Application and Development of Computational Methods for Ligand-Based Virtual Screening

Kumulative Dissertation

zur Erlangung des Doktorgrades (Dr. rer. nat.)

der Mathematisch-Naturwissenschaftlichen Fakultät

der Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Kathrin Heikamp

aus Bonn

Bonn 2014

Angefertigt mit Genehmigung
der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

1. Gutachter: Univ.-Prof. Dr. rer. nat. Jürgen Bajorath
2. Gutachter: Univ.-Prof. Dr. rer. nat. Andreas Weber
Tag der Promotion: 11. April 2014
Erscheinungsjahr: 2014

Abstract

The detection of novel active compounds that are able to modulate the biological function of a target is the primary goal of drug discovery. Different screening methods are available to identify hit compounds having the desired bioactivity in a large collection of molecules. As a computational method, virtual screening (VS) is used to search compound libraries *in silico* and identify those compounds that are likely to exhibit a specific activity. Ligand-based virtual screening (LBVS) is a subdiscipline that uses the information of one or more known active compounds in order to identify new hit compounds. Different LBVS methods exist, e.g. similarity searching and support vector machines (SVMs). In order to enable the application of these computational approaches, compounds have to be described numerically. Fingerprints derived from the two-dimensional compound structure, called 2D fingerprints, are among the most popular molecular descriptors available.

This thesis covers the usage of 2D fingerprints in the context of LBVS. The first part focuses on a detailed analysis of 2D fingerprints. Their performance range against a wide range of pharmaceutical targets is globally estimated through fingerprint-based similarity searching. Additionally, mechanisms by which fingerprints are capable of detecting structurally diverse active compounds are identified. For this purpose, two different feature selection methods are applied to find those fingerprint features that are most relevant for the active compounds and distinguish them from other compounds. Then, 2D fingerprints are used in SVM calculations. The SVM methodology provides several opportunities to include additional information about the compounds in order to direct LBVS search calculations. In a first step, a variant of the SVM approach is applied to the multi-class prediction problem involving compounds that are active against several related targets. SVM linear combination is used to recover compounds with desired activity profiles and deprioritize compounds with other activities. Then, the SVM methodology is adopted for potency-directed VS. Compound potency is incorporated into the SVM approach through potency-oriented SVM linear combination and kernel function design to direct search calculations to the preferential detection of potent hit compounds. Next, SVM calculations are applied to address an intrinsic limitation of similarity-based methods, i.e., the presence of similar compounds having large differences in their

potency. An especially designed SVM approach is introduced to predict compound pairs forming such activity cliffs. Finally, the impact of different training sets on the recall performance of SVM-based VS is analyzed and caveats are identified.

Acknowledgements

I would like to take the opportunity and thank the persons who accompanied me during the last years and contributed to the completion of this dissertation in many different ways.
First of all, I like to thank my supervisor Prof. Dr. Jürgen Bajorath for his invaluable guidance, continuous support and encouragement throughout my PhD study. I also like to thank Prof. Dr. Andreas Weber for being the co-referent of my thesis.
I would like to express my gratitude to all my colleagues of the LSI research group for providing a helpful, friendly and interactive working atmosphere. There have been many joyful and funny moments, both inside and outside the BIT, that I will not forget. Furthermore, many thanks are given to Anne Mai Wassermann, Dagmar Stumpfe, Dilyana Dimova and Jenny Balfer for productive and pleasant collaborations. I would like to thank Martin Vogt for being approachable for my questions at any time.
Moreover, special thanks are given to Dilyana Dimova for her encouragement and understanding, and for becoming a close friend during the last years. I would like to thank Dagmar Stumpfe for being a confidential person and her patient advices in scientific and personal questions. Furthermore, I like to thank Antonio de la Vega de León for his cheerfulness and his sense of humor.
Finally, I am grateful to my family and friends for their encouragement and support during the last years. Especially, I thank André Oeckerath for his reliable and invaluable support and motivation.

Contents

List of abbreviations

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
AUC	Area under the ROC curve
ECFP	Extended-connectivity fingerprint
FP	Fingerprint
HTS	High-throughput screening
IC_{50}	Half maximal inhibitory concentration
K_i	Inhibition dissociation constant
kNN	k-nearest neighbors
LBVS	Ligand-based virtual screening
LC	Linear combination
MACCS	Molecular ACCess System
ML	Machine learning
MMP	Matched molecular pair
MOE	Molecular Operating Environment
QSAR	Quantitative structure–activity relationship
ROC	Receiver operating characteristic
SAR	Structure-activity relationship
SBVS	Structure-based virtual screening
SPP	Similarity property principle
SVM	Support vector machine
SVR	Support vector regression
Tc	Tanimoto coefficient
TGD	Typed graph distance
TGT	Typed graph triangle
VS	Virtual screening

Introduction

Drug discovery is concerned with the detection of small molecules that are active against a biological target and modulate its biological function. The whole process until a drug can be used as a medication for a specific disease requires several years and is very costly [1]. Furthermore, only a small proportion of candidate compounds is approved and brought to market [1, 2]. The drug discovery process involves several preclinical and clinical stages with numerous investigators involved. In the preclinical phase, the major stages include the identification and validation of novel drug targets, the identification of active compounds (so-called *hits*), and the transformation of hits to lead compounds that can be further optimized [2–4].

The major source of hits is high-throughput screening (HTS) [5, 6]. In HTS, a very large collection of compounds is tested for a biochemical or cellular effect [6]. Those compounds having a positive response in the screening are considered as hit compounds. In follow-up screens, these hits are analyzed concerning pharmacological and physicochemical properties and evaluated for their potential to become a lead compound [6]. In general, HTS data are of limited quality due to the presence of false-positive and false-negative activity measurements. The typically large number of false-positives makes subsequent control experiments necessary. False-negative measurements are often a consequence of limited purity and stability of test compounds or of too low concentrations in the screening assay [5, 7].

In order to compensate for such limitations, computational approaches have been developed. Among them, virtual screening (VS) was introduced as a computational, "time-efficient and cost-effective" [5] analog to HTS. In VS, a large compound library is screened *in silico* against a drug target of interest. Test compounds can then be prioritized according to their probability to possess

a specific activity and a reduced list of compounds is submitted to biological experiments [5, 7–9]. Although VS also suffers from false predictions, it was shown that VS often produces higher hit rates than HTS [9].
In general, HTS and VS are considered to be complementary screening methods [5, 7]. Hence, the integration of computational and biological screening is considered in order to reduce the number of candidate compounds and thereby the costs of experimental testing [7, 10].

VS includes two different approaches: *structure-based* virtual screening (SBVS) [11, 12] and *ligand-based* virtual screening (LBVS) [10, 11, 13]. SBVS methods use the three-dimensional (3D) structure of a target in order to make assumptions about the interactions between ligand and target. The most popular SBVS method is docking, where database compounds are docked into the 3D structure of the target in order to predict the hypothetical binding modes. Then, a score is calculated that reflects the estimated binding affinity of the compound and serves as an indication which compounds should be tested [12, 14].
In contrast, LBVS makes use of ligand information only. The LBVS methods require one or more compounds with a specific activity to identify new hits. Conceptually, LBVS is based on the *similarity property principle* (SPP) formulated by Johnson and Maggiora in 1990 [15]. The principle states that "similar molecules should have similar biological properties (activity)". Subdisciplines of LBVS methods include pharmacophore searching [16], shape comparison [17], similarity searching [18, 19], and machine learning [13]. In the following, similarity searching and an example of a machine learning method, the support vector machines, are discussed in more detail.

Similarity searching

Similarity searching is one of the most widely used LBVS approaches in drug discovery. One or multiple active compounds are used as reference compounds (or templates) to screen a large database of compounds with unknown activity. The database compounds are compared to the reference set and ranked in the order of decreasing similarity. According to the SPP, those compounds that

are located at the top positions of the ranking most probably exhibit a similar bioactivity [19].
For similarity searching, three principal components have to be defined: (*i*) a molecular representation for the compounds, (*ii*) a coefficient for determining the similarity, and (*iii*) a search strategy [19, 20]. They are discussed in detail in the following.

Molecular representations

Molecular descriptors are used to numerically describe the molecular structure and compound properties. There are many molecular descriptors of different complexity available that capture different levels of compound information [21]. In general, one can classify molecular descriptors as one-, two- or three-dimensional (1D, 2D, 3D) depending on the structure representation from which they are derived [7]. 1D descriptors are calculated from the molecular formula not considering atom connectivities. Examples of 1D descriptors are atom count and molecular weight. 2D descriptors are derived from the molecular graph and include topological descriptors as well as calculated descriptors approximating compound properties like logP. Molecular conformations are used to determine 3D descriptors. These descriptors comprise, e.g., volume or molecular surface [7].
The most popular molecular descriptors include molecular fingerprints (FPs) that are bit or integer string representations capturing structural features and physicochemical properties of compounds [22]. In binary fingerprints, each bit in the string encodes the presence or absence of a specific feature. If a specific feature is present in the molecule, the bit is set to '1'; otherwise, it is set to '0'. There are also non-binary count fingerprint versions where the bits are replaced with an integer indicating the frequency of the features. Integer-based fingerprints are derived when compound features are hashed.
In analogy to the molecular descriptors, one can distinguish 2D and 3D FPs based on the structure representation of the molecule [19]. Furthermore, different types of fingerprints have been introduced that vary in the encoding of chemical information and how they are calculated. Hence, they capture different aspects of chemical information [22]. Important molecular fingerprints

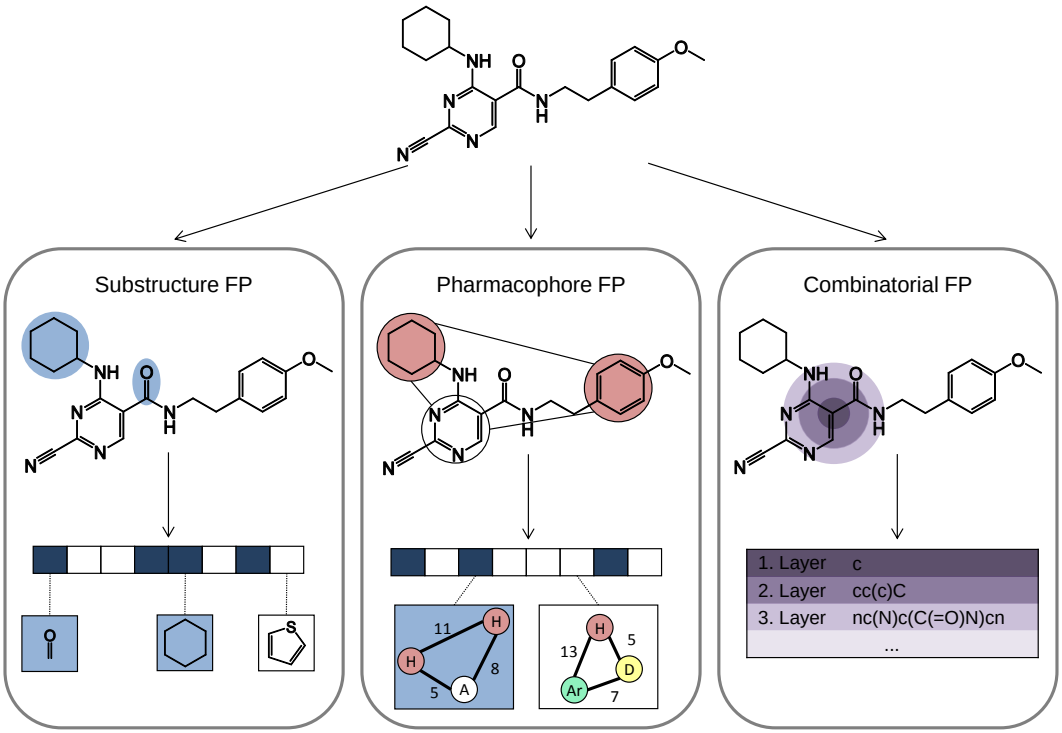


Figure 1: Fingerprint prototypes. Three different fingerprint designs are compared. (*i*) An exemplary keyed substructure FP is shown. Bit positions are set on (i.e., blue) if the corresponding substructure is present and set off otherwise (i.e., white). Two substructures, a carbonyl group and a ring, that account for two bits set are highlighted in the compound structure. (*ii*) In a pharmacophore FP, each bit accounts for one geometrical arrangement of atom types. Here, the three-point pharmacophore pattern "hydrophobic - hydrophobic - hydrogen bond acceptor" with the defined inter-feature distances is available in the structure and the according bit is set on. (*iii*) A combinatorial FP encoding local atom environments is illustrated. Around a central atom, here a carbon, three layers are created up to a diameter of four bonds. The resulting circular environments are hashed. The figure is adapted from [22].

include substructure-based fingerprints, pharmacophore fingerprints, and combinatorial fingerprints like circular atom environments. These fingerprint types are compared in Figure 1 and discussed below.

Two popular substructure-based fingerprints are the Molecular ACCess System (MACCS) structural keys [23] and the BCI fingerprint [24]. The publicly available version of MACCS contains 166 structural fragments and the BCI consists of 1,052 substructures. Both fingerprints are keyed fingerprints with a one-to-one mapping of each bit position to a structural fragment. Each bit in the fingerprint then accounts for the presence or absence of the according substructure.

Pharmacophore fingerprints also belong to the class of keyed fingerprints. Each bit in the string encodes one geometrical arrangement of atom types. The pharmacophore fingerprint of a compound is derived by generating all possible pharmacophore patterns of the compound. That means, in the first step atomic features (e.g. hydrogen bond donor or acceptor) are assigned to individual atoms or groups of atoms. Then, all possible combinations of two to four atomic features and their inter-feature distances are determined. The Molecular Operating Environment (MOE) [25] contains several different 2D and 3D pharmacophore fingerprints. For example, the typed graph distance (TGD) and typed graph triangle (TGT) fingerprints consist of 420 and 1704 bit positions or pharmacophore patterns, respectively, which can be derived from the 2D molecular graph. Thereby, TGD encodes the distance of atomic features and TGT is a three-point pharmacophore encoding feature triangles.
In contrast, combinatorial fingerprints represent another concept as they do not have a predefined length. For example, the extended-connectivity fingerprints (ECFPs) [26] encode circular atom environments. Each non-hydrogen atom in the molecule is assigned to an atom code describing its mass, charge, element type, valence, and the number of neighboring atoms. Then, a local atom environment is created around each atom up to a specific bond depth. The resulting features are hashed to an integer and the final collection of integers forms the fingerprint. Hence, the size of ECFPs is dependent on the given compound. A comparable FP is MOLPRINT2D [27]. It also consists of molecule-specific atom environments. However, different atom encodings are used and the atom environments are encoded as strings of varying size.
Although 2D fingerprints have a lower information content than 3D molecular representations, screening calculations based on 3D descriptors do not principally perform better than similarity searching using 2D fingerprints [28–30]. Important information about target-ligand interactions are implicitly encoded in 2D fingerprint representations [31]. In general, 2D fingerprints are simpler and more robust as they do not require an approximation of the bioactive compound conformation. Therefore, several studies focus on 2D FPs only [20]. Among the 2D FPs, the ECFPs showed the best screening performance in several studies [32, 33].

Similarity coefficients

The similarity between two compounds is typically assessed by a fingerprint comparison. A similarity measure is used to quantify the compound similarity by determining the overlap between the fingerprint strings. The most frequently used similarity measure for binary fingerprints is the Tanimoto (or Jaccard) coefficient (Tc) [18], that is a function $\mathrm{Tc} : \{0,1\}^n \times \{0,1\}^n \to [0,1]$. For two molecular fingerprints A and B, the Tc is defined as

$$\mathrm{Tc}(A,B) = \frac{c}{a+b-c} \tag{1}$$

where a and b are the number of the bits set on in the fingerprints A and B, respectively, and c corresponds to the number of bits set in both fingerprints. It follows that the Tc compares the intersection of fingerprint features with the union of all features present in two compound fingerprints. The Tc values range from zero to one, where zero corresponds to minimal fingerprint similarity and one is the maximal similarity.
The Tc can also be applied to non-binary fingerprints. Then, the Tc calculates the fingerprint overlap by

$$\mathrm{Tc}(A,B) = \frac{\sum_{i=1}^{n} a_i b_i}{\sum_{i=1}^{n} (a_i^2 + b_i^2 - a_i b_i)} \tag{2}$$

Here, the fingerprints have the form $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_n)$ with a length of n. The variables a_i and b_i represent the ith position in the fingerprints A and B, respectively, and $a_i b_i$ is their product. The non-binary Tc has a value range of -0.333 to 1 [18].
Other popular similarity coefficients used in similarity searching include the Tversky coefficient [34], the Forbes coefficient [35], and the Russel-Rao coefficient [35].

Search strategies

Although similarity searching can be applied with only one reference compound, using multiple active compounds usually improves the search performance [36].

There exist different strategies how to make use of the information provided by the multiple references. In general, we can distinguish the two categories data fusion and fingerprint modifications [19].
In the first case, a fusion rule is applied on the similarity values or compound ranks after multiple search calculations have been performed. The k-nearest neighbors (kNN) search method [37, 38] is widely applied in combination with Tc similarity values. In kNN similarity searching, the similarity between all reference compounds to a database compound is individually calculated. The final similarity score for the current database molecule is then the averaged similarity of the k most similar reference compounds. This value is used for ranking. For example, in 10NN searching using Tc similarity the database score is the average of the 10 highest Tc values derived from at least 10 reference compounds. In 1NN search calculations, the maximal similarity yields the database score.
In contrast, fingerprint modification techniques alter the fingerprint representation that is used for searching. In the centroid method, multiple reference compounds are combined by averaging over each bit position in the reference fingerprints [37]. The generated non-binary fingerprint is then used for screening calculations. Furthermore, a consensus fingerprint [39] can be constructed where individual bit positions are set on if the bit frequency in the reference compounds reaches a predefined threshold. In addition, other "fingerprint engineering" methods have been introduced to improve search performance using multiple references [40–43].

Support vector machines

Over the last years, machine learning (ML) methods have become increasingly important to address complex tasks in drug discovery [13]. The general aim of ML is the derivation of a computational model that learns labels from patterns in data, here in compound data. The models can then be applied for property predictions and classification of new, previously not considered compounds. Analogous to similarity searching, the models are also used to rank and/or filter database compounds. Among the different ML methods available,

support vector machines (SVMs) are one of the most popular techniques. They have become increasingly popular during the 1990s based on the work of Vapnik and Cortes [44]. The basic idea of SVMs is the generation of a hyperplane in a high-dimensional space to derive a separation of objects from two classes. Thereby, a key feature of the SVM approach is the attempt to simultaneously address two conflicting objectives: (*i*) minimization of errors on training data and (*ii*) reduction of model complexity in order to avoid overfitting. The result balancing these objectives is a model with high generalization potential.

SVM classification and ranking

SVMs are a supervised machine learning method that makes use of annotated training examples [45–47]. Originally, the SVM approach was developed to solve binary classification problems. During learning, SVM uses a set of n training data $\{\mathbf{x}_i, y_i\}$ $(i = 1, \ldots, n)$ with $\mathbf{x}_i \in \mathcal{X}$ (e.g. $\mathbb{R}^d$) being a feature vector representation and $y_i \in \{-1, 1\}$ the class label (negative or positive) of the training compound i. The SVM derives a hyperplane H that best separates positive from negative instances:

$$H = \{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ is a scalar product, $\mathbf{w}$ the normal vector of the hyperplane and b a scalar.

For linearly separable training data, there exist an infinite number of hyperplanes that correctly classify the data. The optimal hyperplane chosen by the SVM algorithm is the hyperplane that maximizes the distance from the closest training instances to the hyperplane (called *margin*). The optimal, so-called *maximum margin* hyperplane minimizes the "structural risk" of overfitting and enhances the generalization of the classification model. The distance from the hyperplane H to the nearest training instances from the positive and negative class is $1/\|\mathbf{w}\|$ each. Hence, maximizing the distance $1/\|\mathbf{w}\|$ or, correspondingly, minimizing $\|\mathbf{w}\|$ yields the maximum margin hyperplane. The minimiza-

tion problem is a constrained quadratic programming optimization problem and is formulated as

$$\underset{\mathbf{w},b}{\text{minimize:}} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{4}$$

$$\text{subject to:} \quad y_i\left(\langle \mathbf{x}_i, \mathbf{w}\rangle + b\right) \geq 1 \quad \forall i$$

The inequality constraints are defined in order to ensure correct classification of all training examples, which is possible because of the assumed linear separability of the training data.

In case the training data are not linearly separable, the minimization problem has no solution. Then, so-called *slack variables* $\xi_i \geq 0$ can be introduced and a *soft-margin* separating hyperplane is derived. The slack variables relax the constraints defined in equation (4) and allow some training data to be located within the margin or even on the incorrect side of the hyperplane. The value of the slack variables ξ_i correlates with the degree of mispositioning of the training compound i. The minimization problem is reformulated as

$$\underset{\mathbf{w},b}{\text{minimize:}} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \tag{5}$$

$$\text{subject to:} \quad y_i\left(\langle \mathbf{x}_i, \mathbf{w}\rangle + b\right) \geq 1 - \xi_i \quad \text{with} \quad \xi_i \geq 0 \quad \forall i$$

In this formulation, the constant $C > 0$ is introduced to penalize large slack variables. If C is small, large ξ_i are allowed and a less complex model is learned tolerating many errors. However, if C is large, ξ_i has to be small in the minimization and a complex model is learned avoiding large errors. The parameter C can be considered as a trade-off between the size of the margin and the best fit of the classifier to the training data.

In order to solve this optimization problem, Lagrange multipliers α_i are used to reformulate the problem from the primal to a dual form:

$$\underset{\alpha}{\text{maximize:}} \quad L_D = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j\rangle \tag{6}$$

$$\text{subject to:} \quad \sum_i \alpha_i y_i = 0 \quad \text{with} \quad 0 \leq \alpha_i \leq C \quad \forall i$$

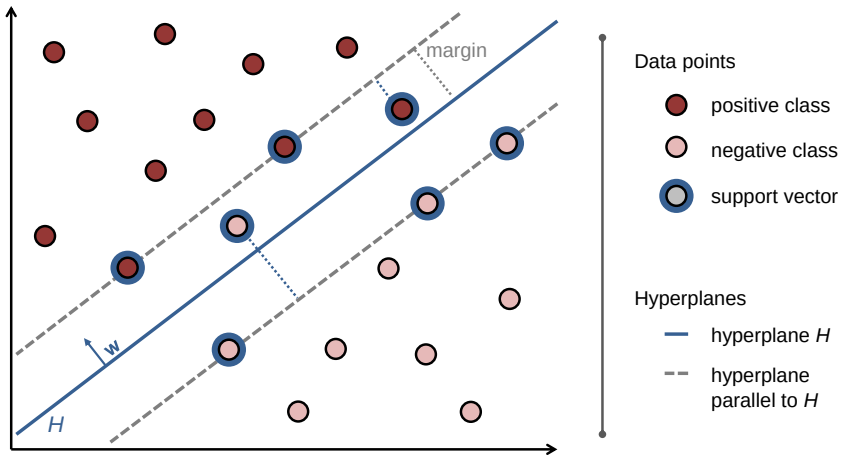


Figure 2: SVM classification. In a binary classification problem, the maximum margin hyperplane H separates two classes (dark and light red dots, respectively). The optimal hyperplane is shown as a blue solid line. Those data points that determine the hyperplane, the support vectors, are encircled in blue. They either lie on the edge of the margin (i.e. on the hyperplanes parallel to H), within the margin, or on the incorrect side of the hyperplane. Misclassified data points obtain values of the slack variables that correlate with the distance to the margin, as illustrated by the dotted lines. The figure is adapted from [48].

This optimization problem is convex and hence has a solution that is the global optimum. Solving the Lagrangian dual formulation results in the normal vector $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$ with non-negative α_i. Those training instances that are associated with factors α_i greater than zero are the so-called *support vectors* and solely determine the position of the hyperplane. These data points lie on the edge, within the margin or even on the incorrect side of the hyperplane. Hence, the generation of the hyperplane only depends on some training examples and not on the dimension of the input space, which allows calculations in a higher dimensional space. A schematic illustration of an SVM classification problem is shown in Figure 2.

Once the normal vector $\mathbf{w}$ has been calculated, the scalar b can be derived from any support vector. Then, the final decision function (or separation rule) for the classification can be formulated as

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b\right) \tag{7}$$

The signum function sgn determines the sign of the prediction value for a test instance $\mathbf{x}$. Geometrically, this is the side of the hyperplane onto which the test

example falls. This means that test points (i.e. compounds) with $f(\mathbf{x}) = 1$ are assigned to the positive class and those with $f(\mathbf{x}) = -1$ to the negative class. In order to adapt the SVM approach for VS and allow ranking of test compounds, the decision function is transformed into a ranking function. This is obtained by removing the signum function from the decision function, thus generating a real value for each test example, i.e.

$$g(\mathbf{x}) = \sum_i \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b \tag{8}$$

Then, test compounds are ranked from the highest to the lowest value. This corresponds to a ranking from the most distant data point on the positive half space to the most distant point on the negative half space [49].

SVM for regression

The SVM approach can also be used to generate real values by estimating a regression function [45, 47, 50]. SVM for regression, also called support vector regression (SVR), has a methodological basis comparable to SVM classification seeking for margin maximization. The training instances for SVR are a set of the form $\{\mathbf{x}_i, y_i\}$ with $\mathbf{x}_i$ being a feature representation from input space $\mathcal{X}$ and $y_i \in \mathbb{R}$ a real value for each training data point i. SVR derives a regression function of the form $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$, mapping training data x_i as close as possible to their real value y_i with a maximum deviation of ϵ. Hence, the SVR approach tolerates errors less than ϵ, but deviations beyond this value are penalized. Therefore, SVR is also termed ϵ-SVR [50]. Similar to SVM classification, the optimization problem is formulated as

$$\underset{\mathbf{w},b}{\text{minimize:}} \quad \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i (\xi_i + \xi_i^*) \tag{9}$$

$$\begin{aligned} \text{subject to:} \quad & y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b \leq \epsilon + \xi_i \quad \forall i \\ & \langle \mathbf{w}, \mathbf{x}_i \rangle + b - y_i \leq \epsilon + \xi_i^* \quad \forall i \\ & \text{with} \quad \xi_i, \xi_i^* \geq 0 \end{aligned}$$

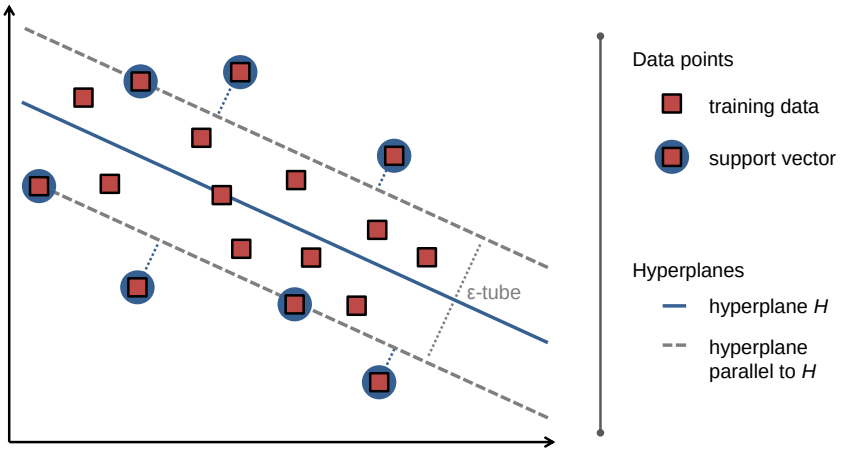


Figure 3: SVM for regression. In SVR, a regression line is fitted to the training data (shown as red squares). The regression function is shown as a solid blue line. It depends on those data points that lie on the edge of the ϵ-tube or outside (shown with dotted lines). These data points are the support vectors and are highlighted by blue circles. All other data points are fitted with sufficient precision and lie within the tube. The figure is adapted from [48].

In SVR, two sets of non-negative slack variables are required to account for positive and negative deviations of the predicted regression value to the true output value. Again, the constant $C > 0$ is used to penalize large slack variables, i.e. deviations from the so-called ϵ-tube. After solving the optimization problem using Lagrangian reformulation, the final regression function is defined as

$$f(\mathbf{x}) = \sum_i (\alpha_i - \alpha_i^*) \langle \mathbf{x}_i, \mathbf{x} \rangle + b \tag{10}$$

Those training data points having either $\alpha_i > 0$ or $\alpha_i^* > 0$ are the support vectors. There are two types of support vectors. In the first case, the support vectors lie exactly on the boundary of the ϵ-tube. They have either $0 < \alpha_i < C$ or $0 < \alpha_i^* < C$ and are used to derive the parameter b. The other support vectors fall outside of the tube and have $\alpha_i = C$ or $\alpha_i^* = C$. All other training data points have both $\alpha_i = \alpha_i^* = 0$ and are fitted with sufficient precision [47]. Figure 3 illustrates an exemplary SVR problem.

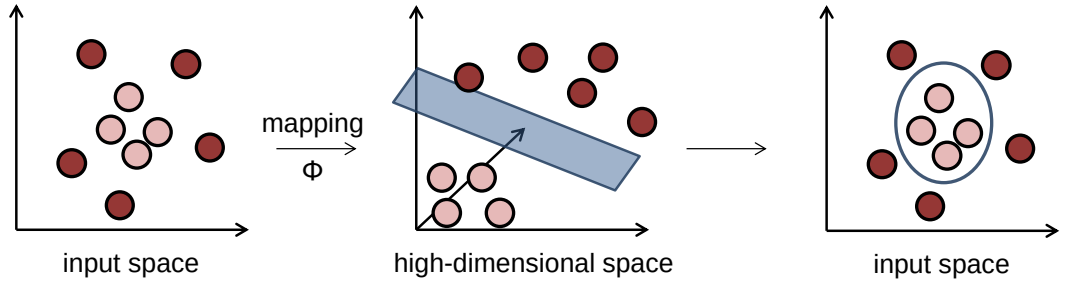


Figure 4: Kernel trick. On the left, data points from two different classes are shown as dark and light red points, respectively. A linear separation of the two classes is not possible in the 2D reference space. However, using a kernel function that implicitly transforms the data into a higher dimensional space (here, a 3D reference space) enables a linear separation. The linear decision function in the 3D space corresponds to an ellipse decision boundary in the input space. The figure is adapted from [48].

Kernel functions

In some cases, a linear separation of training data might not be feasible in space $\mathcal{X}$. In order to allow nonlinear separation rules, the so-called *kernel trick* [51] can be applied to replace the standard scalar product $\langle \cdot, \cdot \rangle$ by a kernel function $K(\cdot, \cdot)$. The kernel transfers the calculations of the scalar product into a higher dimensional space $\mathcal{H}$ by a nonlinear transformation Φ without explicitly calculating the mapping $\Phi : \mathcal{X} \to \mathcal{H}$. The kernel function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by the scalar product between transformed objects $\mathbf{x}_i$ and $\mathbf{x}_j \in \mathcal{X}$ as $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$. The kernel function can be considered as a specialized similarity measure between two arbitrary data points. Because the embedding function Φ does not have to be known, only a valid kernel has to be defined. A valid kernel has to meet two conditions: it must be (i) symmetric, i.e. $K(\mathbf{x}_i, \mathbf{x}_j) = K(\mathbf{x}_j, \mathbf{x}_i)$ for all $\mathbf{x}_i \in \mathcal{X}$, and (ii) positive semi-definite, i.e. the kernel (or Gram) matrix $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{ij}$ is positive semi-definite for $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ [47, 52, 53]. This requirement is met if for all $\mathbf{c} \in \mathbb{R}^n$ $\mathbf{c}^T \mathbf{K} \mathbf{c} \geq 0$, i.e. $\mathbf{K}$ has only non-negative eigenvalues. Here, the condition of positive semi-definiteness guarantees that the optimization problem remains convex.

The kernel trick can be used in SVM classification, ranking and regression. It replaces the standard scalar product of training and test instances in the decision function. Then, the linear model in the new space $\mathcal{H}$ corresponds to a nonlinear model in the input space $\mathcal{X}$. The application of the kernel trick is illustrated in Figure 4.

Popular kernel functions that are used in SVM calculations include, e.g., the linear kernel that corresponds to the standard scalar product:

$$K_{\text{linear}}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle \tag{11}$$

The Gaussian kernel, also known as the radial basis function kernel, is defined as

$$K_{\text{Gaussian}}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2\right) \tag{12}$$

The Gaussian kernel depends on the choice of the inverse-width parameter $\gamma > 0$. Small values of γ result in a smooth decision boundary. However, a large γ increases the flexibility of the decision boundary by incorporating more support vectors raising the risk of overfitting [49, 54].
Finally, the polynomial kernel is given by

$$K_{\text{polynomial}}(\mathbf{x}_i, \mathbf{x}_j) = \left(\langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1\right)^d \tag{13}$$

In the polynomial kernel, the parameter $d \in \mathbb{N}$ determines its degree. Thereby, a kernel with $d = 1$ is the linear kernel with offset 1. Higher values of d result in more flexible decision boundaries [54].
A useful property of kernel functions is that new kernel functions can be built by applying mathematical operations such as multiplication and addition to original kernels [55]. This also allows the design of kernel functions that can be applied on different data types.

Kernel functions in drug discovery

There are several kernel functions available that have been designed to operate on compounds and other bioactivity data. These include compound kernels that accept different compound representations as input and compare diverse properties of molecules. Other kernel functions use multiplication of kernels in order to derive a similarity measure on combined compound-target data.

Compound kernels

Compounds are typically represented as graphs where atoms are shown as labeled vertices and bonds as labeled edges. Several kernel functions have been created to operate on this graph-structured data (see below). These graph kernels allow the comparison of compounds for determining their similarity without ever computing or storing a feature vector representation of the compounds. Gärtner et al. [56] and Kashima et al. [57] introduced graph kernels that measure the overall similarity between two molecular graphs by detecting and counting common walks of equal label sequences in two labeled graphs. In the first case, the kernel is based on the direct graph product [56]. The second kernel belongs to the class of marginalized kernels and uses Markov random walks on the underlying graph structures [57]. In addition to these global graph kernels, there are kernels capturing the local similarity between two graphs [58]. Furthermore, extensions to graph kernels have been introduced [59]. The first extension is a relabeling of vertices so that the new atom labels include information about the topological environment. The second extension covers a modification of the random walk model proposed in [57] in order to prevent irregular loops along an edge, so-called totters.
Although these extensions aim at reducing the computational expense and improving prediction accuracy, graph kernels still are computationally complex and require parameter determination. In order to circumvent these limitations, another direction of kernel design consists of transforming the compound graphs into vectors using molecular descriptors. Ralaivola et al. [60] introduced several new kernel functions that are applied on molecular fingerprints or compound descriptor vectors and are thus computed more efficiently. Among others, the Tanimoto kernel is mentioned that is defined in accordance with the popular Tanimoto coefficient (2) as

$$K_{\text{Tanimoto}}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\langle \mathbf{x}_i, \mathbf{x}_j \rangle}{\langle \mathbf{x}_i, \mathbf{x}_i \rangle + \langle \mathbf{x}_j, \mathbf{x}_j \rangle - \langle \mathbf{x}_i, \mathbf{x}_j \rangle} \tag{14}$$

The Tanimoto kernel is parameter-free. Selecting different compound representations allows the comparison of different compound properties of interest.
Furthermore, kernel functions have been developed that consider the 3D structure of compounds. For example, the pharmacophore kernel [61] focuses on

three-point pharmacophores composed of three atoms, i.e., atom triangles, in 3D space. It can be decomposed into two separate kernels where the first kernel determines the similarity between the atoms and the second kernel assesses the spatial similarity, i.e., the relative locations of the triangle atoms. The similarity between two molecules is finally assessed by summing up the pairwise similarities between all possible pharmacophores in the compounds. The pharmacophore kernel was shown to outperform fingerprint representations of pharmacophores in SVM calculations [61].
Azencott et al. [62] discussed various classes of kernels derived from different levels of molecular representations ranging from 1D to 4D. Molecules were represented as SMILES (1D), bond graphs (2D) or by their 3D atom coordinates. Using the 3D compound structure, additional kernel functions were obtained, e.g. a 2.5D surface kernel and diverse 3D kernels based on Delaunay tetrahedrization, atomic coordinates or pharmacophores. Furthermore, they introduced an additional dimension for kernel functions by averaging over multiple compound configurations leading to 3.5D and 4D kernels. However, Azencott et al. overall showed that the 2D kernel functions for feature vectors outperform kernel functions designed for higher dimensional compound representations [62].

Target-ligand kernels

In addition to the design of different compound kernels, the properties of kernel functions allow the combination of compound data with target information. The so-called target-ligand kernel [63, 64] uses a tensor product to separately determine the similarity between targets and the similarity between compounds. The target similarity is assessed by defining a target kernel on protein information like sequence, structure, or ontology [65]. The similarity between ligands is separately calculated by a ligand kernel in the chemical reference space. The target-ligand kernel is used to separate true from false target-ligand pairings and enables compound classification of any small molecule against multiple targets in parallel [64, 65].
In another study, the target kernel was designed to account for the similarity of the binding sites in two proteins [66]. The descriptors used in the binding site kernel were derived from high-quality X-ray structures of protein-ligand

complexes. A target-ligand kernel using the binding site kernel to account for target similarity was able to correctly predict true protein-ligand pairings [66].

Applications of SVMs in VS

Since their introduction, SVMs have been successfully used in a range of VS tasks. There are studies that only consider the activity against a single target, but increasing efforts are made in the investigation of multi-target activities. Furthermore, SVM-based VS can utilize all variants of the SVM methodology. SVM classification can be used to separate active from inactive compounds. The SVM ranking approach introduced by Jorissen and Gilson [49] can be applied to derive a ranking of database compounds with unknown activity. In addition, SVM regression can be used to predict compound potency. Finally, different kernel functions are applied and enable decisions about the description level of bioactivity data, i.e., using compounds alone or in combination with target information. In the following, a number of applications are discussed.

Detection of active compounds against a single target

One of the first reported screening studies using SVMs was the recovery of dihydrofolate reductase inhibitors [67]. Burbidge et al. demonstrated that SVM-based compound classification outperformed other ML algorithms. In the following, several studies reported the recovery of active compounds against one target of interest including kinases [68], acetylcholinesterase (AChE) [69], or cytochrome P450 (CYP450) [70].

Other studies emphasized the methodological aspect in predicting active compounds and tested their approaches against a range of targets in benchmark calculations. For example, SVM modeling has been used for active learning in a screening study [71]. SVM models were iteratively improved based on the compound predictions from the previous iteration step. SVMs were shown to be able to identify structurally diverse compounds having similar activities [49, 72] and outperformed other fingerprint-based methods [49]. In a comparison of similarity searching and SVM ranking, SVM showed superior performance

when using the same fingerprints as descriptors even when only small training sets were available [73]. Furthermore, a recent study demonstrated good VS performance of linear SVM (i.e. utilizing the linear kernel function) when using high-dimensional, sparsely set fingerprints.
Other studies report the usage of SVMs to derive quantitative structure–activity relationship (QSAR) models that predict compound activity from structure. For example, Sun et al. [74] built a QSAR model based on 2D molecular descriptors to predict phospholipidosis (PLD) activities. The QSAR model introduced by Chen et al. [75] used compound R-group signatures resulting in a more accurate model than the standard Free-Wilson model without losing interpretability.
As discussed above, using the target-ligand kernel enables the inclusion of target information in SVM calculations and further extends the spectrum of available methods in VS. Jacob et al. [64] showed that incorporating additional targets by the target-ligand kernel improves the prediction of single-target activities. Wang et al. [76] proposed to further extend the information content of compound and target data and add drug pharmacological and therapeutic effects to describe drug-target interactions. The fusion of these multiple sources into a kernel function increased prediction accuracy of the SVM classifier. The integration of target and ligand information can also be obtained by the design of a target-ligand vector (i.e. no special kernel function is necessary) [77]. SVM calculations using a target-ligand vector combining small molecule descriptors and target sequence information recovered novel active compounds against four different targets [77].

