Contents

18 Lead Identification by Virtual Screening

Andreas Kämper, Didier Rognan, Thomas Lengauer

18.1 Introduction

The identification of new drugs is a research topic of outstanding interest. Due to recent progress in the determination of several complete genome sequences including the human genome, the structural genomics projects aiming for structure determination of all naturally occurring protein folds, new techniques for target validation, and the advances in bioinformatics, our understanding of the nature of many diseases and their causative facts is constantly increasing. These efforts help to achieve the goal to identify novel small molecules interacting with proteins and in this way find new drug targets. In the year 2000 it has been anticipated that the number of potential drug targets will increase tenfold [47] which was too optimistic from todays view [177]. The number of druggable proteins is more likely to be $\approx$ 2,200 – 3,000 [81, 174]. Of them $\approx$ 600 – 1,500 are disease-related and thus are putative drug targets for small-molecule drugs [81].

The availability of new targets calls for effective systematic procedures for finding putative drugs that bind to these targets. The process of searching through a collection of compounds for molecules showing biological activity against a given target is called *lead identification*. This lead identification is a *screening* procedure (Section 18.1.1) and part of the overall drug discovery process. It can be subdivided into several individual steps (Section 18.1.2). As a prerequisite for screening, the molecules which are tested against the target, the *screening compounds* (Section 18.1.3), have to be preprocessed (Section 18.2). The actual screening can be performed with a variety of methods outlined in Section 18.3. The results obtained from these methods need to be analyzed and interpreted (Section 18.4). The final Sections 18.5 and 18.6 of this chapter provide recent case studies and critical evaluations of structure-based and ligand-based virtual screening techniques.

18.1.1 Screening techniques

Until a few decades ago, the search for drugs was a trial-end-error procedure, with the target proteins being mostly unknown. In the last few decades of the twentieth century, two different systematic techniques for searching for drugs have become accessible. Both of them are based on the fact that, increasingly, the target proteins for drugs or putative drugs have been identified. The two approaches are:

- High-Throughput Screening (HTS) is an experimental technique, where in a fully-automated fashion, a robot tests all molecules from a library against a molecular test system [78].

- Virtual Screening (VS), on the other hand, is a pure computational technique. Here, the computer is used to estimate biological activities, e.g binding affinities. This includes one or more computational techniques.

These techniques can complement each other in the sense that VS guides the experimental setup of HTS, but recently VS is also more and more seen as an alternative to HTS [101]. There are many concepts for the integration of both approaches [7, 69], showing the benefit of including experimental and *in silico* methods in the drug discovery studies. As example, VS methods can be used to select a subset of compounds for HTS or to analyze the results of a HTS experiment.

Due to their different nature, VS and HTS techniques have different advantages and disadvantages. For HTS, the major drawback is the cost of the experiments. The cost is mainly determined by the purchase of compounds of about US$ 1.00 per compound [198]. This has to be multiplied by the number of compounds used per HTS run, typically on the order of a few hundred thousands. In addition, supplies and an assay are needed. For both the cost is highly depending on the type of target. On the other hand, the major limitation for VS is the need of prerequisite knowledge about the binding process. If there are neither known actives which can serve as templates nor a 3D-structure of the target protein, VS cannot be used. Either the three-dimensional structure of the target must be known, then methods of structure-based design can be used (see Chapter 16). The other possibility is that at least one ligand is known that binds to the target, such as the natural substrate or another inhibitor. In the latter case, methods of ligand-based design can be applied. A substantial advantage of VS is its applicability to not yet synthesized, *virtual* compounds. This facilitates screening of virtual combinatorial libraries with up to billions of molecules. It is obvious that VS methods must be very efficient to deal with such large numbers of compounds. Thus, often not only a single technique is used for VS. Instead, screening proceeds in a sequence of steps each of which reduces the number of cinsidered compounds, starting with very fast techniques, followed by more advanced but computationally more expensive techniques.

Within this chapter we will cover all computational aspects of VS. The sections on preprocessing of libraries (Section 18.2) and post-processing of hit lists (Section 18.4) are also valid for HTS. However, we will exclude the more technical aspects of data handling and information storage of HTS. Furthermore, the structure-based design techniques covered in detail by Rarey *et al.* in Chapter 16 of this volume will not be included here.

18.1.2 Drug discovery process

Drug discovery is a time-consuming and expensive process [44] which involves a number of steps. Although the process is not linear — several of the steps have to be repeated iteratively — it is often represented as a pipeline (Figure 18.1). Within the pipeline, screening is performed during hit identification after a suitable drug target has been identified and validated. A hit can be defined as a compound which exhibits a strong binding affinity to the target. In order to perform a screening run, first a collection of compounds (Section 18.1.3) is submitted to the pipeline. This collection has to be prepared for screening using a number of preprocessing steps (Section 18.2.1). Specifically, unsuitable compounds are discarded by filtering steps. Next within the pipeline a structure-based or ligand-based technique is applied

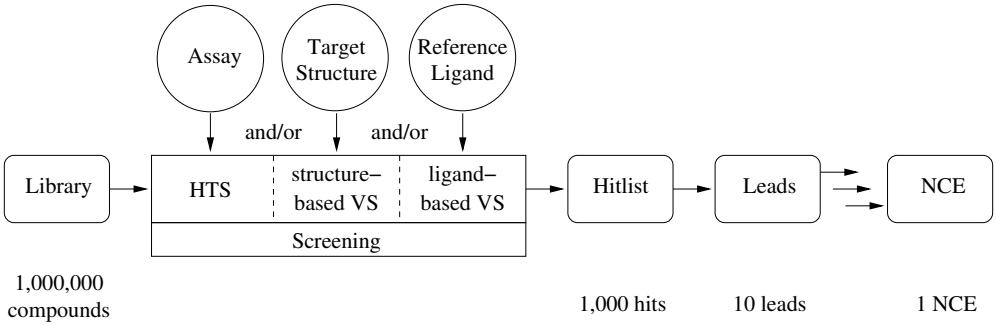

Figure 18.1: The drug development process. Screening is applied for reducing the number of initial compounds to a hitlist of molecules with a high binding affinity. Compounds from the hitlist are subsequently optimized to leads. The final steps (not shown) are then the finding of candidate structures, clinical trials, and finally the approval of the new chemical entity (NCE) by the authorities.

to further restrict the number of compounds (for ligand-based methods see Section 18.3, for structure-based methods see Chapter 16). This is not necessarily performed in a single step. Quite common is also the use of several cascading techniques, starting with a fast but inaccurate method to exclude many compounds, ending with a slow but better method to screen the most promising compounds (see Sections 18.5 and 18.6 for recent examples). Finally, a number of molecules exhibiting a strong binding affinity to the target are obtained, the *hits*. The crude hitlist obtained by these methods needs to be analyzed and compounds need to be sorted to prioritize subsequent lead selection (Section 18.4). In further steps of drug development, top-ranking compounds of the hitlist are refined and a small number of lead structures exhibiting promising properties are obtained (*hit-to-lead*) which may be optimized further to finally become *candidates* used in clinical trials. The respective drug optimization techniques are described in Chapter 19 of this volume.

18.1.3 Compound collections

The number of all possible drug-sized molecules, the virtual chemistry space, is huge. A systematic exploration of a small part of this space with molecules up to eleven heavy atoms was recently performed [57]. After exclusion of unsuitable chemicals with many small rings, over 13 million different compounds remained. A typical drug molecule can be up to twice as large as the compounds investigated in this study (average mass of 340 Da [55], about 24 heavy atoms). Estimates of the number of 'drug-like' molecules accessible to current synthesis procedures are on the order of 10^{60} [18] to 10^{100} [200]. These numbers indicate, that even when combining all compounds ever synthesized (estimated 10^8 molecules), we cover an almost negligible fraction of the virtual chemical space.

Compounds for screening can be obtained from databases of known structures, from combinatorial libraries, or from *de novo* design programs. Due to problems with synthesizability, often only known structures are considered. Typical databases with organic laboratory compounds (e.g. MDL Available Chemicals Directory (ACD, http://www.mdli.com/

products/experiment/available_chem_dir/) or SPRESI (http://www.spresi.com/)) are not suitable sources for screening compounds due to the non-druglike properties of most of the entries. (In fact, these databases are used as references for non-drugs, see below). Much better sources are collections available in-house to pharmaceutical companies or offered by screening compound vendors, containing historical compounds and combinatorial libraries. Within the MDL Screening Compounds Directory (SCD, http://www.mdli.com/products/experiment/screening_compounds/) database, over 3 million screening compounds are listed together with supplier information. Unfortunately, all these compound databases need extensive cleanup to be suitable for drug screening. Very recently, ZINC, a curated large screening library of purchasable compounds has become available [85], in which all necessary preprocessing steps (Section 18.2.1) have been performed already.

Reference data of pharmaceutical compounds at various stages of development can be taken from the MDL Drug Data Report (MDDR, http://www.mdli.com/products/knowledge/drug_data_report/), the World Drug Index (WDI, http://scientific.thomson.com/products/wdi/), or the MDL Comprehensive Medicinal Chemistry (CMC, http://www.mdli.com/products/knowledge/medicinal_chem) database.

18.2 Filtering and preparation of ligands

The data from screening compound collections are usually not suitable for virtual screening off the shelf. On the one hand, this is due to incomplete information (often missing 3D coordinates, stereochemistry, hydrogen atoms). On the other hand, chemical libraries tend to contain a number of undesired compounds and a lot of duplicates. Thus, before a library of compounds can be used in VS, a number of preparatory and filtering steps (Section 18.2.1) have to be applied. Among these filters, bioavailability (Section 18.2.2) and drug-likeness (Section 18.2.3) of the compounds are of special relevance. The overall preparation process is summarized in Figure 18.2.

18.2.1 Library preprocessing

An initial preprocessing step comprises the separation of entries with more than a single molecule (e.g. containing a charged species and its counter ions) into their constituent entries. Next, all non-organic molecules (e.g. chloride ions, water) must be removed. The most obvious and easiest method is the removal of all molecules without any carbon atom. Alternatively, as almost all drugs contain a bond between carbon and a hetero-atom, another strategy is to remove molecules without any of these bonds.

Then the library is subjected to a functional group filter. Here substructure search is performed in order to identify and discard compounds with known undesired groups in the library. This technique can be applied to reactive functional groups [170], groups unlikely to be leads, promiscuous binders, and even functional groups classified as toxic [201]. Strategic pooling, a technique proposed by Hann *et al.* [75], can also be applied by using functional group filters. For instance, if the ligand is required to contain an acidic group, only compounds with this function are integrated into the final library.

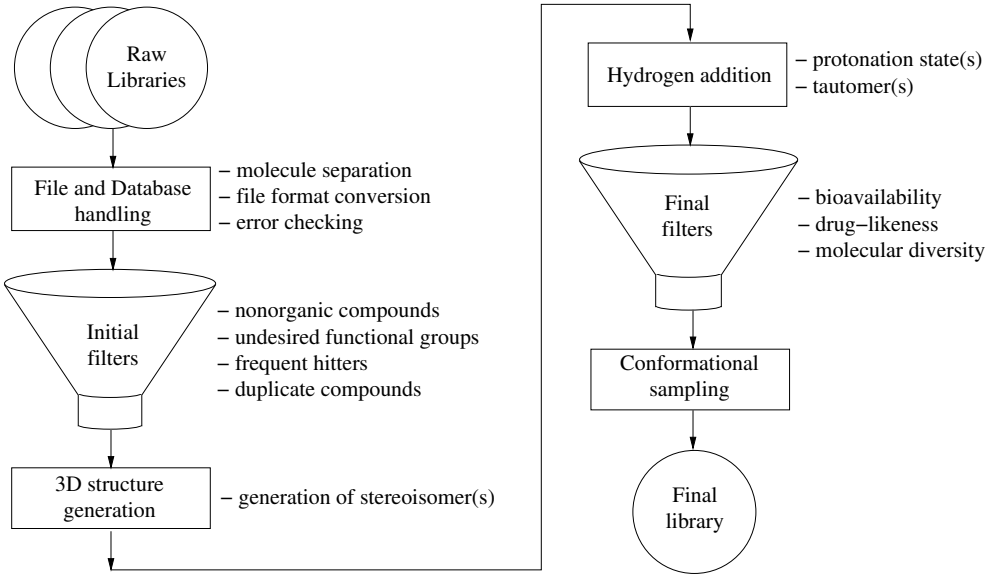

Figure 18.2: Example preprocessing workflow for chemical libraries (see text for details).

A common problem of HTS and VS consists of small molecules binding to many different proteins, resulting in false positive results (so called 'frequent hitters' or 'promiscuous binders') [135, 172]. By statistical methods based on substructures, Roche *et al.* [172] developed a scoring scheme based on a neural network to classify molecules as frequent hitters. Merkwirth *et al.* [141] use an ensemble model for this classification. Molecules covalently binding to proteins can be filtered out using reactive functional group filters (see above).

Finding duplicates in screening libraries significantly reduces the number of compounds to be tested. A one-by-one subgraph matching of connection tables for all pairs of compounds is too expensive for collections with more than a few thousand molecules. Instead, a representation of the molecular structure which can be compared easily is generated. Often chemical hash codes like those provided by Ihlenfeldt and Gasteiger [84] are used. Here, the molecular topology is encoded in a single number, the hash value. If two hash values are different, the molecules are different. If the hash values are identical, the two compounds are most likely identical. Only in extremely rare cases, two different compounds exhibit the same hash code. To be sure, a substructure matching has to be performed if two identical hash codes have been found. An alternative to hash codes is the generation of a unique string for every molecule. This can be done by using the SMILES representation (Simplified Molecular Input Line Entry System) after Weininger [210]. An enhancement of the SMILES representation can generate unique strings [209] by using an atom ordering scheme, which is suitable for molecular comparison.

The methods described so far need the two-dimensional representation of the molecule in form of a connection table only. If three-dimensional information is needed further in the screening pipeline then, as a next step, this information must be generated from either

two-dimensional coordinates (e.g. most Structure Data Files, SDF) or connection tables (e.g. structures in one-dimensional string representations, like SMILES or Sybyl Line Notation, SLN). Due to the still large amount of data this can only be done efficiently by structure generation programs. These convert the connection table into a single low-energy conformation. The two most widely applied tools of this kind are CONCORD [157] and CORINA [176]. CONCORD uses rule sets based on literature values of bond lengths and torsion angles to construct the molecule. Based on these values, acyclic parts of the molecule are constructed. For cyclic parts special rules pertaining to ring geometries are applied. Ring systems are obtained by merging conformations of the individual rings and optimizing the geometry such as to minimize the strain. CORINA works similarly, but includes more literature data and has a backtracking algorithm for generation of strained ring systems.

While the three-dimensional structure is generated, information on the stereochemistry of the compound has to be included. The often incomplete annotation of the stereocenters is a pervasive problem with current compound collections. Each stereocenter offers a choice of two stereoisomers (for the relevant cases asymmetrical carbons and cis-trans-isomerism). Which of these alternatives are explored is up to the user. An exhaustive generation of all possible stereoisomers of the compounds is one extreme, discarding the molecule due to incomplete information is the other. Typically, only a small number of stereoisomers are generated at this stage. Some programs even allow to handle stereocenters as variable during the calculation [62].

Next step in preprocessing, although often combined with structure generation, is the addition of missing hydrogens. This includes the assignment of a protonation state and an assessment of the tautomerism of the molecule. Depending on the application, either the most likely or all protonation states/tautomers are generated. A typical approach for assignment is the use of empirical rules. For example, carbonic acids are usually kept deprotonated and primary aliphatic amines are protonated. A program for generation of protonated forms is LigPrep from Schrödinger LLC (http://www.schroedinger.com/). Tautomers can be generated with TAUTOMER (http://www.mol-net.de/software/tautomer/) by Molecular Networks GmbH.

Depending on the methods used for screening, often the conformational space of a molecule needs to be explored by a VS program. While the structure generation programs (see above) produce a single structure only, the conformational analysis programs produce a set of alternative low-energy conformations for a molecule. Leach has reviewed the available techniques [119] and several available programs have been compared [20], whether they predict bioactive conformations correctly. The OMEGA program (http://www.eyesopen.com/products/applications/omega.html) by OpenEye, using a rule-based algorithm, currently seems to provide the best tradeoff between accuracy and speed [19, 20].

During all the steps of the screening procedure the identity of the molecules (e.g. their registry numbers, order information) has to be maintained and stored, typically in a relational database system.

18.2.2 Bioavailability

It is highly desirable for a drug to be administered by oral ingestion in order to be easily applicable by the patient. Thus, the molecule must have reasonable aqueous solubility and

has to pass the intestinal membrane in order to enter blood circulation. Bioavailability — a transport phenomenon — is often confused with drug-likeness. Here we keep these two entirely different subjects separate. We discuss bioavailability in this section and drug-likeness in the next Section 18.2.3.

By analysis of molecules that have entered clinical trials (and thus, are bioavailable), Lipinski and coworkers established the 'rule of five', which provides a simple heuristic rule for oral bioavailability [128]. It is likely that a molecule exhibits poor absorption, if two or more of the following criteria are fulfilled:

- Number of hydrogen bond donors (counted as number of O—H and N—H groups) > 5
- Number of bond acceptors (counted as number of any O or N atom) > 10
- Molecular weight > 500
- Calculated logP > 5 (if ClogP is used, see below)

Here, logP represents the logarithm base 10 of the octanol-water partitioning coefficient, a property which can easily be calculated by property estimation techniques [143]. Among the estimation techniques for logP, ClogP introduced by Hansch and Leo [76] and available from BioByte Corp. (http://www.biobyte.com/) is widely accepted. The idea of simple property-based rules as rejection criteria for bioavailability was extended by Ghose *et al.* [64]. Here, ranges were calculated for logP, molar refractivity, molecular weight, and number of atoms.

After development of fast estimation methods for the polar surface area (PSA) by Clark [31] and Ertl *et al.* [50], rejection of molecules with a PSA > 140 Å was proposed as the only rejection criterion. Veber *et al.* [193] analyzed a database with drug candidates. They classified compounds as bioavailable, if the PSA was less than 140 Å and the number of rotatable bonds less than 12. More detail on bioavailability and, more generally, on ADME properties (absorption, dissipation, metabolism, excretion) is provided by Baringhaus and Matter in the subsequent Chapter 19 of this volume.

