Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation

Hanna Geppert, Martin Vogt, and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

Received October 29, 2009

1. INTRODUCTORY REMARKS

Data mining methods play a major role in chemoinformatics. In this Perspective, we focus on data mining methodologies that are particularly relevant for ligand-based virtual screening. In order to provide an up-to-date view of the field, we largely concentrate on publications of the past two to three years. We discuss alternative chemical space representations for compound screening, popular data mining methods, and new algorithms. Furthermore, recently developed molecular fingerprints and specialized similarity measures are reviewed. As increasing amounts of public-domain compound bioactivity data become available, data mining approaches are also utilized for new types of virtual screening applications, for example, to search for target-selective molecules or ligands of orphan targets. Moreover, how to best evaluate and compare the performance of different computational screening methodologies has emerged as one of the central questions in chemical data mining. Therefore, method evaluation strategies and approaches for the design of advanced benchmark data sets are also described. Taken together, the survey presented herein makes it possible to highlight a number of trends that can currently be observed in the chemoinformatics field. Furthermore, we also discuss problems associated with conventional benchmark settings for the evaluation of virtual screening methods and the need for community-wide standards.

2. LIGAND-BASED VIRTUAL SCREENING

The search for active compounds using computational methods is a central theme in chemoinformatics. Ligand-based virtual screening extrapolates from known active compounds utilized as input information and aims at identifying structurally diverse compounds having similar bioactivity, regardless of the methods that are applied. While practical applications are frequently reported, much of the virtual screening literature focuses on the evaluation and comparison of methods in benchmark settings, where recall of known active compounds added to background databases is typically considered the major measure of method performance. Yet, to this date, the virtual screening field lacks generally accepted standards for method evaluation. Furthermore, the potential of virtual screening methods indicated in benchmark calculations does not correspond to their performance in practical applications, where the identification of only one or of a few (weakly) active compounds that structurally depart from available reference molecules is usually already considered a success. Thus, despite their widespread use, there are still many open questions concerning the relative performance of different virtual screening approaches and their application potential. Therefore, it is often difficult to judge about method performance, on the basis of the original literature, which also affects a review of the field. However, we also observe a steadily increasing knowledge base of small molecule bioactivity data that can be utilized to train and test virtual screening methods, which provides unprecedented opportunities for the field. Consequently, in recent years, machine learning and data mining methodologies have been increasingly applied to identify active compounds and have become a viable alternative to conventional structure–activity relationship analysis, compound classification, and similarity search methods. In light of this situation, we have surveyed data mining approaches in the context of virtual compound screening. In our analysis, we have paid particular attention to novel algorithms and methods that have evolved into widely applied standards for chemical database mining and that continue to be further developed. Data mining methods generally require chemical reference space representations for analyzing compound sets. Therefore, we initially evaluate current developments in chemical space design.

3. CHEMICAL SPACE REPRESENTATIONS

Computational chemical space representations usually do not aim to accurately reflect the chemical universe but rather provide reference frames for the projection or design of compound data sets of finite size. Chemical reference spaces are typically generated as structural and/or physicochemical feature spaces. Designing and navigating chemical space representations and understanding their topology are important topics in chemoinformatics research. In computational chemical space, compounds can be represented and compared in different ways, for example, by a numerical or binary vector in (possibly very large) *n*-dimensional space.[1] In order to study their relationships, a similarity or a distance metric must be applied to the vector representation.[1] Alternatively,

* Corresponding author. Telephone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

10.1021/ci900419k © 2010 American Chemical Society
Published on Web 01/20/2010

compound similarity might be assessed on the basis of two-dimensional graphs,[2–4] three-dimensional conformations,[5] or through the application of kernel functions.[6] To this date, no chemical space representations have been reported that are generally applicable in chemoinformatics, and the model character of space designs and the underlying approximations are often not considered. In fact, crucial aspects of chemical space design have not been debated much in the literature over the past few years, although the majority of chemoinformatics applications depend on reference spaces. However, recently, there has again been increasing interest in a systematic analysis of chemical space representations that are based on numerical molecular property descriptors or fingerprints.[7,8] Central questions of these investigations include: (i) how well do compound data sets cover biologically relevant sections of chemical space? and (ii) how predictive and relevant are reference spaces utilizing numerical representations of molecules? In most instances, chemical spaces are only of interest if distance relationships between compounds not only reflect molecular but also biological similarity. Hence, chosen descriptors and other molecular representations must be relevant for and respond to different biological activities.

Singh et al.[8] compared four in-house generated combinatorial libraries with known drugs in different space representations. The R-NN curve technique[9] and the descriptor principal component analysis were applied to characterize the population density of chemical spaces and to analyze compound similarity and scaffold diversity using different fingerprint representations. Using known drugs as references, it was assessed how well the compound libraries covered both densely and sparsely populated regions of drug-relevant chemical space and also of uncharted areas.

Fingerprints are a descriptor format that assigns (binary) vector representations to test compounds and, hence, to represent (simplified) reference spaces. Bender et al.[7] reported a systematic analysis of 37 alternative fingerprint representations for their ability to detect various biologically active compounds using standard similarity searching techniques. The study focused on differences in the rank order of active compounds between alternative similarity searches. Using principal component analysis, the alternative fingerprint representations were mapped to a three-dimensional space, revealing a separation of representations into four clusters: (i) circular, (ii) circular count, (iii) three-dimensional pharmacophore, and (iv) connectivity path-based fingerprints. This cluster distribution largely determined which classes of compounds were recovered in a similarity search trial and, to a lesser extent, determined how well the methods performed. Bender et al. carried out simple similarity searching using single reference compounds and calculated Tanimoto and cosine coefficient similarity. In their analysis, binary fingerprints and their count fingerprint analogs displayed different potentials to recover active compounds. In count fingerprints, not only the presence or absence of features in a molecule is recorded but also the feature frequency. Count fingerprints have been increasingly investigated as an alternative to binary fingerprints, especially in combination with Bayesian classification schemes,[10–12] but systematic performance increases over binary fingerprints have not been demonstrated.

The results of studies like the ones discussed above are intrinsically dependent on the data sets that are analyzed and on the applied search strategies. Hence, it is often difficult to draw general conclusions from such studies. However, the findings of Singh et al.[8] and Bender et al.[7] support the view that fingerprints of different design capture complementary and target-dependent structure−activity information.

The often observed target class-dependent predictive performance of molecular representations and search methods has triggered the development of approaches to combine different types of chemical representations in order to improve virtual screening performance. A straightforward way is combining the rank-ordered lists of compounds retrieved in virtual screens employing different molecular representations using data fusion techniques, such as rank averaging or maximum rank scoring.[13–15] Furthermore, full integration of different molecular representations and complementary information into virtual screening has also been attempted, and methods from machine learning and statistics have been adapted for this purpose.[16–18] For example, Simmons et al.[16] have utilized ensemble machine learning methods in order to combine different models for compound ranking. Combinations of individual machine learning models based on different fingerprint representations were found to substantially increase compound recall in virtual screening benchmark calculations. As an alternative to ensemble methods, a probabilistic framework using Bayesian statistics was introduced[17] to combine numerical molecular property descriptors with structural fragment fingerprints. For a large number of compound activity classes, the combination of these different types of descriptors boosted compound recall in Bayesian screening calculations. Moreover, it has been demonstrated that "recombination" of selected features from different fingerprint designs leads to increased compound recall compared to the original fingerprints.[18] A statistical analysis was carried out of the significance of each bit position in a fingerprint to discriminate between active and inactive compounds on a class-by-class basis, and the most discriminating positions were combined into activity class specific "hybrid fingerprints".[18] Thus, generally applicable fingerprints and reference spaces were transformed into compound class-directed space representations. Such efforts also illustrate the variability of reference space design, which provides many opportunities for further exploration. It is likely that generally applicable chemical space representations will be difficult to obtain. Hence, the generation of reference spaces tailored toward specific compound classes or characteristics provides a logical alternative. The limited number of relevant studies published over the past few years in this area indicates that the rational design and exploration of chemical reference spaces is still an underrepresented yet critically important topic in chemoinformatics research.