Orphan screening

In the context of VS, one important subfield is the so-called *orphan screening*. Here, no active ligands against a target of interest are known and additional information about related targets and their active compounds have to be considered in the screening calculations. Wassermann et al. [65] used the target-ligand kernel and combined many different protein kernels comparing protein sequence, structure, or ontology information with ligand similarity to predict new compounds active against orphan targets. Systematic search calculations

showed that the different combinations of target and ligand information did not notably improve performance compared to the standard ligand kernel. The search performance was dominated by the similarity to active compounds of closely related targets to the orphan target [65].
Additionally, SVM linear combination (LC) was introduced for orphan screening [78]. In SVM LC, hyperplanes are generated for each target with a set of known ligands available. The hyperplanes are then linearly combined in order to derive a combined hyperplane for the orphan target. The linear factors for the individual hyperplanes reflect the sequence similarity of each target to the orphan target. The final SVM LC model was demonstrated to successfully predict ligands for orphan targets with high accuracy [78].

Multi-target activity predictions

When considering compound activity against multiple targets two essentially opposing goals can be distinguished. Some studies aim at the recovery of compounds having a specific selectivity against one target over others. Other studies focus on the identification of promiscuous ligands binding multiple targets and predict whole compound profiles.
In a search for target-selective compounds, Wassermann et al. [79] designed different multi-class SVM ranking strategies. The aim was to "purify" the final selection set of a screening experiment and separate selective from non-selective compounds. Hereby, the ranking strategies "preference", "two-step" and "one-versus-all" outperformed the standard SVM ranking approach. Another approach to determine compound selectivity is the multi-label approach "cross-training with SVMs" (ct-SVM) [80]. In this case, a single model was constructed that integrates binary classifiers for individual targets and combines the output values of each classifier to a final compound label. Furthermore, the sequential application of three models with increasingly strict activity levels was applied in order to quantify the activity of test compounds.
The combination of several single-target models can also be used to derive profiles consisting of compound activities against a range of targets. Kawai et al. [81] predicted compound profiles against 100 targets by the sequential application of individual binary SVM classifiers. Additionally, Sato et al. [82]

created a functional profile for marketed drugs containing 125 different molecular functions. The functional profiling was used to detect multifunctionality and adverse effects of small molecules.

Source information: this section follows the text of a recent SVM review from our group [48].

Limitations and challenges of similarity-based prediction methods

LBVS methods have proven to be powerful tools to recover active compounds. However, using only ligand information is often difficult. Besides, the SPP underlying the similarity based approaches has intrinsic limitations, hence complicating the screening calculations and often resulting in failures.
First of all, similarity-based methods depend on how molecular similarity is defined, and this is mainly influenced by the molecular representation chosen. Problems arise if the molecular representation used for screening does not encode the features that are important for the specific activity of interest. In this case, the prediction methods detect similar compounds but probably not those having the desired bioactivity [83]. Furthermore, there is no general rule stating which molecular representation should be used. Instead, preferred search parameters usually depend on the activity data considered [84].
Yet, the main reason for failures to predict active compounds is attributed to the presence of *activity cliffs* [85]. The term activity cliff is used for compound pairs or groups of compounds that have a high structural similarity but show large differences in their potencies. Biologically, this means that a small structural modification changes ligand properties like volume or charge distribution so that a compound cannot bind the target properly anymore and is inactive. This discontinuity in the structure-activity relationship (SAR) of small molecules is frequently observed [86]. However, similarity methods based on the SPP cannot account for SAR discontinuity [19, 87]. As the structural changes are small, they have only slight influences on a molecular representation such as a fingerprint. A similarity method will therefore recover compounds forming

an activity cliff, which will lead to incorrect predictions.
Furthermore, one is generally not interested in those compounds at the top positions of the ranking, i.e., the most similar compounds, but in structurally diverse compounds having similar activity. This is called *scaffold hopping* [88]. 2D fingerprints have often been questioned to be able to detect structurally diverse active compounds, but several studies have proven the opposite [10, 32, 89]. However, it is often not clear where in the ranking these scaffold hops occur, as they depend on the method chosen and the search parameters used [19, 32].
Additionally, all LBVS methods assume that the investigated ligands have the same mode of action. However, active compounds may interact with the same target differently. Two compounds may bind to the active site of the target but occupy different parts of the binding site, or may act on an allosteric site of the target protein [87]. As a consequence, compounds having different binding patterns interact differently with the target and no similarity assumptions for other compounds can be made.

Benchmark calculations

VS methods can be applied prospectively or retrospectively. In prospective VS, candidate compounds are selected after the screening calculations and then experimentally evaluated in order to identify new hit compounds [90]. Alternatively, for method evaluation, many VS calculations are performed retrospectively using benchmark settings [13].
Benchmark calculations require a set of active compounds (i.e., an activity class) and a database containing decoy compounds that are assumed to be inactive. The activity class is assembled from the literature or from databases containing compounds with potency annotations against targets. The decoy compounds are typically randomly selected from a repository like ZINC [91, 92]. The active compounds are split into a reference and a test set. The test set is combined with the database of decoys in order to generate a screening database with "hidden" actives. Then, a VS method (e.g., similarity searching) is applied to the reference set and the screening database. The performance of

the VS approach is measured based on the predictions made for the active test compounds.
There are several statistics available that measure the retrieval of active test compounds. The recall rate (or recovery rate) determines the fraction of identified active compounds at a specific selection set size of the ranking. The hit rate determines the fraction of active test compounds in the selection set. The receiver operating characteristic (ROC) analyzes a compound ranking by plotting the correctly classified actives (true positive rate or sensitivity) against the misclassified actives (false positive rate or 1-specificity). Additionally, the area under the ROC curve (AUC) is used as a performance measure [93]. The AUC value has a range of zero to one, where 0.5 corresponds to a random ranking and values above are preferred. The AUC statistic is recommended to be used in benchmark calculations [94].
However, it is also recognized that the performance is often influenced by the topology of the benchmark data sets [94–97]. Data sets that are affected by "artificial enrichment" and/or "analogue bias" produce artificially high recall statistics [87, 97, 98]. Analogue bias exists when the active compounds are too similar to each other considering simple properties. Artificial enrichment results from actives being too dissimilar to the decoy compounds. For example, in a standard benchmark setting, active compounds are often chemically optimized and hence more complex than the decoys. Simple descriptors like molecular weight or atom count would already enrich actives in the top ranking positions and the general search performance might be overestimated [19]. Therefore, benchmark data sets have been designed to reduce these effects [96, 97].

Public compound repositories

In addition to the benchmark sets available, compounds for LBVS application are typically assembled from public compound repositories. Important databases include BindingDB [99, 100], ChEMBL [101], PubChem [102–104], and ZINC [91, 92], which are discussed in more detail in the following.

BindingDB

BindingDB was developed at the University of Maryland and is probably the first public target-ligand database that was accessible via web in 2000 [99, 100, 105]. It contains small molecules together with their activity measurements and target annotations. Binding affinities against a defined protein target are mainly defined by quantitative data like the inhibition dissociation constant (K_i) or the half maximal inhibitory concentration (IC_{50}) [100]. The main source of target-ligand data is the literature. Additionally, the database includes high-quality data from ChEMBL and PubChem [105].

ChEMBL

Like BindingDB, ChEMBL is an annotated and public database containing activity information for small drug-like bioactive compounds [101, 105]. It is maintained by the European Bioinformatics Institute (EBI), an outpost of the European Molecular Biology Laboratory (EMBL). In addition to binding data and target annotations, ChEMBL provides further information about the compounds like functional information and ADMET properties [101]. ChEMBL contains many bioactivity records against G protein-coupled receptors (GPCRs) and kinases. The main source of bioactivity data are scientific publications in the field of medicinal chemistry that are manually extracted and curated. Furthermore, a subset of PubChem assays is integrated into ChEMBL [101, 105].

PubChem

PubChem is an open database administrated by the US National Institutes of Health (NIH) with the aim to collect bioactivity test data for small molecules and RNA interference (RNAi) reagents [102–105]. PubChem incorporates BindingDB and ChEMBL. It is organized into three related databases: Compounds, Substances and BioAssays. The PubChem BioAssay database contains screening data of chemical structures [103, 104]. These screening data are structured into three different types of records: Summary, Primary, and Confirmatory. The overview of an experiment is provided by the Summary record. The Pri-

mary record contains compounds and their annotations as active and inactive derived by a primary screening experiment at a single concentration. The Confirmatory record reports on confirmatory screening assays that reevaluate the actives from a primary screen and investigate multi-concentration dose-response behavior [105].

ZINC

Finally, ZINC is a large repository containing over 20 million non-annotated molecules derived from chemical vendors. For each compound, ZINC provides generated 3D models for VS calculations [91, 92]. The ZINC database is maintained by the Department of Pharmaceutical Chemistry at the University of California San Francisco (UCSF). ZINC is often used as a screening database in VS applications.

Data from these compound repositories have been used throughout this thesis.

Bibliography

[1] Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.* **2010**, *9*, 203–214.

[2] Bajorath, J. Rational drug discovery revisited: interfacing experimental programs with bio- and chemo-informatics. *Drug Discov. Today* **2001**, *6*, 989–995.

[3] Ratti, E.; Trist, D. Continuing evolution of the drug discovery process in the pharmaceutical industry. *Pure Appl. Chem.* **2001**, *73*, 67–75.

[4] Lombardino, J. G.; Lowe, J. A. The role of the medicinal chemist in drug discovery - then and now. *Nat. Rev. Drug Discov.* **2004**, *3*, 853–862.

[5] Mestres, J. Virtual screening: a real screening complement to high-throughput screening. *Biochem. Soc. Trans.* **2002**, *30*, 797–799.

[6] Smith, A. Screening for drug discovery: the leading question. *Nature* **2002**, *418*, 453–459.

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

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

[9] Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

[10] Stumpfe, D.; Bajorath, J. Applied virtual screening: strategies, recommendations, and caveats. In *Virtual Screening: Principles, Challenges, and Practical Guidelines*, Sotriffer, C., Ed.; Wiley-VCH: Weinheim, 2011, 73–103.

[11] Green, D. V. S. Virtual screening of chemical libraries for drug discovery. *Expert Opin. Drug Discov.* **2008**, *3*, 1011–1026.

[12] Lyne, P. D. Structure-based virtual screening: an overview. *Drug Discov. Today* **2002**, *7*, 1047–1055.

[13] Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

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

[15] Johnson, M. A.; Maggiora, G. M. *Concepts and applications of molecular similarity*; Johnson, M. A., Maggiora, G. M., Eds.; John Wiley & Sons: New York, 1990.

[16] Mason, J. S.; Good, A. C.; Martin, E. J. 3-D pharmacophores in drug discovery. *Curr. Pharm. Des.* **2001**, *7*, 567–597.

[17] Rush, T. S.; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48*, 1489–1495.

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

[19] Stumpfe, D.; Bajorath, J. Similarity searching. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2011**, *1*, 260–282.

[20] Geppert, H.; Bajorath, J. Advances in 2D fingerprint similarity searching. *Expert Opin. Drug Discov.* **2010**, *5*, 529–542.

[21] Todeschini, R.; Consonni, V. *Handbook of molecular descriptors*; Wiley-VCH: Weinheim, 2002.

[22] Heikamp, K.; Bajorath, J. Fingerprint design and engineering strategies: rationalizing and improving similarity search performance. *Future Med. Chem.* **2012**, *4*, 1945–1959.

[23] *MACCS Structural Keys.* Symyx Software, San Ramon, CA, USA, 2002.

[24] Barnard, J. M.; Downs, G. M. Chemical fragment generation and clustering software. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 141–142.

[25] *Molecular Operating Environment (MOE).* Chemical Computing Group Inc., Montreal, Quebec, Canada.

[26] Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

[27] Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.

[28] Brown, R. D.; Martin, Y. C. Use of structure-activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572–584.

[29] Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J. Chem. Inf. Model.* **2010**, *50*, 2079–2093.

[30] Hu, G.; Kuang, G.; Xiao, W.; Li, W.; Liu, G.; Tang, Y. Performance evaluation of 2D fingerprint and 3D shape similarity methods in virtual screening. *J. Chem. Inf. Model.* **2012**, *52*, 1103–1113.

[31] Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand-receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1–9.

[32] Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? Guidelines for virtual screening. *J. Med. Chem.* **2010**, *53*, 5707–5715.

[33] Gardiner, E. J.; Holliday, J. D.; O'Dowd, C.; Willett, P. Effectiveness of 2D fingerprints for scaffold hopping. *Future Med. Chem.* **2011**, *3*, 405–414.

[34] Tversky, A. Features of similarity. *Psychol. Rev.* **1977**, *84*, 327–352.

[35] Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and display of the size dependence of chemical similarity coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819–828.

[36] Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] Nisius, B.; Bajorath, J. Molecular fingerprint recombination: generating hybrid fingerprints for similarity searching from different fingerprint types. *ChemMedChem* **2009**, *4*, 1859–1863.

[44] Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.

[45] Vapnik, V. N. *The nature of statistical learning theory*; Springer: New York, 1995.

[46] Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.* **1998**, *2*, 121–167.

[47] Alpaydin, E. *Introduction to machine learning*, 2nd ed.; The MIT Press: Cambridge, MA, USA, 2010.

[48] Heikamp, K.; Bajorath, J. Support vector machines for drug discovery. *Expert Opin. Drug Discov.* **2014**, *9*, 93–104.

[49] Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.

[50] Smola, A. J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222.

[51] Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proc. 5th Annu. Work. Comput. Learn. Theory*, ACM**, 1992**, 144–152.

[52] Vert, J.-P.; Jacob, L. Machine learning for in silico virtual screening and chemical genomics: new strategies. *Comb. Chem. High Throughput Screen.* **2008**, *11*, 677–685.

[53] Hofmann, T.; Schölkopf, B.; Smola, A. J. Kernel methods in machine learning. *Ann. Stat.* **2008**, *36*, 1171–1220.

[54] Ben-Hur, A.; Weston, J. A user's guide to support vector machines. *Methods Mol. Biol.* **2010**, *609*, 223–239.

[55] Aronszajn, N. Theory of reproducing kernels. *Trans. Am. Math. Soc.* **1950**, *68*, 337–404.

[56] Gärtner, T.; Flach, P.; Wrobel, S. On graph kernels: hardness results and efficient alternatives. In *Proc. 16th Annu. Conf. Comput. Learn. Theory 7th Kernel Work.* **2003**, 129–143.

[57] Kashima, H.; Tsuda, K.; Inokuchi, A. Marginalized kernels between labeled graphs. In *Proc. 20th Int. Conf. Mach. Learn.* The AAAI Press**,** **2003**, 321–328.

[58] Smalter, A.; Huan, J.; Lushington, G. GPM: A graph pattern matching kernel with diffusion for chemical compound classification. In *Proc. IEEE Int. Symp. Bioinforma. Bioeng.* **2008**, 1–6.

[59] Mahé, P.; Ueda, N.; Akutsu, T.; Perret, J.-L.; Vert, J.-P. Graph kernels for molecular structure-activity relationship analysis with support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 939–951.

[60] Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093–1110.

[61] Mahé, P.; Ralaivola, L.; Stoven, V.; Vert, J.-P. The pharmacophore kernel for virtual screening with support vector machines. *J. Chem. Inf. Model.* **2006**, *46*, 2003–2014.

[62] Azencott, C.-A.; 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.

[63] 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.

[64] Jacob, L.; Vert, J.-P. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149–2156.

[65] Wassermann, A. M.; 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.

[66] Meslamani, J.; Rognan, D. Enhancing the accuracy of chemogenomic models with a three-dimensional binding site kernel. *J. Chem. Inf. Model.* **2011**, *51*, 1593–1603.

[67] Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5–14.

[68] 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.

[69] Lv, W.; Xue, Y. Prediction of acetylcholinesterase inhibitors and characterization of correlative molecular descriptors by machine learning methods. *Eur. J. Med. Chem.* **2010**, *45*, 1167–1172.

[70] Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Huang, R. Predictive models for cytochrome p450 isozymes based on quantitative high throughput screening data. *J. Chem. Inf. Model.* **2011**, *51*, 2474–2481.

[71] 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.

[72] Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead hopping using SVM and 3D pharmacophore fingerprints. *J. Chem. Inf. Model.* **2005**, *45*, 1122–1133.

[73] 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.

[74] Sun, H.; Shahane, S.; Xia, M.; Austin, C. P.; Huang, R. Structure based model for the prediction of phospholipidosis induction potential of small molecules. *J. Chem. Inf. Model.* **2012**, *52*, 1798–1805.

[75] Chen, H.; Carlsson, L.; Eriksson, M.; Varkonyi, P.; Norinder, U.; Nilsson, I. Beyond the scope of Free-Wilson analysis: building interpretable QSAR models with machine learning algorithms. *J. Chem. Inf. Model.* **2013**, *53*, 1324–1336.

[76] Wang, Y.-C.; Zhang, C.-H.; Deng, N.-Y.; Wang, Y. Kernel-based data fusion improves the drug-protein interaction prediction. *Comput. Biol. Chem.* **2011**, *35*, 353–362.

[77] Wang, F.; Liu, D.; Wang, H.; Luo, C.; Zheng, M.; Liu, H.; Zhu, W.; Luo, X.; Zhang, J.; Jiang, H. Computational screening for active compounds targeting protein sequences: methodology and experimental validation. *J. Chem. Inf. Model.* **2011**, *51*, 2821–2828.

[78] 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.

[79] 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.

[80] Michielan, L.; Stephanie, F.; Terfloth, L.; Hristozov, D.; Cacciari, B.; Klotz, K.-N.; Spalluto, G.; Gasteiger, J.; Moro, S. Exploring potency and selectivity receptor antagonist profiles using a multilabel classification approach: the human adenosine receptors as a key study. *J. Chem. Inf. Model.* **2009**, *49*, 2820–2836.

[81] 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.

[82] Sato, T.; Matsuo, Y.; Honma, T.; Yokoyama, S. In silico functional profiling of small molecules and its applications. *J. Med. Chem.* **2008**, *51*, 7705–7716.

[83] Maggiora, G. M.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* in press.

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

[85] Maggiora, G. M. On outliers and activity cliffs - why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

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

[87] 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.

[88] Schneider, G.; Schneider, P.; Renner, S. Scaffold-hopping: how far can you jump? *QSAR Comb. Sci.* **2006**, *25*, 1162–1171.

[89] Stumpfe, D.; Bill, A.; Novak, N.; Loch, G.; Blockus, H.; Geppert, H.; Becker, T.; Schmitz, A.; Hoch, M.; Kolanus, W.; Famulok, M.; Bajorath, J. Targeting multifunctional proteins by virtual screening: structurally diverse cytohesin inhibitors with differentiated biological functions. *ACS Chem. Biol.* **2010**, *5*, 839–849.

[90] Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **2010**, *53*, 8461–8467.

[91] 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.

[92] Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: a free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757–1768.

[93] Witten, I. H.; Frank, E. *Data mining: practical machine learning tools and techniques*, 2nd ed.; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2005.

[94] Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput. Aided Mol. Des.* **2008**, *22*, 133–139.

[95] Nicholls, A. What do we know and when do we know it? *J. Comput. Aided Mol. Des.* **2008**, *22*, 239–255.

[96] Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

[97] 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.

[98] Rohrer, S. G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J. Chem. Inf. Model.* **2009**, *49*, 169–184.

[99] Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: data management and interface design. *Bioinformatics* **2002**, *18*, 130–139.

[100] Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.

[101] Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107.

[102] Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H. PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res.* **2009**, *37*, W623–W633.

[103] Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's BioAssay database. *Nucleic Acids Res.* **2012**, *40*, D400–D412.

[104] Wang, Y.; Suzek, T.; Zhang, J.; Wang, J.; He, S.; Cheng, T.; Shoemaker, B. A.; Gindulyte, A.; Bryant, S. H. PubChem BioAssay: 2014 update. *Nucleic Acids Research* **2014**, *42*, D1075–D1082.

[105] Nicola, G.; Liu, T.; Gilson, M. K. Public domain databases for medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 6987–7002.

Thesis outline

The analysis and extension of similarity-based search methods using 2D fingerprints are the objectives of this thesis. For this purpose, search methods utilizing fingerprints are studied in detail. In addition, computational approaches are designed for applications in LBVS that are not addressed by standard methods. The thesis consists of six individual chapters and is structured as follows.

Chapter 1 presents a large-scale similarity search analysis of ChEMBL compound data sets. Similarity searching using 2D fingerprints is applied to a wide range of pharmaceutical targets in order to estimate the performance range of 2D fingerprints.

In *Chapter 2*, 2D fingerprint-based similarity searching is analyzed to identify the mechanism by which fingerprints detect structurally diverse compounds. Therefor, fingerprints are systematically reduced and the recall performance of reduced and unmodified fingerprints is compared.

Chapter 3 reports the application of 2D fingerprints in SVM-based search calculations. A variant of the SVM methodology, SVM linear combination, is used to investigate a multi-class prediction problem involving compounds having closely related and overlapping activity profiles.

Chapter 4 describes the adaption of SVM linear combination and the design of a kernel function with the objective to incorporate compound potency in LBVS. The results of the potency-directed VS calculations applied on HTS data sets are presented.

In *Chapter 5*, kernel functions are introduced for the comparison of pairs of compounds. These kernels capture different structural elements of compound pairs and are used to predict activity cliffs in compound data sets, a principal limitation for which similarity-based search methods cannot account for.

Chapter 6 investigates the influence of the negative training set on compound

recall of SVM-based VS. Compounds of different origin are used as negative training instances and their effect on the search performance is analyzed. Additionally, search calculations are performed in diverse background databases to determine their impact.
Finally, the major findings of the thesis project are summarized and general conclusions are presented.

Chapter 1

Large-scale similarity search profiling of ChEMBL compound data sets

Introduction

Similarity searching is often performed using 2D fingerprints as molecular representation. In this study, we generate a large-scale similarity search profile of the ChEMBL database using two popular, conceptually different 2D fingerprints. Search calculations are performed on well-defined ChEMBL activity classes applying three different kNN search strategies and the fingerprints MACCS and ECFP4. The two fingerprints represent opposite levels of resolution and complexity and allow the derivation of a performance range of 2D fingerprint-based similarity search calculations.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

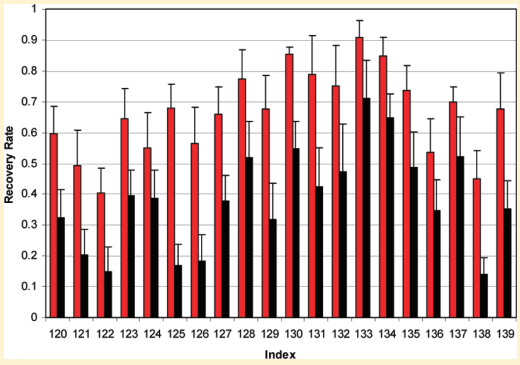
Large-Scale Similarity Search Profiling of ChEMBL Compound Data Sets

Kathrin Heikamp and Jürgen Bajorath*

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

Supporting Information

ABSTRACT: A large-scale similarity search investigation has been carried out on 266 well-defined compound activity classes extracted from the ChEMBL database. The analysis was performed using two widely applied two-dimensional (2D) fingerprints that mark opposite ends of the current performance spectrum of these types of fingerprints, i.e., MACCS structural keys and the extended connectivity fingerprint with bond diameter four (ECFP4). For each fingerprint, three nearest neighbor search strategies were applied. On the basis of these search calculations, a similarity search profile of the ChEMBL database was generated. Overall, the fingerprint search campaign was surprisingly successful. In 203 of 266 test cases (~76%), a compound recovery rate of at least 50% was observed with at least the better performing fingerprint and one search strategy. The similarity search profile also revealed several general trends. For example, fingerprint searching was often characterized by an early enrichment of active compounds in database selection sets. In addition, compound activity classes have been categorized according to different similarity search performance levels, which helps to put the results of benchmark calculations into perspective. Therefore, a compendium of activity classes falling into different search performance categories is provided. On the basis of our large-scale investigation, the performance range of state-of-the-art 2D fingerprinting has been delineated for compound data sets directed against a wide spectrum of pharmaceutical targets.



INTRODUCTION

Molecular fingerprints are usually defined as bit string representations of molecular structure and properties and have for more than two decades been utilized in chemical similarity searching and virtual screening for new active compounds.[1−3] Fingerprints can be classified into 2D and 3D molecular representations, dependent on molecular graph- or conformation-derived features that are utilized for their design.[1,4] Regardless of specific design criteria, fingerprint search calculations involve the comparisons of fingerprints calculated for reference and database compounds and the quantitative comparison of fingerprint (bit string) overlap as a measure of molecular similarity.[1] Accordingly, fingerprint searching is an intrinsically simple similarity method, especially when 2D fingerprints are used that only require the molecular graph as input. Various fingerprint engineering[4−6] and similarity search strategies[7−9] have been introduced to further improve fingerprint performance and/or tune fingerprints for compound class-specific search calculations. Despite their conceptual simplicity, 2D fingerprints have been shown to display significant scaffold hopping potential in benchmark trials[10,11] and practical applications.[12,13]

Virtual screening tools including 2D fingerprints are typically evaluated in retrospective benchmark investigations on activity classes taken from compound databases.[14−16] Popular source databases include, for example, the MDDR,[17] which is, however, license-restricted similar to other commercial database products and, hence, not generally available. Therefore, carefully designed data sets that are made publicly available[10,18] are highly relevant for method evaluation[10,19] as well as public domain compound repositories,[20−22] especially those that collect compound activity and optimization data from medicinal chemistry literature or patent sources.[21,22] These compound databases provide a sound basis for the generation of compound data sets that can be freely shared for method comparison.

Here we have carried out an unconventional fingerprint similarity search investigation on public domain compound data. Rather than comparing the search performance of 2D fingerprints on a limited number of selected activity classes, which is typically done,[10,15,16] we have extracted all compound data sets from ChEMBL[22] that were suitable for fingerprint test calculations in order to generate a similarity search profile of this database. ChEMBL currently is the largest publicly available repository of curated compound activity data taken from medicinal chemistry sources. Furthermore, rather than evaluating different fingerprints on ChEMBL data sets to compare details of their relative search performance, we selected two 2D fingerprints that represent the current performance spectrum of these search tools.

Received: May 5, 2011
Published: July 05, 2011

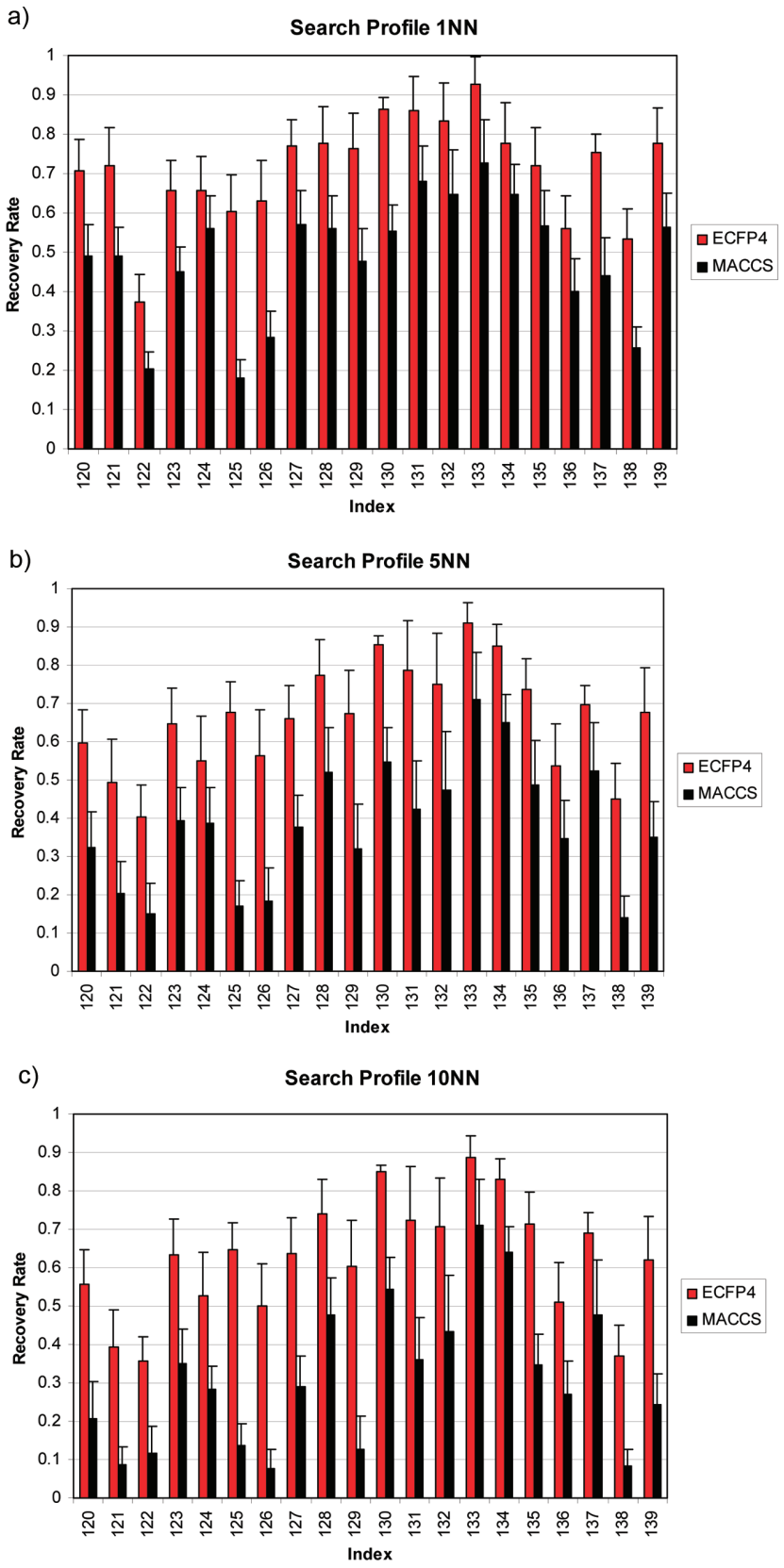


Figure 1. Similarity search profile. Average recovery rates (selection set size equal to the number of ADCs) of a representative subset of 20 activity classes (number 120−139 in Supporting Information Table S1) are reported in a histogram representation for MACCS (black) and ECFP4 (red). Positive standard deviations are displayed as error bars. The index on the *x*-axis reports the consecutively numbered activity classes. Search strategy: (a) 1NN, (b) 5NN, (c) 10NN.

As a well-established fingerprint that marks the basic performance level of conventional 2D fingerprints, we selected MACCS[23] as the prototype of a "low resolution" structural fragment dictionary fingerprint (consisting of 166 predefined structural keys). The MACCS design goes back to the roots of 2D fingerprinting and is often used as a standard to put the performance of different fingerprints into perspective.[10,16] Furthermore, we selected ECFP4 as a representative of a popular "high resolution" class of extended connectivity fingerprints[24] that currently probably represent the top performance level among 2D fingerprints of different design.[10,11] These combinatorial fingerprints systematically monitor circular atom environments up to a given bond diameter in test compounds and assemble these structural features in a molecule-specific manner, rather than based on predefined dictionaries. Hence, MACCS and ECFP4 can be used as markers to represent the current spectrum of 2D fingerprint search performance, which enables the generation of a similarity search profile of a large database and also makes it possible to characterize individual compound activity classes according to the degree of difficulty they represent for 2D fingerprint searching. Importantly, we did not aim to generate individual activity classes with predefined molecular properties for benchmarking or, alternatively, to carry out a standard fingerprint comparison. Rather, the focal points of this study have been to mark the boundaries of 2D similarity search performance on a large scale and, in addition, provide some guidance for the evaluation of similarity search calculations on well-curated publicly available compound classes.

Table 1. Average Recovery Rates[a]

	ECFP4			MACCS		
	1NN	5NN	10NN	1NN	5NN	10NN
average	63.6	58.9	55.2	45.3	37.2	31.7
st dev	7.4	8.6	8.2	7.2	8.4	7.6

[a] Average recovery rates (in percent) and standard deviations (st dev) are reported over all ChEMBL activity classes and search trials.

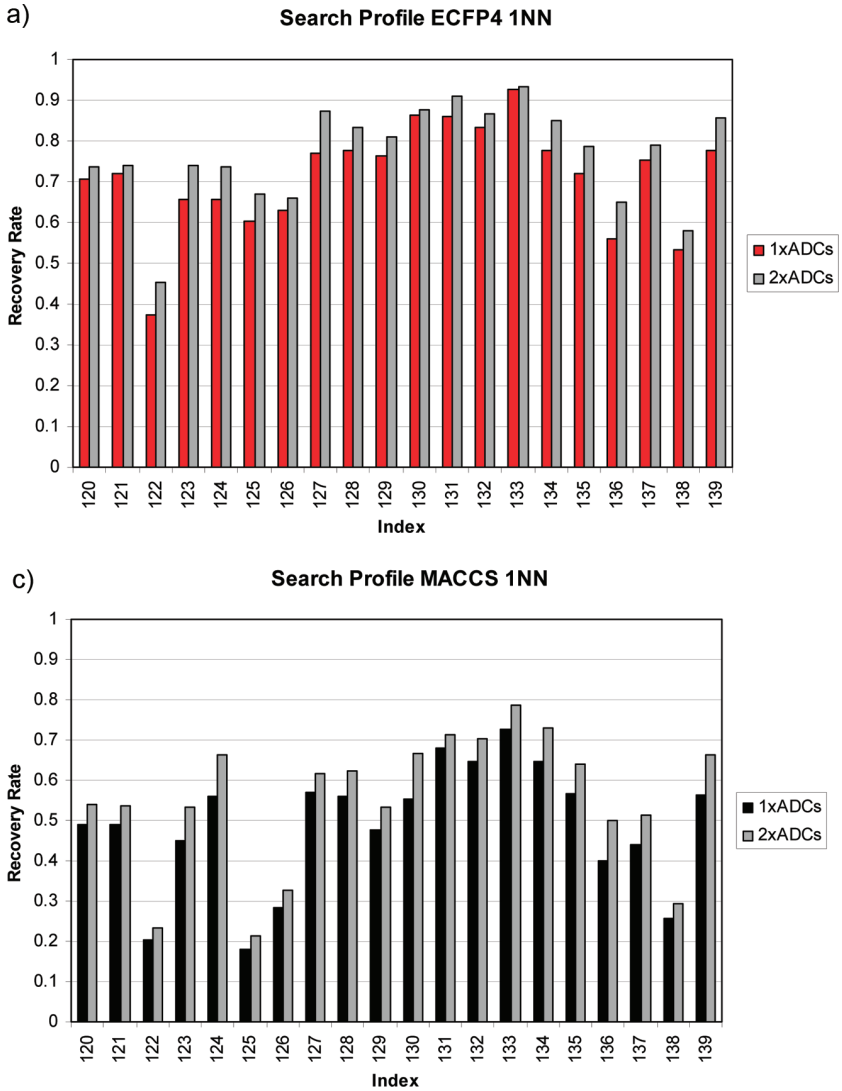


Figure 2. Early enrichment characteristics. Average recovery rates of a representative subset of 20 activity classes (numbers 120−139 in Supporting Information Table S1) are reported for selection set sizes of one or two times the number of ADCs per activity class. The index reports the consecutively numbered activity classes. Fingerprints and search strategies: (a) ECFP4/1NN, (b) MACCS/1NN.

Table 2. Activity Class Yielding Highest Fingerprint Search Performance[a]

			ECFP4		MACCS	
no.	target ID	target name	1NN	10NN	1NN	10NN
256	101174	pituitary adenylate cyclase-activating polypeptide type I receptor	100.0	100.0	100.0	100.0
264	101395	IgG receptor FcRn large subunit p51	100.0	99.2	100.0	100.0
83	10102	5-lipoxygenase activating protein	100.0	100.0	94.7	95.8
180	10144	bone morphogenetic protein 1	97.2	97.2	84.0	88.6
251	12909	ileal bile acid transporter	90.4	91.9	89.5	81.5
253	20130	inhibitor of apoptosis protein 3	90.3	90.1	86.0	86.6
169	275	retinoid X receptor alpha	92.5	94.9	79.2	84.0
228	11061	motilin receptor	94.3	90.7	83.3	82.0
173	10056	DNA-dependent protein kinase	88.8	93.5	81.6	84.5
231	11096	sodium/hydrogen exchanger 1	77.9	89.7	88.5	91.2
214	10845	phospholipase D1	90.4	90.6	81.1	84.3
246	11758	glucagon-like peptide receptor	95.8	80.3	87.7	78.4
189	11402	furin	91.3	81.1	91.0	76.7
31	12725	matriptase	91.4	83.6	87.9	76.0
262	101219	secreted frizzled-related protein 1	99.1	100.0	67.1	72.0
119	176	Purinergic receptor P2Y12	89.2	86.3	81.4	79.7
175	10087	deoxycytidine kinase	95.9	91.8	85.4	63.1
247	100098	serine/threonine-protein kinase WEE1	95.2	96.6	65.9	71.7
74	10624	serotonin 5a (5-HT5a) receptor	84.3	91.4	72.5	79.1
133	12659	prostanoid DP receptor	92.8	88.5	72.8	70.9
203	10582	cytosolic phospholipase A2	95.1	89.5	76.2	62.9
261	100862	metastin receptor	93.9	84.6	75.5	66.1
90	117	somatostatin receptor 2	87.4	86.4	71.1	73.8
6	4	voltage-gated T-type calcium channel alpha-1H subunit	86.5	80.4	74.6	76.3
216	11635	protein kinase C alpha	79.9	85.3	72.5	72.0
235	11242	Focal adhesion kinase 1	91.9	92.3	63.5	61.9
48	34	fibronectin receptor beta	93.2	90.3	68.0	56.9
212	100077	cell division cycle 7-related protein kinase	89.1	89.5	60.8	68.4
33	193	coagulation factor IX	85.6	86.0	60.9	74.7
102	80	FK506-binding protein 1A	91.0	85.6	71.0	59.0

[a] The top 30 activity classes yielding the highest overall search performance are reported and ranked according to decreasing average recovery rate (i.e., top-down) of ECFP4 (1NN and 10NN) and MACCS (1NN and 10NN) calculations for selection sets equal to the number of ADCs. For each activity class, the ChEMBL target ID and target name are provided and average recovery rates are reported (in percent).

The results of our large-scale fingerprint search investigation on ChEMBL are reported herein.

METHODS AND MATERIALS

Compound Data Sets. From ChEMBL, version 9,[22] activity classes were systematically extracted that contained at least 50 compounds active against human target proteins at high confidence level (ChEMBL level 9) for direct (D) interactions (i.e., 9/D[22]) with at least 10 μM potency. On the basis of these selection criteria, a total of 266 activity classes were obtained that contained between 50 and 1793 compounds, with on average ~239 compounds per class, as reported in Table S1 of the Supporting Information. These activity classes consist of designated enzyme or transporter inhibitors or receptor antagonists (with the exception of one class designated as ligands). When reporting activity classes herein, we refer to the target name, as given in ChEMBL.

Fingerprints and Search Strategies. For both MACCS[23] and ECFP4,[24] three *k*-nearest neighbor (*k*NN) search strategies[7] for multiple reference compounds were applied, i.e., 1NN, 5NN, and 10NN. The Tanimoto coefficient (Tc)[1] was calculated as the similarity measure. In 1NN calculations, a database compound is compared to all *k* reference compounds and the highest Tc value is utilized as the final similarity value for the database compound. In 5NN and 10NN calculations, the top 5 and top 10 Tc values are averaged, respectively, to yield the final similarity value for a database compound.

