3 *In Silico* Screening

DAGMAR STUMPFE, HANNA GEPPERT, AND JÜRGEN BAJORATH*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany.
bajorath@bit.uni-bonn.de

CONTENTS

Lead Generation Approaches in Drug Discovery,
Edited by Zoran Rankovic and Richard Morphy.

3.1. INTRODUCTION

Over the past decade, advances in combinatorial chemistry and high throughput screening (HTS) have made it possible to synthesize and test large quantities of chemical compounds, often millions. In computer-aided drug discovery, virtual libraries are also generated that by far exceed the size of synthetic libraries and might in some instances contain up to 10^{12} molecules. In order to efficiently process large virtual libraries, or libraries of virtually formatted synthetic compounds, computational methods are required to automatically evaluate databases, reduce their size, or prioritize compounds for biological evaluation. Computer programs designed for such purposes are subsumed as "virtual screening" (VS) or "*in silico* screening" tools and have become an integral part of pharmaceutical research.

Although the term "virtual screening" was introduced only about 10 years ago [1], many of the concepts behind modern VS approaches are well established, with a history going back over several decades. For example, already in 1947, physicochemical properties were correlated with structures using different descriptors such as substructures or topological indices [2]. In the early 1960s, Hansch et al. [3,4] made the first attempts to analyze drug actions with the conceptual framework of quantitative structure–activity relationship (QSAR) analysis when introducing the use of multiple linear regression to correlate physiochemical molecular properties with quantitative bioactivity data. In the mid-1980s, work carried out at Lederle [5] and Pfizer [6] led to an early form of similarity searching based on structural fragments for large-scale structure–activity relationship (SAR) studies and database searching. Since these early days of computational methods in pharmaceutical research, a multitude of different computational approaches have been developed assisting in a variety of VS tasks. Today, a typical application for VS is filtering of libraries for compounds having undesirable properties (e.g., reactivity, toxicity), which is crucial for early-on weeding out of compounds that might later on fail in the development pipeline [7]. Filtering techniques have become especially popular with the introduction of Lipinski's "rule of five" [8], which estimates the likelihood of compounds to be orally absorbed. Besides filtering, probably the most important application domain for contemporary VS is the derivation of predictive models for the identification of active compounds through database searching. In general, VS attempts to identify those candidates that have the highest probability to possess a desired biological activity.

Virtual screening methods can generally be divided into target structure-dependent and ligand-based approaches (see Fig. 3.1). Structure-based methods (SBVS) [9] are applicable when the 3D structure of the biological target is available. Here, docking of ligands into protein binding sites [10] represents the most prominent technique. In contrast, ligand-based virtual screening (LBVS) [11] does not involve target structure information, but utilizes information derived from available small molecules with known biological activity. For example, in a straightforward approach, known active molecules are used as

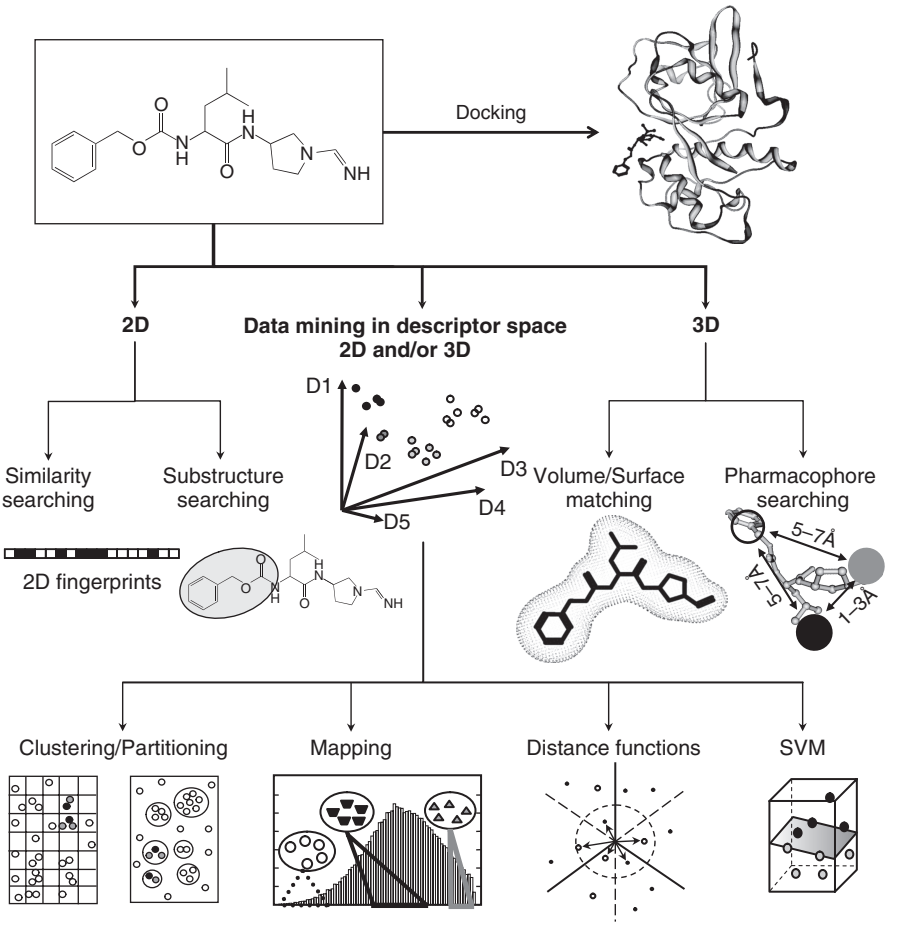

Figure 3.1. Virtual screening methods. Focusing on a cathepsin K inhibitor (PDB entry 1yk7) exemplary computational methods for database screening are shown, distinguishing between structure- and ligand-based approaches. Ligand-based approaches are further divided into techniques that are based on 2D compound representations or 3D conformations or that operate in chemical reference spaces formed by 2D or 3D descriptors.

templates to identify novel molecules that structurally resemble the known ligands and are thus likely to have similar biological properties. Figure 3.1 gives examples of different VS concepts. The design of novel algorithms for VS continues to be an active area of research, and it is hardly possible to provide a comprehensive overview. Hence, in this chapter, we will focus on three categories of VS algorithms, each having a long history in pharmaceutical research: similarity searching using molecular fingerprints, data mining techniques operating in chemical descriptor spaces, and 3D VS approaches relying

on compound conformations. For the latter category, we will focus on long-established pharmacophore searching and molecular docking. Before we start with the description of these methods, fundamental principles underlying VS will be introduced, that is, molecular similarity concepts, molecular descriptors, and chemical reference spaces. The chapter will conclude with practical aspects of VS.

3.2. MOLECULAR SIMILARITY CONCEPTS

Molecular similarity analysis provides the foundation of many current ligand-based VS techniques and was introduced originally during the 1980s and early 1990s. Methods to explore molecular similarity can generally be distinguished dependent on whether they focus on "local" or "global" similarity. "Local" similarities such as the presence of specific functional groups in a given topological and/or geometric arrangement are central to pharmacophores [12] or QSAR [13]. For example, QSAR analysis is traditionally applied to series of analogous structures in order to study molecular features that are responsible for biological activity and predict compound potency as a function of local structural modifications. However, the majority of LBVS methods rely on a "global" or "holistic" molecular perception that originates from the "similarity property principle" (SPP) described from different perspectives in a seminal publication by Johnson and Maggiora [14]. The core of the SPP states that structurally similar molecules are likely to have similar biological activity. Thus, global molecular similarity is associated with biological response behavior of small molecules. The advantage of the "global" molecular view is that it does not rely on hypotheses about activity-determining features. At the same time, one of its disadvantages is that structural features that are not relevant for biological activity are also considered and might negatively affect molecular similarity calculations [15]. Moreover, not all SARs are globally determined. Rather, it is common that small local changes to compound structure dramatically alter biological activity and strictly locally determined SARs fall outside the applicability domain of the SPP and holistic similarity methods [15].

The opposite of molecular similarity analysis is dissimilarity analysis. Computationally, both techniques rely on the calculation of molecular distances in chemical reference spaces defined by mathematical descriptors of molecular structure and physicochemical properties (see Section 3.3). However, while similarity analysis aims at the identification of compounds that are "closest" to each other in chemical space (and might thus be similar even in biological terms), dissimilarity analysis attempts to find compounds with the largest distance to one or several reference molecules. The latter procedure is especially useful for the selection of diverse subsets of compounds, for example, in the context of diverse library design. A detailed discussion of many molecular similarity, dissimilarity, and diversity methods is provided by the recent review of Maldonado et al. [16].

3.2.1. Structure–Activity Relationships

Although the SPP is very intuitive and supported by the success of many LBVS methods, it is also contradicted by general medicinal chemistry experience, as already referred to earlier: a minor structural modification might lead to a substantial change in physicochemical properties and compound potency [17]. While this "similarity paradox" [18] provides the basis for effective lead optimization, it can be a limiting factor for similarity-based LBVS. Why does the SPP not always apply? One major reason is that SARs characterizing biologically active molecules have intrinsic differences [19,20]. Thus, in order to estimate opportunities and limitations of given LBVS methods, a good understanding of SAR characteristics is essential.