4. DATA MINING METHODS

In recent years virtual screening has evolved from traditional similarity searching, using single reference compounds, into an advanced application domain for data mining and machine learning methods that require compound reference sets of increasing size and of high-information content for training. As already stated above, this is, at least, in part a consequence of increasing amounts of publicly

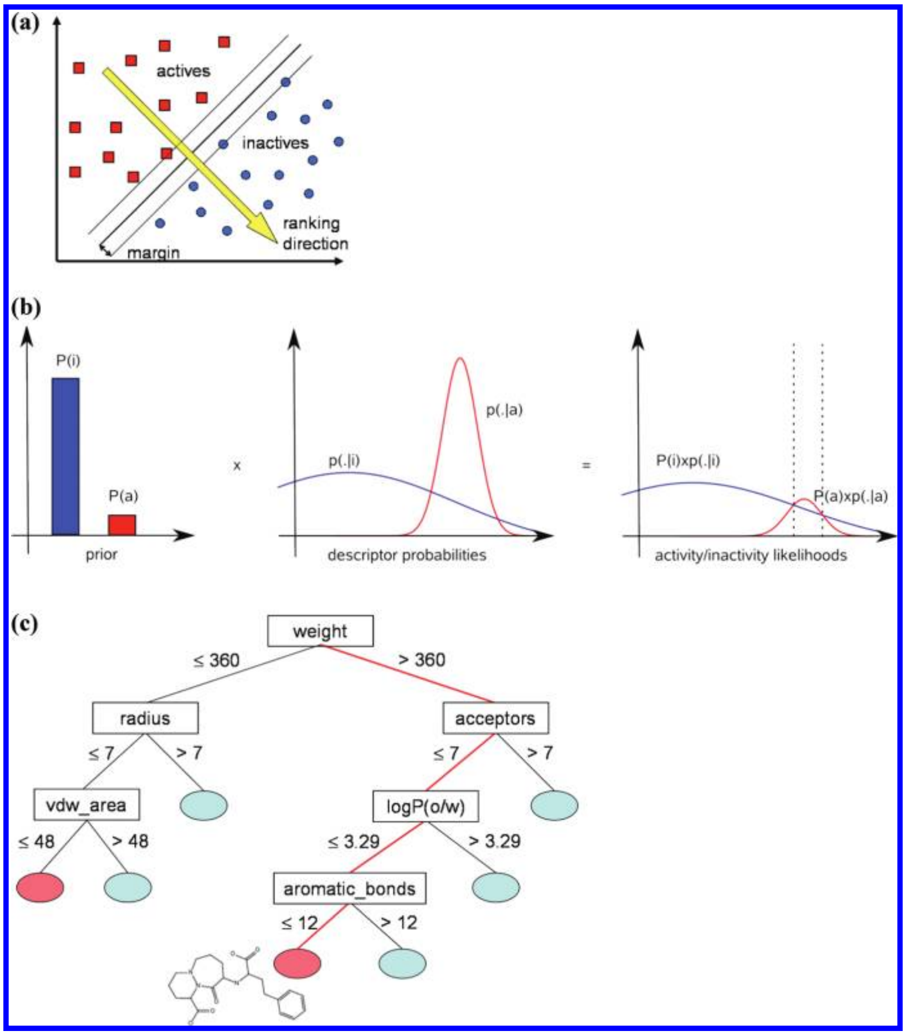

Figure 1. Popular data mining algorithms. (a) Support vector machines. Following the simple SVM-based ranking strategy, active and inactive compounds are separated by the maximum-margin hyperplane. Screening database compounds are ranked by their signed distance from that hyperplane (indicated by the arrow). (b) Bayesian classifiers. On the basis of Bayes theorem, the prior probabilities for activity and inactivity, P(a) and P(i), respectively, and the conditional distributions of descriptor or feature values, $p(\cdot|\text{a})$ and $p(\cdot|\text{i})$, can be used to predict the likelihood of activity:

$$L(\text{a}|\cdot) = P(\text{a})p(\cdot|\text{a}) \quad \text{and} \quad L(\text{i}|\cdot) = P(\text{i})p(\cdot|\text{i})$$

The descriptor range indicated by the two dashed lines marks the region where the posterior probability for activity exceeds the posterior probability for inactivity. (c) Decision trees. In order to classify a database compound, a path is followed through the tree according to the compound's values of the descriptor assigned to each node until a classifying terminal node is reached.

available bioactivity data, which favors the application of these methodologies. Essentially, data mining provides the basis for the derivation of predictive models. While new concepts and methods are being introduced specifically for virtual screening applications, established data mining approaches have recently gained much in popularity in virtual screening and are currently widely applied. Of the different data mining methods that have been explored, three approaches have evolved to largely dominate the field at present. These methodologies include: (i) support vector machines (SVMs), (ii) Bayesian methods, and (iii) decision trees, as illustrated in Figure 1. We first discuss these popular approaches and then provide an overview of new data mining concepts developed or adapted for virtual screening.

4.1. Support Vector Machines. Compared to Bayesian statistics or decision trees, support vector machines represent a relatively new data mining methodology that has become popular during the 1990s based on the works of Vapnik and Cortes.[19] The SVM approach is based on the idea of constructing a hyperplane in a high-dimensional reference space in order to classify data. However, the idea of a separating hyperplane itself is not new and dates back to the perceptron concept of Rosenblatt.[20] The SVM methodology has a strong theoretical foundation in statistical learning theory[21] by taking into account two generally conflicting objectives in machine learning: (i) fit the training data well and ii) sufficiently generalize so that external test data can be classified. That is, the more complex the structure of a trained model is, the better it might fit the training data, but the less likely it is to generalize to unknown data (because complex models tend to simply "learn the training data"). In typical machine learning settings, the complexity of the model is controlled by the number of parameters learned by the method. Bayesian classification (vide infra), for instance, typically estimates only a few parameters that determine probability distributions. Decision trees (vide infra), on the other hand, can be readily constructed to perfectly classify the training data. However, in order for decision trees to sufficiently generalize, they must typically be "pruned" to reduce their complexity. Because SVMs operate in high-dimensional spaces, the construction of a hyperplane has many degrees of freedom, which might by far exceed the number of training data points. By constructing a maximum-margin hyperplane, i.e., a hyperplane that maximizes the distance to the nearest training data points, SVMs control the so-called "structural risk" of overfitting. However, overfitting might still occur when optimizing SVM performance with respect to "meta-parameters", such as the penalty factor for misclassification or for kernel functions. For a detailed account of SVM theory, the reader is referred to the work of Vapnik.[21]

Importantly, in SVM learning, the generalization ability is not dependent on the dimensionality of the data representation but only on the number of support vectors, which renders the approach capable of navigating high-dimensional (descriptor) spaces. Another important aspect of SVM classification is that high-dimensional descriptor spaces do not need to be explicitly formulated. Instead, it is sufficient to utilize a kernel function that correctly accounts for the degree of similarity between any two molecules (by measuring the degree of orthogonality as defined by the respective inner product). Alternative kernel functions have been explored for virtual screening. For example, Azencott et al.[6] have proposed different kernel functions based on one- to four-dimensional molecular representations (where four-dimensional refers to the application of a kernel function to multiple conformations of a compound). In addition, Rupp et al.[2] have developed a kernel function based on two-dimensional graph representations, which avoids the intermediate step of explicitly calculating a numerical or binary representation. In the context of quantitative structure−activity relationship (QSAR) analysis, Mohr et al.[22] have introduced a kernel function based on global three-dimensional similarity of molecules, hence, providing a "descriptor-free" QSAR modeling approach. Following the works of Warmuth et al.[23] and Jorissen et al.[24] (who probably first introduced SVMs to virtual screening), a variety of SVM-based screening studies have been reported. Gaussian kernels in combination with physicochemical property descriptors have been widely used for the classification of biologically active and inactive (or randomly selected) compounds.[25–28] In one of these studies, drug-like compounds were targeted using fingerprints as descriptors.[25] Geppert et al.[29] then investigated a variety of different fingerprint designs for SVM-based prioritization of compounds for different biological activity classes and found increased performance over standard similarity search methods. Here, linear kernels performed very well, thus, demonstrating that SVMs are capable of handling the high-dimensional data representations produced by fingerprints without the need for advanced kernel functions, even if small training sets of only 5 active and 14 inactive compounds were used. Prioritization of active compounds was achieved by ranking them according to their signed distance from the hyperplane, as illustrated in Figure 1a.

SVMs have also opened up a new avenue in virtual screening by utilizing not only compound but also target-ligand information for training. For example, target information might simply be added as sequence data. The resulting models can then be used to search for ligands of novel targets,[30–32] a process often referred to as "orphan target screening", which will be further discussed in Section 6.3.