Similarity Searching. From each activity class, 100 reference sets of 10 compounds each were randomly selected and used for individual MACCS and ECFP4 search trials. In each case, all remaining active compounds were added as *active database compounds* (ADCs) to a background database containing one million molecules randomly selected from ZINC.[25] The choice of 10 reference compounds meant that the 10NN search strategy equally took similarity contributions from all reference compounds into account when calculating the similarity score for a database molecule. Initially, rather than using database selection sets of constant size, activity class-specific selection sets were utilized of a size equal to the number of ADCs. For each activity

Table 3. Activity Classes Yielding Lowest Fingerprint Search Performance[a]

			ECFP4		MACCS	
no.	target ID	target name	1NN	10NN	1NN	10NN
5	165	HERG	21.1	9.9	13.6	1.5
37	10193	carbonic anhydrase I	17.6	11.3	17.4	6.2
24	15	carbonic anhydrase II	17.6	14.3	15.8	7.0
96	11489	11-beta-hydroxysteroid dehydrogenase 1	25.1	16.8	15.2	2.8
67	121	serotonin transporter	26.5	17.5	14.7	5.5
62	72	dopamine D2 receptor	27.0	19.2	13.0	7.3
104	259	cannabinoid CB2 receptor	30.0	17.8	15.9	4.9
22	10188	MAP kinase p38 alpha	29.5	19.4	16.1	4.2
70	108	serotonin 2c (5-HT2c) receptor	30.7	19.8	18.7	5.4
34	12952	carbonic anhydrase IX	27.8	21.5	19.9	8.0
36	93	acetylcholinesterase	33.3	22.9	17.7	4.8
20	10980	vascular endothelial growth factor receptor 2	35.5	23.2	16.7	3.5
73	19905	melanin-concentrating hormone receptor 1	29.0	19.6	16.9	14.6
66	107	serotonin 2a (5-HT2a) receptor	35.2	20.2	20.9	4.6
103	87	cannabinoid CB1 receptor	31.8	31.3	14.9	9.9
91	17045	cytochrome P450 3A4	39.5	23.7	23.2	4.7
89	11140	dipeptidyl peptidase IV	36.9	27.3	17.8	9.5
117	114	adenosine A1 receptor	31.9	26.9	21.1	12.6
68	90	dopamine D4 receptor	33.4	20.6	24.3	14.2
255	100166	kinesin-like protein 1	40.4	28.4	19.0	6.9
26	13001	matrix metalloproteinase-2	33.1	27.6	24.9	9.4
92	104	monoamine oxidase B	38.5	24.8	21.5	10.6
40	65	cytochrome P450 19A1	31.8	30.1	21.3	12.4
55	61	muscarinic acetylcholine receptor M1	38.1	27.4	25.9	6.5
99	10280	histamine H3 receptor	36.6	29.5	20.3	12.2
53	51	serotonin 1a (5-HT1a) receptor	34.8	27.0	24.8	12.8
77	100	norepinephrine transporter	41.2	28.8	25.9	5.6
155	10260	vanilloid receptor	40.0	31.1	20.8	10.7
76	52	alpha-2a adrenergic receptor	40.5	28.3	28.0	6.8
153	11365	cytochrome P450 2D6	42.0	24.6	28.8	8.7

[a] The bottom 30 activity classes yielding the lowest overall search performance are reported and ranked according to increasing average recovery rate (i.e., bottom-up) of ECFP4 (1NN and 10NN) and MACCS (1NN and 10NN) calculations for selection sets equal to the number of ADCs. For each activity class, the ChEMBL target ID and target name are provided and average recovery rates are reported (in percent).

class, compound *recovery rates* (RRs) were then calculated by determining the ratio of active compounds contained in each class-specific selection set over all available ADCs. For example, if an activity class contained 200 compounds, 190 ADCs were available. If a search trial recovered 95 of these active compounds within a selection set of 190 database compounds (equal to the number of ADCs), the recovery rate would be 50%. Individual RRs were then averaged over all 100 trials for each activity class. Receiver operating characteristic (ROC) curves[26] and ROC area under the curve (AUC) values[26] were also calculated for averaged search trials. The initial use of ADC-based selection sets ensured that selection set sizes were balanced with respect to the size of an activity class and that in each case, a perfect similarity search outcome with 100% recovery rate (and 100% search specificity of the calculations) was principally possible. Subsequently, larger selection sets of two to three times the number of ADCs (e.g., 380 or 570 compounds for the example given above) were also considered. Finally, average RRs were also calculated for all classes for a constant database selection set size equal to the largest number of ADCs among all activity classes, i.e., 1783 compounds, which corresponded to ~0.18% of the screening database. MACCS and ECFP4 were generated using the Molecular Operating Environment[27] and Pipeline Pilot,[28] respectively, and all search calculations were carried out with in-house generated Java scripts.

RESULTS AND DISCUSSION

Similarity Search Profile. All 266 activity classes extracted from ChEMBL were subjected to systematic fingerprint search calculations in order to generate a similarity search profile of the database. From each class, 100 compound reference sets were randomly selected and for each combination of a fingerprint and a search strategy, 100 independent search trials were carried out in order to obtain statistically relevant data, which amounted to a total of ~160 000 search trials with multiple reference compounds. Figure S1 of the Supporting Information reports the resulting similarity search profiles for the three alternative nearest neighbor search strategies, and Figure 1 shows three representative profile subsets (for 20 activity classes, 120−139). In addition, Supporting Information Table S1 also reports 1NN recall

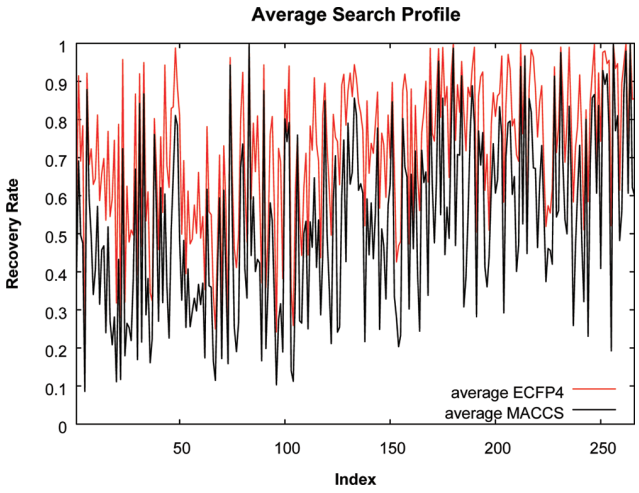


Figure 3. Similarity search profile for large database selection sets. For MACCS (black) and ECFP4 (red), recovery rates averaged over all three search strategies and for a constant database selection set size of 1783 molecules (see text) are plotted for all 266 activity classes. Index reports the consecutively numbered activity classes according to Supporting Information Table S1.

rates for each class. The profiles revealed the anticipated differences in global search performance between ECFP4 and MACCS. With only two exceptions, ECFP achieved consistently higher recovery rates (RRs). Furthermore, the profiles also illustrated the general compound class-dependence of fingerprint/similarity search calculations, with in part significantly varying RRs for each fingerprint. Importantly, however, the profiles revealed a perhaps unexpected success rate of 2D fingerprint searching on this large array of activity classes. In 203 of 266 test cases (~76%), an RR of at least 50% was obtained with at least the better performing fingerprint and at least one of the three different search strategies. It should be noted that these RRs were achieved for generally small selection set sizes equal to the number of ADCs for each class. For all search calculations, the average RR was 59.2% for ECFP4 and 38.1% for MACCS, which delineates a global performance range between approximately 40% and 60% compound recall achieved by a low-resolution (MACCS) and a high-resolution (ECFP4) 2D fingerprint. Given the large-scale character of these search calculations, these findings provide a realistic expectation value for 2D fingerprint searching on diverse compound classes.

Similarity Search Strategies. The 1NN, 5NN, and 10NN search strategies take contributions of reference compounds in different ways into account (see Methods and Materials). Because 1NN calculations only consider the match between a database molecule and the most similar reference compound, this strategy displays the tendency to select database molecules that are very similar to individual reference compounds. By contrast, because 10NN calculations take contributions of 10 reference compounds equally into account, this strategy shows a greater tendency to select database molecules that structurally differ from individual reference compounds. In Figure S2 of the Supporting Information, search profiles are compared for the 1NN, 5NN, and 10NN search strategies, and Table 1 reports the average RRs and standard deviations for each combination of a strategy and fingerprint. For both fingerprints, we observed that the global search performance decreased with increasing numbers of reference compound contributions (i.e., from 1NN over 5NN to 10NN), with an overall decline of ~8% for ECFP4 and 13% for MACCS. For all search calculations, standard deviations of ~7% or 8% were observed, which reflected the (limited) influence of reference set composition on the search results. Thus, we found that 1NN was the globally preferred nearest neighbor search strategy, yielding an average RR of 63.6% and 45.3% for ECFP4 and MACCS, respectively. Although the differences between individual search strategies were not very large, maximally on the order of 10%, selecting database molecules that were most similar to individual reference compounds globally produced highest RRs on the ChEMBL activity classes. These findings were consistent with the notion that compound data sets from medicinal chemistry typically contain different series of analogs, which are often easier to detect when applying the 1NN rather than other *k*NN search strategies.

Enrichment Behavior. We also studied the enrichment characteristics in database selection sets of increasing size. Figure S3 of the Supporting Information shows similarity search profiles for the original selection set sizes and selection sets that were doubled in size, and Figure 2 shows representative profile subsets for 20 activity classes and two fingerprint/search strategy combinations. Profile subsets for the remaining four fingerprint/strategy combinations are shown in Figure S4 of the Supporting Information. For both fingerprints, we consistently observed only slight increases in RRs of a few percent when selection sets were doubled or tripled in size (data not shown). Thus, these fingerprint search calculations were generally characterized by an early enrichment of active compounds in database selection sets. This meant that correctly identified active compounds often appeared at relatively high positions in the Tc-based similarity rankings. These finding were also consistent with the observation that active compounds were preferentially detected by matching the most similar reference compound (1NN), which typically yields higher similarity values than Tc average calculations and hence increases the probability of higher ranking positions. Figure S5 of the Supporting Information shows representative ROC curves for ECFP4 and MACCS 1NN calculations at different levels of search performance and ROC AUC values for all activity classes are reported in Table S2 of the Supporting Information.

Prioritization of Activity Classes. Our low/high resolution fingerprint similarity search strategy also made it possible to categorize activity classes according to their relevance for fingerprint benchmarking. We first identified particularly "easy" and "difficult" classes for 2D fingerprinting. In Table 2, the top 30 classes with overall highest search performance are reported. For ECFP4, the search performance was consistently very high in these cases, at or above the 90% levels, for both 1NN and 10NN calculations. For MACCS, RRs of close to or above 80% were also observed for 16 classes and all remaining RRs were above 60%. For the first three classes, almost perfect search results were obtained for both ECFP4 and MACCS. Taken together, the activity classes listed in Table 2 consistently yielded high to very high search performance for our prototypic low- and high-resolution 2D fingerprints. Thus, these classes are not suitable for fingerprint benchmarking because they yield RRs that go much beyond the typical performance range of 2D fingerprints, even for relatively small database selection sets. Importantly, the classes in Table 2 include a number of popular targets, for example, phospholipases, serine proteases, protein kinases, purinergic receptors, and other G protein coupled receptors that might often be attractive for benchmark trials. However, the uncritical choice of such data sets would provide artificially good results for fingerprint methods.

Table 4. Activity Classes Preferred for Evaluating 2D Fingerprints[a]

					average RR	
no.	target ID	target name	BMS	cpds per CSK	ECFP4	MACCS
4	11359	phosphodiesterase 4D	60	3.30	78.4	47.5
8	28	thymidylate synthase	44	4.29	72.3	49.4
9	11536	ghrelin receptor	228	3.52	63.0	34.1
10	8	tyrosine-protein kinase ABL	64	4.47	64.5	40.5
12	10434	tyrosine-protein kinase SRC	229	3.48	58.7	31.5
13	12670	tyrosine-protein kinase receptor FLT3	49	3.30	65.7	45.7
14	20014	serine/threonine-protein kinase Aurora-A	66	3.65	69.7	46.8
16	234	insulin-like growth factor I receptor	124	4.09	76.9	51.8
21	12261	c-Jun N-terminal kinase 1	51	5.94	78.7	43.3
35	12209	carbonic anhydrase XII	60	3.40	61.0	38.2
42	25	glucocorticoid receptor	169	4.41	55.6	31.9
44	36	progesterone receptor	99	5.79	67.9	36.2
52	43	beta-2 adrenergic receptor	88	2.11	69.2	48.2
54	219	muscarinic acetylcholine receptor M3	140	2.60	61.2	40.8
57	130	dopamine D3 receptor	214	3.23	57.5	33.0
59	105	serotonin 1d (5-HT1d) receptor	45	1.81	66.1	36.7
81	11336	neuropeptide Y receptor type 5	182	6.33	63.3	40.8
86	20174	G protein-coupled receptor 44	132	5.21	66.3	40.1
95	126	cyclooxygenase-2	117	5.92	56.0	32.7
98	11225	renin	183	5.34	68.8	31.6
105	12252	beta-secretase 1	246	3.31	61.7	37.3
112	11682	glycine transporter 1	66	3.95	78.3	53.1
113	134	vasopressin V1a receptor	110	2.54	72.0	46.5
115	116	oxytocin receptor	55	4.03	73.7	41.2
120	11265	somatostatin receptor 5	67	2.50	73.0	50.0
121	10475	neuropeptide Y receptor type 1	66	4.70	62.8	36.9
129	12679	C5a anaphylatoxin chemotactic receptor	67	3.54	78.6	42.7
140	10579	C−C chemokine receptor type 4	87	2.73	65.9	44.5
142	11575	C−C chemokine receptor type 2	178	6.11	71.2	43.5
143	18061	sodium channel protein type IX alpha subunit	58	5.26	78.8	55.3
146	237	leukotriene A4 hydrolase	87	3.20	76.4	51.3
147	276	phosphodiesterase 4A	38	2.61	73.3	46.7
148	11534	cathepsin S	298	3.61	59.6	32.9
152	10198	voltage-gated potassium channel subunit Kv1.5	97	3.94	67.0	33.6
163	10498	cathepsin L	67	3.29	65.7	40.2
168	12911	cytochrome P450 2C9	31	2.27	63.6	33.8
171	12968	orexin receptor 2	43	4.55	74.3	47.6
181	100579	nicotinic acid receptor 1	80	4.47	74.6	46.9
186	100126	serine/threonine-protein kinase B-raf	73	2.94	71.8	38.7
195	10378	cathepsin B	56	2.56	61.4	41.3
196	10417	P2X purinoceptor 7	69	3.26	70.9	36.1
210	10752	inhibitor of nuclear factor kappa B kinase beta subunit	46	3.81	70.8	40.1
211	10773	interleukin-8 receptor B	76	6.85	69.0	47.1
213	11631	sphingosine 1-phosphate receptor Edg-1	51	3.59	76.3	52.6
220	10927	urotensin II receptor	74	3.00	75.4	46.4
230	11085	melatonin receptor 1B	52	3.61	78.4	56.2
234	11442	liver glycogen phosphorylase	104	5.10	79.3	50.0
238	11279	metabotropic glutamate receptor 1	84	4.37	72.6	46.7
241	11488	estradiol 17-beta-dehydrogenase 3	39	5.30	76.3	48.4
250	12840	macrophage colony stimulating factor receptor	59	5.57	74.3	40.9

[a] Listed are 50 activity classes that met our selection criteria for benchmarking relevance (as described in the text). These classes are ordered by their consecutive numbers. Average RRs over all search strategies are reported (in percent) for ECFP4 and MACCS and a database selection set size of 1783 compounds (see text for details). For each class, the total number of Bemis and Murcko scaffolds (BMS)[29] and the compound-to-carbon skeleton (CSK)[30] ratio (cpds per CSK) are reported.

dx.doi.org/10.1021/ci200199u |J. Chem. Inf. Model. 2011, 51, 1831–1839

In Table 3, we report the opposite end of the similarity search spectrum. Here the 30 activity classes with lowest search performance are listed. For ECFP4, these classes mostly resulted in RRs of ~20% to ~30%. For MACCS, many 10NN RRs were lower than 10% or even 5%, but 1NN RRs were still close to or above 20% in many instances. Therefore, fingerprint searching on none of these classes could *per se* be considered a failure. However, given their overall low search performance, the 2D fingerprints clearly approached their detection limits in these cases that also included a number of popular enzyme and G protein coupled receptor targets. Hence, these activity classes might be more appropriate for the evaluation of similarity methods that employ more elaborate molecular representations or utilize 3D information.

In virtual screening benchmark calculations, compound recall is typically evaluated on the basis of larger selection set sizes than the variably balanced selection set sizes that we utilized for our analysis up to this point. Often, 0.1−1% of the screening/background database are selected for recovery rate analysis. Therefore, we also calculated average RRs over all search strategies for a selection set of constant size, i.e., the largest individual selection set utilized in our study, which contained 1783 compounds corresponding to ~0.18% of our background database. On average, this constant selection set corresponded to an approximately 6-fold increase in selection set size for the ChEMBL activity classes. The resulting similarity search profile for all 266 classes is displayed in Figure 3. As expected, for this comparably large selection set, average RRs were higher than originally observed, with 71.9% and 52.7% for ECFP4 and MACCS, respectively (again with standard deviations of ~8%). However, the increase relative to the originally observed RRs was also limited with approximately 11% for ECFP4 and 14% for MACCS, consistent with the generally observed early enrichment characteristics.

On the basis of these results, we then prioritized activity classes that were considered particularly suitable for benchmarking of 2D fingerprints. Therefore, in light of the observed search performance range for our fingerprint prototypes, we selected activity classes that minimally yielded more than 30% compound recall for MACCS (thus ensuring a meaningful base performance) and maximally less than 80% recall for ECFP4 (thus leaving room for further improvements) and that differed by more than 20% in relative search performance (thus reflecting the overall performance range). On the basis of these selection criteria, we identified a total of 50 activity classes that we would assign a high priority for the evaluation and comparison of alternative 2D fingerprints. As reported in Table 4, these classes covered a variety of different target families including a number of prominent therapeutic targets and were generally characterized by the presence of large numbers of distinct scaffolds and low compound-to-carbon skeleton ratios (i.e., structural heterogeneity). These activity classes can also be obtained via the following URL (please, see the "Downloads" section): http://www.lifescienceinformatics.uni-bonn.de.

CONCLUSIONS

Here we have reported a large-scale similarity search investigation to systematically analyze compound activity classes extracted from the ChEMBL database, a major public domain repository of compounds originating from medicinal chemistry sources. For similarity search profiling of ChEMBL, we selected two prototypic 2D fingerprints that represent markers for current performance levels of popular fingerprints. On the basis of systematic search calculations, we also determined the global performance range defined by these fingerprints covering compound data sets directed against the spectrum of current pharmaceutical targets. Overall, the search results were rather encouraging, more so than we anticipated, indicating that many activity classes can be well treated using 2D fingerprints, despite their relative simplicity. Other general trends emerged concerning preferred search strategies and early enrichment characteristics of active compounds that corroborated earlier findings. Furthermore, by comparing the search performance of our low- and high-resolution fingerprint standards, we have identified activity classes that were unsuitable for 2D fingerprint evaluation, because they yielded artificially high search performance, and other classes that represented rather difficult test cases where 2D fingerprints approached their limits. We also prioritized 50 activity classes as particularly useful for 2D fingerprint evaluation in light of the search characteristics we observed. Taken together, these findings should also aid in the design of meaningful benchmark investigations.

ASSOCIATED CONTENT

Supporting Information. Supplementary Table S1 listing all activity classes extracted from the ChEMBL database and reports recovery rates for each class, Supplementary Table S2 reporting ROC AUC values, and Supplementary Figures S1−S5 showing similarity search profiles, comparisons of similarity search strategies, enrichment characteristics for all activity classes, enrichment characteristics for activity class subsets, and exemplary ROC curves for activity classes at different similarity search performance levels, respectively. This information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

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

REFERENCES

(1) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.

(2) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

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

(4) Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdiscip. Rev.: Comput. Molec. Sci.* **2011**, *1*, 260–282.

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

(6) Nisius, B.; Bajorath, J. Reduction and recombination of fingerprints of different design increase compound recall and the structural diversity of hits. *Chem. Biol. Drug Des.* **2010**, *75*, 152–160.

(7) Hert, J.; Willet, P.; Wilton, D. J. Comparison of Fingerprint-based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(8) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information. *J. Med. Chem.* **2005**, *48*, 7049–7054.

(9) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New Methods for Ligand-Based Virtual Screening:

Use of Data Fusion and Machine Learning to Enhance the Effectiveness of Similarity Searching. *J. Chem. Inf. Model* **2006**, *46*, 462–470.

(10) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two- Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010**, *53*, 5707–5715.

(11) Gardiner, E. J.; Holliday, J. D.; O'Dowd, C.; Willett, P. Effectiveness of 2D Fingerprints for Scaffold Hopping. *Future Med. Chem.* **2011**, *3*, 405–414.

(12) Stumpfe, D.; Bill, A.; Novak, N.; Loch, G.; Blockus, H.; Geppert, H.; Becker, T.; Hoch, M.; Schmitz, A.; Kolanus, W.; Famulok, M.; Bajorath, J. Targeting Multi-Functional Proteins by Virtual Screening: Structurally Diverse Cytohesin Inhibitors with Differentiated Biological Functions. *ACS Chem. Biol.* **2010**, *5*, 839–849.

(13) Stumpfe, D.; Bajorath, J. Applied virtual screening: strategies, recommendations, and caveats. In *Methods and Principles in Medicinal Chemistry. Virtual Screening. Principles, Challenges, and Practical Guidelines*; Sotriffer, C., Ed.; Wiley-VCH: Weinheim, 2011; pp 73−103.

(14) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model* **2010**, *50*, 205–216.

(15) 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.

(16) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-Scale Systematic Analysis of 2D Fingerprint Methods to Improve Virtual Screening Enrichments. *J. Chem. Inf. Model* **2010**, *50*, 771–784.

(17) *MDL Drug Data Report*; Accelrys: San Diego, CA, 2011.

(18) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model* **2009**, *49*, 169–184.

(19) 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.

(20) *PubChem*; National Center for Biotechnology Information: Bethesda, MD, 2010; http://pubchem.ncbi.nlm.nih.gov/ (accessed December 10, 2010).

(21) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.

(22) *ChEMBL*; European Bioinformatics Institute (EBI): Cambridge, 2011; http://www.ebi.ac.uk/chembl/ (accessed March 2, 2011).

(23) *MACCS Structural keys*; Accelrys: San Diego, CA, 2011.

(24) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model* **2010**, *50*, 742–754.

(25) 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.

(26) Bradley, A. P. The Use of the Area under the ROC Curve for the Evaluation of Machine Learning Algorithms. *Pattern Recog.* **1997**, *30*, 1145–1159.

(27) *Molecular Operating Environment*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2010.

(28) *Scitegic Pipeline Pilot*; Accelrys, Inc.: San Diego, CA, 2010.

(29) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(30) Xu, Y.-J.; Johnson, M. Using Molecular Equivalence Numbers to Visually Explore Structural Features that Distinguish Chemical Libraries. *J. Med. Chem.* **2002**, *42*, 912–926.

Summary

A large-scale similarity search analysis of compound activity classes from the ChEMBL database was presented. The search investigation showed that 2D fingerprints are effective molecular representations for similarity searching despite their simplicity. Most of the active compounds have been recovered at small database selection set sizes revealing general early enrichment potential. The large-scale character of the study and the fingerprints used enabled us to define a performance range of 2D fingerprints in search calculations on pharmaceutical targets. In addition, activity classes have been grouped depending on how difficult they are for standard 2D fingerprint searching.
The supporting information of this publication can be found under the following URL: http://dx.doi.org/10.1021/ci200199u.

Next, we further analyze similarity searching and focus on the mechanistical aspect of 2D fingerprint searching. The aim of the follow-up study is to determine on a large-scale how fingerprints recover active compounds, even if they have only little similarity to the reference compounds.

Chapter 2

How do 2D fingerprints detect structurally diverse active compounds? Revealing compound subset-specific fingerprint features through systematic selection

Introduction

In the previous study, it was shown that similarity searching using 2D fingerprints yields high search performance. This study follows up on the analysis of 2D fingerprint searching and determines how 2D fingerprints work mechanistically. Key focus of this study is the identification of the mechanism by which 2D fingerprints facilitate scaffold hopping. Therefore, two different feature selection methods, namely Kullback-Leibler divergence and gain ratio, are applied to systematically reduce two different atom environment fingerprints that are used to search compound activity classes of different structural diversity. The recall of reduced and unmodified fingerprints is analyzed and fingerprint features are identified that are responsible for the recovery of distinct subsets of active compounds.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

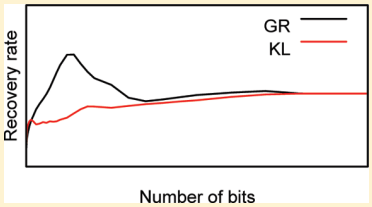
How Do 2D Fingerprints Detect Structurally Diverse Active Compounds? Revealing Compound Subset-Specific Fingerprint Features through Systematic Selection

Kathrin Heikamp and Jürgen Bajorath*

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

Supporting Information

ABSTRACT: In independent studies it has previously been demonstrated that two-dimensional (2D) fingerprints have scaffold hopping ability in virtual screening, although these descriptors primarily emphasize structural and/or topological resemblance of reference and database compounds. However, the mechanism by which such fingerprints enrich structurally diverse molecules in database selection sets is currently little understood. In order to address this question, similarity search calculations on 120 compound activity classes of varying structural diversity were carried out using atom environment fingerprints. Two feature selection methods, Kullback−Leibler divergence and gain ratio analysis, were applied to systematically reduce these fingerprints and generate alternative versions for searching. Gain ratio is a feature selection method from information theory that has thus far not been considered in fingerprint analysis. However, it is shown here to be an effective fingerprint feature selection approach. Following comparative feature selection and similarity searching, the compound recall characteristics of original and reduced fingerprint versions were analyzed in detail. Small sets of fingerprint features were found to distinguish subsets of active compounds from other database molecules. The compound recall of fingerprint similarity searching often resulted from a cumulative detection of distinct compound subsets by different fingerprint features, which provided a rationale for the scaffold hopping potential of these 2D fingerprints.



INTRODUCTION

Molecular fingerprints have for long been used for chemical similarity searching.[1,2] These descriptors typically consist of bit string representations of structural features or other molecular properties. Fingerprint representations calculated from molecular graphs, thus termed two-dimensional (2D) fingerprints, were among the early descriptors for similarity searching.[1] Original 2D fingerprint designs, such as structural keys,[3] were based on fragment dictionaries. In such fingerprints, each bit accounts for the presence or the absence of a predefined substructure in a compound. In addition to dictionary-based fingerprints, the introduction of topological 2D fingerprints that assemble connectivity pathways through molecules and represent them in a hashed format[4] has been another milestone event in this field. To this date, most—but not all—available 2D fingerprints account for structural and/or topological features.[2,5]

In similarity searching, the overlap between fingerprints of reference and database compounds is quantified as a measure of molecular similarity,[1] and database compounds are ranked in the order of decreasing fingerprint similarity to reference molecule(s) such that the structurally most similar compounds are highest on the list. As is generally the case with structural descriptors, 2D fingerprints do not capture biological activity information, and hence there is no well-defined relationship between (calculated) fingerprint similarity and (observed) biological activity similarity.[5,6] Of course, because fingerprints detect compounds that are structurally most similar to active reference molecules, these compounds have a certain probability to exhibit a similar activity. However, as structural similarity decreases between reference and ranked database compounds, calculated similarity and activity similarity are not related to each other, and it is generally difficult to select active compounds.[6]

Nevertheless, 2D fingerprints have a history of successful applications in virtual screening for novel active compounds.[2,5,6] Here the identification of structurally highly similar or analogous compounds, which one can easily accomplish using fingerprints, is much less interesting than the search for structurally diverse molecules having similar activity, a challenge often referred to as scaffold hopping.[7,8] However, although the scaffold hopping ability of relatively simple structural descriptors, such as 2D fingerprints, has often been questioned, it has clearly been demonstrated that 2D fingerprints are capable of enriching structurally diverse active compounds in small database selection sets, both in benchmark trials[9,10] and prospective virtual screening applications.[11,12] Methodological foundations for the rather surprising virtual screening potential of 2D fingerprints have, however, largely remained unclear.

The virtual screening performance of 2D fingerprints has been much improved over the years through the introduction of search

Received: June 16, 2011
Published: July 27, 2011

strategies for multiple reference compounds[13−15] and various new fingerprint designs.[6,16] Currently, fingerprints that capture atom environment information, such as Molprint2D[17,18] and especially extended connectivity fingerprints (ECFPs),[19] often produce the highest compound recall in comparative benchmark trials.[9,10,16] These types of fingerprints also capture topological information, similar to (yet algorithmically distinct from) the prototypic atom pathway fingerprints.[4] ECFPs systematically determine circular atom environments up to a given bond diameter in compounds and assemble these structural/topological features in a molecule-specific manner.[19] Feature arrays resulting from different compounds are then also quantitatively compared on the basis of Tanimoto similarity[1] or other standard similarity metrics.

Furthermore, the virtual screening performance of 2D fingerprints has also been increased through the introduction of fingerprint engineering strategies that modify fingerprint formats in specific ways, for example, by eliminating fingerprint bits (features) that are not critical for detecting a specific biological activity (fingerprint reduction)[20] or by combining the most important bit segments from fingerprints of different design (fingerprint hybridization).[21] Such modifications convert generally applicable fingerprints into compound class-specific versions with increased class-specific recall performance.[6,16] Such fingerprint engineering techniques have revealed that individual bit positions influence the outcome of similarity search calculations in different ways, depending on the compound classes under investigation.[6,16]

Fingerprint reduction techniques depend on the application of feature ranking and selection methods that make it possible to evaluate the importance of individual bits (features) for detecting compounds belonging to a given activity class. For fingerprint reduction, Kullback−Leibler (KL) divergence analysis[22] from information theory[23] has been originally applied[20] and has thus far been a method of choice.[16]

In this study, we have carried out comparative feature selection analysis for atom environment fingerprints and a large number of activity classes to revisit the scaffold hopping potential of 2D fingerprints and address the question why such fingerprints are capable of recognizing active compounds having little structural and topological resemblance to reference molecules. We have compared KL divergence with gain ratio (GR)[24] analysis, another information−theoretic approach, for activity classes of varying degrees of structural diversity and monitored compound recall characteristics of systematically reduced fingerprints. On the basis of this analysis, we have found that for structurally diverse activity classes, small numbers of fingerprint features are responsible for distinguishing different subsets of active compounds from the background database, resulting in a cumulative detection of such compound subsets. Taken together, these findings suggest a plausible mechanism for scaffold hopping by state-of-the-art 2D fingerprints.

METHODS AND MATERIALS

Fingerprints. For our analysis, two atom environment fingerprints were selected, Molprint2D[17,18] and an ECFP with bond diameter four (ECFP4).[19] These 2D fingerprints have often yielded high compound recall rates in comparative fingerprint benchmark investigations and are considered state-of-the-art.[16] Molprint2D was calculated using public domain software tools[17] and ECFP4 using Pipeline Pilot.[26]

Activity Classes. A total of 120 activity classes, each containing at least 200 compounds active against human targets, was extracted from BindingDB.[25] Selected compounds were required to be rule-of-five compliant, have at least 1 μM potency (K_i or IC_{50} values), and consist of atom types compatible with the calculation of the Molprint2D fingerprint.[17] Although fingerprint search calculations do not take potency information into account, the potency threshold was applied to avoid the inclusion of very weakly or borderline active compounds in similarity searching. Atom typing and rule-of-five calculations were carried out using Pipeline Pilot.

The relative structural diversity of activity classes was assessed by determining the number of Bemis−Murcko scaffolds (BMS)[27] and corresponding carbon skeletons (CSK), also referred to as cyclic skeletons,[27,28] per class and by calculating the average ratio of compounds per BMS and CSK. BMS consist of all rings and linker fragments between rings after removal of R-groups from compounds. They are further reduced to CSK by setting all bond orders to one and by converting all heteroatoms to carbon.

Feature Ranking Methods. For fingerprint feature ranking, KL divergence[22] and GR analysis[24] were carried out. Both methods determine the ability of individual fingerprint bits/features to differentiate between active and background database (inactive) compounds. In the following, the terms bit and feature are synonymously used.

The KL divergence measures the difference between two value distributions $p(x)$ and $q(x)$. When $p(x)$ describes the probability of a value (0 or 1) of bit x in active and $q(x)$ the probability of a value of this bit in inactive compounds, the KL divergence D is defined as

$$D(p(x)||q(x)) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$$

Hence, bit positions can be ranked on the basis of D to assign high priority to bits that are preferentially set on in active compounds. KL divergence calculations have thus far been applied for the identification of bit settings that are characteristic of activity classes and fingerprint reduction.[20,21]

Furthermore, GR is a statistical feature ranking approach that is based on normalized mutual information[23] (MI). MI measures the amount of information a variable X contributes to the values of another variable Y. Given the definition that X describes the value of a fingerprint bit (0 or 1) and Y the activity states (active or inactive), MI determines how much information about the correct activity state of test compounds is specified by a bit:

$$\begin{aligned} MI(X;Y) &= H(Y) - H(Y|X) \\ &= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)} \end{aligned}$$

where H is the information entropy, $p(x)$ and $p(y)$ are the probability functions of X and Y, and $p(x,y)$ the joint probability function. Thus, MI represents the difference between the entropy of the activity states and the entropy of the activity states under the condition that the value of a specific bit is known. GR is then obtained by dividing MI by the overall entropy of this bit:

$$\begin{aligned} GR(X;Y) &= \frac{MI(X;Y)}{H(X)} \\ &= \frac{-\sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}}{\sum_x p(x) \log_2 p(x)} \end{aligned}$$

dx.doi.org/10.1021/ci200275m |J. Chem. Inf. Model. 2011, 51, 2254–2265

For a given activity class, bits are ranked on the basis of their KL or GR values, which reflect the ability of each individual bit to distinguish between active and database compounds. For example, if a bit would occur in all active but no database compounds, then this ability would be maximal.

Reduced Fingerprints. For each activity class, reduced Molprint2D and ECFP4 fingerprints were generated on the basis of the KL and GR feature rankings by incrementally extending sets of highly ranked bits. Bit probability distributions within active and database compounds were derived on the basis of the active reference set and the database (without active compounds), respectively. In order to adjust probabilities of zero, which would lead to nondefined KL divergence and GR calculations, an *m*-estimate correction was applied by adding a (hypothetical) fingerprint reflecting the average fingerprint bit settings of the database compounds to the fingerprints of each reference set (and an equivalent correction was applied to the database). In addition, prior to feature ranking, bits were removed that only occurred in very few databases but no active compounds because their calculated feature weights were negligible.

Following feature ranking, reduced fingerprints were generated as follows, beginning with the smallest possible versions: The top five bits were added one-by-one, hence generating five minimal fingerprint versions consisting of only one to five bits. Then, bits from rank 6 to 20 were added in 5-bit increments, hence generating three further extended fingerprint versions. Furthermore, bit positions ranked from 21 to 100 were added in 10-bit increments, bits ranked from 101 to 200 in 20-bit, bits ranked from 201 to 500 in 50-bit, and bit positions ranked from 501 to 1000 in 100-bit increments. Thus, in the design of reduced fingerprints, the most highly ranked bits were utilized on an individual basis, and bit positions of decreasing significance were considered in increments of increasing size. For each activity class and original fingerprint, 32 reduced fingerprint versions were generated. Thus, in total, 7680 reduced fingerprints were investigated, in addition to the unmodified Molprint2D and ECFP4 fingerprints.

Fingerprint Similarity Searching. From each activity class, 100 different subsets of 20 compounds each were randomly selected as reference sets for 100 independent similarity search trials. In each case, the remaining active compounds were added as potential hits to a background database consisting of 1.44 million compounds randomly selected from ZINC.[29] As a search strategy, 10 nearest neighbor (10NN) calculations[13,14] were carried out. Following this strategy, each database compound is compared to each individual reference molecule by calculating pairwise Tanimoto similarity,[1] and the final similarity score of a database compound is obtained by averaging the 10 highest individual similarity values. Compound recall rates were determined for a selection set of 5000 database compounds (i.e., ~0.35% of the background database). This selection set size was chosen because several of the activity classes contained more than 1000 compounds. Because of the differences in activity class size, there were different probabilities for the random enrichment of active compounds in database selection sets. However, since we did not compare search results across different activity classes but analyzed the performance of unmodified and reduced fingerprints on individual activity classes, differences in random enrichment probabilities across different classes did not need to be considered in the context of our analysis. Importantly, for ranking of active compounds, we applied a pessimistic ranking strategy such that all background database compounds were included in the ranking prior to the last recovered active compound within a selection set. This means that compounds having the same similarity value were not given the same but subsequent ranks. Otherwise, the top 5000 ranks might yield more than 5000 compounds, which would bias the search results.

Given the number of different fingerprint versions and reference sets for each activity class, a total of 792 000 similarity search trials were carried out for our analysis.

RESULTS AND DISCUSSION

Study Objective. We have been interested in exploring the question of how compound recall characteristics of fingerprint search calculations might be rationalized, with a particular focus on scaffold hopping ability. In order to generate a substantial body of primary similarity search data for further analysis, we have initially carried out systematic fingerprint search calculations on 120 different activity classes using 2 state-of-the-art 2D fingerprints. To obtain statistically meaningful results, 100 different reference sets per activity class were utilized. On the basis of these data, compound recall characteristics have then been explored in detail. We have systematically applied and compared feature selection methods to identify fingerprint features that were responsible for the recall of different compound classes. In this context, we have then focused on the question why 2D fingerprints that emphasize structural/topological resemblance have the potential to enrich structural diverse active compounds in database selection sets. In the following, we describe the composition of data sets used for our analysis, present the results of large-scale fingerprint searching, discuss the findings of comparative feature selection, compare the search performance of unmodified and reduced fingerprint representations, and attempt to rationalize the scaffold hopping potential of the 2D fingerprints studied here.

Composition of Data Sets. The composition of the 120 activity classes utilized for systematic fingerprint similarity searching is reported in Table S1 of the Supporting Information. In the text, activity classes are referred to by numbers given in Table S1, Supporting Information. These compound data sets contained between 200 and 1425 different enzyme inhibitors or receptor ligands with varying degrees of intraclass structural diversity, reflected by differences in BMS and CSK distributions and compound-to-BMS and -CSK ratios, as also reported in Table S1, Supporting Information. For our analysis, we distinguish between structurally more homogeneous and heterogeneous (diverse) compound classes. Therefore, compound-to-scaffold ratios were determined to estimate intraclass structural diversity. This approach was considered more robust than, for example, the calculation of average compound similarity values. For example, if a data set would consist of a few topologically distinct scaffolds, each of which would be represented by a larger number of analogs, then average pairwise similarity values might be relatively low, although intraclass structural diversity would be limited in this case. However, the compound-to-scaffold ratio would be rather high, which would better account for limited intraclass diversity. In our set of activity classes reported in Table S1, Supporting Information, structural homogeneous classes are characterized by the presence of comparably small numbers of BMS and CSK and large compound-to-BMS and -CSK ratios, with about 5−10 or more compounds per CSK. For activity classes of increasing structural diversity, the compound-to-CSK ratio is decreasing to about 2−3 compounds per CSK.