In principle, SARs might be *continuous* in nature, *discontinuous*, or *heterogeneous*. In the case of a *continuous* SAR, moderate changes in molecular structure cause small effects on activity, which is consistent with the SPP and the "holistic" view of molecular similarity. For LBVS methods, this provides an opportunity to identify a spectrum of increasingly diverse structures having similar activity, a process also referred to as "lead hopping" [21]). An extreme example of a continuous SAR would be a so-called flat SAR: many chemical modifications do not substantially affect biological activity. In terms of receptor–ligand interactions, such SAR behavior might be the consequence of an adaptable receptor binding site that is characterized by a high degree of structural plasticity and permits small molecules to adopt a spectrum of suboptimal binding modes. This SAR phenotype is often feared in medicinal chemistry because it might remain unclear for a long time if it is possible to optimize active compounds to the extent that they might become viable candidates. A contrasting behavior is presented by a *discontinuous* SAR. In this case, minor modifications lead to a significant enhancement or decrease in activity or even a complete loss. This situation applies when a few critical receptor–ligand interactions that are targeted determine binding. Finally, *heterogeneous* SARs consist of both continuous and discontinuous SAR components. Based on a systematic study of compound activity classes and binding sites [20], heterogeneous SARs are expected to play a role for most biological activities because active sites are usually capable of binding at least a limited number of analogs of active compounds. The presence of differently balanced heterogeneous SARs poses a challenge for the development and application of LBVS methods. A crucial determinant of VS success is how a method responds to specific SARs and one generally observes a strong compound class-dependence of VS method performance. However, to further complicate matters, this dependence is in general not only a consequence of algorithmic details. Rather, the performance of VS methods is equally influenced by the molecular representations and chemical reference spaces that are utilized. Dependent on the choice of these variables, "activity landscapes" of compound classes might strongly differ and limit the applicability of one or the other methodology. Accordingly, two compounds that are regarded

as being highly similar by one LBVS method might be distant in chemical space when another similarity metric or LBVS approach is applied. Thus, the design of novel LBVS algorithms and chemical reference spaces continues to be of crucial importance for applications in pharmaceutical research.

3.2.2. Structure–Selectivity Relationships

In the context of complex SARs, molecular similarity analysis is often a difficult task, as discussed in the previous section. However, the situation can be further complicated when biological activity for a second target is taken into account and compound selectivity information is considered, resulting from differences in binding affinities to two targets. If compound potency differs in a significant way, a ligand is considered to be selective for one target over another. If differential potency of a compound against multiple targets is systematically evaluated, structure–selectivity relationships (SSRs) can be formulated. The study of SSRs adds another level of complexity to SAR analysis because structural modifications of ligands are likely to lead to different changes in potency against individual targets, which often leads to complex SSR phenotypes. This makes the analysis of SSRs and the identification of small molecules that show differential selectivity patterns for individual targets within a protein family an important task in chemical biology [22], chemogenomics [23], and drug discovery because such molecules can serve as selective probes for exploring biological functions and also as starting points for the development of potent and selective lead compounds.

Two recent computational analyses [24–26] of especially designed compound selectivity test systems have indicated that SSRs can be very different in nature, as observed for SARs and discussed above. The first analysis considered 26 compound series with bidirectional (i.e., opposite) selectivity for different pairs of two closely related target proteins. In the second study, 18 compound sets consisting of selective and non-selective molecules for one target compared to a closely related one were investigated. In these two studies, no obvious correlation between structural similarity and compound selectivity could be detected. Structurally similar as well as structurally diverse compounds were found to be selective for the same target, and ligands with distinct selectivity were related to each other by different degrees of structural similarity. Such relationships are illustrated in Figure 3.2, which shows the variety of SSRs covered by one exemplary compound series. As one would intuitively assume, analogs have the same selectivity (down left) and diverse structures are found to exhibit different selectivity (top right). However, also inverse SSRs occur with high frequency [26], that is, closely related structures show significant differences in their selectivity profile (top left) and diverse structures have the same selectivity (down right). These findings suggest that there are no simple structural rules that govern selectivity differences and that the analysis of global molecular similarity alone is not sufficient to predict selectivity profiles.

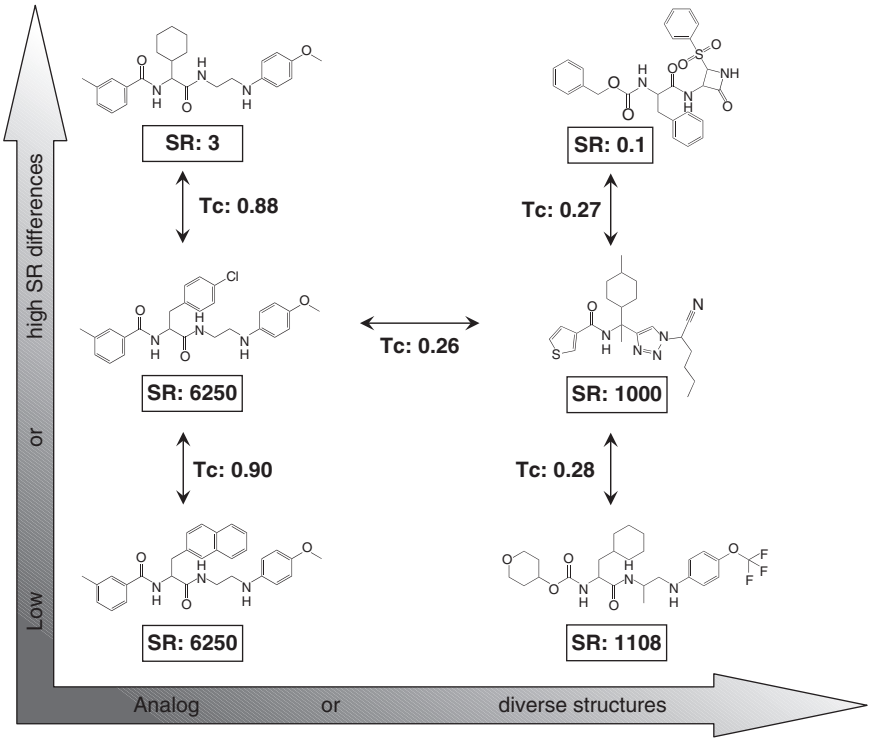

Figure 3.2. Structure–selectivity relationships (SSRs). Different types of SSRs are illustrated that are present in a set of cathepsin enzyme inhibitors. Structurally similar as well as diverse molecules display different selectivity patterns. "SR" means selectivity ratio and is calculated as the quotient of the potency values of a ligand for cathepsin S over cathepsin K. As a measure of structural similarity, "Tc" reports pairwise MACCS Tanimoto coefficient values. Structures are taken from the "cat S/K 7-8 set" of Stumpfe et al. [26].

3.3. MOLECULAR DESCRIPTORS AND CHEMICAL SPACES

Computational methods for *in silico* screening rely on the extraction, analysis, and comparison of physicochemical and structural features of small molecules. The basis for such tasks is a machine-readable compound representation, which can describe "1D," "2D," or "3D" molecular information. The simplest representation (1D) is the chemical formula that reports the different atom types and their number of occurrences in a molecule. The 2D representation of a molecule defines the connectivity of atoms in terms of the presence and nature of chemical bonds. Typical 2D representations are molecular graphs (stored as connection tables) or linear notation systems such as the SMILES language [27]. Finally, the 3D representation of a molecule adds information about

individual compound conformations captured by Cartesian atomic coordinates.

Typically, VS methods do not directly operate on rudimentary molecular representations, but rely on a more computationally efficient or information-rich meta-level. This meta-level is generated through the application of so-called molecular descriptors, which represent highly specific data formats and/or mathematical functions to estimate chemical properties. A multitude of chemical descriptors are available that vary dramatically in complexity, design, and applicability [28]. Principally, they can be divided into *structural*, *numerical property*, and *composite* descriptors, although the boundaries between these categories are fluid. Structural descriptors are usually predefined substructures or arrays of structural fragments (fingerprints) that capture detailed structural information of molecules. This type of descriptor originates from substructure analysis [29] and is mostly applied in similarity searching (see Section 3.5). Numerical property descriptors [30] are used to express physicochemical properties of molecules by means of scalar values (natural or real numbers) that allow a mathematical treatment of the chemical information contained in a compound. Property descriptors are usually applied in QSAR investigations, diversity analysis, combinatorial library design, and VS because most of the methodologies utilized for these purposes depend on the generation of chemical descriptor spaces. Finally, the remaining descriptors that are not represented by single numbers but vectors, matrices, lattices of grid points, or even more complex formulations can be considered as composite descriptors.

We will further discuss numerical property descriptors and chemical space design, which provide the basis for the data mining techniques introduced in Section 3.4. Property descriptors can usually be classified according to the dimensionality of the molecular representation from which they are derived. Accordingly, 1D descriptors are calculated from chemical formulae and capture bulk properties (molecular weight) or simple atom counts such as the number of carbon/oxygen/nitrogen atoms in a molecule. Examples of 2D descriptors are bond counts (e.g., single, double, or aromatic), connectivity and shape indexes [31,32], partial charge descriptors [33], or computational approximations to experimental measurements (e.g., molar refractivity, solubility, partition coefficient, or dipole moment). Furthermore, 3D descriptors can be used to determine characteristics of molecular surfaces and volumes (e.g., solvent-accessible surface area, van der Waals volume) or conformations. For the application of 3D descriptors, a critical aspect is that they can only be reliably used for SAR studies if they are calculated from biologically active compound conformations. However, for database compounds, active conformations, if they exist, are usually not known and need to be predicted, and the uncertainties associated with such predictions are the major limiting factors for modeling based on 3D descriptors. Clearly, compounds are active in three dimensions, and hence 3D representations should in principle have higher predictive value, provided this intrinsic advantage is not overridden by inaccuracies in predicting bioactive compound conformations. From this point

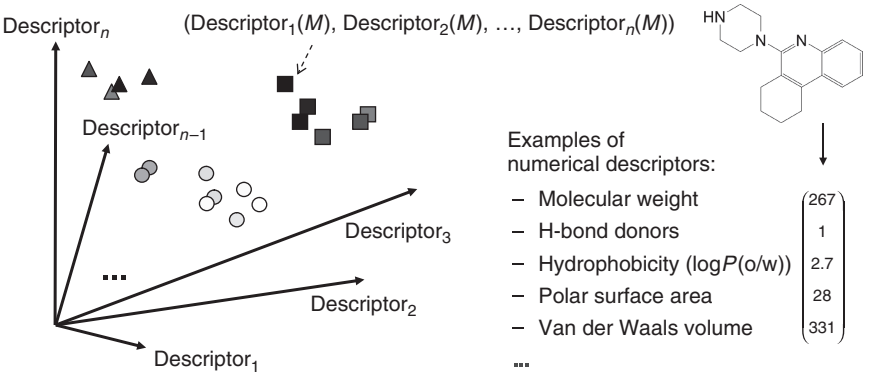

Figure 3.3. Numerical property descriptors and chemical space design. On the right, examples of numerical property descriptors are given that are derived from molecular representations of different dimensionality (1D: molecular weight; 2D: number of H-bond donors, hydrophobic character; 3D: polar surface area, van der Waals volume). Different descriptors are combined to form a chemical reference space, as shown on the left. In n-dimensional space, a molecule M is represented by an n-dimensional descriptor coordinate vector. According to the similarity property principle, compounds having similar activity (indicated by geometric form and shading) should map to the same areas in chemical space.

of view, it is not surprising that there has been, and continues to be, much debate in the literature as to whether 2D or 3D descriptors should generally be preferred [34,35]; currently, there is no general answer to this question, although 2D descriptors have been proven surprisingly successful in many applications. As an advanced alternative to 3D descriptors, several types of 2D descriptors were developed that approximate 3D information from molecular graphs and connectivity tables [36].