Although SVM-based classification has originally been binary in nature (e.g., active versus inactive), SVMs have also been adapted for multicategory classification problems, for instance, by Kawai et al.,[33] who used multiple SVM models derived from topological fragment spectra for the generation of compound activity profiles. In addition, different SVM multiclass systems have been investigated for the detection of target-selective molecules (see the Applications Section).[34]

4.2. Bayesian Methods. Bayes theorem and statistics have laid the foundation for Bayesian modeling methods (Figure 1b). Despite algorithmic differences, Bayesian methods utilized in virtual screening generally have in common that they derive a probability of compound activity. Because Bayesian classifiers typically yield numerical estimates for likelihoods of activity, they are well suited to rank compound databases with respect to these probabilities. Bayesian methods generally rely on the estimation of probability distributions of numerical representations of compounds based on property descriptors or fingerprints. Popular methods include naïve Bayesian classifiers[11,12] and binary kernel discrimination,[35,36] which can be applied to both binary and count fingerprints.

Extensions of Bayesian classification have also been reported. For example, Kullback−Leibler (KL) divergence analysis from information theory has been combined with Bayesian compound screening as a statistical measure for the ability of property descriptors or fingerprint features to discriminate between active and inactive compounds.[37] The KL divergence between an activity class and a screening database was shown to scale with the recall of active compounds in Bayesian screening using property descriptors[37] or fingerprints.[38] This has made it possible to predict the recall rates for different compound classes. Furthermore, on the basis of KL divergence analysis, it was also possible to select molecular representations for different activity classes that maximized compound recall in virtual screening[39] and identify small subsets of fingerprint features that largely determined the recall rates.[40,41] Muchmore et al.[42] have applied a different strategy, "belief theory", to predict virtual screening performance by creating probability assignment curves that depend on the pairwise similarity of active and database compounds calculated using different fingerprints and similarity measures. Similarity values are converted to a probability of activity, which makes it possible to fuse different compound rankings by computing the "joint belief" and which yields a quantitative estimate for the probability of activity. Predicting virtual screening performance on the basis of KL divergence analysis involves the identification of activity class-specific feature representations that best distinguish between active and inactive compounds. In contrast, by applying belief theory, no discriminatory compound representations are identified on a per-class basis, but similarity value distributions for a variety of representations are transformed into probabilities and subjected to data fusion, thereby yielding a general estimate of activity.

4.3. Decision Trees. Among conventional data mining methods, decision trees (Figure 1c) continue to be applied in chemoinformatics, especially in the form of ensemble-based random forest models, which are utilized, for example, in the context of QSAR analysis,[43] for prediction of aqueous solubility[44] or mutagenicity,[45] and in virtual screening.[46] Furthermore, Zhou et al.[43] introduced "modified particle swarm optimization" for the classification of regression trees in QSAR analysis. Modified particle swarm optimization is a stochastic optimization technique adapted for regression trees to account for the discrete structures of the trees. Regression trees and Bayesian modeling have also been combined. Angelopoulos et al.[47] have integrated the classification and regression trees (CART) methodology with Markov chain Monte Carlo simulations to sample CART "space". Then, Bayesian averaging was applied in order to derive a virtual screening model on the basis of sparse activity data. Currently, fewer regression tree models appear in the literature than SVM or Bayesian classifiers.

4.4. New Methodologies. Other machine learning and data mining methods, originally developed in computer science, have recently also been adapted for virtual screening. For example, the "Winnow" algorithm derives a linear classifier (i.e., a separating hyperplane) for high-dimensional data spaces and is particularly suitable for data distributions where dimensions are redundant or irrelevant (hence the name), which corresponds to a typical dimension reduction problem. Therefore, the Winnow algorithm is well suited for the processing of complex (high-dimensional) fingerprints. It has been applied for compound classification and for multiclass ligand target predictions using Molprint2D[48,49] and extended connectivity fingerprints[50] and has displayed a classification performance comparable to Bayesian methods.[51,52]

"Influence relevance voting" might be regarded as an extension of *k*-nearest-neighbor methods, where the influence of the neighbors on the classification of a test compound is determined by a neural network. Swamidass et al.[53] successfully applied this approach to virtual compound screening and proposed chemical interpretability of the model as an advantage over alternative machine learning methods, such as SVMs.

"Inductive logic programming"[54] has also been evaluated for the recognition of structurally diverse compounds having similar activity (i.e., scaffold hopping).[55,56] The method derives a number of interpretable structural rules from active and inactive training compounds. Given a set of rules, a scaffold is represented as a binary fingerprint, where each bit position indicates whether an individual rule is satisfied or not. These fingerprints are then subjected to standard similarity search methods or to SVM classification. Tsunoyama et al. have compared this approach to other methods in searching for structurally diverse compounds and reported comparable performance and identification of scaffolds not detected by other methods.[56]

5. SIMILARITY SEARCHING USING FINGERPRINTS

As described above, two-dimensional fingerprints are among the most popular molecular representations for ligand-based virtual screening. In addition, fingerprint similarity searching continues to be a widely applied approach. However, the use of fingerprint types is increasingly shifting from intuitive structural fingerprints, such as dictionaries of structural keys,[57,58] to "combinatorial" fingerprints calculated from the molecular graphs. These fingerprints monitor, for example, systematically derived circular atom environments,[48–50] atom paths,[50,59] or typed graph distances or triangles.[60] Such fingerprints are implemented in the widely used chemoinformatics platforms Pipeline Pilot[50] and Molecular Operating Environment,[60] which might also contribute to their popularity. Because many combinatorial fingerprints rely on feature sets of, in part, very large size, individual features are often not assigned to specific bit positions, as in structural key-type fingerprints but are string-coded[48,49] or mapped to bit segments using hash functions.[50,59] In addition, fingerprint size can be reduced using simple fingerprint compression schemes, such as modulo-based folding, with the drawback that this might have negative effects on compound recall. Therefore, the Baldi group investigated modifications of similarity measures to better account for similarity of uncompressed fingerprints on the basis of compressed representations.[61] In addition, an alternative compression scheme based on statistical fingerprint models and integer coding techniques was introduced that improved compound recall over other compression techniques.[62]

Although fingerprint similarity searching is a computationally efficient approach, several recent investigations have attempted to further improve search efficiency, given the large numbers of features encoded in combinatorial fingerprints and the large size of current screening databases. For example, mathematical bounds were derived for standard similarity measures, such as the Tanimoto or Tversky coefficient,[63] to prune search space and, hence, reduce search time.[64–66] These bounds usually make use of compact fingerprint signatures that are stored as small header vectors with each fingerprint. The header vectors contain precomputed fingerprint information, such as the overall number of 1-bits,[64] the number of 1-bits within modulo-derived fingerprint components,[65] or logical XOR-derived values.[66] Another approach to accelerate similarity searching makes use of a new data structure, the compressed binary bit tree, as opposed to bit strings, in combination with a pruning technique.[67]

The current popularity of combinatorial fingerprints with large feature sets has triggered recent investigations to analyze which fingerprint features are responsible for the identification of active compounds, whether activity class-specific features exist, and how they might be selected. With the introduction of the Molprint2D fingerprint,[48] Bender et al.[49] proposed an activity-oriented feature selection based on the information gain measure of Quinlan,[68] and it was found that approximately 40 selected features produced the highest enrichment factors for active compounds. For the conceptually similar extended connectivity fingerprints, feature space was reduced to subsets of features present in active reference compounds.[69,70] Applying a simple feature counting strategy instead of conventional similarity coefficients maximized both the compound recall and the structural diversity of active compounds.[69,70] In a substructure-based approach, a feature space was generated through random fragmentation of active and inactive reference molecules, and a hierarchy that captured conditional probabilities of fragment co-occurrence was utilized to select combinations of activity−class characteristic substructures.[71] Between 10 and 30 of these substructures were then used as compound class-directed "mini-fingerprints" for similarity searching and performed as well or better than standard fingerprints.[71,72] Kullback−Leibler divergence analysis has been applied to analyze feature distributions in active and database compounds and to select those fingerprint features that determine the search performance on given activity classes.[41] This type of "fingerprint anatomy" is applicable to any fingerprint design. Often, only small subsets of bit positions were found to make the largest contributions to compound recall rates.[41]

Feature selection is strongly related (and sometimes equivalent) to emphasizing fingerprint bit positions during similarity searching through activity class-oriented feature weights. Traditionally, such weights were derived on the basis of feature frequency analysis in active and inactive molecules,[73] for example, through the calculation of consensus fingerprints,[74] through fingerprint profiling and scaling,[75,76] or through fingerprint averaging.[77] A recently introduced "bit silencing" approach differs from these strategies by directly monitoring the change in compound recall when omitting individual features from fingerprints of active reference molecules.[78] From the resulting recall rate profile, another form of fingerprint anatomy, an activity-oriented weight vector, can be derived and incorporated in conventional Tanimoto coefficient calculations.[78]

Other recently introduced feature weighting schemes that are applicable to two-dimensional structural fingerprints make indirect use of three-dimensional ligand−target interaction information.[79,80] For example, a pool of extended connectivity fingerprint features was assembled from ligands in complex crystal structures. For each feature, a weight was derived on the basis of force field energy scores, reflecting atomic contributions to interaction energies computed for crystallographic complexes. As a similarity value for a database compound relative to an X-ray ligand, its fingerprint features present in the pool were determined, and their weights were summed.[79] Following a different approach, interacting fragments were extracted from X-ray structures of multiple protein−ligand complexes and encoded as structural keys. From these fingerprints, scaling factors were derived by conventional bit frequency analysis and applied to noncrystallographic reference compounds.[80]

Thus, taken together, recent research activities focusing on fingerprint representations have mainly concentrated on increasing computational efficiency, on feature selection and anatomy, or on similarity evaluation of combinatorial fingerprints, rather than that of the design of new two- or three-dimensional fingerprint types.