Table 1. Average Recovery Rates for ECFP4[a]

	unmodified	GR		KL	
no.	RR	RR	no. bits	RR	no. bits
1	45.5	69.3	120	45.3	800
2	50.8	75.4	140	53.8	700
3	37.6	61.1	140	37.6	700
4	49.4	65.0	140	46.9	700
5	96.6	98.7	600	98.5	60
6	51.8	67.2	120	50.6	800
7	52.4	67.7	120	50.3	800
8	81.9	84.1	90	79.9	700
9	72.1	84.6	80	71.4	350
10	60.1	60.5	700	59.6	700
11	74.2	74.0	700	73.6	700
12	51.6	66.0	120	50.8	700
13	71.7	76.6	160	69.7	800
14	93.8	95.8	120	94.4	700
15	68.0	67.0	140	65.6	700
16	58.4	77.7	120	64.0	140
17	86.9	87.5	600	87.9	700
18	93.3	97.3	80	95.8	40
19	48.1	61.0	140	45.1	900
20	58.2	66.1	120	55.4	800
21	58.2	62.4	700	60.9	700
22	49.5	55.6	700	53.7	700
23	65.9	70.0	600	67.9	700
24	79.6	92.8	80	80.2	60
25	64.2	87.9	140	67.1	700
26	62.5	88.5	120	70.7	120
27	73.6	91.1	120	81.9	80
28	67.6	80.3	120	68.1	700
29	72.4	80.4	100	73.2	700
30	85.2	85.9	140	85.0	600
31	52.5	65.6	140	54.7	800
32	44.0	64.8	140	50.3	800
33	79.1	92.6	120	82.3	90
34	76.3	76.3	700	76.7	700
35	72.9	74.7	800	74.3	800
36	87.7	93.1	70	88.0	350
37	71.6	92.0	160	78.5	40
38	78.2	82.2	100	78.7	600
39	38.6	42.5	120	40.0	700
40	34.6	35.7	700	36.3	700
41	30.0	34.4	120	34.0	700
42	53.0	64.7	120	49.2	700
43	71.3	72.1	140	71.0	600
44	85.1	85.1	90	83.0	800
45	69.6	71.2	60	65.1	600
46	79.5	78.4	800	78.1	700
47	79.3	78.0	800	77.6	700
48	55.6	77.8	120	65.5	700
49	77.0	83.8	120	76.2	700
50	53.2	80.7	200	58.0	800
51	85.4	85.4	90	84.0	600
52	63.8	75.1	120	62.2	600

Table 1. Continued

	unmodified	GR		KL	
no.	RR	RR	no. bits	RR	no. bits
53	87.0	92.2	160	87.0	700
54	27.7	39.3	160	29.4	800
55	48.6	63.0	100	48.5	700
56	64.7	93.4	70	85.2	15
57	75.0	81.9	100	72.7	700
58	82.2	92.0	90	80.2	700
59	85.6	89.9	600	90.0	600
60	88.2	90.8	250	89.9	600
61	81.2	84.1	90	82.5	80
62	64.8	77.1	180	68.3	800
63	61.3	81.7	120	60.3	140
64	82.9	84.5	600	84.9	600
65	36.9	54.7	120	43.2	600
66	64.2	93.8	140	85.5	10
67	67.9	83.5	140	71.4	700
68	53.8	82.3	160	67.4	3
69	73.7	90.2	100	78.2	10
70	56.3	86.5	140	76.8	1
71	39.4	56.0	120	41.4	800
72	78.5	81.0	90	74.0	800
73	71.6	75.2	90	66.4	800
74	66.2	70.6	90	62.7	700
75	78.4	93.5	60	82.8	40
76	66.9	82.3	90	69.1	350
77	66.3	71.3	180	69.3	800
78	45.7	64.7	140	44.7	800
79	50.6	67.6	160	48.4	800
80	69.7	79.3	120	67.8	800
81	84.5	93.9	120	85.4	50
82	49.6	64.4	120	46.0	700
83	55.7	82.4	100	58.0	600
84	72.8	80.5	140	72.4	800
85	61.4	79.1	140	61.0	800
86	50.9	70.1	160	48.0	700
87	62.5	76.2	70	57.5	700
88	42.7	68.8	120	43.4	50
89	80.1	88.0	60	80.5	700
90	63.0	71.1	120	61.2	800
91	67.0	82.0	140	65.3	700
92	64.1	82.9	100	64.9	700
93	69.2	76.0	160	67.5	800
94	71.3	82.1	160	75.5	800
95	77.0	88.8	80	77.0	30
96	94.1	95.4	60	93.5	140
97	37.1	48.0	140	39.0	800
98	75.9	76.0	80	75.5	700
99	76.8	80.4	90	76.6	700
100	28.9	45.9	120	30.5	800
101	52.8	68.1	120	46.3	700
102	33.8	53.5	120	34.7	700
103	65.0	70.0	120	66.7	700
104	41.1	44.4	100	40.1	800
105	39.2	61.7	140	38.8	800

dx.doi.org/10.1021/ci200275m |J. Chem. Inf. Model. 2011, 51, 2254–2265

Table 1. Continued

	unmodified	GR		KL	
no.	RR	RR	no. bits	RR	no. bits
106	35.6	40.2	120	33.9	800
107	85.9	87.2	80	85.0	600
108	41.9	73.1	160	54.0	160
109	85.2	87.0	700	86.5	700
110	77.3	84.4	100	77.5	700
111	80.3	90.4	100	87.3	600
112	78.5	97.5	120	91.5	20
113	55.2	68.6	120	56.3	200
114	58.5	68.0	140	59.2	800
115	64.8	73.6	80	63.7	700
116	73.1	72.5	450	72.4	600
117	81.0	91.7	80	89.6	40
118	64.7	62.9	500	64.0	700
119	59.4	67.3	140	58.8	700
120	27.8	45.6	140	29.8	450

[a] Average recovery rates (RR) over 100 independent trials are reported for full-length (unmodified) ECFP4 and the best-performing reduced fingerprints derived by GR and KL divergence. The length of each reduced fingerprint (no. bits) is specified. Recovery rates are calculated for a database selection size of 5000 compounds.

Large-Scale Similarity Searching. We first compared the overall search performance of unmodified ECFP4 and Molprint2D. Average recall rates over 120 activity classes were 63.0% and 64.5% for Molprint2D and ECFP4, respectively. ECFP4 produced higher recall rates for 70 classes and Molprint2D for 47 classes (with equal performance in three cases). Hence, both fingerprints reached comparable performance levels, but ECFP4 produced overall slightly better results. All search results for ECFP4 and Molprint2D are provided in Table 1 and Table S2 of the Supporting Information, respectively. In addition, histogram representations of recall rates according to Table 1 are provided in Figure 1a and b for 10 activity classes with the highest and lowest increase in recall, respectively, as a consequence of GR feature selection.

Global Effects of Feature Selection. Both Molprint2D and ECFP4 are combinatorial fingerprints that can, in principle, yield exceedingly large feature numbers (although this is typically not the case for small organic compounds). For feature selection studies, we considered the top 1000 features on the basis of GR and KL divergence ranking. We generally observed that reduced fingerprints, often of only small size, met or exceeded the search performance of unmodified ECFP4 and Molprint2D, as reported in Table 1 and Table S2 of the Supporting Information. However, there were notable global differences between the two alternative feature selection approaches. On average, the best-performing reduced fingerprints selected on the basis of KL divergence consisted of 252 and 578 bits for Molprint2D and ECFP4, respectively. For GR selection, the corresponding numbers were 182 bits for Molprint2D and 196 bits for ECFP4. Thus, the best-performing reduced fingerprints generated on the basis of GR contained fewer features than those generated on the basis of KL divergence. Furthermore, for Molprint2D and ECFP4, the best KL divergence-based fingerprints produced average recall rates of 65.3% and 65.9%, respectively, which slightly increased the recall rates of the full-length fingerprints (63.0% and 64.5%, respectively). However, GR-based reduced fingerprints achieved average recall rates of 74.3% and 75.4% for Molprint2D and ECFP4, respectively. Thus, the top-performing GR-based fingerprints consisted of fewer than 200 features and further increased the average recall rates of the unmodified fingerprints by approximately 10%. Given these differences observed in feature selection, we compared KL divergence and GR approaches in more detail.

Comparison of Feature Probabilities. Next we studied the probabilities of top-ranked bits selected by KL divergence and GR to occur in active and database compounds. For this purpose, many individual search trials on our activity classes were analyzed (averages over different references sets leading to different bit selections would not be meaningful to calculate). In Figure 2, the corresponding probabilities of the 100 top-ranked KL divergence and GR bits of ECFP4 are compared for two different activity classes. These results are representative of many comparisons we carried out and reflect a clear general trend we observed. It should be noted that the probabilities in Figure 2 are reported on a logarithmic scale. Because of the large number of database compounds, the probabilities of bit settings in the database might become very small. In Figure 2a, the probability histogram for the GR selection of ECFP4 features for a reference set of activity class no. 24 is shown, and in Figure 2b, the corresponding histogram for the KL divergence selection is shown. In Figure 2a, the top 25 bits in the GR-based histogram have a decreasing probability of occurrence in active compounds but no probability of occurrence in database compounds. Over the remaining bit positions, the active probability remains high, and the database probability slightly increases. Thus, the GR-based bit ranking reflects strong emphasis on probability differences between active and database compounds. By contrast, the corresponding KL divergence-based bit ranking in Figure 2b reveals a less systematic profile. In this case, bit positions are also highly ranked that have a detectable probability to occur in database compounds. In these cases, however, the active probabilities are very high. Other bit positions with lower active probability but no database probability occur at intermediate ranks. This profile phenotype can be rationalized on the basis of the KL divergence formula presented in the Methods and Materials Section. Thus, KL divergence calculations not only emphasize probability differences but also the magnitude of the active probability, ultimately leading to a selection compromise. Equivalent observations concerning the GR- and KL divergence-based probability histograms are made in Figure 2c and d, respectively, for a reference set of activity class 22 (that is structurally more homogeneous than class 24).

Hence, there were significant differences between the GR and KL divergence feature selections, which also applied to the actual features that were prioritized. We generally observed that the overlap between GR- and KL divergence-based rankings considerably varied for different activity classes. This is illustrated in Table 2 where the average GR versus KL divergence overlap of features selected for the 100 individual reference sets of activity classes 22 and 24 is reported. For the 20 most significant bits, the overlap was only small for class 22 but large for class 24. Taking bits 30−100 into account, the overlap is ultimately increasing to ∼89% and ∼84%. Thus, most significant differences were observed for top-ranked bit positions.

Considering the differences between the GR- and KL divergence-based histograms in Figure 2, we conclude that GR provides a more stable feature selection approach for fingerprint reduction. This is the case because GR yields bit rankings that

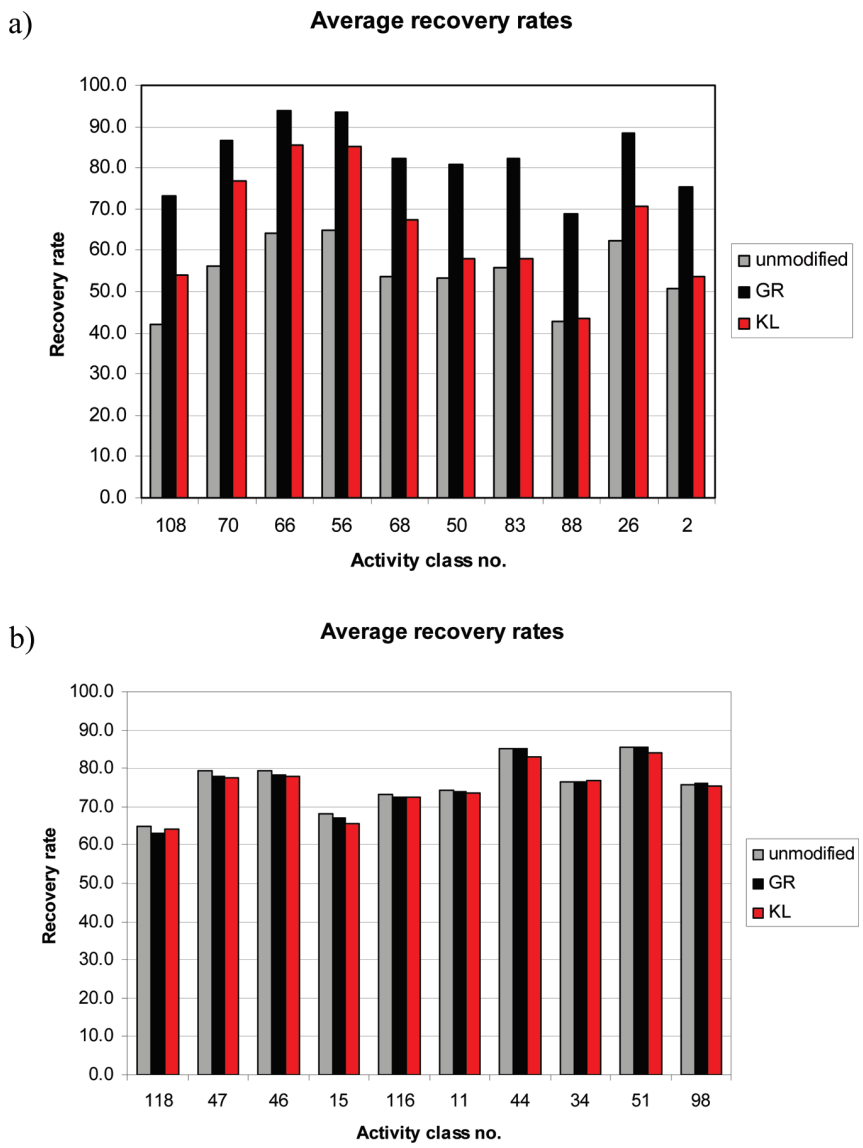


Figure 1. Average recovery rates for ECFP4. Shown are 10 activity classes with the highest (a) and lowest (b) increase in recovery rates as a consequence of GR feature selection. Also shown are the corresponding recovery rates for unmodified fingerprints and KL-based selection.

predominantly reflect probability differences between active and database compounds and are more intuitive than KL divergence rankings. This conclusion is also consistent with the overall better search performance we observed for fingerprints reduced on the basis of GR compared to KL divergence and their smaller size, as discussed above. However, on the basis of KL divergence analysis, reduced fingerprints, in part, with significantly increased search performance compared to the original fingerprints have also been generated in a number of cases,[20,21] thus demonstrating its feature selection potential. In addition, KL divergence analysis is generally less dependent on background database composition than GR because it emphasizes the magnitude of active probabilities, which should also be taken into account in cases where background databases are relatively small.

Compound Recall Characteristics. We next analyzed recovery rates for reduced fingerprints consisting of up to 1000 bit positions. For each activity class, averages were calculated for all 100 reference sets. Representative examples are shown in Figure 3. We generally observed that recovery rates peaked at relatively small feature numbers and then decreased and/or remained constant as feature numbers increased. The recovery rate profiles were overall comparable for both ECFP4 and Molprint2D, although details differed in many cases. For both fingerprints, a notable difference between GR and KL divergence selection was also observed at the level of recall curves. GR-based feature selection often led to a sharper increase in recall performance, resulting in a clear peak followed by a reduction in recovery rates for further increasing feature numbers. Then the recovery rates essentially remained constant. These recall characteristics were frequently also observed for KL divergence selection but were generally less obvious. Figure 3a−h illustrates these effects. The number of features required to reach the top search performance was often found to differ between structurally diverse and homogeneous activity classes. In Figure 3a, ECFP4 search results are shown for activity class 1 (with a compound-to-CSK ratio of 3.36), and in Figure 3c, ECFP4

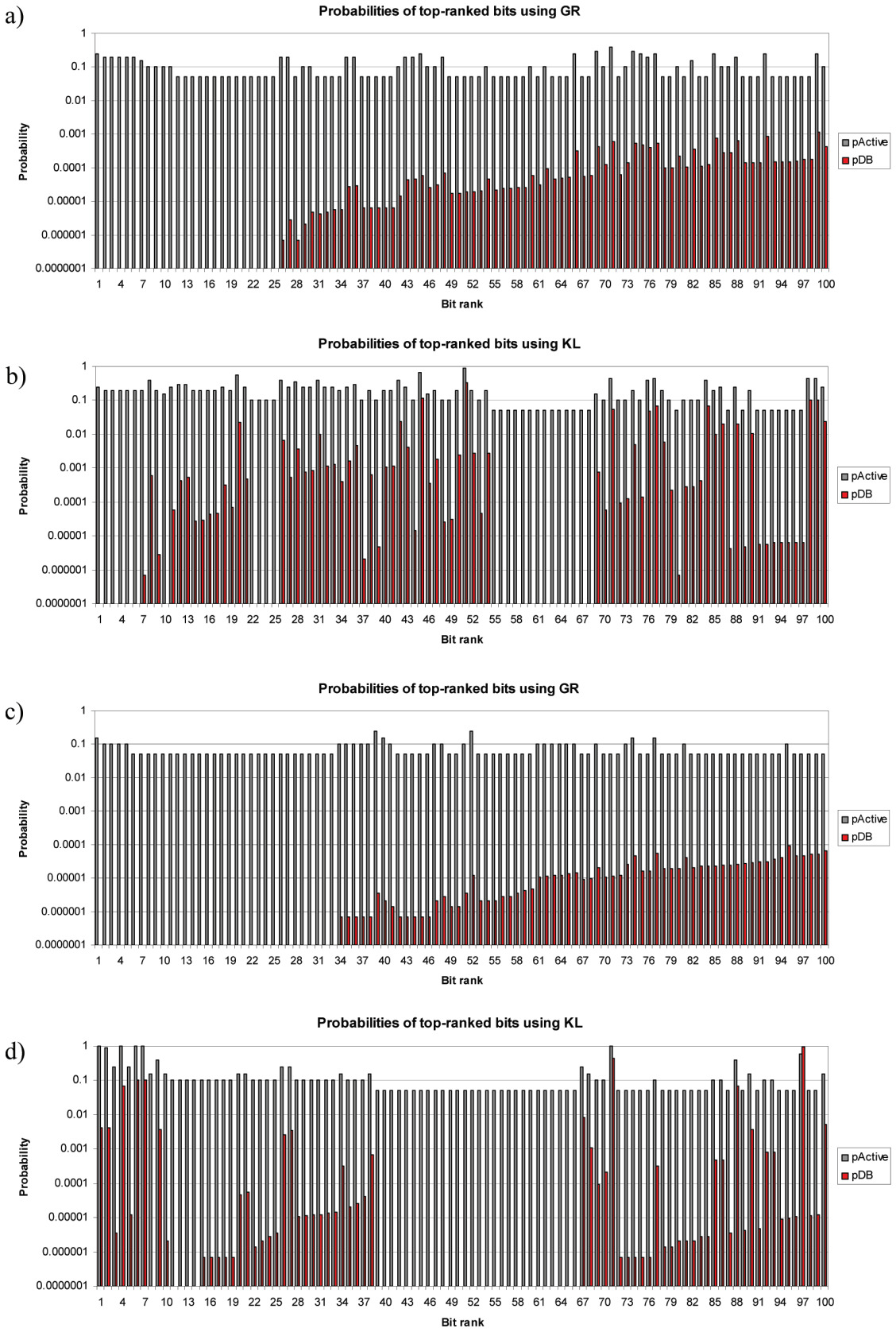


Figure 2. Probabilities of top-ranked bits. For two exemplary activity classes of different structural diversity (no. 24, diverse; no. 22, homogeneous) and an individual reference set, the probabilities of the 100 top-ranked bits to occur in active (pActive) and database compounds (pDB) are reported for different feature selection methods. Probabilities (without *m*-estimate correction) are plotted on a logarithmic scale. (a) 24/GR, (b) 24/KL divergence, (c) 22/GR, (d) 22/KL divergence.

dx.doi.org/10.1021/ci200275m |*J. Chem. Inf. Model.* 2011, 51, 2254–2265

Table 2. Overlap of Bits Between GR and KL Divergence[a]

	overlap	
no. bits	no. 22	no. 24
1	0.0	100.0
2	2.5	90.0
3	6.7	84.0
4	7.5	79.3
5	7.2	75.4
10	20.8	61.4
15	27.7	51.8
20	35.5	49.8
30	53.3	53.3
40	70.9	58.2
50	76.8	60.6
60	80.8	70.4
70	82.7	75.9
80	86.7	78.3
90	88.3	80.5
100	88.8	84.1

[a] For two exemplary activity classes of different structural diversity (24, diverse; 22, homogeneous), the average overlap (in %) between features in reduced fingerprints of different size (no. bits) selected by GR and KL divergence is reported. The average overlap was calculated for 100 independent search trials with different reference sets, and in each case, the 100 top-ranked bit positions were compared.

results are shown for activity class 54 (ratio 2.13). Here GR-based reduced ECFP4 representations consisting of 120 bits for class 1 and 160 bits for class 54 reached a clear performance peak. In Figure 3e and f, corresponding ECFP4 search profiles are shown for activity class 30 (ratio 10.33) and class 75 (ratio 7.17), respectively. These structurally more homogeneous activity classes yielded higher compound recovery rates than the more diverse classes 1 and 54; class 30 required 140 ECFP4 features to reach the top performance and class 75 only 60 features.

Taken together, despite the fingerprint, selection method, and activity class dependent differences we observed, three general conclusions could be drawn for both fingerprints and GR selection from the findings discussed above. First, reduced fingerprints, rather than unmodified versions, generally produced highest recall rates. Second, structurally diverse activity classes often—but not always—required more features to reach top performance than structurally more homogeneous classes. Third, in all instances, the recall curves revealed a nearly linear increase in recovery rates over increasing bit numbers until the top search performance was reached. Depending on the activity classes, the slope of these pseudolinear curve intervals often differed, as illustrated in Figure 3.

Recovery of Activity Class Subsets. In order to further rationalize the observations discussed above, we analyzed the number of active database compounds (correctly identified hits) and other database molecules that were detected by reduced fingerprints of increasing size of up to 100 bits, beginning with the smallest representations containing the most highly ranked features. In Table 3, three representative examples are shown for ECFP4 features and GR selection. In Table 3a, activity class 1 (compound-to-CSK ratio 3.36) contained 554 compounds and yielded a recovery rate of 45.5% with full-length ECFP4 and 69.3% with the best reduced fingerprint. In Table 3b, class 24 (ratio 3.82) consisted of 294 compounds and produced an ECFP4 recovery rate of 79.6% and of 92.8% for the best reduced version. In Table 3c, class 36 (ratio 7.06) contained 374 compounds and yielded a recovery rate of 87.7% for ECFP4 and 93.1% for the best reduced fingerprint. Hence, these activity classes had different composition and displayed partly different recall characteristics. As reported in Table 3a−c, a varying number of the most highly ranked fingerprint features consistently detected a significant number of the active compounds, without selecting any other database molecules. For activity class 1, 24, and 36, the top 40, 20, and 10 bits exclusively recognized 206, 211, and 104 active compounds, respectively. Hence, the exclusive recognition of active compounds by small feature sets significantly contributed to the overall compound recall. As also revealed in Table 3, the inclusion of additional bits resulted in further detection of active compounds accompanied by a steady and, in part, dramatic increase in the number of other database molecules that were detected, corresponding to a substantial loss of the specificity of the search calculations. From these observations, one can infer that the generally observed gain in search performance through fingerprint reduction can largely be attributed to the elimination of fingerprint features that predominantly increase the background noise of the calculations (i.e., preferentially detect other database compounds).

Moreover, Table 3 reveals another trend that is highly relevant for our analysis. The most important fingerprint features were specific for subsets of active compounds. For example, in Table 3a, the first bit detected 45 active compounds, the second 80 other compounds, the third another 38 previously unrecognized actives, and so on. There was a steady increase in active compounds up to a size of 100 bits. Starting with 50 bits, database molecules were beginning to be retrieved. In the presence of 100 bits, the number of detected active compounds approximately doubled compared to 40 bits (when still no database compounds were detected), but 2090 other database molecules were also selected. In Table 3b, bit 1 detected 80 compounds, bit 2 selected 40 more, bit 3 did not add more active compounds, but bit 4 detected 37 additional ones, which then remained essentially unchanged for up to 15 bits, until the inclusion of the next 5 bits led to the detection of another 53 active compounds, still without retrieval of other database molecules. Beginning with 30 bits, other database molecules were detected, and the increments of newly recognized active compound became smaller. Corresponding effects also occurred for the structurally more homogeneous activity class in Table 3c, although there was overall less variation in the number of active compounds detected by increasing numbers of bits, with the exception of notable increases in active compounds within the range of 10−30 bits, where other database compounds were also selected.

Figure 4 illustrates some of the results discussed above. Shown are five structurally diverse hits that were recognized by the top three fingerprint bits according to Table 3a. Each of these compounds contains only one of the atom environments encoded by these three bits, i.e., it responds to only one of the three top-ranked bits. The corresponding substructures are mapped on the hits. The comparison reveals that a substructure match provided by a single fingerprint feature has been sufficient in these instances to facilitate a scaffold hop. With only the first three fingerprint bits, a total of 163 active compounds were retrieved that yielded 35 distinct CSKs.

Taken together, on the basis of GR selection, compound subset detection by bit subsets was consistently observed for

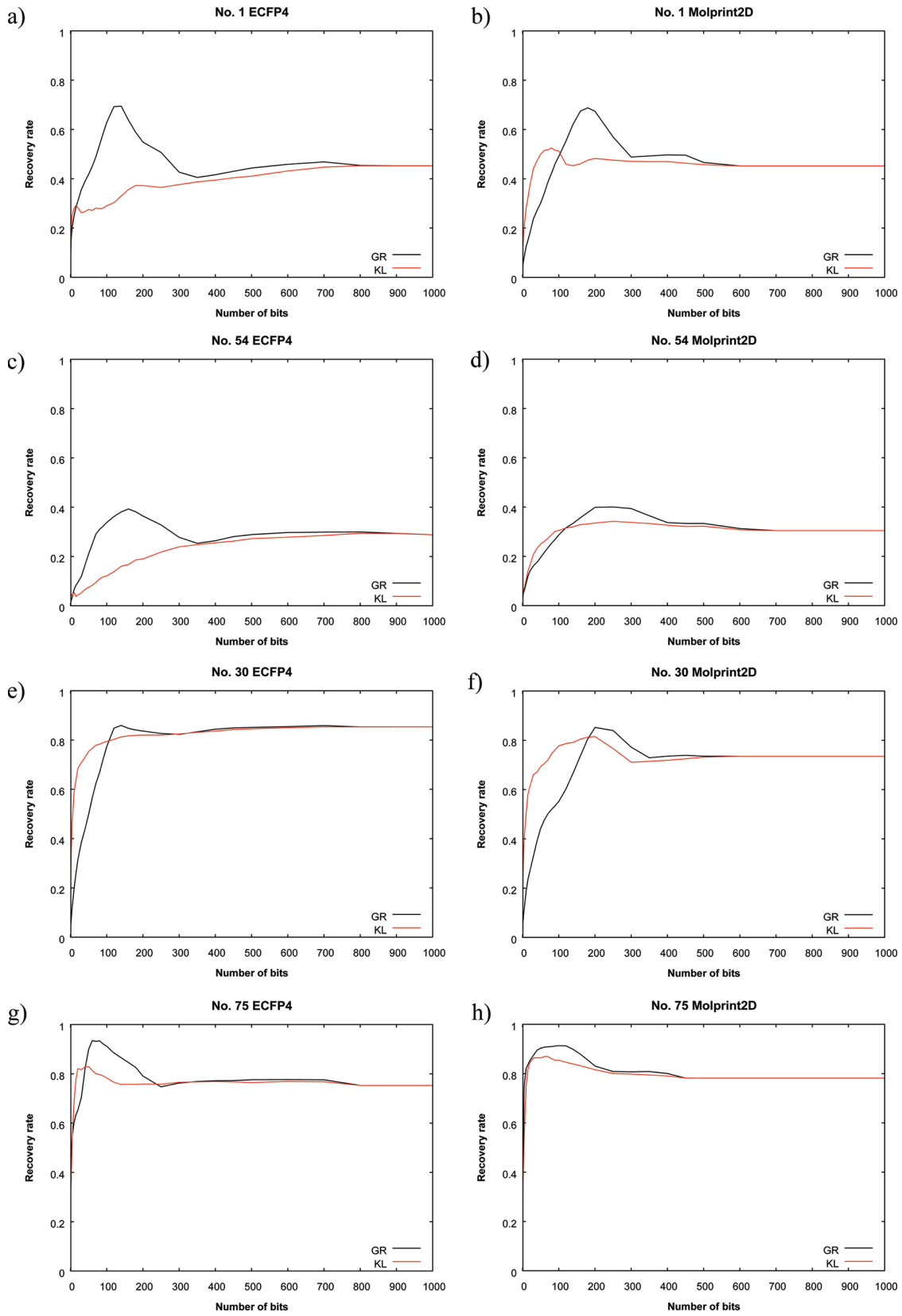


Figure 3. Recovery rates for reduced fingerprints. For four exemplary activity classes of different structural diversity (no. 1 and 54, diverse; no. 30 and 75, homogeneous), average recovery rates are reported for reduced fingerprint representations with increasing number of bits selected by GR or KL divergence. (a) 1/ECFP4, (b) 1/Molprint2D, (c) 54/ECFP4, (d) 54/Molprint2D, (e) 30/ECFP4, (f) 30/Molprint2D, (g) 75/ECFP4, and (h) 75/Molprint2D.

Table 3. Detection of Activity Class Subsets[a]

	GR	
no. bits	no. ADC	no. DC
(a) Activity Class 1		
1	45	0
2	125	0
3	163	0
4	176	0
5	176	0
10	176	0
15	184	0
20	187	0
30	206	0
40	206	0
50	212	9
60	243	38
70	293	132
80	308	379
90	352	860
100	425	2090
(b) Activity Class 24		
1	80	0
2	120	0
3	120	0
4	157	0
5	157	0
10	157	0
15	158	0
20	211	0
30	226	15
40	228	151
50	241	589
60	261	790
70	268	1936
80	271	4557
90	256	4503
100	243	4017
(c) Activity Class 36		
1	86	0
2	87	0
3	87	0
4	90	0
5	90	0
10	104	0
15	110	3
20	158	52
30	319	177
40	319	236
50	319	383
60	319	1648
70	319	4240
80	306	109
90	306	104
100	306	104

[a] For representative activity classes, the numbers of detected active database compounds (no. ADC) and other database compounds (no. DC) retrieved prior to the last recovered active are reported for varying numbers of ECFP4 features (no. bits) selected by GR. In each case, the results are shown for an individual reference set: (a) 1, diverse; (b) 24, diverse; and (c) 36, homogeneous.

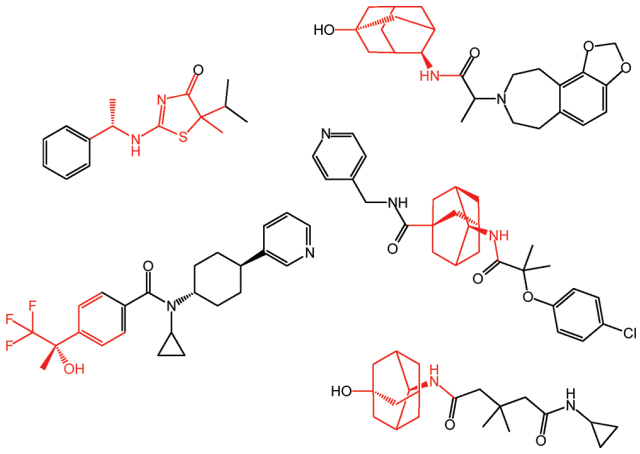


Figure 4. Structurally diverse hits. Shown are five exemplary hits detected with the top three fingerprint bits selected by GR, as reported in Table 3a, that represent scaffold hops. In the fingerprint of each of these compounds, only one of the three bits was set on, and the corresponding structural features are mapped on the hits (red).

reduced ECFP4 and Molprint2D fingerprints. On the basis of KL divergence selection, similar subsets effects were also found, in particular for structurally diverse activity classes but much less so for structurally homogeneous classes. Representative examples for KL divergence selection are provided in Table S3 of the Supporting Information. For structurally homogeneous activity classes, KL divergence selection often yielded top-ranked bits that were not specific for active compounds but also detected other database molecules, different from many structurally diverse classes. By contrast, GR always yielded bit positions that were specific for active compounds.

Scaffold Hopping Potential Revisited. The incremental recognition of different subsets of active compounds by small fingerprint feature sets observed in our analysis, especially for structurally diverse activity classes, provided a rationale for the scaffold hopping potential of the investigated fingerprints. On the basis of our findings, the overall recovery rates of active compounds achieved by these 2D fingerprints largely resulted from the cumulative detection of distinct subsets of active compounds by different fingerprint features. Typically, small numbers of key features specifically selected active compounds over other database molecules. Often, single bit positions were responsible for the detection of relatively large compound subsets. Other less specific bits also retrieved additional subsets of active compounds and also rapidly increased the number of other database molecules. Thus, identifying those bit positions

that were most important for recognizing different activity classes and analyzing their contribution to the recovery of active compounds revealed a plausible mechanism for scaffold hopping using the 2D fingerprints studied here.

CONCLUDING REMARKS

In this study, we have investigated in detail the compound recall characteristics of state-of-the-art 2D fingerprints. Beginning with a large-scale fingerprint search campaign, feature selection methods were applied to systematically reduce original fingerprint and identify the most important fingerprints bits/features for each activity class. Fingerprint reduction generally improved compound recovery rates, consistent with earlier findings. In many instances, small numbers of bits were sufficient to yield the highest search performance. In addition, our results indicated that fingerprint reduction mostly improved compound recall by omitting features that predominantly recognized other database compounds (and thus reduced the specificity of the search calculations). By comparing GR- and KL divergence-based fingerprint feature selection, we identified different characteristics of these approaches, assigning overall higher confidence to GR-based bit rankings. On the basis of feature selection, we observed, in particular, for structurally diverse activity classes, that small numbers of highly ranked fingerprint features (often individual bits) distinguished subsets of active compounds from other database molecules. Additional features also recognized distinct compound subsets but were not specific for active compounds. These cumulative subset contributions to compound recovery rationalized the scaffold hopping potential of 2D atom environment fingerprints and revealed a mechanism for the recognition of structurally diverse hits. It is anticipated that the feature selection approaches introduced herein will be useful for additional mechanistic studies on fingerprints of different design and for further fingerprint engineering applications.

ASSOCIATED CONTENT

Supporting Information. Tables S1, S2, and S3 report the composition of compound activity classes, average recovery rates for Molprint2D, and activity class subset detection on the basis of KL divergence, respectively. This information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

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

ACKNOWLEDGMENT

The authors thank Britta Nisius and Martin Vogt for many helpful discussions.

REFERENCES

(1) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(2) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

(3) *MACCS Structural keys*; Symyx Software: San Ramon, CA.

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

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

(6) Stumpfe, D.; Bajorath, J. Similarity Searching. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1*, 260–282.

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

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

(9) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two- Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010**, *53*, 5707–5715.

(10) Gardiner, E. J.; Holliday, J. D.; O'Dowd, C.; Willett, P. Effectiveness of 2D Fingerprints for Scaffold Hopping. *Future Med. Chem.* **2011**, *3*, 405–414.

(11) Stumpfe, D.; Bill, A.; Novak, N.; Loch, G.; Blockus, H.; Geppert, H.; Becker, T.; Hoch, M.; Schmitz, A.; Kolanus, W.; Famulok, M.; Bajorath, J. Targeting Multi-Functional Proteins by Virtual Screening: Structurally Diverse Cytohesin Inhibitors with Differentiated Biological Functions. *ACS Chem. Biol.* **2010**, *5*, 839–849.

(12) Stumpfe, D.; Bajorath, J. Applied virtual screening: strategies, recommendations, and caveats. In *Methods and Principles in Medicinal Chemistry. Virtual Screening. Principles, Challenges, and Practical Guidelines*; Sotriffer, C., Ed.; Wiley-VCH: Weinheim, Germany, 2011; pp 73−103.

(13) Hert, J.; Willet, P.; Wilton, D. J. Comparison of Fingerprint-based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. J. *Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.

(14) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information. *J. Med. Chem.* **2005**, *48*, 7049–7054.

(15) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.

(16) Geppert, H.; Vogt, M.; Bajorath, J. Current Trends in Ligand-Based Virtual Screening: Molecular Representations, Data Mining Methods, New Application Areas, and Performance Evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205–216.

(17) Bender, A.; Glen, R. C. *MOLPRINT 2D*; Center for Molecular Science informatics, University of Cambridge: Cambridge, U.K.; http://www.molprint.com/. Accessed October 1, 2009.

(18) Bender, 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.

(19) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(20) 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.

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

(22) Kullback, S. *Information Theory and Statistics*; Dover Publications: Mineola, MN, 1997, pp 1−11.

(23) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley: New York, 1991.

(24) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann: San Mateo, CA, 1993.

(25) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198–D201.

(26) *Scitegic Pipeline Pilot*; Accelrys, Inc.: San Diego, CA, 2010.

(27) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(28) Xu, Y.-J.; Johnson, M. Using Molecular Equivalence Numbers to Visually Explore Structural Features that Distinguish Chemical Libraries. *J. Med. Chem.* **2002**, *42*, 912–926.

(29) 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.

Summary

Two feature selection methods from information theory have been used to systematically reduce atom environment fingerprints. The search calculations revealed an improved search performance for reduced fingerprints over unmodified versions due to the elimination of fingerprint features that mainly detected database compounds. Those atom environments that were most relevant for the recovery of active compounds were specific for distinct subsets of the compound activity classes. Individual features were responsible for the detection of different active scaffolds and combinations of these features resulted in a cumulative recall of structurally diverse active compounds.
The supporting information of this publication can be obtained via the following URL: http://dx.doi.org/10.1021/ci200275m.

This and the previous study have shown that 2D fingerprints are well suited for VS tasks despite their relative simplicity. In the following, the fingerprint representation is used as a descriptor for another similarity-based search method in order to address a complex multi-class prediction task involving compounds with different activity profiles. The SVM methodology is applied to recover compounds with specific activities against combinations of targets.

Chapter 3

Prediction of compounds with closely related activity profiles using weighted support vector machine linear combinations

Introduction

The general aim of VS is the recovery of compounds having a specific bioactivity. However, many compounds are active against multiple targets in addition to the one of interest. Standard SVM-based ranking does not consider these multiple activities. Therefore, we apply a variant of the SVM methodology in order to treat compounds having activities against different combinations of targets. Weighted SVM linear combination is used to recover compounds with specific single- or dual-target activities. The use of positive and negative linear weighting factors results in the prioritization and deprioritization of compounds with desired and undesired activity profiles, respectively.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

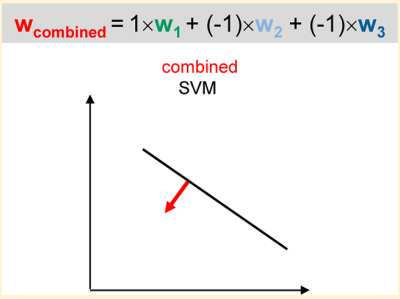
pubs.acs.org/jcim

Prediction of Compounds with Closely Related Activity Profiles Using Weighted Support Vector Machine Linear Combinations

Kathrin Heikamp 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

ABSTRACT: Using support vector machine (SVM) ranking, a complex multi-class prediction task has been investigated involving sets of compounds that were active against related targets and represented all possible combinations of single-, dual-, and triple-target activities. Standard SVM models were not capable of differentiating compounds with overlapping yet distinct activity profiles. To address this problem, we designed differentially weighted SVM linear combinations that were found to preferentially detect compounds with desired activity profiles and deprioritize others. Hence, combining independently derived SVM models using negative and positive linear weighting factors balanced relative contributions from individual reference sets and successfully distinguished between compounds with overlapping activity profiles.



INTRODUCTION

Predicting biological activities of small molecules from chemical structure is one of the major focal points of the chemoinformatics field. For this purpose, a plethora of computational methodologies have been introduced. In recent years, machine learning approaches have become increasingly popular for activity predictions, especially Bayesian classifiers[1−4] and support vector machines (SVMs).[5−9] These supervised learning methods have typically been used for binary classification and prediction of class labels of compounds (e.g., active vs inactive) focusing on different compound activity classes.[2,4,8,9]

Given the increasing interest in chemogenomics,[10,11] these machine learning approaches have also been considered for complex activity predictions. In chemogenomics, the systematic analysis of compound-target annotations and the study of multi-target activities of small molecules take center stage. Accordingly, machine learning approaches have been used for applications such as computational profiling of compounds against arrays of classifiers for individual targets,[12] prediction of ligand−receptor pairings,[13,14] or searching for target-selective compounds.[15] For selectivity predictions, SVM modeling was carried out to distinguish target-selective compounds from others that were active against multiple members of a given target family and also from inactive compounds.

In machine learning terms, the chemogenomics-oriented applications outlined above translate into multi-class prediction problems. For multi-class modeling, SVM-based compound ranking schemes,[15,16] rather than pairwise binary classifications, are particularly suitable. In general, multi-class predictions,[15,17] require the combination or sequential consideration of different SVM models. As an approach for model combination, SVM linear combination has previously been introduced,[18] which was originally applied to the prediction of ligands for orphan targets,[18] another chemogenomics application.

In our current study, we have investigated SVM modeling for another multi-class prediction problem involving compounds with overlapping activities against related targets. This application was found to be challenging for SVM ranking. Furthermore, a standard (unweighted) linear combination was not applicable in this case. Given the difficulties observed in distinguishing between compounds with in part overlapping activities using individual SVM models, we have combined independently derived models in differentially weighted SVM linear combination using positive and negative factors. The application of this strategy led to the preferential detection of compounds with desired activity profiles.

