J3.3 Similarity Searching in Database

The topic Similarity Searching in Database contains the following 11 pages:

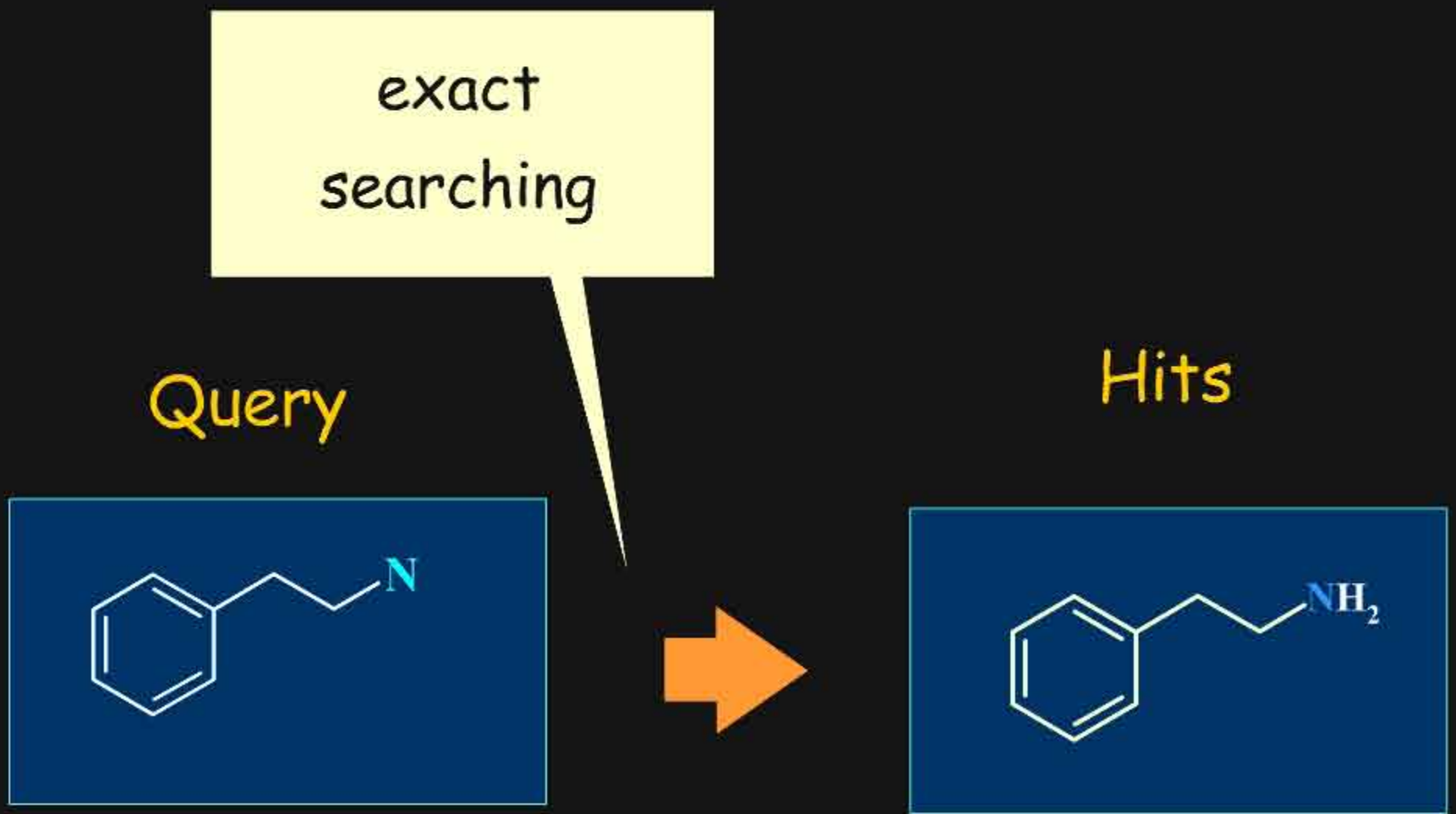
- Exact and Substructure Searching
- Similarity Searching
- Semi-Manual Similarity Searching
- Similarity Concept and Creativity
- Output of Similarity Searching
- Broad Range of Applications
- Substructure & Similarity Searching Complementarities
- General Requirements of a Method
  - Make Molecules Accessible to the Computer
  - Need of Methods to Measure Similarity
  - Apply an Algorithm

J3.3.1 Exact and Substructure Searching

Until the mid 1980s database searching was limited to exact and substructure search, i.e. the identification of all the molecules in the database that contain a specified substructure. Substructure searching is based on sub-graph algorithms that are computationally lengthy.