6. APPLICATIONS

6.1. Scaffold Hopping. The identification of different chemotypes having comparable activity continues to be the major task in ligand-based virtual screening, which has much influenced method development in the past.[81,82] The exploration of scaffold hopping potential continues to be an active area of research, with increasing emphasis on data mining approaches. Wale et al.[83] have attempted to improve the scaffold hopping potential of conventional similarity searching by utilizing "indirect" compound similarities. These indirect similarities were obtained via the analysis of networks formed by a *k*-nearest-neighbor graph representation of the reference and the database compounds. In addition, as already mentioned above, inductive logic programming has been adapted as a scaffold hopping tool by learning from structural examples encoded as logical relationships.[56] Furthermore, a tool termed CORUS has been developed[84] to postprocess top-ranked compounds from similarity searching by splitting compounds into fragments that are subsequently annotated. Using simple filter rules on these annotations, false-positive scaffold hops can be eliminated, and molecules of interest identified through interactive visualization.

A fundamental problem associated with evaluating scaffold hopping analyses is that they are often not comparable, similar to many virtual screening benchmark studies. First, the definition of what constitutes a scaffold hop is highly subjective and often differs, and there currently is no accepted metric available for the evaluation of the scaffold hopping potential. For example, in the three recent publications discussed above, three different scaffold hopping definitions are found. In the work of Wale et al.,[83] active database molecules are ranked according to their (path-based) fingerprint similarity to the query molecule, and the lowest-ranked 50% of them are regarded as scaffold hops relative to the query. Thus, in this case, the decision whether a compound transition is considered a scaffold hop or not depends on the composition of the active compound test set. By contrast, Tsunoyama et al.[56] have clustered active compounds using path-based fingerprints. From each cluster, one compound was selected, and compounds from different clusters were considered to represent different scaffolds. Finally, Senger[84] applied a scaffold hopping definition based on core structural differences rather than that of whole-molecule similarity values. Regardless of which definition one might prefer in this context, it would be very difficult to directly compare the results and estimate the relative method performance.

The question of how to best quantify the recall of different scaffolds in virtual screening studies has been investigated by Mackey and Melville.[85] As a basis for their theoretical considerations, it was assumed that active compounds were partitioned into disjoint clusters such that each cluster represented a separate scaffold. This was done because such "clustered" virtual screening data are often utilized in scaffold hopping analyses. This investigation demonstrated that the "first found" strategy, where only the first detected active of a scaffold cluster is taken into account for scoring, has significant drawbacks when assessed with quantitative compound recall metrics.[85] By contrast, the so-called "cluster averaging", where the contribution of each active compound to the score is proportional to the number of compounds in its scaffold cluster, was strongly preferred because it avoids the pitfalls of extreme value statistics and is not influenced by factors such as cluster size and cluster number.[85]

In conclusion, to further advance research activities directed at scaffold hopping and to make the performance of different methods comparable, there is a need to establish generally applicable scaffold definitions and retrieval metrics.

6.2. Data Mining for Selective Compounds. Over the past few years, evolving interdisciplinary research fields, such as chemogenomics and chemical biology, have opened new application areas for data mining methods and for virtual compound screening.[86] In pharmaceutical research, virtual screening has traditionally been applied as a hit identification tool. For applications in chemical biology, compounds having the potential to become drug leads are not necessarily required. Here, small molecules are primarily utilized as probes for biological functions, and hence, it is desirable to identify compounds that are selective against individual targets or subfamilies within target families. Because the identification of such molecular probes typically requires large-scale biological compound screening, the ability to predict selective compounds through database mining is currently of increasing interest.[86] In order to support the development of computational approaches for chemical biology applications, publicly available molecular benchmark systems have been designed consisting of compound selectivity sets for different pairs of closely related target proteins, including both target-selective and active but nonselective compounds.[87,88] On these data sets, conventional two-dimensional similarity searching was found to enrich database selection sets with target-selective molecules.[88,89] Building upon these studies, various SVM-based strategies were designed to further advance the search for target-selective compounds.[34] Different from conventional binary SVM classification, SVMs were adapted for three class predictions and for compound ranking to distinguish between selective, active but nonselective, and inactive compounds. SVM-based methods further improved the performance of "selectivity searching" using fingerprints[89] by effectively removing nonselective molecules from high-ranking positions, while retaining the recall of selective compounds.[34] Going beyond selectivity analysis on the basis of binary target relationships, Lounkine et al.[90] introduced molecular formal concept analysis (MolFCA), a variant of formal concept analysis from information theory,[91] to systematically compare the selectivity of compounds against multiple targets. This approach enabled the mining of complex compound selectivity profiles in public domain databases.[90]

6.3. Orphan Screening. Another computational task with high relevance for both chemoinformatics and chemical biology is the prediction of ligands for novel members of protein families or for proteins for which no ligand information is available, so-called "orphan screening". Conventional de novo ligand design methods require target structure and/or pharmacophore information[92,93] and are as such only applicable to already well-characterized targets. Schuffenhauer et al.[77] introduced a conceptually different approach to predict the ligands for orphan targets that only required target sequence information and the availability of ligands for targets homologous to an orphan. Then, known ligands of homologous targets were used as reference molecules for "homology-based similarity searching".[77] The idea underlying this approach is that homologous proteins have similar binding sites and, hence, bind structurally similar ligands. Recently, multitask learning using SVMs has been successfully applied to orphan screening in benchmark calculations[31,32,94] and was found to further increase the performance of homology-based similarity searching.[32]

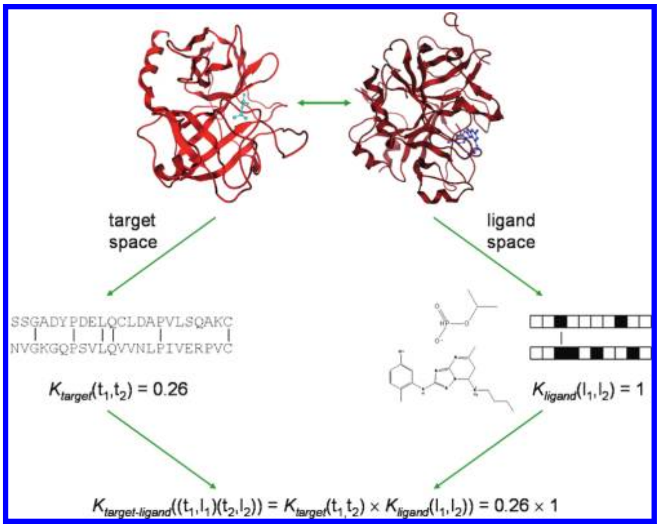
target
space
ligand
space
SSGADYPDELQCLDAPVLSQAKC
NVGKGQPSVLQVVNLPIVERPVC
$K_{target}(t_1,t_2) = 0.26$
$K_{ligand}(l_1,l_2) = 1$
$K_{target\text{-}ligand}((t_1,l_1)(t_2,l_2)) = K_{target}(t_1,t_2) \times K_{ligand}(l_1,l_2)) = 0.26 \times 1$

Figure 2. Target-ligand kernels. In order to determine the distance between two different target-ligand pairs, a target-ligand kernel function is applied that can be split into a separate target kernel and a ligand kernel component. In this case, the distance between protein targets is measured by sequence comparison, while ligand distance is assessed using the dot product of their fingerprint representations. The product of the separate target and ligand kernel recombines target and ligand information.