MATERIALS AND METHODS

Basic SVM Theory. As a supervised machine learning technique, SVMs[5] are primarily used for binary object classification and ranking. For learning, "positive" and "negative" training data (e.g., active and inactive compounds) are projected into a feature (descriptor) space χ. By solving a convex quadratic optimization problem, a hyperplane H is derived that best separates objects with different class labels. During the optimization, the trade-off parameter C is adjusted to balance errors due to misclassification of training data and the generalization of the classification. The separating hyperplane H is defined by the normal weight vector $\mathbf{w}$ and a bias b

$$H = \{\mathbf{x} | \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\}, \text{ with } \langle \cdot, \cdot \rangle \text{ being a scalar product}$$

Test data are mapped into the same feature space χ and classified depending on which side of the generated hyperplane they fall. For SVM-based ranking compounds are sorted on the basis of the signed distance from the hyperplane (from the positive to the negative half-space):[16] $g(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$.

For training data that are nonlinearly separable in the feature space χ, which is usually the case, the so-called Kernel trick[19,20] is applied. For this purpose, kernel functions are utilized that replace an

Received: February 5, 2013
Published: March 21, 2013

explicit projection of the data into a high-dimensional feature space H where linear separation might be possible. Therefore, kernel functions $K(\cdot,\cdot)$ replace the standard scalar product.

Weighted SVM Linear Combination. The SVM linear combination (LC) was introduced as an extension of standard SVM classification to enable the prediction of ligands that are active against different targets.[18] In SVM LC, a hyperplane is generated for each individual target t_i under consideration using known active compounds of t_i as positive and inactive compounds as negative training data. The normal vectors of all hyperplanes are then linearly combined into a single vector

$$\mathbf{w}_{\text{combined}} = \sum_{i=1}^{n} f_i \mathbf{w}_i$$

where $\mathbf{w}_{\text{combined}}$ is the single combined normal vector, n the number of original hyperplanes, and f_i and $\mathbf{w}_i$ are the individual linear factors and normal vectors of each hyperplane, respectively. Herein, the LC approach is extended through the use of positive and negative linear factors. Test compounds are then ranked using a global ranking function

$$g(\mathbf{x}) = K(\mathbf{w}_{\text{combined}}, \mathbf{x})$$

Factors can be applied to adjust the relative contributions of individual weight vectors to the linear combination. Such weighting factors were previously applied to an SVM linear combination in similarity search calculations taking potency information of reference compounds into account.[21] In this case, SVM models derived for reference compounds at different potency levels were linearly combined, and their potency values were used as weighting factors.[21]

For the multi-class prediction task addressed in our current analysis, we introduce a modification of the weighted SVM LC approach that uses negative linear factors, as illustrated in Figure 1. The use of

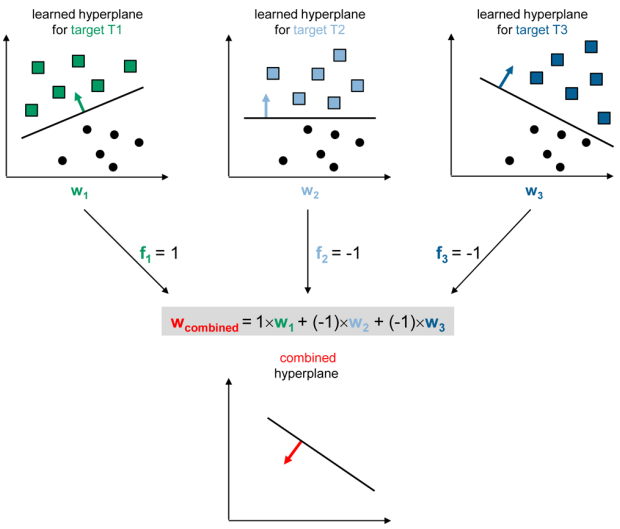


Figure 1. Weighted SVM linear combination. Normal weight vectors are derived from SVM calculations for sets of compounds active against three different targets. Weight vectors are linearly combined to yield a single vector used for global ranking. Positive and negative factors are applied to individual weight vectors to adjust their relative contributions to class label prediction or ranking.

negative factors effectively inverts the contributions of positive and negative training data for individual models during linear combination, a strategy that has not yet been considered in SVM modeling. This modification deprioritizes compounds with undesired activity relative to confirmed inactive compounds and compounds with desired activity. Balancing SVM LC through the application of positive as well as negative weighting factors is shown herein to effectively distinguish between compounds with different activity profiles.

In order to consistently prioritize or deprioritize compounds with given activity profiles, weighting factors of 1, 2, −1, and −2 were systematically varied. Single-target SVM classifiers were built as a control for each individual activity.

Compound Data Sets. Compound data sets with single-, dual-, and triple-target activities have been assembled from PubChem[22] confirmatory bioassays for three cytochrome P450 isoforms and three different dehydrogenases. The first set of three assays identified compounds active against cytochrome P450 2C19 (CYP2C19; assay id (AID) 899), cytochrome P450 2D6 (CYP2D6; AID 891), and cytochrome P450 3A4 (CYP3A4; AID 884). The second set of three assays contained inhibitors of aldehyde dehydrogenase 1 (ALDH1A1; AID 1030), hydroxyacyl-coenzyme A dehydrogenase type II (HADH2; AID 886), and 15-hydroxy-prostaglandin dehydrogenase (HPGD; AID 894). From all assays confirmed active and inactive compounds were extracted and compared. Compounds with confirmed single-, dual-, and triple-target annotations were identified as well as compounds inactive against all targets. For cytochrome P450s, a total of 1906 active compounds with different target profiles were obtained and 2901 inactive compounds. For the dehydrogenases, 5085 active and 39,355 inactive compounds were obtained. The composition of the two data sets is reported in Figure 2.

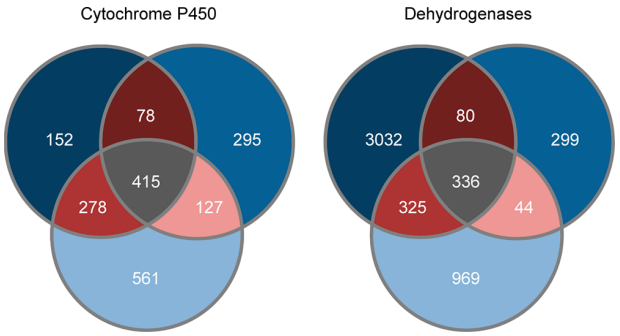


Figure 2. Active compounds. For inhibitors of three cytochrome P450 isoforms (data set 1) and three different dehydrogenases (data set 2), the number of compounds with single and multiple target annotations is reported in a Venn diagram.

Calculations. Compounds were represented using the extended-connectivity fingerprint[23] with bond diameter 4 (ECFP4) calculated using the Molecular Operating Environment.[24] For compound comparison during SVM calculations, the Tanimoto kernel[20] was applied.

In calculations searching for compounds with single-target annotations, compounds with the desired activity were used as positive training data and confirmed inactive compounds as negative data. In calculations searching for compounds with dual-target annotations, compounds with the desired dual-target activity were used as positive training data and confirmed inactive compounds as negative data. To assess search performance, SVM ranking was carried out using target-specific models and their weighted linear combination.

For training, 500 inactive compounds were used in each case as inactive training data. In order to use size-balanced compound sets for the generation of hyperplanes for linear combination, the number of

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

positive training compounds was set to half of the smallest available single- or dual-target compound subset for each series of search calculations.

For each search calculation, 100 different trials with randomly assembled positive and negative training data and test sets were carried out. In each case, test data consisted of all active compounds with different activity profiles plus all confirmed inactive compounds not used for training. Active and inactive compounds used for training were never included in test sets.

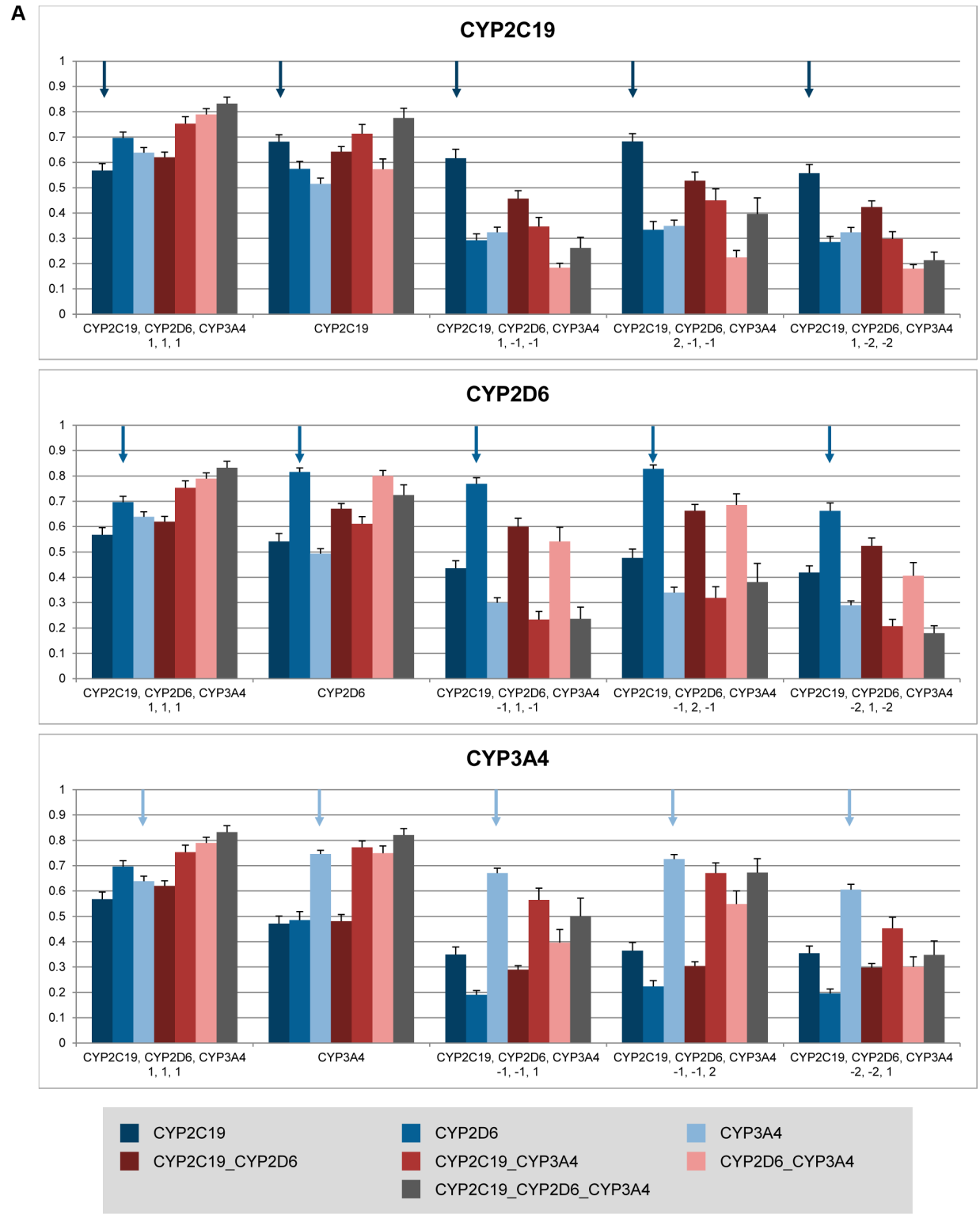


Figure 3. continued

dx.doi.org/10.1021/ci400090t | *J. Chem. Inf. Model.* 2013, 53, 791−801

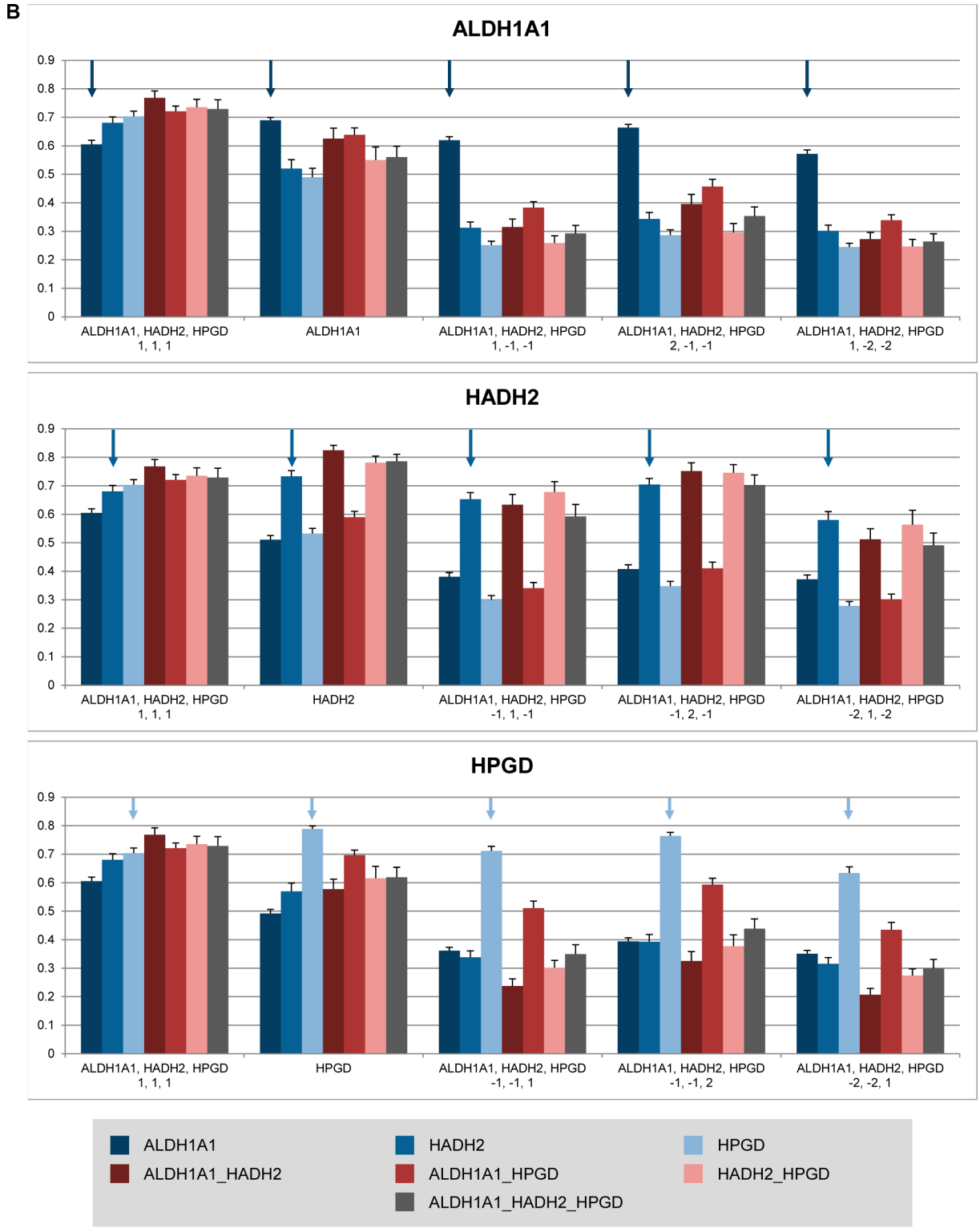


Figure 3. Searching compounds with single-target activities. ROC_AUC values are reported for compounds active against all possible target combinations in search calculations for compounds with single-target activities. (A) cytochrome P450s, (B) dehydrogenases. Dual- and triple-target combinations are indicated by underscores (e.g., CYP2C19_CYP2D6). Individual targets and target combinations are color-coded as in Figure 2. Search results are reported for standard SVM LC, SVM training using compounds active against the designated target, and differently weighted SVM LCs. In each case, color-coded arrows indicate the desired search result. In addition, standard deviations over 100 independent search trials are reported above each bar.

Search performance was measured using the area under the receiver operating characteristic curve (ROC_AUC)[25] averaged over all 100 trials.

SVM calculations were performed using SVM^{light},[26] a freely available SVM implementation. SVM parameters were suggested SVM^{light} default settings. SVM LC calculations were carried out using in-house generated Perl scripts.

RESULTS AND DISCUSSION

Compounds with Overlapping Activity Profiles. The compound data sets analyzed herein were assembled to investigate a multi-class prediction problem that we considered rather challenging: differentiating between compounds with related and overlapping activity profiles. The cytochrome P450 (CYP) and dehydrogenase data sets contained compounds with all possible combinations of single-, dual-, and triple-target activities, as illustrated in Figure 2. Because many compounds shared activities against individual targets in different combinations, this classification problem was anticipated to be difficult to solve using conventional SVM strategies.

Confirmed Inactive Compounds. With these data sets, we also addressed a potential caveat for machine learning in chemoinformatics that is often pointed out: usually randomly chosen sets of database compounds assumed to be inactive are used as negative training examples. By contrast, by assembling our data sets from PubChem confirmatory bioassays, we were able to obtain sets of confirmed inactive compounds for training against all possible activity combinations, hence providing a sound basis for model building. Our data sets containing all active and inactive compounds are made freely available via http://www.lifescienceinformatics.uni-bonn.de/downloads.

Single-Target Classifiers vs SVM Linear Combination. It should be stressed that SVM models built for individual activities, i.e., single-target classifiers, are conceptually distinct from SVM LC models. An SVM LC represents a model for multi-class predictions that integrates individual classifiers (and is as such distinct from them) and weights their relative contributions prior to predicting test data (but not posthoc). Hence, there is no retrospective fitting of individual classifiers comprising an SVM LC. Weighting factors and their combinations must be systematically explored, as discussed in the following.

Searching for Compounds with Single-Target Activity. We first generated SVM models on the basis of reference compounds that were active against individual targets to search for compounds with single-target activity. Using these models, compounds with all activity combinations were ranked and ROC_AUC values calculated for all compound categories. The results for compounds active against CYP isoforms and dehydrogenases are reported in Figure 3A andB (second bar charts from the left), respectively. For all three CYP targets, the recall of compounds with desired single-target activity was met or exceeded by compounds with activity against other targets or target combinations. Equivalent observations were made for dehydrogenase target HADH2, while for targets ALDH1A1 and HPGD at least slightly higher recall of specifically active compounds was observed. Overall, however, single-target SVM models failed to yield a clear separation in recall between compounds with the desired single-target activity and other activity profiles. Figure 3 also shows that single-target models produced comparably high recall of active compounds in most cases, with ROC_AUC values of ~0.7 to ~0.8, but failed to discriminate between compounds with different activity profiles. In addition, the standard SVM LC of single-target models (with factor setting "1,1,1" in Figure 3) preferentially detected compounds with triple- or dual-target activity. Differences in recall performance were small in all cases.

On the basis of these findings, we then investigated combinations of negative and positive factors for SVM linear combination, as rationalized in the Methods section. We first used a factor setting of "1,−1,−1" to deprioritize compounds with undesired single-target activities compared to the desired activity. Because undesired single-target activities were a part of all activity combinations, this factor setting was also anticipated to deprioritize compounds with dual- and triple-target activity. The results of weighted SVM LC calculations using positive and negative factors are reported in Figure 3 and confirmed the utility of this strategy. For all CYP targets, the recall of compounds with correct target activity was retained or only slightly reduced, whereas the recall of all other categories of compounds was significantly reduced, frequently to ROC_AUC values close to or below 0.5 (corresponding to random selection), as shown in Figure 3A. For compounds active against two dehydrogenases (ALDH1A1 and HPGD), equivalent observations were made. By contrast, in the case of HADH2, the recall of two dual- and the triple-target combinations remained comparable to compounds with the desired single-target activity, and no separation was observed (Figure 3B).

We then tested additional weighting factor combinations. For factor setting "2,−1,−1", which put additional weight on positive training examples with desired activity, a further increase in recall was generally observed for the desired compounds as expected, while the recall rates for other compound categories essentially remained constants for two of three CYP and two of three dehydrogenase targets. In the remaining two cases (CYP3A4 and HADH2), the recall for compounds with dual- and triple-target activities increased relative to the recall of desired compounds, which notably reduced the separation.

In addition, for factor setting "1,−2,−2", which more strongly deprioritized compounds with undesired activity, a significant reduction of recall of compounds with desired single-target activity (and in part other activity combinations) was observed, very likely because too much weight was put on negative training examples, hence rendering these calculations overall less sensitive to active compounds.

Among the differently weighted SVM LCs including negative factors, the factor settings "1,−1,−1" and "2,−1,−1" yielded a clear separation in recall between compounds with desired single-target activity and compounds with other activity profiles for four of six targets, whereas standard SVM calculations consistently failed to do so, despite reaching overall high recall performance on active compounds.

For calculations using SVM models trained on compounds with single-target activity, the recall of compounds with desired single-target activity varied between 0.68 for CYP2C19 and 0.82 for target CYP2D6. For standard SVM LC, the recall of compounds with desired single-target activity varied between 0.57 for CYP2C19 and 0.7 for HPGD. In differentially weighted SVM LCs, recall of the single-target activities ranged from 0.62 for CYP2C19 to 0.77 for CYP2D6.

As a consequence of SVM LC weighting, recall of undesired targets and target combination was reduced to values between 0.18 (CYP2D6_CYP3A4 compounds in single-target CYP2C19 calculation using factor "1,−1,−1") and 0.69 (CYP2D6_CYP3A4 compounds in single-target CYP2D6 calculation with factor "2,−1,−1"), with the majority of compound categories falling to ROC_AUC value below 0.5 (random selection).

Searching for Compounds with Dual-Target Activity. We next searched for compounds with dual-target activity. In these calculations, compounds with desired dual-activity were used as reference compounds for the generation of individual SVM models or SVM LCs. The linear combination strategy was adjusted accordingly. In this case, only two models were combined including the SVM model trained on the basis of the desired dual-target activity and the model derived for compounds with single-target activity against the third (undesired) target. This linear combination scheme enabled a meaningful application of

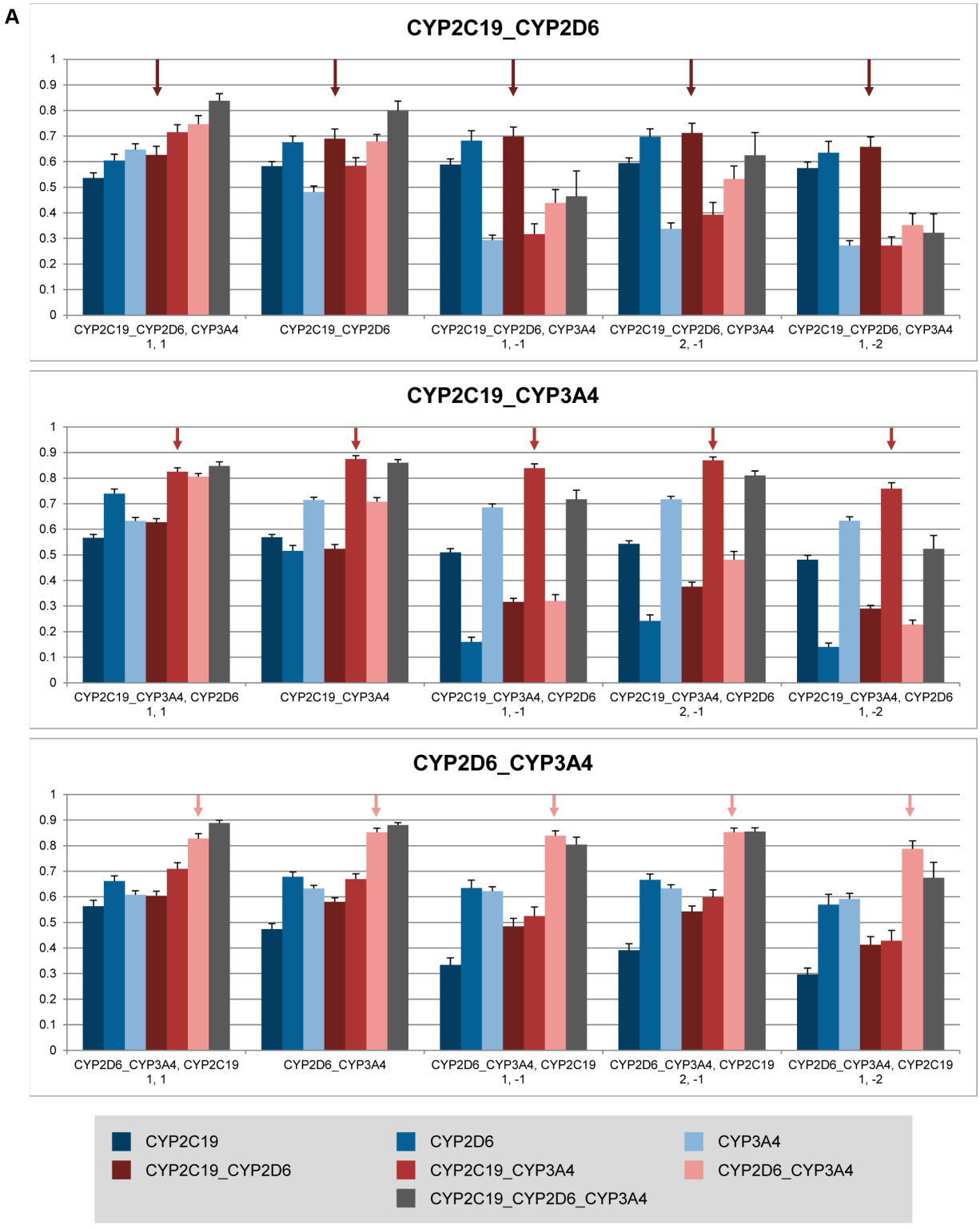


Figure 4. continued

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

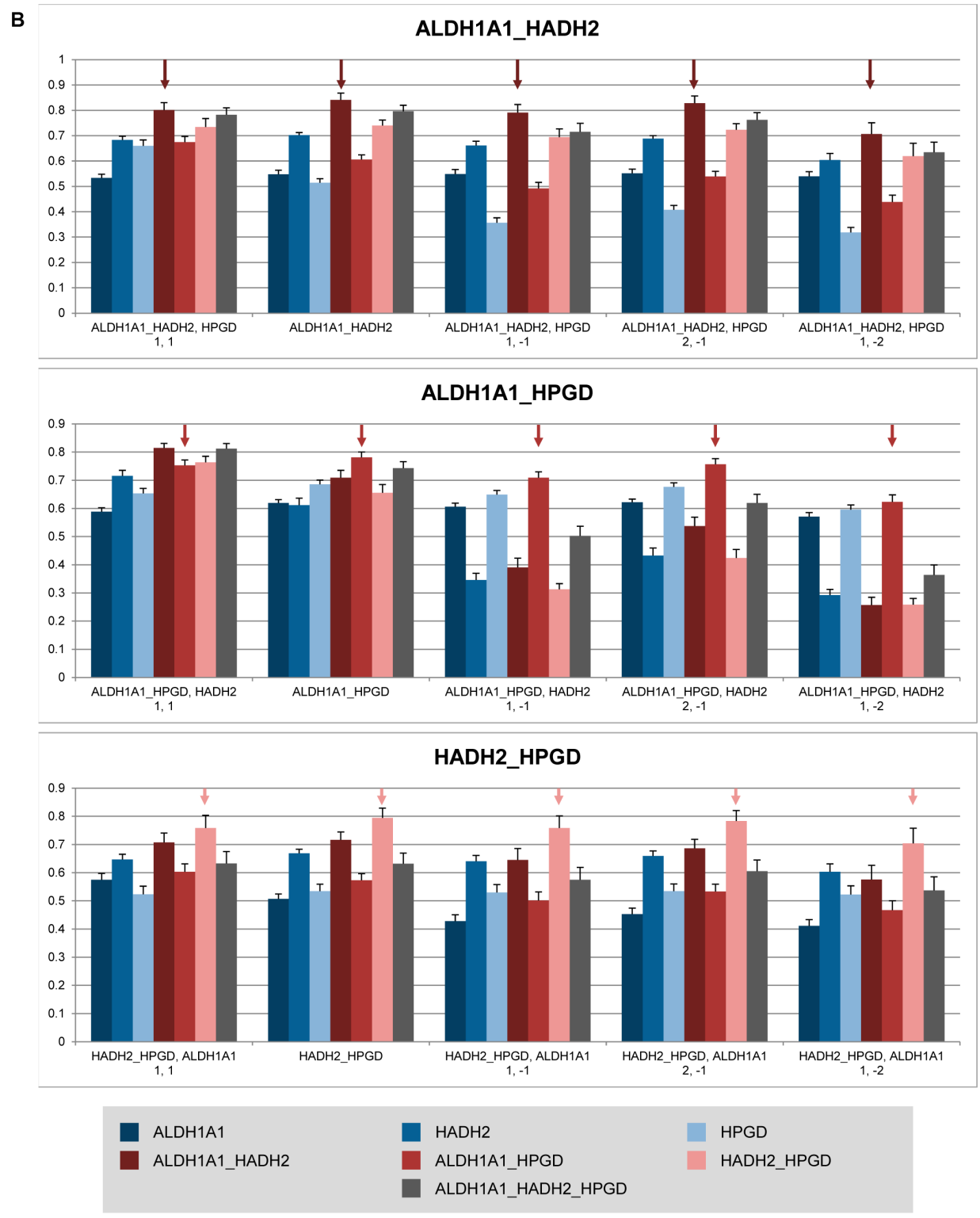


Figure 4. Searching compounds with dual-target activities. ROC_AUC values are reported for compounds active against all possible target combinations in search calculations for compounds with dual-target activities. (A) cytochrome P450s, (B) dehydrogenases. The presentation is according to Figure 3. Search results are reported for standard SVM LC, SVM training using compounds active against the designated target combination, and differently weighted SVM LCs.

negative and positive weighting factors. The search results for CYP isoforms and dehydrogenases are reported in Figure 4A and B, respectively.

Results obtained for SVM models generated for compounds with the desired dual-target activity and the standard SVM LC essentially paralleled the observations discussed above.

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

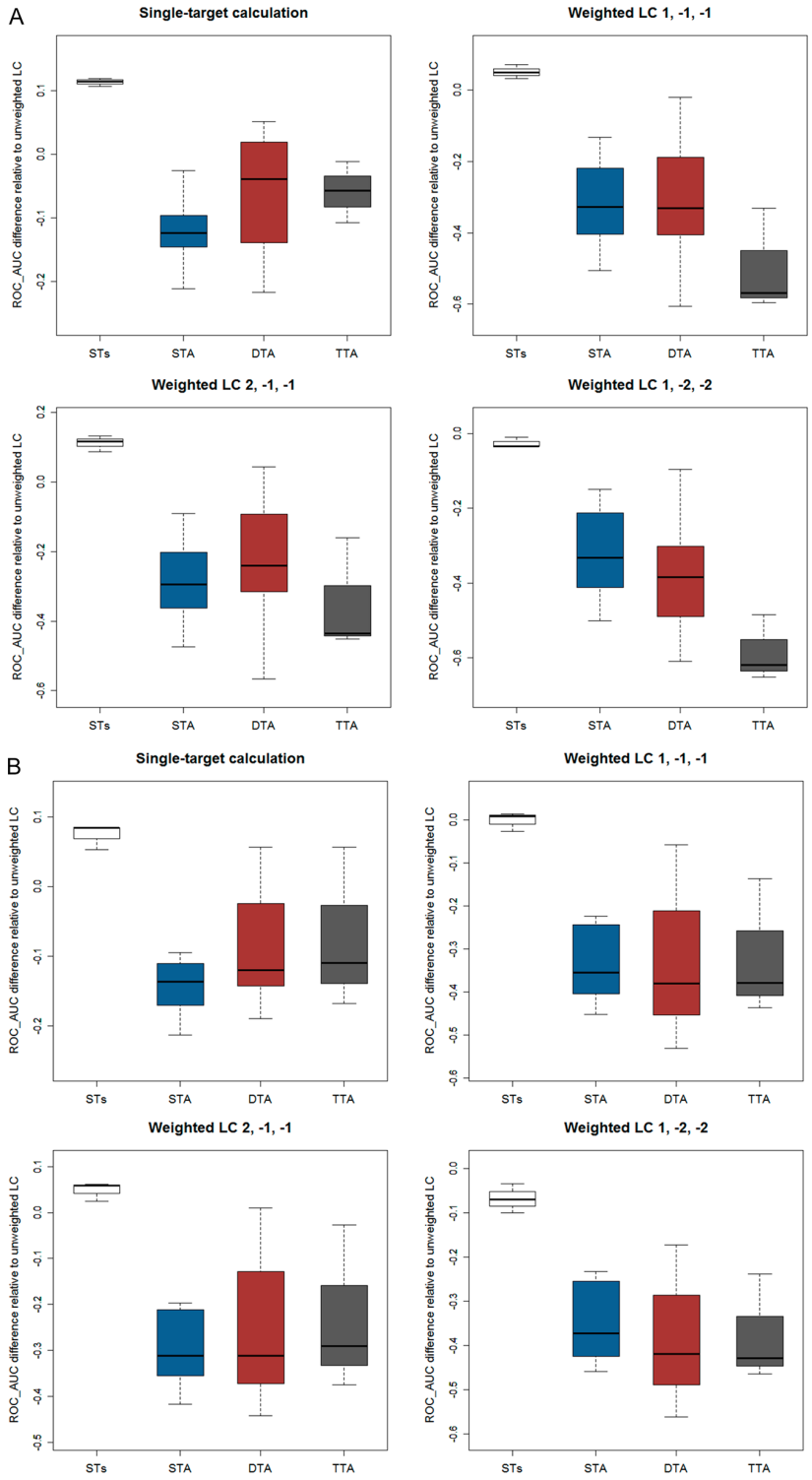


Figure 5. Recall differences for compounds with single-target activities. Differences between ROC_AUC values are reported for single-target calculations (Figure 3) and differently weighted SVM LCs relative to the standard (unweighted) LC calculation. (A) cytochrome P450s, (B) dehydrogenases. Boxplots are shown for four different types of targets or target combinations: desired single targets (STs, i.e., monitoring calculations for all three individual targets), other single-target activities (STA; blue), dual-target activities (DTA; red), and triple-target activity (TTA; gray).

The recall was high for many compound categories and no notable separation was observed. In addition, in this case, recall was high for compounds with triple-target activity in most calculations, as one would expect. Thus, standard SVM calculations also failed here to distinguish between compounds with different activity profiles.

When SVM LCs were tested with factor settings of "1,−1", which deprioritized compounds with activity against the undesired target,

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

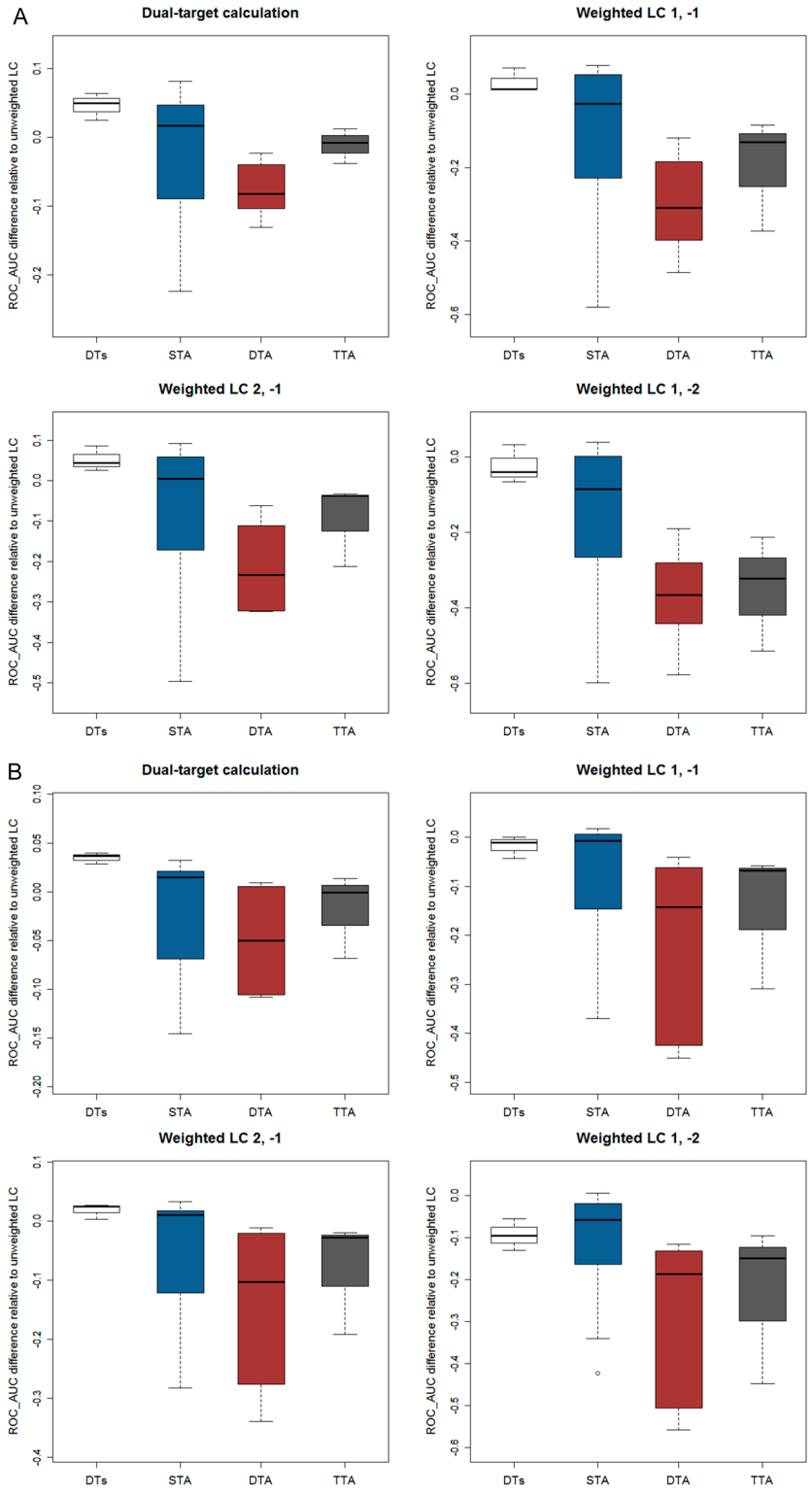


Figure 6. Recall differences for compounds with dual-target activities. Differences between ROC_AUC values are reported for dual-target calculations (Figure 4) and differently weighted SVM LCs relative to the standard LC calculation. (A) cytochrome P450s, (B) dehydrogenases. Boxplots are shown for four different types of targets or target combinations: desired dual-target combination (DTs, i.e., all calculations for the desired dual target combinations), STA, DTA, and TTA. Abbreviations and colors are used according to Figure 5.

and with setting "2,−1", which prioritized compounds with desired dual-target activity and deprioritized compounds with activity against the undesired target, a comparable improvement in recall separation was observed for four of six target combinations, except CYP2D6_CYP3A4 (Figure 4A) and HADH2_HPGD (Figure 4B). However, recall of compounds with activity against individual targets of the

dx.doi.org/10.1021/ci400090t | J. Chem. Inf. Model. 2013, 53, 791−801

desired combination and/or the triple-target combination remained high in most instances. Hence, a partial separation was observed in these cases, different from the calculations focusing on single-target activities discussed above. This was not unexpected given the dual-target nature of the desired activity profiles.

Furthermore, the application of factor setting "1,−2", which strongly deprioritized compounds active against the undesired target, led to a general reduction in recall, similar to the findings discussed for single-target activities under equivalent SVM LC weighting conditions because the influence of negative training examples was further emphasized in these cases.

Taken together, the results of search calculations obtained for compounds with dual-target activities indicated that deprioritization of the undesired target using weighting factor "−1" led to a preferred recall separation.

For standard SVM LC, the recall of compounds with desired dual-target activity varied between 0.63 for the CYP2C19_CYP2D6 and 0.83 for the CYP2C19_CYP3A4 combination. Because of SVM LC weighting, recall of undesired targets or target combinations including the undesired target was reduced to values between 0.16 (CYP2D6 compounds in calculations for CYP2C19_CYP3A4 using factor "1,−1") and 0.86 (for the triple-target set of all CYP isoforms in calculations for CYP2D6_CYP3A4 with factor "2,−1").