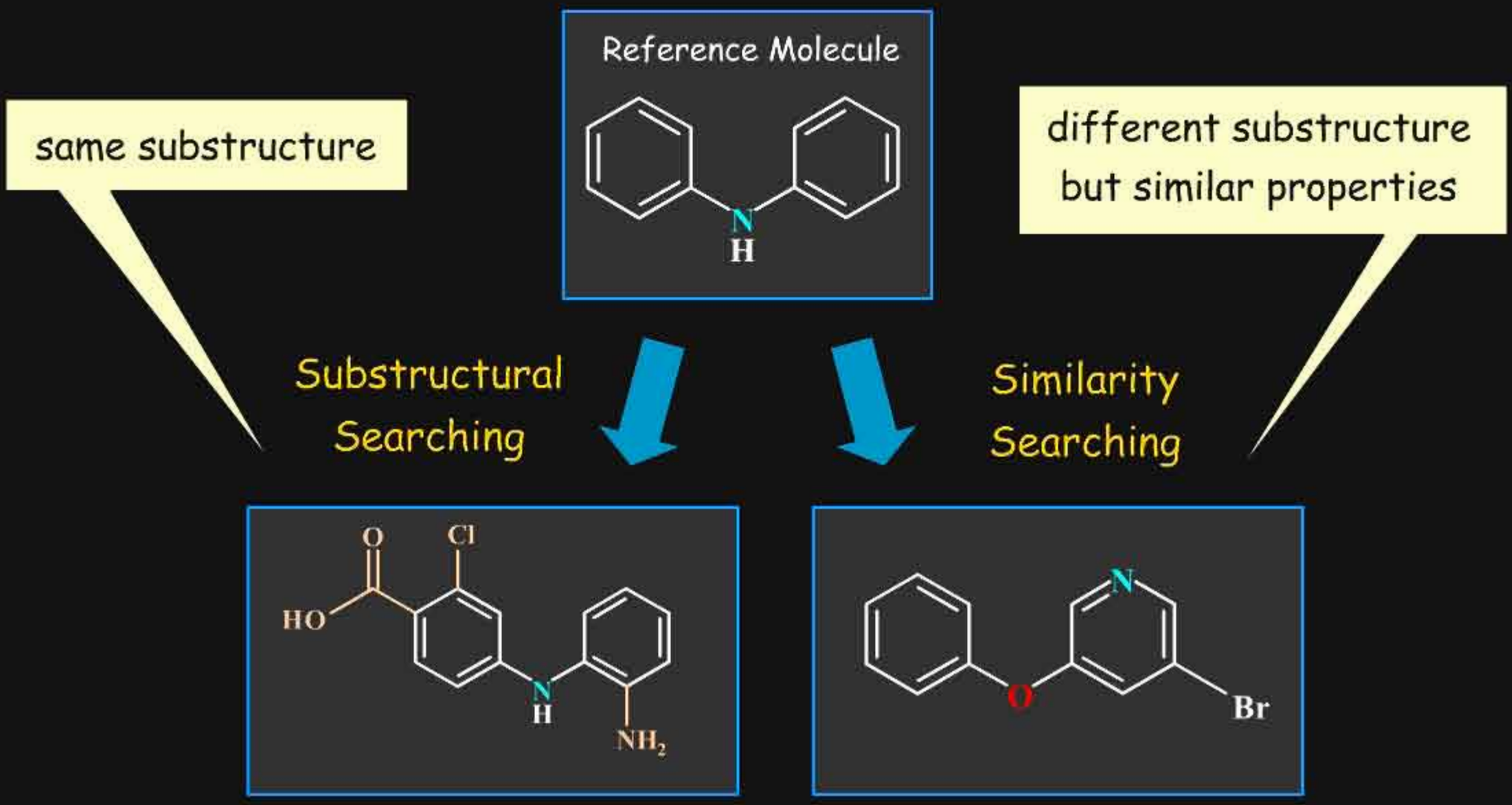
Exact Search

Substructure Search



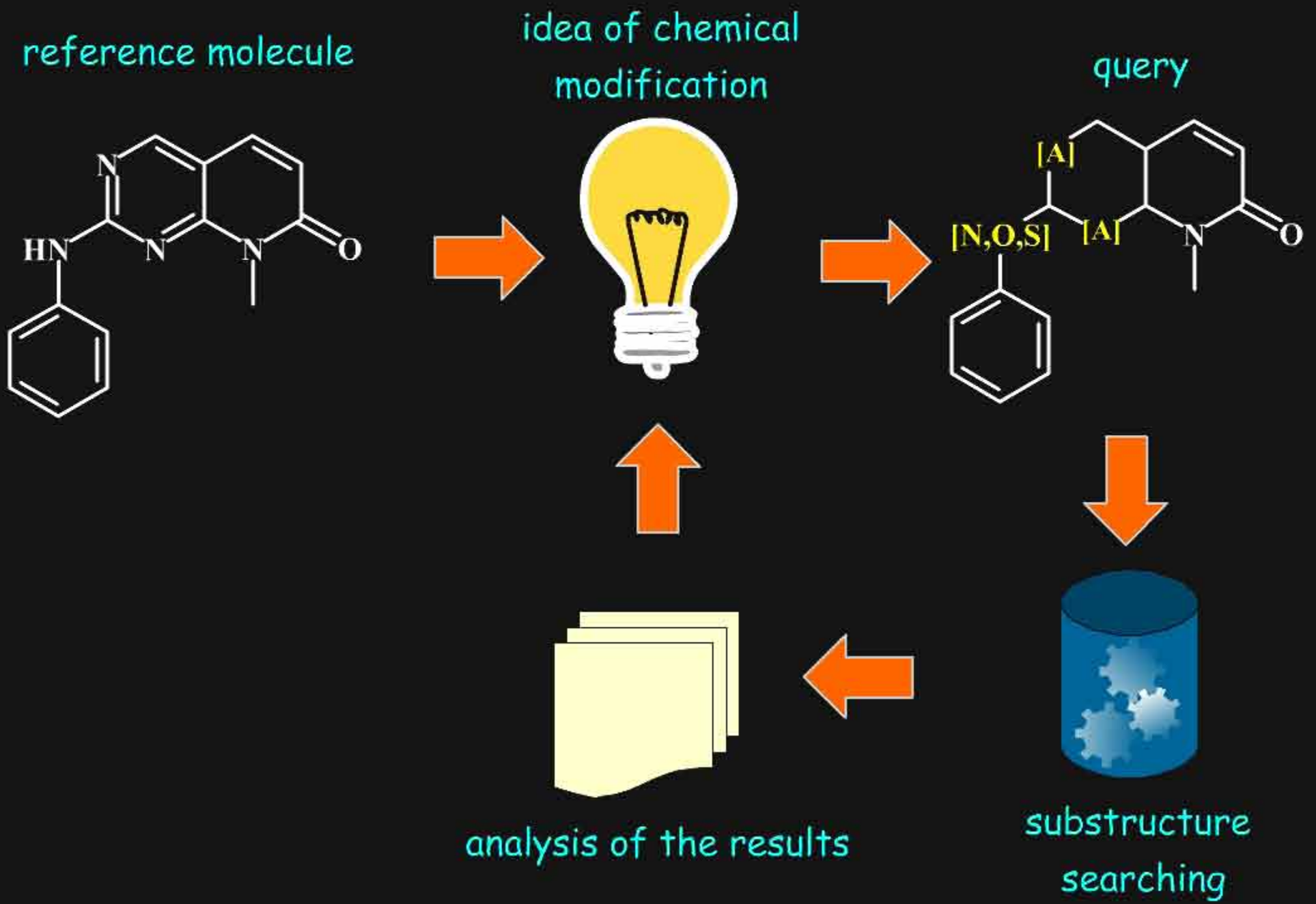
J3.3.2 Similarity Searching

The implementation of similarity principles in database searching was first introduced in the late 1980s. While substructural searching attempts exact retrieval of a certain substructure, in similarity searching, compounds (or parts of compounds) with different substructures but similar physico-chemical properties are identified.



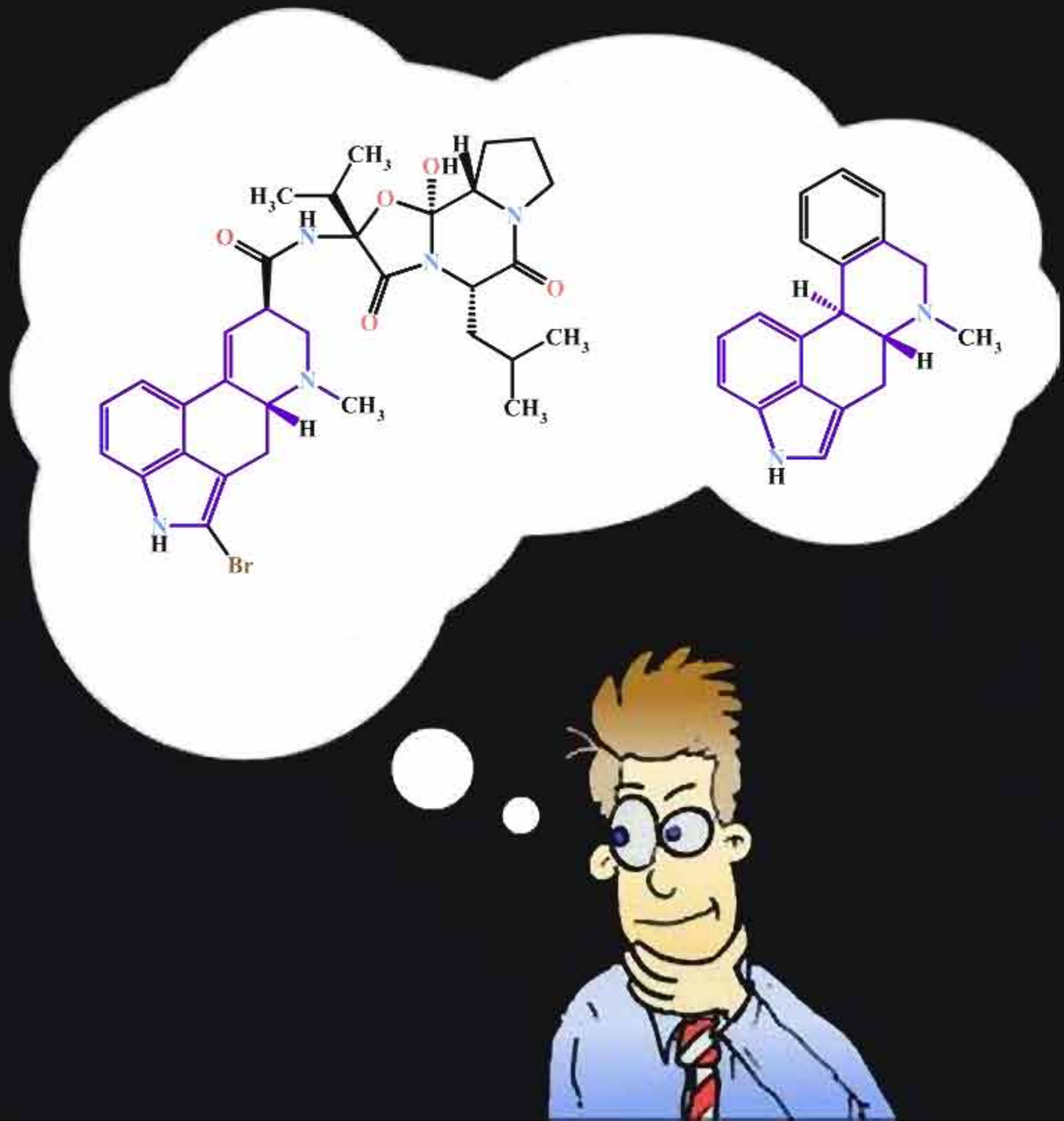
J3.3.3 Semi-Manual Similarity Searching

Note that the medicinal chemist did not wait for advanced computerized programs to start searching for similar molecules to a reference compound with different sub-structural patterns. For example, starting from a reference compound, the chemist imagined several possible chemical modifications he explored and analyzed by traditional substructural searching. This is a cumbersome and sometimes discouraging process (low hit rate). Moreover, the approach is biased by the chemist's imagination and is far from being systematic.



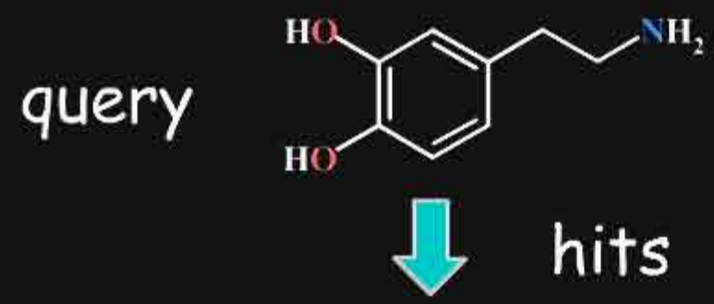
J3.3.4 Similarity Concept and Creativity

Since similarity searching is a fuzzy concept, the medicinal chemist does not know exactly what he is looking for. This in fact enables the chemist to extend and control creativity by focusing on new ideas and concepts.

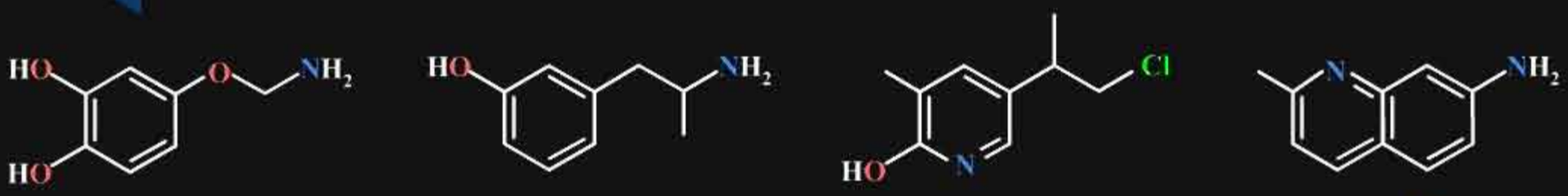


J3.3.5 Output of Similarity Searching

One of the big advantages of similarity searching is that the results are well structured, with scoring indicating the degree of similarity with the reference molecule. This makes it possible to analyze the results by order of importance. This is different from traditional substructure searches whose outputs were chaotic, consisting of the accumulation of results obtained from different sessions.



high similarity



decreasing priority

J3.3.6 Broad Range of Applications

Because of its simplicity of use and a computational speed that can routinely compare millions of compounds in minutes, database similarity searching is widely used in a great number of applications. Some of these applications are listed below and will be presented later in more detail.

- Searching molecules with similar properties
- Searching information from similar molecules
- Validation of novelty
- Knowing a pharmacophore, search for molecules
- Analysis of the diversity
- Peptidomimetics
- Filtering undesired hits
- Clustering

