nucleic Acids Res.* 37, e125.
- Gelpi, J., Hospital, A., Goñi, R., Orozco, M., 2015. Molecular dynamics simulations: advances and applications. *Adv Appl Bioinforma Chem [Internet]* 37.
- Georgieva, P., Himo, F., 2010. Quantum chemical modeling of enzymatic reactions: The case of histone lysine methyltransferase. *J. Comput. Chem.* 31, 1707–1714.
- Geyer, C.J., Thompson, E.A., 1995. Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Stat. Assoc.* 90, 909–920.
- Goldenzweig, A., Goldsmith, M., Hill, S.E., Gertman, O., Laurino, P., Ashani, Y., et al., 2016. Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell* 63, 337–346.
- Goldsmith, M., Tawfik, D.S., 2013. Enzyme engineering by targeted libraries. In: *Methods in Enzymology*, pp. 257–283.
- Gora, A., Brezovsky, J., Damborsky, J., 2013. Gates of enzymes. *Chem. Rev.* 113, 5871–5923.
- Grisewood, M.J., Hernández-Lozada, N.J., Thoden, J.B., Gifford, N.P., Mendez-Perez, D., Schoenberger, H.A., et al., 2017. Computational redesign of Acyl-ACP thioesterase with improved selectivity toward medium-chain-length fatty acids. *ACS Catal.* 7, 3837–3849.
- Guedes, I.A., de Magalhães, C.S., Dardenne, L.E., 2014. Receptor–ligand molecular docking. *Biophys. Rev.* 6, 75–87.
- Guedes, I.A., Pereira, F.S.S., Dardenne, L.E., 2018. Empirical scoring functions for structure-based virtual screening: applications, critical aspects, and challenges. *Front. Pharmacol.* 9.
- Guerois, R., Nielsen, J.E., Serrano, L., 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 320, 369–387.
- Hall, R.J., Hindle, S.A., Burton, N.A., Hillier, I.H., 2000. Aspects of hybrid QM/MM calculations: The treatment of the QM/MM interface region and geometry optimization with an application to chorismate mutase. *J. Comput. Chem.* 21, 1433–1441.
- Hamelberg, D., Mongan, J., McCammon, J.A., 2004. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* 120, 11919–11929.
- Hammes-Schiffer, S., 2013. Catalytic efficiency of enzymes: a theoretical analysis. *Biochemistry*. 52, 2012–2020.
- Hanson-Smith, V., Johnson, A., 2016. PhyloBot: a web portal for automated phylogenetics, ancestral sequence reconstruction, and exploration of mutational trajectories. *PLoS Comput. Biol.* 12, e1004976.
- Harvey, M.J., Giupponi, G., De Fabritiis, G., 2009. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. *J. Chem. Theory Comput.* 5, 1632–1639.

- Hassan, N.M., Alhossary, A.A., Mu, Y., Kwok, C.K., 2017. Protein-ligand blind docking using quickvina-W with inter-process spatio-temporal integration. *Sci. Rep.* 7, 15451.
- Hehre, W.J., Lathan, W.A., Ditchfield, R., Newton, M.D., Pople, J.A., 1970. Gaussian 70. Quantum Chemistry Program Exchange. Program No. 237.
- Hellings, H.W., Richards, F.M., 1991. Construction of new ligand binding sites in proteins of known structure. *J. Mol. Biol.* 222, 763–785.
- Hellings, H.W., Caradonna, J.P., Richards, F.M., 1991. Construction of new ligand binding sites in proteins of known structure. *J. Mol. Biol.* 222, 787–803.
- Henzler-Wildman, K., Kern, D., 2007. Dynamic personalities of proteins. *Nature*. 450, 964–972.
- Himo, F., 2017. Recent Trends in Quantum Chemical Modeling of Enzymatic Reactions. *J. Am. Chem. Soc.* 139, 6780–6786.
- Hirose, S., Noguchi, T., 2013. ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. *Proteomics*. 13, 1444–1456.
- Hon, J., Marusiak, M., Martinek, T., Zendulka, J., Bednar, D., Damborsky, J., 2021. SoluProt: prediction of protein solubility. *Bioinformatics*. <https://doi.org/10.1093/bioinformatics/btaa1102> (In press).
- Hoppe, C., Schomburg, D., 2005. Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. *Protein Sci.* 14, 2682–2692.
- Houk, K.N., Cheong, P.H., 2008 Sep 18. Computational prediction of small-molecule catalysts. *Nature*. 455 (7211), 309–313.
- <https://www.nobelprize.org/prizes/chemistry/2013/summary/> (cited 2019 May 29).
- <https://www.nobelprize.org/prizes/chemistry/2018/summary/> (cited 2019 May 29).
- Huang, B., 2009. MetaPocket: a meta approach to improve protein ligand binding site prediction. *Omi A J Integr Biol.* 13, 325–330.
- Huang, L.-T., Gromiha, M.M., Ho, S.-Y., 2007. iPTREE-STAB: interpretable decision tree based method for predicting protein stability changes upon mutations. *Bioinformatics*. 23, 1292–1293.
- Huelsbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*. 294, 2310–2314.
- Hughes, G., Lewis, J.C., 2018. Introduction: Biocatalysis in Industry. *Chem. Rev.* 118, 1–3.
- Hutter, J., 2012. Car-Parrinello molecular dynamics. *Wiley Interdiscip Rev Comput Mol Sci.* 2, 604–612.
- Irwin, J.J., Sterling, T., Mysinger, M.M., Bolstad, E.S., Coleman, R.G., 2012. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* 52, 1757–1768.
- Izaguirre, J.A., Hampton, S.S., 2004. Shadow hybrid Monte Carlo: an efficient propagator in phase space of macromolecules. *J. Comput. Phys.* 200, 581–604.
- Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J.F., Honig, B., Shaw, D.E., et al., 2004. A hierarchical approach to all-atom protein loop prediction. *Proteins*. 55, 351–367.
- Jamieson, C.S., Ohashi, M., Liu, F., Tang, Y., Houk, K.N., 2019. The expanding world of biosynthetic pericyclases: cooperation of experiment and theory for discovery. *Nat. Prod. Rep.* 36, 698–713.
- Jansen, J.M., Koehler, K.F., Hedberg, M.H., Johansson, A.M., Hacksell, U., Nordvall, G., Snyder, J.P., 1997. Molecular design using the minireceptor concept. *J. Chem. Inf. Comput. Sci.* 37, 812–818.