Recall Differences. The recall trends and separation effects discussed above are further quantified for single-target and dual-target calculations in Figures 5 and 6, respectively. As a reference point for all comparisons, standard SVM LC recall was used. In search calculations for compounds with single-target activities, median ROC_AUC value differences in recall between compounds with desired activity and other activity profiles were within 0.2 for SVM models trained on compounds with single target activity for both CYP (Figure 5A) and dehydrogenase targets (Figure 5B). In both cases, SVM LCs with negative weighting factors increased the recall separation to median ROC_AUC value differences of ∼0.3 to ∼0.6, depending on the model and compound category. In search calculations for compounds with dual-target activities, maximal median ROC_AUC value separations of ∼0.15 and ∼0.1 were observed for CYP (Figure 6A) and dehydrogenase targets (Figure 6B), respectively, when SVM models trained on compounds with dual-target activities were utilized. For weighted SVM LCs, the median recall for compounds with single-target activity was very similar to recall for compounds with activity against the desired target combination (due to the influence of shared targets), but the separation relative to compounds with other dual-target or triple-target activity was increased to median values of ∼0.2 to ∼0.4 for CYP and ∼0.05 to ∼0.2 for dehydrogenase targets. For individual compound sets, much larger recall separations were also observed in the latter case, as shown in Figure 6B.

■ CONCLUSIONS

In this study, we have investigated a multi-class prediction task involving compounds with activity against different combinations of targets. Given the overlap in activity profiles between these compounds, we anticipated that it might be difficult to address this task. Initially, individual SVM models were trained for all compound categories. Searching for compounds with desired single- or dual-target activity in the presence of confirmed inactive compounds using standard SVM calculations confirmed our expectations. SVM-based compound ranking was found to produce reasonable to high compound recall for different compound categories but essentially failed to distinguish compounds with desired activity from compounds with other activity profiles. Therefore, we designed an SVM linear combination strategy that involved weighting of different models using positive and negative factors. The combination of models and use of positive and negative weighting factors made it possible to prioritize and deprioritize compounds with desired and undesired activity profiles, respectively. Differentially weighted SVM LC calculations yielded in part significant recall separation effects. Especially for compounds with desired single-target activity, the weighted SVM LC approach consistently reduced the recall of compounds with different activity profiles while essentially retaining the recall of compounds with desired target activity. Hence, the SVM LC weighting strategy introduced herein to investigate a complex activity prediction task should also be of interest for other multi-class SVM applications.

■ AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*, 2nd ed.; Wiley-Interscience: New York, 2000, pp 20−83.

(2) Bender, 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. Comput. Sci.* **2004**, *44*, 170−178.

(3) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124−1133.

(4) Watson, P. Naive Bayes classification using 2D pharmacophore feature triplet vectors. *J. Chem. Inf. Model.* **2008**, *48*, 166−178.

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

(6) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discovery* **1998**, *2*, 121−167.

(7) Burbidge, R.; Trotter, M.; Holden, S.; Buxton, B. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.

(8) 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.

(9) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−561.

(10) Bredel, M.; Jacoby, E. Chemogenomics: An emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, *5*, 262−275.

(11) Stockwell, B. R. Exploring biology with small organic molecules. *Nature* **2004**, *432*, 846−854.

(12) Paolini, G. V.; Shapland, R. H. B.; van Hoorn, W. P.; Mason, J. S.; Hopkins, A. L. Global mapping of pharmacological space. *Nat. Biotechnol.* **2006**, *24*, 805−815.

(13) Bock, J. R.; Gough, D. A. Virtual screens for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1402−1414.

(14) Jacob, L.; Vert, J.-P. Protein−ligand interaction prediction: An improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149−2156.

(15) Wassermann, A. M.; Geppert, H.; Bajorath, J. Searching for target-selective compounds using different combinations of multi-class support vector machine ranking methods, kernel functions, and fingerprint descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 582−592.

(16) 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.
(17) 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.
(18) 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.
(19) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*; Pittsburgh, PA, 1992, ACM: New York, 1992; pp 144−152.
(20) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093−1110.
(21) Wassermann, A. M.; Heikamp, K.; Bajorath, J. Potency-directed similarity searching using support vector machines. *Chem. Biol. Drug Des.* **2011**, *77*, 30−38.
(22) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's bioassay database. *Nucleic Acids Res.* **2012**, *40*, D400−D412.
(23) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(24) *Molecular Operating Environment (MOE)*; Chemical Computing Group, Inc.: Montreal, Quebec, Canada.
(25) Witten, I. H.; Frank, E. *Data Mining − Practical Machine Learning Tools and Techniques*, ed. 2; Morgan Kaufmann: San Francisco, 2005, pp 161−176.
(26) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods − Support Vector Learning*; Schölkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT-Press: Cambridge, MA, 1999; pp 169−184.

Summary

SVM calculations were performed on compound data sets with single-, dual- and triple-target activities. Standard SVM-based compound ranking was shown to result in a high recall of active compounds, but failed to distinguish compounds having different activity profiles. However, differently weighted SVM linear combinations derived clear separations in the recall of compounds having related and overlapping target activity annotations. The combination of individual SVM models using positive and negative linear factors into a single multi-class prediction model essentially retained the recall of compounds with desired activities and simultaneously reduced the detection of those having other activity profiles.

In this study, the SVM search calculations were influenced by the composition of the compound data sets used for training, without considering activity explicitly as a search parameter. In the following study, compound activity is considered directly and incorporated into the SVM method through the design of a potency-oriented SVM linear combination approach and a structure-activity kernel.

Chapter 4

Potency-directed similarity searching using support vector machines

Introduction

Several different similarity-based methods exist to identify novel active compounds. However, similarity searching and machine learning methods usually do not consider compound potency as a search parameter. Here, we introduce two SVM approaches that incorporate potency information as a parameter in order to direct search calculations towards the preferential detection of highly potent compounds. A structure-activity kernel function and a potency-oriented SVM linear combination are designed that take potency annotations of reference compounds into account. On public high-throughput screening sets, these approaches show an enrichment of highly potent compounds at the top positions of database rankings.

Chem Biol Drug Des 2011; **77:** 30–38

Research Article

doi: 10.1111/j.1747-0285.2010.01059.x

Potency-Directed Similarity Searching Using Support Vector Machines

Anne M. Wassermann†, Kathrin Heikamp† and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstr. 2, D-53113 Bonn, Germany
**Corresponding author: Jürgen Bajorath, bajorath@bit.uni-bonn.de*
† The contributions of these authors should be considered equal.

Support vector machine modeling has become increasingly popular in chemoinformatics. Recently, several advanced support vector machine applications have been reported including, among others, multitask learning for ligand–target prediction. Here, we introduce another support vector machine approach to add compound potency information to similarity searching and enrich database selection sets with potent hits. For this purpose, we introduce a structure–activity kernel function and a potency-oriented support vector machine linear combination approach. Using fingerprint descriptors, potency-directed support vector machine searching has been successfully applied to four high-throughput screening data sets, and different support vector machine strategies have been compared. For potency-balanced compound reference sets, potency-directed support vector machine searching meets or exceeds recall rates of standard support vector machine calculations but detects many more potent hits.

Key words: compound classification, compound potency, kernel functions, molecular similarity, similarity searching, support vector machines

Received 30 August 2010, revised 20 October 2010 and accepted for publication 20 October 2010

Support vector machine (SVM) learning is currently widely applied in chemoinformatics for a variety of applications including compound database searching (1,2). In addition to SVM, there are many ligand similarity-based methods that are utilized to mine databases for novel active compounds (1,2). However, with the exception of QSAR models (3), these approaches typically do not consider compound potency as search information (2). Thus far, it has only rarely been attempted to incorporate potency information into search algorithms (4), although the ability to direct search calculations toward the recognition of potent hits would certainly be attractive for practical applications.

Support vector machine is a supervised machine learning methodology for object classification and class label prediction (5,6). In the training phase, learning sets with different class labels are projected into feature space, and a hyperplane is constructed in this reference space to best separate objects belonging to different classes. These linear models are then applied for predictions, for example, of active compounds contained in screening databases. For SVM learning, kernel functions (7,8) are applied to project objects into reference spaces of increasing dimensionality to solve classification problems through hyperplane construction that are non-linear in lower dimensional spaces.

In chemoinformatics and compound screening, SVM has steadily gained in popularity over the past few years, in particular, because of its generally high classification performance (1). Here, SVM learning has originally been applied to build predictive models for binary classification of active and inactive compounds (9,10). In addition, further advanced SVM strategies have recently been introduced. For example, SVM has been adapted for similarity-based ranking of molecular databases (11–13) and multitask learning applied, for example, to predict ligands for orphan targets (14–16). Furthermore, SVM strategies have been developed to build SAR models from multiple assays (17) and perform multi-class label predictions, for example, for target fishing (18) or selectivity profiling (12,19,20). To predict selectivity toward human adenosine receptors (hARs), a recent study (20) used a multi-label approach (termed ct-SVM) to construct a single model integrating binary classifiers for four different hARs subtypes. Furthermore, three models based on increasingly strict criteria for threshold activity (i.e., K_i threshold values of 500, 250, and 100 nM) were applied sequentially to quantify the biological affinity of test compounds. This analysis demonstrated that SVM-based classification provides an interesting alternative to traditional regression-based QSAR modeling. This study aimed at the annotation of test compounds with predefined potency ranges. It currently represents the only non-QSAR SVM-based classification of different biological activity levels.

In this study, we present SVM strategies for potency-directed similarity searching. For this purpose, a new structure–activity kernel function is introduced, and potency-oriented SVM linear combinations (LCs) are constructed. In contrast to standard SVM learning based on binary class labels (active or inactive), these SVM techniques distinguish between highly, intermediately, or weakly active compounds by incorporating categorized potency labels of reference molecules into training. In test calculations on different public domain screening data sets, a preferential enrichment of top-ranked

positions with highly potent hits has been observed for potency-directed SVM searching compared to conventional SVM ranking.

Materials and Methods

Data sets

The four confirmatory high-throughput screening (HTS) assays used in this study have been extracted from PubChem[a] and include inhibition assays for enzyme targets hydroxyacyl-coenzyme A dehydrogenase type II [assay id (AID) 886], 15-human lipoxygenase (AID 887), 15-hydroxyprostaglandin dehydrogenase (AID 894), and aldehyde dehydrogenase 1 (AID 1030). Their composition is summarized in Table 1. Compound potencies are reported as half-maximal inhibitory concentrations (IC_{50} values). After standardization of structural representations using the Molecular Operating Environment[b], compounds with incomplete or ambiguous activity annotations or with fewer than five non-hydrogen atoms were removed from these data sets. Then, a 2D unique version of each compound set was generated, i.e., of molecules having the same 2D molecular graph (i.e., the same 2D structure), only the one with highest potency was retained. Statistics for the so-prepared compound data sets are reported in Table 1.

It should be emphasized that active compounds in all four data sets covered wide potency ranges of more than three orders of magnitude. Furthermore, in each data set, there were many more weakly than highly potent compounds (Table 1). Thus, for potency-directed similarity searching, these data sets provided challenging test cases.

For each data set, potency intervals were defined to divide active compounds into four potency categories, termed *C1–C4*, with potency values decreasing from C1 to C4. For each category, the negative decadic logarithm of the potency value of its lower potency threshold was calculated and used as its annotation, *pAct*, as also reported in Table 1.

Support vector machine search strategies

Standard SVM

As a supervised learning technique, SVM utilizes training sets for model building. For training objects projected into feature spaces, a linear decision function is built that divides the objects into two classes. Mathematical details are provided as Appendix S1. In the case of linearly inseparable training classes, the scalar product calculated to construct the hyperplane (see Appendix S1) is generally replaced by a kernel function to project compounds into a higher dimensional space where a linear separation might be possible. Test compounds are then ranked according to their distance from the separating hyperplane (11). Thus, compounds are ranked from the most distant object on the positive half-space to the most distant object on the negative half-space.

For standard SVM calculations, compound training sets of different composition were used. First, active compounds from all potency categories were pooled to provide the positive training class, and a randomly selected subset of inactive molecules was used as the negative training class, as shown in Figure 1A on the left. This SVM strategy is referred to as *SVMpooled*. Furthermore, for control calculations, reference sets exclusively containing highly potent compounds were also utilized, as illustrated in Figure 1A on the right.

Linear combination

The SVM LC strategy was originally introduced by Geppert *et al.* (15) in the context of ligand predictions for orphan targets and is adapted herein for potency-directed SVM searching. Therefore, for each potency category, a hyperplane is constructed using known

Table 1: Data sets

AID	Target	# Active	# Inactive	Potency categories	pAct	Cat	# Mol	# Ref
886	Hydroxyacyl-coenzyme A dehydrogenase type II	2409	68 845	10–100 nM	7	C1	20	5
				100 nM–1 μM	6	C2	128	32
				1–10 μM	5	C3	803	200
				10–100 μM	4	C4	1458	364
887	15-human lipoxygenase	998	70 822	2–200 nM	6.7	C1	16	5
				200 nM–2 μM	5.7	C2	93	29
				2–20 μM	4.7	C3	711	222
				20–200 μM	3.7	C4	178	55
894	15-hydroxy-prostaglandin dehydrogenase	6318	139 805	1–100 nM	7	C1	12	5
				100 nM–1 μM	6	C2	115	47
				1–10 μM	5	C3	1452	605
				10–100 μM	4	C4	4739	1974
1030	Aldehyde dehydrogenase 1	15 817	197 666	10–100 nM	7	C1	138	5
				100 nM–1 μM	6	C2	946	34
				1–10 μM	5	C3	6091	220
				10–100 μM	4	C4	8642	313

AID, assay id.

For four confirmatory high-throughput screening data sets, the numbers of active (# active) and inactive (# inactive) compounds are reported. For each data set, the ranges of the four potency categories (cat) C1–C4 into which active compounds were divided and the potency threshold values pAct are given. Furthermore, the numbers of molecules per potency category (# mol) and reference compounds (# ref) taken from each category are reported.

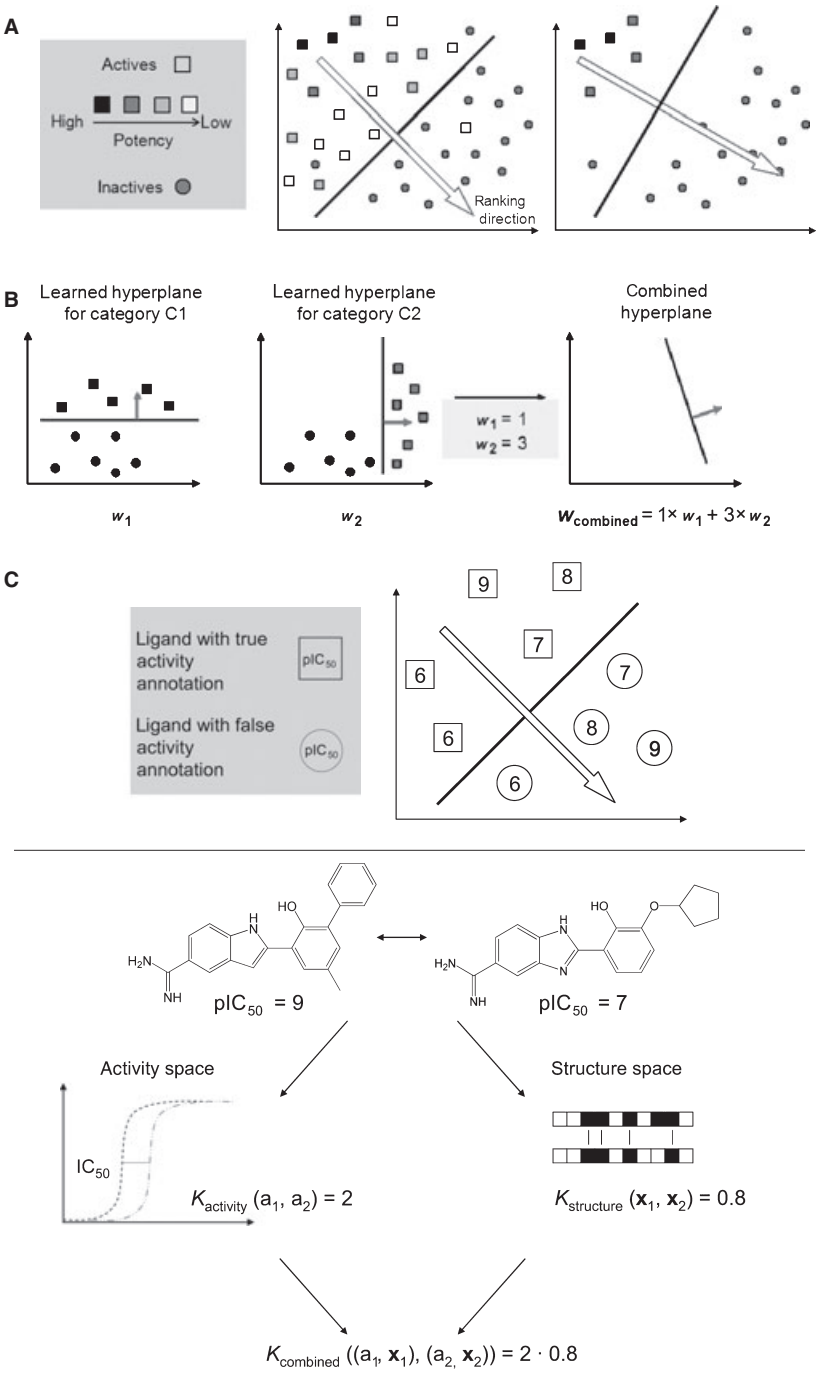


Figure 1: Support vector machine (SVM) strategies. (A) In *SVMpooled* (left), active compounds from all potency categories are pooled to form the positive training set, and randomly selected inactive compounds constitute the negative training class. Test compounds are ranked according to their signed distance from the hyperplane represented by the arrow. For control calculations (right), only reference compounds from the two highest potency categories are used as positive training examples. (B) In *SVM linear combination*, weight vectors are first generated for each potency category. The weight vectors are then linearly combined to a final weight vector used for compound ranking. (C) In *SVM structure–activity kernel*, a hyperplane is constructed to separate true potency category–feature vector pairings from false pairings (see text). The comparison of two potency category–feature vector pairs is divided into two independent tasks such that the structural similarity and activity similarity of two compounds are first separately determined and then combined.

active ligands of this category as positive training objects and randomly selected inactive compounds as negative examples. To obtain a ranking function, the individual hyperplanes are linearly combined to yield a single hyperplane. The weights given to individual hyperplanes in the LC increase with the potency of the active training molecules (for further details, see Appendix S1). Analogous to *SVMpooled*, test compounds are then ranked based on their signed distance to the combined hyperplane. This strategy is termed *LCsimple*. To further increase weights on highly active compounds, the *LCsquared* strategy is introduced that utilizes the square product of the linear factor used in *LCsimple* as the potency category-specific weight for the LC (Appendix S1). The SVM LC approach is exemplarily illustrated for the combination of two potency categories with arbitrarily chosen linear factors in Figure 1B.

Structure–activity kernel

We also introduce a *structure–activity kernel* (SAK) that is conceptually related to kernels for the comparison of different target–ligand pairs (14,15). To calculate the scalar products for target–ligand pairs during SVM optimization and ranking, the target–ligand kernel was defined as the product of two separate kernels for a target pair and a ligand pair. Here, we represent each compound as

a potency category–feature vector pair (a_i, $\mathbf{x}_i$). Accordingly, the comparison of two compounds is divided into a separate assessment of their activity similarity and their structural similarity by two different kernel functions K_{activity} and $K_{\text{structure}}$ that are then combined to build the SAK:

$$K((a_i, \mathbf{x}_i)(a_k, \mathbf{x}_k)) = K_{\text{activity}}(a_i, a_k) \times K_{\text{structure}}(\mathbf{x}_i, \mathbf{x}_k)$$

In analogy to SVM LC (see Appendix S1), the activity kernel is defined as

$$K_{\text{activity}}(a_i, a_k) = \max_{j=1\ldots n}(a_j) - \min_{j=1\ldots n}(a_j) + 1 - |a_i - a_k|$$

for the approach *SAKsimple*

or as

$$K_{\text{activity}}(a_i, a_k) = \left(\max_{j=1\ldots n}(a_j) - \min_{j=1\ldots n}(a_j) + 1 - |a_i - a_k|\right)^2$$

for the approach *SAKsquared*.

The design principle of SAK is illustrated in Figure 1C. The hyperplane is, in this case, designed to separate true compound fingerprint–potency pairings from false pairings. For SVM training, positive training objects (true pairings) were obtained by combining the fingerprint representation of each active compound with its potency category threshold value and negative training examples (false pairings) by randomly selecting inactive compounds and combining their fingerprint representations with all possible potency category threshold values. For the classification of molecules with unknown activity (i.e., active versus inactive) or potency, test compounds were assigned the highest potency category c_{high}. Then, the hyperplane was utilized to assess the probability of true or false assignments. A ranking of test compounds was then generated by determining the signed distance from the pairs (c_{high}, $\mathbf{x}$) to the hyperplane H derived in structure–activity reference space.

Test calculations

The performance of the alternative SVM ranking strategies was evaluated in search calculations on the four PubChem HTS data sets. Compounds were represented by MACCS structural keys[c] or the ECFP4 fingerprint[d] (21). To compare fingerprint representations, the Tanimoto kernel (22) was utilized. For test calculations, reference compound sets were assembled to reflect the potency distribution in each data set. Accordingly, five compounds belonging to the highest potency category were randomly selected in each case, and reference compounds of the other categories were chosen such that the reference-to-test molecule ratio was approximately the same for all potency categories, as reported in Table 1. Control calculations were carried out with reference sets containing only highly potent compounds. For all assays and SVM strategies, 1000 inactive compounds were taken as negative training examples. All remaining molecules from each data set were utilized as the screening database. For each combination of a search strategy and fingerprint, 10 different trials with randomly assembled reference and test sets were carried out. As a measure of performance, recovery rates (RR: number of correctly identified active molecules divided by their total number) were calculated for database selection sets of increasing size and averaged over the 10 independent trials per target.

All calculations were carried out using SVM$^{\text{light}}$ (23), a freely available SVM implementation.[e] Calculation parameters were suggested SVM$^{\text{light}}$ default settings to ensure reproducibility of the calculations. Perl scripts were applied to calculate SVM LCs.

Results

With our study, we aimed at investigating whether compound potency could be incorporated as a search parameter into SVM-based similarity searching to direct the calculations toward the identification of potent hits.

HTS data and reference compounds

High-throughput screening data were selected for benchmark calculations to provide a practically relevant search scenario. Two types of compound reference sets were assembled for our analysis including 'potency-balanced' reference sets that mirrored the potency distribution of each data set, as reported in Table 1, and also reference sets only consisting of compounds falling into the potency ranges C1 (*1Cat*) or C1 and C2 (*2Cat*). These biased reference sets were used to evaluate whether potent reference compounds would lead to the preferential detection of potent hits.

Advanced SVM strategies

We have compared standard SVM calculations with two potency-directed SVM techniques including a structure–activity kernel taking reference compound potency differences directly into account and, in addition, the LC of different SVM hyperplanes derived for reference compounds falling into different potency ranges. Both the SAK and LC strategies were also tested with 'squared' weights, i.e., by further emphasizing the contributions of potent reference compounds. In the following, SAK and LC are also referred to as *advanced SVM strategies*.

Search performance

We first compared alternative SVM strategies for potency-balanced reference sets. The results for the ECFP4 and MACCS fingerprints are reported in Figures 2 and S1, respectively. In these figures, compound recall of all active compounds (regardless of their potency) and of the most potent compounds (categories C1, C2) is separately monitored. Compound recall was generally higher for ECFP4 than for MACCS. Overall, the average recovery rates of all active compounds were comparable for standard SVM and advanced SVM strategies. However, in all cases, SAK and LC calculations were found to retrieve a higher percentage of highly potent compounds than standard SVM calculations. Although search results for SAK and LC were very similar, some underlying trends and characteristics of the individual methods were detected. Independent of the fingerprint representation, SAK usually identified more active compounds belonging to the highest potency category than LC. Because the overall compound recall was comparable for all advanced strategies using the ECFP4 fin-

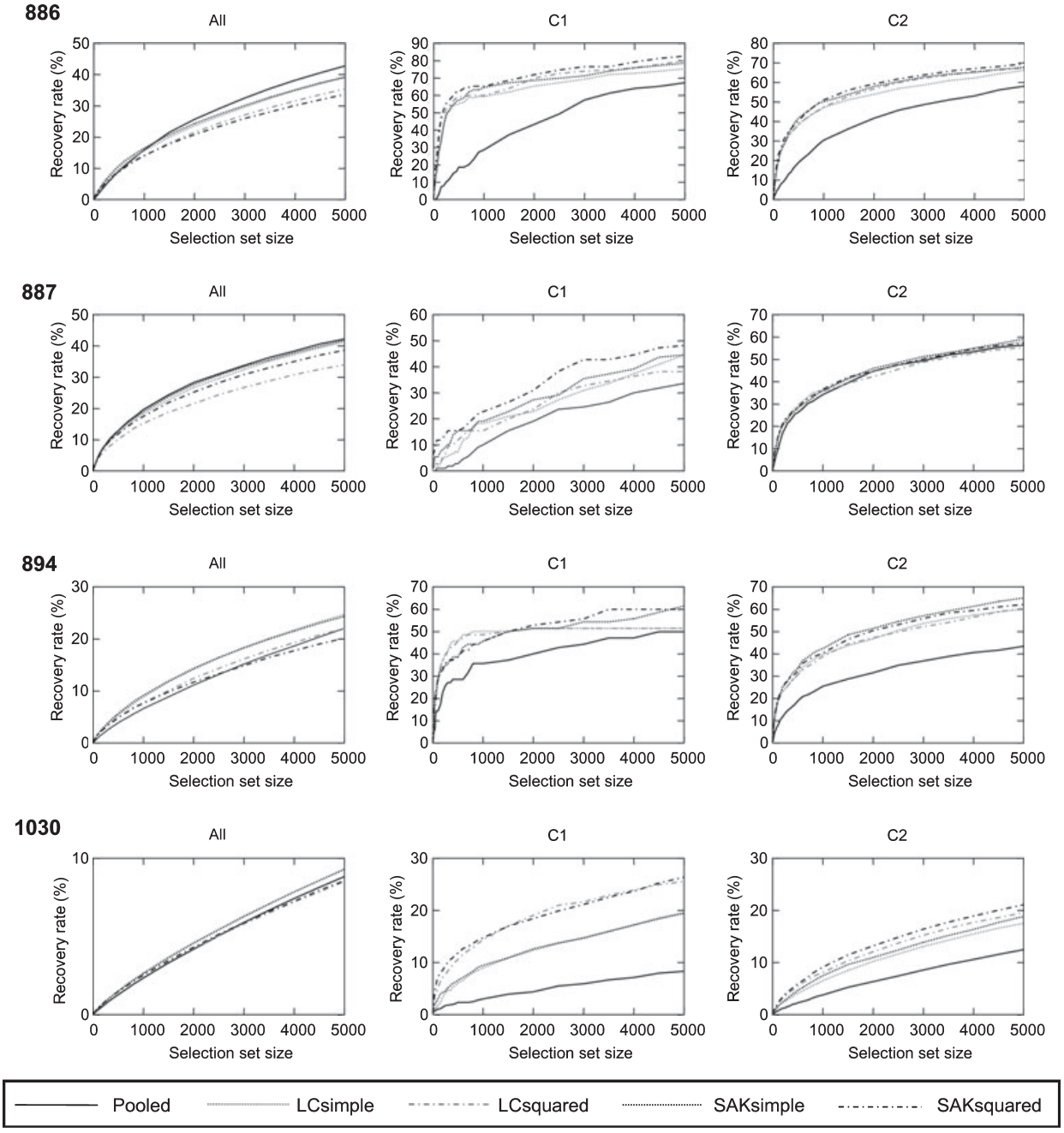


Figure 2: Cumulative recall curves for potency-balanced reference sets. For each bioassay, cumulative recall curves are shown for all active compounds and the highest potency categories (C1 and C2) and different support vector machine strategies using the ECFP4 fingerprint. Recall curves represent the average of 10 independent trials using different reference sets. Potency-balanced reference sets consist of compounds spanning the entire potency range in a data set.

gerprint, SAK was considered as the preferred strategy for this fingerprint. However, for the MACCS fingerprint, overall compound recall was consistently higher for LC than for SAK such that there was no clear advantage of one over the other method. Furthermore, the search results for simple and squared weights were also comparable. However, the use of simple weights often led to slightly higher recovery rates for all active compounds, whereas squared weights favored the recovery of highly potent molecules.

In Figures 3 and S2, average recovery rates are reported for a constant selection set size of 1000 database compounds. Depending on the HTS data set, recovery rates for all active compounds ranged from ~3% to ~20%. For potent (C1, C2) compounds, higher recovery rates were observed ranging, on average, from ~10% to ~60%. Here, it should be taken into account that many more weakly than highly potent compounds were available in each data set. The comparison of the recall rates of alternative SVM strate-

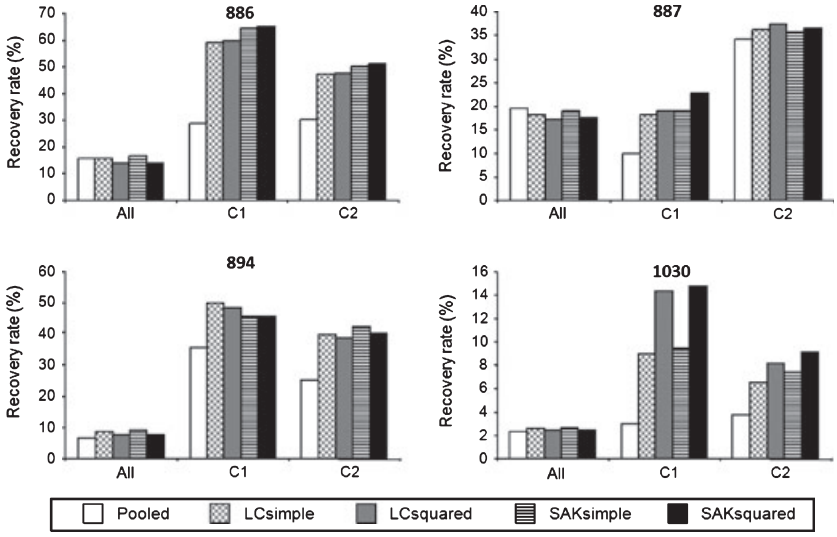


Figure 3: Support vector machine performance for database selection sets of constant size. Recovery rates are shown for the ECFP4 fingerprint, potency-balanced reference sets, and database selection sets of 1000 database compounds. The results are averaged over 10 independent trials per data set.

gies for selection sets of 1000 database compounds further illustrated that SAK and LC calculations consistently detected more potent compounds than standard SVM calculations ('Pooled'). Thus, potency-directed SVM searching reached the recall performance of standard SVM classification but led to the desired preferential detection of hits having higher potency.

Control calculations

We next carried out standard SVM calculations on reference sets exclusively consisting of potent compounds. The results were then compared to standard SVM and SAK calculations for potency-balanced reference sets. These control calculations were carried out to determine to what extent the potency of reference compounds determined the outcome of the search calculations relative to advanced SVM strategies. The results for the ECFP4 and MACCS fingerprints are reported in Figures 4 and S3, respectively. It can be seen that standard SVM using only the five most potent reference compounds as positive training examples (strategy 1Cat) produced recovery rates of potent compounds that were significantly lower compared to advanced strategies, which was especially obvious for the recall of potent compounds belonging to category C2. In most cases, even standard SVM using potency-balanced reference sets recognized more C2 compounds. Of course, the C1 reference set was the smallest of all and hence contained the least information about active molecules. Accordingly, when adding reference compounds falling into potency category C2 to the positive training class (strategy 2Cat), recovery rates of potent compounds increased and were found to be overall comparable to advanced strategies (or even slightly better in case of the MACCS fingerprint). Thus, the exclusive use of highly potent reference compounds in standard SVM calculations also led to the preferential detection of potent screening hits. However, there was a price to pay, because in this case, the recovery rates of all active compounds were substantially reduced for standard SVM calculations. Thus, overall, much better recall rates of active compounds were obtained for balanced reference sets where potency-directed SVM searching provided a clear enrichment of potent screening hits.

Discussion

In ligand similarity-based database searching, one typically attempts to distinguish active from inactive compounds but rarely considers compound potency information to further refine the search calculations. This sets conventional similarity searching apart from QSAR approaches. However, the inclusion of available potency information would certainly be meaningful for practical similarity search applications. For many similarity-based search methods, the incorporation of potency as a search parameter is a difficult problem. However, in the context of SVM learning, the use of kernel functions and their combination provides a basis for the design and implementation of a multi-parametric search approach. Support vector machine LC learns separate hyperplanes for training sets of different activity ranges and then combines them by associating a potency-dependent weighting scheme. By contrast, the SAK approach introduced herein compares compound pairs simultaneously in activity and structure space by evaluating structural similarity on the basis of whole-molecule fingerprint descriptors and multiplying it with an assessment of activity similarity for pairs of ligands. Using balanced (unbiased) compound reference sets, both advanced SVM techniques met the active compound-recall performance of conventional SVM calculations but achieved a clear enrichment of potent hits. In addition, we demonstrated that reference sets biased toward compounds having high potency also led to an enrichment of potent hits in standard SVM calculations, but only at the cost of recall performance.

These findings have a number of implications for practical SVM database search applications. We deliberately performed our analy-

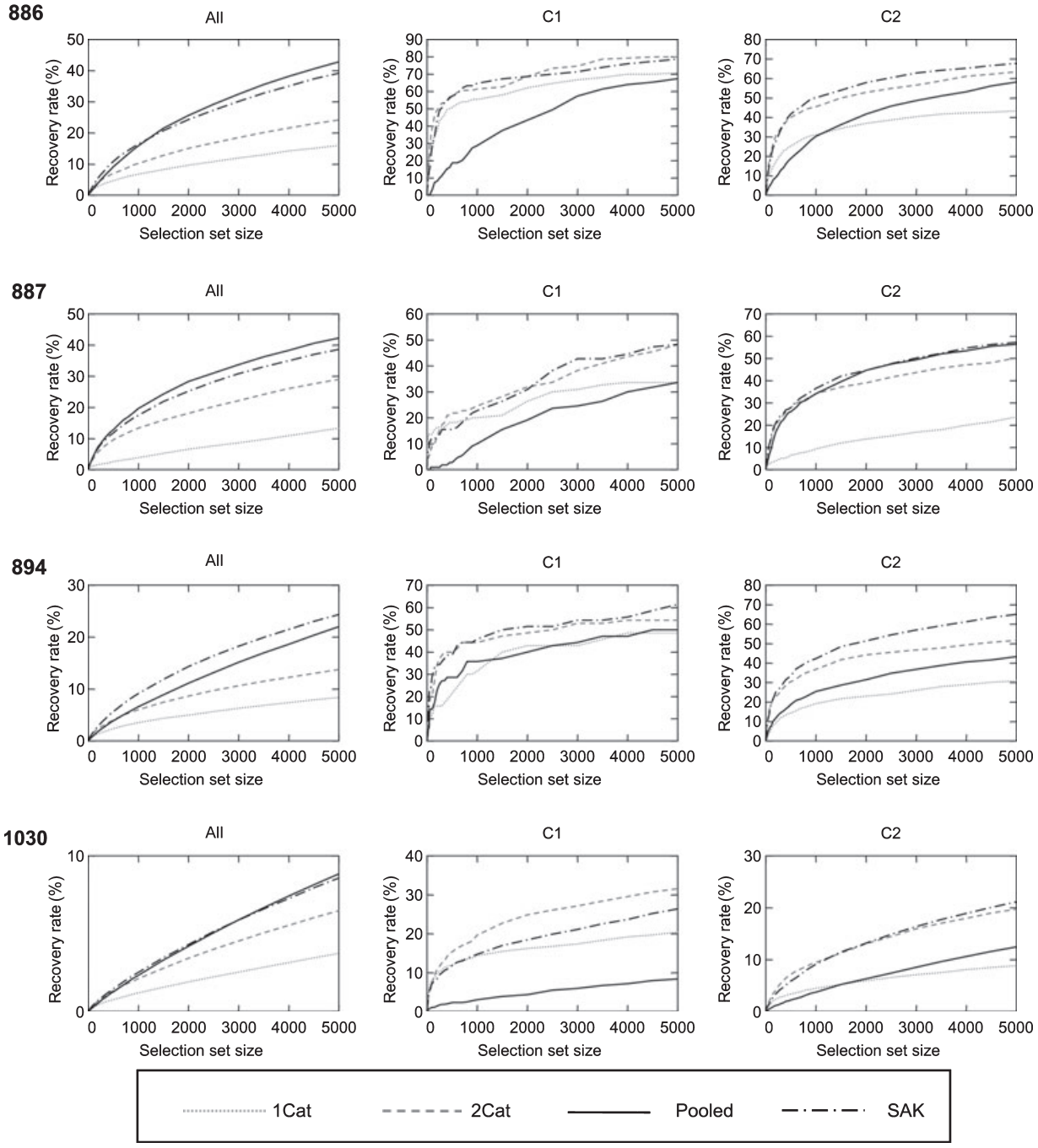


Figure 4: Cumulative recall curves for potency-balanced and highly potent reference compounds. For each bioassay, recall curves are shown for all active compounds and the two highest potency categories (C1 and C2). Compound recall is monitored for different support vector machine (SVM) strategies using the ECFP4 fingerprint averaged over 10 independent trials. The following strategies are compared: standard SVM with reference compounds from potency category 1 ('1Cat'), 1 and 2 ('2Cat'), and all categories ('Pooled') and SVM structure–activity kernel (SAK). *SAKsimple* is shown for sets 886 and 894, *SAKsquared* for sets 887 and 1030.

sis on HTS data to eliminate the influence of molecular complexity effects (24) on the search results. In typical screening libraries, hits with different potency usually have comparable molecular weight and topological complexity because they are not (yet) chemically optimized with respect to a specific biological activity. This avoids complications that are often associated with benchmark calculations and also practical applications. In typical benchmark settings, highly optimized and potent compounds are usually added to screening databases consisting of lower complexity compounds, which generally yields artificially high recall rates (1), because highly complex reference and active database compounds are relatively easy to distinguish from screening molecules having lower complexity. How-

ever, the situation is completely different when highly complex reference compounds are utilized to search for hits having average screening database complexity, which has been shown to provide the by far most difficult practical search scenario for fingerprint-based methods (25). These considerations would suggest to better focus on screening hits as reference compounds, even if many of them might only be weakly potent (24). For SVM learning, we now introduce techniques that take relative compound potency into account and are particularly well suited for this task. Selecting a spectrum of available screening hits for learning, the SVM SAK and LC techniques would be expected to detect many active compounds and direct the search toward potent hits, if available in a screening database. Because potency-directed SVM searching ultimately detects active compounds on the basis of a (weighted) 'structure/potency similarity compromise', such calculations should be particularly promising if reference compounds and potential hits would originate from the same screening collection (where many active compounds might have similar chemical properties). For example, this would make the application of these methods attractive in the context of sequential screening (26) where initial screening HTS hits from a fraction of the database are used as reference compounds for search calculations to prioritize another subset of the database (with a putative enrichment of additional hits) for the next round of experimental screening.

Conclusions

In conclusion, herein, we have introduced and evaluated SVM-based techniques for potency-directed similarity searching, for which alternative methods currently are not available. Potency-directed SVM searching further extends the current spectrum of advanced SVM approaches for different chemoinformatics applications. Both the SVM LC and structure–activity kernel showed a notable enrichment of potent compounds relative to standard SVM ranking. Different from SAK, the LC approach requires the availability of distinct learning sets with sufficient numbers of compounds at different potency levels, which might often be difficult to obtain for practical applications. However, even if only very small numbers of highly potent compounds are available, the application of SAK is still feasible. One of the attractive features of potency-directed LC and SAK calculations is that high recall rates of active compounds are obtained and that the searches are not strongly focused on the exclusive recognition of highly potent compounds. This provides a meaningful compromise between the recall of active compounds and the enrichment of potent hits relative to standard SVM ranking schemes.

Acknowledgments

The authors thank Martin Vogt for help with SVM^{light}.