J3.7 Examples of Direct Use of Similarity Coefficients

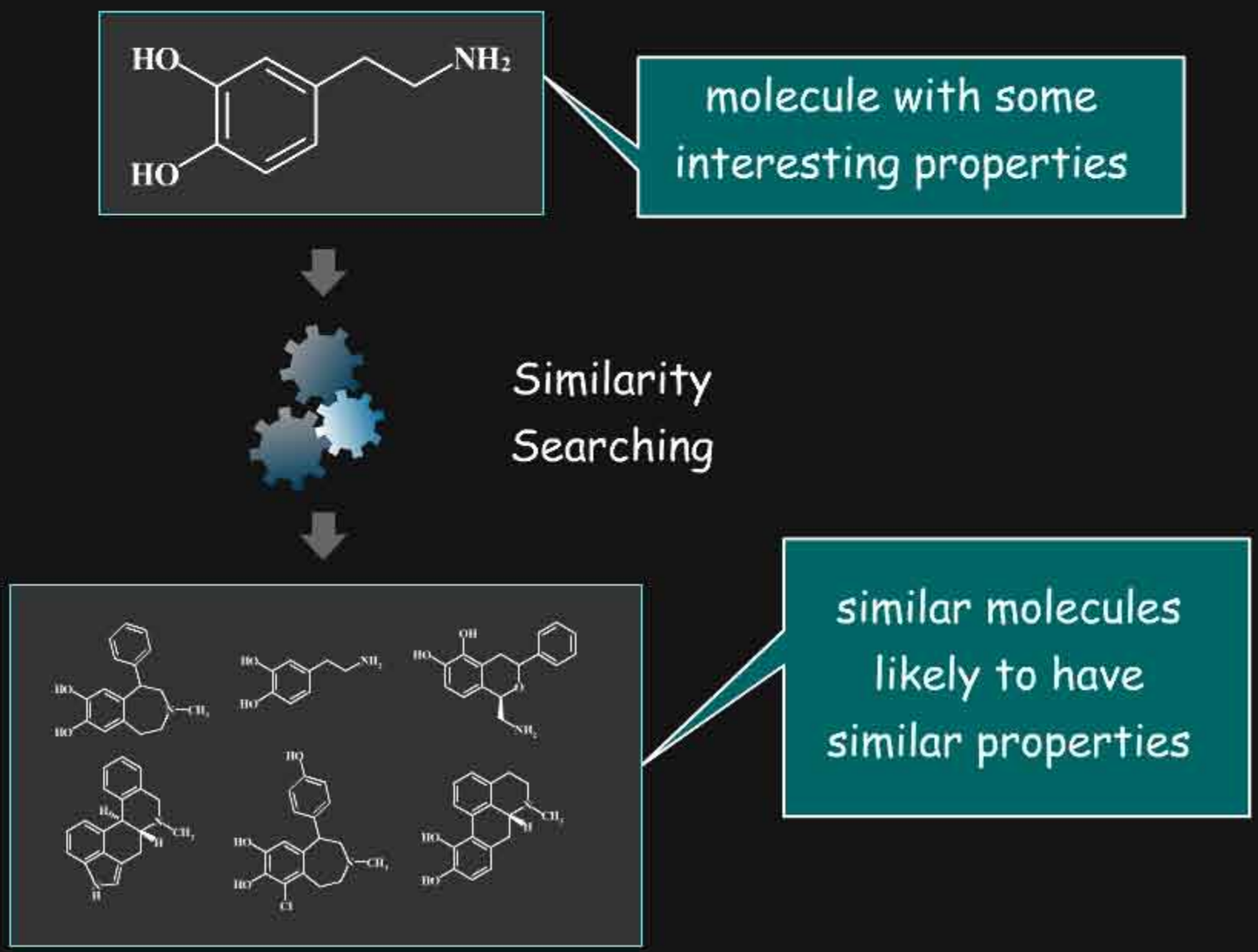
The topic Examples of Direct Use of Similarity Coefficients contains the following 11 pages:

- Searching Molecules with Similar Properties
- Searching Information from Similar Molecules
- Knowing a Pharmacophore, Search for Novel Molecules
- Example of a Fuzzy Pharmacophore
- Validation of Novelty
- Reducing a Virtual Library to a Practical Size
- Peptidomimetics
- Compounds that Fit the Shape of an Active Site
- Find a Synthetic Route
- Filtering Undesired Hits
- Clustering of Molecules

J3.7.1 Searching Molecules with Similar Properties

One of the most prominent applications of the similarity approach is the task of screening a database for molecules which are similar to a known active compound. In this case, individual similarities between the query compound and the library compounds are calculated and the most similar compounds are returned to the user. For example an active compound with undesired properties (e.g. toxicity, poor solubility) can be used as a query to retrieve molecules which are similar to the original molecule in that they show the desired activity, but dissimilar with respect to the undesired properties.

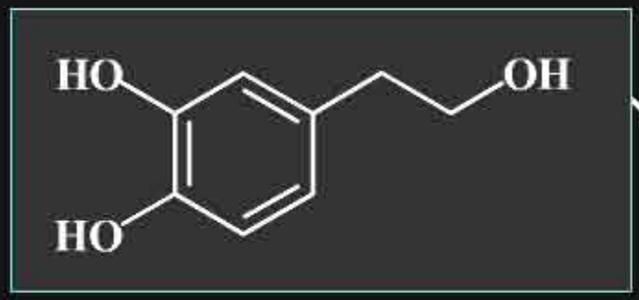
Scheme Space properties



J3.7.2 Searching Information from Similar Molecules

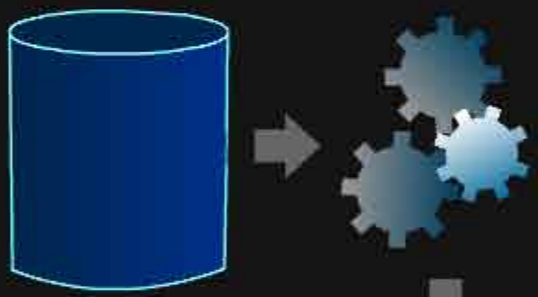
Similarity searches can be used to derive information on the potential properties of a given structure. By using the structure of this compound as a query, it is possible to develop a similarity search in order to find information on the properties that were observed on similar compounds.