The basic idea of multitask learning is to train a classifier on related problems in parallel and to benefit from the commonalities among the learning tasks. By doing so, the generalization performance of an algorithm can be improved as a consequence of inductive transfer, using a common representation of the different learning tasks. For the purpose of orphan screening, multitask learning must consider multiple target classes in parallel, using true/false target-ligand pairs as positive/negative training examples. Using SVMs as a kernel-based machine learning algorithm, distances between different target-ligand pairs in the target-ligand feature space can be elegantly determined by applying separate kernel functions for pairs of proteins and ligands,[30,31] as illustrated in Figure 2. For SVM applications, two-dimensional fingerprints have been found to be particularly effective ligand descriptors,[29] and conventional kernel functions like the linear, Tanimoto or Gaussian kernel accurately determine fingerprint similarity.[94] In order to quantitatively measure target similarities, in part very different target representations and protein kernels have been investigated for orphan screening. Erhan et al.[30] used "binding pocket fingerprints" for which standard kernel functions were applied. Jakob and Vert[31] introduced a hierarchy kernel for proteins based on protein classification schemes, such as the EC numbers or the KEGG database hierarchy[95] that was shown to outperform the Dirac and multitask kernels, and Geppert et al.[32] applied pairwise sequence identity values as the target kernel. Finally, Wassermann et al.[94] systematically investigated the question of how different target kernels influence SVM-based ligand prediction. In their study, the performance of 11 different target kernels was compared that captured protein information in rather different ways (including sequence, secondary and tertiary structures, biophysical amino acid properties, ontologies, or structural taxonomy). An unexpected result has been that alternative kernel functions, despite different complexity and information content, had only a minor effect on ligand prediction performance. Rather, the success rates of orphan screening were determined by the availability of ligand information from nearest neighbors of orphan targets.[94] Thus, ligand prediction was ultimately determined by related ligand information, and consequently, learning conditions for orphan screening might be substantially simplified. Orphan screening also requires the availability of carefully assembled ligand/target systems for benchmarking. Accordingly, the assessment of these advanced computational screening approaches is as much affected by the specifics of benchmark settings as the conventional virtual compound screening.

7. BENCHMARKING, PERFORMANCE EVALUATION, AND COMMUNITY STANDARDS

The chemoinformatics and virtual screening literature continues to be dominated by benchmark studies. From the multitude of data mining and machine learning methods developed in computer science, new methodologies are constantly adapted for compound classification and for virtual screening and are compared to established methodologies in, often more or less, unique benchmark settings, for example, as discussed above for scaffold hopping analyses. In light of this situation, a major problem is that method comparisons are often not reproducible, and that it is difficult to judge the relative method performance on the basis of the original literature. In the past few years, the chemoinformatics community has become increasingly aware of these problems, and the first community-wide initiatives can be observed. Recently, the Journal of Computer-Aided Molecular Design devoted a special issue to the topic of evaluation of computational methods.[96] In this issue, Jain and Nicholls note that "a serious weakness within the field is a lack of standards with respect to statistical evaluation of methods, data set preparation, and data set sharing".[96] Furthermore, Nicholls described the situation that standard procedures for benchmark evaluation of data mining methods, which are common in other areas of science, are not or are rarely followed in reports on virtual screening or, more generally, molecular modeling. As a consequence, benchmark studies often become "anecdotal instead of systematic".[97] Key issues include the study and calculation design, the data set preparation, the performance metrics, and the variance analysis, as discussed in the following, Sections 7.1−7.3.

7.1. Study Design and Data Set Preparation. In reviewing the literature, the observation can be made that many benchmarking or molecular design exercises do not clearly state the goal of the study. Is it all about which method is supposed to be "better" than others? Is there a hypothesis that should be investigated? What should we expect to learn? Furthermore, the system setup is often not sufficiently described or is even arbitrary. In virtual screening, this is often reflected by the way background databases of putative "decoys" are assembled.[97] Often inactivity of background database compounds can only be assumed as the data is not available. However, more serious complications are due to the use of nondrug or lead-like decoys that are easily separated from activity classes, consisting of highly optimized compounds. In such cases, even simple approaches, such as molecular weight-based screens, might result in significant enrichments of active compounds in database selection sets,[98] thus, rendering benchmark conditions highly artificial. Therefore, virtual screening results using new methodologies might be normalized relative to structure-unaware descriptors, such as molecular weight and simple atom counts.[98] Community dedicated activities to help standardize decoy sets were initiated by pioneering the development of the Shoichet group, including the generation of the ZINC database[99] and the Directory of Useful Decoys[100] (DUD) for docking and for other virtual screening applications. These databases have already become widely applied benchmark sets.

The assembly of compound activity classes (that are typically divided into reference molecules and potential database hits) is as important (if not more important) for assessing virtual screening performance as is the choice of decoys. Expressions, such as "artificial enrichment" or "analogue bias", have become recurrent in the literature to point at virtual screening artifacts associated with certain activity classes, as analyzed, for example, in a series of papers by Good and others.[101−103] In ligand-based virtual screening, standard activity classes that are utilized throughout the community are not available to this date, with a few exceptions. For example, 11 activity classes originally reported by Hert and Willet[104,105] are frequently utilized, a caveat being that these sets were assembled from the license-protected MDL/Symyx Drug Data Report.[106] Furthermore, these sets have not been filtered for close analogs and are, thus, susceptible to analogue bias.

The effect of the analogue bias and the ensuing artificial enrichment problem (Figure 3a−c) was recently analyzed in detail by Rohrer and Baumann,[107] who applied spatial statistics to analyze the topology of sets of active compounds and of quantitatively related topological features to that of virtual screening performance. Based on their analysis, Rohrer and Baumann developed publicly available compound data sets termed "maximum unbiased validation data sets" (MUV), another pioneering effort for virtual screening.

By utilizing a "simple" descriptor space based on simple atom counts, benchmark sets of active and inactive compounds were composed based on the premise that active−active distances in the descriptor space should be at least as large as active-decoy distances. All bioactivity data were taken from PubChem,[108] hence, ensuring accessibility. In addition, decoys were confirmed to be inactive (at least within the error margins of high-throughput screening). Filters were also applied to remove "frequent hitters" and other potentially problematic compounds. The resulting MUV data sets contain 17 activity classes with 30 active and 15 000 inactive compounds each.

Recently, the MUV data sets have been applied in a benchmark study comparing two pharmacophore elucidation tools and five similarity search methods.[109] The authors have regarded the performance of the similarity search methods as "disappointing", stating as potential reasons the presence of activity cliffs, the variable binding modes, and the false negatives, and suggest further improvement of these benchmark sets. However, in this context "disappointing" performance might also be rationalized as more "realistic" performance, similar to what is often observed in practical virtual screening publications. Hence, at the very least, such investigations begin to move the field into a better direction.

Analogue bias is not the only chemical feature that might lead to artificial benchmarking performance. The underlying problem might, in fact, be more general. For machine learning, "inductive bias" is a prerequisite for successful classification. For example, in chemoinformatics applications, inductive bias is provided by the conjecture that chemically similar molecules often have similar activity. This relationship is reflected in compound data sets because molecules are generally not made in an "unbiased" manner but rather to be similar to compounds or chemotypes already known to be active.[110] Hence, the experience and the intuition of synthetic and medicinal chemists propagate and lead to the preferential generation of a compound series that is not random samples of chemical space but rather directed toward

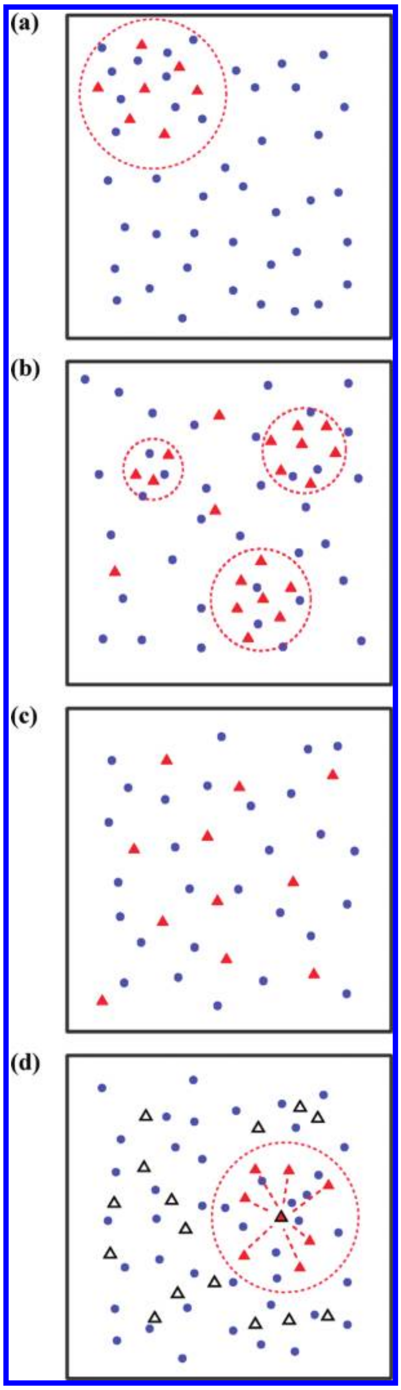

Figure 3. Analogue bias, inductive bias, and artificial enrichment. Shown is a schematic representation of the distribution of sets of active compounds in simple chemical space representations. (a) If the screening library does not reflect the chemical character of active compounds, as assessed using molecular descriptors, then the actives will only occupy a small fraction of the chemical space that is covered by the library. This makes it easy to separate active from inactive compounds, leading to "artificial enrichment" in virtual screens. (b) If active compounds consist of an analogue series, then the virtual screening using subsets of these analogs as reference compounds will easily detect remaining analogs. This phenomenon is known as "analogue bias". (c) Appropriate compound data sets for virtual screening that mimic "real life" application scenarios should be composed of diverse active compounds that are widely distributed over the space covered by the screening library. (d) Inductive bias means that of possible chemical space (blue dots) and of potentially available biologically relevant chemical space (black triangles), one generally only explores regions already known to be activity relevant (black/red triangle) and chemically extrapolates from such regions (red triangles).