- Jiang, L., Althoff, E.A., Clemente, F.R., Doyle, L., Rothlisberger, D., Zanghellini, A., et al., 2008. De novo computational design of retro-aldol enzymes. *Science*. 319, 1387–1391.
- Joachimski, L.A., Kortemme, T., Stoddard, B.L., Baker, D., 2006. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein–protein interface. *J. Mol. Biol.* 361, 195–208.
- Jochens, H., Bornscheuer, U.T., 2010. Natural diversity to guide focused directed evolution. *ChemBioChem*. 11, 1861–1866.
- Johansson, K.E., Tidemand-Johansen, N., Christensen, S., Horowitz, S., Bardwell, J.C.A., Olsen, J.G., et al., 2016. Computational redesign of thioredoxin is hypersensitive toward minor conformational changes in the backbone template. *J. Mol. Biol.* 428, 4361–4377.
- Jones, G., Willlett, P., Glen, R.C., Leach, A.R., Taylor, R., 1997. Development and validation of a genetic algorithm for flexible docking 1 Edited by F. E. Cohen. *J. Mol. Biol.* 267, 727–748.
- Jurcik, A., Bednar, D., Byska, J., Marques, S.M., Furmanova, K., Daniel, L., et al., 2018. CAVER Analyst 2.0: analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories. *Bioinformatics* 34, 3586–3588.
- Kamerlin, S.C.L., Warshel, A., 2011. The empirical valence bond model: theory and applications. *Wiley Interdiscip Rev Comput Mol Sci.* 1, 30–45.
- Kapp, G.T., Liu, S., Stein, A., Wong, D.T., Remenyi, A., Yeh, B.J., et al., 2012. Control of protein signaling using a computationally designed GTPase/GEF orthogonal pair. *Proc. Natl. Acad. Sci.* 109, 5277–5282.
- Karplus, M., McCammon, J.A., 2002. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* 9, 646–652.
- Kästner, J., 2011. Umbrella sampling. *Wiley Interdiscip Rev Comput Mol Sci.* 1, 932–942.
- Kaushik, S., Prokop, Z., Damborsky, J., Chaloupkova, R., 2017 Jan. Kinetics of binding of fluorescent ligands to enzymes with engineered access tunnels. *FEBS J.* 284, 134–148.
- Kellogg, E.H., Leaver-Fay, A., Baker, D., 2011. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*. 79, 830–838.
- Kepp, K.P., 1854. Towards a “Golden Standard” for computing globin stability: Stability and structure sensitivity of myoglobin mutants. *Biochim. Biophys. Acta* 2015, 1239–1248.
- Khare, S.D., Fleishman, S.J., 2013. Emerging themes in the computational design of novel enzymes and protein-protein interfaces. *FEBS Lett.* 587, 1147–1154.
- Khurana, S., Rawi, R., Kunji, K., Chuang, G.-Y., Bensmail, H., Mall, R., 2018. DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* 34, 2605–2613.
- Kim, J.-K., Cho, Y., Lee, M., Laskowski, R.A., Ryu, S.E., Sugihara, K., et al., 2015. BetaCavityWeb: a webserver for molecular voids and channels. *Nucleic Acids Res.* 43, W413–W418.
- Kim, S., Thiessen, P.A., Bolton, E.E., Chen, J., Fu, G., Gindulyte, A., et al., 2016. PubChem substance and compound databases. *Nucleic Acids Res.* 44, D1202–D1213.
- Kingsley, L.J., Lill, M.A., 2015. Substrate tunnels in enzymes: Structure-function relationships and computational methodology. *Proteins Struct Funct Bioinforma.* 83, 599–611.
- Kiss, G., Çelebi-Ölçüm, N., Moretti, R., Baker, D., Houk, K.N., 2013. Computational enzyme design. *Angew. Chem. Int. Ed.* 52, 5700–5725.
- Klesmith, J.R., Bacik, J.-P., Wrenbeck, E.E., Michalczyk, R., Whitehead, T.A., 2017. Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl. Acad. Sci. U. S. A.* 114, 2265–2270.
- Kokh, D.B., Amaral, M., Bomke, J., Grädler, U., Musil, D., Buchstaller, H.-P., et al., 2018. Estimation of drug-target residence times by τ -random acceleration molecular dynamics simulations. *J. Chem. Theory Comput.* 14, 3859–3869.
- Kokkonen, P., Sykora, J., Prokop, Z., Ghose, A., Bednar, D., Amaro, M., et al., 2018. Molecular gating of an engineered enzyme captured in real time. *J. Am. Chem. Soc.* 140, 17999–18008.
- Kokkonen, P., Bednar, D., Pinto, G., Prokop, Z., Damborsky, J., 2019. Engineering enzyme access tunnels. *Biotechnol. Adv.* 37, 107386.
- Korendovych, I.V., 2018. Rational and Semirational Protein Design, pp. 15–23.
- Korendovych, I.V., DeGrado, W.F., 2014. Catalytic efficiency of designed catalytic proteins. *Curr. Opin. Struct. Biol.* 27, 113–121.
- Kotsiantis, S., Kanellopoulos, D., Pintelas, P., 2005. Handling imbalanced datasets: A review. *GESTS Int Trans Comput Sci Eng.* 30, 25–36.
- Kreß, N., Halder, J.M., Rapp, L.R., Hauer, B., 2018. Unlocked potential of dynamic elements in protein structures: channels and loops. *Curr. Opin. Chem. Biol.* 47, 109–116.
- Kries, H., Blomberg, R., Hilvert, D., 2013. De novo enzymes by computational design. *Curr. Opin. Chem. Biol.* 17, 221–228.
- Kuipers, R.K., Joosten, H.-J., van Berkel, W.J.H., Leferink, N.G.H., Rooijen, E., Ittmann, E., et al., 2010. 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Proteins*. 78, 2101–2113.
- Kumar, S., Rosenberg, J.M., Bouzida, D., Swendsen, R.H., Kollman, P.A., 1992. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.* 13, 1011–1021.
- Kumari, I., Sandhu, P., Ahmed, M., Akhter, Y., 2017. Molecular dynamics simulations, challenges and opportunities: a biologist's prospective. *Curr. Protein Pept. Sci.* 18, 1163–1179.