Query



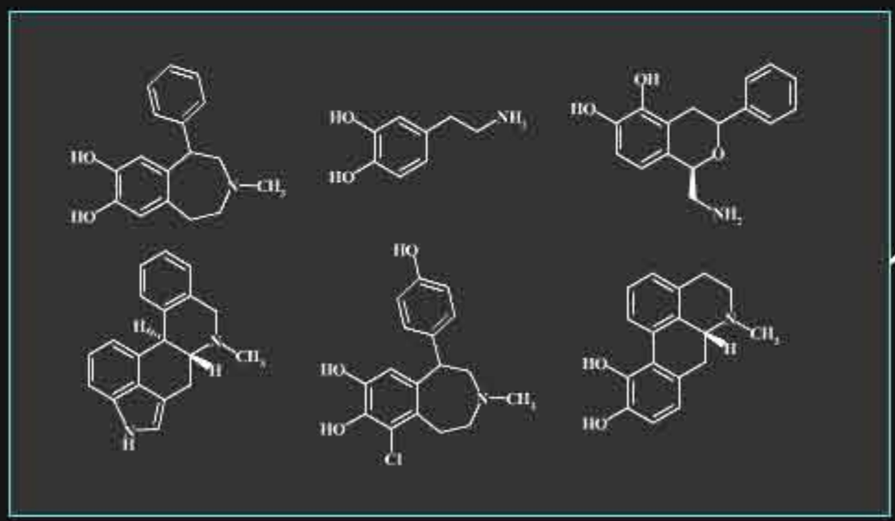
molecule with unknown properties

database of bioactive molecules



Similarity Searching

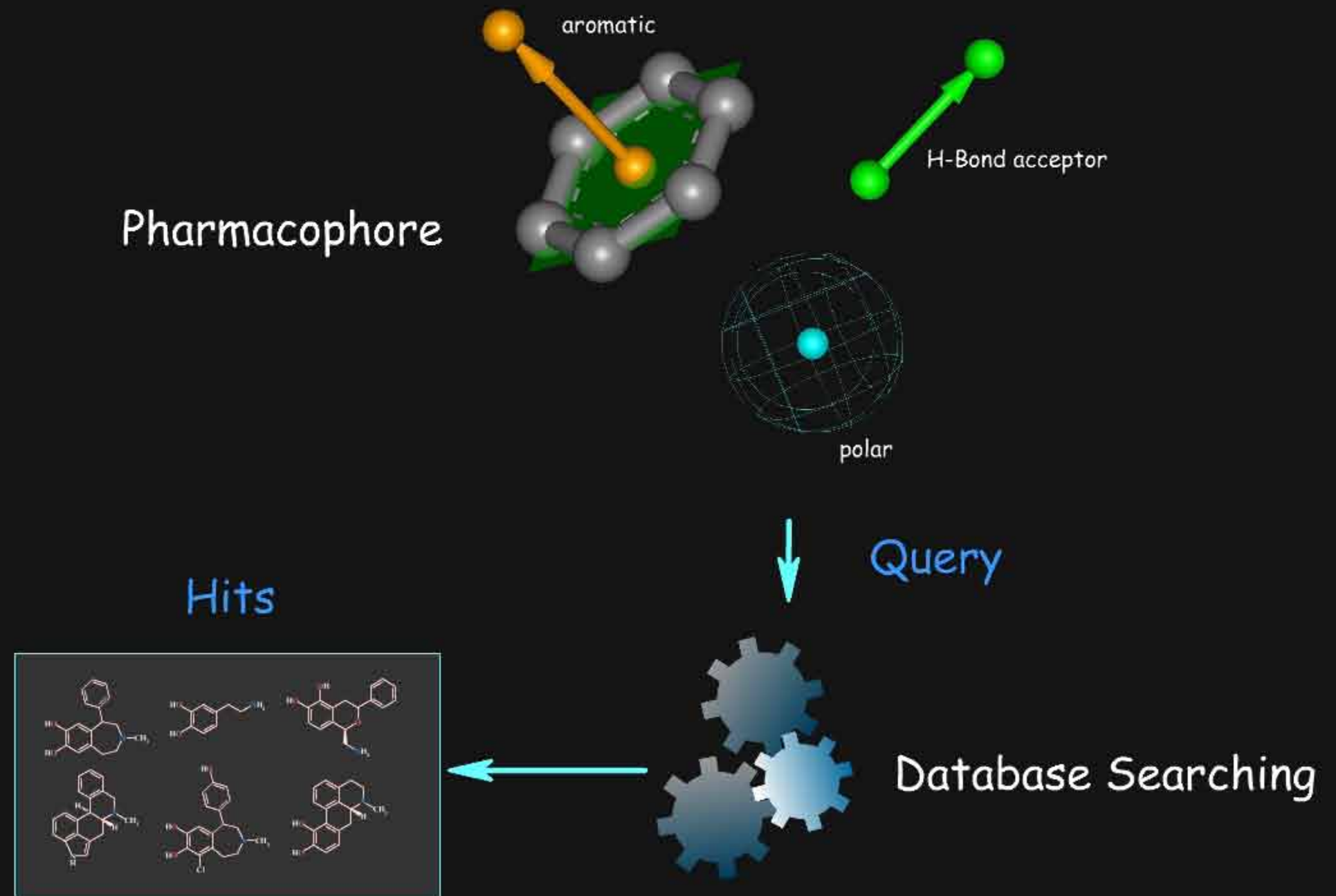
Hits



similar molecules that bring information on the query

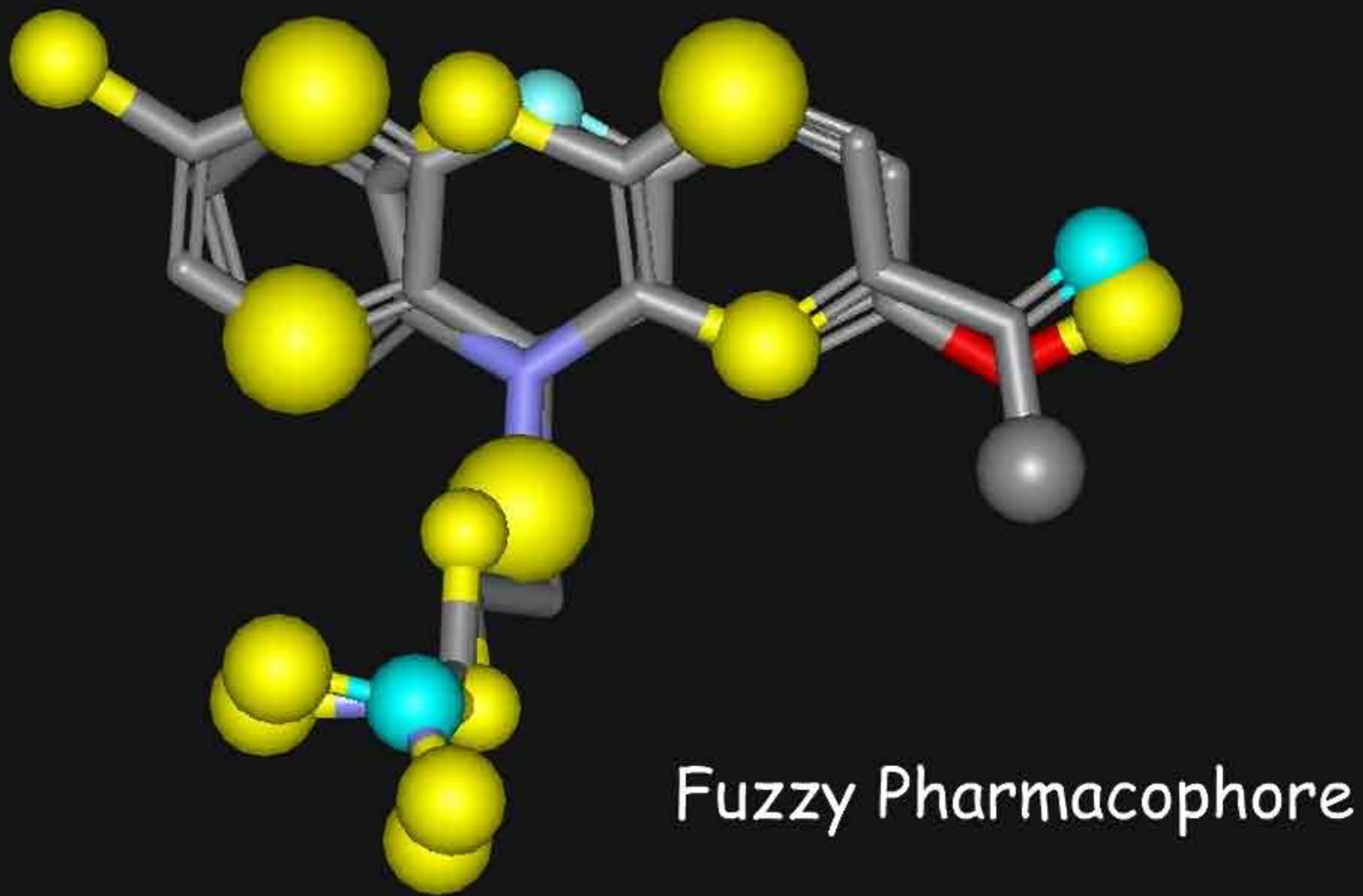
J3.7.3 Knowing a Pharmacophore, Search for Novel Molecules

Working on a given structure having some limitations (chemical, biological or physical), it is possible to use similarity searching to find novel molecules that have a similar pharmacophore to that of the reference compound. The search will retrieve similar molecules, (possibly chemically unrelated) carrying this pharmacophore. Note that the pharmacophore query can be defined in a fuzzy way; an example of such pharmacophore is illustrated in the next page.



J3.7.4 Example of a Fuzzy Pharmacophore

The following illustrates an example of a fuzzy pharmacophore. Working with TAR-RNA as a target for AIDS therapy, the authors employed a fuzzy pharmacophore to discover novel ligands of the target by similarity searching. In vitro testing of the hits obtained confirmed the predictions of the method.



J3.7.5 Validation of Novelty

Before embarking upon a costly synthetic program, it is important to make sure that the molecule considered is novel for the project. Has someone already thought of such a structure? Is that an original idea? Has someone synthesized this molecule? Answers to these questions can be addressed by similarity searching, which helps to assess or modify the initial idea.

Exact matching

'Has someone thought of this molecule?'

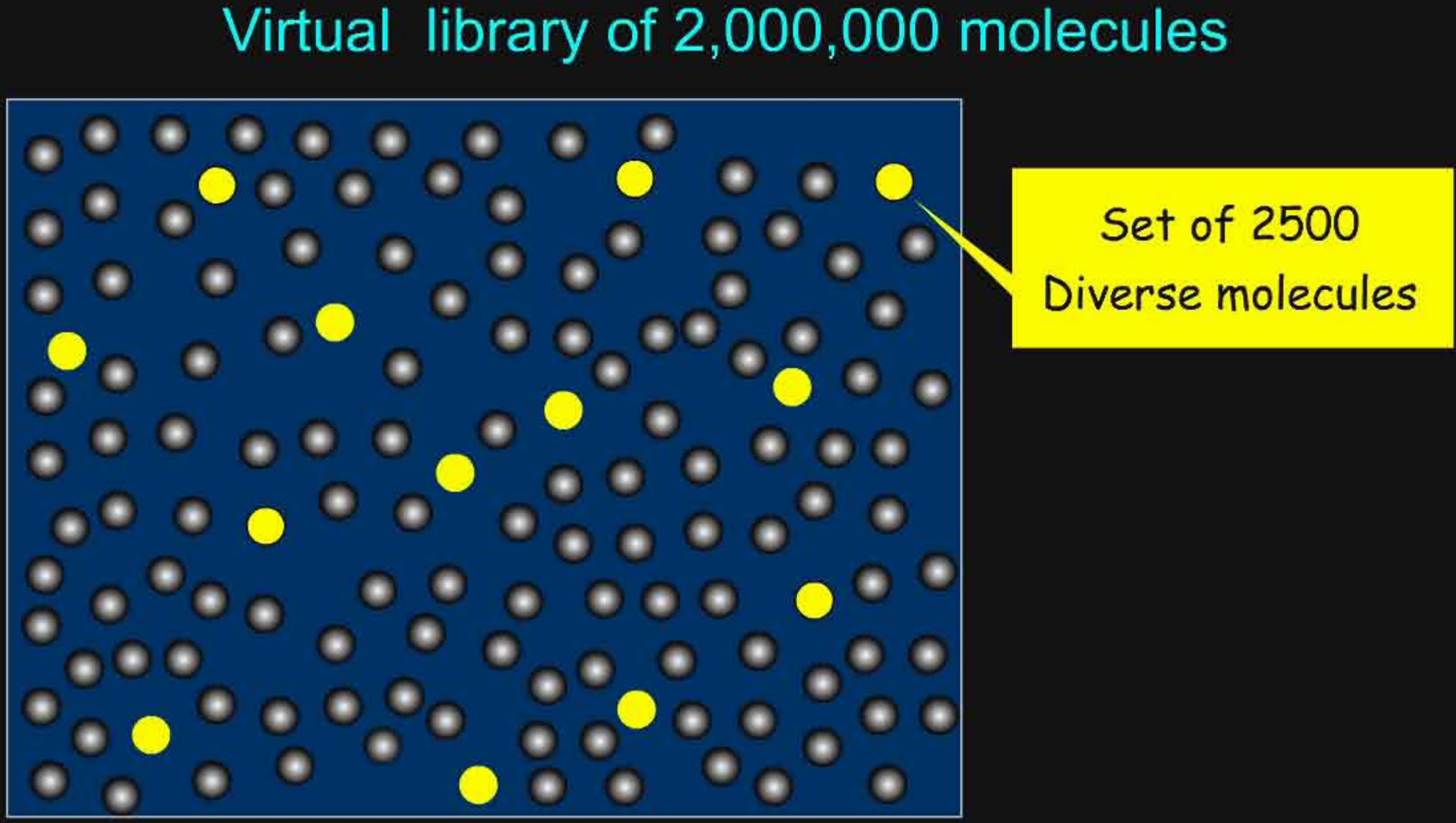
Fuzzy matching

'Is my idea original?'
'Is there any chance that someone has synthesized this molecule?'

J3.7.6 Reducing a Virtual Library to a Practical Size

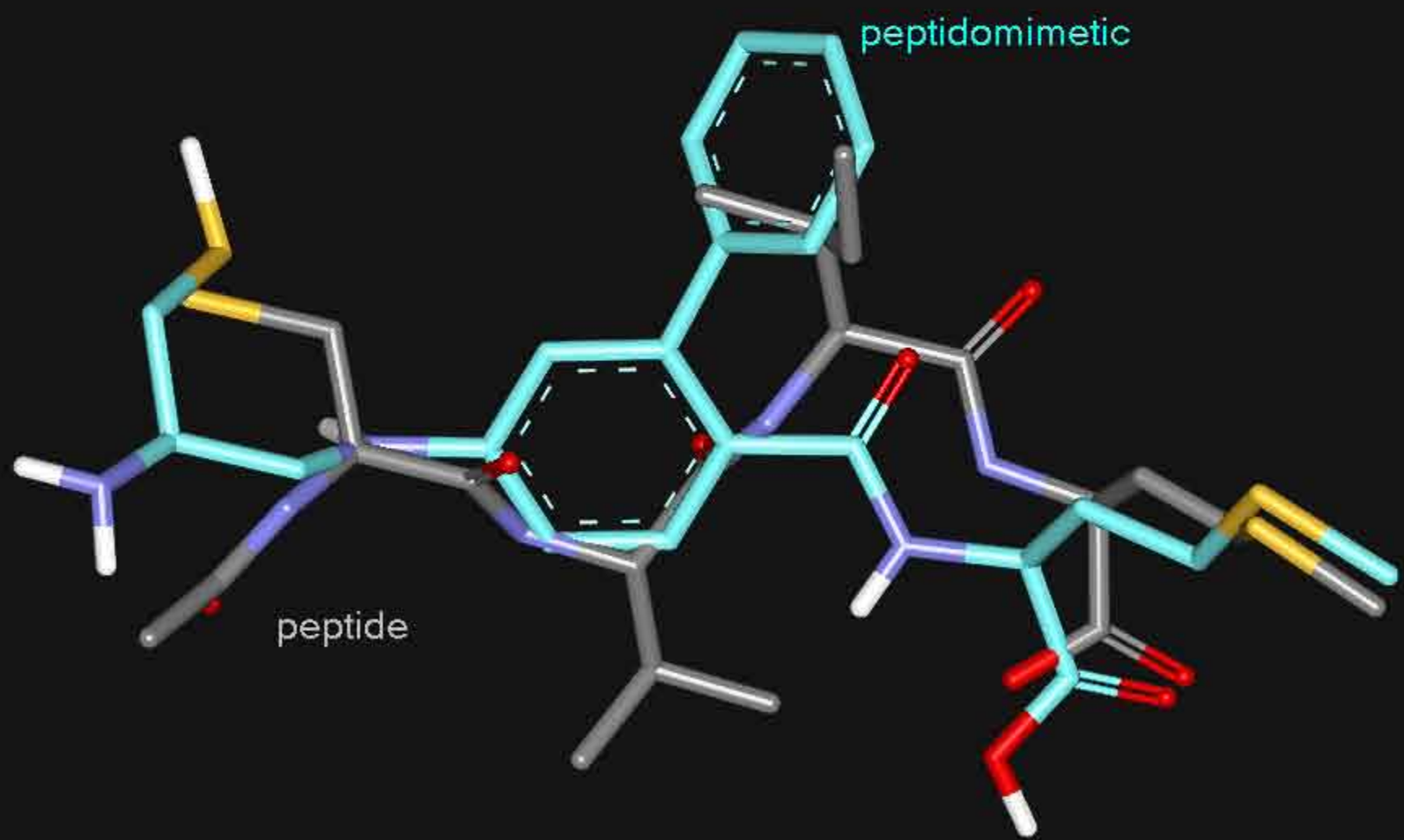
Another application of the "molecular similarity principle" is to reduce a virtual library to a practical size of diverse molecules to be synthesized. In this case the focus is on dissimilarity rather than on similarity.