References

1. Geppert H., Vogt M., Bajorath J. (2010) Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model;50:205–216.
2. Eckert H., Bajorath J. (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov Today;12:225–233.
3. Esposito E.X., Hopfinger A.J., Madura J.D. (2004) Methods for applying the quantitative structure-activity relationship paradigm. Methods Mol Biol;275:131–214.
4. Vogt I., Bajorath J. (2007) Analysis of a high-throughput screening data set using potency-scaled molecular similarity algorithms. J Chem Inf Model;47:367–375.
5. Vapnik V.N. (2000) The Nature of Statistical Learning Theory, 2nd edn. New York: Springer.
6. Boser B.E., Guyon I.M., Vapnik V. (1992) A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual Workshop on Computational Learning Theory. Pittsburgh, Pennsylvania. New York: ACM; p. 144–152.
7. Müller K.-R., Mika S., Rätsch G., Tsuda K., Schölkopf B. (2001) An introduction to kernel-based learning algorithms. IEEE Neural Netw;12:181–201.
8. Schölkopf B., Smola A. (2002) Learning with Kernels. Cambridge, MA: MIT Press.
9. Burbidge R., Trotter M., Buxton B., Holden S. (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput Chem;26:5–14.
10. 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. J Chem Inf Comput Sci;43:667–673.
11. Jorissen R.N., Gilson M.K. (2005) Virtual screening of molecular databases using a support vector machine. J Chem Inf Model;45:549–561.
12. Wassermann A.M., Geppert H., Bajorath J. (2009) Searching for target-selective compounds using different combinations of multiclass support vector machine ranking methods, kernel functions, and fingerprint descriptors. J Chem Inf Model;49:582–592.
13. Agarwal S., Dugar D., Sengupta S. (2010) Ranking chemical structures for drug discovery: a new machine learning approach. J Chem Inf Model;50:716–731.
14. Jacob L., Vert J.-P. (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics;24:2149–2156.
15. Geppert H., Humrich J., Stumpfe D., Gärtner T., Bajorath J. (2009) Ligand prediction from protein sequence and small molecule information using support vector machines and fingerprint descriptors. J Chem Inf Model;49:767–779.
16. Wassermann A.M., Geppert H., Bajorath J. (2009) Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. J Chem Inf Model;49:2155–2167.
17. Ning X., Rangwala H., Karypis G. (2009) Multi-assay-based structure–activity relationship models: improving structure–activity relationship models by incorporating activity information from related targets. J Chem Inf Model;49:2444–2456.
18. Wale N., Karypis G. (2009) Target fishing for chemical compounds using target-ligand activity data and ranking based methods. J Chem Inf Model;49:2190–2201.

19. Michielan L., Terfloth L., Gasteiger J., Moro S. (2009) Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome P450 substrates. J Chem Inf Model;49:2588–2605.
20. Michielan L., Stephanie F., Terfloth L., Hristozov D., Cacciari B., Klotz K.-N., Spalluto G., Gasteiger J., Moro S. (2009) Exploring potency and selectivity receptor antagonist profiles using a multilabel classification approach: the human adenosine receptors as a key study. J Chem Inf Model;49:2820–2836.
21. Rogers D., Hahn M. (2010) Extended-connectivity fingerprints. J Chem Inf Model;50:742–754.
22. Ralaivola L., Swamidass S.J., Saigo H., Baldi P. (2005) Graph kernels for chemical informatics. Neural Netw;18:1093–1110.
23. Joachims T. (1999) Making large-scale SVM learning practical. In: Schölkopf B., Burges C., Smola A., editors. Advances in Kernel Methods – Support Vector Learning. Cambridge, MA: MIT-Press: p. 169–184.
24. Wang Y., Bajorath J. (2010) Advanced fingerprint methods for similarity searching: balancing molecular complexity effects. Comb Chem High Throughput Screen;13:220–228.
25. Wang Y., Bajorath J. (2008) Balancing the influence of molecular complexity on fingerprint similarity searching. J Chem Inf Model;48:75–84.
26. Parker C.N., Shamu C.E., Kraybill B., Austin C.P., Bajorath J. (2006) Measure, mine, model, and manipulate: the future for HTS and chemoinformatics? Drug Discov Today;11:863–865.

Notes

[a]Pubchem, http://pubchem.ncbi.nlm.nih.gov.
[b]Molecular Operating Environment (MOE), Chemical Computing Group Inc., Montreal, Quebec, Canada, http://www.chemcomp.com.
[c]MACCS Structural Keys, Symyx Technologies, Inc., Sunnyvale, CA, USA, http://www.symyx.com.
[d]Scitegic Pipeline Pilot, Accelrys Inc., San Diego, CA, USA, http://accelrys.com/products/scitegic/index.html.
[e]SVM[light], http://svmlight.joachims.org.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. Methods.

Figure S1. Cumulative recall curves for potency-balanced reference sets.

Figure S2. SVM performance for database selection sets of constant size.

Figure S3. Cumulative recall curves for potency-balanced and highly potent reference compounds.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Summary

Two SVM-based approaches have been introduced that incorporate compound potency information for potency-directed LBVS. The structure-activity kernel is designed to compare compounds by separately assessing structural similarity and the similarity of the activity annotations of the ligands. In the potency-oriented SVM linear combination, hyperplanes were derived for compounds from different potency categories and then combined using linear factors reflecting the potency level. Both approaches have been applied to potency-balanced data sets and compared to standard SVM-based compound ranking. The potency-directed SVM approaches were found to meet or exceed the active compound recall performance of standard SVM calculations and furthermore showed a clear early enrichment of potent compounds.
The supporting information of this publication can be obtained via the following URL: http://dx.doi.org/10.1111/j.1747-0285.2010.01059.x.

The idea of designing new kernel functions to account for specific data types and addressing questions in VS is adapted in the following study. Here, kernel functions are designed that compare compound pairs in order to predict activity cliffs, a problem not considered by standard similarity-based search methods.

Chapter 5

Prediction of activity cliffs using support vector machines

Introduction

Similarity-based search methods cannot account for discontinuous SARs. It is therefore of high interest to identify those pairs of structurally similar compounds in a data set that have large differences in their potency. In the following study, newly designed kernel functions are introduced that enable comparisons of compound pairs with the objective to predict activity cliffs. Additionally, a substructure representation is designed to encode substructural differences between molecule pairs. SVM calculations using the new kernel functions and substructure representation are applied to predict activity cliffs in several compound data sets.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

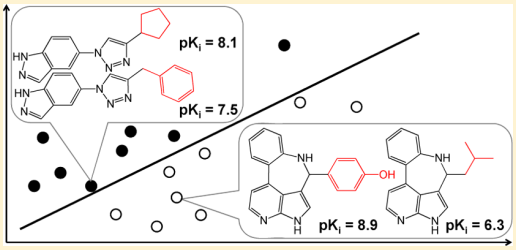
Prediction of Activity Cliffs Using Support Vector Machines

Kathrin Heikamp,[†,§] Xiaoying Hu,[†,‡,§] Aixia Yan,[‡] 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

[‡]State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, P.O. Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road, Beijing 100029, People's Republic of China

ABSTRACT: Activity cliffs are formed by pairs of structurally similar compounds that act against the same target but display a significant difference in potency. Such activity cliffs are the most prominent features of activity landscapes of compound data sets and a primary focal point of structure−activity relationship (SAR) analysis. The search for activity cliffs in various compound sets has been the topic of a number of previous investigations. So far, activity cliff analysis has concentrated on data mining for activity cliffs and on their graphical representation and has thus been descriptive in nature. By contrast, approaches for activity cliff prediction are currently not available. We have derived support vector machine (SVM) models to successfully predict activity cliffs. A key aspect of the approach has been the design of new kernels to enable SVM classification on the basis of molecule pairs, rather than individual compounds. In test calculations on different data sets, activity cliffs have been accurately predicted using specifically designed structural representations and kernel functions.



1. INTRODUCTION

In medicinal chemistry and chemoinformatics, the study of activity cliffs has experienced increasing interest in recent years.[1] Activity cliffs are generally defined as pairs or groups of chemically similar compounds with large potency differences (i.e., usually at least 2 orders of magnitude).[1,2] In chemoinformatics, the exploration of activity cliffs is a topic of interest because qualifying compound pairs can be identified through mining of compound data sets, hence enabling large-scale SAR analysis.[3] In medicinal chemistry, activity cliffs and their structural neighborhoods are considered a prime source of SAR information, given that small chemical differences lead to large bioactivity effects.[1] In traditional medicinal chemistry, activity cliffs are often analyzed in individual compound series. However, they are also systematically explored. For example, in independent studies, activity cliffs were systematically identified and characterized in different data sets.[3−6] In these investigations, cliffs were often defined and represented in rather different ways. Furthermore, activity cliff distributions in current bioactive compounds have been determined through systematic data mining.[7] Moreover, many compound data sets have also been searched for higher-order activity cliff arrangements such as activity ridges[8] and coordinated cliffs.[9] Taken together, these investigations were primarily focused on compound data mining for and visualization of activity cliffs, as mentioned above. Clearly, activity cliff analyses available at present are descriptive in nature, as an integral part of large-scale SAR exploration.[3] By contrast, no attempts have thus far been reported to develop computational models for activity cliff prediction. Here, we present a first step in this direction. Support vector machine (SVM)[10] models have been developed to screen compound data sets and predict activity cliffs. In the following, the derivation and evaluation of these models and the underlying methodology are described in detail.

2. ACTIVITY CLIFF REPRESENTATION AND DATA SETS

2.1. Compound Pairs. For any study of activity cliffs, a molecular representation must be selected that provides a basis for pairwise compound similarity assessment. For our analysis, we have applied the concept of matched molecular pairs (MMPs)[11] to represent activity cliffs, following the recent introduction of MMP-based cliffs.[12] An MMP is defined as a pair of compounds that differ only at a single site, i.e., a substructure such as a ring or an R group. Hence, two compounds forming an MMP share a common core and are distinguished by a molecular transformation, i.e., the exchange of a pair of substructures, which converts one compound into the other. The exchange of a substructure can also induce changes in physicochemical properties such as, for example, lipophilicity, charge, or hydrogen bond potential. Compared to other similarity measures, an advantage of the MMP formalism in the context of activity cliff analysis is that the structural difference between compounds in a pair is well-defined and limited to a single substructure. This represents a clearly defined and chemically intuitive criterion for cliff formation that does not rely on calculated similarity values. Furthermore, this approach is consistent with the basic idea of the activity cliff

Received: July 2, 2012
Published: August 15, 2012

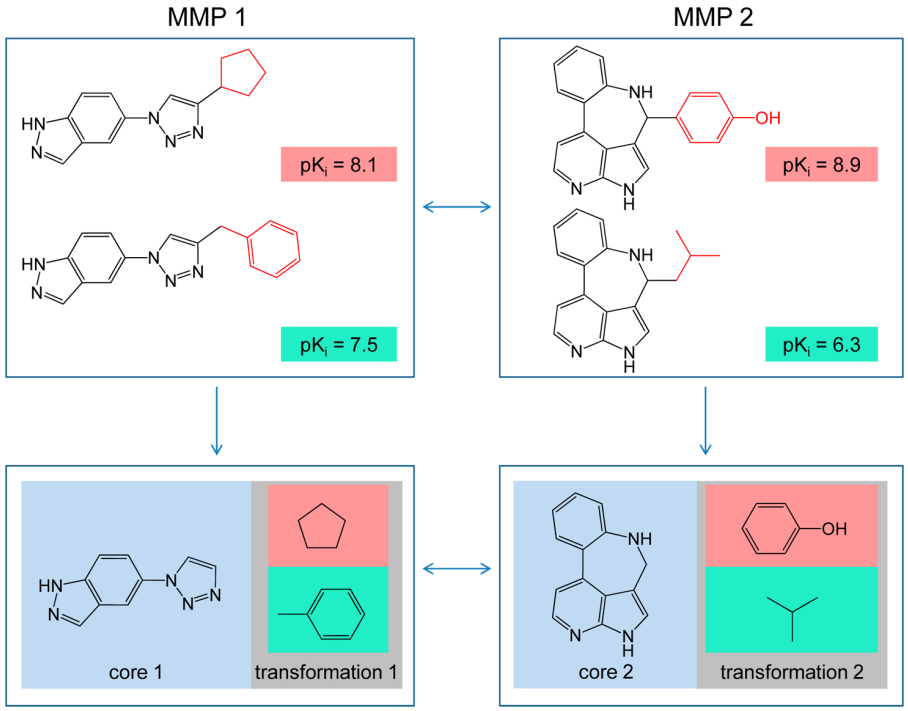


Figure 1. MMP comparison. Two MMPs are compared. On the basis of compound potency differences, MMP 1 is an MMP-nonCliff, whereas MMP 2 represents an MMP-cliff. In the upper panels, transformation substructures are shown in red. In the lower panels, the MMPs are divided into the common core (blue background) and the molecular transformation (gray background). Substructures originating from the compounds with higher and lower potency are highlighted (red and green background, respectively).

concept that compounds must be similar; i.e., structural differences must be limited.

MMPs were derived using an in-house Java implementation of the Hussain and Rea algorithm.[13] MMP generation was restricted to molecular transformations of terminal groups; i.e., only single bond cuts were considered. Furthermore, the maximal size of exchanged substructures was restricted to 13 heavy atoms, and the maximal size difference was limited to eight heavy atoms.[12] Furthermore, we concentrated on the smallest of all possible transformations to define a given MMP. Consequently, MMP core structures consisted of coherent fragments, for which other molecular representations could be calculated, and typically small substituents. For model building, as described below, MMPs were either represented as pairs of complete compounds or, alternatively, only by the transformations defining them.

In the following, we use the term 'substructures' to refer to fragments exchanged during a transformation and 'core structure' to refer to the common core of MMPs.

2.2. Compound Data Sets. Nine compound data sets were extracted from BindingDB.[14,15] The data sets were selected because they yielded large numbers of MMP-cliffs that were exclusively formed by compounds with at least 10 μM potency on the basis of K_i measurements. If several K_i values were available for a compound, the geometric mean was calculated as the final potency annotation. For fingerprint calculations, only compounds in which all atoms were assigned to Sybyl atom types were considered.[16] These atom types were used to enable calculations with a combinatorial feature fingerprint, as described below. Additionally, an MMP was omitted from the calculations if a chosen molecular representation (see below) did not unambiguously specify the underlying transformation.

For each data set, the resulting MMPs were divided into MMPs forming activity cliffs (MMP-cliffs), MMP-nonCliffs, and other MMPs based on the following potency difference criteria: To qualify as an MMP-cliff, compounds forming the pair were required to have a potency difference of at least 2 orders of magnitude. To control the potential influence of potency boundary effects on activity cliff prediction, the potency difference of compounds forming an MMP-nonCliff was limited to at most 1 order of magnitude. Accordingly, MMPs with compounds having a potency difference between 1 and 2 orders of magnitude were not further considered for SVM modeling and were assigned to the class of 'other MMPs'.

The partition of an MMP into its common core and transformation is illustrated in Figure 1 for two exemplary MMPs forming an MMP-cliff and MMP-nonCliff, respectively. Compound sets and MMP statistics are reported in Table 1. The top five data sets in Table 1 contained the largest number of MMP-cliffs. These data sets were relatively unbalanced because the ratio of MMP-nonCliffs to MMP-cliffs varied between 6 and 21. Because data sets of unbalanced composition typically present a difficult scenario of SVM modeling,[17,18] we also selected four more balanced data sets (ranks six to nine in Table 1). In these cases, the MMP-nonCliff/MMP-cliff ratio was less than 4.

3. SUPPORT VECTOR MACHINE MODELING

3.1. Motivation and Strategy. SVMs[10] are supervised machine learning algorithms for binary object classification and ranking. The prediction of activity cliffs requires the comparison of compound pairs instead of individual compounds, which presents an off-the-beaten path scenario for machine learning and classification methods. For SVM

dx.doi.org/10.1021/ci300306a | J. Chem. Inf. Model. 2012, 52, 2354−2365

Table 1. Data Sets[a]

target name	target code	no. of cpds	no. of MMPs	no. of MMP-cliffs	no. of MMP-nonCliffs	no. of other MMPs
factor Xa	fxa	2202	14493	1161	10108	3224
melanocortin receptor 4	mcr4	1159	13053	449	9618	2986
kappa opioid receptor	kor	1645	10104	649	7190	2265
thrombin	thr	2037	9585	1103	6390	2092
adenosine a3 receptor	aa3	1862	9575	681	6752	2142
calpain 2	cal2	121	1206	387	718	101
cathepsin b	catb	150	681	120	451	110
dipeptidyl peptidase 8	dpp8	44	602	141	421	40
janus kinase 2	jak2	58	366	109	186	71

[a]For each of the nine compound sets, the target name, a target code (abbreviation), the number of compounds (cpds), and the number of MMPs are reported. MMPs are divided into the number of compound pairs forming activity cliffs (MMP-cliffs), no activity cliffs (MMP-nonCliffs), and other MMPs, following the potency difference-based definition detailed in the Methods section. The data sets are sorted by decreasing numbers of MMPs.

modeling, kernel functions can be designed to account for specific relationships between objects and facilitate classification on the basis of these relationships. Our focus on the SVM approach for activity cliff prediction was largely motivated by the design of new kernel functions to facilitate comparisons of compound pairs, as illustrated in Figure 2. Our approach to facilitate compound pair-based predictions included, as a basis, the generation of training sets of MMP-cliffs and MMP-nonCliffs. In addition, an integral part of our approach was to attempt a systematic analysis of structural differences between compounds in cliff and noncliff pairs. The underlying hypothesis was that there should be structural features among compounds sharing a specific activity that are responsible for high and low potency and thus, ultimately, for the formation of activity cliffs. Although this hypothesis was intuitive, its potential utility for activity cliff prediction remained to be evaluated. Methodologically, this was not a trivial task because it required, first, relating features of compounds forming pairs to each other and, second, comparing feature differences across pairs.

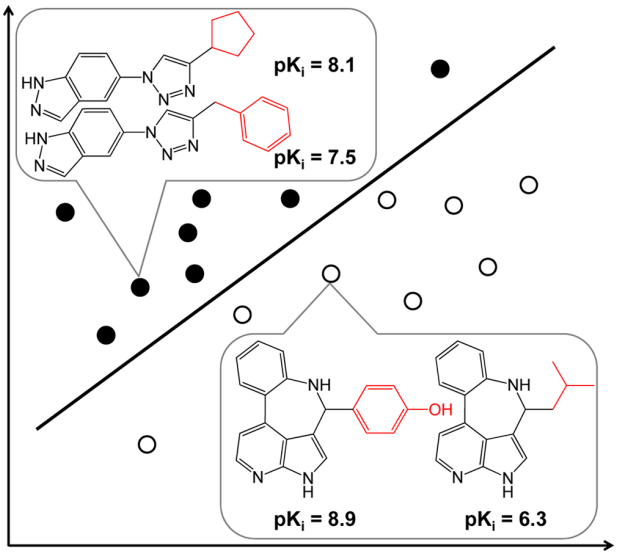


Figure 2. Activity cliff prediction using SVMs. The schematic figure illustrates the principal idea of SVM-based activity cliff prediction. In this case, the basic classification unit is a compound pair, different from standard compound classification tasks. Compound pairs forming MMP-cliffs (nonfilled circles) and MMP-nonCliffs (black circles) are separated by a hyperplane. In molecular graphs, transformation substructures are colored red. For each compound, its pK_i value is reported.

3.2. SVM Theory in Brief. SVMs make use of labeled training data that are mapped into a feature space to build a linear classification model. A set of n training objects $\{\mathbf{x}_i, y_i\}$ (i = 1, ..., n) are represented by a feature vector $\mathbf{x}_i \in \chi$ (e.g., $\mathbb{R}^d$) and an associated class label $y_i \in \{-1, 1\}$ corresponding to the 'negative' and 'positive' classes, respectively. By solving a convex quadratic optimization problem, a hyperplane H is derived that best separates positive from negative training data (Figure 2). During training, the cost parameter C penalizes the misclassification of training data and achieves a balance between minimizing the training error and maximizing the generalization of the classification.

The hyperplane H is defined by the normal weight vector $\mathbf{w}$ and the bias b, so that $H = \{\mathbf{x}|\langle \mathbf{w},\mathbf{x}\rangle + b = 0\}$, where $\langle\cdot,\cdot\rangle$ is a scalar product. Test data are mapped into the same feature space χ and classified by the linear decision function $f(\mathbf{x}) = \mathrm{sgn}(\langle \mathbf{x}, \mathbf{w}\rangle + b)$, i.e., depending on which side of the hyperplane they fall. In our calculations, the positive class consisted of the MMP-cliffs and the negative class of MMP-nonCliffs.

If the training data are not linearly separable in the feature space χ, the so-called *Kernel trick*[19] can be applied to replace the scalar product $\langle\cdot,\cdot\rangle$ by a kernel function $K(\cdot,\cdot)$. Kernel functions are used to calculate the scalar product of two feature vectors in a higher dimensional space $\mathcal{H}$ without explicitly calculating the mapping $\Phi:\chi \rightarrow \mathcal{H}$. In the higher dimensional space $\mathcal{H}$, a linear separation of the training data might be feasible. Kernel functions are of the form $K(\mathbf{u},\mathbf{v}) = \langle\Phi(\mathbf{u}), \Phi(\mathbf{v})\rangle$, where $\mathbf{u}$ and $\mathbf{v}$ are feature vector representations.

3.3. Standard Kernel Functions. The following four popular kernels are often used in SVM calculations:

$$K_{\mathrm{linear}}(\mathbf{u}, \mathbf{v}) = \langle \mathbf{u}, \mathbf{v}\rangle$$

$$K_{\mathrm{Gaussian}}(\mathbf{u}, \mathbf{v}) = \exp(-\gamma |\mathbf{u} - \mathbf{v}|^2)$$

$$K_{\mathrm{polynomial}}(\mathbf{u}, \mathbf{v}) = (\langle \mathbf{u}, \mathbf{v}\rangle + 1)^d$$

$$K_{\mathrm{Tanimoto}}(\mathbf{u}, \mathbf{v}) = \frac{\langle \mathbf{u}, \mathbf{v}\rangle}{\langle \mathbf{u}, \mathbf{u}\rangle + \langle \mathbf{v}, \mathbf{v}\rangle - \langle \mathbf{u}, \mathbf{v}\rangle}$$

The linear kernel corresponds to the standard scalar product. The Gaussian kernel is also known as the radial basis function kernel and depends on an adjustable parameter γ. In the polynomial kernel, the parameter d determines the degree of the polynomial function. The Tanimoto kernel[20] was introduced given the popularity of the Tanimoto coefficient for quantifying compound similarity. On the basis of these kernels, new kernel functions were designed specifically for activity cliff prediction, as described in the following.

dx.doi.org/10.1021/ci300306a | J. Chem. Inf. Model. 2012, 52, 2354−2365

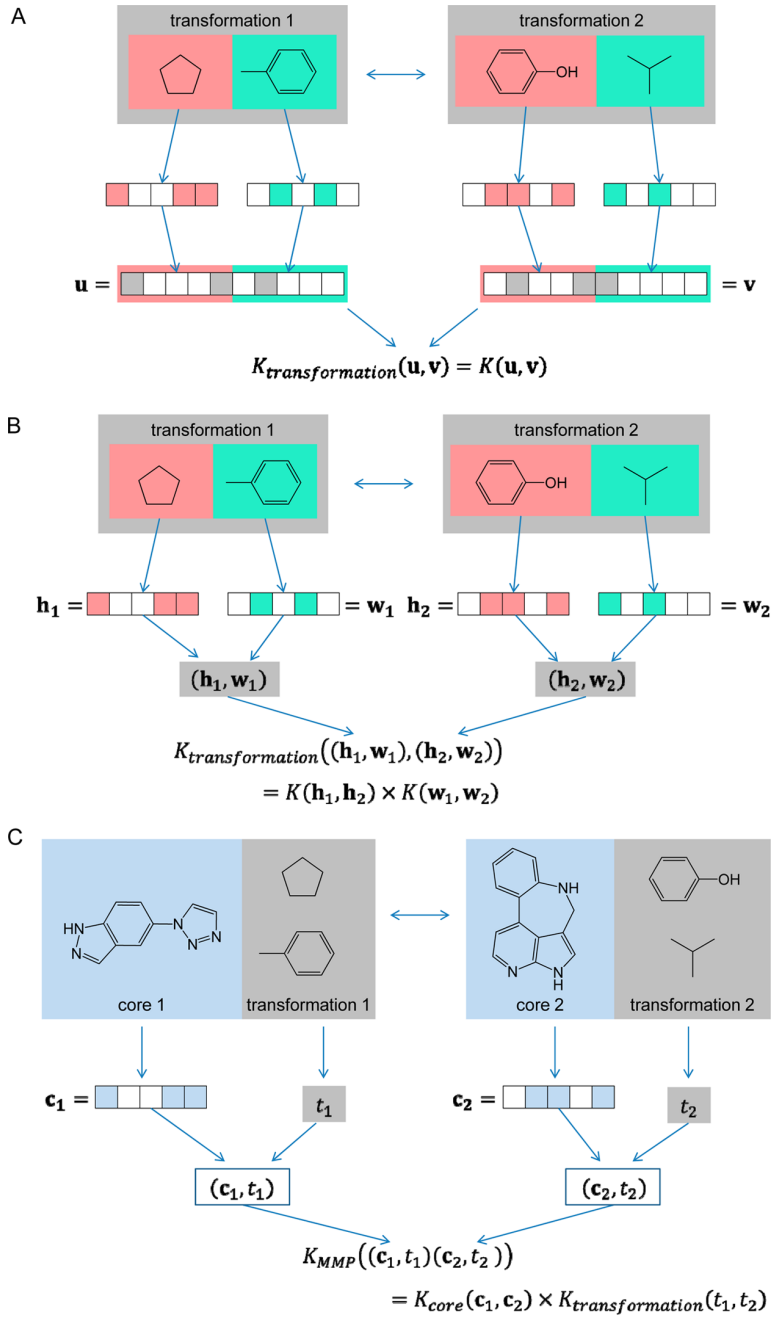


Figure 3. Kernel functions. The design of new kernel functions for SVM-based activity cliff prediction is illustrated. (A) Substructure-difference kernel. A fingerprint representation is generated for each substructure representing a given transformation. Then, a difference vector is calculated for the two substructure fingerprints. In kernel calculations, difference vectors for different transformations are compared. (B) Substructure-pair kernel. Fingerprint representations of substructures representing a transformation are combined to yield substructure pairs. Kernel calculations then compare the substructure pairs of different transformations. (C) MMP kernel. A fingerprint representation is calculated for the common core of each MMP. The corresponding transformations are represented by a transformation object that is either the substructure-difference vector or the substructure-pair representation (according to A and B, respectively). The core structure vector and the transformation object are then combined for kernel calculations.

4. DESIGN OF KERNEL FUNCTIONS FOR ACTIVITY CLIFF PREDICTION

4.1. Substructure-Difference Kernel. In order to use MMPs for SVM calculations, a feature vector representation of MMPs must be generated. As discussed above, MMPs consist of a common core structure and two differentiating substructures (substituents) that constitute the molecular transformation. We first designed a kernel that utilized only the transformation to create a single feature vector. The substituents were classified according to the highly potent partner in the MMP, termed 'highly potent substructure', and the weakly potent MMP compound, referred to as 'weakly potent substructure' (for MMP-nonCliffs, these potency

dx.doi.org/10.1021/ci300306a | J. Chem. Inf. Model. 2012, 52, 2354−2365

differences were within an order of magnitude). For both substructures, a keyed fingerprint of size n was calculated. Then, a difference fingerprint of size $2n$ was created that contained as the first n positions only those features present in the highly but not weakly potent substructure. If a feature was present in both substructures, the corresponding bit in the difference fingerprint was set off. The last n positions in the difference fingerprint contained features present only in the weakly but not the highly potent substructure. Accordingly, this difference vector uniquely described the transformation defined by the substructures on the basis of fingerprint features. The design of the difference fingerprint and substructure-difference kernel is illustrated in Figure 3A. Because this compound pair representation only comprises a single vector, kernel calculations can be performed as follows:

$$K_{\text{transformation}}(\mathbf{u}, \mathbf{v}) = K(\mathbf{u}, \mathbf{v})$$

where $\mathbf{u}$ and $\mathbf{v}$ are the substructure-difference vectors of two MMPs and $K(\mathbf{u},\mathbf{v})$ might, for example, be the Tanimoto or the polynomial kernel.

Because substructures were sorted on the basis of potency relationships between MMP compounds, this process is referred to as potency-based ordering. This information was taken into account during the learning and test phase. In addition, ordering of substructures by size (without considering potency relationship) was also investigated.

4.2. Substructure-Pair Kernel. Another new kernel represented a transformation as a pair of substructures. Again, substructures were classified according to the potency of the compounds from which they originate, and fingerprint representations were calculated. However, in this case, a transformation was represented as the pair $(\mathbf{h},\mathbf{w})$, where $\mathbf{h}$ is the fingerprint vector of the highly potent substructure and $\mathbf{w}$ the feature vector of the weakly potent substructure. Given two substructure pairs $(\mathbf{h}_i,\mathbf{w}_i)$ and $(\mathbf{h}_j,\mathbf{w}_j)$, a 'transformation kernel' was defined as the product of two separate kernels for the highly potent and weak potent substructures:

$$K_{\text{transformation}}((\mathbf{h}_i, \mathbf{w}_i), (\mathbf{h}_j, \mathbf{w}_j)) = K(\mathbf{h}_i, \mathbf{h}_j) \times K(\mathbf{w}_i, \mathbf{w}_j)$$

Thus, two independent kernels for highly potent and weakly potent substructures were combined to account for pairwise transformation similarities. The two kernels could again be implemented using standard kernel functions. The design of the substructure-pair kernel is illustrated in Figure 3B.

4.3. MMP Kernel. So far, we only considered molecular transformations to represent structural changes in MMPs that potentially lead to the formation of activity cliffs. However, the common core of an MMP might add further information for the classification of MMP-cliffs and MMP-nonCliffs because it defines the structural environment of a transformation. A potential caveat associated with considering the common core was that a given core structure might appear in both the positive and the negative class. This might be the case if a compound formed an MMP-cliff and an MMP-nonCliff with different partners. Hence, it was difficult to predict how the inclusion of the core might influence the classification calculations.

In order to generate a kernel that contains core information, an MMP was represented by combining the common core and the transformation, i.e., $(\mathbf{c},t)$, where $\mathbf{c}$ is the feature vector representation from the common core and t is a transformation object that can either be described by the substructure-difference vector or the substructure pair. Thus, the MMP kernel is defined by

$$K_{\text{MMP}}((\mathbf{c}_i, t_i)(\mathbf{c}_j, t_j)) = K_{\text{core}}(\mathbf{c}_i, \mathbf{c}_j) \times K_{\text{transformation}}(t_i, t_j)$$

The kernel function for pairs is again separated into independent kernels for each data type. The design of the MMP kernel is illustrated in Figure 3C. Because the common core was represented by a single feature vector, standard kernel functions could replace the core kernel (see above).

5. CALCULATION SETUP

5.1. Cost Factor. All SVM calculations were carried out using SVM$^{\text{light}}$,[21] a freely available SVM implementation. With two exceptions, suggested default parameters of SVM$^{\text{light}}$ were used to render the calculations reproducible. Apart from adjustable parameters in kernel functions, as specified above, we only modified the cost factor for the treatment of unbalanced data sets. SVM calculations on significantly unbalanced data sets often result in the generation of a hyperplane that is proximal to under-represented training examples,[17,18] here the positive examples (MMP-cliffs). As a consequence, positive instances are often predicted at only low rates. The cost factor defines the ratio of training error costs on the positive class (C^+) to penalties on the negative class (C^-):[22]

$$\text{cost factor} = \frac{C^+}{C^-}$$

The default value of the cost factor is 1; i.e., the same penalty is applied to positive and negative examples. However, increasing the error cost C^+, i.e., the penalty to predict a false-negative, repositions the hyperplane farther away from the positive examples. An often recommended cost-factor adjustment[17,21] can be expressed as

$$\text{cost factor} = \frac{\text{NTE}}{\text{PTE}}$$

where NTE and PTE are the number of negative and positive training examples, respectively. Thus, the potential total cost of false negative errors and the potential total cost of false positive errors are the same.[22]

5.2. Statistics. We performed 10-fold cross-validation as a reasonable compromise between data perturbation and training data size.[23] The MMP-cliff and MMP-nonCliff classes were randomly partitioned into 10 samples such that the global ratio between positive and negative training examples was constant. Nine of 10 samples were utilized for SVM learning and model building including all positive and negative training examples, and the remaining sample was used as a test set for prediction. In systematic classification calculations, each sample was used once as a test set. Average statistics were calculated over all 10 trials and used for performance evaluation. The following statistics were calculated:

$$\text{accuracy} = \text{AC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}$$

$$\text{recall} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{specificity} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Table 2. Cost-Factor Settings[a]

	cost factor = 1					cost factor = NTE/PTE				
target	AC	TPR	TNR	P	F score	AC	TPR	TNR	P	F score
fxa	92.47	32.38	99.38	85.82	46.89	89.82	72.69	91.79	50.45	59.51
mcr4	97.01	35.62	99.88	94.17	51.34	95.90	78.15	96.72	53.02	63.01
kor	92.69	13.56	99.83	86.75	23.08	87.09	66.87	88.92	35.44	46.26
thr	91.69	52.07	98.53	85.93	64.64	88.90	81.15	90.23	58.99	68.29
aa3	92.47	25.84	99.19	76.81	38.51	86.95	71.95	88.46	38.63	50.19
cal2	95.65	95.11	95.95	92.83	93.86	96.10	97.44	95.40	92.14	94.65
catb	95.80	81.67	99.56	98.57	88.67	95.62	88.33	97.56	91.10	89.39
dpp8	99.82	99.29	100.00	100.00	99.63	99.82	99.29	100.00	100.00	99.63
jak2	90.33	88.18	91.58	86.24	87.01	90.00	92.73	88.42	82.82	87.30

[a]For each data set, the average accuracy (AC), true positive rate (TPR), true negative rate (TNR), precision (P), and F score are reported (in %) for SVM calculations with different cost-factor settings. In these calculations, transformations were represented using the substructure-difference vector generated with MACCS, and the Tanimoto kernel was used as part of the substructure-difference kernel.

$$\text{precision} = \text{P} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where TP, FN, TN, and FP define the number of predicted true positives, false negatives, true negatives, and false positives, respectively. TPR and TNR denote the true positive rate and true negative rate, respectively, and are used in the following to account for recall and specificity. Because so assessed prediction accuracy is not a very informative measure when the number of negative examples is much larger than the number of positive examples,[24] we also calculated the (balanced) F score that accounts for both precision and recall (and ranges from 0% to 100%):[24]

$$\text{F score} = 2 \times \frac{\text{P} \times \text{TPR}}{\text{P} + \text{TPR}}$$

5.3. Fingerprints. For representing subtructures, two fragment-type fingerprints were used. The bonded-atom pair fingerprint (BAP)[25] encodes 117 different atom pairs with a focus on short-range connectivity information. In order to account for substructures comprising single atoms, we added three features describing carbon atoms, heteroatoms, and hydrogens resulting in a final fingerprint consisting of 120 structural descriptors. In addition, the MACCS[26] fingerprint was used that consists of 166 structural keys encoding substructures with one to 10 non-hydrogen atoms. An additional feature for a single hydrogen atom was also added in this case (because it might participate in a transformation). Furthermore, we evaluated the combination of both fingerprints (MACCS+BAP), resulting in a descriptor with 284 (117 + 166 + 1) features (two features added to the BAP fingerprint correspond to MACCS structural keys).

As a molecular representation of the common core, we used MACCS and Molprint2D.[27] The Molprint2D fingerprint requires the use of Sybyl atom types and encodes circular atom environments by fusing each atom in the structure with its neighboring atoms until a specific bond radius is reached. Here, we used features with a maximal bond radius of 2.

6. INITIAL TRIALS, COST-FACTOR ADJUSTMENT, AND SUBSTRUCTURE REPRESENTATION

6.1. Basic Classification Performance. To evaluate the potential of our SVM-based approach, we first determined the classification performance in calculations in which substructures were represented using the substructure-difference vector generated with MACCS, and the Tanimoto kernel was used as part of the substructure-difference kernel. In addition, a constant cost factor of 1 was applied. The results of these calculations are reported in Table 2. To present comprehensive statistics for performance evaluation, we report for all cross-validated calculations the average accuracy (AC), true positive (TPR) and true negative (TNR) rates, the precision (P), and the F score. In the following discussion, most emphasis is put on TPR, P, and F score values. The results in Table 2 for a cost factor of 1 mirror overall successful activity cliff predictions, with notable compound class dependence. In particular, the (un)balance of positive and negative training examples affected the calculations. For the first five data sets in Table 2, which contained many more MMP-nonCliffs than MMP-cliffs, prediction accuracy was lower than for the remaining more balanced sets (i.e., cal2, catb, dpp8, and jak2), as to be expected (see above). MMP-nonCliffs were generally predicted with very high accuracy, leading to TNRs of nearly 100% in all but one (jak2; 91.6%) case. This also led to an overall accuracy of 90−100% of the calculations and to a precision of 76−100%. Significant differences were observed between the rates with which MMP-cliffs were correctly predicted. Here, TPRs ranged from 13.6% to 99.3%, leading to F scores between 23.1% and 99.6%. For the unbalanced data sets, TPRs and F scores ranged from 13.6% to 52.1% and 23.1% to 64.6%, respectively. By contrast, for the balanced data sets, TPRs and F scores of 81.7−99.3% and 87.0−99.6% were observed, respectively. Thus, the results of initial activity cliff predictions were considered encouraging, at least for balanced data sets, and we thus further refined the approach, as discussed in the following.

6.2. Cost Factor. We first attempted to address the low TPRs and resulting F scores observed for unbalanced data sets in our initial calculations. Therefore, the default cost factor of 1 was replaced by the adjusted cost factor = NTE/PTE, which introduced a higher penalty on misclassification of positive training instances, i.e., activity cliffs. We repeated cross-validated SVM calculations under these conditions and observed a significant increase in TPRs for unbalanced data sets, as reported in Table 2. For balanced data sets, classification performance remained essentially unchanged, but for unbalanced sets, TPRs and F scores further increased to 66.9−81.2% and 46.3−68.3%, respectively. A trade-off has been a reduction in precision because the TNRs were reduced from on average 99.4% to 91.2%, due to the adjusted cost factor. However, this relatively small reduction in TNRs was clearly overcompensated for by an average TPR increase of 42.3% for unbalanced sets, yielding reasonably to highly accurate activity cliff

dx.doi.org/10.1021/ci300306a | J. Chem. Inf. Model. 2012, 52, 2354−2365

Table 3. Comparison of Fingerprints for Substructure Representation[a]

	BAP					MACCS					MACCS+BAP				
target	AC	TPR	TNR	*P*	F score	AC	TPR	TNR	*P*	F score	AC	TPR	TNR	*P*	F score
fxa	83.72	72.18	85.04	35.66	47.73	89.82	72.69	91.79	50.45	59.51	90.11	73.21	92.06	51.50	60.39
mcr4	89.58	77.28	90.15	26.90	39.85	95.90	78.15	96.72	53.02	63.01	95.91	77.25	96.78	53.26	62.84
kor	79.36	65.33	80.63	23.37	34.39	87.09	66.87	88.92	35.44	46.26	87.31	67.95	89.05	36.19	47.15
thr	85.11	78.61	86.23	49.66	60.84	88.90	81.15	90.23	58.99	68.29	89.15	81.43	90.49	59.71	68.85
aa3	81.21	68.15	82.52	28.18	39.83	86.95	71.95	88.46	38.63	50.19	86.98	71.37	88.55	38.63	50.04
cal2	92.67	94.87	91.50	86.08	90.16	96.10	97.44	95.40	92.14	94.65	96.11	97.44	95.40	92.23	94.69
catb	93.17	90.83	93.79	80.24	84.91	95.62	88.33	97.56	91.10	89.39	96.15	88.33	98.22	93.38	90.47
dpp8	99.82	99.29	100.00	100.00	99.63	99.82	99.29	100.00	100.00	99.63	99.82	99.29	100.00	100.00	99.63
jak2	82.20	80.82	83.02	74.53	76.97	90.00	92.73	88.42	82.82	87.30	89.67	90.00	89.47	83.74	86.56

[a]The performance of the BAP, MACCS, and MACCS+BAP fingerprints for substructure representation is compared. Calculation statistics are reported according to Table 2. The transformations were represented using the substructure-difference vector. The Tanimoto kernel was used as part of the substructure-difference kernel, and the adjusted cost factor was applied.