The definition of chemical reference spaces using numerical property descriptors is conceptually simple, as illustrated in Figure 3.3. Each of n selected descriptors adds a dimension to n-dimensional chemical space, and coordinates are assigned to each test molecule in this space based on calculated descriptor values. If appropriate descriptors are selected for chemical space design, decreasing distances between molecules correlate with increasing compound similarities and neighboring molecules often exhibit similar biological activity.

3.4. DATA MINING IN CHEMICAL DESCRIPTOR SPACES

The design of chemical descriptor spaces for compound data set representation, diversity analysis, or chemical data mining is one of the most critical tasks in chemoinformatics and LBVS. There are no generally accepted solutions to the

problem of designing chemical reference spaces, and space design is often highly subjective. There are thousands of molecular descriptors available, and key to the design of reference spaces is the selection of descriptors that are responsive to the compound classification or data mining problem at hand. For example, if active compounds should be identified in large data sets, then feature combinations must be identified that are capable of distinguishing between active and inactive compounds. Alternatively, if diversity selection is carried out, chosen descriptors must maximize the diversity distributions within compound libraries. Thus, the task of space design cannot be separated from the specific applications that are performed.

In general, intercompound distance in chemical spaces of any design is utilized as a measure of molecular similarity, as mentioned earlier. However, distance relationships in chemical space are substantially influenced by the dimensionality of reference spaces and correlation effects between feature variables. Consequently, another critical factor in chemical space design is the dimensionality issue. Over the past decade, the generation of low-dimensional reference spaces has dominated space design, which was catalyzed by the introduction of cell-based partitioning methods and information-rich orthogonal descriptors [37]. Compared to high-dimensional space representations, low-dimensional (e.g., 6D or 8D) reference spaces have a number of apparent advantages including, for example, the ability to remove descriptor correlation effects and provide orthogonal reference spaces, control the distribution of compounds over subregions of chemical space, or visualize and interpret compound distributions without significant loss of information. For standard data mining techniques such as cluster analysis or partitioning [38,39], limiting the number of feature variables and controlling descriptor correlation effects typically improve the quality of compound classification results.

Clustering has a particularly long history in chemical database mining [38], and clustering techniques are generally divided into nonhierarchical and hierarchical approaches. Nonhierarchical clustering methods organize compounds into a predefined number of clusters on the basis of nearest-neighbor analysis in descriptor spaces. Cluster membership of compounds is determined by shortest distance to a cluster center. Cluster centers are recalculated and compounds are reassigned until cluster distributions become stable. Hierarchical clustering, on the other hand, builds relationships between clusters in subsequent steps, and the composition of each cluster depends on the one from which it originated. Furthermore, one distinguishes between hierarchical-agglomerative or -divisive methods, dependent on whether the clustering process starts from singletons or a megacluster containing all compounds, respectively.

In recent years, clustering methods have become less popular for chemical database mining because of the dramatically increasing size of available compound libraries and databases. It must be remembered that clustering methods generally involve systematic comparisons of pairwise compound distances in chemical space, regardless of their algorithmic details, which

makes clustering increasingly computationally demanding when compound data sets grow in size. Therefore, partitioning methods have become popular in recent years because they do not rely on pairwise compound comparison, and statistical classification techniques, that is, decision tree methods such as recursive partitioning [40] that can effectively process many descriptors and compounds.

For compound classification and selection from chemical reference spaces, the paradigm of low dimensionality has been widely applied. There are two principal possibilities to generate low dimensional spaces: a priori design utilizing complex descriptors [37] or, alternatively, the reduction of high dimensional feature spaces to minimal descriptor sets carrying most information. For the latter so-called dimension reduction approach, different methodologies are available including principal component analysis [41], nonlinear mapping [42], multidimensional scaling [43], or self-organizing maps [44].

However, it has also been shown that low dimensional chemical spaces are not essential for compound classification and VS. Several methodologies have been introduced that were specifically designed for navigating high dimensional chemical spaces and have proven to be effective in identifying active compounds. For example, support vector machines (SVMs) [45] have become increasingly popular for compound classification (e.g., active versus inactive) [46,47]. Here compounds are projected into high dimensional feature spaces, as visualized in Figure 3.4. Then a hyperplane is constructed by linear combination of training set vectors to optimally separate active and inactive compounds. The margin of the hyperplane is defined as the minimum distance of

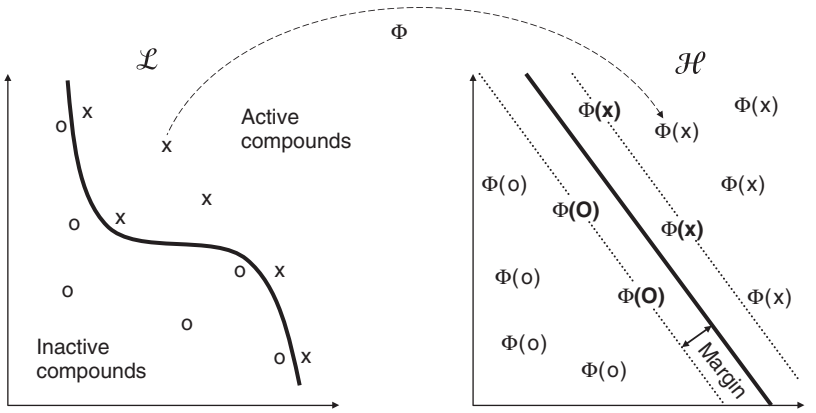

Figure 3.4. Projection into high dimensional feature space. Using a mapping function Φ, active and inactive compounds that are not linearly separable in low dimensional feature space $\mathcal{L}$ (left side) are projected into high dimensional feature space $\mathcal{H}$ (right side), where a linear classification model can be established. The separating hyperplane in $\mathcal{H}$ is defined as a linear combination of a limited number of support vectors (shown in bold) that map to the margin.

any data vector to the hyperplane. A limited number of compound descriptor vectors that map to the margin are called support vectors and are used to construct the hyperplane. This model is then applied to classify test compounds dependent on which side of the hyperplane they fall. The basic idea underlying the SVM approach is that a linear classification model might be established in high dimensional feature space that correctly accounts for SARs that are nonlinear in low dimensional descriptor space. SVMs have achieved high classification accuracy, also in combination with fingerprint descriptors [48].

Furthermore, a simple Euclidian distance function has been introduced to quantify similarity relationships in high dimensional spaces that are centered on regions preferentially populated with compounds having a desired activity [49]. The larger the distance of a test compound from the center of such activity islands, the lower the likelihood that the compound shares the same activity. Accordingly, the distance function has also been reformulated according to Bayesian principles and expressed as a likelihood estimate [50]. In addition, a class of so-called mapping algorithms has been introduced that assigns test compounds to consensus positions of active reference compounds in chemical space based on combination of activity class-specific descriptor value settings or ranges [51–53]. The currently most advanced mapping algorithm, dynamic mapping of activity class-specific descriptor value ranges (DynaMAD) [53], iteratively maps database compounds to consensus positions in chemical reference spaces of increasing dimensionality in order to effectively discriminate between active and irrelevant database compounds. Consensus positions are defined via descriptor value ranges of active reference molecules. The methodology is highly effective for *in silico* screening of very large databases and automatically determines and selects the most relevant descriptors, thereby eliminating subjective descriptor selection. The DynaMAD algorithm has been further refined to enable continuous dimension extension [54]. Thus, with SVM, Bayesian distance functions, and mapping algorithms, conceptually different approaches are currently available to navigate high dimensional chemical space representations in the search for novel active compounds.

3.5. SIMILARITY SEARCHING

Similarity searching also has a long history in pharmaceutical research, similar to clustering, and continues to be one of the most widely used LBVS approaches [55]. Reasons for its popularity might in part be due to its computational efficiency and intuitive search strategy, which is well in accord with the SPP [14]. In a typical similarity search application, computational tools are used to transform molecules into a binary bit string, termed molecular fingerprint, and the resemblance of database compounds to one or a few active reference structures is assessed based on the quantification of fingerprint overlap using a similarity metric or coefficient.

A variety of fingerprint designs have been introduced over the past decades ranging from simple 2D structural representations, such as atom pairs [56], topological torsional fragments [57], and structural keys (MACCS [58], BCI [59]) with hundreds to thousands of bit positions, to complex 3D pharmacophore fingerprints [60] (see next section) consisting of millions of bits. Other more recent fingerprints describe local atom environments (ECFP [61], MOLPRINT [62]) or abstract from structural features through the transformation of property descriptor value ranges into binary formats (MP-MFP [63], PDR-FP [64]). Regardless of the type of molecular information that is represented in a fingerprint, in so-called keyed designs each bit position is associated with exactly one molecular feature and monitors its presence (''1'' bit) or absence (''0'' bit), whereas in ''hashed'' and/or ''folded'' designs different molecular features are mapped to overlapping bit segments with the primary purpose of reducing fingerprint size [65]. The latter approach produces highly characteristic bit patterns, but is chemically less intuitive or interpretable.