Diversity Rationale



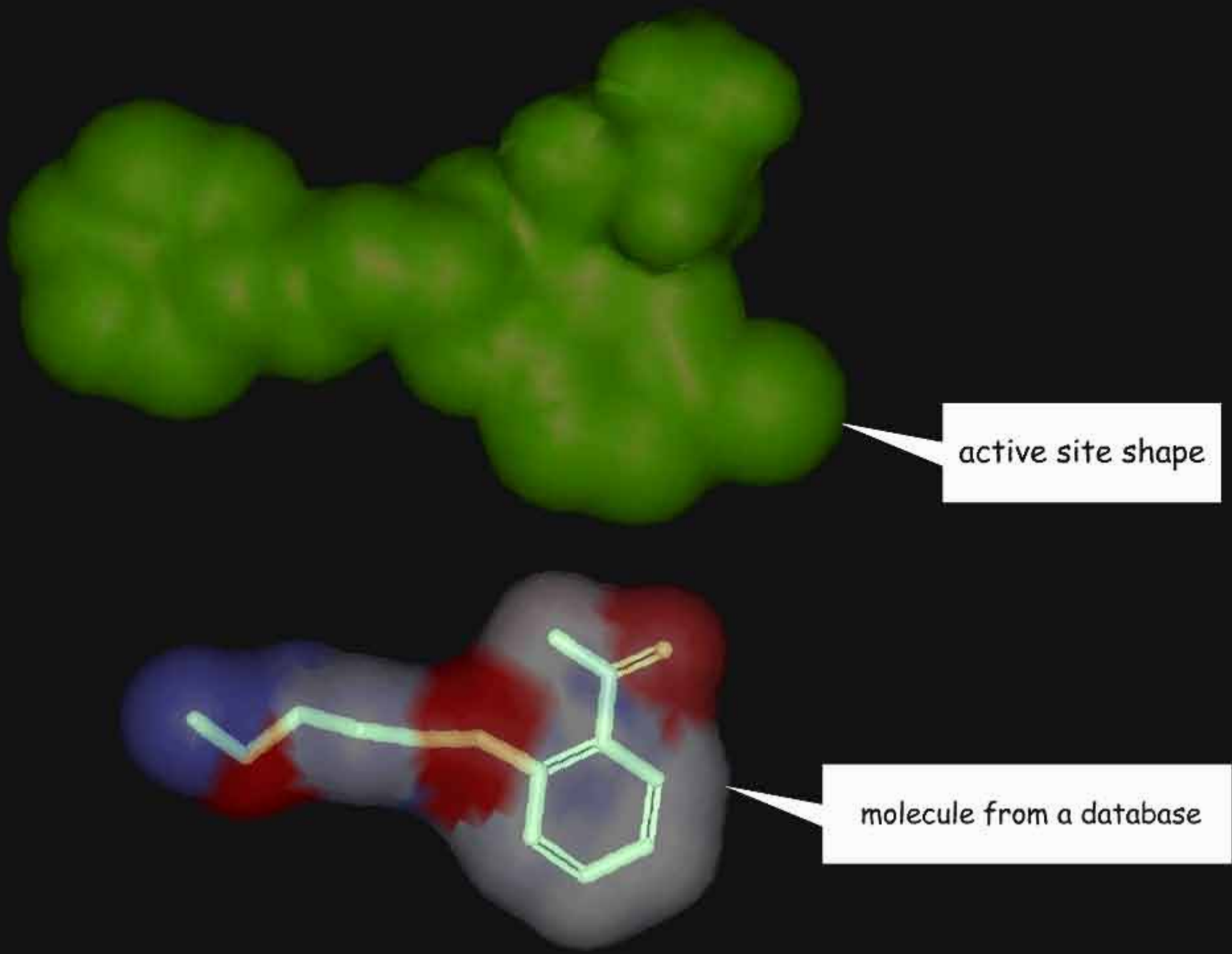
J3.7.7 Peptidomimetics

Due to their ubiquity in living systems, many natural ligands of receptors in the human body are small peptides. While it may be straightforward to apply them also directly as drugs, they have shortcomings (discussed in detail in the chapter on peptidomimetics) that make it advantageous to "hop" from peptides to non-peptide compounds. Similarity searching can be used to find small molecule compounds that carry the important structural elements of a peptide and mimic its bioactivity profile.



J3.7.8 Compounds that Fit the Shape of an Active Site

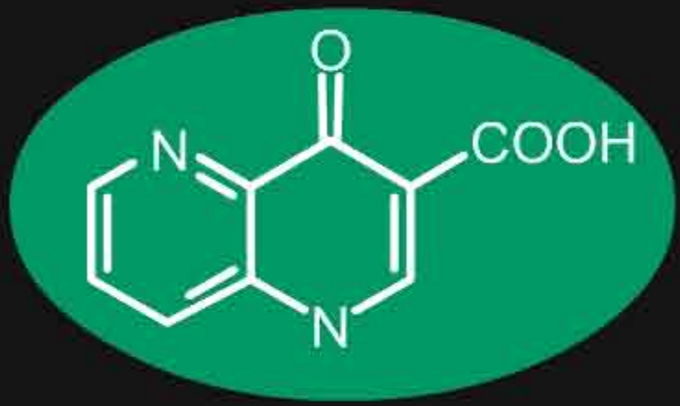
The "molecular similarity principle" can be used for searching compounds that fit the shape of the active site of a target protein. We know that physico-chemical properties are also involved in interaction with the receptor, however similarity searching is useful for first identifying molecules that satisfy the condition of good shape complementarity.



J3.7.9 Find a Synthetic Route

Similarity searching can be used for finding a synthetic route for a given molecule. It serves to identify similar molecules that were previously synthesized and adapt a particular synthetic scheme to the molecule intended to be prepared.