18.2.3 Drug-likeness

With the filtering methods described above, the knowledge of medicinal chemists on whether a compound might be a good drug or not, is not taken into account. It is desirable that the compounds in a screening library have the typical properties of drugs. Thus, a binary classification, whether a compound is 'drug-like' or not must be performed on all compounds. The challenge regarding this problem is that 'drug-likeness' is a property which is not easily evaluated and not related in a simple fashion to other chemical, physical, or biological properties. In order to solve the decision problem, the knowledge of medicinal chemists for assessment of 'drug-likeness' is used. The implicit knowledge on drug-likeness is inherent in the databases of known drugs and can be extracted by comparison with databases of non-drugs [5, 65, 175].

Implicit information on drugs is contained in the MDDR, WDI, and CMC databases (see Section 18.1.3). Within these databases, not all entries are existing drugs, but the majority of entries were designed by medicinal chemists with the intention of developing a drug. In contrast, databases like the ACD or SPRESI of general organic compounds are supposed to

contain very few drugs. A statistical classification technique can be used to extract the knowledge from the databases to decide whether a given compound is drug or non-drug.

Gillet and Bradshaw [65] used structural features (including number of hydrogen-bond donors and acceptors, number of rotatable bonds, number of aromatic rings, molecular weight) and a shape descriptor. Compounds of the WDI and SPRESI databases were analyzed with a genetic algorithm to derive a weighting scheme, which calculates the drug-likeness of a given compound. Ajay *et al.* [5] used MDL keys (see Section 18.3.2), in addition. They applied both decision trees and an artificial neural network (ANN) for classification, trained on the MDDR and ACD databases. Sadowski and Kubinyi [175] also used a feed-forward neural network. The classification scheme was trained on atom type descriptors of compounds from the WDI and ACD databases.

Support vector machines (SVM) can also be applied to the drug/non-drug classification problem. In a comparison of SVM to neural networks, Byvatov *et al.* could obtain slightly more accurate classifications on the same data as used in [175]. Overall, the predictive power of the methods presented so far reaches a typical 80 % correctly classified test molecules. Recently, Müller and coworkers [144] were able to reduce the error rate to 7 % in a blind-test using SVMs. This performance was achieved by careful model selection after comparison of several learning methods. In summary, the results show that accurate drug-likeness filters can be constructed which use the knowledge on drug-likeness obtained by medicinal chemists over decades.

The technique presented can be tailored to specific screening problems. Thus, not the phenomenon of drug-likeness is assessed but the likelihood of a drug to belong to a certain class of compounds. Ajay *et al.* [4] used neural networks to classify ligands regarding their CNS activity while Manallack *et al.* [130] used them to screen for candidates binding to kinase and G protein-coupled receptors. More recently, Briem and Günter [23] presented a SVM method also for 'kinase-inhibitor likeness'. All these target-specific drug-likeness classification techniques have a prediction accuracy of about 80 %. Thus, if there is a sufficient number (several hundred) of compounds of a certain class available as training set, machine learning techniques can be trained to predict the likeness to be a certain inhibitor with high accuracy.

18.2.4 Molecular diversity

Compound collections, especially those in pharmaceutical companies, contain many series of analogous compounds. It is desirable to screen for only a subset of 'maximally diverse' compounds, reducing redundant structural information. This diverse subset is hoped to cover the range of chemical structures and physical properties to a sufficient extent. Unfortunately, there is no generally accepted single definition of similarity or dissimilarity for this purpose [114]. Thus, selecting a set of diverse compounds can be performed in a number of ways. These differ not only in the method but also in the selection criteria used. These techniques have been reviewed on several occasions [1, 39, 126].

Among the methods for selecting diverse compounds, clustering methods can be considered as the standard technique. For clustering, descriptors (see Section 18.3.1) are calculated for each compound. Then a cluster analysis algorithm divides a group of compounds into clusters. Compounds within a cluster are similar and compounds from different clusters are dissimilar. After clustering of a library, a representative molecule is taken from each cluster

to belong to the diverse compound collection. For clustering of chemical libraries, typically methods generating disjoint clusters are used in which each molecule belongs to a single cluster only. Both, hierarchical [9, 211] and nonhierarchical [213] methods, have been applied. Evaluations of different clustering algorithms demonstrate the better performance of hierarchical clustering methods for several test cases [12, 24, 46]. Naïve implementations of hierarchical clustering need an initial $N \cdot N$ similarity matrix and have space and time complexities of $\mathrm{O}(N^2)$ and $\mathrm{O}(N^3)$, respectively. By use of the minimum variance method (also called Ward's method) [206] which aims on minimizing the total variance of a cluster, however, a computationally more efficient implementation is possible [146], having space and time complexities of $\mathrm{O}(N)$ and $\mathrm{O}(N^2)$, respectively. For large numbers of compounds ($N > 10^6$) hierarchical clustering is not applicable. Instead non-hierarchical clustering (e.g. K-means clustering [139]) is used. For these algorithms the time complexity for K generated clusters is $\mathrm{O}(KN)$ per iteration for efficient implementations. For descriptions of clustering algorithms, see also Chapters 25 and 28.

A different approach on selecting diverse compounds is provided by partitioning methods. These use a low-dimensional property space, each property represented by a continuous number that is categorized into a discrete set of value ranges, forming a set of cells in property space [37, 125]. Compounds from a library are then partitioned into these pre-computed cells. Two special kinds of descriptors, BCUT descriptors by Pearlman and Smith [155] and pharmacophore keys by Davies [38] can be used for partitioning chemical property space. BCUT comprises molecular descriptors based on the eigenvalues of a matrix representation of the molecules. These descriptors are designed specifically to define a low-dimensional property space. For the description of the properties, axes are chosen that span property space in such a way that the compounds exhibit maximal variance along the axes and compounds are evenly distributed in property space [155, 156]

Davies [38] developed ChemDiverse, a program using pharmacophore keys for the selection of diverse compound sets. The basic idea is to calculate the pharmacophore key for the first molecule, add the molecule to the diverse compound selection, and store the pharmacophore in a list. Subsequently, for the next molecule in the library, the pharmacophore keys are calculated. If this molecule has a pharmacophore key not yet represented in the list, the molecule is added to the selection.

A completely different attempt to find diverse compounds is by dissimilarity-based compound selection. Among them, maxmin by Lajiness [118] is most used. The maxmin algorithm first selects a compound randomly and adds it to the selection. Iteratively, the compound most dissimilar to the already selected set is identified and added to the selection, until a desired number of compounds has been found. A stochastic variant, OptiSim has been proposed by Clark [32]. The initial random compound is compared to a set of K other randomly chosen compounds. Here, only compounds with a dissimilarity greater than a defined threshold to the already selected compounds are considered. The most dissimilar compound among the K compounds is added to the selection. In the next iteration a new set of K candidates is generated and the process continues.

Taylor [187] proposed a method based on the stepwise elimination of the most similar molecule from the collection. Initially the similarity matrix between all molecules is calculated. In a stepwise fashion, the two currently most similar compounds are identified, as long as there is more than a single compound left.

While diversity is desired in initial screening runs for identification of hits, once a lead is found, compounds should be similar to the lead, i.e. the library should be 'focused'. The clustering and partitioning methods can be used directly for generation of focused libraries by switching the selection criteria from one of each kind to all of one kind.

18.3 Ligand-based virtual screening

The methods of ligand-based VS can be divided into two classes. One class of methods tries to match compounds with identical parts (substructure or pharmacophore) as the active molecule. For these methods, initially, a number of active molecules are analyzed for a common substructure or pharmacophore, which is then used for searching exact matches. The other class tries to find molecules which are 'similar' to a known active molecule. The underlying assumption is that if a molecule is structurally similar, it has similar properties, binds in a similar binding mode, and exhibits similar activity. This assumption is known as 'similar property principle' [40, 87] or 'neighborhood behavior' [153]. Methods for similarity search are applicable if only a single active compound is known. In contrast to substructure and pharmacophore searches, the compounds are not only partitioned with regard to whether they are matching the query or not. Instead, a complete ranking of compounds according to their similarity scores is obtained. Similarity — like dissimilarity — is not clearly defined [114] and there are some remarkable exceptions from the similar property principle [114, 131]. Nevertheless, similarity search is the most widely used method in VS and numerous similarity measures have been developed [181]. Figure 18.3 illustrates the most common similarity search techniques, detailed below.

The actual methods for defining molecular similarity in this context are quite diverse. Often these methods are classified as one-, two- or three-dimensional, depending on the type of molecular representation used. In this section we focus on those methods first, which use the information of a single reference molecule. Then techniques that need a set of input molecules are discussed. The latter are also often used for the selection of compounds from hitlists. These techniques are described in Section 18.4.

18.3.1 Descriptor-based similarity measures

Similarity search methods have been used for a long time. The field started with counting the numbers of substructures common to a pair of molecules [27, 212]. This figure provides an initial effective way of quantifying similarity between molecules with very low computational cost. Since then, it became a standard retrieval technique for chemical databases.

Methods for calculating quantitative measures for the similarity between a reference molecule and a set of molecules have been studied in detail (see Willett *et al.* [211] and references therein). Common to all these techniques is the requirement to provide a set of attributes of the molecules being compared and a similarity coefficient, to provide a quantitative numerical measure for similarity between the molecules. The individual importance of different attributes (e.g. logP, molecular mass, presence of functional groups, ...) has to be accounted for by definition of a weighting scheme.

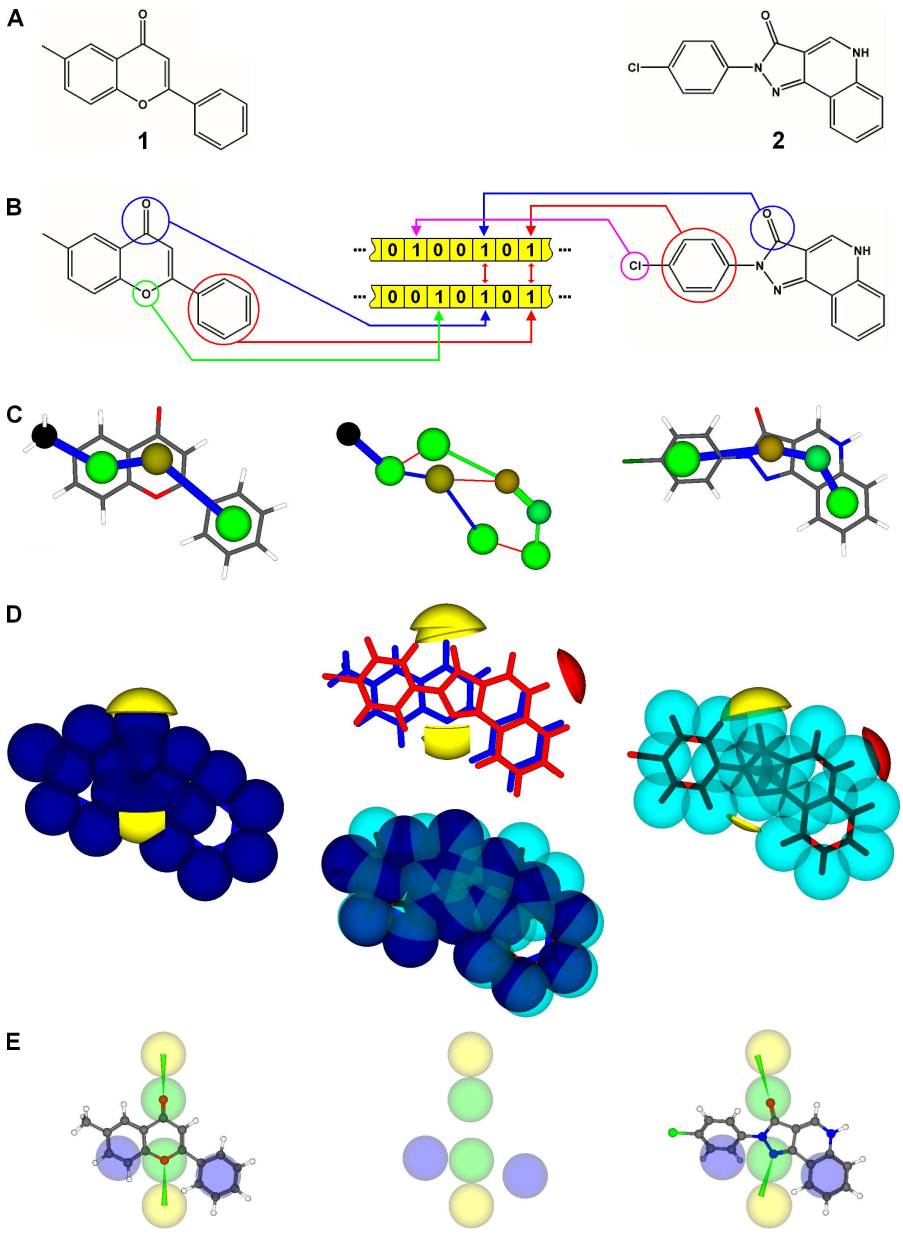

Figure 18.3: Comparison of common ligand-based screening techniques (see text for details). (A) Example molecules: Flavonoid molecule (**1**, test molecule) and a known binder (CGS-9896, **2**, reference molecule) for the benzodiazepine site of the $GABA_A$ receptor [89]. (B) Bitstring generation and comparison. (C) Generation and comparison of Feature Trees. (D) Comparison by molecular superimposition using FlexS (center top: molecular interaction surfaces, center bottom: molecular volume represented by Gaussian functions. (E) Comparison of both molecules to a pharmacophore model (center).

Many different similarity coefficients have been proposed in the literature. Some are measures of dissimilarity, while others measure similarity directly. In this review, we focus on the most widely used similarity and distance coefficients only, more details can be found in another review [8]. One of the most frequently used distance measures is the Euclidean distance $D_{A,B}^{Euclidean}$ between two molecules A and B described by properties $x_1 ... x_n$. For continuous property variables, the distance is defined by

$$D_{A,B}^{Euclidean} = \sqrt{\sum_{i=1}^{n} (x_{i_A} - x_{i_B})^2}. \tag{18.1}$$

The most often used similarity measure is the Tanimoto coefficient $S_{A,B}^{Tanimoto}$. This coefficient can be interpreted as the fraction of the number of features present in both molecules divided by the number of features present in at least one of the compounds. For continuous variables it is defined by

$$S_{A,B}^{Tanimoto} = \frac{\sum_{i=1}^{n} x_{i_A} x_{i_B}}{\sum_{i=1}^{n} x_{i_A}^2 + \sum_{i=1}^{n} x_{i_B}^2 - \sum_{i=1}^{n} x_{i_A} x_{i_B}}. \tag{18.2}$$

Willett *et al.* [212] assessed the performance of several distance and similarity coefficients for predicting a measured activity value. The Tanimoto coefficient was performing best and, since then, it has become the 'standard' coefficient for chemical similarity comparison. Furthermore, this coefficient was shown to be the most appropriate for similarity searches in 2D databases [212].

In order to compare two molecules in a quantitative fashion, numerical values of attributes of the molecules are needed. The term 'molecular descriptor' subsummarizes all numerical representations of chemical information about molecules obtained by a defined mathematical procedure. As example, the molecular descriptor 'molecular weight' is defined exactly as the sum of all atomic weights of a molecule

$$MW = \sum_{i=1}^{n_{atoms}} m_i. \tag{18.3}$$

The number of different chemical descriptors which have been proposed for the field of quantitative structure-activity relationships (QSAR) and VS is large (several thousands are described in detail in the handbook by Todeschini and Consonni [189]) and an extensive discussion is out of the scope of this review. However, some descriptors have become very important in VS. Among them, logP, MW, and molar refractivity are typically used in the preprocessing of libraries (see Section 18.2.2). BCUT descriptors are useful in diversity analysis (see Section 18.2.4). A second important class of descriptors are bit strings, used in molecular similarity search, as detailed in the next Section 18.3.2.

18.3.2 Bit string descriptors

For VS the most popular descriptors are based on binary vectors (also called bit strings). The idea of using binary vector representations has its origin in chemical database systems. Bit

strings are used in substructure queries to efficiently discard large fractions of the database, before subsequently a much slower subgraph isomorphism algorithm is used. The length of the bit strings varies from roughly a hundred bits [218, 219, 220] to several million bits [136], depending on the type of information stored. The two most widely used approaches are substructure keys and hashed fingerprints.

Substructure keys (MDL keys available from Elsevier MDL (http://www.mdli.com/) and BCI structure fingerprints available from Barnard Chemical Information Ltd. (http://www.bci.gb.com/)) represent a description of the substructures present in a molecule (Figure 18.3 (B)). Within the binary vector, each of the positions is 1, if the corresponding substructure is present in the molecule, 0 otherwise. All substructures for the bit string are predefined in a fragment dictionary. Entries can encode the presence of certain atoms (e.g. there is at least one N present in the molecule) or common functional groups (e.g. ester function). Furthermore, electronic (e.g. O with double-bond) and structural features (e.g. six-membered ring) are described by bits.

In hashed fingerprints (Daylight Chemical Information Systems, Inc., http://www.daylight.com/), substructure information is encoded by an algorithm. All possible paths of atoms and bonds through the 2D formula of the molecule, up to a certain certain path length are generated by systematic search, e.g. for a path length of three the path C=C-N-C. These patterns are then converted to integer numbers and hashed. The hash code is generated using these numbers as seeds for a pseudo-random number generator, resulting in the fingerprint of this path. The advantage of hashed fingerprints is their applicability to any type of structure without the need of a pre-computed fragment dictionary. A disadvantage is the possibility of the same bit string being calculated from different atom paths which may result in false positives, since these hash conflicts are not solved. Unity available from Tripos (http://www.tripos.com) uses a combined approach of substructure keys and hashed fingerprints.

For the use with binary variables, the two similarity coefficients from above can be reformulated. If a denotes the number of bits set (to 1) in molecule A, b for molecule B, and c the number of bits set in both molecules, the Euclidean distance becomes

$$D_{A,B}^{Euclidean} = \sqrt{a + b - 2c} \tag{18.4}$$

and the Tanimoto coefficient is given by

$$S_{A,B}^{Tanimoto} = \frac{c}{a + b - c}. \tag{18.5}$$

The substantial advantage of linear descriptions for chemical structures is their speed. The counting of numbers of bits set (a, b, c) can be done computationally very fast, resulting in several hundred thousand molecule comparisons per minute. This allows for fast comparison of millions of compounds in minutes to hours. A disadvantage of the methods described so far is that they cannot detect the similarity between two compounds, which behave similar with respect to binding to the target protein, but are structurally quite different [35]. Thus, the techniques cannot find a scaffold different from the scaffold of the reference molecule (scaffold-hopping).