Similarity searching using molecular fingerprints has the attractive feature that it can be applied in situations where only a single reference structure is available. This sets similarity searching apart from compound classification or machine learning techniques that rigorously depend on the availability of sets of multiple active molecules. When only a single reference molecule is available, the bit string representation of the template and database compound are compared using a similarity metric that usually produces values between 0 and 1 for minimal and maximal similarity, respectively. A multitude of different similarity metrics have been extensively tested and compared [66–68], and the Tanimoto coefficient (Tc) has de facto evolved to be the most widely accepted. For two molecules A and B, Tc is defined by

$$\mathrm{Tc}(A,B) = \frac{c}{a+b-c}$$

where a and b represent the numbers of bit positions set ''on'' in molecule A and B, respectively, and c represents the number of bits set ''on'' in both molecules. After the calculation of similarity values, database compounds are ranked in the order of decreasing similarity to the active reference molecule, which provides the basis for compound selection.

When multiple active reference molecules are available, traditional similarity searching can be extended through different strategies, which often results in improved search performance [69,70]. This improvement is probably due to an increase in information content of the calculations. Such multiple-template search approaches include fingerprint averaging [71], profiling [72], or scaling techniques [69], as well as various data fusion approaches [55], especially group fusion [73], which is currently very popular. In group fusion, pairwise similarity values are determined between each reference and database molecule. The final score for a database compound is then obtained by summing up the individual pairwise

similarity values (sum fusion rule) or by using the maximum value only (max fusion rule), which has often produced best results in comparative studies [73].

Recently, machine learning approaches have also been adopted for similarity searching using fingerprints. For example, binary kernel discrimination [74] has been extensively studied and promoted by the Willett group [70,73,75,76]. Moreover, a ranking strategy based on SVMs has been evaluated on 2D fingerprints and found to significantly outperform data fusion and fingerprint averaging techniques [48].

As a general limitation of fingerprint-based similarity searching, the so-called size effect has been recognized [77], which refers to the fact that fingerprint search calculations can significantly be affected by differences in molecular complexity and size of database and reference molecules [78]. For conventional fingerprints coding for structural features or pharmacophore arrangements, the more complex and large molecules become, the more bits are usually set "on" (see Fig. 3.5). Since many similarity coefficients such as the Tc only focus on the presence of features ("1" bits) that are common to molecules, but not the common absence ("0" bits) of features, larger molecules are preferentially detected in database searching as being more similar, as illustrated in Figure 3.5*a*. The six compounds shown at the bottom of the figure represent molecules of increasing structural complexity, size, and bit density. When using each template compound separately in order to determine the distribution of MACCS Tc similarity values relative to molecules taken from the ZINC database [79], Tc distributions are shifted toward higher similarity values for larger template molecules. These observations might suggest, for example, that ZINC contains more M6-like compounds than M3- or M4-like ones. However, this is not the case because ZINC molecules are filtered and have an average molecular weight of only about 350 Da (whereas the weight of M6 is about 730 Da).

There are several ways to compensate for the "size effect" in similarity searching. First, a fingerprint having a constant bit density has been designed. This so-called PDR-FP fingerprint has exactly 93 of 500 bits set "on" for each molecule, independent of its structural complexity [64]. As can be seen in Figure 3.5*b*, distributions of PDR-FP similarity values between reference and ZINC compounds are very similar and thus not determined by molecular size. Second, similarity metrics can be modified to equally weight bits that are set "on" and "off" [80,81]. Also, a recent study has shown that even random reduction in fingerprint bit density of highly complex reference molecules to the average level of database compounds can balance complexity effects and increase hit rates [82].

3.6. PHARMACOPHORE SEARCHING

A pharmacophore is defined as the spatial arrangement of atoms or functional groups that are known or predicted to be essential for the specific activity of a small molecule (i.e., specific interactions with its biological target). Chemical

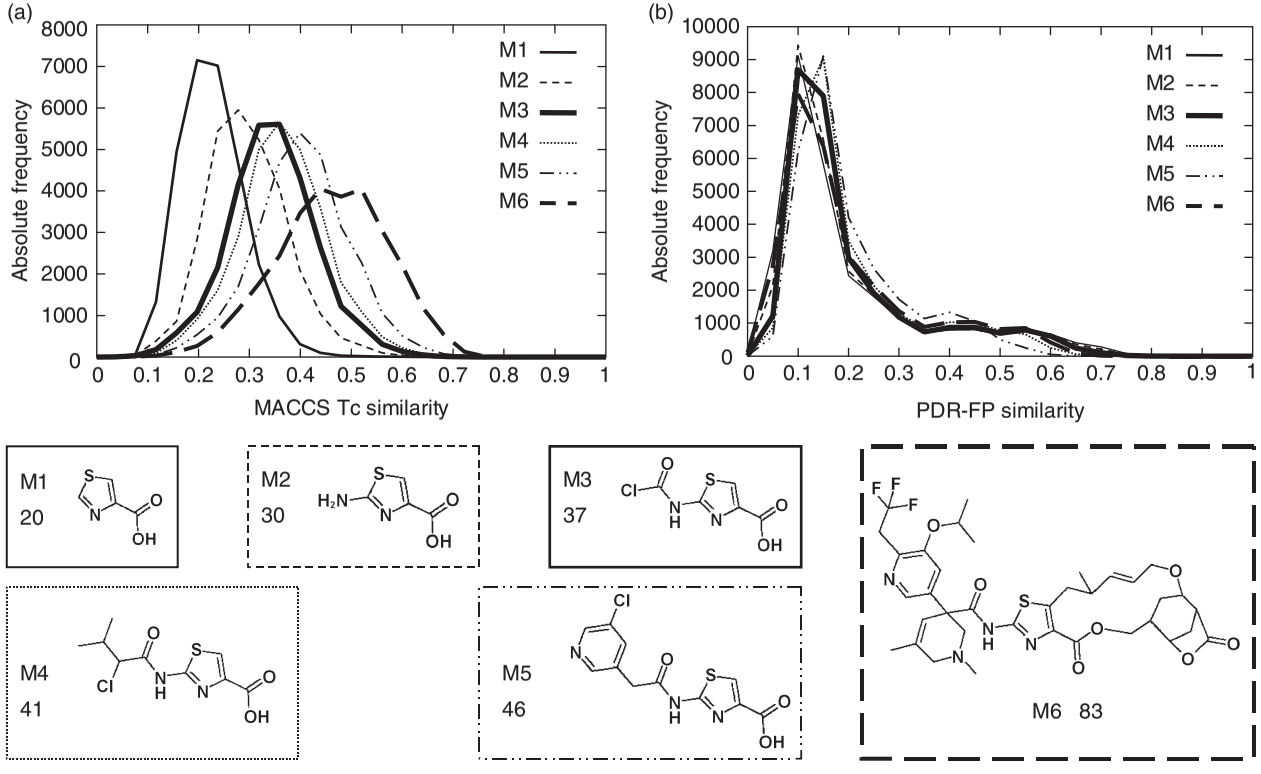

Figure 3.5. Molecular size and complexity effect. The figure shows the distributions of similarity values produced by similarity searches using a single reference molecule of increasing molecular size and complexity against the ZINC database [79]. Different reference molecules (M1–M6) are shown at the bottom and labeled with the number of bits set to "1" in the MACCS fingerprint. In **(a)** the distributions of MACCS Tc values are presented. It can be seen that the larger and more complex the reference molecules become, the more the distributions are shifted to higher similarity values. By contrast, in **(b)** the distributions of similarity values are shown that result from the comparison of molecules using the PDR-FP fingerprint format [64] that sets for each test molecule 93 bits to "1", regardless of its size. Accordingly, these similarity value distributions are very similar. Reference structures M1–M6 were taken from Flower [77].

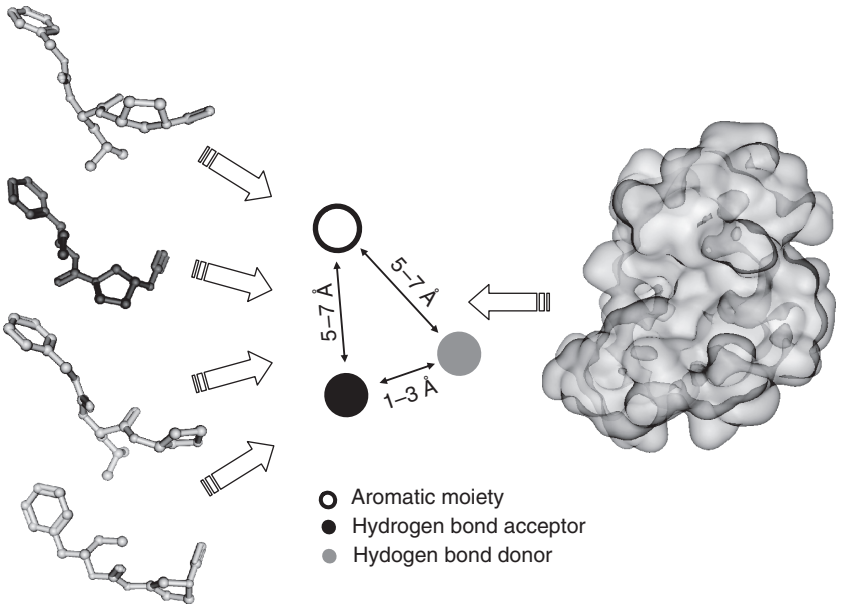

Figure 3.6. Pharmacophore model. These models can be derived in different ways. Using algorithms for "pharmacophore mapping," the model is obtained by combining features common to low energy conformations of a set of known active compounds. Alternatively, a 3D structure-based model can be built to complement the shape and chemical features of a binding site.

features that are frequently utilized in pharmacophore models include hydrogen bond acceptors or donors, positively or negatively charged groups, and hydrophobic or aromatic moieties. Usually, three or four features are combined to yield a pharmacophore model and the center of each feature is represented as a point. A conventional three- or four-point pharmacophore model is unambiguously defined by chosen feature points and pairwise interfeature distances. Pharmacophores can be derived from one or several known ligands or from 3D structures of receptor–ligand complexes, as illustrated in Figure 3.6.