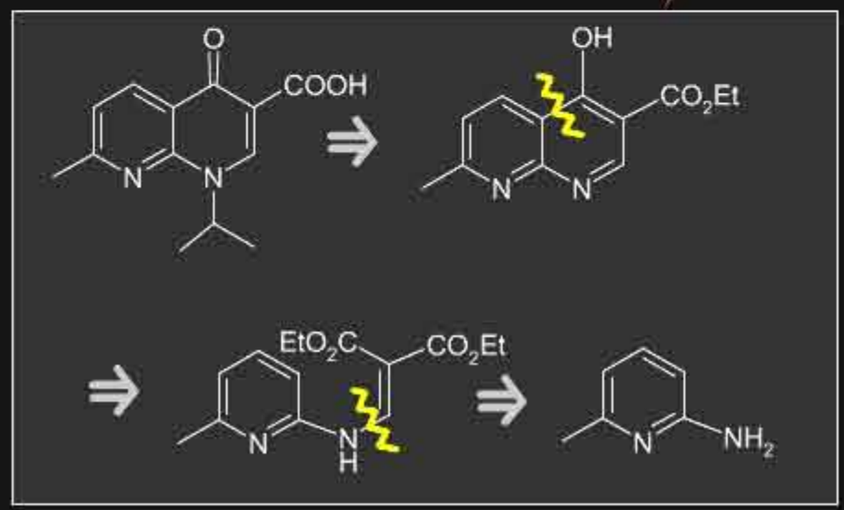
molecule to synthesize



exact match

fuzzy match

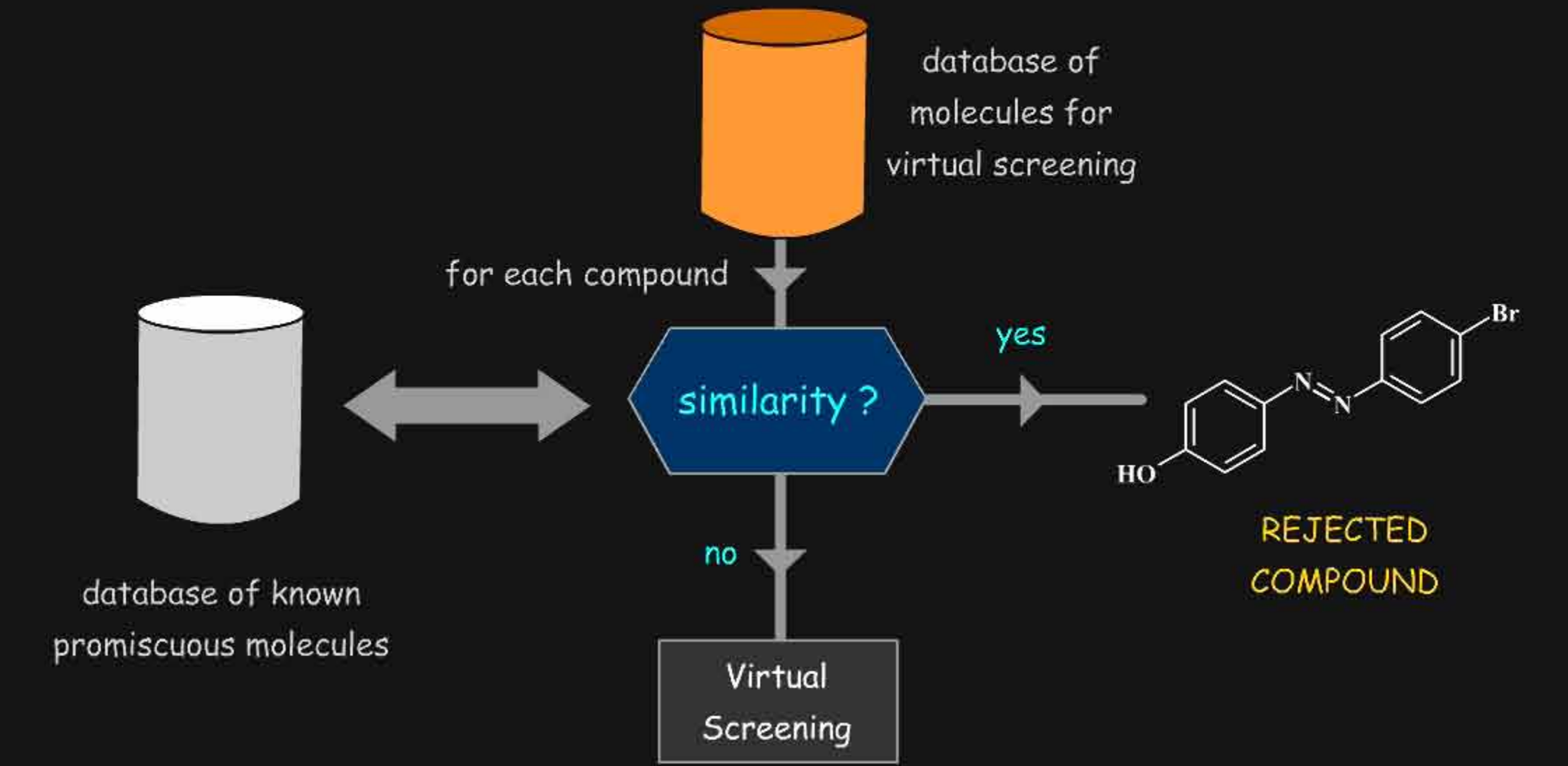
Adapt this known synthetic scheme to the query



the synthesis of this molecule was never described in the literature

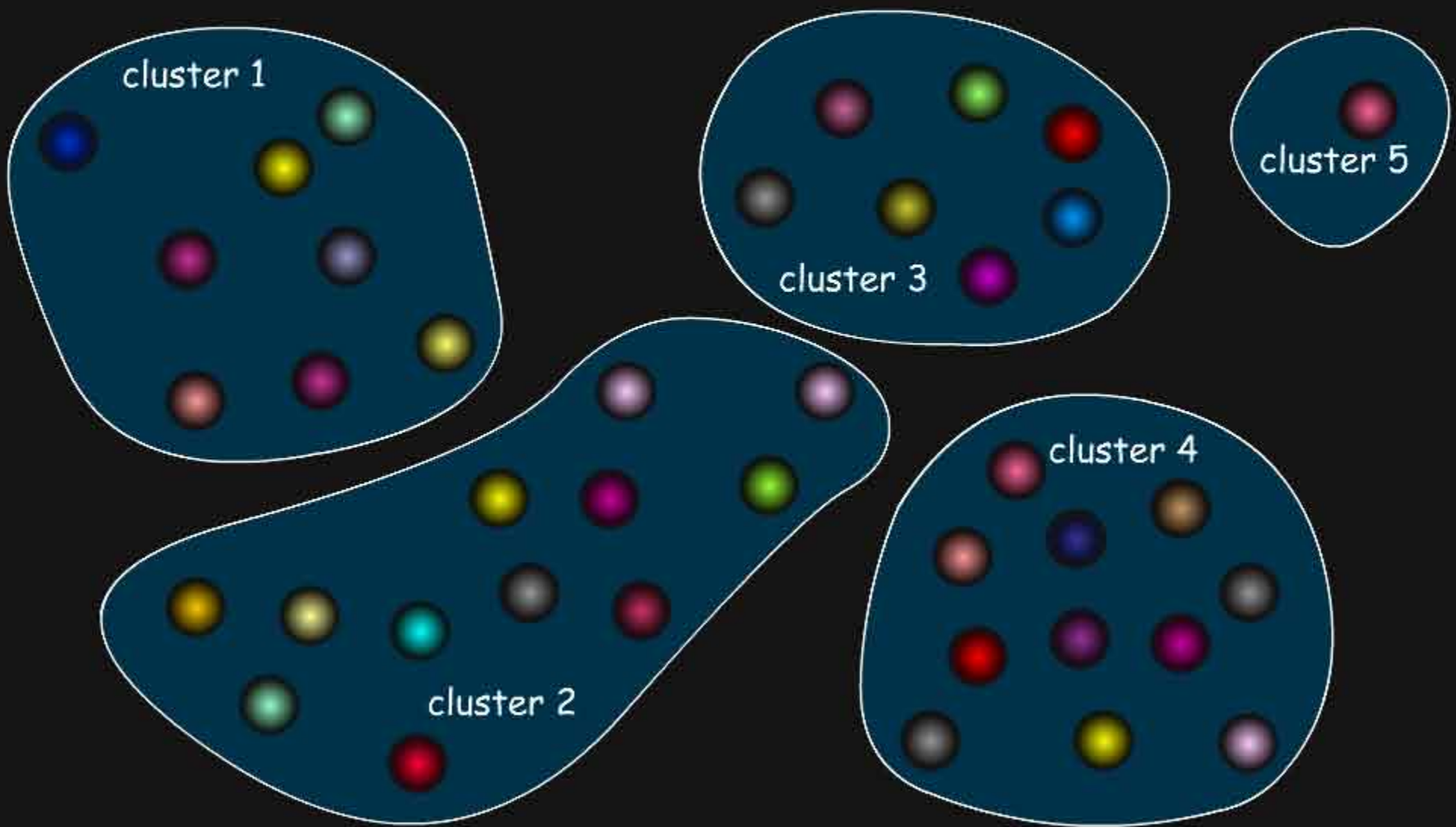
J3.7.10 Filtering Undesired Hits

In drug discovery many compounds bind indiscriminately to a multitude of targets. "Ligand-based virtual screening", which is built upon the molecular similarity principle, can thus also be employed to eliminate those that are "promiscuous" - and those that are similar, from screening results. In this case the structures of promiscuous compounds are treated in a fuzzy manner by the similarity search treatments.



J3.7.11 Clustering of Molecules

Clustering is a procedure in which objects (molecules) are divided into groups. Based on the calculation of distance coefficients each compound is associated with a cluster. Clustering can be used in similarity and dissimilarity analysis, and is useful in many applications such as data mining, subset selection in library design etc...



J1.1 What is 3D Searching?

The topic What is 3D Searching? contains the following 7 pages:

- Importance of the 3D
- What is 3D Searching?
- Components of a 3D Searching Program
  - 3D Database
  - Search Hypothesis
  - Converting a Search Hypothesis into a Query
  - Processing the Query

J1.4.1 Constructing 3D Databases

For 3D database searching one must decide the compounds to be searched, how the 3D conformations will be generated, what will be stored (and how) and the program that will be used for the searching. All of these components are discussed in the following pages.

- compounds to introduce in the 3D database (2D structures)
- How the 3D will be generated
- What will be stored in the 3D database
- What program will be used for searching

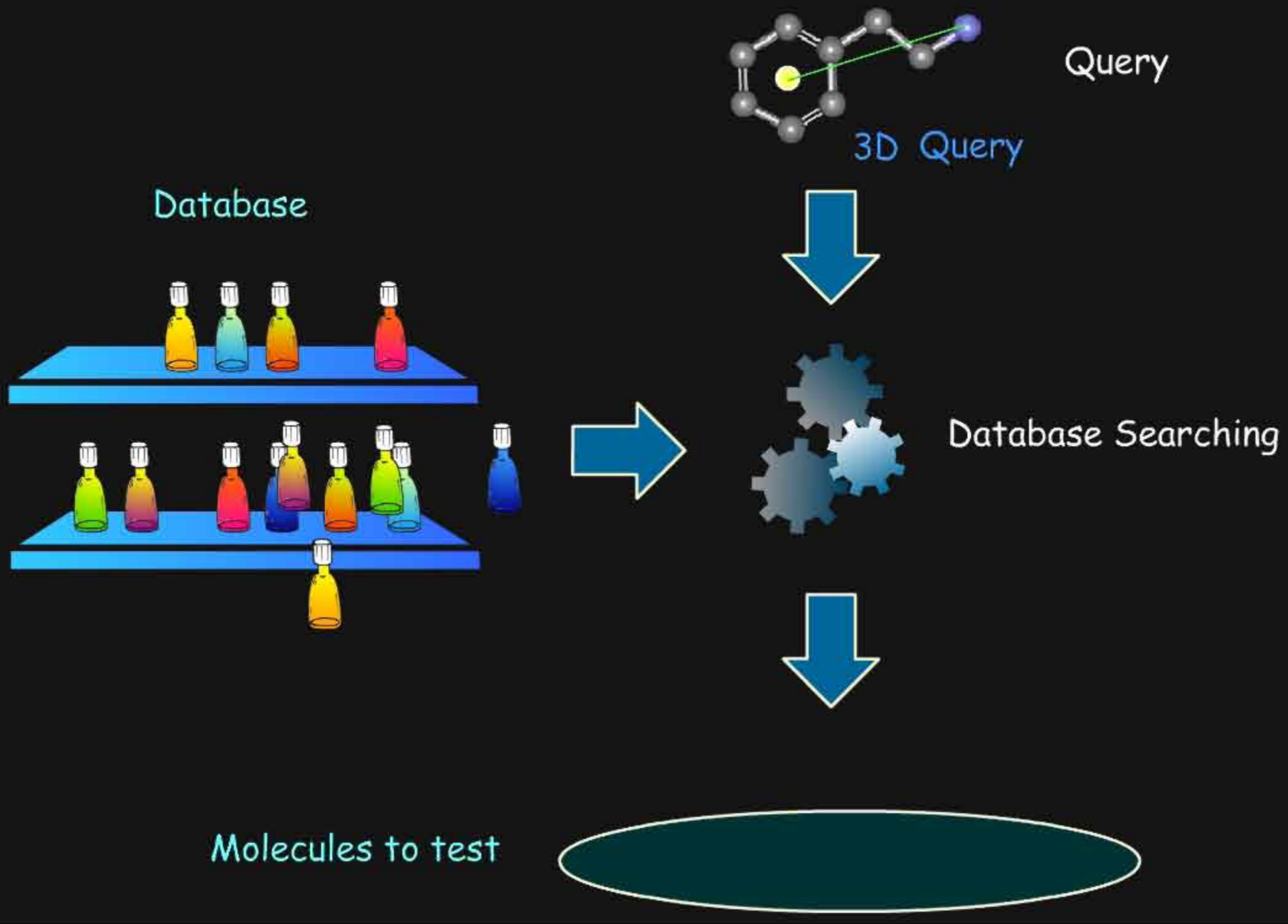
J1.4.2 Sources of Compounds for Searching

Because database preparation and screening is done on the computer, there is no need for the database to contain molecules that are in your stockroom, or indeed that are in anyone's stockroom. The source of compounds for the database depends on how you intend to use the results. Some typical uses of 3D database searching are indicated below.