18.3.3 Feature trees

Feature trees [166] comprise a class of descriptors which are in between the classical linear descriptors described in Section 18.3.1 and molecular superimposition techniques (Section 18.3.4). In this technique (Figure 18.3 (C)), a molecule is described as a tree, that represents its overall topology. Nodes of the tree represent fragments of the molecule. The nodes are connected by edges if the fragments are also connected by covalent bonds or sharing of atoms. A set of features is assigned to each of the nodes, representing physicochemical properties of the respective fragment. Steric features comprise the number of atoms in the fragment and the approximated van der Waals volume. Chemical features include the interaction profile of the fragment, i.e. whether it acts as hydrogen donor or acceptor, is an aromatic ring, and also represent the hydrophobicity of the fragment.

Similarity between molecules is calculated by matching the two corresponding trees, while preserving their topology. A similarity score quantifies the quality of the fit. The advantage of feature trees provides a more accurate description of chemical properties than by linear descriptors. However, the tree matching procedure is slower, due to the higher computational complexity of the tree comparison. Typically, thousands of molecules can be compared per minute.

18.3.4 Molecular superimposition approaches

Molecular superimposition techniques structurally align a compound to a reference ligand in three-dimensional space. During the alignment matching parts of both molecules are placed on top of each other. The large variety of algorithms for molecular superimposition has been reviewed by Lemmen *et al.* [123]. The application of superposition techniques for VS has also been reported [121].

In general, the molecular superimposition can be achieved in two ways: Either a field-based approach is used, in which properties of the molecules are projected onto a common surface or into three-space. The other method aligns pairs of atoms directly. Early approaches achieved this goal by rigid-body superimposition. Newer programs can handle one or both molecules as flexible on the fly. Nevertheless, the rigid-body techniques are much faster and are thus often preferred. An intermediate technique is to address molecular flexibility by considering a set of alternative conformations of the molecule.

A rigid-body superimposition program reads a reference ligand and a test ligand, then performs an optimization of the position and orientation of the test ligand in space. Early attempts using combinatorial approaches to enumerate efficiently possible matches (correspondences) of chemical features of the two molecules [91, 133] are too compute-time intensive. With the program SEAL (Steric and Electrostatic ALignment) [92] and later enhancements [98], for the first time, Gaussian functions were used for describing the physicochemical properties of the molecules and a new algorithm was applied to tackle the rigid superpositioning problem efficiently. The description of chemical features by Gaussian functions has several advantages [68]. First, a Fourier transform of a Gaussian is again a Gaussian. Second, there is no boundary which helps in the initial steps of alignment. Furthermore, derivatives can be calculated easily (even symbolically), and finally, the overlap between two Gaussians increases when their maxima approach each other. An improvement of the search algorithm was proposed by

Lemmen *et al.* [120] in their program RigFit. The optimization is split into two independent optimizations for rotation and translation. Thus, one six-dimensional search is separated into a sequence of two three-dimensional searches.

Current state of the art are programs for flexible superpositioning of two molecules. Sheridan *et al.* [180] uses distance geometry for superposition, while Itai *et al.* [86] proposed a technique, in which all possible matchings between pharmacophoric points are evaluated in a combinatorial matching. For both techniques, still the definition of the pharmacophore is needed as prerequisite. The combined Monte-Carlo and energy-minimization based technique by McMartin and Bohacek [137] also needs manual intervention. The program GASP [88] was the first method available that was able to handle the structural flexibility of both molecules involved and was not constrained by predefined relationships between functional groups assumed to be similar. GASP is based on a genetic algorithm which mimics the process of evolution. The conformations of each molecule and the correspondences between intramolecular features are coded via so-called chromosomes. In order to modify the superimposition, the chromosomes are subjected to the operations of mutation (local changes) and crossover (splitting and merging of two chromosomes). A population of chromosomes is repeatedly subjected to these modifications and then evaluated with a fitness function. Only the fittest chromosomes survive to the next round. The fitness function used in the selection process of each superposition is calculated by volume overlay, intermolecular matching energy and the conformational energy.

A different technique for superposition, FlexS, uses incremental construction [122]. Here the reference molecule is handled as rigid, the test molecule is flexible. The test molecule is partitioned into fragments which are connected by rotatable bonds. A number of relatively rigid fragments is selected and aligned to the reference molecule. Then the next fragment is attached to the previously placed fragment in all allowed torsion angles. The list of admissible torsion angles is derived by statistical analysis [100] of the Cambridge Structural Database (CSD) [6]. All generated placements are scored by paired intermolecular interactions and overlap, the latter being described by Gaussian functions (see Figure 18.3 (D)). The best partial solutions are subjected to the next incremental construction cycle until the complete test molecule is build up. The mean computing time is in the order of 30 seconds per superpositioning for typical test cases.

Krämer *et al.* [107] developed fFLASH, using a fragmentation-reassembly approach. The tool is based on earlier work on FLASHFLOOD [164]. fFLASH describes the query molecule as rigid, the test molecules are handled flexibly. All test molecules are partitioned into fragments by severing rotatable bonds, expanded to a set of conformations, and all conformers stored in a database. Pairs of adjacent fragments are joined and a set of conformations of the fragment-pair is generated by varying the dihedral angle at the connecting bond. Molecular interaction features are then calculated and stored in a lookup table. By use of a clique detection algorithm, patterns of features of fragment-pairs of the test molecule are geometrically matched on the reference molecule. These matches are subsequently joined, based on the pairwise compatibility of two matches, by a graph-algorithm.

18.3.5 Pharmacophore searches

A pharmacophore is usually defined as a set of molecular features and their rigid spatial arrangement, which is necessary for ligand-receptor binding [73]. A pharmacophore is typically composed by three to four pharmacophoric centers and their respective distances (Figure 18.3 (E)). Pharmacophores are applied in three-dimensional database searches after they have been determined from a set of active ligands. In VS, pharmacophores can also be used as constraints in structure-based screening. Pharmacophores can also act as three-dimensional descriptors. Often pharmacophores are encoded in the form of bitstrings, known as pharmacophore fingerprints, which are directly applicable for screening (see Section 18.3.2).

In ligand-based VS, the true pharmacophore is unknown and must be determined first. Pharmacophore perception is closely related to molecular superpositioning, for example, the program GASP [88] can perform both tasks. Nevertheless, since the programs of both groups are tailored to their specific research areas they are described separately. For automatic determination of a pharmacophore hypothesis a set of active ligands as training set is needed. The pharmacophore perception is then performed in a number of steps: First, three-dimensional structures of the molecules must be generated with one of the methods described in Section 18.2.1. Then the molecules are analyzed in order to identify atoms that can interact with a protein in a characteristic way. Commonly these pharmacophoric features are acidic and basic groups, hydrogen acceptor and donor sites, aromatic, and hydrophobic groups. In the next step, conformations of the molecules are passed to the pharmacophore perception algorithm. Here, conformations of the molecules are compared in order to identify pharmacophoric features common to all molecules. A number of programs have been developed for pharmacophore identification [73], among them the commercially available tools Catalyst/HipHop [33], DISCO [132], and GASP [88]. For a recent review and a comparison of these three, see Patel *et al.* [152]. These programs differ with respect to how the conformations are handled and how the molecules are aligned and compared. GASP uses a genetic algorithm (see Section 18.3.4) to describe the molecules as flexible. DISCO uses a set of low-energy conformations which are kept rigid throughout the calculation and a clique-detection algorithm is used for rigid-body alignment. Catalyst/HipHop also uses a set of rigid low-energy conformations of the molecules but then performs a pruned exhaustive search to identify configurations common to all molecules. Once the pharmacophore is identified, it can be used to screen a three-dimensional database.

An extension of the ligand-based pharmacophores described so far, is the use of structural information of the receptor for pharmacophore generation. Two recent examples of these structure-based pharmacophores are the works of Wolber and Langer [214] and Griffith *et al.* [72].

18.3.6 Quantitative structure-activity relationships (QSAR)

The structure and the physicochemical properties of a molecule can be used to model its biological activity. The mathematical description of this relationship in a quantitative way is the aim of QSAR techniques. In order to model structure-activity relationships, first, for each compound in the library a number of molecular descriptors have to be calculated. In a second step, a quantitative relationships between these descriptors and the activity is derived. This

section covers only some selected techniques from the field of QSAR which has grown in terms of using more and more sophisticated descriptors and also more sophisticated statistical tools for finding correlations between structure and activity.

The classical technique is Hansch analysis which correlates activity with physicochemical properties by use of regression analysis. Hansch *et al.* [77] described the dependency of the concentration C needed for a certain biological response in terms of the hydrophobicity (expressed by the logP value) and electronic effects (using the Hammett constant σ) by the equation

$$\log \frac{1}{C} = k_1 \cdot \mathrm{logP} + \mathrm{k}_2 \cdot \sigma + \mathrm{k}_3. \tag{18.6}$$

Here, the k_i are the coefficients to be fitted by the regression. Using this type of QSAR analysis, today, several thousand successful applications have been reported and a database of QSAR equations is electronically available (http://www.cqsar.com/medchem/chem/qsar-db/). The descriptors applied include steric, electronic, and hydrophobic effects as well as indicator variables. These values are obtained either by computer prediction techniques or experimentally. Due to the large number of descriptors available the dependency between them has to be studied in order to find the relevant ones. This ”feature selection” is usually done based on Principal Components Analysis (PCA) [61] or its extension Partial Least Squares Analysis (PLS) [82].

An extension of the classic approach was the introduction of three-dimensional information of the ligands to reflect the geometry of their binding to receptors, including their chirality. The first of this 3D-QSAR techniques was the Comparative Molecular Field Analysis (CoMFA) [34] which turned out to be very successful (many examples can be found in [113, 111, 112]). In CoMFA a set of molecules is selected which have an identical binding mode, i.e., they bind to the same site in the same relative geometry. To derive the CoMFA model, for all training set molecules, first, partial charges are assigned and low-energy conformations are generated. Then the molecules are aligned by use of a pharmacophore hypothesis and positioned inside a 3D grid. For each grid point and for each molecule separately, ’field’ values (interaction energies) are calculated for charged and uncharged probe atoms. Finally, PLS analysis is used to correlate the fields with biological activity data. The result of this analysis is typically represented as a set of contour maps showing favorable and unfavorable regions for certain substituents. Several techniques have been proposed to obtain better fields. By calculating fields with GRID [71] or HINT [94] more different probes can be used which allows modelling of a wider range of interactions. By replacing CoMFA potentials with SEAL (see Section 18.3.4) similarity fields (Comparative Molecular Similarity Indices Analysis, CoMSIA [99]) the results become more stable. A frequent problem for PLS can be the high number of noise variables not contributing to the description. With GOLPE (Generating Optimal Linear PLS Estimations) [10] the meaningful variables can be selected and the predictive ability of the model is checked by cross-validation. A cause of error for CoMFA, CoMSIA, GRID/GOLPE and related techniques is the mutual alignment of all molecules. There are methods available that retain the 3D information but are independent of the alignment. Examples for these techniques are WHIM (Weighted Holistic Invariant Molecular indices) [188], which uses the moments of atomic properties as descriptors, and the related technique MS-WHIM [21], which uses molecular surface points instead of the atoms as descriptors.

In the classical 3D-QSAR methods described above, only information on the geometry of the ligands is used. In 4D techniques multiple conformations or orientations of the ligands are considered simultaneously. It is even possible to include information on the protein structure to which the ligands are bound in the QSAR model. The program Quasar by Vedani *et al.* [195] is a method which constructs a receptor-surface model and bridges between 3D-QSAR and receptor modeling, taking induced fit into account. Currently, multidimensional QSAR studies are extended up to six dimensions to allow for the simultaneous consideration of different solvation models [194].

18.3.7 Other techniques

The interaction of a ligand with a target molecule can be described in terms of the respective molecular surfaces, that have to be complementary with respect to both physicochemical properties and shape. Finding optimal surface complementarity is the main aim of docking procedures (cf. Chapter 16). Thus, the comparison of different ligands in terms of their molecular surfaces and the properties mapped to them is a valuable similarity criterion.

Among the many techniques of molecular surface comparison, we focus on the recent graph-based method SURFCOMP of Hofbauer *et al.* [79] and the gnomonic projection method [17] as examples. The comparison of two surfaces, each described by a point set in three-space is not an easy task. The problem can only be solved efficiently if the surface model is simplified. In SUFRCOMP, first a representation of the surface via overlapping circular patches is calculated. Then the centers of these patches, representing critical points, are reduced in number using a number of filters and matched via maximal common subgraph comparison.

Blaney *et al.* [17] use gnomonic projection of the molecular surface properties onto equispaced points on the surface of an enclosing sphere. To do so, the points in space at which vectors from the sphere's surface to the 'center' of the molecule cut the molecular surface are calculated. The physicochemical properties on the cutpoint farthest from the 'center' are then projected onto the sphere. The comparison of the projections of two molecules is then performed after mapping of the property values on two dimensions.

A different type of measuring similarity between molecules is the use of 'virtual affinity fingerprints' [124, 208]. In the Flexsim-X method by Lessel and Briem [124] ligands are flexibly docked into a carefully selected reference set of protein binding sites using the FlexX docking program [165]. The highest-ranking solution of each docking run is selected. The virtual affinity fingerprint of a ligand is then defined as the vector of docking scores obtained for the different binding pockets. Molecules are compared by the Euclidean distance between their affinity fingerprints. The technique was shown to detect molecules with similar biological affinity without prior knowledge of the target protein structure. An extension of this work to calculate similarities of functional groups is Flexsim-R [208].

18.4 Post-processing of hitlists

HTS or VS runs of a compound collection with up to millions of entries results in a huge volume of data. The obtained list of hits is rather crude and needs substantial clean-up. There are a number of computational methods for the post-processing and analysis of screening

data. First, the output is simplified by removing data points where the screening failed (e.g. no docking solution, failure during experiment). Here, also data not needed in post-processing (e.g. intermediate results, log files) are discarded. Second, the most promising hits have to be selected mostly on the basis of their rank in the hitlists or by criteria based on the scores, in order to reduce the dataset to manageable size. To identify leads among the screening data is a challenging problem, addressed by a number of different computational methods. The often concealed information can be extracted by data mining procedures (Section 18.4.1). A general problem of screening data consists of false positive results. Especially for results of structure-based VS runs, some techniques have been developed to identify and discard false positives (Section 18.4.2). Whenever a combination of different techniques is used in screening, each technique result in a different hitlist. Here, consensus techniques help in picking hits (Section 18.4.3). Nevertheless, the most important method is still the visual inspection of the results by an experienced medicinal chemist, assisted by visualization tools (Section 18.4.4).

18.4.1 Data mining

A common approach for mining hitlists is the search for families with similar chemical structure among the active compounds. Here active compounds (actives) are those with high affinity, high scores, or high similarity after screening. Chemical families can be identified by grouping the compounds with similar chemical structure. A chemical family is characterized by a common scaffold. Substructure search among the results can be applied to identify these families. Roberts *et al.* developed LeadScope [171], a structural classification technique. The method classifies compounds into a collection of predefined chemical families. The predefined families are arranged hierarchically, starting with a major structural class on top, which is subdivided further. For example, a 3-methoxy-pyridine derivative, is found in the pyridine $\rightarrow$ pyridine, 3-R $\rightarrow$ pyridine, 3-alkoxy class of the hierarchy. For each structural class, activity data and frequency in the data set is depicted in an intuitive bar plot.

As an alternative, techniques for similarity search (Section 18.3) can be applied to identify families. In this case, the families are defined by a high degree of similarity. For grouping the families, often clustering techniques are used (as described in Section 18.2.4). Due to the importance of hitlist mining, a number of dedicated clustering techniques have been developed [49, 185].

Another approach to data mining using classification techniques is recursive partitioning (RP) [173, 221]. RP is a nonparametric classification technique (as opposed to the many parameters in QSAR models), in which the whole set of compounds is recursively classified into disjoint subsets using statistically determined rules. In this manner, a tree is constructed, in which some terminal nodes (leaves) are enriched with actives, while other leaves contain mostly inactive molecules. If the path from a leaf with actives is traced back to the root node, the molecular descriptors used for partitioning at the inner nodes can be used to characterize or to search for actives.

Nicolaou *et al.* developed a classification method using a phylogenetic-like tree (PGLT) [147]. This tree is constructed using a combination of techniques. Each node has bins for active and inactive compounds. First, all active molecules are stored in the active bin of the tree's root node. Then, in an iterative fashion, a clustering of the molecules of the current leaf is performed, using a criterion based on chemical descriptors. In a next step, cluster

level selection is performed to select a set of 'natural' clusters. Each of the natural clusters is then subjected to a maximum common subgraph (MCS) search. Common substructures are evaluated by a set of rules to evaluate each and to discard all those, not providing new knowledge. The rules, for example, discard substructures already found in other nodes or those identical or subsets of the parent node. Then, for each of the remaining substructures, all molecules from the parent node containing the respective MCS are added to a newly created tree node. Finally, a node is selected at which the iteration proceeds. After the actives have been used to construct the tree, a post-processing procedure is performed in order to prune the tree and reduce it to contain only nodes with structurally homogeneous families. This is done by adding inactive compounds to the inactive bins of the PGLT using the substructure rules derived with the actives. For each node, the similarity between actives and inactives is calculated and nodes with dissimilarities are eliminated. The technique described has been implemented in the program ClassPharmer (Bioreason S.A.R.L, http://www.bioreason.com/).