For ligand-based derivation of pharmacophores, 2D or 3D information can be utilized. 2D pharmacophores are calculated from the 2D molecular graph representation of active molecules. Here, distances between two pharmacophore points are expressed through the shortest connecting path in the molecular graph (i.e., the number of bonds separating these points). By contrast, 3D pharmacophores are based on 3D conformations of active compounds, geometric in nature, and contain more information than 2D pharmacophores. The process of deriving a pharmacophore model from multiple active compounds is often termed pharmacophore mapping. The key assumption underlying this technique is that given active molecules share the same mode of action. Most 3D pharmacophore models are derived from

compounds whose bioactive conformations are not known, which requires systematic conformational exploration and prioritization of compound conformations. If bioactive conformations cannot be accurately predicted, corresponding pharmacophore models are intrinsically flawed, similar to the situation in 3D QSAR analysis. If bioactive conformations are unknown, a major objective of pharmacophore mapping techniques is to prioritize pharmacophore models that contain the largest number of features common to low energy conformations of given active molecules. Algorithms often applied in pharmacophore mapping include, among others, the constrained systematic search approach [83], clique detection (as used in the DISCO program) [84,85], a maximum likelihood approach (Catalyst/HipHop) [86], or genetic algorithms (GASP) [87,88]. Ideally, several highly potent compounds belonging to different structural series should be available as a starting point for automatic pharmacophore mapping. The more restricted the conformational space available to these test molecules is, the more likely it becomes to identify important pharmacophore elements and produce sufficiently accurate models. If available, structural information about the target binding site or ligand–target complexes can also be utilized for the generation and refinement of pharmacophore models. In these cases, ligand–target interaction information can be directly taken into account to aid in the definition of pharmacophore features shared by series of active compounds (most of which might not have been studied crystallographically).

Pharmacophore models are generally applied as 3D queries to identify active compounds in database searching, to aid in the design or optimization of active molecules, assemble target-focused libraries, or derive 3D QSAR models. However, the original application of pharmacophores has been 3D database searching [89]. Compounds matching a pharmacophore query are considered potential hits. Because exact matches of database compounds to a model are only rarely seen, pharmacophore models are usually relaxed by replacing exact interfeature distances by distance ranges. For example, a calculated interfeature distance of 6.3 Å might be replaced by a 5–7 Å interval. Furthermore, different types of features (e.g., hydrophobic or aromatic moieties or charged and polar groups) might be permitted to map individual feature points. Pharmacophore searching is often applied with the intention to identify active compounds belonging to different structural series [89]. However, ''lead hops'' can only be accomplished if a pharmacophore model abstracts from the exact features of known active compounds and becomes ''fuzzy.''

In order to circumvent the need to predict exact bioactive conformations of reference compounds, ''pharmacophore fingerprinting'' has been introduced [60]. Following this approach, test molecules are subjected to systematic conformational search, and all possible arrangements of predefined pharmacophores are recorded. In a pharmacophore fingerprint, each bit position represents a different (three- or four-point) pharmacophore model resulting from the systematic combination of a preselected set of alternative pharmacophore features and distance intervals. For each test molecule, the fingerprint

stores all alternative pharmacophore models it produces by setting the corresponding bit positions to "1". Pharmacophore fingerprints are then compared in the same way as 2D fingerprints. The underlying idea is that the more potential pharmacophore models are shared by active reference and test molecules, the larger the fingerprint overlap becomes, and the more likely these molecules are to have similar activity. Thus, potential pharmacophore resemblance is utilized as an indicator of biological activity, rather than matches of individual pharmacophore queries.

3.7. MOLECULAR DOCKING

With the increasing number of experimental protein structures that are publicly available in the Protein Data Bank (PDB) [90], the interest in structure-based VS methods, especially protein–ligand docking, has continuously grown. The molecular docking process includes the prediction of the binding conformation and orientation (i.e., the "pose") of a ligand in an active site and the interactions it engages in and, in addition, the energy-based scoring and ranking of alternative ligands. Docking is often applied as a hit identification approach when no or only few active ligands are known, but also at later stages during lead optimization when potential binding modes are studied in detail and interactions optimized.

Docking algorithms also have a long history beginning with pioneering efforts by the Kuntz group who developed the DOCK program in 1982 [91], which is under continuous development to this date [92]. Currently, there is a plethora of in part rather different docking algorithms available, and other popular docking programs include Autodock [93], Gold [94], or FlexX [95]. Although methodological complexity, prediction accuracy, and computational efficiency have substantially increased over the years, docking algorithm generally involves the same tasks, that is, posing and scoring, as mentioned above. Posing refers to the process of exploring different orientations and conformations of a small molecule to fit an active site. Scoring is applied to identify the most probable pose for each molecule and predicts relative binding energies of different compounds as a basis for database ranking.

Regardless of the algorithms used, posing and scoring are nontrivial tasks. Difficulties in posing arise from the translational and rotational degrees of freedom that a small molecule has within a binding site and, in particular, from conformational flexibility. Since computational resources were limited in the early 1980s, pioneering docking programs only considered translational and rotational degrees of freedom of ligands with precomputed 3D conformations [91], a process referred to as rigid body docking. The next generation of docking algorithms often relied on multiconformation docking [96], whereas current state-of-the-art approaches usually fully explore ligand flexibility [10,97,98] and, to a lesser extent, also protein flexibility [99]. Posing often involves rapid scoring based on the calculation of approximate shape and electrostatic complementarity. Early stage scoring is used to preselect small

subsets of preferred conformers that are then subjected to more elaborate scoring schemes to more precisely treat electrostatic and van der Waals interactions, predict interaction energies, and rank database compounds [100].

A variety of scoring functions has been introduced that can roughly be assigned to three categories: force field-based, knowledge-based, and empirical. Force field-based scoring employs molecular mechanics-type treatment of intra- and intermolecular interactions, whereas knowledge-based and empirical scoring attempts to adjust energy terms by fitting experimental complex structures and/or binding data of known active compounds. Thus, knowledge-based and empirical scoring schemes are typically tailored toward specific targets or target classes and often not readily transferable. Extensive computational studies comparing different docking programs and scoring functions have been carried out and indicate that limited scoring accuracy still represents the major shortcoming in structure-based VS, whereas posing is often capable of accurately predicting binding modes [10,101,102]. Current scoring functions are generally limited in their ability to accurately account for electrostatic and long-range interactions as well as solvation and entropic effects and to balance enthalpic and entropic effects. Simply put, a reliable calculation of the free energy of binding is not yet feasible, irrespective of the details of scoring functions. In order to balance imperfections of individual scoring functions, consensus scoring has been introduced that combines information from different scoring schemes and often improves search accuracy [103].

For a more detailed discussion of available algorithms, scoring functions, and docking techniques see, for example, excellent reviews by Leach and colleagues [98] or Klebe [104].

3.8. PRACTICAL ASPECTS OF VIRTUAL SCREENING

At the beginning of a VS campaign, two questions need to be considered: first, what is the aim of the virtual screen, and second, what are the available data. Possible purposes of a virtual screen might include the identification of novel hits, structurally diverse compounds having similar activity (lead hopping [21]), or target-selective molecules. For LBVS, the availability of single or multiple known active compounds and their structural relationships ultimately determine which methods can be applied and what might be accomplished. The design of VS protocols will usually differ, depending on the goals. However, LBVS generally consists of four steps: database preparation, assembly of compound reference sets, application of search algorithms, and compound selection, which are discussed in the following. Figure 3.7 illustrates alternative VS scenarios.

3.8.1. Database Preparation

A VS campaign begins with the selection or creation of an appropriate screening database. In principle, three types of databases can be distinguished: targeted or

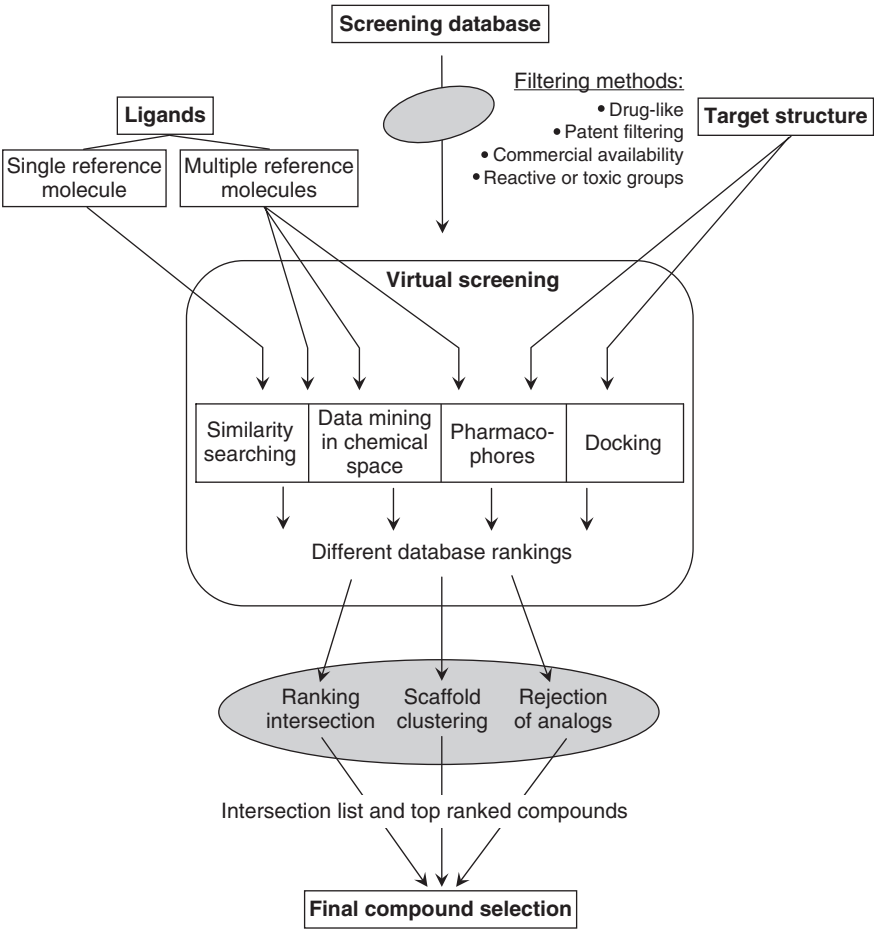

Figure 3.7. Virtual screening (VS) process. Dependent on the available ligand- and/or protein information and the goal of the VS campaign, different methods and strategies can be applied during the four major steps of a typical virtual screen: database preparation, selection of reference molecules or target structures, database ranking, and compound selection.