Use	Database
Confirm a hypothesis	compounds readily available
Geometrical survey	experimental structures
Search for ideas	Any
Search for simple structures	commercial or corporate compounds
Template for CombiChem	Virtual database of compounds
New core	diverse molecules (commercial, corporate)
Proprietary scaffold	proprietary templates

J1.4.3 Database of the Corporate Collection

If you work for a company and your objective is to select compounds for screening, then a good source of compounds for a 3D database is your corporate collection. Samples of many of these compounds are readily available and their biological properties will also be available. Note that these compounds might already have been tested for the activity of interest using high throughput screening!

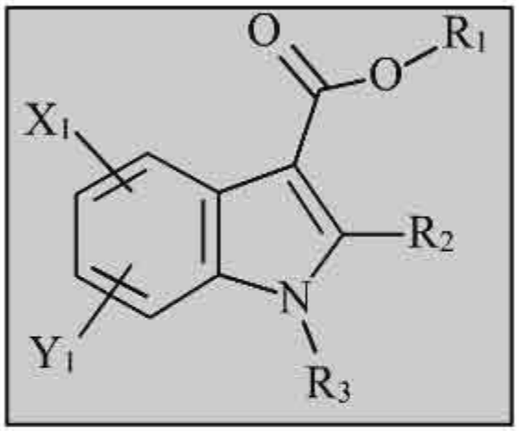


J1.4.4 Databases of Vendor Compounds

If you are looking for additional or different compounds to screen or to test your pharmacophore or bioactive conformation hypothesis, then a 3D database of vendor compounds is useful. Dozens of vendors of compounds for screening are happy to send CDs of their structures. One disadvantage of this source of compounds is that it takes time for the compounds to arrive and some will not be available. A second disadvantage is that typically anyone can purchase these compounds, hence your competitor might have already purchased samples of those that you identify.

J1.4.5 Database of Virtual Compounds

You might search a 3D database of virtual compounds to see which of those proposed for synthesis match a pharmacophore hypothesis. Alternatively you could generate and search one or more combinatorial libraries to have easily synthesized compounds available should you make a hit in a search. If the libraries are derived from proprietary chemistry, then one would expect a higher likelihood of patentability of any hits found active compared to hits found from vendor libraries.

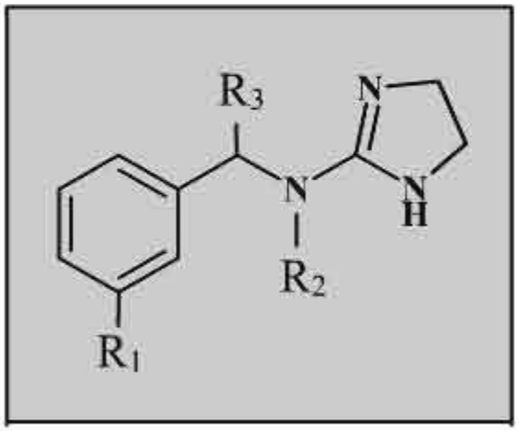


TEMPLATE No TRT-98716432

Current Library Size: 500 compounds

Virtual Library Size: 350,000 compounds

Purity of Compounds: 94% minimum



TEMPLATE No TRT-98716433

Current Library Size: 1000 compounds

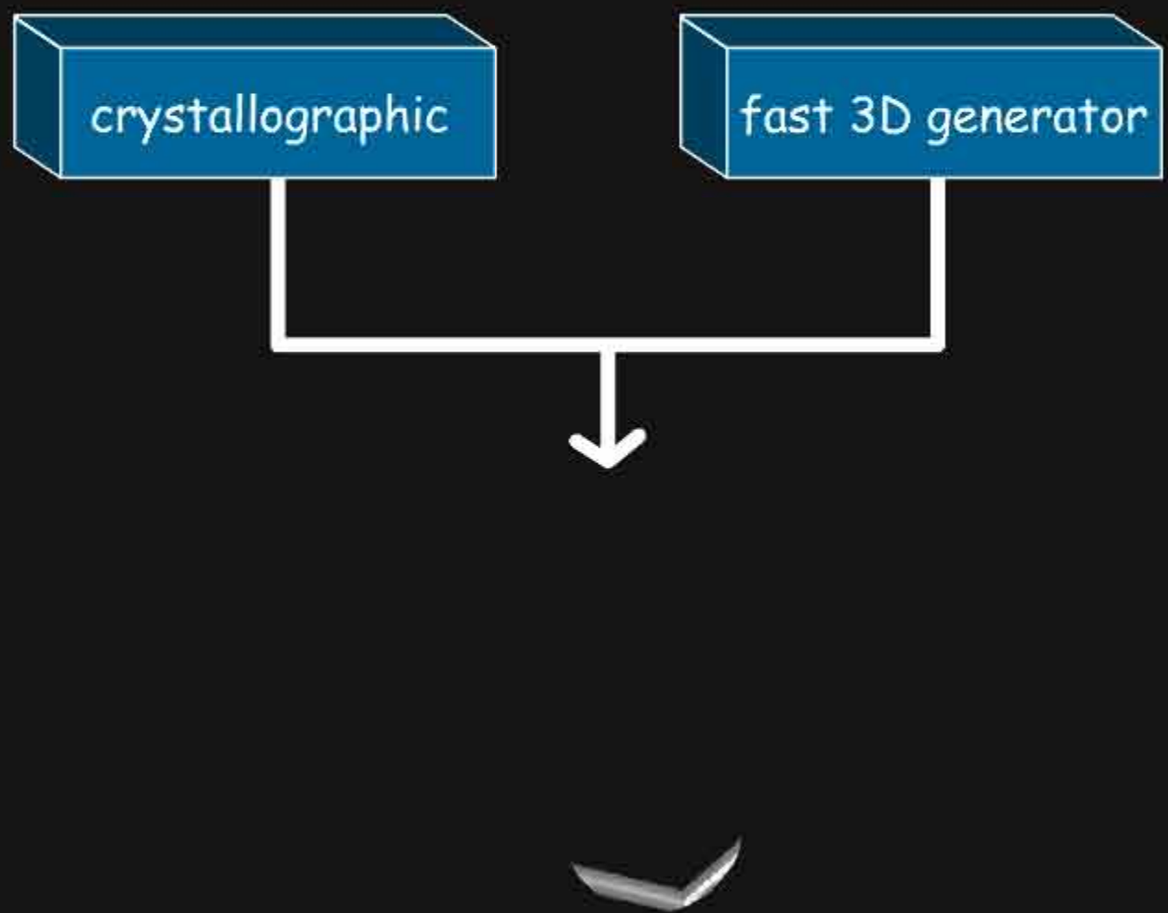
Virtual Library Size: 120,000 compounds

Purity of Compounds: 94% minimum

TEMPLATE No TRT-98716434

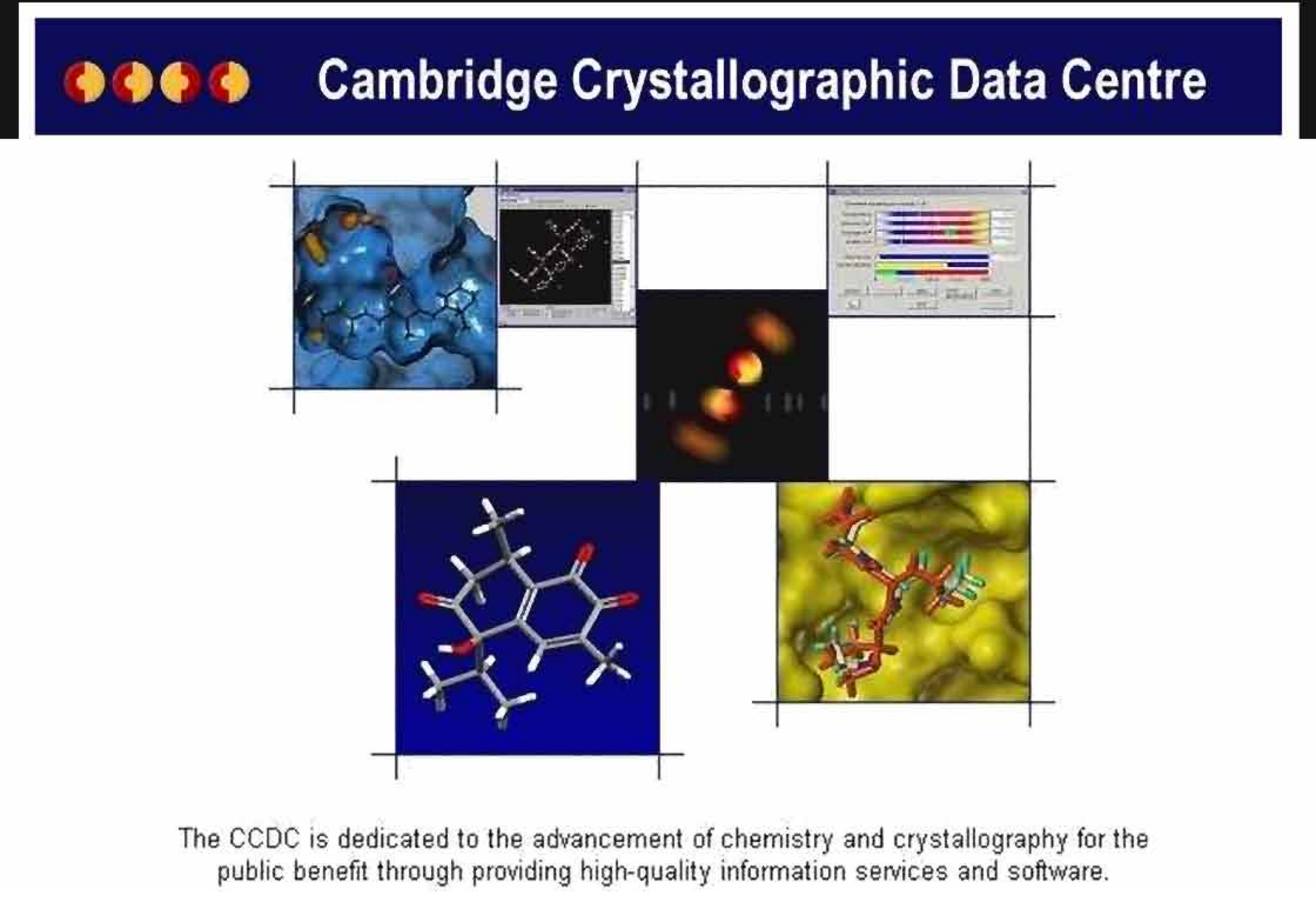
J1.4.6 Generating 3D Structures

Although some of the earliest 3D searching programs depended on carefully modeled 3D structures, these days the structures for searching are either determined crystallographically or generated by a fast 2D->3D structure converter program. Low energy conformers are collected and stored in the database however they cannot exceed a reasonable number because of limitations in storage capacity.