- Kundert, K., Kortemme, T., 2019. Computational design of structured loops for new protein functions. *Biol. Chem.* 400, 275–288.
- Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R., Ferrin, T.E., 1982. A geometric approach to macromolecule–ligand interactions. *J. Mol. Biol.* 161, 269–288.
- Laimer, J., Hofer, H., Fritz, M., Wegenkittl, S., Lackner, P., 2015. MAESTRO–multi agent stability prediction upon point mutations. *BMC Bioinformatics*. 16, 116.
- Laio, A., Parrinello, M., 2002. Escaping free-energy minima. *Prot Natl Acad Sci U S A* 99, 12562–12566.
- LaValle, S.M., Kuffner, J.J., 2001 May 2. Randomized kinodynamic planning. *Int. J. Robot. Res.* 20, 378–400.
- Lazaridis, T., Karplus, M., 2000. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* 10, 139–145.
- Le Guilloux, V., Schmidtke, P., Tuffery, P., 2009. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*. 10, 168.
- Leaver-Fay, A., Jacak, R., Stranges, P.B., Kuhlman, B., 2011. A generic program for multistate protein design. *Uversky VN, editor. PLoS One* 6 e20937.
- Lee, F.S., Chu, Z.T., Warshel, A., 1993. Microscopic and semimicroscopic calculations of electrostatic energies in proteins by the POLARIS and ENZYMIK programs. *J. Comput. Chem.* 14, 161–185.
- Lee, P.-H., Kuo, K.-L., Chu, P.-Y., Liu, E.M., Lin, J.-H., 2009. SLITHER: a web server for generating contiguous conformations of substrate molecules entering into deep active sites of proteins or migrating through channels in membrane transporters. *Nucleic Acids Res.* 37, W559–W564.
- Lehmann, M., Loch, C., Middendorp, A., Studer, D., Lassen, S.F., Pasamontes, L., et al., 2002. The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.* 15, 403–411.
- Lence, E., van der Kamp, M.W., Gonzalez-Bello, C., Mulholland, A.J., 2018. QM/MM simulations identify the determinants of catalytic activity differences between type II dehydroquinase enzymes. *Org. Biomol. Chem.* 16, 4443–4455.
- Levitt, M., Warshel, A., 1975. Computer simulation of protein folding. *Nature*. 253, 694–698.
- Li, Y., Fang, J., 2012. PROTS-RF: a robust model for predicting mutation-induced protein stability changes. *PLoS One* 7, e47247.
- Li, J., Fu, A., 2019. Zhang L. An overview of scoring functions used for protein–ligand interactions in molecular docking, *Interdiscip Sci Comput Life Sci.*
- Li, H., Chang, Y.-Y., Lee, J.Y., Bahar, I., Yang, L.-W., 2017. DynOmics: dynamics of structural proteome and beyond. *Nucleic Acids Res.* 45, W374–W80.
- Liao, M.L., Somero, G.N., Dong, Y.W., 2019. Comparing mutagenesis and simulations as tools for identifying functionally important sequence changes for protein thermal adaptation. *Proc. Natl. Acad. Sci. U. S. A.* 116, 679–688.
- Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. *R news.* 2, 18–22.

- Lionta, E., Spyrou, G., Vassiliadis, D., Cournia, Z., 2014. Structure-based virtual screening for drug discovery: principles, applications and recent advances. *Curr. Top. Med. Chem.* 14, 1923–1938.
- Liskova, V., Bednar, D., Prudnikova, T., Rezacova, P., Koudelakova, T., Sebestova, E., et al., 2015. Balancing the stability-activity trade-off by fine-tuning dehalogenase access tunnels. *ChemCatChem* 7, 648–659.
- Liu, H., 2015. On statistical energy functions for biomolecular modeling and design. *Quant Biol.* 3, 157–167.
- Liu, Y., Kuhlman, B., 2006. RosettaDesign server for protein design. *Nucleic Acids Res.* 34, W235–W238.
- Liu, T., Lin, Y., Wen, X., Jorissen, R.N., Gilson, M.K., 2007. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* 35, D198–D201.
- Liu, Y., Du, J., Yan, M., Lau, M.Y., Hu, J., Han, H., et al., 2013. Biomimetic enzyme nanocomplexes and their use as antidotes and preventive measures for alcohol intoxication. *Nat. Nanotechnol.* 8, 187–192.
- Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., et al., 2015. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* 31, 405–412.
- Löffler, P., Schmitz, S., Hupfeld, E., Sterner, R., Merkl, R., 2017. Rosetta:MSF: a modular framework for multi-state computational protein design. *PLoS Comput. Biol.* 13, e1005600.
- Lu, G., Moriyama, E.N., 2004. Vector NTI, a balanced all-in-one sequence analysis suite. *Brief. Bioinform.* 5, 378–388.
- Lu, X., Fang, D., Ito, S., Okamoto, Y., Ovchinnikov, V., Cui, Q., 2016. QM/MM free energy simulations: recent progress and challenges. *Mol. Simul.* 42, 1056–1078.
- Lüdemann, S.K., Carugo, O., Wade, R.C., 1997. Substrate access to cytochrome p450cam: a comparison of a thermal motion pathway analysis with molecular dynamics simulation data. *J. Mol. Model.* 3, 369–374.
- Lüdemann, S.K., Lounnas, V., Wade, R.C., 2000. How do substrates enter and products exit the buried active site of cytochrome P450cam? *J. Mol. Biol.* 303, 797–811.
- Luzhkov, V., Åqvist, J., 1998. Computer simulation of phenyl ester cleavage by β -cyclodextrin in solution. *J. Am. Chem. Soc.* 120, 6131–6137.
- Lyne, P.D., Hodosecek, M., Karplus, M., 1999. A Hybrid QM–MM Potential Employing Hartree–Fock or Density Functional Methods in the Quantum Region. *J. Phys. Chem. A* 103, 3462–3471.
- MacKerell, A.D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., et al., 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* 102, 3586–3616.
- Magnan, C.N., Randall, A., Baldi, P., 2009. SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics* 25, 2200–2207.
- Mak, W.S., Siegel, J.B., 2014. Computational enzyme design: transitioning from catalytic proteins to enzymes. *Curr. Opin. Struct. Biol.* 27, 87–94.