18.4.2 Analysis of the protein-ligand interface

A particularly interesting type of strategy, which can be applied to results of structure-based screening, is the analysis of structural properties of the bound protein-ligand complex. Although this method also belongs to docking techniques (see Chapter 16), we describe it here as a representative example for an important class of post-processing techniques. Current scoring functions favor the formation of many protein-ligand hydrogen bonds and salt bridges, even if the structures exhibit only limited steric complementarity overall due to holes along the interface or larger parts of the ligand being exposed to the solvent. Stahl and Böhm [184] propose a post-processing procedure of docking results. For a set of generated docking poses, first, all poses with close contacts between polar atoms that do not take part in hydrogen bonds are discarded. Then the fraction of ligand volume located inside the cavity is calculated. Poses with less than average buried volume are discarded. The size of lipophilic cavities at the protein-ligand interface also acts as filter criterion: Poses exceeding the minimum value by more than 25 Å are discarded. Finally, the solvent-accessible surface of nonpolar parts of the ligand is calculated and used for rescoring.

Giordanetto *et al.* [67] also propose the use of solvent-accessible surface areas. These authors perform a classification of all receptor and ligand atoms into classes, depending on the physicochemical properties hydrophilicity, charge, and hybridization. Then descriptors are calculated that describe the energetic cost of burying the atoms. In addition, conformational entropy differences between holo and apo form of the protein are calculated. Here, an amino acid-based conformational entropy contribution of the protein after Murphy and Freire [145] to the binding affinity is used. By use of these techniques, affinity predictions could be improved on the cost of less accurate binding mode prediction.

Results from docking studies can also be analyzed by structural interaction fingerprints as proposed by Deng *et al.* [41]. These interaction fingerprints are a translation of the structural information of a protein-ligand complex into a binary vector. The technique can be applied for identification and clustering of similar docking poses.

18.4.3 Consensus techniques

The combination of several different computational methods is another approach to reducing the number of false positives and prioritizing molecules for further study. Some of these methods are only applicable to structure-based techniques, while some use mixtures of different computational methods, including ligand-based techniques. Prototype of the structure-based methods in this field is consensus scoring [28]. Here, one docking program is used to generate a docking pose. Then, the highest ranking structure is reevaluated with different scoring functions. If the compound is not among the top-scoring compounds for all scoring functions applied it is discarded. In a computer experiment by Wang and Wang [205] it has been shown that hit rates improve significantly after consensus scoring if three or four scoring functions are used.

Methods using not only different scoring functions but different docking techniques go a step further [154]. In the ConsDock approach, docking is performed with three different docking programs and a set of 30 top ranking poses is stored obtained with each of them. Then a hierarchical clustering is performed on each set and the highest-ranking pose within each cluster is defined as its 'leader'. Consensus pairs are defined, where two of the docking program result in similar leaders. Each of these pairs is then described by its mean and clustered again into classes. Finally, the mean pose of the clusters is subjected to re-ranking according to the number of entries in each class.

The use of entirely different computational techniques for investigation of hitlists has been proposed by some groups. Klon and coworkers [102, 103] use a combination of docking and machine learning. First, docking of a library is performed with three different docking programs. Then a näive Bayesian classifier is trained on the docking scores of the top-scoring compounds, which are labeled as 'good', if their score is better than a threshold. The compounds themselves are described by an extended-connectivity fingerprint as structural descriptor (Pipeline Pilot program available from SciTegic, http://www.scitegic.com/). Application of the Bayesian classifier for re-ranking the hitlists improved the enrichment in most of the test cases, without any *a priory* knowledge of the activity of the compounds.

Especially in docking, the high-dimensional search space can be explored a bit further to re-rank hitlists. On the one hand, a multi-conformer description of the protein can be used [199]. On the other hand, not only the top-ranking pose but several poses can be used for calculating the score [104].

Ginn *et al.* [66] proposed the use of data fusion for combining molecular similarity measures. In this procedure, a similarity search is performed with at least two different similarity measures i. The rank positions r_i of each individual structure in the hit lists are then combined to a new score. With the fusion rule $\sum_{i=1}^{n} r_i$ the performance is at least as good as the best individual measure.

18.4.4 Visualization

For the simultaneous display of screening-result data in several dimensions, a number of techniques are available [3, 63, 116]. The techniques have been implemented in several tools for display of screening data using highly sophisticated graphical data representations for visual data mining (DecisionSite (Spotfire, Inc., http://www.spotfire.com), ClassPharmer (Biorea-

son S.A.R.L, http://www.bioreason.com/), LeadNavigator (LION Bioscience AG, http://www.lionbioscience.com/)). Results are plotted in multiple dimensions, combining data from different databases. The data points in the plots are linked to the corresponding chemical structures and vice versa. This enables the medicinal chemists to identify patterns within the results. A technique for visualization of the multidimensional screening data is the non-linear mapping of the data to a lower-dimensional space with just 2 or 3 dimension. The usual techniques for non-linear mapping is multidimensional scaling [110]. This technique aims at keeping points close together in low-dimensional space if they are also close together in the original data-space. With recent enhancements [2, 216], multidimensional scaling is applicable to large screening data sets. Despite all efforts in visualization techniques, it has been pointed out that visual data mining tools are not applicable to extremely large and complex data sets [147]. Furthermore, due to their 'interactive' approach, these tools cannot readily be integrated into fully-automated screening procedures.

18.5 Critical evaluation of structure-based virtual screening

Nowadays a large collection of docking/scoring tools is available for high-throughput virtual screening. Out of the flow of information generated over the last five years, a computational chemist entering into a virtual screening project will have to make a few decisions about the screening strategy and the tools which are the most suited to its project. The first part of this section is aimed at pinpointing some good practices in order to avoid classical failures. The second part of the section will review some recent success stories which could inspire the reader for future work.

18.5.1 Influence of parameter settings

Several input parameters may affect the effectiveness of a VS run. Depending on the computational tool that has been chosen, the number of parameters may vary from a dozen to over one hundred. It is therefore crucial to select the best possible input settings which unfortunately are not always known in advance. However, a few robust guides based on current knowledge can be derived.

Which library?

As reported above (Section 18.1.3), several commercially available compound collections are available. There is usually no reasons to favor one particular compound collection over another one. As most of them are easily accessible [11, 182], the best possible approach for an academic user is to start from a unified and filtered dataset [85]. Of course, corporate and focussed/targeted libraries may also be used. They are particularly interesting for screening targets belonging to deeply-investigated families (e.g. kinases, GPCRs) and containing a high percentage of true positives.

Whatever the database selected, it is generally advisable to downsize the number of molecules which will be submitted to 3-D docking. Beside some important filters (chemical

reactivity (see Section 18.2.1), drug-likeness (see Section 18.2.3), etc.), it is important to remove molecules which do not fulfill simple 2-D or/and 3-D pharmacophoric features. This simple strategy aids in dramatically reducing the number of potentially interesting compounds without losing many true positives [51, 129]. If one is simply interested in setting-up optimal screening conditions (e.g. discriminating a few true actives from randomly-chosen decoys), it remains important to carefully set-up the test dataset in order to avoid artificial enrichments in true actives by making sure that chemical spaces covered by actives and inactives/random compounds largely overlap [197].

Which ligand conformation(s)?

Most docking programs only requires a single low-energy conformation for each ligand of the dataset, provided by automated 3-D converting utilities [157, 176]. For docking tools requiring a multi-conformer ligand library, it is important to start from biologically-relevant conformations. Several studies agree to conclude that the most reliable conformations are not necessarily produced by the most accurate and cpu-demanding methods. A safe start is to use fast conformer generators like Omega (http://www.eyesopen.com/products/applications/omega.html) or Catalyst [33] which accurately sample the biologically-relevant conformational space for a wide array of chemotypes [19, 70, 93, 96].

Which protein coordinates?

When screening a high-resolution X-ray structure, several input coordinates might be available describing either ligand-bound (holo) or a ligand-free (apo) structures. A systematic survey over nine enzymes unambiguously demonstrates that the holo form if it exists should be the first choice [19]. Furthermore, X-ray structures appear to clearly outperform the corresponding homology models in discriminating known inhibitors from random decoys [134, 150]. However, if the sequence identity (on binding site-lining residues) to the X-ray template is higher than 50 %, comparable enrichment rates in true inhibitors can be found [150]. This encouraging results suggests that genomic-scale VS might be feasible, provided that an accurate description of the binding sites can be drawn from existing X-ray templates.

Which docking tool?

Starting from the pioneering work of Kuntz and coworkers [115], numerous docking programs based on very different physicochemical approximations have been reported (see Chapter 16). All docking tools combine a docking engine with a fast scoring function, and the recent literature is full of benchmarks addressing the accuracy of one or few docking/scoring scenarios. The three following issues are usually investigated: (i) the capability of a docking algorithm to reproduce the X-ray pose of selected small molecular-weight ligands [93, 105, 159], (ii) the propensity of fast scoring functions to recognize near-native poses among a set of decoys [56, 203] and to predict absolute binding free energies [56] (iii) the discrimination of known binders from randomly-chosen molecules in virtual screening experiments [36, 93, 159]. However, analyzing all these data for a comparative analysis of available docking tools

is very difficult. First, many tools are not easily available. Second, independent studies assessing the relative performance of docking algorithms/scoring functions are still rare and focus on the usage of few methods. Third, the quality judgment may vary depending on the examined properties (quality of the top-ranked pose, quality of all plausible poses, binding free energy prediction, virtual screening utility). Fourth, most docking programs assume approximation levels that can vary considerably [74] and lead for example to very inhomogeneous docking paces ranging from few seconds to few hours. Last, many docking programs have been calibrated and validated on small protein-ligand datasets. Hence, detailed benchmarks (> 100 PDB-ligand complexes) are only reported for few docking tools [28, 43, 108, 149, 151, 196]. The most recent validation studies on different datasets agree to conclude that the accuracy of a docking tool is largely target-dependent [74, 96, 159, 203] and should be examined on a case-by-case basis. Glide and Gold seem to be the most robust programs for their propensity to generate near-native poses in ca. 75-80 % [96, 159], provided that several solutions are stored. A major problem is that the scoring function does not always predict the correct solution as the most probable one (only in ca. 40–50 % of the cases). This considerably complicates the analysis of docking results. Numerous reasons explain this limited accuracy [93]. Some are easy to correct (e.g. incorrect atom type for either the ligand or the protein), some are more difficult (e.g. accuracy of the protein 3-D structure, flexibility of the ligand, accuracy of the scoring function), and some are really tricky to overcome (protein flexibility, role of bound water). The accuracy of a docking program to predict the protein-bound ligand pose is reflected in its virtual screening efficacy, that is the ability to discriminate true binders from inactives and/or randomly-chosen compounds [36, 93, 159]. However, predicting which docking program will be the most suited for a research project is still problematic. If known ligands are available, a pragmatic approach is to try a systematic combination of docking/scoring parameters and select for productive screening the one that best segregates true actives from true inactives. If no or very few ligands are available, some guides may be followed to choose the tool that seems the most appropriate regarding the physicochemical properties of the protein cavity [93].

Which scoring function?

The scoring function still remains the Achilles' heel of structure-based virtual screening. Several recent and independent studies conclude that many fast scoring functions can indeed distinguish near-native poses (rmsd lower than 2.0 from the X-ray pose) from decoys for ca. 70 % of high-resolution protein-ligand X-ray structures [56, 203]. However, when docking is applied to a large database, the corresponding scoring function should be robust enough to rank putative hits by increasing binding free energy values [97]. Unfortunately, an accurate prediction of absolute binding free energies is still impossible whatever the method [36, 56, 204]. Predicting binding free energy changes is possible at the condition that a customized scoring function is applied to a series of congeneric ligands. However, for a database containing a large diversity of compounds, and for targets which have not been traditionally used for calibrating scoring functions, the obtained accuracy is usually limited (ca. 7 kJ/mol or 1.5 pK unit) [75]. From this observation, two sources of improvement are possible: (i) design more accurate scoring functions [204], (ii) design smarter strategies to post-process docking outputs (see next section). Many computational chemists actually favor the second option. The accuracy of scoring functions has levelled off several years ago, for the simple reason that some

unknown parameters (e.g. role of bound water, protein flexibility) remain extremely difficult to predict whatever the physical principles used to derive a scoring function.

Which post-processing?

Acknowledging that scoring functions are far form being perfect, the easiest way to retrieve true positives from a virtual screen is to first detect false positives. Many strategies are possible. The simplest consists in rescoring poses with additional scoring functions; hoping that a consensus scoring [15, 28] will better identify true hits (top-ranked by several scoring functions) from decoys (see Section 18.4.3. Comparing hit rates between simple and consensus scoring should however be realized on hit lists of comparative size [217]. Moreover, customizing a consensus scoring scheme requires first the knowledge of several and chemically-diverse true hits. Such data are not always available. Therefore, for less well-investigated targets, other strategies have to be designed. Topological filters can be used to filter out poses exhibiting steric or electrostatic mismatches between the ligand and its target [184]. Poses can also be minimized by a more accurate force field [90, 186], hierarchically clustered [154], analyzed by Bayesian statistics [103]. In any case, the post-processing treatment should be simple enough to be reproducible for a wide array of targets. The influence of different post-processing strategies on the hit rate and the percentage of true hits recovered is shown in Figure 18.4. In this figure, the top-right corner with a hit rate of 100% and all true hits recovered would be the optimum.

An alternative strategy for post-processing is to look at enrichment among true hits in pre-computed substructures/scaffolds [147]. This presents the advantage of focusing more on scaffolds and the distribution of docking scores among them, and less on individual molecules. The effect is evident from Figure 18.4, where the results of such a post-processing are closer to the optimal corner. Therefore, false negatives may be recovered if they share a scaffold with true positives. Last but not least; selected hits should be browsed in 3-D target space for the ultimate selection: no algorithms yet outperform the brain of an experienced modeler for such a task!

18.5.2 Recent success stories

Only recent reports from the literature (2003–2005) will be reviewed herein. Most of them still make use of high-resolution X-ray structures (next three subsections). However, encouraging data begin to emerge from homology models (last subsection in this section) and thus broaden the application of structure-based screening methods to a wider array of pharmaceutically-interesting targets.

Some privileged targets

Macromolecular targets presenting a well-defined hydrophilic pocket for which the directionality of intermolecular interactions play a key role in ligand recognition are particularly well suited for virtual screening for the simple reason that most docking tools and scoring functions have been calibrated for such situations [56]. Thus, it is no surprise that some protein families

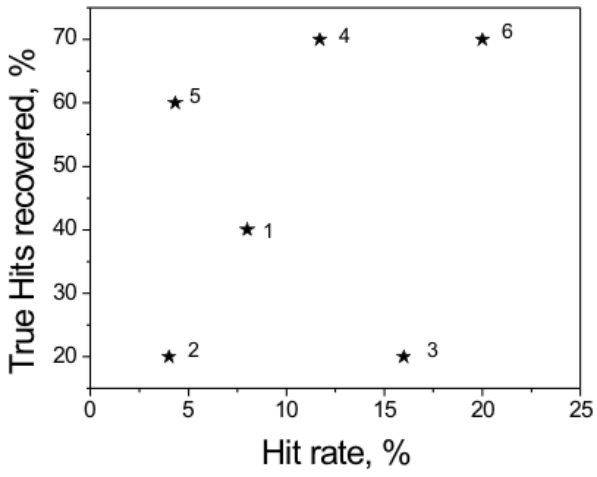

Figure 18.4: Influence of post-processing strategies in retrieving true vasopressin V1a receptor antagonists by structure-based screening of a database of 990 randomly-chosen 'drug-like' compounds seeded with 10 true actives [16]. 1) top 5% ligands as scored by FlexX; 2) top 5% ligands as scored by Gold; 3) hits common to 1) and 2); 4) ClassPharmer (Bioreason S.A.R.L., http://www.bioreason.com/) prioritization of scaffolds for which 60% of the representatives have a FlexXscore lower than $-22\,\mathrm{kJ/mol}$; 5) ClassPharmer prioritization of scaffolds for which 60% of the representatives have a Goldscore higher than 37.5; 6) ClassPharmer prioritization of scaffolds for which 60% of the representatives have a FlexX score lower than $-22\,\mathrm{kJ/mol}$ and a Goldscore higher than 37.5.

(e.g. kinases) are overrepresented in targets for which true inhibitors have been discovered by database docking (Table 18.1).

Protein kinases have been deeply investigated by structure-based virtual screening [59, 83, 129, 158, 190, 191] to identify novel inhibitors for three major reasons: i) kinases are among the most relevant target families for the pharmaceutical industry, ii) a wide array of high-resolution protein-ligand X-ray structures are available for validation purpose, iii) a canonical H-bonding to the so-called 'hinge region' of the kinase is a typical hallmark of ATP-competitive inhibitors. Two recent studies [129, 191] are representative of the results which might be expected for kinase inhibitors. Vangrevelinghe *et al.* reported a knowledge-based virtual screening protocol for identifying casein kinase II (CK2) inhibitors, in which post-docking filters were designed to downsize the hitlist [191]. Starting from ca. 400,000 compounds which were docked using Dock4.01 to the 3-D structure of human CK2, 12,000 molecules were first retrieved by score. This primary hitlist was then reduced to 1,592 molecules by selecting only hits which were H-bonded to the 'hinge segment' of the protein and well scored by a consensus scoring function. Visual check of the remaining hits afforded a hit list of only 12 compounds out of which three molecules inhibited the enzyme with an IC_{50} lower than $10\,\mu\mathrm{M}$.

Pre-docking filters may be useful as well in selecting the most interesting compounds by similarity to known chemotypes present in kinase inhibitors. A good illustration of this strategy has recently been reported by Lyne *et al.* in the discovery of checkpoint kinase-1

Table 18.1: Successful structure-based screening data from the recent literature (2003–2005).