biologically relevant chemical space, as illustrated in Figure 3. This situation generally plays into the strength of computational approaches that are designed to exploit such relationships.

7.2. Performance Metrics. There also is an ongoing discussion in the literature about proper "figures of merit" (or performance metrics) to evaluate retrospective virtual screening trials. A popular metric is the "enrichment factor", which is intuitive and straightforward to interpret. A problem associated with the calculation of simple enrichment factors is the dependence on a chosen cutoff value, typically 1 or 5% of the screening database. Nicholls[97] strongly advocates the use of standard measures, including the receiver−operator characteristic[111] (ROC) and the area under the ROC[111] (AUC), which are commonly applied in other fields employing statistical analysis, data mining, or machine learning techniques. However, AUC does not explicitly take into account the so-called "early recognition problem", i.e., the property of a method to retrieve active compounds "early", i.e., at the top of the ranking. Therefore, Truchon and Bayly[112] developed the Boltzmann-enhanced discrimination of ROC (BedROC) metric, which uses an exponential weighting to assign higher weights to early recognition. This metric is essentially a normalized version of the robust initial enhancement (RIE) measure.[113] Similarly, semilogarithmic scaling of ROC has also been suggested.[114] However, Nicholls also presents evidence for a strong correlation between AUC and BedROC, suggesting AUC as a sufficient measure for virtual screening performance.[97]

7.3. Variance and Meta Analysis. An important part of benchmarking is the analysis of variance. The comparison of different statistical analysis methods with the DUD data sets indicated that the compound class-dependent variance of recall dominates other the variance factors.[97] This makes it difficult to evaluate the overall performances of virtual screening methods with a significant level of statistical confidence, and hence, might require the use of many different activity classes, perhaps on the order of 100. Statistical analysis of virtual screens is also affected by the situation that the recognition of active compounds from chemically related series tends to be correlated, another form of analogue bias.

Furthermore, "meta analysis", i.e., the combination of results from different methods, is currently nearly impossible in the field of virtual screening due to the lack of data analysis standards and of standardized primary data.[85] Hence, the use of publicly available benchmark data sets, like DUD or MUV, is a first step toward standardization, reproducibility, and method evaluation, yet commonly applied performance measures and analysis methods are still required, which presents further challenges for the chemoinformatics community.

8. CONCLUSIONS

In this Perspective, we have primarily focused on the increasing relevance and on the use of data mining methods in virtual compound screening. In particular, we have highlighted three approaches that currently dominate the field: (i) SVM learning, (ii) Bayesian methods, and (iii) decision trees. There are certainly other concepts in the machine learning field that might be adaptable for chemoinformatics applications and are yet to be explored. However, it is difficult to pinpoint approaches that might be of particular interest. A few new and interesting adaptations have been discussed, for example, inductive logic programming or belief theory. In order to better understand the potential and limitations of data mining techniques, it is also beneficial to reflect on the foundations of such approaches. Therefore, foundations for the application of data mining methods to chemical problems have been discussed, including chemical reference spaces and different molecular representations. In addition to reviewing currently popular data mining approaches and describing algorithms recently adopted from computer science, new application areas for virtual screening have also been introduced that go beyond the conventional use. Also, difficulties associated with the performance evaluation of virtual screening methods have been reviewed that continue to present major bottlenecks for further development in this field. However, a number of attempts are currently being made to establish community-wide standards and performance measures and to ensure reproducibility and comparability of virtual screening studies. Clearly, these are basic scientific requirements, and many investigators in the chemoinformatics arena are aware of the fact that only the introduction of noncompromising scientific standards will ensure the further development of this field. The many different methodologies that are currently explored for various virtual screening applications provide an exciting playground for further investigations.

ACKNOWLEDGMENT

The authors thank Anne Mai Wassermann for her help in preparing Figure 2.

REFERENCES AND NOTES

(1) Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233–245.

(2) Rupp, M.; Proschak, E.; Schneider, G. Kernel Approach to Molecular Similarity Based on Iterative Graph Similarity. *J. Chem. Inf. Model.* **2007**, *47*, 2280–2286.

(3) Raymond, J. W.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching of Chemical Structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521–533.

(4) Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Comput. J.* **2002**, *45*, 631–644.

(5) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.

(6) Azencott, C.; Ksikes, A.; Swamidass, S. J.; Chen, J. H.; Ralaivola, L.; Baldi, P. One- to Four-Dimensional Kernels for Virtual Screening and the Prediction of Physical, Chemical, and Biological Properties. *J. Chem. Inf. Model.* **2007**, *47*, 965–974.

(7) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49*, 108–119.

(8) Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. *J. Chem. Inf. Model.* **2009**, *49*, 1010–1024.

(9) Guha, R.; Dutta, D.; Jurs, P. C.; Chen, T. R-NN Curves: An Intuitive Approach to Outlier Detection Using a Distance Based Method. *J. Chem. Inf. Model.* **2006**, *46*, 1713–1722.

(10) Ewing, T.; Baber, J. C.; Feher, M. Novel 2D Fingerprints for Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 2423–2431.

(11) Klon, A. E.; Diller, D. J. Library Fingerprints: A Novel Approach to the Screening of Virtual Libraries. *J. Chem. Inf. Model.* **2007**, *47*, 1354–1365.

(12) Watson, P. Naive Bayes Classification Using 2D Pharmacophore Feature Triplet Vectors. *J. Chem. Inf. Model.* **2008**, *48*, 166–178.

(13) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The Use of Consensus Scoring in Ligand-Based Virtual Screening. *J. Chem. Inf. Model.* **2006**, *46*, 277–288.

(14) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of Data Fusion Methods in Virtual Screening: Theoretical Model. *J. Chem. Inf. Model.* **2006**, *46*, 2193–2205.

(15) Whittle, M.; Gillet, V. J.; Willett, P.; Loesel, J. Analysis of Data Fusion Methods in Virtual Screening: Similarity and Group Fusion. *J. Chem. Inf. Model.* **2006**, *46*, 2206–2219.

(16) Simmons, K.; Kinney, J.; Owens, A.; Kleier, D. A.; Bloch, K.; Argentar, D.; Walsh, A.; Vaidyanathan, G. Practical Outcomes of Applying Ensemble Machine Learning Classifiers to High-Throughput Screening (HTS) Data Analysis and Screening. *J. Chem. Inf. Model.* **2008**, *48*, 2196–2206.

(17) Vogt, M.; Bajorath, J. Bayesian Screening for Active Compounds in High-Dimensional Chemical Spaces Combining Property Descriptors and Fingerprints. *Chem. Biol. Drug Des.* **2008**, *71*, 8–14.

(18) Nisius, B.; Bajorath, J. Fingerprint Recombination - Generating Hybrid Fingerprints for Similarity Searching From Different Fingerprint Types. *ChemMedChem* **2009**, *4*, 1859–1863.

(19) Cortes, C.; Vapnik, V. Support-Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.

(20) Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychol. Rev.* **1958**, *65*, 386–408.

(21) Vapnik, V. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer-Verlag: New York, 2000.

(22) Mohr, J. A.; Jain, B. J.; Obermayer, K. Molecule Kernels: A Descriptor- and Alignment-Free Quantitative Structure-Activity Relationship Approach. *J. Chem. Inf. Model.* **2008**, *48*, 1868–1881.

(23) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active Learning with Support Vector Machines in the Drug Discovery Process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.