- Malisi, C., Schumann, M., Toussaint, N.C., Kageyama, J., Kohlbacher, O., Höcker, B., 2012. Binding pocket optimization by computational protein design. *PLoS One* 7, e52505.
- Mardt, A., Pasquali, L., Wu, H., Noé, F., 2018. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* 9, 5.
- Maria-Solano, M.A., Serrano-Hervás, E., Romero-Rivera, A., Iglesias-Fernández, J., Osuna, S., 2018. Role of conformational dynamics in the evolution of novel enzyme function. *Chem. Commun. (Camb.)* 54, 6622–6634.
- Marques, S.M., Brezovsky, J., Damborsky, J., 2016. Role of tunnels and gates in enzymatic catalysis. editor. In: Svendsen, A. (Ed.), *Understanding Enzymes: Function, Design, Engineering, and Analysis*. Pan Stanford Publishing, pp. 421–463.
- Marques, S.M., Bednar, D., Damborsky, J., 2019. Computational study of protein-ligand unbinding for enzyme engineering. *Front Chem.* 6, 650.
- Maurer-Stroh, S., Debulpaep, M., Kuemmerer, N., Lopez de la Paz M., Martins, I.C., Reumers, J., et al., 2010. Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods* 7, 237–242.
- Maximova, T., Moffatt, R., Ma, B., Nussinov, R., Shehu, A., 2016. Principles and overview of sampling methods for modeling macromolecular structure and dynamics. *PLoS Comput. Biol.* 12, e1004619.
- Mazurenko, S., Prokop, Z., Damborsky, J., 2020. Machine learning in enzyme engineering. *ACS Catal.* 10, 1210–1223.
- McCammon, J.A., Gelin, B.R., Karplus, M., 1977. Dynamics of folded proteins. *Nature* 267, 585–590.
- Mendes, J., Guerois, R., Serrano, L., 2002. Energy estimation in protein design. *Curr. Opin. Struct. Biol.* 12, 441–446.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.* 21, 1087–1092.
- Mills, M., Andricioaei, I., 2008. An experimentally guided umbrella sampling protocol for biomolecules. *J. Chem. Phys.* 129, 114101.
- Miton, C.M., Jonas, S., Fischer, G., Duarte, F., Mohamed, M.F., van Loo, B., et al., 2018. Evolutionary repurposing of a sulfatase: A new Michaelis complex leads to efficient transition state charge offset. *Proc. Natl. Acad. Sci.* 115, E7293–E7302.
- Mitra, P., Shultz, D., Zhang, Y., 2013. EvoDesign: de novo protein design based on structural and evolutionary profiles. *Nucleic Acids Res.* 41, W273–W280.
- Monticelli, L., Tieleman, D.P., 2013. Force fields for classical molecular dynamics. In: *Biomolecular Simulations*, pp. 197–213.
- Morris, G.M., Goodsell, D.S., Halliday, R.S., Huey, R., Hart, W.E., Belew, R.K., et al., 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* 19, 1639–1662.
- Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., et al., 2009. AutoDock4 and AutoDockTools4: automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791.
- Murakami, S., Shimamoto, T., Nagano, H., Tsuruno, M., Okuhara, H., Hatanaka, H., et al., 2015. Producing human ceramide-NS by metabolic engineering using yeast *Saccharomyces cerevisiae*. *Sci. Rep.* 5, 16319.
- Murphy, P.M., Bolduc, J.M., Gallaher, J.L., Stoddard, B.L., Baker, D., 2009. Alteration of enzyme specificity by computational loop remodeling and design. *Proc. Natl. Acad. Sci. U. S. A.* 106, 9215–9220.
- Musil, M., Stourac, J., Bendl, J., Brezovsky, J., Prokop, Z., Zendulka, J., et al., 2017. FireProt: web server for automated design of thermostable proteins. *Nucleic Acids Res.* 45, W393–W399.
- Musil, Milos, Khan, Rayyan Tariq, Beier, Andy, Stourac, Jan, Konegger, Hannes, Damborsky, Jiri, Bednar, David, 2020. FireProt^{ASR}. A web server for fully automated ancestral sequence reconstruction. *Brief. Bioinformatics*, bbaa337. <https://doi.org/10.1093/bib/bbaa337>.
- Naqvi, A.A.T., Mohammad, T., Hasan, G.M., Hassan, M.I., 2018. Advancements in docking and molecular dynamics simulations towards ligand-receptor interactions and structure-function relationships. *Curr. Top. Med. Chem.* 18, 1755–1768.
- Neun, S., Zurek, P.J., Kaminski, T.S., Hoffelder, F., 2020. Chapter Thirteen - Ultrahigh throughput screening for enzyme function in droplets. Tawfik DSBT-M in E, editor. In: *Enzyme Engineering and Evolution: General Methods*. Academic Press, pp. 317–343.
- Nguyen, V., Wilson, C., Hoemberger, M., Stiller, J.B., Agafonov, R.V., Kutter, S., et al., 2017. Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science* 355, 289–294.
- Nick Pace, C., Scholtz, J.M., Grimsley, G.R., 2014. Forces stabilizing proteins. *FEBS Lett.* 588, 2177–2184.
- Nosrati, G.R., Houk, K.N., 2012. SABER: A computational method for identifying active sites for new reactions. *Protein Sci.* 21, 697–706.
- Nov, Y., 2014. Probabilistic methods in directed evolution: library size, mutation rate, and diversity. In: *Directed Evolution Library Creation*, pp. 261–278.
- Nussinov, R., Tsai, C.-J., 2013. Allostery in disease and in drug discovery. *Cell* 153, 293–305.
- Okabe, A., Boots, B., Sugihara, K., Chiu, S.N., 2000. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2nd edition. Wiley.
- Ollikainen, N., de Jong, R.M., Kortemme, T., 2015. Coupling protein side-chain and backbone flexibility improves the re-design of protein-ligand specificity. *PLoS Comput. Biol.* 11, e1004335.
- Oostenbrink, C., Villa, A., Mark, A.E., van Gunsteren, W.F., 2004. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* 25, 1656–1676.
- Osuna, S., 2020. The challenge of predicting distal active site mutations in computational enzyme design. *WIREs Comput. Mol. Sci.*, e1502 <https://doi.org/10.1002/wcms.1502> (In press).
- Ozen, A., Gönen, M., Alpaydan, E., Haliloğlu, T., 2009. Machine learning integration for predicting the effect of single amino acid substitutions on protein stability. *BMC Struct. Biol.* 9, 66.