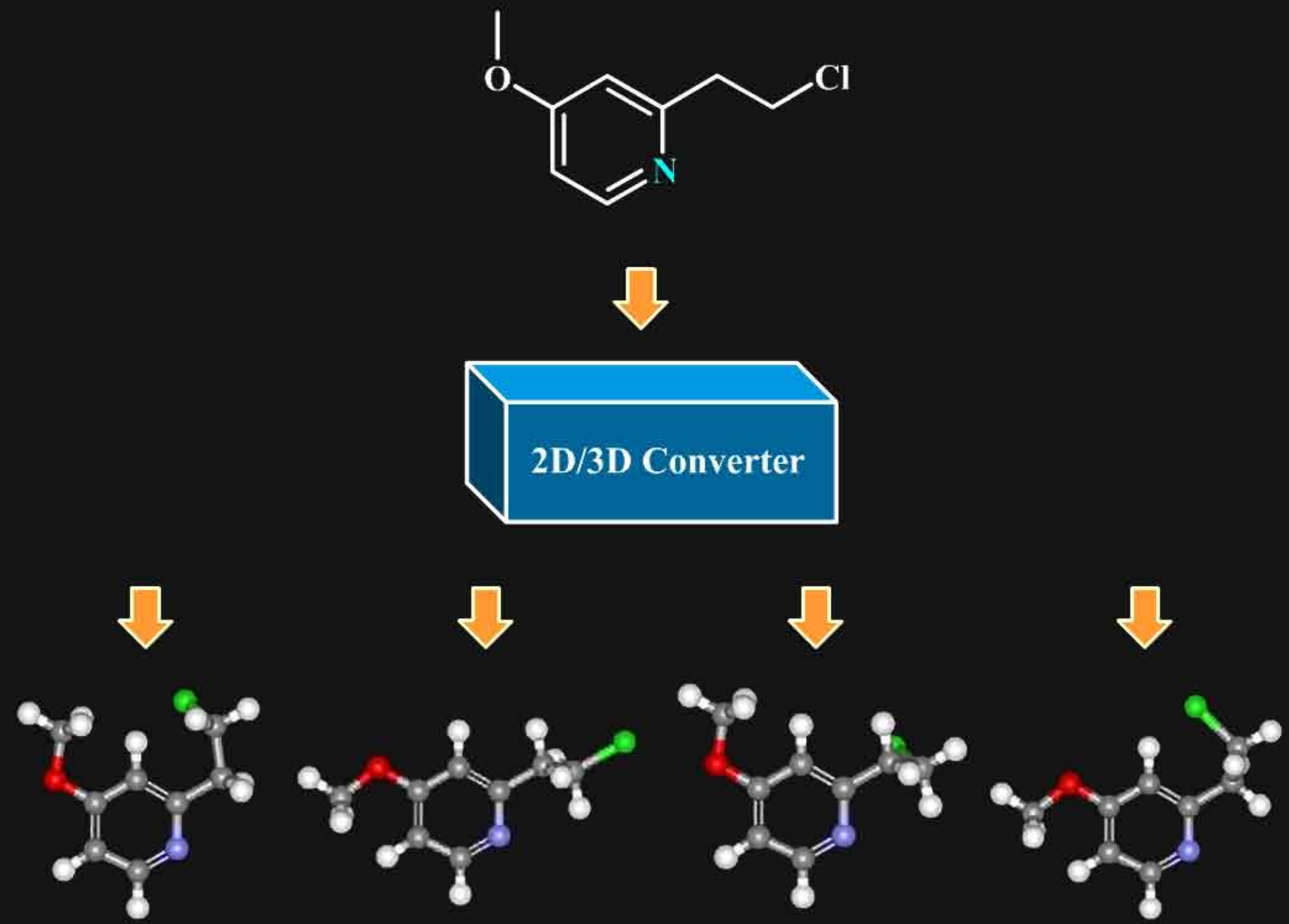
J1.4.7 Experimental 3D Structures

The Cambridge Structural Database contains the 3D coordinates of more than 260,000 compounds. Most of these were determined by X-ray diffraction, the remainder by neutron diffraction. The ConQuest software provides geometric 3D searching but also searching by name, substructure, and various crystallographic features.



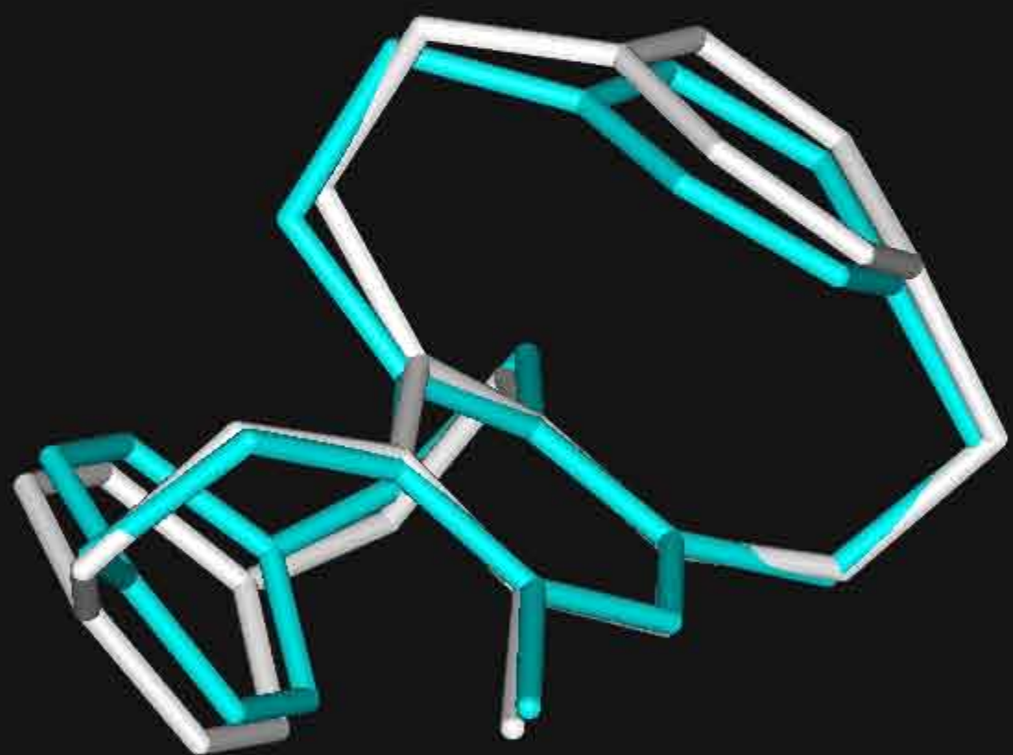
J1.4.8 Computationally Generated 3D Structures

Converting the molecules contained in a database of compounds stored in 2D to 3D has become a routine operation. Many 2D->3D converter programs now exist; a brief outline of the most popular ones is given in the following pages.



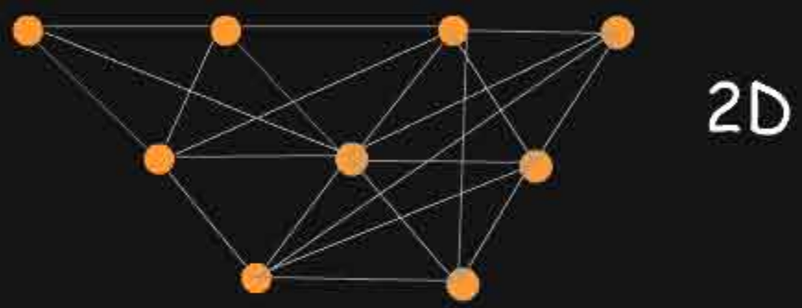
J1.4.11 CORINA

CORINA is a popular program that automatically generates three-dimensional atomic coordinates from the constitution of a molecule as expressed by a connection table or linear code. The program has proven to be reliable to convert large databases, for example a database with more than six million compounds has been converted with a conversion rate of 99%. The following view visualizes the comparison of the structure of cyclophane as generated by CORINA and experimentally determined by X-ray crystallography (RMS=0.26).



J1.4.12 ConFirm and Omega

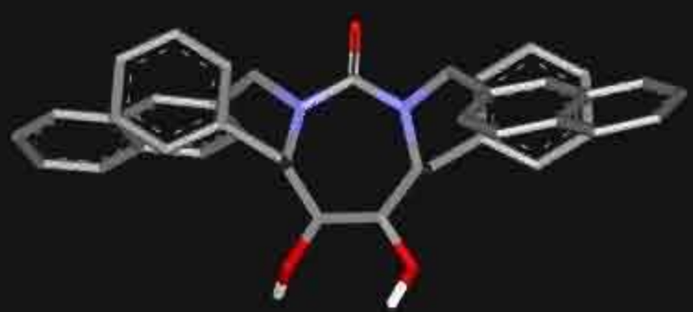
ConFirm and Omega generate and store many 3D conformations of a molecule for use in 3D database searching or for pharmacophore generation. The structures are generated by advanced distance geometry algorithms that eliminate duplicate conformations and attempt to cover all the conformational space.



2D

distance geometry algorithms

3D



J1.5 Programs for 3D Searching

The topic Programs for 3D Searching contains the following 8 pages:

- Programs for 3D Searching
- ConQuest
- CAVEAT
- UNITY
- ISIS
- Catalyst
- FlexS
- ROCS

J1.5.1 Programs for 3D Searching

Each program has its particular advantages and disadvantages. However, it is usually possible to use any of the programs described to identify the structures of interest. The programs are listed in the approximate order of their publication date.