(24) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.

(25) Li, Q.; Bender, A.; Pei, J.; Lai, L. A Large Descriptor Set and a Probabilistic Kernel-Based Classifier Significantly Improve Druglikeness Classification. *J. Chem. Inf. Model.* **2007**, *47*, 1776–1786.

(26) Ma, X. H.; Wang, R.; Yang, S. Y.; Li, Z. R.; Xue, Y.; Wei, Y. C.; Low, B. C.; Chen, Y. Z. Evaluation of Virtual Screening Performance of Support Vector Machines Trained by Sparsely Distributed Active Compounds. *J. Chem. Inf. Model.* **2008**, *48*, 1227–1237.

(27) Liew, C. Y.; Ma, X. H.; Liu, X.; Yap, C. W. SVM Model for Virtual Screening of Lck Inhibitors. *J. Chem. Inf. Model.* **2009**, *49*, 877–885.

(28) Liu, X. H.; Ma, X. H.; Tan, C. Y.; Jiang, Y. Y.; Go, M. L.; Low, B. C.; Chen, Y. Z. Virtual Screening of Abl Inhibitors from Large Compound Libraries by Support Vector Machines. *J. Chem. Inf. Model.* **2009**, *49*, 2101–2110.

(29) Geppert, H.; Horváth, T.; Gärtner, T.; Wrobel, S.; Bajorath, J. Support-Vector-Machine-Based Ranking Significantly Improves the Effectiveness of Similarity Searching Using 2D Fingerprints and Multiple Reference Compounds. *J. Chem. Inf. Model.* **2008**, *48*, 742–746.

(30) Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. Collaborative Filtering on a Family of Biological Targets. *J. Chem. Inf. Model.* **2006**, *46*, 626–635.

(31) Jacob, L.; Vert, J.-P. Protein-Ligand Interaction Prediction: An Improved Chemogenomics Approach. *Bioinformatics* **2008**, *24*, 2149–2156.

(32) Geppert, H.; Humrich, J.; Stumpfe, D.; Gärtner, T.; Bajorath, J. Ligand Prediction from Protein Sequence and Small Molecule Information Using Support Vector Machines and Fingerprint Descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 767–779.

(33) Kawai, K.; Fujishima, S.; Takahashi, Y. Predictive Activity Profiling of Drugs by Topological-Fragment-Spectra-Based Support Vector Machines. *J. Chem. Inf. Model.* **2008**, *48*, 1152–1160.

(34) Wassermann, A. M.; Geppert, H.; Bajorath, J. Searching for Target-Selective Compounds Using Different Combinations of Multiclass Support Vector Machine Ranking Methods, Kernel Functions, and Fingerprint Descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 582–592.

(35) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295–1300.

(36) Willett, P.; Wilton, D.; Hartzoulakis, B.; Tang, R.; Ford, J.; Madge, D. Prediction of Ion Channel Activity Using Binary Kernel Discrimination. *J. Chem. Inf. Model.* **2007**, *47*, 1961–1966.

(37) Vogt, M.; Bajorath, J. Introduction of an Information-Theoretic Method to Predict Recovery Rates of Active Compounds for Bayesian in Silico Screening: Theory and Screening Trials. *J. Chem. Inf. Model.* **2007**, *47*, 337–341.

(38) Vogt, M.; Bajorath, J. Introduction of a Generally Applicable Method to Estimate Retrieval of Active Molecules for Similarity Searching using Fingerprints. *ChemMedChem* **2007**, *2*, 1311–1320.

(39) Vogt, M.; Nisius, B.; Bajorath, J. Predicting the Similarity Search Performance of Fingerprints and their Combination with Molecular Property Descriptors Using Probabilistic and Information-Theoretic Modeling. *Stat. Anal. Data Min.* **2009**, *2*, 123–134.

(40) Vogt, M.; Bajorath, J. Bayesian Similarity Searching in High-Dimensional Descriptor Spaces Combined with Kullback-Leibler Descriptor Divergence Analysis. *J. Chem. Inf. Model.* **2008**, *48*, 247–255.

(41) Nisius, B.; Vogt, M.; Bajorath, J. Development of a Fingerprint Reduction Approach for Bayesian Similarity Searching Based on Kullback-Leibler Divergence Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 1347–1358.

(42) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of Belief Theory to Similarity Data Fusion for Use in Analog Searching and Lead Hopping. *J. Chem. Inf. Model.* **2008**, *48*, 941–948.

(43) Zhou, Y.-P.; Tang, L.-J.; Jiao, J.; Song, D.-D.; Jiang, J.-H.; Yu, R.-Q. Modified Particle Swarm Optimization Algorithm for Adaptively Configuring Globally Optimal Classification and Regression Trees. *J. Chem. Inf. Model.* **2009**, *49*, 1144–1153.

(44) Palmer, D. S.; O'Boyle, N. M.; Glen, R. C.; Mitchell, J. B. O. Random Forest Models To Predict Aqueous Solubility. *J. Chem. Inf. Model.* **2007**, *47*, 150–158.

(45) Zhang, Q.; Aires-de-Sousa, J. Random Forest Prediction of Mutagenicity from Empirical Physicochemical Descriptors. *J. Chem. Inf. Model.* **2007**, *47*, 1–8.

(46) Ehrman, T. M.; Barlow, D. J.; Hylands, P. J. Virtual Screening of Chinese Herbs with Random Forest. *J. Chem. Inf. Model.* **2007**, *47*, 264–278.

(47) Angelopoulos, N.; Hadjiprocopis, A.; Walkinshaw, M. D. Bayesian Model Averaging for Ligand Discovery. *J. Chem. Inf. Model.* **2009**, *49*, 1547–1557.

(48) Bender, A.; Glen, R. C. *MOLPRINT 2D*; Unilever Cambridge, Centre for Molecular Informatics: University of Cambridge, U.K.; http://www.molprint.com/; Accessed 10/01/2009.

(49) Berder, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Model.* **2004**, *44*, 170–178.

(50) Scitegic Pipeline Pilot; Accelrys, Inc.: San Diego, CA, 2008.

(51) Nigsch, F.; Mitchell, J. B. O. How to Winnow Actives from Inactives: Introducing Molecular Orthogonal Sparse Bigrams (MOSBs) and Multiclass Winnow. *J. Chem. Inf. Model.* **2008**, *48*, 306–318.

(52) Nigsch, F.; Bender, A.; Jenkins, J. L.; Mitchell, J. B. O. Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 2313–2325.

(53) Swamidass, S. J.; Azencott, C.-A.; Lin, T.-W.; Gramajo, H.; Tsai, S.-C.; Baldi, P. Influence Relevance Voting: An Accurate And Interpretable Virtual High Throughput Screening Method. *J. Chem. Inf. Model.* **2009**, *49*, 756–766.

(54) Muggleton, S. H. Inductive Logic Programming. *New Generat. Comput.* **1991**, *8*, 295–318.

(55) Cannon, E. O.; Amini, A.; Bender, A.; Sternberg, M. J. E.; Muggleton, S. H.; Glen, R. C.; Mitchell, J. B. O. Support Vector Inductive Logic Programming Outperforms the Naïve Bayes Classifier and Inductive Logic Programming for the Classification of Bioactive Chemical Compounds. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 269–280.

(56) Tsunoyama, K.; Amini, A.; Sternberg, M. J. E.; Muggleton, S. H. Scaffold Hopping in Drug Discovery Using Inductive Logic Programming. *J. Chem. Inf. Model.* **2008**, *48*, 949–957.

(57) MACCS Structural Keys; Symyx Software: San Ramon, CA, 2005.

(58) Barnard, J. M.; Downs, G. M. Chemical Fragment Generation and Clustering Software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–142.

(59) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems Inc.: Irvine, CA, 2009.

(60) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.

(61) Swamidass, S. J.; Baldi, P. Mathematical Correction for Fingerprint Similarity Measures to Improve Chemical Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 952–964.

(62) Baldi, P.; Benz, R. W.; Hirschberg, D. S.; Swamidass, S. J. Lossless Compression of Chemical Fingerprints Using Integer Entropy Codes Improves Storage and Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 2098–2109.

(63) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(64) Swamidass, S. J.; Baldi, P. Bounds and Algorithms for Fast Exact Searches of Chemical Fingerprints in Linear and Sublinear Time. *J. Chem. Inf. Model* **2007**, *47*, 302–317.

(65) Baldi, P.; Hirschberg, D. S. An Intersection Inequality Sharper than the Tanimoto Triangle Inequality for Efficiently Searching Large Databases. *J. Chem. Inf. Model.* **2009**, *49*, 1866–1870.