- Pagadala, N.S., Syed, K., Tuszynski, J., 2017. Software for molecular docking: a review. *Biophys. Rev.* 9, 91–102.
- Paladin, L., Piovesan, D., Tosatto, S.C.E., 2017. SODA: prediction of protein solubility from disorder and aggregation propensity. *Nucleic Acids Res.* 45, W236–W240.
- Palazzesi, F., Salvaggio, M., Barducci, A., Parrinello, M., 2016. Communication: role of explicit water models in the helix folding/unfolding processes. *J. Chem. Phys.* 145, 121101.
- Pantazes, R.J., Grisewood, M.J., Li, T., Gifford, N.P., Maranas, C.D., 2015. The iterative protein redesign and optimization (IPRO) suite of programs. *J. Comput. Chem.* 36, 251–263.
- Paquet, E., Viktor, H.L., 2015. Molecular dynamics, Monte Carlo simulations, and langevin dynamics: a computational review. *Biomed. Res. Int.* 2015, 1–18.
- Parthiban, V., Gromiha, M.M., Schomburg, D., 2006. CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Res.* 34, W239–W242.
- Pei, J., 2008. Multiple protein sequence alignment. *Curr. Opin. Struct. Biol.* 18, 382–386.
- Pellegrini-Calace, M., Maiwald, T., Thornton, J.M., 2009. PoreWalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. *PLoS Comput. Biol.* 5, e1000440.
- Petrovic, D., Risso, V.A., Kamerlin, S.C.L., Sanchez-Ruiz, J.M., 2018. Conformational dynamics and enzyme evolution. *J. R. Soc. Interface* 15, 20180330.
- Pey, A.L., Rodriguez-Larrea, D., Bomke, S., Dammers, S., Godoy-Ruiz, R., Garcia-Mira, M., et al., 2008. Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins* 71, 165–174.
- Phillips, J.C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., et al., 2005. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* 26, 1781–1802.
- Pietrucci, F., 2017. Strategies for the exploration of free energy landscapes: unity in diversity and challenges ahead. *Rev Phys* 2, 32–45.
- Pines, G., Pines, A., Garst, A.D., Zeitoun, R.I., Lynch, S.A., Gill, R.T., 2015. Codon compression algorithms for saturation mutagenesis. *ACS Synth. Biol.* 4, 604–614.
- Pinto, G., Mazzone, G., Russo, N., Toscano, M., 2017. Trimethylphosphate and dimethylphosphate hydrolysis by binuclear Cd II, Mn II, and Zn II-Fe II promiscuous organophosphate-degrading enzyme: reaction mechanisms. *Chem - A Eur J.* 23, 13742–13753.
- Pires, D.E.V., Ascher, D.B., Blundell, T.L., 2014. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 30, 335–342.
- Prokop, Z., Gora, A., Brezovsky, J., Chaloupkova, R., Stepankova, V., Damborsky, J., 2012. Engineering of protein tunnels: Keyhole-lock-key model for catalysis by the enzymes with buried active sites. editors. In: Lutz, S., Bornscheuer, U.T. (Eds.), *Protein Engineering Handbook*. Weinheim: Wiley-VCH, p. 421.

- Pucci, F., Bernaerts, K.V., Kwasigroch, J.M., Rooman, M., 2018. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics*. 34, 3659–3665.
- Pujadas, G., Vaque, M., Ardevol, A., Blade, C., Salvado, M., Blay, M., et al., 2008. Protein-ligand docking: a review of recent advances and future perspectives. *Curr. Pharm. Anal.* 4, 1–19.
- Purg, M., SCL, Kamerlin, 2018. Chapter One - Empirical Valence Bond Simulations of Organophosphate Hydrolysis: Theory and Practice. Allen KNTB-M in E, editor. In: *Phosphatases*. Academic Press, pp. 3–51.
- Quan, L., Lv, Q., Zhang, Y., 2016. STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*. 32, 2936–2946.
- Rarey, M., Kramer, B., Lengauer, T., Klebe, G., 1996. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* 261, 470–489.
- Rauschel, F.M., Thoden, J.B., Holden, H.M., 2003. Enzymes with molecular tunnels. *Acc. Chem. Res.* 36, 539–548.
- Reetz, M.T., 2012. Artificial metalloenzymes as catalysts in stereoselective diels-alder reactions. *Chem. Rec.* 12, 391–406.
- Reetz, M.T., Carballeira, J.D., 2007. Iterative saturation mutagenesis (ISM) for rapid directed evolution of functional enzymes. *Nat. Protoc.* 2, 891–903.
- Reetz, M.T., Wu, S., 2008. Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. *Chem. Commun.* 5499–5502.
- Rice, P., Longden, I., Bleasby, A., 2000. EMBOS: the European molecular biology open software suite. *Trends Genet.* 16, 276–277.
- Richter, F., Leaver-Fay, A., Khare, S.D., Bjelic, S., Baker, D., 2011. De novo enzyme design using Rosetta3. *PLoS One* 6 (5), e19230.
- Ringe, D., Petsko, G.A., 2008. Biochemistry: how enzymes work. *Science*. 320, 1428–1429.
- Rognan, D., 2017. The impact of in silico screening in the discovery of novel and safer drug candidates. *Pharmacol. Ther.* 175, 47–66.
- Rohl, C.A., Strauss, C.E.M., Chivian, D., Baker, D., 2004. Modeling structurally variable regions in homologous proteins with rosetta. *Proteins Struct Funct Bioinforma.* 55, 656–677.
- Romero, P.A., Arnold, F.H., 2009. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.* 10, 866–876.
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., et al., 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. A.
- Roston, D., Cui, Q., 2016. Chapter Nine - QM/MM Analysis of Transition States and Transition State Analogues in Metalloenzymes. Voth GABT-M in E, editor. In: *Computational Approaches for Studying Enzyme Mechanism Part A*. Academic Press, pp. 213–250.
- Röthlisberger, D., Khersonsky, O., Wollacott, A.M., Jiang, L., DeChance, J., Betker, J., et al., 2008. Kemp elimination catalysts by computational enzyme design. *Nature*. 453, 190–195.
- Rueda, M., Chacón, P., Orozco, M., 2007. Thorough validation of protein normal mode analysis: a comparative study with essential dynamics. *Structure*. 15, 565–575.