Target	Docking	Library	Size	Hit rate[a]	Ref.
Chk-1 kinase	FlexX	AstraZeneca	550,000	36%@ 68 μM	[197]
Casein kinase II	Dock	Novartis	450,000	33%@ 10 μM	[191]
BCR-ABL	Dock	Chemdiv	200,000	13%@ 30 μM	[158]
p56 Lck	Dock	n.a.[b]	2,000,000	17%@100 μM	[83]
EphB 2	Gold	Chemdiv	50,452	5%@ 10 μM	[190]
Protein Kinase B	FlexX	Chembridge	50,000	10%@ 20 μM	[59]
DHFR (*S. aureus)*	FlexX	Roche	9,448	21%@ 25 μM	[215]
DHFR	Dock	ACD[c]	n.a.	33%@ 20 μM	[167]
Aldose reductase	FlexX	ACD	260,000	55%@ 20 μM	[106]
XIAP	Dock	TCM[d]	8,000	3%@ 5μM	[148]
Stat3β	Dock	4 collections	429,000	1%@ 20 μM	[183]
Ribosomal A-site	RiboDock	Vernalis collection	1,000,000	26%@500 μM	[58]
IMPDH	FlexX	Roche	3,425	8%@100 μM	[162]
L-xylose reductase	Dock	NCI[e]	249,071	5%@100 μM	[26]
PDE4D	FlexX	Combinatorial library	320	55%@100 nM	[109]
Thymidine phosphorylase	Dock	NCI	250,000	7%@ 20 μM	[138]
t-RNA guanine transglycosylase	FlexX	7 collections	827,000	55%@ 10 μM	[22]
P450 2D6	Gold	NCI subset	111	39%@ 10 μM	[95]
SHBG	Glide	Natural compounds	23,836	7%@ 25 μM	[29]
TMPKmt	FlexX	CMC[f] + KEGG[g]	7,986	10%@ 20 μM	[45]
AICAR transformylase	AutoDock	NCI	1,990	51%@ 20 μM	[127]
5-HT_{1A} receptor	Dock	$>$ 20 suppliers	1,600,000	21%@ 5 μM	[13]
NK_1 receptor	Dock	$>$ 20 suppliers	1,600,000	15%@ 5 μM	[13]
D_2 receptor	Dock	$>$ 20 suppliers	1,600,000	17%@ 5 μM	[13]
CCR_3 receptor	Dock	$>$ 20 suppliers	1,600,000	12%@ 5 μM	[13]
5-HT_4 receptor	Dock	$>$ 20 suppliers	1,600,000	21%@ 5 μM	[13]
α_{1a} receptor	Gold	Aventis	n.a.	30%@ 1 μM	[54]
NK_1 receptor	FlexX	7 collections	827,000	14%@ 1 μM	[53]
D_3 receptor	LigandFit	NCI	250,000	40%@ 1 μM	[192]

[a] Hit rate at a concentration threshold. The hit rate is the ratio of the number of active compounds to the total number of compounds tested.

[b] not available

[c] Available Chemicals Directory (http://www.mdli.com/products/experiment/available_chem_dir/)

[d] Traditional Chinese Medicine Database (http://www.tcm3d.com/)

[e] National Cancer Institute (http://129.43.27.140/ncidb2/)

[f] Comprehensive Medicinal Chemistry Database (http://www.mdli.com/products/knowledge/medicinal_chem)

[g] KEGG database (http://www.genome.jp/kegg/ligand.html)

inhibitors [129]. A hierarchical screening protocol involving filters of increasing complexity (simple molecular descriptors, 3-D pharmacophore search, FlexX-Pharm constrained docking, knowledge-based consensus scoring) decreases the number of virtual hits from 400,000 to 103, and allowed to identify novel inhibitors in four chemical series. Interestingly, most true inhibitors were not recovered among the top-ranked poses but by rescoring at least the top 50 poses by a consensus scoring protocol designed from a surrogate kinase (Cdk-2) and a test dataset. Post-docking filtering by similarity to well-defined intermolecular interactions may also be a reliable option as it was recently shown to outperform consensus scoring in identifying protein kinase B inhibitors [59]. In the above-cited cases, a precise knowledge-based selection of the most reliable compounds has been achieved thanks to the large information available for related compounds.

The same remark applies to three recent studies aimed at discovering inhibitors of two reductases (dihydrofolate reductase, aldose reductase) [106, 167, 215], extensively studied in the past. Wyss *et al.* [215] docked a library of 2,4-diaminopyrimidines to the X-ray structure of DHFR from *S.aureus* complexed with an in-house inhibitor. 252 out of the 300 top-ranked compounds could be synthesized and tested for DHFR inhibition. 21 % of the proposed compounds inhibited DHFR from either *S. aureus* or *S. pneumoniae* with IC_{50} values lower 10 μM. Remarkably, a structure-based screening protocol was found to be much superior to a ligand-based diversity selection in enriching a hit list in true inhibitors.

Rastelli *et al.* [167] screened a subset of the ACD for inhibitors of the DHFR from *P. falciparum* which would be insensitive to specific active site mutations. The full dataset was first filtered by Catalyst (Accelrys Software Inc., http://www.accelrys.com/products/catalyst/) to retrieve molecules satisfying a set of 3-D pharmacophores generated from known protein-inhibitors X-ray structures and potentially able to bind to some enzyme mutants. Docking the focussed dataset using Dock, then selecting the top-ranked molecules interacting with a key residue and clustering by chemotypes afforded a final list of 24 molecules. 12 compounds truly inhibited DHFR wild type as well as active site mutants at micromolar concentrations.

Kraemer *et al.* [106] identified, from the ACD, aldose reductase inhibitors by a series of hierarchical filters implying substructure similarity search to known inhibitors, 2-D pharmacophore filtering, FlexX docking and DrugScore scoring. Compounds able to bind to the anionic pocket of the enzyme were prioritized for purchase and experimental evaluation. Out of the nine compounds tested, six exhibited micromolar inhibition of the target. Interestingly, DrugScore values were weighted according to the molecular weight and number of rotatable bonds of the corresponding molecules to favor the selection of lead-like compounds.

First-in-class compounds

Not all targets are suited for experimental high-throughput screening. However, if 3-D coordinates are available, VS is still a cheap alternative to HTS. Two recent studies [148, 183] demonstrate the power of VS for quasi-orphan targets (XIAP, Stat3) of interest for discovering new antitumoral drugs. The X-ray structure of XIAP complexed to a peptidic inhibitor was used to identify, within a database of 8,000 compounds derived from traditional Chinese medicinal herbs, a nonpeptidic micromolar XIAP inhibitor [148]. Likewise, 429,000 compounds from various screening collections were docked to the X-ray structure of Stat3, a signal transducer and activator of transcription. Rescoring the top 10 % scored compounds

from each dataset with X-score [202] yielded 200 compounds out of which 100 could be purchased and tested for Stat3 inhibition [183]. As in the previous study, obtained hit rates at micromolar concentrations were rather low (a single hit out of 100 compounds tested) but a totally novel compound could be discovered and used as a basis for further improvement.

Nucleic acids have not been widely investigated in structure-based screening approaches mainly because of the lack of accurate scoring functions. Foloppe *et al.* [58] recently reported the successful discovery of bacterial ribosomal A-site ligands by using a docking tool (RiboDock) specifically designed for that purpose [142]. An electronic catalogue of 1 million commercially-available compounds was first filtered to select lead-like compounds and father docked to the crystal structure of the *E. coli* ribosomal A-site. Visual inspection of the top 2,000 best scored compounds yielded a list of 129 molecules which were evaluated by a FRET binding assay. Five compounds, unrelated to the aminoglycoside series, exhibited an apparent inhibition constant lower than $50\ \mu$M. This study is promising by widening the scope of application of high-throughput docking to non-protein targets and more successful applications are expected in a near future thanks to a better parameterization of common docking tools for predicting ligand binding to nucleic acids [42].

Fragment screening

Fragment screening by X-ray or NMR [197] is becoming an increasingly popular method for identifying low molecular weight leads which usually shows a greater optimization potential than drug-like compounds [80]. Because of the difficulty to correctly rank docking poses of small fragments, computational screening of low-molecular weight compounds is still in its infancy. Two recent reports [26, 162] indicate however that this approach might be promising.

Pickett *et al.* reported the discovery of low-molecular inhibitors of inosine 5'-monophosphate dehydrogenase (IMPDH) by virtual needle screening [162]. A test set of 21 true IMPDH inhibitors and two in-house X-ray structures was first used to select the most adequate docking/scoring combination (FlexX docking/ScreenScore scoring). A corporate database of 3,425 low-molecular weight reagents was then docked to both X-ray structures to retrieve, among top-ranked compounds, 100 virtual hits satisfying a visual check Out of the 74 compounds evaluated for IMPDH inhibition, three molecules exhibited an IC_{50} lower than $35\ \mu$M

Carbone *et al.* [26], although not explicitly looking for fragments, also discover low-moecular weight inhibitors of L-xylose reductase by structure-based screening. Hence, this enzyme is characterized by a very shallow active site and most known XR inhibitors are short chain fatty acids. By screening with the Dock program ca. 240,000 compounds from the NCI dataset (National Cancer Institute, Enhanced NCI Database browser, http://129.43.27.140/ncidb2/) against the X-ray structure of Xylose reductase (XR), a limited number of putative hits (ca. 1,000) could be prioritized by score and known interactions to key catalytic residues. Out of 39 molecules which were purchased and evaluated for XR inhibition, two carboxylic acids (nicotinic acid, benzoic acid) inhibited the target with IC_{50} values under $100\ \mu$M.

Chapter 16 discusses methodical aspects of fragment-based drug design.

Lead optimization

A large majority of structure-based screening projects are aimed at identifying hits. However, lead optimization might be possible at the condition that the binding mode of the starting lead can be unambiguously recovered and that a rationale exists for selecting the next compounds to synthesize and test. Krier *et al.* [109] recently proposed a straightforward approach for exploring a lead series by enumerating small-sized libraries (a few hundred compounds) in which all combinatorial assemblies of a few linkers and pharmacophoric moieties to a given scaffold are probed. The selection of the best analogues was based on FlexX docking to the X-ray structure of the phosphodiesterase target and topological filtering. A single-round screening campaign on nine synthesized analogues yielded to a subnanomolar inhibitor and a 900-fold improvement in affinity over the starting lead. Lead optimization is discussed in detail in Chapter 19.

Homology models as virtual screening targets

All above-reported applications have used high-resolution X-ray structures to represent the 3-D coordinates of the target under study. However, enzymes for which a crystal or an NMR structure are still missing but which shows enough sequence homology (ca. 50 %) in the active site to a X-ray template, can also be used for database docking approaches with reasonable success [138]. However, there is still a debate whether targets ranging in a much lower homology range ($< 30\,\%$) might be reliable starting points. This observation is particularly relevant for G-protein coupled receptors (GPCRs) a target family of outmost pharmaceutical interest for which a single X-ray structure (bovine rhodopsin) might be used for comparative modelling. Several recent reports [13, 16, 53, 54, 192] demonstrated that GPCRs might be suitable indeed for structure-based screening. In all above-cited successful cases, preliminary knowledge about known ligands was necessary to fine tune the receptor model. Moreover, the choice of a relevant pharmacophore hypothesis was a key factor to downsize the number of molecules for docking. Last, a visual inspection was necessary to ensure that key intermolecular interactions were established with selected hits. Although the derived homology models remain crude with respect to high-resolution X-ray structures, drug-like sub-micromolar antagonists for rhodopsin-like receptors [13, 16, 53, 54, 192] have already been discovered by VS.

18.5.3 Concluding remarks

Virtual screening of compound libraries by high-throughput docking is nowadays a routinely-used computational technique for identifying bioactive ligands with numerous proofs of record. One should however keep in mind that the method is highly sensitive to the 3-D coordinates of the target and is likely to generate numerous false positives. As important as the docking itself are the pre- and post-processing steps which are key factors to optimize the hit rate. The number of new validated chemotypes amenable to optimization is therefore a better descriptor than the simple hit rate which considerably vary with regard to the current knowledge on a particular target. VS is a natural complement to traditional medicinal chemistry and particularly well suited for proposing new molecular scaffolds that can be easily converted

into focussed ligand libraries of higher values. Both methodological improvements (scoring, hit triage, prediction of ADMET properties) and better screening collections (focussed and targeted libraries) should contribute to improve the value of this powerful tool in a near future.

18.6 Critical evaluation of ligand-based virtual screening

The choice of tools for ligand-based virtual screening is at least as complicated as for structure-based techniques. While for structure-based techniques mainly the 'right' docking program has to be chosen, for ligand-based VS also the method itself, e.g., similarity search or pharmacophore search, has to be chosen. The first part of this Chapter will evaluate and compare several methods and give guidance for selecting appropriate methods and/or tools for the screening. The second part will then — as in Section 18.5 — review some recent success stories and, finally, will draw some conclusions on which methods to apply.

18.6.1 Influence of parameter settings

For ligand-based VS the considerations about the choice of the library and its preprocessing are identical to those for structure-based screening (see Section 18.5). The main selection process is then, which method among those of Section 18.3 is chosen. This depends mainly on the number of known active ligands available. If at least some (more than 5, better more than 20) ligands are known, a ligand-based pharmacophore model can be derived. If additionally their activity is known, QSAR techniques are possible. If, however, less actives are known, only similarity searches can be performed at this stage.

18.6.2 Recent success stories

Here, we review recent studies from the literature (2003 – 2005) covering the entire field of ligand-based techniques (Table 18.2), ranging from pharmacophore searching over similarity searching to QSAR studies.

The studies of Laggner *et al.* [117] and Peukert *et al.* [161] demonstrate the application of ligand-based pharmacophore models in VS. Laggner *et al.* [117] built pharmacophore models for ERG2, the emopamil binding protein (EBP), and the σ_1 receptor using Catalyst. The training set comprised 23 structurally diverse ligands with a broad activity range from picomolar to micromolar affinity. The pharmacophore models were assessed using cost analysis and randomization tests. Furthermore, on a testset of 9 molecules with binding affinities from sub-nanomolar to micromolar, from 26 measured affinities, 14 were predicted within 1 order of magnitude. The pharmacophore models were then used for VS of the WDI. From the WDI previously known binders were found as expected but also a number of new hits. Among them, 11 were experimentally tested and hitrates between 55% and 75% were obtained for the three targets. Subsequently, the pharmacophore models were altered to perform a search in a subset of the KEGG database of 3,525 metabolites. Peukert *et al.* [161] described the discovery of novel blockers of the Kv1.5 potassium ion channel based on pharmacophore search. The authors used DISCO for pharmacophore elucidation using a training set of 7 known Kv1.5 blockers. The pharmacophore model obtained was consistent with published

Table 18.2: Successful ligand-based screening data from the recent literature (2003–2005).

Target	Method	Library	Size	Hit rate[a]	Ref.
ERG2	Pharmacophore	WDI[b]	48,405	55%@ 1 μM	[117]
σ_1 receptor	Pharmacophore	WDI	48,405	63%@ 1 μM	[117]
Emopamil binding protein	Pharmacophore	WDI	48,405	73%@ 1 μM	[117]
Kv1.5	Pharmacophore	Aventis	n.a.[c]	6%@ 6 μM	[161]
A_{2A} purinergic receptor	Pharmacophore similarity	Combinatorial library	192	53%@ 10 nM	[179]
mGluR5	Pharmacophore similarity	Asinex	194,563	33%@ 70 μM	[168]
Tat-TAR RNA interaction	Pharmacophore similarity	SPECS library	229,659	11%@500 μM	[169]
H_1 receptor	QSAR	Combinatorial library	9,000	87% Watanabe[d]	[48]
Trichomonas vaginalis	QSAR	in-house	100	25% *in vitro*[e]	[140]
Kv1.5	Similarity	Aventis	n.a.	1%@ 10 μM	[160]
MCH-1R	Similarity[f]	24 collections	650,000	2%@ 30 μM	[30]
D_3 receptor	Pharmacophore fingerprints	2 collections	255,286	40%@100 nM	[25]
COX-2	Pharmacophore fingerprints	Commercial collections	2,700,000	15%@ 10 μM	[60]

[a] Hit rate at a concentration threshold. The hit rate is the ratio of the number of active compounds to the total number of compounds tested.
[b] World Drug Index (http://scientific.thomson.com/products/wdi/)
[c] not available
[d] Antihistaminic activity according to the protocol of Watanabe *et al.* [207]
[e] Cytocidal activity of 100% after 48 h at a concentration of 100 μg/ml.
[f] A combination of 2D and 3D substructure search, 2D and 3D similarity, as well as clustering was used.

SAR data and was able to retrieve 58% of a testset of 423 known Kv1.5 blockers. A 3D search was performed on the Aventis compound collection resulting in 1,975 hits after filtering. In a subsequent clustering 27 clusters were obtained and representatives of 18 clusters were screened *in vitro*. One active compound was found with an IC_{50} of $5.6\,\mu$M belonging to a new class exhibiting a favorable pharmacokinetic profile.

Schneider and Nettekoven [179] demonstrated the use of a topological pharmacophore similarity model named CATS [178]. This approach was applied to the prediction of selective purinergic receptor (A_{2A}) antagonists from a virtual combinatorial library. From a preliminary SAR model an artificial neural network (self-organizing map, SOM) was trained. Molecules were encoded by the CATS descriptor and the features were mapped from 150-dimensional space onto the plane of a SOM. Each field of the SOM has thus certain pharmacophore features in common. With this technique, the library was reduced from 192 to 17 combinatorial

products. These 17 molecules exhibit 3-fold higher binding affinities and 3.5-fold higher selectivities than the initial library. The most selective antagonist displays 121-fold selectivity and an affinity of 2.4 nM. The CATS3D descriptor, a 3D extension of the CATS approach, was used by Renner *et al.* [168] to identify metabotropic glutamate receptor 5 (mGluR5) modulators. From the original library of 194,563 molecules, first, the 20,000 most 'drug-like' compounds were selected and screened by similarity of the CATS3D vectors with each of 7 active molecules. Of the obtained 27 top-scoring molecules 9 exhibited an activity below $70\,\mu$M. The authors validate, that the method used allows for pharmacophore-based similarity searching with 'scaffold-hopping'. This descriptor was also reported to be successful for identification of new inhibitors of the Tat-TAR RNA interaction [169]. In addition, also a 'fuzzy' pharmacophore approach (SQUID) was used. Again, the 20,000 most 'drug-like' compounds of an initial library of 229,658 compounds were screened. In the VS the similarities were calculated by the Manhattan-distance for the CATS3D and a similarity score for the SQUID, respectively. Both techniques revealed 10 hits, with one molecule overlap. Two molecules among them had IC_{50} values of $500\,\mu$M and $46\,\mu$M, respectively.

A screening for antihistaminic compounds blocking the H_1 receptor was performed by Duart *et al.* [48] using a QSAR model based on molecular topology descriptors. From the initial virtual library of 9,000 compounds, 236 molecules were predicted as active. Of the selected 7 most promising compounds, experimental testing exhibited antihistaminic activity in 87%. The discovery of trichomonacidal compounds was reported by Meneses-Marcel *et al.* [140]. A linear discrimination analysis (LDA) QSAR model was trained to classify molecules using atom-based quadratic indices as descriptors. Since validation of the model revealed 88% good classification, a virtual screening was performed. Biological assays of 8 compounds selected by screening gave good classification. Two molecules maintained their efficacy against *Trichosomas vaginalis* even at $10\,\mu\mathrm{g/ml}$ and one of them did not show cytotoxic effects in macrophage cultivations.