(66) Baldi, P.; Hirschberg, D. S.; Nasr, R. J. Speeding Up Chemical Database Searches Using a Proximity Filter Based on the Logical Exclusive OR. *J. Chem. Inf. Model* **2008**, *48*, 1367–1378.

(67) Smellie, A. Compressed Binary Bit Trees: A New Data Structure For Accelerating Database Searching. *J. Chem. Inf. Model.* **2009**, *49*, 257–262.

(68) Quinlan, J. R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106.

(69) Hu, Y.; Lounkine, E.; Bajorath, J. Improving the Performance of Extended Connectivity Fingerprints through Activity-Oriented Feature Filtering and Application of a Bit Density-Dependent Similarity Function. *ChemMedChem* **2009**, *4*, 540–548.

(70) Hu, Y.; Lounkine, E.; Bajorath, J. Filtering and Counting of Extended Connectivity Fingerprint Features Maximizes Compound Recall and the Structural Diversity of Hits. *Chem. Biol. Drug Des.* **2009**, *74*, 92–98.

(71) Batista, J.; Bajorath, J. Similarity Searching Using Compound Class-Specific Combinations of Substructures Found in Randomly Generated Molecular Fragment Populations. *ChemMedChem* **2008**, *3*, 67–73.

(72) Hu, Y.; Lounkine, E.; Batista, J.; Bajorath, J. RelACCS-FP: A Structural Minimalist Approach to Fingerprint Design. *Chem. Biol. Drug Des.* **2008**, *72*, 341–349.

(73) Ormerod, A.; Willett, P.; Bawden, D. Comparison of Fragment-Weighting Schemes for Substructural Analysis. *Quant. Struct.-Act. Relat. (QSAR)* **1989**, *8*, 115–129.

(74) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: an Algorithm to Determine Structural Commonalities in Diverse Datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.

(75) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint Scaling Increases the Probability of Identifying Molecules with Similar Activity in Virtual Screening Calculations. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 746–753.

(76) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile Scaling Increases the Similarity Search Performance of Molecular Fingerprints Containing Numerical Descriptors and Structural Keys. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218–1225.

(77) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity Metrics for Ligands Reflecting the Similarity of the Target Proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.

(78) Wang, Y.; Bajorath, J. Bit Silencing in Fingerprints Enables the Derivation of Compound Class-Directed Similarity Metrics. *J. Chem. Inf. Model.* **2008**, *48*, 1754–1759.

(79) Crisman, T. J.; Sisay, M. T.; Bajorath, J. Ligand-Target Interaction-Based Weighting of Substructures for Virtual Screening. *J. Chem. Inf. Model.* **2008**, *48*, 1955–1964.

(80) Tan, L.; Vogt, M.; Bajorath, J. Three-Dimensional Protein-Ligand Interaction Scaling of Two-Dimensional Fingerprints. *Chem. Biol. Drug Des.* **2009**, *74*, 449–456.

(81) Schneider, G.; Schneider, P.; Renner, S. Scaffold-Hopping: How Far Can You Jump. *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.

(82) Brown, J.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini. Rev. Med. Chem.* **2006**, *6*, 1217–1229.

(83) Wale, N.; Watson, I. A.; Karypis, G. Indirect Similarity Based Methods for Effective Scaffold-Hopping in Chemical Compounds. *J. Chem. Inf. Model* **2008**, *48*, 730–741.

(84) Senger, S. Using Tversky Similarity Searches for Core Hopping: Finding the Needles in the Haystack. *J. Chem. Inf. Model.* **2009**, *49*, 1514–1524.

(85) Mackey, M. D.; Melville, J. L. Better than Random? The Chemotype Enrichment Problem. *J. Chem. Inf. Model.* **2009**, *49*, 1154–1162.

(86) Bajorath, J. Computational Approaches in Chemogenomics and Chemical Biology: Current and Future Impact on Drug Discovery. *Expert Opin. Drug Discovery* **2008**, *3*, 1371–1376.

(87) Stumpfe, D.; Ahmed, H.; Vogt, I.; Bajorath, J. Methods for Computer-Aided Chemical Biology, Part 1: Design of a Benchmark System for the Evaluation of Compound Selectivity. *Chem. Biol. Drug Des.* **2007**, *70*, 182–194.

(88) Stumpfe, D.; Geppert, H.; Bajorath, J. Methods for Computer-Aided Chemical Biology, Part 3: Analysis of Structure-Selectivity Relationships through Single- or Dual-Step Selectivity Searching and Bayesian Classification. *Chem. Biol. Drug Des.* **2008**, *71*, 518–528.

(89) Vogt, I.; Stumpfe, D.; Ahmed, H.; Bajorath, J. Methods for Computer-Aided Chemical Biology, Part 2: Evaluation of Compound Selectivity Using 2D Fingerprints. *Chem. Biol. Drug Des.* **2007**, *70*, 195–205.

(90) Lounkine, E.; Stumpfe, D.; Bajorath, J. Molecular Formal Concept Analysis for Compound Selectivity Profiling in Biologically Annotated Databases. *J. Chem. Inf. Model.* **2009**, *49*, 1359–1368.

(91) Priss, U. Formal Concept Analysis in Information Science. *Annu. Rev. Inf. Sci. Technol.* **2006**, *40*, 521–543.

(92) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-Like Molecules. *Nat. Rev. Drug Discovery* **2005**, *4*, 649–663.

(93) Soichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862–865.

(94) Wassermann, A.; Geppert, H.; Bajorath, J. Ligand Prediction for Orphan Targets Using Support Vector Machines and Various Target-Ligand Kernels is Dominated by Nearest Neighbor Effects. *J. Chem. Inf. Model.* **2009**, *49*, 2155–2167.

(95) Kanehisa, M.; Goto, S.; Kawashima, S.; Nakaya, A. The KEGG Databases at GenomeNet. *Nucleic Acids Res.* **2002**, *30*, 42–46.

(96) Jain, A.; Nicholls, A. Recommendations for Evaluation of Computational Methods. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 133–139.

(97) Nicholls, A. What Do We Know and When Do We Know It? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.

(98) Bender, A.; Glen, R. C. A Discussion of Measures of Enrichment in Virtual Screening: Comparing the Information Content of Descriptors with Increasing Levels of Sophistication. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 1369–1375.

(99) Irwin, J. J.; Shoichet, B. K. ZINC - A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.

(100) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(101) Good, A. C.; Hermsmeier, M. A.; Hindle, S. A. Measuring CAMD Technique Performance: A Virtual Screening Case Study in the Design of Validation Experiments. *J. Comput.-Aided. Mol. Des.* **2004**, *18*, 529–536.

(102) Good, A. C.; Hermsmeier, M. A. Measuring CAMD Technique Performance. 2. How "Druglike" Are Drugs? Implications of Random Test Set Selection Exemplified Using Druglikeness Classification Models. *J. Chem. Inf. Model* **2007**, *47*, 110–114.

(103) Good, A.; Oprea, T. Optimization of CAMD Techniques 3. Virtual Screening Enrichment Studies: A Help or Hindrance in Tool Selection. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 169–178.

(104) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(105) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Topological Descriptors for Similarity-Based Virtual Screening Using Multiple Bioactive Reference Structures. *Org. Biomol. Chem.* **2004**, *2*, 3256–3266.

(106) *MDL Drug Data Report*; Symyx Technologies, Inc.: Santa Clara, CA, 2009.

(107) Rohrer, S. G.; Baumann, K. Impact of Benchmark Data Set Topology on the Validation of Virtual Screening Methods: Exploration and Quantification by Spatial Statistics. *J. Chem. Inf. Model.* **2008**, *48*, 704–718.

(108) Pubchem; National Center for Biotechnology Information (NCBI): Bethesda, MD; http://pubchem.ncbi.nlm.nih.gov. Accessed February 14, 2008.

(109) Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical Comparison of Virtual Screening Methods Against the MUV Data Set. *J. Chem. Inf. Model.* **2009**, *49*, 2168–2178.

(110) Cleves, A. E.; Jain, A. N. Effects of Inductive Bias on Computational Evaluations of Ligand-based Modeling and on Drug Discovery. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 147–159.

(111) Witten, I. H.; Frank, E. *Data Mining - Practical Machine Learning Tools and Techniques*, 2nd ed.; Morgan Kaufmann: San Francisco, CA, 2005, pp. 161−176.

(112) Truchon, J.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

(113) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Model.* **2001**, *41*, 1395–1406.

(114) Clark, R.; Webster-Clark, D. Managing Bias in ROC Curves. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 141–146.

CI900419K