- Saier, M.H., 2000. Families of proteins forming transmembrane channels. *J. Membr. Biol.* 175, 165–180.
- Saikia, S., Bordoloi, M., 2019. Molecular docking: challenges, advances and its use in drug discovery perspective. *Curr. Drug Targets* 20, 501–521.
- Salomon-Ferrer, R., Case, D.A., Walker, R.C., 2013. An overview of the Amber biomolecular simulation package. *Wiley Interdiscip Rev Comput Mol Sci.* 3, 198–210.
- Sammond, D.W., Eletr, Z.M., Purbeck, C., Kimple, R.J., Siderovski, D.P., Kuhlman, B., 2007. Structure-based protocol for identifying mutations that enhance protein–protein binding affinities. *J. Mol. Biol.* 371, 1392–1404.
- Schymkowitz, J., Borg, J., Stricher, F., Nys, R., Rousseau, F., Serrano, L., 2005. The FoldX web server: an online force field. *Nucleic Acids Res.* 33, W382–W388.
- Sehnal, D., Svoboda Varekova, R., Berka, K., Pravda, L., Navratilova, V., Banas, P., et al., 2013. MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *Aust. J. Chem.* 5, 39.
- Shao, Y., Gan, Z., Epifanovsky, E., Gilbert, A.T.B., Wormit, M., Kussmann, J., et al., 2015. Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* 113, 184–215.
- Shirke, A.N., Basore, D., Butterfoss, G.L., Bonneau, R., Byströf, C., Gross, R.A., 2016. Toward rational thermostabilization of *Aspergillus oryzae* cutinase: insights into catalytic and structural stability. *Proteins*. 84, 60–72.
- Siegbahn, P.E.M., Him, F., 2009. Recent developments of the quantum chemical cluster approach for modeling enzyme reactions. *J. Biol. Inorg. Chem.* 14, 643–651.
- Siegbahn, P.E.M., Him, F., 2011. The quantum chemical cluster approach for modeling enzyme reactions. *Wiley Interdiscip Rev Comput Mol Sci.* 1, 323–336.
- Siegel, J.B., Zanghellini, A., Lovick, H.M., Kiss, G., Lambert, A.R., St.Clair, J.L., et al., 2010. Computational design of an enzyme catalyst for a stereoselective bimolecular diels-alder reaction. *Science*. 329, 309–313.
- Sinha, R., Shukla, P., 2019. Current trends in protein engineering: updates and progress. *Curr. Protein Pept. Sci.* 20, 398–407.
- Skovstrup, S., David, L., Taboureau, O., Jørgensen, F.S., 2012. A steered molecular dynamics study of binding and translocation processes in the GABA transporter. *PLoS One* 7, e39360.
- Skyner, R.E., McDonagh, J.L., Groom, C.R., van Mourik, T., Mitchell, J.B., 2015. A review of methods for the calculation of solution free energies and the modelling of systems in solution. *Phys. Chem. Chem. Phys.* 17, 6174–6191.
- Smialowski, P., Doose, G., Torkler, P., Kaufmann, S., Frishman, D., 2012. PROSO II—a new method for protein solubility prediction. *FEBS J.* 279, 2192–2200.
- Sormanni, P., Aprile, F.A., Vendruscolo, M., 2015. The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* 427, 478–490.
- Sousa, S.F., Fernandes, P.A., Ramos, M.J., 2006. Protein-ligand docking: current status and future challenges. *Proteins Struct Funct Bioinforma.* 65, 15–26.
- Sousa, S.F., Ribeiro, A.J.M., Neves, R.P.P., Brás, N.F., NMFS, Cerqueira, Fernandes, P. A., et al., 2017. Application of quantum mechanics/molecular mechanics methods in the study of enzymatic reaction mechanisms. *Wiley Interdiscip Rev Comput Mol Sci* 7, e1281.
- Spivok, V., Sucur, Z., Hosek, P., 2015. Enhanced sampling techniques in biomolecular simulations. *Biotechnol. Adv.* 33, 1130–1140.
- Stahura, F., Bajorath, J., 2005. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* 11, 1189–1202.
- Stamatakis, A., 2006. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22, 2688–2690.
- Stamatakis, A., 2014. RAXML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30, 1312–1313.
- Stone, J.E., Phillips, J.C., Freddolino, P.L., Hardy, D.J., Trabuco, L.G., Schulten, K., 2007. Accelerating molecular modeling applications with graphics processors. *J. Comput. Chem.* 28, 2618–2640.
- Stourac, J., Vavra, O., Kokkonen, P., Filipovic, J., Pinto, G., Brezovsky, J., et al., 2019. Caver Web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. *Nucleic Acids Res.* 47, W414–W422.
- Stryer, L., Berg, J., Tymoczko, J., 2002. Biochemistry, 5th ed. W.H. Freeman, San Francisco. (8.1).
- Suchard, M.A., Redelings, B.D., 2006. BALI-Phy: simultaneous Bayesian inference of alignment and phylogeny. *Bioinformatics*. 22, 2047–2048.
- Sugita, Y., Okamoto, Y., 1999. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314, 141–151.
- Sullivan, B.J., Nguyen, T., Durani, V., Mathur, D., Rojas, S., Thomas, M., et al., 2012. Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J. Mol. Biol.* 420, 384–399. A.
- Sultan, M., Pande, V.S., 2017. tICA-metadynamics: accelerating metadynamics by using kinetically selected collective variables. *J. Chem. Theory Comput.* 13, 2440–2447.
- Sumbalova, L., Stourac, J., Martinek, T., Bednar, D., Damborsky, J., 2018. HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information. *Nucleic Acids Res.* 46, W356–W362.
- Swendsen, R.H., Wang, J.-S., 1986. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* 57, 2607–2609.
- Świderek, K., Tuñón, I., Moliner, V., Bertran, J., 2015. Computational strategies for the design of new enzymatic functions. *Arch. Biochem. Biophys.* 582, 68–79.
- Tantillo, D.J., Chen, J., Houk, K.N., 1998. Theozymes and compozymes: theoretical models for biological catalysis. *Curr. Opin. Chem. Biol.* 2, 743–750.
- Teng, S., Srivastava, A.K., Wang, L., 2010. Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics* 11, S5.