A 2-D similarity search with Unity was performed by Peukert *et al.* [160] for blockers of the Kv1.5 ion channel. Using a compound with an IC_{50} of $0.1\,\mu$M as reference molecule, 75 compounds with a similarity value of ≥ 0.8 were found in the Aventis compound library and experimentally tested. In this step a moderately active compound (IC_{50} of $9.5\,\mu$M) was discovered. Although, this compound was rejected due to problems with its stability and properties, a compound with similar side chains but a different scaffold (naphthene spacer replaced by a biphenyl group) was identified as lead ($IC_{50} = 4.8\,\mu$M).

Clark *et al.* performed substructure and similarity searches, both in two and three dimensions, among other techniques, for discovering MCH-1R antagonists. As query compounds 11 known MCH-1R antagonists were selected. The combined hits from all searches were selected (3,015 molecules) and assessed for drug-likeness, synthetic tractability, and molecular properties. After duplicate removal, 1,490 compounds remained which were clustered using Daylight fingerprints. After final visual inspection 795 compounds were purchased and biochemically screened, resulting in 19 compounds with IC_{50} values below $1\,\mu$M and the best having an IC_{50} of 50 nM. Clark *et al.* analyzed, which of the searches revealed which of the 19 compounds. Six compounds were found by 3D similarity search with FlexS only, also six were found by 3D substructure search only, two were found by a clustering approach only, and one was revealed only by 2D similarity search. Just four compounds were discovered by more than a single technique. The hit rates were in the range of 0.0% (2D substructure) to

5.6% (2D similarity).

18.6.3 Comparison of structure-based and ligand-based techniques

Recently groups at Roche [14], at Aventis [52], and at Argenta Discovery [30] compared structure-based and ligand-based techniques for virtual screening for GPCR targets. Bissantz *et al.* [14] performed a comparative evaluation of the techniques for searching 5-HT_{2C} agonists, while Evers *et al.* [52] performed the comparison on four different biogenic amine-binding GPCRs (α1A, 5-HT_{2A}, D2, and M1 receptors) and Clark *et al.* [30] used a number of ligand-based techniques (see above) and compared them to structure-directed pharmacophores.

In the work of Bissantz *et al.* the results of docking into homology models with FRED were compared to results from Daylight fingerprints, FeatureTrees, and the program Phacir. The performance was assessed by hit rate, enrichment factor, and the diversity of the structures retrieved. Test database was a collection of actives and inactives from the Roche compound depository, with high similarity between actives and inactives. Four molecules were used as reference for the three similarity search programs (so in total 12 similarity searches were performed) and the top 20% of the ranking lists were analyzed. Each of the 207 actives was retrieved with at least one of methods by combining the results for each of the reference molecules. When looking at the 12 screening runs, in each individual search many compounds were not retrieved, or even worse, not a single compound with some of the scaffolds was found. Furthermore, the results show that the success of the methods depends strongly on the choice of the reference ligand. While all three similarity measures obtained hit rates of at least 4.8% and enrichment factors of ≥ 2.3 for one of the ligands, for another reference ligand the best hit rate was only 2.8%. Some combinations of method and reference ligand did not perform better than random selection. For comparison the docking program FRED was applied using different scoring functions. The hit rate was between 3.0% and 4.5% and the enrichment factor between 1.5 and 2.2. Thus, while the top-performing ligand-based techniques reached better hit rates than docking, docking always performed better than half of the ligand-based screening runs. Furthermore, the compounds retrieved by structure-based techniques were more diverse on average than those from ligand-based screening. The authors conclude that the results of structure-based screening are more stable than those of ligand-based screening. The latter can yield higher hit rates, but only for some of the reference ligands. In addition, the actives retrieved by docking were more diverse. Based on these results, the authors propose to combine at least one similarity search with a docking technique.

Evers *et al.* [52] also compared docking into homology models to ligand-based protocols. For the latter, ligand-based pharmacophores, multiple Feature Trees (MTrees) as well as 3D-similarity by FlexS, and QSAR models were applied. Pharmacophore and a MTree models were compared on two different reference ligands (one for each class of ligand molecules) for each GPCR. In this study, ligand-based pharmacophore, MTrees, and 2D QSAR techniques received higher enrichment factors than docking into the homology model with GOLD and FlexX-Pharm. However, the results with GOLD were still satisfying. The authors conclude that docking into GPCR homology models can be useful if no or only a few active ligands are known. In this study the hit rates obtained with FlexS are worse than those obtained from the other virtual screening techniques applied. This is in contrast to results of other studies (e.g.

Clark *et al.* [30], see above and below) where FlexS gives respectable results. Evers *et al.* conclude that a 'fair' comparison can be made only by using several reference structures for the queries.

Clark *et al.* [30] compared a set of different ligand-based methods (see above) to searches using structure-directed pharmacophores. The pharmacophores were generated by docking one ligand into an homology model, then aligning nine other molecules with GASP on the docked conformation and refining the complexes by simulated annealing. Based on this alignment three different pharmacophore hypotheses were derived and used as queries, but none of them gave rise to a hit.

Ligand-based and structure-based virtual screening has not only been compared for GPCR targets. Another group at Aventis used a number of techniques for the search for Kv1.5 ion channel blockers. Besides the ligand-based work (ligand-based pharmacophore and two-dimensional similarity, see above) also a structure-based screening was performed in which a protein-derived pharmacophore based on a homology model was used [163]. The structure-based VS gave a higher hit rate (7.8%) than the screenings based on ligand-based pharmacophore (5.5%) and similarity searches (2.7%). Furthermore, the structure-based technique yielded more active compounds and more chemotypes. Even more important is the result that there was no overlap between the hit lists obtained by ligand-based and structure-based approaches.

18.6.4 Concluding remarks

From the large number of successful applications of ligand-based virtual screening two general and simple rules can be derived. These rules help to reduce the false negative rate (ligands being active but not found) of the screening.

1. *Use as many query ligands as possible.* Several authors have reported that some ligands perform very poorly not giving any hit at all, while with other reference ligands many hits were found. Unfortunately, it cannot be determined in advance, which of the ligands will be successful.

2. *Use as many different techniques as possible.* While some of the hits are 'easy' to find by many different techniques, often valuable compounds (e.g. unique scaffold) are found only by one of the techniques. Again, it cannot be predicted, which of the techniques will be successful. It is important to note, that not necessarily the most sophisticated techniques yield the most hits. In some cases a very simple search technique can find an interesting compound.

Comparing ligand-based and structure-based techniques is difficult since the effectiveness of ligand-based and structure-based techniques depends strongly on the screening project. For some targets the ligand-based techniques perform better than structure-based methods while for other targets they perform worse. From this finding a third rule can be derived:

3. *Use both ligand-based and structure-based techniques if possible.* In this combined scenario the maximal benefit of the different starting points can be obtained and the best

compromise between the strengths and limitations of the various methods can be obtained. In other words, make use of the complementarity between ligand-based and structure-based techniques [14].

18.7 Acknowledgements

The authors are grateful to Matthias Rarey (Hamburg) and Andreas Steffen (Saarbrücken) for valuable suggestions. A large part of this chapter is based on the course ”Computer-aided Drug Design” taught by A. K. He would like to thank the Max Planck Institute for Informatics and the Center for Bioinformatics of Saarland University for the opportunity to give this course.

References

[1] **Agrafiotis, D. K., V. S. Lobanov, D. N. Rassokhin and S. Izrailev.** 2000. The measurement of molecular diversity. Böhm, H.-J. and Schneider, G. Virtual screening for bioactive molecules Wiley-VCH 265–300.

[2] **Agrafiotis, D. K. and V. S. Lobanov.** 2000. Nonlinear mapping networks. J. Chem. Inf. Comput. Sci. **40**:1356–1362.

[3] **Ahlberg, C.** 1999. Visual exploration of HTS databases: bridging the gap between chemistry and biology. Drug Discov. Today **4**:370–376.

[4] **Ajay, G. W. Bemis and M. A. Murcko.** 1999. Designing Libraries with CNS activity. J. Med. Chem. **42**:4942–4951.

[5] **Ajay, W. P. Walters and M. A. Murcko.** 1998. Can we learn to distinguish between 'drug-like' and 'nondrug-like' molecules? J. Med. Chem. **41**:3314–3324.

[6] **Allen, F. H.** 2002. The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Cryst. **B58**:380–388.

[7] **Bajorath, J.** 2002. Integration of virtual and high-throughput screening. Nat. Rev. Drug Discov. **1**:882–894.

[8] **Barnard, J. M., G. M. Down and P. Willett.** 2000. Descriptor-based similarity measures for screening chemical databases. Böhm, H.-J. and Schneider, G. Virtual screening for bioactive molecules Wiley-VCH 59–80.

[9] **Barnard, J. M. and G. M. Downs.** 1997. Chemical fragment generation and clustering software. J. Chem. Inf. Comput. Sci. **37**:141.

[10] **Baroni, M., G. Costantini, G. Cruciani, D. Riganelli, R. Valigi and S. Clementi.** 1993. Generating optimal linear PLS estimation (GOLPE): An advanced chemometric tool for handling 3D-QSAR problems. Quant. Struct.-Act. Relat. **12**:9–20.

[11] **Baurin, N., R. Baker, C. Richardson, I. Chen, N. Foloppe, A. Potter, A. Jordan, S. Roughley, M. Parrat, P. Greaney, D. Morley and R. E. Hubbard.** 2004. Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. J. Chem. Inf. Comput. Sci. **44**:643–651.

[12] **Bayada, D. M., H. Hamersma and V. J. van Geerestein.** 1999. Molecular diversity and representativity in chemical databases. J. Chem. Inf. Comput. Sci. **39**:1–10.

[13] **Becker, O. M., Y. Marantz, S. Shacham, B. Inbal, A. Heifetz, O. Kalid, S. Bar-Haim, D. Warshaviak, M. Fichman and S. Noiman.** 2004. G protein-coupled receptors: In silico drug discovery in 3D. Proc. Natl. Acad. Sci. U.S.A. **101**:11304–11309.

[14] **Bissantz, C., C. Schalon, W. Guba and M. Stahl.** 2005. Focused library design in GPCR projects on the example of 5-HT2c agonists: Comparison of structure-based virtual screening with ligand-based methods. Proteins Struct, Func, Bioinf. **61**:938–952.

[15] **Bissantz, C., G. Folkers and D. Rognan.** 2000. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J. Med. Chem. **43**:4759–4767.

[16] **Bissantz, C., P. Bernard, M. Hibert and D. Rognan.** 2003. Protein-Based Virtual Screening of Chemical Databases. II. Are Homology Models of G-Protein Coupled Receptors Suitable Targets? Proteins Struct. Func. Genet. **50**:5–25.

[17] **Blaney, F. E., P. Finn, R. W. Phippen and M. Wyatt.** 1993. Molecular surface comparison: application to molecular design. J. Mol. Graph. **11**:98–105.

[18] **Bohacek, R. S., C. McMartin and W. C. Guida.** 1996. The art and practive of structure-based drug design: A molecular modeling perspective. Med. Res. Rev. **1**:3–50.

[19] **Boström, J., J. R. Greenwood and J. Gottfries.** 2003. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. J. Mol. Graph. Model. **21**:449–462.

[20] **Boström, J.** 2001. Reproducing the conformations of protein-bound ligands: a critical evaluation of several popular conformational searching tools. J. Comput. Aided Mol. Des. **15**:1137–1152.

[21] **Bravi, G., E. Gancia, P. Mascagni, M. Pegna, R. Todeschini and A. Zaliani.** 1997. MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: a comparative 3D QSAR study in a series of steroids. J. Comput.-Aided Mol. Des. **11**:79–92.

[22] **Brenk, R., L. Naerum, U. Grädler, H.-D. Gerber, G. A. Garcia, K. Reuter, M. T. Stubbs and G. Klebe.** 2003. Virtual Screening fur Submicromolar Leads of tRNA-guanine Transglycosylase Based on a New Unexpected Binding Mode Detected by Crystal Structure Analysis. J. Med. Chem. **46**:1133–1143.

[23] **Briem, H. and J. Günther.** 2005. Classifying 'kinase inhibitor-likeness' by using machine-learning methods. ChemBioChem **6**:558–566.

[24] **Brown, R. D. and Y. C. Martin.** 1996. Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. J. Chem. Inf. Comput. Sci. **36**:572–584.

[25] **Byvatov, E., B. C. Sasse, H. Stark and G. Schneider.** 2005. From virtual to real screening form D_3 dopamine receptor ligands. ChemBioChem **6**:997–999.

[26] **Carbone, V., S. Ishikura, A. Hara and O. El-Kabbani.** 2005. Structure-based discovery of human L-xylulose reductase inhibitors from database screening and molecular docking. Bioorg. Med. Chem. **13**:301–312.

[27] **Carhart, R. E., D. H. Smith and R. Venkataraghavan.** 1985. Atom pairs as molecular features in structure-activity studies: definition and applications. J. Chem. Inf. Comput. Sci. **25**:64–73.

[28] **Charifson, P. S., J. J. Corkery, M. A. Murcko and W. P. Walters.** 1999. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. J. Med. Chem. **42**:5100–5109.

[29] **Cherkasov, A., Z. Shi, M. Fallahi and G. L. Hammond.** 2005. Successful in Silico Discovery of Novel Nonsteroidal Ligands for Human Sex Hormone Binding Globulin. J. Med. Chem. **48**:3203–3213.

[30] **Clark, D. E., C. Higgs, S. P. Wren, H. J. Dyke, M. Wong, D. Norman, P. M. Lockey and A. G. Roach.** 2004. A virtual screening approach to finding novel and potent antagonists at the melanin-concentrating hormone 1 receptor. J. Med. Chem. **47**:3962–3971.

[31] **Clark, D. E.** 1999. Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption. J. Pharm. Sci. **88**:807–814.

[32] **Clark, R. D.** 1997. OptiSim: an extended dissimilarity selection method for finding diverse representative subsets. J. Chem. Inf. Comput. Sci. **37**:1181–1188.

[33] **Clement, O. O. and A. T. Mehl.** 2000. HipHop: pharmacophores based on multiple common-feature alignments. Güner, O. F. Pharmacophore perception, development, and use in drug design International University Line 69–84.

[34] **Cramer, R. D., I., D. E. Patterson and J. D. Bunce.** 1988. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. J. Am. Chem. Soc. **110**:5959–5967.

[35] **Cruciani, G., M. Pastor and R. Mannhold.** 2002. Suitability of molecular descriptors for database mining. A comparative analysis. J. Med. Chem. **45**:2685–2694.

[36] **Cummings, M. D., R. L. DesJarlais, A. C. Gibbs, M. Venkatraman and E. P. Jaeger.** 2005. Comparison of automated docking programs as virtual screening tools. J. Med. Chem. **48**:962–976.

[37] **Cummins, D. J., C. W. Andrews, J. A. Bentley and M. Cory.** 1996. Molecular diversity in chemical databases: comparison of medicinal chemistry knowledge bases and databases of commercially available compounds. J. Chem. Inf. Comput. Sci. **36**:750–763.

[38] **Davies, K.** 1996. Using pharmacophore diversity to select molecules to test from commercial catalogues. Chaiken, I. M. and Janda, K. D. Molecular Diversity and Combinatorial Chemistry: Libraries and Drug Discovery American Chemical Society 309–316.

[39] **Dean, P. M. and R. A. Lewis.** 1999. Molecular diversity in drug design. Kluwer Amsterdam.

[40] **Dean, P. M.** 1994. Molecular similarity in drug design. Chapman and Hall Glasgow.

[41] **Denk, Z., C. Chuaqui and J. Singh.** 2004. Structural Interaction Fingerprint (SIFt): A novel method for analyzing three-dimensional protein-ligand binding interactions. J. Med. Chem. **47**:337–344.

[42] **Detering, C. and G. Varani.** 2004. Validation of Automated Docking Programs for Docking and Database Screening against RNA Drug Targets. J. Med. Chem. **47**:4188–4201.

[43] **Diller, D. J. and J. Merz, Kenneth M..** 2001. High throughput docking for library design and library prioritization. Proteins Struct. Func. Genet. **43**:113–124.

[44] **DiMasi, J. A., R. W. Hansen and H. G. Grabowski.** 2003. The price of innovation: new estimates of drug development costs. J. Health Econ. **22**:151–185.

[45] **Douguet, D., H. Munier-Lehmann, G. Labesse and S. Pochet.** 2005. LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design. J. Med. Chem. **48**:2457–2468.

[46] **Downs, G. M. and P. Willett.** 1994. Clustering of chemical structure databases for compound selection. van de Waterbeemd, H. Advanced Computer-Assisted Techniques in Drug Discovery VCH 111–130.

[47] **Drews, J.** 2000. Drug discovery today – and tomorrow. Drug Discov. Today **5**:2–4.

[48] **Duart, M. J., G. M. Antón-Fos, P. A. Alemán, J. B. Gay-Roig, M. E. González-Rosende, J. Gálvez and R. Garcia-Domenech.** 2005. New potential antihistaminic compounds. Virtual combinatorial chemistry, computational screening, real synthesis, and pharmacological evaluation. J. Med. Chem. **48**:1260–1264.

[49] **Engels, M. F. M., T. Thielemans, D. Verbinnen, J. P. Tollenaere and R. Verbeeck.** 2000. CerBeruS: a system supporting the sequential screening process. J. Chem. Inf. Comput. Sci. **40**:241–245.

[50] **Ertl, P., B. Rohde and P. Selzer.** 2000. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. J. Med. Chem. **43**:3714–3717.

[51] **Evensen, E., J. E. Eksterowicz, R. V. Stanton, C. Oshiro, P. D. J. Grootenhuis and E. K. Bradley.** 2003. Comparing performance of computational tools for combinatorial library design. J. Med. Chem. **46**:5125–5128.

[52] **Evers, A., G. Hessler, H. Matter and T. Klabunde.** 2005. Virtual screening of biogenic amine-binding G-protein coupled receptors: Comparative evaluation of protein- and ligand-based virtural screening protocols. J. Med. Chem. **48**:5448–5465.