focused, general, and diverse. Targeted and focused libraries are designed to have a high probability to be enriched with compounds that are active against individual targets or a target family. For the design of such compound libraries, detailed knowledge about active molecules and/or the target structures must be taken into account. General screening libraries are intended to be used against many targets and often contain millions of molecules. These libraries are typically assembled from many different sources that are available and are "pragmatic" in nature. Diverse libraries are typically the product of a computational design process that attempts to cover as much chemical space as possible

with a limited number of compounds. Such diversity design has guided the generation of many combinatorial compound libraries. Diversity design and diversity selection are conceptually distinct because diversity selection is based on the principles of molecular dissimilarity discussed earlier. In diversity selection, one typically attempts to complement an existing database with sets of compounds that are not yet covered and further increase diversity. This type of diversity selection has determined compound acquisition strategies of pharmaceutical companies for many years. For most LBVS methods that employ 2D molecular representations, computational time is no longer a limiting factor, and hence very large compound database can be readily processed. The ZINC collection introduced by the Shoichet group [79] is an example of a publicly available general compound database, which has become very popular for VS applications. The current version of ZINC (version 8) contains 8.5 million molecules from various vendor sources in modeled 3D conformations.

Prior to VS calculations, a compound database is usually filtered to remove compounds with undesired properties or flawed computational representations. Popular filters include drug- or lead-like filters, the "rule of five" [8], or filters designed to detect unstable, reactive, or toxic molecules. If necessary, the size of a compound database can be further reduced by using application-dependent information about known inactive or other irrelevant compounds. Using a fast similarity searching algorithm, compounds being very similar to known inactive structures can be identified and omitted, a technique introduced as "data shaving" [105].

3.8.2. Assembly and Selection of Reference Molecules

Typically, there are two sources of active reference molecules, hits from experimental screening or known active molecules from the scientific or patent literature. While screening hits usually have molecular size and complexity comparable to database compounds, structures taken from the literature often represent chemically optimized and highly potent molecules that are larger and more complex than average database molecules. As described in Section 3.5, this difference in complexity provides a difficult search scenario and often biases LBVS (especially similarity searching) toward the preferential selection of other large (but mostly inactive) compounds. In principle, reference molecules should be preferred that are comparable in size and chemical complexity to an "average" database compound.

It is generally not advisable to use all known active compounds as reference molecules. The complete set may contain analog series (e.g., from previous lead optimization efforts), which might overemphasize certain chemotypes and bias VS calculations. Thus, especially for the purpose of lead hopping, clustering of active compounds by core structures [106] is usually helpful in order to avoid the use of chemically unbalanced reference sets. If only one reference molecule is available, the search radius might still be expandable by applying methods such as "turbo similarity searching" [107], which includes molecules in the

reference set that were found to be most similar to the active compound in an initial similarity search, although their potential activity is not known.

Several LBVS approaches, especially data mining algorithms such as SVM, require a set of inactive training compounds in addition to active reference molecules. If the active reference set was selected from the literature, usually no inactive molecules are available. In this case, a random subset of the screening database can be used to represent features of inactive compounds because, in practice, the ratio of active to inactive molecules is usually close to zero in large databases. A recent comparative study [48] has shown that the use of only 14 or 144 random database compounds as inactive training set for an SVM was sufficient for significantly increasing the search performance of traditional similarity searching, which exclusively relies on active reference molecules.

3.8.3. Strategies for Database Searching

In order to detect as many new hits as possible, different VS approaches should be applied in concert. The reason for this is the well-known compound class dependence and complementarity of VS algorithms that rely on different molecular representations and take different molecular properties and input information into account [34]. As a consequence, top candidate compounds suggested by different methods usually overlap only very little, which was observed, for example, when various molecular representations were applied for database searching in order to select candidate compounds for followup evaluation after HTS [108]. Candidate compound lists had only about 15 percent overlap and most of the experimentally verified hits were detected by only one of several alternative search methods.

For the combination or parallel execution of multiple search methods, different strategies such as serial application, data fusion, or parallel compound selection exist. Serial application is increasingly used in the context of molecular docking, where large virtual libraries are reduced to a manageable size on the basis of fast 2D and/or 3D similarity searching prior to computationally expensive flexible docking calculations [109–111]. Data fusion techniques have extensively been studied in the context of similarity searching and combine similarity scores from different similarity metrics (similarity fusion) [68], results of different molecular representations by averaging rank positions (rank fusion) [112], or similarity scores produced using different active reference molecules (group fusion) [73]. In protein–ligand docking, consensus scoring [103] essentially represents an analogous approach to similarity fusion in similarity searching. A recent study compared the performance of alternative strategies for combination of molecular docking and 2D similarity searching [113]. Parallel selection of candidate compounds from individual docking and similarity search rankings was found to be superior to rank fusion. These findings were rationalized by a potentially unfavorable averaging effect of rank fusion, whereas parallel selection was more effective due to complementarity of structure- and ligand-based screening [113].

3.8.4. Compound Selection

As discussed earlier, the overlap between top-ranked molecules selected by different LBVS methods is usually very limited. Thus, if a compound is placed consistently high on different rankings, it should have considerable probability to be active and thus be selected for experimental evaluation. In order to identify preferred candidates, the overlap in, for example, the top 1000 rank positions might be determined between all pairs, triplets, etc. of compound rankings. In addition, the molecules occurring at the top (e.g., 10–100 positions) of each individual ranking should be considered. Ultimately, top candidates from different virtual screens and compounds occurring in intersections of these rankings should be selected in order to maximize the probability of identifying novel active molecules.

Depending on the desired size of the final compound selection set, the number of preselected molecules must often be further reduced. For this purpose, two strategies can be applied. First, preselected molecules with a high degree of structural similarity to one or more reference compounds are removed by calculating pairwise fingerprint similarity values (e.g., MACCS Tc values) for a candidate and each reference molecule, and removing compounds that exceed a predefined similarity threshold value. Second, the occurrence of different scaffolds or chemotypes in the final selection set can be balanced, for example, by including as many alternative core structures as possible and avoiding the inclusion of analog series. For this purpose, preselected compounds can be reduced to heteroatom-containing scaffolds [106] or even cyclic carbon skeletons [114]. Of course, depending on the underlying SAR characteristics, "active scaffolds" might still be missed when including only one or two inactive analogs in the final selection set.

3.9. CONCLUDING REMARKS

The drug discovery process can roughly be divided into an early phase, which mainly focuses on target and lead discovery, and a late phase, where clinical evaluation and development take center stage [115]. *In silico* screening approaches are exclusively applied during very early stages of this process: hit identification and hit-to-lead transition. In this chapter, we have discussed various VS approaches, characteristics of structure–activity or structure–selectivity relationships that are relevant for VS, and practical screening aspects. The methodological diversity that characterizes the VS field has been emphasized, as well as some intrinsic limitations. *In silico* screening as presented herein is not a search for development candidates, but rather a complex and iterative process whose primary goal is the identification of structurally diverse and weakly to moderately active molecules that might be transformed into viable leads. Typically, different computational methods are sequentially utilized to filter databases or reduce their size and enable the application of

increasingly sophisticated and time-consuming analyses. An important point to consider is that VS cannot be separated from the experimental evaluation of compounds. In practice, its major attraction is to reduce the experimental effort involved in identifying and validating novel active compounds. Furthermore, VS and biological screening can also be iteratively performed and integrated [116], which is expected to have considerable potential for future applications.

REFERENCES

1. Walters, W. P., Stahl, M. T., Murcko, M. A. (1998). Virtual screening – an overview. *Drug Discovery Today*, *3*, 160–178.
2. Wiener, H. (1947). Structural determination of Paraffin boiling points. *Journal of the American Chemical Society*, *69*, 17–20.
3. Hansch, C., Maloney, P. P., Fujita, T., Muir, R. M. (1962). Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. *Nature*, *194*, 178–180.
4. Hansch, C., Fujita, T. (1964). p–σ–π analysis. A method for the correlation of biological activity and chemical structure. *Journal of the American Chemical Society*, *86*, 1616–1626.
5. Carhart, R. E., Smith, D. H., Venkataraghavan, R. (1985). Atom pairs as molecular features in structure–activity studies: definitions and applications. *Journal of Chemical Information and Computer Sciences*, *25*, 64–73.
6. Willett, P., Winterman, V., Bawden, D. (1986). Implementation of nearest neighbor searching in an online chemical structure search system. *Journal of Chemical Information and Computer Sciences*, *26*, 36–41.
7. DiMasi, J. A., Hansen, R. W., Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, *22*, 151–185.
8. Lipinski, C. A., Lombardo, F., Dominy, B. W., Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, *23*, 3–25.
9. Shoichet, B. K. (2004). Virtual screening of chemical libraries. *Nature*, *432*, 862–865.
10. Kitchen, D. B., Decornez, H., Furr, J. R., Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, *3*, 935–949.
11. Stahura, F. L., Bajorath, J. (2005). New methodologies for ligand-based virtual screening. *Current Pharmaceutical Design*, *11*, 1189–1202.
12. Mason, J. S., Good, A. C., Martin, E. J. (2001). 3-D pharmacophores in drug discovery. *Current Pharmaceutical Design*, *7*, 567–597.
13. Eposito, E. X., Hopfinger, A. J., Madura, J. D. (2004). Methods for applying the quantitative structure–activity relationship paradigm. *Methods in Molecular Biology*, *275*, 131–214.
14. Johnson, M. A., Maggiora, G. M. eds. (1990). *Concepts and applications of molecular similarity*, Wiley, New York.