- Tian, J., Wu, N., Chu, X., Fan, Y., 2010a. Predicting changes in protein thermostability brought about by single- or multi-site mutations. *BMC Bioinformatics*. 11, 370.
- Tian, Y., Deutsch, C., Krishnamoorthy, B., 2010b. Scoring function to predict solubility mutagenesis. *Algorithms Mol Biol.* 5, 33.
- Tinberg, C.E., Khare, S.D., Dou, J., Doyle, L., Nelson, J.W., Schena, A., et al., 2013. Computational design of ligand-binding proteins with high affinity and selectivity. *Nature*. 501, 212–216.
- Tirion, M.M., 1996. Large amplitude elastic motions in proteins from a single-parameter. *Atomic Analysis. Phys Rev Lett.* 77, 1905–1908.
- Tiwari, M.K., Singh, R., Singh, R.K., Kim, I.-W., Lee, J.-K., 2012. Computational approaches for rational design of proteins with novel functionalities. *Comput Struct Biotechnol J.* 2, e201204002.
- Tokuriki, N., Stricher, F., Serrano, L., Tawfik, D.S., 2008. How protein stability and new functions trade off. *PLoS Comput. Biol.* 4, e1000002.
- Torrie, G.M., Valleau, J.P., 1977. Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J. Comput. Phys.* 23, 187–199.
- Trott, O., Olson, A.J., 2009. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31, 455–461.
- Tsallis, C., Stariolo, D.A., 1996. Generalized simulated annealing. *Phys A Stat Mech its Appl.* 233, 395–406.
- Vaissier Welborn, V., Head-Gordon, T., 2018. Computational design of synthetic enzymes. *Chem. Rev.* 119, 6613–6630.
- Valiev, M., Bylaska, E.J., Govind, N., Kowalski, K., Straatsma, T.P., Van Dam, H.J.J., et al., 2010. NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* 181, 1477–1489.
- Valsson, O., Tiwary, P., Parrinello, M., 2016. Enhancing important fluctuations: rare events and metadynamics from a conceptual viewpoint. *Annu. Rev. Phys. Chem.* 67, 159–184.
- Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A.E., Berendsen, H.J.C., 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26, 1701–1718.
- Van Durme, J., De Baets, G., Van Der Kant, R., Ramakers, M., Ganesan, A., Wilkinson, H., et al., 2016. Solubis: a webserver to reduce protein aggregation through mutation. *Protein Eng Des Sel* 29, 285–289. A.
- Vasina, M., Vanacek, P., Damborsky, J., Prokop, Z., 2020. Chapter Three - Exploration of enzyme diversity: High-throughput techniques for protein production and microscale biochemical characterization. Tawfik DSBT-M in E, editor. In: *Enzyme Engineering and Evolution: General Methods*. Academic Press, pp. 51–85.
- Vavra, O., Filipovic, J., Plhak, J., Bednar, D., Marques, S.M., Brezovsky, J., et al., 2019. A molecular docking-based tool to analyse ligand transport through protein tunnels and channels. editor. In: Ponty, Y. (Ed.), *Bioinformatics. CaverDock*.

- Verdonk, M.L., Cole, J.C., Hartshorn, M.J., Murray, C.W., Taylor, R.D., 2003. Improved protein-ligand docking using GOLD. *Proteins Struct Funct Bioinforma.* 52, 609–623.
- Verma, R., Schwaneberg, U., Roccatano, D., 2012. Computer-aided protein directed evolution: a review of web serverS, databases and other computational tools for protein engineering. *Comput Struct Biotechnol J* 2, e201209008.
- Vlachakis, D., Bencurova, E., Papangelopoulos, N., Kossida, S., 2014. Current state-of-the-art molecular dynamics methods and applications. In: *Advances in Protein Chemistry and Structural Biology*, pp. 269–313.
- Volkamer, A., Kuhn, D., Grombacher, T., Rippmann, F., Rarey, M., 2012. Combining global and local measures for structure-based druggability predictions. *J. Chem. Inf. Model.* 52, 360–372.
- Wainreb, G., Wolf, L., Ashkenazy, H., Dehouck, Y., Ben-Tal, N., 2011. Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. *Bioinformatics.* 27, 3286–3292.
- Walsh, I., Seno, F., Tosatto, S.C.E., Trovato, A., 2014. PASTA 2.0: an improved server for protein aggregation prediction. *Nucleic Acids Res.* 42, W301–W307.
- Wang, F., Landau, D.P., 2001. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* 86, 2050–2053.
- Wang, X., Li, R., Cui, W., Li, Q., Yao, J., 2018. QM/MM free energy simulations of an efficient gluten hydrolase (Kuma030) implicate for a reactant-state based protein-design strategy for general acid/base catalysis. *Sci. Rep.* 8, 7042.
- Wang, S., He, J., Shen, C., Manefield, M.J., 2019. Editorial: Organohalide Respiration: New Findings in Metabolic Mechanisms and Bioremediation Applications. *Front. Microbiol.* 10, 526.
- Warshel, A., 1976. Bicycle-pedal model for the first step in the vision process. *Nature.* 260, 679–683.
- Warshel, A., Levitt, M., 1976. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.* 103, 227–249.
- Warshel, A., Weiss, R.M., 1981. Empirical valence bond calculations of enzyme catalysis. *Ann. N. Y. Acad. Sci.* 367, 370–382.
- Westmaier, Y., Barril, X., Scapozza, L., 2015. Virtual screening: an in silico tool for interlacing the chemical universe with the proteome. *Methods.* 71, 44–57.
- Westesson, O., Barquist, L., Holmes, I., 2012. HandAlign: Bayesian multiple sequence alignment, phylogeny and ancestral reconstruction. *Bioinformatics* 28, 1170–1171.
- Wijma, H.J., Janssen, D.B., 2013. Computational design gains momentum in enzyme catalysis engineering. *FEBS J.* 280, 2948–2960.
- Wijma, H.J., Floor, R.J., Jekel, P.A., Baker, D., Marrink, S.J., Janssen, D.B., 2014. Computationally designed libraries for rapid enzyme stabilization. *Protein Eng Des Sel.* 27, 49–58.
- Wijma, H.J., Floor, R.J., Bjelic, S., Marrink, S.J., Baker, D., Janssen, D.B., 2015. Enantioselective enzymes by computational design and in silico screening. *Angew. Chem. Int. Ed. Eng.* 54, 3726–3730.