[53] **Evers, A. and G. Klebe.** 2004. Successful Virtual Screening for a Submicromolar Antagonist of the Neurokinin-1 Receptor Based on a Ligand-Supported Homology Model. J. Med. Chem. **47**:5381–5392.

[54] **Evers, A. and T. Klabunde.** 2005. Structure-based Drug Discovery Using GPCR Homology Modeling: Successful Virtual Screening for Antagonists of the Alpha1A Adrenergic Receptor. J. Med. Chem. **48**:1088–1097.

[55] **Feher, M. and J. M. Schmidt.** 2003. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemstry. J. Chem. Inf. Comput. Sci. **43**:218–227.

[56] **Ferrara, P., H. Gohlke, D. J. Price, G. Klebe and I. Brooks, Charles L..** 2004. Assessing scoring funcitons for protein-ligand interactions. J. Med. Chem. **47**:3032–3047.

[57] **Fink, T., H. Bruggesser and J.-L. Reymond.** 2005. Virtual exploration of the small-molecule chemical universe below 160 daltons. Angew. Chem. Int. Ed. **44**:1504–1508.

[58] **Foloppe, N., I.-J. Chen, B. Davis, A. Hold, D. Morley and R. Howes.** 2004. A structure-based strategy to identify new molecular scaffolds targeting the bacterial ribosomal A-site. Bioorg. Med. Chem. **12**:935–947.

[59] **Forino, M., D. Jung, J. B. Easton, P. J. Houghton and M. Pellechia.** 2005. Virtual Docking Approaches to Protein Kinase B Inhibition. J. Med. Chem. **48**:2278–2281.

[60] **Franke, L., E. Byvatov, O. Werz, D. Steinhilber, P. Schneider and G. Schneider.** 2005. Extraction and visualization of potential pharmacophore points using support vector machines: Application to ligand-based virtual screening for COX-2 inhibitors. J. Med. Chem. **48**:6997–7004.

[61] **Franke, R. and A. Gruska.** 1995. Principal Component and Factor Analysis. van de Waterbeemd, Han Chemometric Methods in Molecular Design VCH 113–163.

[62] **Gastreich, M., H. Briem, C. Lemmen and M. Rarey.** Addressing the virtual screening challenge - The Flex* approach. Alvarez, J. and Shoichet, B. Virtual screening in drug discovery Decker/CRC Press 2005. 25–46.

[63] **Gedeck, P. and P. Willet.** 2001. Visual and computational analysis of structure-activity relationships in high-throughput screening data. Curr. Op. Chem. Biol. **5**:389–395.

[64] **Ghose, A. K., V. N. Viswanadhan and J. J. Wendoloski.** 1999. A knowledge-based approach in desigining combinatorial or medicinal libraries for drug discovery. 1. A qualitative and quantitative characterization of known drug databases. J. Comb. Chem. **1**:55–68.

[65] **Gillet, V. J., P. Willett and J. Bradshaw.** 1998. Identification of biological activity profiles using substructural analysis and genetic algorithms. J. Chem. Inf. Comput. Sci. **38**:165–179.

[66] **Ginn, C. M. R., P. Willett and J. Bradshaw.** 2000. Combination of molecular similarity measures using data fusion. Persp. Drug Discov. Des. **20**:1–16.

[67] **Giordanetto, F., S. Cotesta, C. Catana, J.-Y. Trosset, A. Vulpetti, P. F. W. Stouten and R. T. Kroemer.** 2004. Novel scoring functions comprising QXP, SASA, and protein side-chain entropy terms. J. Chem. Inf. Comput. Sci. **44**:882–893.

[68] **Good, A. C., E. E. Hodgkin and W. G. Richard.** 1992. Utilization of Gaussian functions for the rapid evaluation of molecular similarity. J. Chem. Inf. Comput. Sci. **32**:188–191.

[69] **Good, A. C., S. R. Krystek and J. S. Mason.** 2000. High-throughput and virtual screening: core lead discovery technologies move towards integration. Drug Discov. Today **12**:S61–S69.

[70] **Good, A. C. and D. L. Cheney.** 2003. Analysis and optimization of structure-based virtual screening protocola (1): exploration of ligand conformational sampling techniques. J. Mol. Graph. Model. **22**:23–30.

[71] **Goodford, P.** 1996. Multivariate characterization of molecules for QSAR analysis. J. Chemometrics **10**:107–117.

[72] **Griffith, R., T. T. T. Luu, J. Garner and P. A. Keller.** 2005. Combining structure-based drug design and pharmacophores. J. Mol. Graph. Model. **23**:439–446.

[73] **Güner, O. F.** 2000. Pharmacophore perception, development and use in drug design. International University Line La Jolla, CA.

[74] **Halperin, I., B. Ma, H. Wolfson and R. Nussinov.** 2002. Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins Struc. Func. Genet. **47**:409–443.

[75] **Hann, M., B. Hudson, X. Lewell, R. Lifely, L. Miller and N. Ramsden.** 1999. Strategic pooling of compounds for high-throughput screening. J. Chem. Inf. Comput. Sci. **39**:897–902.

[76] **Hansch, C. and A. Leo.** 1979. Substituent constants for correlation analysis in chemistry. Wiley New York.

[77] **Hansch, C. and T. Fujita.** 1964. ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J. Am. Chem. Soc. **86**:1616–1626.

[78] **Hertzberg, R. P. and A. J. Pope.** 2000. High-throughput screening: new technology for the 21st century. Curr. Op. Chem. Biol. **4**:445–451.

[79] **Hofbauer, C., H. Lohninger and A. Aszódi.** 2004. SURFCOMP: a novel graph-based approach to molecular surface comparison. J. Chem. Inf. Comput. Sci. **44**:837–847.

[80] **Hopkins, A. L., C. R. Groom and A. Alex.** 2004. Ligand efficiency: a useful metric for lead selection. Drug Discov. Today **9**:430–431.

[81] **Hopkins, A. L. and C. R. Groom.** 2002. The druggable genome. Nat. Rev. Drug Discov. **1**:727–730.

[82] **Höskuldsson, A.** 1988. PLS regression methods. J. Chemometrics **2**:211–228.

[83] **Huang, N., A. Nagarsekar, G. Xia, J. Hayashi and A. D. MacKerell, Jr..** 2004. Identification of non-phosphate-containing small molecular weight inhibitors of the tyrosine kinase p56 Lck SH2 domain via in silico screening against the pY + 3 Binding Site. J. Med. Chem. **47**:3502–3511.

[84] **Ihlenfeldt, W.-D. and J. Gasteiger.** 1994. Hash codes for the identification and classification of molecular structure elements. J. Comput. Chem. **15**:793–813.

[85] **Irwin, J. J. and B. K. Shoichet.** 2005. ZINC – a free database of commercially available compounds for virtual screening. J. Chem. Inf. Model. **45**:177–182.

[86] **Itai, A., N. Tomioka, M. Yamada, A. Inoue and Y. Kato.** 1993. Molecular superposition for rational drug design. Kubinyi, H. 3D QSAR in Drug Design: Theory, Methods and Applications ESCOM 173–199.

[87] **Johnson, M. A. and G. M. Maggiora.** 1990. Concepts and applications of molecular similarity. Wiley New York.

[88] **Jones, G., P. Willet and R. C. Glen.** 1995. A genetic algorithm for flexible molecular overlay and pharmacophoere elucidation. J. Comput. Aided Mol. Des. **9**:532–549.

[89] **Kahnberg, P., M. H. Howard, T. Liljefors, M. Nielsen, E. Ø. Nielsen, O. Sterner and I. Pettersson.** 2004. The use of a pharmacophore model for identification of novel ligands for the benzodiazepine binding site of the $GABA_A$ receptor. J. Mol. Graph. Model. **23**:253–261.

[90] **Kalyanaraman, C., K. Bernacki and M. P. Jacobson.** 2005. Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. Biochemistry **44**:2059–2071.

[91] **Kato, Y., A. Inoue, M. Yanada, N. Tomicka and A. Itai.** 1992. Automatic superposition of drug molecules based on their common receptor site. J. Comput.-Aided Mol. Des. **6**:475–486.

[92] **Kearsley, S. K. and G. M. Smith.** 1990. An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlap. Tetrahedron Comput. Methodol. **3**:615–633.

[93] **Kellenberger, E., J. Rodrigo, P. Muller and D. Rognan.** 2004. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins Struct. Func. Bioinf. **57**:225–242.

[94] **Kellogg, G. E., S. F. Semus and D. J. Abraham.** 1991. HINT - A New Method of Empirical Hydrophobic Field Calculation for CoMFA. J. Comput.-Aided Mol. Des. **5**:545–552.

[95] **Kemp, C. A., J. U. Flanagan, A. J. van Eldik, J.-D. Maréchal, C. R. Wolf, C. K. Roberts, Gordon, M. J. I. Paine and M. J. Sutcliffe.** 2004. Validation of Model of Cytochrome P450 2D6: An in Silico Tool for Predicting Metabolism and Inhibition. J. Med. Chem. **47**:5340–5346.

[96] **Kirchmair, J., C. Laggner, G. Wolber and T. Langer.** 2005. Comparative analysis of protein-bound ligand conformations with respect to Catalyst's conformational space subsampling algorithm. J. Chem. Inf. Model. **45**:422–430.

[97] **Kitchen, D. B., H. Decornez, J. R. Furr and J. Bajorath.** 2004. Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Discov. **3**:935–949.

[98] **Klebe, G., T. Mietzner and F. Weber.** 1994. Different approaches toward an automatic structural alignment of drug molecules: applications to sterol mimics, thrombin and thermolysin inhibitors. J. Comput. Aided Mol. Des. **8**:751-78.

[99] **Klebe, G., U. Abraham and T. Mietzner.** 1994. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict their Biological Activity. J. Med. Chem. **37**:4130–4146.

[100] **Klebe, G. and T. Mietzner.** 1994. A fast and efficient method to generate biologically relevant conformations. J. Comput. Aided Mol. Des. **8**:583–606.

[101] **Klebe, G.** 2000. Virtual screening: An alternative or complement to high throughput screening. Kluwer Dordrecht.

[102] **Klon, A. E., M. Gick, M. Thoma, P. Ackling and J. W. Davies.** 2004. Finding more needles in the haystack: a simple and efficient method for improving high-throughput docking results. J. Med. Chem. **47**:2743–2749.

[103] **Klon, A. E., M. Glick and J. W. Davies.** 2004. Appliation of machine learning to improve the results of high-throughput docking against HIV-1 protease. J. Chem. Inf. Comput. Sci. **44**:2216–2224.

[104] **Kontoyianni, M., G. S. Sokol and L. M. McClellan.** 2005. Evaluation of library ranking efficacy in virtual screening. J. Comput. Chem. **26**:11–22.

[105] **Kontoyianni, M., L. M. McClellan and G. S. Sokol.** 2004. Evaluation of docking performance: comparative data on docking algorithms. J. Med. Chem. **47**:558–565.

[106] **Kraemer, O., I. Hazemann, A. D. Podjarny and G. Klebe.** 2004. Virtual Screening for Inhibitors of Human Aldose Reductase. Proteins **55**:814–823.

[107] **Krämer, A., H. W. Horn and J. E. Rice.** 2003. Fast 3D molecular superposition and similarity search in databases of flexible molecules. J. Comput. Aided Mol. Des. **17**:13–38.

[108] **Kramer, B., M. Rarey and T. Lengauer.** 1999. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. Proteins Struct. Func. Genet. **37**:228–241.

[109] **Krier, M., J. X. de Araújo-Júnior, M. Schmitt, J. Duranton, H. Justiano-Basaran, C. Lugnier, J.-J. Bourguignon and D. Rognan.** 2005. Design of Small-Sized Libraries by Combinatorial Assembly of Linkers and Functional Groups to a Given Scaffold: Application to the Structure-Based Optimization of a Phosphodiesterase 4 Inhibitor. J. Med. Chem. **48**:3816–3822.

[110] **Kruskal, J. B.** 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypotheses. Psychometrika **29**:1–27.

[111] **Kubinyi, H., G. Folkers and Y. C. Martin.** 1998. 3D QSAR in Drug Design, Volume 2. Kluwer/ESCOM Dordrecht.

[112] **Kubinyi, H., G. Folkers and Y. C. Martin.** 1998. 3D QSAR in Drug Design, Volume 3. Kluwer/ESCOM Dordrecht.

[113] **Kubinyi, H.** 1993. 3D QSAR in Drug Design: Theory, Methods and Applications. ESCOM Leiden.

[114] **Kubinyi, H.** 1998. Similarity and dissimilarity: a medicinal chemist's view. Persp. Drug Discov. Des. **9–11**:225–252.

[115] **Kuntz, I. D., J. M. Blaney, S. J. Oatley, R. Langridge and T. E. Ferrin.** 1982. A geometric approach to macromolecule-ligand interactions. J. Mol. Biol. **161**:269–288.

[116] **Ladd, B.** 2000. Intuitive data analysis: The next generation. Mod. Drug Discov. **3**:46–52.

[117] **Laggner, C., C. Schieferer, B. Fiechtner, G. Poles, R. D. Hoffman, H. Glossmann, T. Langer and F. F. Moebius.** 2005. Discovery of high-affinity ligands of σ_1 receptor, ERG2, and emopamil binding protein by pharmacophore modeling and virtual screening. J. Med. Chem. **48**:4754–4764.

[118] **Lajiness, M. S.** 1991. Evaluation of the performance of dissimilarity performance methodology. Silipo, C. and Vittoria, A. QSAR: Rational Aproaches to the Design of Bioactive Compounds Elsevier 201–204.

[119] **Leach, A. R.** 1991. A survey of methods for searching the conformational space of small and medium-sized molecules. Lipkowitz, Kenny B. and Boyd, Donald B. Reviews in computational chemistry VCH 1–55.

[120] **Lemmen, C., C. Hiller and T. Lengauer.** 1998. RigFit: a new approach to superimposing ligand molecules. J. Comput. Aided Mol. Des. **12**:491–502.

[121] **Lemmen, C., M. Zimmermann and T. Lengauer.** 2000. Multiple molecular superpositioning as an effective tool for virtual database screening. Persp. Drug Discov. Des. **20**:43–62.

[122] **Lemmen, C., T. Lengauer and G. Klebe.** 1998. FlexS: a method for fast flexible ligand superposition. J. Med. Chem. **41**:4502–4520.

[123] **Lemmen, C. and T. Lengauer.** 2000. Computational methods for the structural alignment of molecules. J. Comput. Aided Mol. Des. **14**:215–232.

[124] **Lessel, U. F. and H. Briem.** 2000. Flexsim-X: a method for the detection of molecules with similar biological activity. J. Chem. Inf. Comput. Sci. **40**:246–253.

[125] **Lewis, R. A., J. S. Mason and I. M. McLay.** 1997. Similarity measures for rational set selection and analysis of combinatorial libraries: the diverse property-derived (DPD) approach. J. Chem. Inf. Comput. Sci. **37**:599–614.

[126] **Lewis, R. A., S. D. Pickett and D. E. Clark.** 2000. Computer-aided molecular diversity analysis and combinatorial library design. Lipkowitz, Kenny B. and Boyd, Donald B. Reviews in Computational Chemistry Wiley-VCH 1–51.

[127] **Li, C., L. Xu, D. W. Wolan, I. A. Wilson and A. J. Olson.** 2004. Virtual Screening of Human 5-Aminoimidazole-4-carboxamide Ribonucleotide Transformylase against the NCI Diversity Set by Use of AutoDock to Identify Novel Nonfolate Inhibitors. J. Med. Chem. **47**:6681–6690.

[128] **Lipinski, C. A., F. Lombardo, B. W. Dominy and P. J. Feeney.** 1997. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Delivery Rev. **23**:3–25.

[129] **Lyne, P. D., P. W. Kenny, D. A. Cosgrove, C. Deng, S. Zabludoff, J. J. Wendoloski and S. Ashwell.** 2004. Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening. J. Med. Chem. **47**:1962–1968.

[130] **Manallack, D. T., W. R. Pitt, E. Gancia, J. G. Montana, D. J. Livingstone, M. G. Ford and D. C. Whitley.** 2002. Selecting screening candidates for kinase and G protein-coupled ceceptor targets using neural networks. J. Chem. Inf. Comput. Sci. **42**:1256–1262.

[131] **Martin, Y. C., J. L. Kofron and L. M. Traphagen.** 2002. Do structurally similar molecules have similar biological activities? J. Med. Chem. **45**:4350–4358.

[132] **Martin, Y. C., M. G. Bures, E. A. Danaher, J. DeLazzer, I. Lico and P. A. Pavlic.** 1993. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. J. Comput.-Aided Mol. Des. **7**:83–102.

[133] **Martin, Y. C.** 1992. 3D database searching in drug design. J. Med. Chem. **35**:2145–2154.

[134] **McGovern, S. L., B. T. Helfand, B. Feng and B. K. Shoichet.** 2003. A specific mechanism of nonspecific inhibition. J. Med. Chem. **46**:4265–4272.

[135] **McGovern, S. L., E. Caselli, N. Grigorieff and B. K. Shoichet.** 2002. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. J. Med. Chem. **45**:1712–1722.

[136] **McGregor, M. J. and S. M. Muskal.** 1999. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. J. Chem. Inf. Comput. Sci. **39**:569–574.

[137] **McMartin, C. and R. S. Bohacek.** 1995. Flexible matching of test ligands to a 3D pharmacophore using a molecular superposition force field: Comparison of predicted and experimental conformations of inhibitors of three enzymes. J. Comput.-Aided Mol. Des. **9**:237–250.

[138] **McNally, V. A., A. Gbaj, K. T. Douglas, L. J. Stratford, M. Jaffar, S. Freeman and R. A. Bryce.** 2003. Identification of a Novel Class of Inhibitor of Human and Escherichia coli Thymidine Phosphorylase by In Silico Screening. Bioorg. Med. Chem. Lett. **13**:3705–3709.

[139] **McQueen, J.** Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability 1967. 281–297.