15. Auer, J., Bajorath, J. (2008). Molecular similarity concepts and search calculations. *Methods in Molecular Biology*, *453*, 327–347.

16. Maldonado, A. G., Doucet, J. P., Petitjean, M., Fan, B.-T. (2006). Molecular similarity and diversity in chemoinformatics: from theory to applications. *Molecular Diversity*, *10*, 39–79.

17. Kubinyi, H. (1998). Similarity and dissimilarity. A medicinal chemist's view. *Perspectives in Drug Discovery and Design*, *9–11*, 225–252.

18. Bajorath, J. (2002). Virtual screening: methods, expectations, and reality. *Current Drug Discovery*, *2*, 24–28.

19. Eckert, H., Bajorath, J. (2007). Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today*, *12*, 225–233.

20. Peltason, L., Bajorath, J. (2007). Molecular similarity analysis uncovers heterogeneous structure–activity relationships and variable activity landscapes. *Chemistry and Biology*, *14*, 489–497.

21. Cramer, R. D., Jilek, R. J., Guessregen, S., Clark, S. J., Wendt, B., Clark, R. D. (2004). "Lead hopping". Validation of topomer similarity as a superior predictor of biological activities. *Journal of Medicinal Chemistry*, *47*, 6777–6791.

22. Spring, D. R. (2005). Chemical genetics to chemical genomics: small molecules offer big insights. *Chemical Society Reviews*, *34*, 472–482.

23. Bredel, M., Jacoby, E. (2004). Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nature Reviews Genetics*, *5*, 262–275.

24. Stumpfe, D., Ahmed, H. E. A., Vogt, I., Bajorath, J. (2007). Methods for computer-aided chemical biology. Part 1: design of a benchmark system for the evaluation of compound selectivity. *Chemical Biology and Drug Design*, *70*, 182–194.

25. Vogt, I., Stumpfe, D., Ahmed, H. E. A., Bajorath, J. (2007). Methods for computer-aided chemical biology. Part 2: evaluation of compound selectivity using 2D fingerprints. *Chemical Biology and Drug Design*, *70*, 195–205.

26. Stumpfe, D., Geppert, H., Bajorath, J. (2008). Methods for computer-aided chemical biology. Part 3: analysis of structure–selectivity relationships through single- or dual-step selectivity searching and Bayesian classification. *Chemical Biology and Drug Design*, *71*, 518–528.

27. Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, *28*, 31–36.

28. Todeschini, R., Consonni, V. (2000). Handbook of molecular descriptors. In: Mannhold, R., Kubinyi, H., Timmerman, H. eds., *Methods and principles in medicinal chemistry* (Vol. 11), Wiley, New York.

29. Barnard, J. M. (1993). Substructures searching methods: old and new. *Journal of Chemical Information and Computer Sciences*, *33*, 532–538.

30. Livingstone, D. J. (2000). The characterization of chemical structures using molecular properties. A survey. *Journal of Chemical Information and Computer Sciences*, *40*, 195–209.

31. Hall, L. H., Kier, L. B. (1977). The nature of structure–activity relationships and their relation to molecular connectivity. *European Journal of Medicinal Chemistry*, *12*, 307–312.

32. Hall, L. H., Kier, L. B. (1991). The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling. In: Lipkowitz, K. B., Boyd, D. B. eds., *Reviews in computational chemistry* (Vol. *2*), Wiley, New York, pp. 367–422.
33. Gasteiger, J., Marsili, M. (1980). Iterative partial equalization of orbital electronegativity – a rapid access to atomic charges. *Tetrahedron*, *36*, 3219–3228.
34. Sheridan, R. P., Kearsley, S. K. (2002). Why do we need so many chemical similarity search methods? *Drug Discovery Today*, *7*, 903–911.
35. Bajorath, J. (2001). Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *Journal of Chemical Information and Computer Sciences*, *41*, 233–245.
36. Labute, P. (2004). Derivation and applications of molecular descriptors based on approximate surface area. In: Bajorath, J. ed., *Chemoinformatics – concepts, methods, and tools for drug discovery*, Humana Press, Totowa, NJ, pp. 261–278.
37. Pearlman, R. S., Smith, K. M. (1998). Novel software tools for chemical diversity. *Perspectives in Drug Discovery and Design*, *9*, 339–353.
38. Willett, P. (1987). *Similarity and clustering in chemical information systems*, Research Studies Press, Letchworth.
39. Stahura, F. L., Bajorath, J. (2003). Partitioning methods for the identification of active molecules. *Current Medicinal Chemistry*, *8*, 707–715.
40. Rusinko, A. 3rd, Farmen, M. W., Lambert, C. G., Brown, P. L., Young, S. S. (1999). Analysis of a large structure/biological activity data set using recursive partitioning. *Journal of Chemical Information and Computer Sciences*, *39*, 1017–1026.
41. Xue, L., Bajorath, J. (2000). Molecular descriptors for effective classification of biologically active compounds based on principal component analysis identified by a genetic algorithm. *Journal of Chemical Information and Computer Sciences*, *40*, 801–809.
42. Agrafiotis, D. K., Lobanov, V. S. (2000). Nonlinear mapping networks. *Journal of Chemical Information and Computer Sciences*, *40*, 1356–1362.
43. Agrafiotis, D.K., Rassokhin, D.N., Lobanov, V.S. (2001). Multi-dimensional scaling and visualization of large molecular similarity tables. *Journal of Computational Chemistry*, *22*, 488–500.
44. Agrafiotis, D. K., Xu, H. (2002). A self-organizing principle for learning nonlinear manifolds. *Proceedings of the National Academy of Sciences of the United States of America*, *99*, 15869–15872.
45. Cristianini, N., Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*, Cambridge University Press, Cambridge, UK.
46. Warmuth, M. K., Liao, J., Rätsch, G., Mathieson, M., Putta, S., Lemmen, C. (2003). Active learning with support vector machines in the drug discovery process. *Journal of Chemical Information and Computer Sciences*, *43*, 667–673.
47. Jorissen, R. N., Gilson, M. K. (2005). Virtual screening of molecular databases using a support vector machine. *Journal of Chemical Information and Modeling*, *45*, 549–561.
48. Geppert, H., Horváth, T., Gärtner, T., Wrobel, S., Bajorath, J. (2008). Support-vector-machine-based ranking significantly improves the effectiveness of similarity

searching using 2D fingerprints and multiple reference compounds. *Journal of Chemical Information and Modeling*, *48*, 742–746.

49. Godden, J. W., Bajorath, J. (2006). A distance function for retrieval of active molecules from complex chemical space representations. *Journal of Chemical Information and Modeling*, *46*, 1094–1097.
50. Vogt, M., Godden, J. W., Bajorath, J. (2007). Bayesian interpretation of a distance function for navigating high-dimensional descriptor spaces. *Journal of Chemical Information and Modeling*, *47*, 39–46.
51. Godden, J. W., Furr, J. R., Xue, L., Stahura, F. L., Bajorath, J. (2004). Molecular similarity analysis and virtual screening in binary-transformed chemical descriptor spaces with variable dimensionality. *Journal of Chemical Information and Computer Sciences*, *44*, 21–29.
52. Eckert, H., Bajorath, J. (2006). Determination and mapping of activity-specific descriptor value ranges for the identification of active compounds. *Journal of Medicinal Chemistry*, *49*, 2284–2293.
53. Eckert, H., Vogt, I., Bajorath, J. (2006). Mapping algorithms for molecular similarity analysis and ligand-based virtual screening: design of DynaMAD and comparison with MAD and DMC. *Journal of Chemical Information and Modeling*, *46*, 1623–1634.
54. Vogt, I., Bajorath, J. (2008). Design and exploration of target-selective chemical space representations. *Journal of Chemical Information and Modeling*, *48*, 1389–1395.
55. Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today*, *11*, 1046–1053.
56. Carhart, R. E., Smith, D. H., Venkataraghavan, R. (1985). Atom pairs as molecular features in structure–activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, *25*, 64–73.
57. Nilakantan, R., Bauman, N., Dixon, J. S., Venkataraghavan, R. (1987). Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *Journal of Chemical Information and Computer Sciences*, *27*, 82–85.
58. McGregor, M. J., Pallai, P. V. (1997). Clustering of large databases of compounds: using the MDL "keys" as structural descriptors. *Journal of Chemical Information and Computer Sciences*, *37*, 443–448.
59. Barnard, J. M., Downs, G. M. (1997). Chemical fragment generation and clustering software. *Journal of Chemical Information and Computer Sciences*, *37*, 141–142.
60. Mason, J. S., Morize, I., Menard, P. R., Cheney, D. L., Hulme, C., Labaudiniere, R. F. (1999). New 4-point pharmacophore method for molecular similarity and diversity applications: overview over the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *Journal of Medicinal Chemistry*, *42*, 3251–3264.
61. Klon, A. E., Glick, M., Thoma, M., Acklin, P., Davies, J. W. (2004). Finding more needles in the haystack: a simple and efficient method for improving high-throughput docking results. *Journal of Medicinal Chemistry*, *47*, 2743–2749.
62. Bender, A., Mussa, H. Y., Glen, R. C., Reiling, S. (2004). Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *Journal of Chemical Information and Computer Sciences*, *44*, 1708–1718.

63. Xue, L., Godden, J. W., Stahura, F. L., Bajorath, J. (2003). Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *Journal of Chemical Information and Computer Sciences*, *43*, 1151–1157.