- Witvliet, D.K., Strokach, A., Giraldo-Forero, A.F., Teyra, J., Colak, R., Kim, P.M., 2016. ELASPIC web-server: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. *Bioinformatics.* 32, 1589–1591.
- Wolfenden, R., Snider, M.J., 2001. The depth of chemical time and the power of enzymes as catalysts. *Acc. Chem. Res.* 34, 938–945.
- Woolfson, D.N., Bartlett, G.J., Burton, A.J., Heal, J.W., Niitsu, A., Thomson, A.R., et al., 2015. De novo protein design: how do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.* 33, 16–26.
- Workalemahu, G., Wang, H., Puan, K.-J., Nada, M.H., Kuzuyama, T., Jones, B.D., et al., 2014. Metabolic engineering of *Salmonella* vaccine bacteria to boost human Vγ2Vδ2 T cell immunity. *J. Immunol.* 193, 708–721.
- Xie, Z.-R., Hwang, M.-J., 2015. Methods for predicting protein-ligand binding sites. In: *Molecular Modeling of Proteins*, pp. 383–398.
- Xie, T., France-Lanord, A., Wang, Y., Shao-Horn, Y., Grossman, J.C., 2019. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nat. Commun.* 10, 2667.
- Xu, B., Yang, Z., 2013. PAMLX: a graphical user interface for PAML. *Mol. Biol. Evol.* 30, 2723–2724.
- Xu, T., Zhang, L., Wang, X., Wei, D., Li, T., 2009. Structure-based substrate screening for an enzyme. *BMC Bioinformatics* 10, 257.
- Xu, Z., Xue, Y.-P., Zou, S.-P., Zheng, Y.-G., 2020. In: Singh, S.P., Pandey, A., Singhanian, R. R., Larroche, C., Biofuels, Li, ZBT-B, Biochemicals (Eds.), Chapter 5 - Enzyme engineering strategies to confer thermostability. Elsevier, pp. 67–89 editors.
- Xue, L., Ko, M.-C., Tong, M., Yang, W., Hou, S., Fang, L., et al., 2011. Design, preparation, and characterization of high-activity mutants of human butyrylcholinesterase specific for detoxification of cocaine. *Mol. Pharmacol.* 79, 290–297.
- Yaffe, E., Fishelovitch, D., Wolfson, H.J., Halperin, D., Nussinov, R., 2008. MolAxis: a server for identification of channels in macromolecules. *Nucleic Acids Res.* 36, W210–W215.
- Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591.
- Yang, Y., Niroula, A., Shen, B., Vihinen, M., 2016. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. *Bioinformatics.* 32, 2032–2034.
- Yang, Y.I., Shao, Q., Zhang, J., Yang, L., Gao, Y.Q., 2019. Enhanced sampling in molecular dynamics. *J. Chem. Phys.* 151, 070902.
- Yin, S., Ding, F., Dokholyan, N.V., 2007. Eris: an automated estimator of protein stability. *Nat. Methods* 4, 466–467.
- Yu, H., Dalby, P.A., 2018a. Exploiting correlated molecular-dynamics networks to counteract enzyme activity-stability trade-off. *Proc. Natl. Acad. Sci. U. S. A.* 115, E12192–E12200.
- Yu, H., Dalby, P.A., 2018b. Coupled molecular dynamics mediate long- and short-range epistasis between mutations that affect stability and aggregation kinetics. *Proc. Natl. Acad. Sci. U. S. A.* 115, E11043–E11052.
- Yuan, Y., Pei, J., Lai, L., 2013. Binding site detection and druggability prediction of protein targets for structure-based drug design. *Curr. Pharm. Des.* 19, 2326–2333.
- Zambrano, R., Jamroz, M., Szczasiuk, A., Pujols, J., Kmiecik, S., Ventura, S., 2015. AGGRESAN3D (A3D): server for prediction of aggregation properties of protein structures. *Nucleic Acids Res.* 43, W306–W313.
- Zanghellini, A., Jiang, L., Wollacott, A.M., Cheng, G., Meiler, J., Althoff, E.A., Röthlisberger, D., Baker, D., 2006. New algorithms and an in silico benchmark for computational enzyme design. *Protein Sci.* 15, 2785–2794.
- Zeng, J., Li, Y., Zhang, J.Z.H., Mei, Y., 2016. Examination of the quality of various force fields and solvation models for the equilibrium simulations of GA88 and GB88. *J. Mol. Model.* 22, 177.
- Zhang, J., Klinman, J.P., 2011. Enzymatic methyl transfer: role of an active site residue in generating active site compaction that correlates with catalytic efficiency. *J. Am. Chem. Soc.* 133, 17134–17137.
- Zhang, Y., Voth, G.A., 2011. Combined metadynamics and umbrella sampling method for the calculation of ion permeation free energy profiles. *J. Chem. Theory Comput.* 7, 2277–2283.
- Zhang, H., Yin, C., Jiang, Y., van der Spoel, D., 2018. Force field benchmark of amino acids: I. Hydration and diffusion in different water models. *J. Chem. Inf. Model.* 58, 1037–1052.
- Zhang, Y., Ma, C., Dischert, W., Soucaille, P., Zeng, A., 2019. Engineering of phosphoserine aminotransferase increases the conversion of l-homoserine to 4-hydroxy-2-ketobutyrate in a glycerol-independent pathway of 1,3-propanediol production from glucose. *Biotechnol. J.* 1900003.
- Zheng, H., Reetz, M.T., 2010. Manipulating the stereoselectivity of limonene epoxide hydrolase by directed evolution based on iterative saturation mutagenesis. *J. Am. Chem. Soc.* 132, 15744–15751.
- Zheng, F., Xue, L., Hou, S., Liu, J., Zhan, M., Yang, W., et al., 2014. A highly efficient cocaine-detoxifying enzyme obtained by computational design. *Nat. Commun.* 5, 3457.
- Zhou, S., Liu, Z., Xie, W., Yu, Y., Ning, C., Yuan, M., et al., 2019. Improving catalytic efficiency and maximum activity at low pH of *Aspergillus* neoniger phytase using rational design. *Int. J. Biol. Macromol.* 131, 1117–1124.
- Zipse, H., Wang, L.H., Houk, K.N., 1996. Polyether catalysis of ester aminolysis – a computational and experimental study. *Liebigs Ann/Recl.* 1996, 1511–1522.