[140] **Meneses-Marcel, A., Y. Marrero-Ponce, Y. Machado-Tugores, A. Montero-Torres, D. M. Pereira, J. A. Escario, J. J. Nogal-Ruiz, C. Ochoa, V. J. Arán, A. R. Martinez-Fernández and R. N. G. Sánchez.** 2005. A linear discrimination analysis based virtual screening of trichomonacidal lead-like compounds: Outcomes of in silico studies supported by experimental results. Bioorg. Med. Chem. Lett. **15**:3838–3843.

[141] **Merkwirth, C., H. Mauser, T. Schulz-Gasch, O. Roche, M. Stahl and T. Lengauer.** 2004. Ensemble Methods for Classification in Cheminformatics. J. Chem. Inf. Comput. Sci. **44**:1971–1978.

[142] **Morley, S. D. and M. Afshar.** 2004. Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock. J. Comput.-Aided Mol. Des. **18**:189–208.

[143] **Morris, J. J. and P. P. Bruneau.** 2000. Prediction of physicochemical properties. Böhm, H.-J. and Schneider, G. Virtual screening for bioactive molecules Wiley-VCH 33–58.

[144] **Müller, K.-R., G. Rätsch, S. Sonnenburg, S. Mika, M. Grimm and N. Heinrich.** 2005. Classifying 'drug-likeness' with kernel-based learning methods. J. Chem. Inf. Comput. Sci. **45**:249–253.

[145] **Murphy, K. P. and E. Freire.** 1992. Thermodynamics of structural stability and cooperative folding behavior in proteins. Adv. Protein Chem. **43**:313–361.

[146] **Murtaugh, F.** 1983. A survey of recent advances in hierarchical clustering alrorithms. Computer J. **26**:354–359.

[147] **Nicolaou, C. A., S. Y. Tamura, B. P. Kelley, S. I. Bassett and R. F. Nutt.** 2002. Analysis of large screening data sets via adaptively grown phylogenetic-like trees. J. Chem. Inf. Comput. Sci. **42**:1069–1079.

[148] **Nikolovska-Coleska, Z., L. Xu, Z. Hu, Y. Tomita, P. Li, P. P. Roller, R. Wang, X. Fang, R. Guo, M. Zhang, M. E. Lippman, D. Yang and S. Wang.** 2004. Discovery of Embelin as a Cell-Permeable, Small-Molecular Weight Inhibitor of XIAP through Structure-Based Computational Screening of a Traditional Herbal Medicine Three-Dimensional Structure Database. J. Med. Chem. **47**:2430–2440.

[149] **Nissink, J. W. M., C. Murray, M. Hartshorn, J. C. Verdonk, Marcel L.and Cole and R. Taylor.** 2002. A new test set for validating predictions of proteion-ligand interaction. Proteins Struct. Func. Genet. **49**:457–471.

[150] **Oshiro, C., E. K. Bradley, J. Eksterowicz, E. Evensen, M. L. Lamb, J. K. Lanctot, S. Putta, R. Stanton and P. D. J. Grootenhuis.** 2004. Performance of 3D-database molecular docking studies into homology models. J. Med. Chem. **47**:764–767.

[151] **Pang, Y.-P., E. Perola, K. Xu and F. G. Prendergast.** 2001. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. J. Comput. Chem. **22**:1750–1771.

[152] **Patel, Y., V. J. Gillet, G. Bravi and A. R. Leach.** 2002. A comparison of the pharmacophore identification programs: Catalyst, DISCO and GASP. J. Comput. Aided Mol. Des. **16**:653–681.

[153] **Patterson, D. E., R. D. Cramer, A. M. Ferguson, R. D. Clark and L. E. Weinberger.** 1996. Neighborhood behavior: a useful concept for validation of 'molecular diversity' descriptors. J. Med. Chem. **39**:3049–3059.

[154] **Paul, N. and D. Rognan.** 2002. ConsDock: a new program for the consensus analysis of protein-ligand interactions. Proteins Struct. Func. Genet. **47**:521–533.

[155] **Pearlman, R. S. and K. M. Smith.** 1998. Novel software tools for chemical diversity. Persp. Drug Discov. Des. **9–11**:339–353.

[156] **Pearlman, R. S. and K. M. Smith.** 1999. Metric validation and the receptor-relevant subspace concept. J. Chem. Inf. Comput. Sci. **39**:28–35.

[157] **Pearlman, R. S.** 1993. 3D molecular structures: generation and use in 3D-searching. Kubinyi, H. 3D QSAR in drug design: theory, methods and applications ESCOM Science Publishers 21–58.

[158] **Peng, H., N. Huang, J. Qi, P. Xie, C. Xu, J. Wang and C. Yang.** 2003. Identification of novel inhibitors of BCL-ABL tyrosine kinase via virtual screening. Bioorg. Med. Chem. Lett. **13**:3693–3699.

[159] **Perola, E., W. P. Walters and P. S. Charifson.** 2004. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins **56**:235–249.

[160] **Peukert, S., J. Brendel, B. Pirard, A. Brüggemann, P. Below, H.-W. Kleemann, H. Hemmerle and W. Schmidt.** 2003. Identification, synthesis, and activity of novel blockers of the voltage-gated potassium channel Kv1.5. J. Med. Chem. **46**:486–498.

[161] **Peukert, S., J. Brendel, B. Pirard, C. Strübing, H.-W. Kleemann, T. Böhme and H. Hemmerle.** 2004. Pharmacophore-based search, synthesis, and biological evaluation of anthranilic amides as novel blockers of the Kv1.5 channel. Bioorg. Med. Chem. Lett. **14**:2823–2827.

[162] **Pickett, S. D., B. S. Sherborne, T. Wilkinson, J. Bennett, N. Borkakoti, M. Broadhurst, D. Hurst, I. Kilford, M. McKinnell and P. S. Jones.** 2003. Discovery of Novel Low Molecular Weight Inhibitors of IMPDH Via Virtual Needle Screening. Bioorg. Med. Chem. Lett. **13**:1691–1694.

[163] **Pirard, B., J. Brendel and S. Peukert.** 2005. The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches. J. Chem. Inf. Model. **45**:477–485.

[164] **Pitman, M. C., W. K. Huber, H. Horn, A. Krämer, J. E. Rice and W. C. Swope.** 2001. FLASHFLOOD: a 3D field-based similarity search and alignment method for flexible molecules. J. Comput. Aided Mol. Des. **15**:587–612.

[165] **Rarey, M., B. Kramer, T. Lengauer and B. Klebe.** 1996. A fast flexible docking method using an incremental construction algorithm. J. Mol. Biol. **261**:470–489.

[166] **Rarey, M. and J. S. Dixon.** 1998. Feature trees: a new molecular similarity measure based on tree matching. J. Comput. Aided Mol. Des. **12**:471–490.

[167] **Rastelli, G., S. Pacchioni, W. Sirawaraporn, R. Sirawaraporn, M. D. Parenti and A. M. Ferrari.** 2003. Docking and Database Screening Reveal New Classes of Plasmodium falciparum Dihydrofolate Reductase Inhibitors. J. Med. Chem. **46**:2834–2845.

[168] **Renner, S., T. Noeske, C. G. Parsons, P. Schneider, T. Weil and G. Schneider.** 2005. New allosteric modulators of metabotropic glutamate receptor 5 (mGluR5) found by ligand-based virtual screening. ChemBioChem **6**:620–625.

[169] **Renner, S., V. Ludwig, O. Boden, U. Scheffer, M. Göbel and G. Schneider.** 2005. New inhibitors of the Tat-TAR RNA interaction found with a 'fuzzy' pharmacophore model. ChemBioChem **6**:1119–1125.

[170] **Rishton, G. M.** 1997. Reactive compounds and in vitro false positives in HTS. Drug Discov. Today **2**:382–384.

[171] **Roberts, G., G. J. Myatt, W. P. Johnson, K. P. Cross and J. Blower, Paul E..** 2000. LeadScope: software for exploring large sets of screening data. J. Chem. Inf. Comput. Sci. **40**:1302–1314.

[172] **Roche, O., P. Schneider, J. Zuegge, W. Guba, M. Kasny, A. Alanine, F. Bleicher, Konrad andn Danel, E.-M. Gutknecht, M. Rogers-Evans, W. Neidhart, H. Stalder, E. Dillon, Michael andn Sjögren, P. Fotouhi, Nader andn Gillespie, R. Goodnow, P. Harris, William andn Jones, M. Taniguchi, S. Tsujii, W. von der Saal, G. Zimmermann and G. Schneider.** 2002. Development of a virtual screening method for identification of 'frequent hitters' in compound libraries. J. Med. Chem. **45**:137–142.

[173] **Rusinko, Andres, I., M. W. Farmen, C. G. Lambert, P. L. Brown and S. S. Young.** 1999. Analysis of a large structure/biological activity data set using recursive partitioning. J. Chem. Inf. Comput. Sci. **39**:1017–1026.

[174] **Russ, A. P. and S. Lampel.** 2005. The druggable genome: an update. Drug. Discov. Today **10**:1607–1610.

[175] **Sadowski, J. and H. Kubinyi.** 1998. A scoring scheme for discriminating between drugs and nondrugs. J. Med. Chem. **41**:3325–3329.

[176] **Sadowski, J. and J. Gasteiger.** 1993. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. Chem. Rev. **93**:2567–2581.

[177] **Sams-Dodd, F.** 2005. Target-based drug discovery: is something wrong? Drug Discov. Today **10**:139–147.

[178] **Schneider, G., W. Neidhart, T. Giller and G. Schmid.** 1999. 'Scaffold-hopping' by topological pharmacophore search: a contribution to virtual screening. Angew. Chem. Int. Ed. **38**:2894–2896.

[179] **Schneider, G. and M. Nettekoven.** 2003. Ligand-based combinatorial design of selective purinergic receptor A_{2A} antagonists using self-organizing maps. J. Comb. Chem. **5**:233–237.

[180] **Sheridan, R. P., R. Nilakantan, J. S. Dixon and R. Venkataraghavan.** 1986. The ensemble approach to distance geometry: application to the nicitinic pharmacophore. J. Med. Chem. **29**:899–906.

[181] **Sheridan, R. P. and S. K. Kerarley.** 2002. Why do we need so many chemical similarity search methods? Drug Discov. Today **7**:903–911.

[182] **Sirois, S., G. Hatzakis, D. Wei, Q. Du and C. Kuo-Chem.** 2005. Assessment of chemical libraries for their druggability. Comput. Biol. Chem. **29**:55–67.

[183] **Song, H., R. Wang, S. Wang and J. Lin.** 2005. A low-molecular-weight compound discovered through virtual database screening inhibits Stat3 function in breast cancer cells. Proc. Natl. Acad. Sci. U.S.A. **102**:4700–4705.

[184] **Stahl, M. and H.-J. Böhm.** 1998. Development of filter functions for protein-ligand docking. J. Mol. Graph. Model. **16**:121–132.

[185] **Stanton, D. T., T. W. Morris, S. Roychoudhury and C. N. Parker.** 1999. Application of nearest-neighbor and cluster analysis in pharmaceutical lead discovery. J. Chem. Inf. Comput. Sci. **39**:21–27.

[186] **Taylor, R. D., P. J. Jewsbury and J. W. Essex.** 2003. FDS: flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. J. Comput. Chem. **24**:1637–1656.

[187] **Taylor, R.** 1995. Simulation of experimental design strategies for screening random compounds as potential new drugs and agrochemicals. J. Chem. Inf. Comput. Sci. **35**:59–67.

[188] **Todeschini, R., M. Lasagni and E. Marengo.** 1994. New molecular descriptors for 2D and 3D structures. J. Chemometrics **8**:263–272.

[189] **Todeschini, R. and V. Consonni.** 2000. Handbook of Molecular Descriptors. Wiley-VCH Weinheim.

[190] **Toledo-Sherman, L., E. Deretey, J. J. Slon-Uskiewiecz, W. Ng, J.-R. Dai, J. E. Foster, P. R. Redden, D. Uger, Marni, L. C. Liao, A. Pasternak and N. Reid.** 2005. Frontal Affinity Chromatography with MS Detection of EphB2 Tyrosine Kinase Receptor. 2. Identification of Small-Molecule Inhibitors via Coupling with Virtual Screening. J. Med. Chem. **48**:3221–3230.

[191] **Vangrevelinghe, E., K. Zimmermann, J. Schoepfer, R. Portmann, D. Fabbro and P. Furet.** 2003. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. J. Med. Chem. **46**:2656–2662.

[192] **Varady, J., X. Wu, X. Fang, J. Min, Z. Hu, B. Levant and S. Wang.** 2003. Molecular Modeling of the Three-Dimensional Structure of Dopamine 3 (D_3) Subtype Receptor: Discovery of Novel and Potent D_3 Ligands through a Hybrid Pharmacophore- and Structure-Based Database Searching Approach. J. Med. Chem. **46**:4377–4392.

[193] **Veber, D. F., S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple.** 2002. Molecular properties that influence the oral bioavailability of drug candidates. J. Med. Chem. **45**:2615–2623.

[194] **Vedani, A., M. Dobler and M. A. Lill.** 2005. Combining protein modeling and 6D-QSAR - Simulating the binding of structurally diverse ligands to the estrogen receptor. J. Med. Chem. **48**:3700–3703.

[195] **Vedani, A., M. Dobler and P. Zbinden.** 1998. Quasi-atomistic receptor surface models: a bridge between 3-D QSAR and receptor modeling. J. Am. Chem. Soc. **120**:4471–4477.

[196] **Verdonk, M. L., J. C. Cole, M. J. Hartshorn, C. W. Murray and R. D. Taylor.** 2003. Improved protein-ligand docking using GOLD. Proteins Struct. Func. Genet. **52**:609–623.

[197] **Verdonk, M. L., V. Berdini, M. J. Hartshorn, W. T. M. Mooij, C. W. Murray, R. D. Taylor and P. Watson.** 2004. Virtual screening using protein-ligand docking: Avoiding artificial enrichment. J. Chem. Inf. Comput. Sci. **44**:793–806.

[198] **Verkman, A. S.** 2004. Drug discovery in academia. Am. J. Physiol. Cell Physiol. **286**:C465–C474.

[199] **Vigers, G. P. A. and J. P. Rizzi.** 2004. Multiple active site corrections for docking and virtual screening. J. Med. Chem. **47**:80–89.

[200] **Walters, W. P., M. T. Stahl and M. A. Murcko.** 1998. Virtual screening – an overview. Drug Discov. Today **3**:160–178.

[201] **Wang, J., L. Lai and Y. Tang.** 1999. Structural features of toxic chemicals for specific toxicity. J. Chem. Inf. Comput. Sci. **39**:1173–1189.

[202] **Wang, R., L. Lai and S. Wang.** 2002. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. J. Comut.-Aided Mol. Des. **16**:11–26.

[203] **Wang, R., Y. Lu and S. Wang.** 2003. Comparative evaluation of 11 scoring functions for molecular docking. J. Med. Chem. **46**:2287–2303.

[204] **Wang, R., Y. Lu, X. Fang and S. Wang.** 2004. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. J. Chem. Inf. Comput. Sci. **44**:2114–2125.

[205] **Wang, R. and S. Wang.** 2001. How does consensus scoring work for virtual library screening? An idealized computer experiment. J. Chem. Inf. Comput. Sci. **41**:1422–1426.

[206] **Ward, Joe H., J.** 1963. Hierarchical grouping to optimize an objective function. J. Am. Statistical Assoc. **58**:236–244.

[207] **Watanabe, K., H. Nakacawa and S. Tsurufuji.** 1986. A new sensitive fluorometric method for measurement of plasma exudation in the inflammatory skin reaction. J. Pharmacol. Meth. **15**:255–261.

[208] **Weber, A., A. Teckentrup and H. Briem.** 2002. Flexsim-R: a virtual affinity fingerprint descriptor to calculate similarities of functional groups. J. Comput. Aided Mol. Des. **16**:903–916.

[209] **Weininger, D., A. Weininger and J. L. Weininger.** 1989. Algorithm for generation of unique SMILES notation. J. Chem. Inf. Comput. Sci. **29**:97–101.

[210] **Weininger, D.** 1988. SMILES, a chemical language for information szstems. 1. Introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. **28**:31–36.

[211] **Willett, P., J. M. Barnard and G. M. Downs.** 1998. Chemical similarity searching. J. Chem. Inf. Comput. Sci. **38**:983–996.

[212] **Willett, P., V. Winterman and D. Bawden.** 1986. Implementation of nearest-neighbor searching in an online chemical structure search. J. Chem. Inf. Comput. Sci. **26**:36–41.

[213] **Willett, P., V. Winterman and D. Bawden.** 1986. Implementation of nonhierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering substructure search output. J. Chem. Inf. Comput. Sci. **26**:109–118.

[214] **Wolber, G. and T. Langer.** 2005. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. J. Chem. Inf. Model. **45**:160–169.

[215] **Wyss, P. C., P. Gerber, P. G. Hartman, C. Hubschwerlen, H. Locher, H.-P. Marty and M. Stahl.** 2003. Novel Dihydrofolate Reductase Inhibitors. Structure-Based versus Diversity-Based Library Design and High-Throughput Synthesis and Screening. J. Med. Chem. **46**:2304–2312.

[216] **Xie, D., A. Tropsha and T. Schlick.** 2000. An efficient projection protocol for chemical databases: singular value decomposition combined with truncated-Netwon minimisation. J. Chem. Inf. Comput. Sci. **40**:167–177.

[217] **Xing, L., E. Hodgkin, Q. Liu and D. Sedlock.** 2004. Evaluation and application of multiple scoring functions for a virtual screening experiment. J. Comput.-Aided Mol. Des. **18**:333–344.

[218] **Xue, L., F. L. Stahura, J. W. Godden and J. Bajorath.** 2001. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. J. Chem. Inf. Comput. Sci. **41**:394–401.

[219] **Xue, L., J. W. Godden and J. Bajorath.** 1999. Database searching for compounds with similar biological activity using short binary bit string representation of molecules. J. Chem. Inf. Comput. Sci. **39**:881–886.

[220] **Xue, L., J. W. Godden and J. Bajorath.** 2000. Evaluation of descriptors and mini-fingerprints for the identification of molecules with similar activity. J. Chem. Inf. Comput. Sci. **40**:1227–1234.

[221] **Young, S. S. and D. M. Hawkins.** 1995. Analysis of a 2^9 full factorial chemical library. J. Med. Chem. **38**:2784–2788.