64. Eckert, H., Bajorath, J. (2006). Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *Journal of Chemical Information and Modeling*, *46*, 2515–2526.

65. Ihlenfeldt, W. D., Gasteiger, J. (1994). Hash codes for the identification and classification of molecular structure elements. *Journal of Computational Chemistry*, *15*, 793–813.

66. Willett, P., Winterman, V. (1986). A comparison of some measures of inter-molecular structural similarity. *Quantitative Structure–Activity Relationships*, *5*, 18–25.

67. Holliday, J. D., Hu, C. Y., Willett, P. (2002). Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Combinatorial Chemistry and High Throughput Screening*, *5*, 155–166.

68. Salim, N., Holliday, J., Willett, P. (2003). Combination of fingerprint-based similarity coefficients using data fusion. *Journal of Chemical Information and Computer Sciences*, *43*, 435–442.

69. Xue, L., Stahura, F. L., Godden, J. W., Bajorath, J. (2001). Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *Journal of Chemical Information and Computer Sciences*, *39*, 699–704.

70. Hert, J., Willet, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A. (2006). New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *Journal of Chemical Information and Modeling*, *46*, 462–470.

71. Schuffenhauer, A., Floersheim, P., Acklin, P., Jacoby, E. (2003). Similarity metrics for ligands reflecting the similarity of the target proteins. *Journal of Chemical Information and Computer Sciences*, *43*, 391–405.

72. Godden, J. W., Xue, L., Stahura, F. L., Bajorath, J. (2000). Searching for molecules with similar biological activity: analysis by fingerprint profiling. *Pacific Symposium on Biocomputing*, *5*, 563–572.

73. Hert, J., Willett, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A. (2004). Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *Journal of Chemical Information and Computer Sciences*, *44*, 1177–1185.

74. Harper, G., Bradshaw, J., Gittins, J. C., Green, D. V. S., Leach, A. R. (2001). Prediction of biological activity for high-throughput screening using binary kernel discrimination. *Journal of Chemical Information and Computer Sciences*, *41*, 1295–1300.

75. Wilton, D., Willett, P., Lawson, K., Mullier, G. (2003). Comparison of ranking methods for virtual screening in lead-discovery programs. *Journal of Chemical Information and Computer Sciences*, *43*, 469–474.

76. Wilton, D. J., Harrison, R. F., Willett, P., Delaney, J., Lawson, K., Mullier, G. (2006). Virtual screening using binary kernel discrimination: analysis of pesticide data. *Journal of Chemical Information and Modeling*, *46*, 471–477.

77. Flower, D. R. (1998). On the properties of bit string-based measures of chemical similarity. *Journal of Chemical Information and Computer Sciences*, *38*, 379–386.

78. Wang, Y., Eckert, H., Bajorath, J. (2007). Apparent asymmetry in fingerprint similarity searching is a direct consequence of differences in bit densities and molecular size. *ChemMedChem*, *2*, 1037–1042.

79. Irwin, J. J., Shoichet, B. K. (2005). ZINC – a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, *45*, 177–182.

80. Fligner, M., Verducci, J., Blower, P. (2002). A modification of the Jaccard–Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics*, *44*, 110–119.

81. Wang, Y., Bajorath, J. (2008). Balancing the influence of molecular complexity on fingerprint similarity searching. *Journal of Chemical Information and Modeling*, *48*, 75–84.

82. Wang, Y., Geppert, H., Bajorath, J. (2008). Random reduction in fingerprint bit density improves compound recall in search calculations using complex reference molecules. *Chemical Biology and Drug Design*, *71*, 511–517.

83. Dammkoehler, R. A., Karasek, S. F., Shands, E. F. B., Marshall, G. R. (1989). Constrained search of conformational hyperspace. *Journal of Computer-Aided Molecular Design*, *3*, 3–21.

84. Martin, Y. C., Bures, M. G., Danaher, A. A., DeLazzer, J., Lico, I., Pavlik, P. A. (1993). A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *Journal of Computer-Aided Molecular Design*, *7*, 83–102.

85. Martin, Y. C. (2000). DISCO: what we did right and what we missed. In: Guner, O. F. ed., *Pharmacophore perception, development, and use in drug design*, International University Line, La Jolla, CA, pp. 51–66.

86. Barnum, D., Greene, J., Smellie, A., Sprague, P. (1996). Identification of common functional configurations among molecules. *Journal of Chemical Information and Computer Sciences*, *36*, 563–571.

87. Jones, G., Willett, P., Glen, R. C. (1995). A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *Journal of Computer-Aided Molecular Design*, *9*, 532–549.

88. Jones, G., Willett, P., Glen, R. C. (2000). GASP: genetic algorithm superposition program. In: Guner, O. F. ed., *Pharmacophore perception, development, and use in drug design*, International University Line, La Jolla, CA, pp. 85–106.

89. Langer, T., Krovat, E. M. (2003). Chemical feature-based pharmacophores and virtual library screening for discovery of new leads. *Current Opinion in Drug Discovery and Development*, *6*, 370–376.

90. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28*, 235–242.

91. Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology*, *161*, 269–288.

92. Moustakas, D. T., Lang, P. T., Pegg, S., Pettersen, E. T., Kuntz, I. D., Brooijmans, N., Rizzo, R. C. (2006). Development and validation of a modular, extensible docking program: DOCK 5. *Journal of Computer-Aided Molecular Design*, *20*, 601–609.

93. Goodsell, D. S., Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins*, *8*, 195–202.

94. Jones, G., Willet, P., Glen, R. C. (1995). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *Journal of Molecular Biology*, *245*, 43–53.

95. Rarey, M., Kramer, B., Lengauer, T., Klebe, G. (1996). A fast flexible docking method using an incremental construction algorithm. *Journal of Molecular Biology*, *261*, 470–489.

96. Miller, M. D., Kearsley, S. K., Underwood, D. J., Sheridan, R. P. (1994). FLOG: A system to select "quasi-flexible" ligands complementary to a receptor of known three-dimensional structure. *Journal of Computer-Aided Molecular Design*, *8*, 153–174.

97. Brooijmans, N., Kuntz, I. D. (2003). Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure*, *32*, 335–373.

98. Leach, A. R., Shoichet, B. K., Peishoff, C. E. (2006). Prediction of protein–ligand interactions. Docking und scoring: successes and gaps. *Journal of Medicinal Chemistry*, *49*, 5851–5855.

99. Carlson, H. A., McCammon, J. A. (2000). Accommodating protein flexibility in computational drug design. *Molecular Pharmacology*, *57*, 213–218.

100. Gohlke, H., Klebe, G. (2002). Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie, International Edition*, *41*, 2644–2676.

101. Kontoyianni, M., McClellan, L. M., Sokol, G. S. (2004). Evaluation of docking performance: comparative data on docking algorithms. *Journal of Medicinal Chemistry*, *47*, 558–565.

102. Bissantz, C., Folkers, G., Rognan, D. (2000). Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *Journal of Medicinal Chemistry*, *43*, 4759–4767.

103. Charifson, P. S., Corkery, J. J., Murcko, M. A., Walters, W. P. (1999). Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *Journal of Medicinal Chemistry*, *42*, 5100–5109.

104. Klebe, G. (2006). Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today*, *11*, 580–594.

105. Schreyer, S. K., Parker, C. N., Maggiora, G. M. (2004). Data shaving: a focused screening approach. *Journal of Chemical Information and Computer Sciences*, *44*, 470–479.

106. Xue, L., Bajorath, J. (1999). Distribution of molecular scaffolds and R-groups isolated from large compound databases. *Journal of Molecular Graphics and Modeling*, *5*, 97–102.

107. Hert, J., Willet, P., Wilton, D. J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A. (2005). Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. *Journal of Medicinal Chemistry*, *48*, 7049–7054.

108. Shanmugasundaram, V., Maggiora, G. M., Lajiness, M. S. (2004). Hit-directed nearest-neighbor searching. *Journal of Medicinal Chemistry*, *48*, 240–248.

109. Wei, D., Zhang, R., Du, Q. S., Gao, W. N., Li, Y., Gao, H., Wang, S. Q., Zhang, X., Li, A. X., Sirois, S., Chou, K. C. (2006). Anti-SARS drug screening by molecular docking. *Amino Acids*, *31*, 73–80.

110. Barreiro, G., Guimaraes, C. R. W., Tubert-Brohman, I., Lyons, T. M., Tirado-Rives, J., Jorgensen, W. L. (2007). Search for non-nucleoside inhibitors of HIV-1 reverse transcriptase using chemical similarity, molecular docking, and MM-GB/SA scoring. *Journal of Chemical Information and Modeling*, *47*, 2416–2428.

111. Tikhonova, I. G., Sum, C. S., Neumann, S., Engel, S., Raaka, B. M., Costanzi, S., Gershengorn, M. C. (2008). Discovery of novel agonists and antagonists of the free fatty acid receptor 1 (FFAR1) using virtual screening. *Journal of Medicinal Chemistry*, *51*, 625–633.

112. Ginn, C. M. R., Turner, D. B., Willett, P., Ferguson, A. M., Heritage, T. W. (1997). Similarity searching in files of three-dimensional chemical structures: evaluation of the EVA descriptor and combination of rankings using data fusion. *Journal of Chemical Information and Computer Sciences*, *37*, 23–37.

113. Tan, L., Geppert, H., Sisay, M. T., Gütschow, M., Bajorath, J. (2008). Integrating structure- and ligand-based virtual screening: comparison of individual, parallel, and fused molecular docking and similarity search calculations on multiple targets. *ChemMedChem*, *3*, 1566–1571.

114. Xu, Y.-J., Johnson, M. (2002). Using molecular equivalent numbers to visually explore structural features that distinguish chemical libraries. *Journal of Chemical Information and Computer Sciences*, *42*, 912–926.

115. Terstappen, G. C., Reggiani, A. (2001). In silico research in drug discovery. *TRENDS in Pharmacological Sciences*, *22*, 23–26.

116. Bajorath, J. (2002). Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery*, *1*, 882–894.