Table 4. Comparison of Standard Kernels[a]

	Tanimoto					linear				
target	AC	TPR	TNR	P	F score	AC	TPR	TNR	P	F score
fxa	89.82	72.69	91.79	50.45	59.51	80.70	71.66	81.74	31.14	43.38
mcr4	95.90	78.15	96.72	53.02	63.01	87.30	80.82	87.61	23.44	36.28
kor	87.09	66.87	88.92	35.44	46.26	79.93	67.49	81.06	24.39	35.82
thr	88.90	81.15	90.23	58.99	68.29	84.20	76.97	85.45	47.80	58.93
aa3	86.95	71.95	88.46	38.63	50.19	77.88	75.19	78.16	25.90	38.50
cal2	96.10	97.44	95.40	92.14	94.65	95.30	96.41	94.71	91.06	93.55
catb	95.62	88.33	97.56	91.10	89.39	94.92	85.83	97.33	90.46	87.63
dpp8	99.82	99.29	100.00	100.00	99.63	99.82	99.29	100.00	100.00	99.63
jak2	90.00	92.73	88.42	82.82	87.30	92.33	95.45	90.53	85.94	90.24
	Gaussian (γ = 1/numFeatures)					Gaussian (γ = 0.1)				
target	AC	TPR	TNR	P	F score	AC	TPR	TNR	P	F score
fxa	84.83	71.40	86.37	37.65	49.25	91.19	64.77	94.22	56.37	60.20
mcr4	91.76	82.16	92.20	33.10	47.11	96.26	70.80	97.44	56.89	62.86
kor	83.31	69.19	84.59	28.92	40.77	87.86	62.72	90.13	36.66	46.17
thr	86.73	78.87	88.09	53.37	63.63	89.82	78.71	91.74	62.32	69.51
aa3	80.84	77.10	81.22	29.46	42.57	87.50	68.14	89.46	39.50	49.91
cal2	95.48	96.92	94.71	91.09	93.81	95.56	96.67	94.98	91.47	93.90
catb	94.92	85.83	97.33	90.46	87.63	95.98	85.00	98.89	95.92	89.48
dpp8	99.82	99.29	100.00	100.00	99.63	99.82	99.29	100.00	100.00	99.63
jak2	92.67	95.45	91.05	86.54	90.60	90.33	90.91	90.00	84.50	87.40
	polynomial (d = 2)					polynomial (d = 3)				
target	AC	TPR	TNR	P	F score	AC	TPR	TNR	P	F score
fxa	90.67	70.45	93.00	53.74	60.88	92.12	65.20	95.21	61.04	62.97
mcr4	96.03	75.02	97.01	54.04	62.69	97.02	72.13	98.18	65.41	68.39
kor	87.84	65.50	89.86	37.03	47.20	89.35	59.95	92.00	40.60	48.30
thr	90.35	78.97	92.32	64.15	70.74	91.03	74.89	93.82	67.73	71.11
aa3	87.45	70.93	89.11	39.70	50.84	88.69	66.07	90.97	42.44	51.58
cal2	96.11	95.38	96.52	93.81	94.53	95.66	91.52	97.90	96.02	93.55
catb	95.97	82.50	99.56	98.57	89.19	95.62	80.00	99.78	99.17	88.04
dpp8	99.82	99.29	100.00	100.00	99.63	99.82	99.29	100.00	100.00	99.63
jak2	90.27	88.09	91.58	86.46	86.77	88.93	82.64	92.63	87.07	84.42

[a]The performance of different kernels as part of the substructure-difference kernel is compared. Performance statistics are reported. The Gaussian kernel was used with two different γ values (γ = 1/numFeatures and γ = 0.1) and the polynomial kernel with two different exponents d (d = 2 and d = 3). The parameter numFeatures describes the number of features present in the substructure-difference vector. The substructures were represented using the substructure-difference vector with MACCS, and the adjusted cost factor was applied.

predictions for all nine different data sets (Table 2). Accordingly, the adjusted cost factor was used in all subsequent calculations.

6.3. Substructure Representation. Next, we compared different fingerprint representations of transformation substructures. Table 3 reports search results for the comparison of the BAP, MACCS, and (MACCS+BAP) fingerprints used for the generation of the substructure-difference vector. Calculations with the BAP substructure-difference vector resulted in consistently high TPRs but low precision for unbalanced data

dx.doi.org/10.1021/ci300306a | J. Chem. Inf. Model. 2012, 52, 2354−2365

Table 5. Comparison of Transformation Kernels[a]

	substructure-difference vector					substructure pairs				
target	AC	TPR	TNR	P	F score	AC	TPR	TNR	P	F score
fxa	89.82	72.69	91.79	50.45	59.51	91.01	74.85	92.87	54.68	63.17
mcr4	95.90	78.15	96.72	53.02	63.01	96.30	74.14	97.34	56.94	64.16
kor	87.09	66.87	88.92	35.44	46.26	88.00	60.55	90.47	36.73	45.61
thr	88.90	81.15	90.23	58.99	68.29	90.03	82.07	91.41	62.34	70.80
aa3	86.95	71.95	88.46	38.63	50.19	88.83	67.11	91.02	43.23	52.45
cal2	96.10	97.44	95.40	92.14	94.65	96.29	96.41	96.24	93.39	94.81
catb	95.62	88.33	97.56	91.10	89.39	95.44	86.67	97.78	91.63	88.85
dpp8	99.82	99.29	100.00	100.00	99.63	99.82	99.29	100.00	100.00	99.63
jak2	90.00	92.73	88.42	82.82	87.30	91.20	94.45	89.33	84.48	88.91

[a]The performance of the substructure-difference vector is compared to the substructure-pair representation. The substructures were encoded using MACCS. The Tanimoto kernel was used as part of the two transformation kernels, and the adjusted cost factor was applied.

Table 6. Comparison of Transformation and MMP Kernels[a]

	transformation kernel					MMP kernel (core structure: MACCS)					MMP kernel (core structure: Molprint2D)				
target	AC	TPR	TNR	P	F score	AC	TPR	TNR	P	F score	AC	TPR	TNR	P	F score
fxa	89.82	72.69	91.79	50.45	59.51	94.11	81.30	95.58	67.91	73.95	94.68	82.17	96.11	70.92	76.03
mcr4	95.90	78.15	96.72	53.02	63.01	98.02	81.04	98.81	76.50	78.57	98.35	83.05	99.06	80.82	81.82
kor	87.09	66.87	88.92	35.44	46.26	93.21	72.57	95.08	57.38	63.99	94.86	72.58	96.87	67.88	70.04
thr	88.90	81.15	90.23	58.99	68.29	93.07	84.41	94.57	72.93	78.21	93.75	84.05	95.43	76.20	79.85
aa3	86.95	71.95	88.46	38.63	50.19	93.52	74.74	95.41	62.60	67.91	95.12	74.45	97.20	73.23	73.57
cal2	96.10	97.44	95.40	92.14	94.65	97.55	97.69	97.49	95.54	96.57	97.64	97.69	97.63	95.79	96.70
catb	95.62	88.33	97.56	91.10	89.39	96.85	90.00	98.67	95.24	92.30	97.02	90.83	98.67	95.43	92.76
dpp8	99.82	99.29	100.00	100.00	99.63	99.82	99.29	100.00	100.00	99.63	99.82	99.29	100.00	100.00	99.63
jak2	90.00	92.73	88.42	82.82	87.30	90.67	91.82	90.00	84.74	87.89	91.00	91.82	90.53	85.55	88.28

[a]The performance of the transformation kernel is compared to the MMP kernel. The substructure-difference kernel was used as the transformation kernel. Substructures were represented using the substructure-difference vector with MACCS. The MACCS and Molprint2D fingerprints were compared as core structure representations in the MMP kernel. The Tanimoto kernel was used as part of the substructure-difference kernel as well as the MMP kernel, and the adjusted cost factor was applied.

sets, due to a reduction in TNRs, leading to low F scores. MACCS-based calculations yielded comparably high TPRs but higher precision and F scores. No further increases in these rates and scores were observed when the (MACCS+BAP) combination was used. Consequntly, MACCS was used in subsequent calculations.

7. KERNEL COMPARISON

Kernel design for the treatment of compound pairs as a basic classification object has been a key aspect of our approach to activity cliff prediction. We first compared the performance of standard kernels that provided a basis for the generation of substructure and MMP kernels.

7.1. Standard kernels. Table 4 summarizes the results of SVM calculations using different kernel functions as a component of the substructure-difference kernel. The use of the Tanimoto kernel resulted in TPRs that were consistently above 66% for all targets. In these calculations, the precision was low for two data sets (kor and aa3). For unbalanced sets, F scores varied from 46.3% to 68.3%. By contrast, for balanced sets, F scores were consistently higher than 87%. The linear kernel essentially paralleled the results of the Tanimoto kernel for balanced data sets but displayed consistently lower precision for unbalanced sets. The use of the Gaussian kernel with small γ parameter values (0.0034−0.0065), depending on the number of features (γ = 1/numFeatures) used in the calculations, also resulted in comparable TPRs but lower precision for unbalanced sets. For a larger γ value of 0.1, increased precision was observed but TPRs were reduced, yielding F scores comparable to the Tanimoto kernel. Furthermore, the polynomial kernel (with d = 2 and d = 3) also produced rates and scores that were similar to those obtained for the Tanimoto kernel. Thus, taken together, differences in prediction performance for different standard kernels were by and large insignificant. Since the Tanimoto kernel was parameter-free, it was selected for further calculations.

7.2. Transformation and MMP Kernels. An interesting initial finding was that promising classification results were obtained using the Tanimoto substructure-difference kernel (see section 6.1). This kernel only accounted for differences between transformation substructures, rather than entire MMPs. We then compared the substructure-difference and substructure-pair kernels (based on the Tanimoto kernel). Calculation requirements for these kernels differed. The substructure-difference kernel only required one kernel calculation, but the difference vector must be precalculated. By contrast, for the substructure-pair kernel, no precalculations were required, but the kernel calculation must be carried out for two functions. The results of search calculations using these alternative transformation-only kernels are reported in Table 5. No clear preference for one or the other kernel was detectable. Overall, the substructure-pair kernel produced slightly lower TPRs but slightly higher precision than the substructure-difference kernel (except for the dpp8 set yielding P = 100% in both instances), which resulted in similar F scores.

We then included the MMP kernel in the comparison, which was designed to combine the core structure representation of an MMP with its substructure-difference vector. For this

dx.doi.org/10.1021/ci300306a | J. Chem. Inf. Model. 2012, 52, 2354−2365

Table 7. Comparison of Potency- and Size-Based Substructure Ordering[a]

	potency-based ordering					size-based ordering				
target	AC	TPR	TNR	P	F score	AC	TPR	TNR	P	F score
fxa	89.82	72.69	91.79	50.45	59.51	87.89	68.38	90.13	44.30	53.73
mcr4	95.90	78.15	96.72	53.02	63.01	95.21	70.15	96.38	47.80	56.71
kor	87.09	66.87	88.92	35.44	46.26	84.17	62.10	86.16	28.81	39.33
thr	88.90	81.15	90.23	58.99	68.29	87.32	77.81	88.97	54.95	64.36
aa3	86.95	71.95	88.46	38.63	50.19	83.79	68.13	85.37	32.06	43.50
cal2	96.10	97.44	95.40	92.14	94.65	91.94	93.03	91.36	85.38	88.89
catb	95.62	88.33	97.56	91.10	89.39	93.17	85.00	95.33	84.27	83.77
dpp8	99.82	99.29	100.00	100.00	99.63	99.82	99.29	100.00	100.00	99.63
jak2	90.00	92.73	88.42	82.82	87.30	86.27	87.18	85.79	79.10	82.44

[a]The performance of potency-based ordering of transformation substructures is compared to size-based ordering. The transformations were represented using the substructure-difference vector. The Tanimoto kernel was used as part of the substructure-difference kernel, and the adjusted cost factor was applied.

purpose, the core structure can be represented using different fingerprints. For substructure representations, fragment fingerprints such as BAP or MACCS are in principle a preferred choice, but for core structure representation, other types of fingerprints might also be used. In Table 6, search results for the substructure-difference kernel are compared to those obtained for two versions of the MMP kernel including one in which the common core was represented using MACCS and another that utilized Molprint2D instead (i.e., a topological atom environment fingerprint). We found that application of the MMP kernel further improved classification performance. For both versions of the MMP kernel, an increase in TPRs and precision was observed compared to the transformation kernel, leading to higher F scores. For the MMP kernel, TPRs were very similar for MACCS and Molprint2D, but F scores were slightly higher for Molprint2D, due to a minor increase in TNRs. On average, F scores were 82.1% for the MACCS- and 84.3% for the Molprint2D-based MMP kernel. Compared to the substructure-difference kernel, which yielded TPRs and F scores of 66.9−99.3% and 46.3−99.6%, respectively, the Molprint2D-based MMP kernel produced TPRs and F scores of 72.6−99.3% and 70.0−99.6%, respectively. Improvements were observed for balanced and unbalanced data sets but were of larger magnitude for the latter. On average, TPRs slightly increased from 83.2% (transformation kernel) to 86.2% (MMP kernel) and F scores (reflecting both recall and precision) from 73.1% to 84.3%. Thus, the incorporation of core structure contributions of MMP-cliffs and MMP-nonCliffs into the kernel function further increased the accuracy of activity cliff predictions.

We also investigated the influence of substructure ordering on the calculations. In the classification scheme underlying our analysis, the ordering of transformation substructures in MMPs was potency-based. As a control, we also evaluated size-based ordering of substructures. The results for the substructure difference are presented in Table 7. With the exception of one set (dpp8; with consistently 100% precision), both TPRs and F scores decreased for size-based ordering. Comparable trends were observed when the MMP kernel was used (data not shown). Hence, potency-based ordering of substructures was generally preferred, but size-based ordering also yielded accurate predictions.

8. STRUCTURAL PATTERNS

Given the results of our calculations, we also investigated whether successful predictions of MMP-cliffs might be rationalized in structural terms. Therefore, we analyzed correctly identified MMP-cliffs and MMP-nonCliffs for the presence of characteristic transformations and structural features. In a number of instances, structural patterns were identified that could be attributed to activity cliff formation. In the following, representative examples are discussed.

Figure 4A shows a number of transformations leading to the formation of MMP-cliffs in the dpp8 set. Most MMP-cliffs were characterized by transformations in which a substructure containing a carbonyl group was replaced by a substituted phenyl group. The carbonyl group was predominantly involved in the formation of amide bonds, but there were also ketone and ether linkages proximal to the carbonyl group. With one exception, all of these MMP-cliffs were correctly classified. The only exception was a structurally very different transformation observed in an MMP-cliff (shown on a gray background), which might present an interesting test case for further analysis of activity cliffs among dpp8 inhibitors. Apart from this exception, the typical MMP-cliff transformation patterns observed in the dpp8 set resulted in perfect classifications, independent of chosen kernel functions and SVM calculations settings.

Figure 4B shows transformations of MMP-cliffs from the cal2 set that were correctly identified or misclassified as MMP-nonCliffs (gray background). The weakly potent substructures of correctly classified MMP-cliffs were linear alkyl chains, oxygen containing (alkyl) substituents, or groups containing phenyl rings. Replacement of these substructures with nitrogen- or oxygen-containing substructures or substituted ring systems caused a strong increase in potency, leading to the formation of activity cliffs. The transformation of misclassified pairs exclusively consisted of small nitrogen- and oxygen-containing substructures. These examples illustrate that clearly defined structural signatures of activity cliffs were not always obvious in the cal2 set, making this case a difficult classification problem. Despite this structural variability, 64.2% of all MMP-cliffs in the cal2 set displayed similar structural patterns and were correctly classified. Figure 4C shows examples of MMP-nonCliff transformations, which further illustrate the presence of complex transformation−potency relationships. In these cases, more potent compounds in pairs contained substructures that were found in weakly potent MMP-cliff compounds shown in Figure 4B. Consequently, these pairs were correctly classified as MMP-nonCliffs. Nevertheless, calculations on the cal2 set yielded accurate predictions using our SVM models, with a TPR and F score of 97.4% and 94.7%.

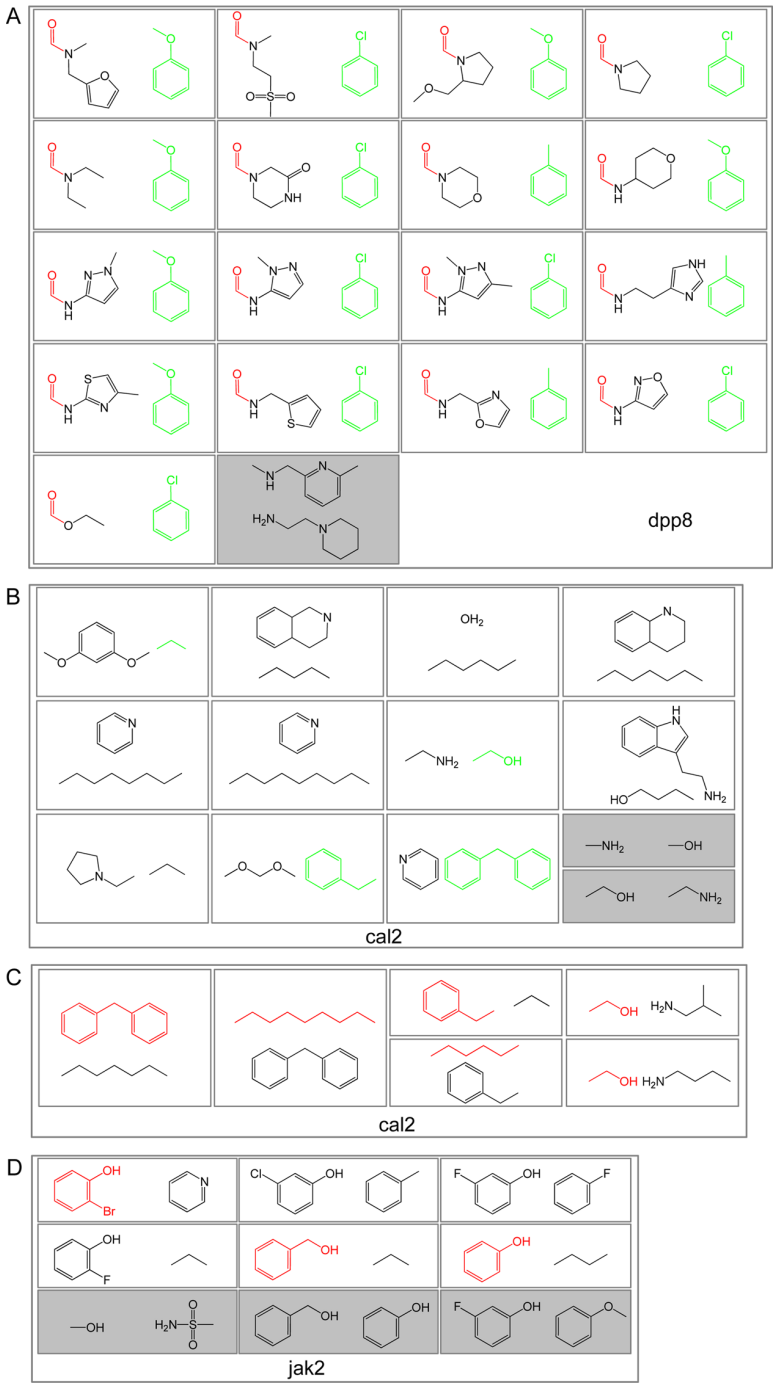


Figure 4. Exemplary MMP transformations. Exemplary MMP-cliff and MMP-nonCliff transformations are shown for different compound data sets (according to Table 1). Highly potent substructures are positioned on the left or at the top of cells and weakly potent ones on the right or at the bottom. Correctly classified MMP-cliff transformations are shown on a white background and misclassified transformations on a gray background. Selected structural features/patterns are colored red (in highly potent substructures) and green (in weakly potent substructures). (A) Data set dpp8/MMP-cliffs, (B) cal2/MMP-cliffs, (C) cal2/MMP-nonCliffs, (D) jak2/MMP-cliffs.

Figure 4D shows MMP-cliff transformations from the jak2 set. Here, replacements of alkyl or phenyl groups with benzyl alcohol or halogen substituted benzyl alcohol groups often led to activity cliffs that were correctly detected. Examples of transformations in misclassified MMP-cliffs included the exchange of methoxy and sulfonamide groups and of pairs of different benzyl alcohol derivatives that departed from the prevalent substructure patterns among MMP-cliffs.

9. CONCLUDING REMARKS

Herein, we have presented a first approach to predict activity cliffs in compound data sets. From a conceptual and methodological point of view, the prediction of activity cliffs

dx.doi.org/10.1021/ci300306a | J. Chem. Inf. Model. 2012, 52, 2354−2365

represents a nontrivial task. The underlying assumption is that structural differences in pairs of similar compounds can be directly related to potency differences and then compared across pairs representing cliffs and noncliffs. We have approached the task of activity cliff prediction using SVM modeling because the SVM formalism provides the opportunity to design kernel functions specifically tailored towards this task. To represent activity cliffs and noncliffs, the concept of MMP-cliffs is applied that yields a structurally well-defined representation of activity cliffs on the basis of common core structures and distinguishing substructure transformations. Another general difficulty in activity cliff prediction is the assembly of compound data sets with a balanced composition of positive (cliffs) and negative (noncliff) training examples, which typically is an important prerequisite for effective machine learning. Because activity cliffs are relatively rare among bioactive compounds, data sets are generally unbalanced. We have systematically searched for compound data sets that contained MMP-cliffs at a relatively high frequency and determined all MMP-cliffs and MMP-nonCliffs in these sets. A total of nine data sets were obtained for our analysis that contained significant numbers of MMP-cliffs. However, all of these data sets contained many more MMP-nonCliffs than MMP-cliffs, as expected. For the purpose of our analysis, we considered data sets balanced if the MMP-nonCliff/MMP-cliff ratio was not larger than 4, which was the case for four of our sets. The remaining five sets were characterized by much larger ratios and hence considered unbalanced. However, using newly introduced kernel functions, activity cliffs were predicted with reasonable to high accuracy on the basis of SVM learning and classification. During learning, unbalanced training example distributions were effectively handled by adjusting the cost factor of the SVM calculations. We designed alternative kernel functions that only took transformation substructure differences or transformation and core structure features into account. Interestingly, overall accurate predictions were already obtained when transformation kernels were applied, but prediction accuracy was further improved through the use of MMP kernel functions that considered transformation and core differences. However, much structural information relevant for the formation of activity cliffs was often encoded by transformations, without a critical influence of their specific structural environment. In our analysis, best predictions were obtained when cross-validated SVM calculations with adjusted cost factors were carried out using the Tanimoto kernel-based MMP kernel with a MACCS substructure-difference vector (on the basis of potency-based ordering of substructures) and a Molprint2D representation of common MMP cores. Under these conditions, average true positive rates and F scores of 86.2% and 84.3%, respectively, were achieved in activity cliff predictions, with an average precision of 82.9% and accuracy of 95.8% of the calculations. In many instances, it was possible to rationalize successful predictions of activity cliffs on the basis of structural features of corresponding transformations. Taken together, given the results presented herein, we anticipate that the SVM-based approach to activity cliff prediction should be of considerable interest in the search for cliffs in large compound data sets.

AUTHOR INFORMATION

Corresponding Author

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

Author Contributions

[§]The contributions of these authors should be considered equal.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

X.H. is supported by the *China Scholarship Council*. The authors thank Martin Vogt for helpful discussions.

REFERENCES

(1) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.

(2) Maggiora, G. M. On Outliers and Activity Cliffs − Why QSAR often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

(3) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

(4) Guha, R.; Van Drie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.

(5) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: Consensus Activity Cliffs. *J. Chem. Inf. Model.* **2009**, *49*, 477−491.

(6) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure-Activity Landscapes. *Drug Discovery Today* **2009**, *14*, 698−705.

(7) Wassermann, A. M.; Dimova, D.; Bajorath, J. Comprehensive Analysis of Single- and Multi-Target Activity Cliffs Formed by Currently Available Bioactive Compounds. *Chem. Biol. Drug Des.* **2011**, *78*, 224−228.

(8) Vogt, M.; Huang, Y.; Bajorath, J. From Activity Cliffs to Activity Ridges: Informative Data Structures for SAR Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1848−1856.

(9) Namasivayam, V.; Bajorath, J. Searching for Coordinated Activity Cliffs Using Particle Swarm Optimization. *J. Chem. Inf. Model.* **2012**, *52*, 927−934.

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

(11) Kenny, P. W.; Sadowski, J. Structure Modification in Chemical Databases. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: Weinheim, Germany, 2004; pp 271−285.

(12) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.

(13) Hussain, J.; Rea, C. Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.

(14) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.

(15) BindingDB. http://www.bindingdb.org/ (accessed February 8, 2012).

(16) Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.

(17) Akbani, R.; Kwek, S.; Japkowicz, N. Applying Support Vector Machines to Imbalanced Datasets. In *Machine Learning: ECML 2004*; Boulicaut, J.-F., Esposito, F., Giannotti, F., Pedreschi, D., Eds.; Springer: Berlin/Heidelberg, 2004; pp 39−50.

(18) Tang, Y.; Zhang, Y.-Q.; Chawla, N. V.; Krasser, S. SVMs Modeling for Highly Imbalanced Classification. *IEEE Trans. Syst. Man. Cybern. B: Cybern.* **2009**, *39*, 281−288.

(19) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual*

Workshop on Computational Learning Theory; Pittsburgh, PA, 1992; ACM: New York, 1992; pp 144−152.
(20) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph Kernels for Chemical Informatics. *Neural Networks* **2005**, *18*, 1093−1110.
(21) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods − Support Vector Learning*; Schölkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT-Press: Cambridge, MA, 1999; pp 169−184.
(22) Morik, K.; Brockhausen, P.; Joachims, T. Combining Statistical Learning with a Knowledge-based Approach - A Case Study in Intensive Care Monitoring. In *Proceedings of the 16th International Conference on Machine Learning (ICML-99)*; Morgan Kaufmann: Burlington, MA, 1999.
(23) McLachlan, G. J.; Do, K.-A.; Ambroise, C. *Analyzing Microarray Gene Expression Data*; Wiley & Sons: Hoboken, NJ, 2004.
(24) Kubat, M.; Matwin, S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In *Proceedings of the Fourteenth International Conference on Machine Learning*, San Francisco, 1997; Morgan Kaufmann: Burlington, MA, 1997; pp 179−186.
(25) Ahmed, H. E. A.; Vogt, M.; Bajorath, J. Design and Evaluation of Bonded Atom Pair Descriptors. *J. Chem. Inf. Model.* **2010**, *50*, 487−499.
(26) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2002.
(27) Bender, A.; Mussa, H. Y.; Glen, R. C. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170−178.

Summary

SVM classification using kernel functions designed for compound pairs was applied to predict activity cliffs. The kernel functions compared pairs of compounds on the basis of matched molecular pairs and captured molecular transformations and common core structures. SVM calculations with these kernels showed high prediction accuracy when applied to compound data sets with different activity cliff content. Thereby, the use of transformations alone already resulted in accurate predictions indicating that essential information about the cliffs is encoded by the transformation substructures and that these structural patterns were recognized in other compound pairs in the data set. The results were further improved by considering both the transformation and the core structure through the use of the MMP kernel.

As a supervised machine learning method, SVM calculations are influenced by the underlying training data. In this study, SVM classification was affected by different activity cliff content of the data sets. Therefore, we asked the question how SVM-based virtual screening is influenced by the data set composition and size. In the following study, we analyze the compound recall of SVMs under alternative benchmark settings using different negative training examples, varying background databases, and differently sized training sets.

Chapter 6

Comparison of confirmed inactive and randomly selected compounds as negative training examples in support vector machine-based virtual screening

Introduction

The quality of an SVM model directly depends on the quality of the data set used for learning. This is especially important for virtual screening, where usually only active compounds are available and confirmed inactive compounds are rare. Therefore, we analyze how the selection of negative training data influences the screening performance. SVM modeling is performed using assumed inactive compounds randomly selected from the ZINC database and confirmed inactives from PubChem with varying sizes of the training data. The respective recall of active compounds is evaluated and compared. Additionally, the models are applied to different background databases.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

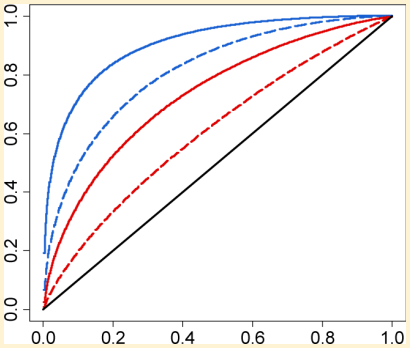
Comparison of Confirmed Inactive and Randomly Selected Compounds as Negative Training Examples in Support Vector Machine-Based Virtual Screening

Kathrin Heikamp 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

Supporting Information

ABSTRACT: The choice of negative training data for machine learning is a little explored issue in chemoinformatics. In this study, the influence of alternative sets of negative training data and different background databases on support vector machine (SVM) modeling and virtual screening has been investigated. Target-directed SVM models have been derived on the basis of differently composed training sets containing confirmed inactive molecules or randomly selected database compounds as negative training instances. These models were then applied to search background databases consisting of biological screening data or randomly assembled compounds for available hits. Negative training data were found to systematically influence compound recall in virtual screening. In addition, different background databases had a strong influence on the search results. Our findings also indicated that typical benchmark settings lead to an overestimation of SVM-based virtual screening performance compared to search conditions that are more relevant for practical applications.



INTRODUCTION

For ligand-based virtual screening, selected machine-learning methodologies are increasingly utilized, given their usually good performance in predicting active compounds, at least in benchmark settings.[1,2] Methods like Bayesian modeling[3−6] and support vector machines (SVMs)[7−11] currently are among the most widely used methodologies for supervised learning to predict candidate compounds for different targets and to rank them according to their proposed likelihood of activity. These machine-learning algorithms rely on already known active compounds to derive computational models for compound classification and activity prediction.

The availability of sufficient amounts of relevant training data is a prerequisite for the applicability of these approaches. If no, only one, or very few known active compounds are available for a new target, scientifically sound classification models cannot be derived. For orphan screening, compound information from related targets (if available) must be utilized. Hence, knowledge of active compounds is generally considered the most critical requirement for model building. Consequently, major databases such as ChEMBL[12] and BindingDB,[13] which store active compounds from medicinal chemistry together with their activity data, are prime sources of positive training examples for machine learning.

A general caveat for training set assembly is that confirmed inactive compounds are often not available for a given target. Therefore, it is common practice in Bayesian or SVM modeling to randomly select compounds from databases that are not annotated with biological activities as negative training examples and assume that these random selections are inactive against the target of interest.[2,14−16] Scientifically, this represents an approximation, which is only very little investigated in chemoinformatics machine-learning applications.

The potential influence of negative training data on the quality of machine-learning models is just beginning to be addressed in the chemoinformatics field. In a recent study, Smusz et al.[17] tested alternative selection methods for assumed inactive training compounds. Using different fingerprints, machine-learning algorithms, and targets, the authors compared random and diverse selection of negative training examples from the ZINC database,[18] the Molecular Drug Data Report (MDDR),[19] and from compound libraries that were designed following the principles underlying the Directory of Useful Decoys (DUD) approach.[20] In these benchmark calculations, overall best compound classification results were achieved with different methods when negative training compounds were randomly selected from ZINC.[17] This database represents the largest public collection of compounds from chemical vendor sources. ZINC compounds, which are typically not biologically annotated, currently are the most popular source for random compound selection in machine-learning and ligand-based virtual screening.

Compounds confirmed to be inactive against a virtual screening target should generally have highest priority as negative training data for machine learning. This is the case

Received: May 8, 2013
Published: June 25, 2013

dx.doi.org/10.1021/ci4002712 | J. Chem. Inf. Model. 2013, 53, 1595−1601

because experimentally confirmed inactive compounds no longer rely on the assumption of inactivity that needs to be made for randomly chosen database compounds, which might or might not be true for individual molecules. Although it is difficult to obtain confirmed inactive compounds for many targets, the problem can be addressed by taking biological screening data into account. For example, confirmatory screening assays follow up on compounds with an activity signal in a primary screen (usually at a single concentration), investigate dose−response behavior, determine IC_{50} values for active compounds, and identify false positives. Thus, such follow-up assays confirm the activity of initial hits and also yield confirmed inactive compounds. Alternatively, dose−response behavior and IC_{50} titration curves might also be determined in a primary screen using multiple compound concentrations. In this case, large numbers of inactive compounds can be directly obtained.

In this study, we have investigated the influence of different negative training data and background databases on compound recall in SVM-based virtual screening. By assembling data sets from the PubChem Confirmatory Bioassays,[21] we were able to compare the search performance for training data sets comprising experimentally confirmed active and inactive compounds and training data sets consisting of experimentally confirmed active and randomly chosen "inactive" compounds. In addition, the size and relative composition of training data sets were varied and SVM-based virtual screening was carried out using either the biological screening database or ZINC compounds as a background. The results of these systematic calculations are reported in the following.

MATERIALS AND METHODS

Support Vector Machines. SVMs[7] are a supervised machine-learning technique for binary object classification and ranking. In the training phase, a set of "positive" and "negative" data are projected into a feature space χ. In ligand-based virtual screening, training objects are known active and assumed/known inactive compounds that are represented by a feature vector $\mathbf{x}_i \in \chi$. During optimization, a convex quadratic optimization problem is solved, and a hyperplane H is derived that best separates the positive and negative training objects from each other. Maximizing the margin (i.e., distance from the nearest training examples) and minimizing training errors are basic requirements to achieve model generalization and high prediction performance. The hyperplane H is defined by the normal weight vector $\mathbf{w}$ and a bias b, so that $H = \{\mathbf{x}|\langle \mathbf{w},\mathbf{x}\rangle + b = 0\}$, with $\langle\cdot,\cdot\rangle$ being a scalar product.

Test data, i.e., compounds with unknown activity, are also mapped into the feature space χ. Depending on which side of the hyperplane the test compounds fall, they are classified either as positive (active) or negative (inactive). In SVM ranking, compounds are sorted from the position most distant to the hyperplane on the positive half-space to the most distant position on the negative half-space using their score $g(\mathbf{x}) = \langle \mathbf{w},\mathbf{x}\rangle$.

In order to allow model building in the case of nonlinearly separable training data in the feature space χ, the so-called *Kernel trick*[22] is applied to replace the standard scalar product $\langle\cdot,\cdot\rangle$ by a kernel function $K(\cdot,\cdot)$. The kernel transfers the calculation of the scalar product into a higher dimensional space $\mathcal{H}$, where a linear separation might be feasible, without explicitly calculating the mapping into $\mathcal{H}$. This operation is at the core of SVM modeling.

As descriptors for SVM modeling, sets of numerical descriptors or fingerprints (i.e., bit string representations of molecular structure and properties) can be used.

Compound Data Sets. Six confirmatory high-throughput screening (HTS) assays have been extracted from PubChem, as reported in Table 1. These inhibitor assays were chosen to select targets from diverse families and maximize the number of confirmed active compounds available for modeling. From all assays, confirmed active and inactive compounds consisting of at least five non-hydrogen atoms were extracted. For each data set, a set of assumed inactive compounds was randomly selected from ZINC (version 12) that had the same size as the set of experimentally confirmed inactive compounds.

Table 1. Compound Data Sets[a]

AID	target	target code	# actives	# inactives
504332	euchromatic histone-lysine N-methyltransferase 2	EHMT2	30,170	262,493
1030	aldehyde dehydrogeanse 1	ALDH1A1	15,822	143,429
504333	bromodomain adjacent to zinc finger domain 2B	BAZ2B	15,539	307,386
504444	nuclear factor erythroid 2-related factor 2 isoform 2	Nrf2	7,284	280,339
588855	transforming growth factor beta	SMAD3	4,800	343,727
588591	polymerase eta	POLH	4,664	366,364

[a]Data sets were extracted from PubChem Confirmatory Bioassays. For each of the six data sets, the PubChem assay id (AID), the target name, and a target code are given. In addition, the numbers of confirmed active (# actives) and confirmed inactive compounds (# inactives) are reported. The data sets are sorted by decreasing numbers of active compounds.

Training Set Composition and Background Databases. SVM models were built using active training compounds from PubChem and either confirmed inactive PubChem compounds (P/P) or assumed inactive compounds randomly selected from ZINC (P/Z). In each case, training set sizes were varied to include all possible combinations of 100, 200, 500, or 1000 active and inactive training compounds for a total of 16 combinations with different proportions of active and inactive compounds. In each case, the remaining active compounds according to Table 1 were then added as potential hits to the background/screening database. Two background databases were explored in each case. The first database contained all remaining inactive compounds (not used for model building) from each PubChem screening set. Thus, the size of this background database varied in each case according to Table 1 but always contained hundreds of thousands of compounds. As the second background database, the same number of randomly selected ZINC compounds was used in each case. Combination of the two training set categories (P/P, P/Z) and the two background databases (P, Z) resulted in four distinct training/test categories designated as P/P−P, P/P-Z, P/Z-P, and P/Z-Z. These four training/test categories in combination with 16 different training set compositions then gave rise to 64 alternative screening setups.

Calculations and Performance Criteria. Compounds were represented using the extended-connectivity fingerprint[23] with bond diameter 4 (ECFP4) or MACCS structural keys[24] calculated with the Molecular Operating Environment.[25] For SVM model building, the Tanimoto kernel[26] was used. With

dx.doi.org/10.1021/ci4002712 | J. Chem. Inf. Model. 2013, 53, 1595−1601

Table 2. Average AUC Values and Standard Deviations for ECFP4[a]

(A)								
EHMT2	P/Z-Z		P/P−P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.800	0.011	0.647	0.017	0.641	0.013	0.692	0.042
100A_200I	0.820	0.008	0.666	0.014	0.661	0.009	0.706	0.026
100A_500I	0.840	0.003	0.679	0.006	0.671	0.006	0.708	0.015
100A_1000I	0.846	0.004	0.681	0.007	0.674	0.004	0.702	0.015
200A_100I	0.819	0.008	0.653	0.014	0.655	0.010	0.704	0.029
200A_200I	0.837	0.008	0.682	0.008	0.668	0.014	0.730	0.022
200A_500I	0.855	0.005	0.706	0.008	0.682	0.007	0.744	0.020
200A_1000I	0.865	0.004	0.711	0.009	0.686	0.005	0.740	0.016
500A_100I	0.833	0.005	0.676	0.013	0.664	0.009	0.721	0.032
500A_200I	0.851	0.006	0.702	0.007	0.678	0.010	0.750	0.013
500A_500I	0.873	0.004	0.728	0.005	0.695	0.006	0.776	0.010
500A_1000I	0.885	0.004	0.743	0.005	0.705	0.005	0.783	0.012
1000A_100I	0.837	0.006	0.679	0.011	0.663	0.009	0.721	0.023
1000A_200I	0.859	0.005	0.708	0.008	0.680	0.009	0.756	0.025
1000A_500I	0.882	0.002	0.739	0.004	0.702	0.006	0.794	0.018
1000A_1000I	0.896	0.002	0.761	0.002	0.713	0.004	0.809	0.008
(B)								
ALDH1A1	P/Z-Z		P/P−P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.813	0.009	0.606	0.012	0.619	0.010	0.671	0.031
100A_200I	0.825	0.010	0.623	0.009	0.623	0.009	0.690	0.028
100A_500I	0.844	0.004	0.638	0.008	0.630	0.007	0.686	0.014
100A_1000I	0.847	0.008	0.646	0.011	0.628	0.009	0.689	0.021
200A_100I	0.830	0.008	0.624	0.010	0.631	0.009	0.687	0.027
200A_200I	0.843	0.010	0.645	0.010	0.635	0.010	0.709	0.016
200A_500I	0.863	0.008	0.669	0.011	0.644	0.007	0.722	0.019
200A_1000I	0.869	0.007	0.674	0.009	0.643	0.008	0.718	0.018
500A_100I	0.841	0.005	0.640	0.014	0.642	0.006	0.676	0.032
500A_200I	0.861	0.004	0.671	0.010	0.653	0.005	0.718	0.024
500A_500I	0.884	0.003	0.696	0.007	0.663	0.005	0.755	0.011
500A_1000I	0.893	0.004	0.712	0.005	0.666	0.005	0.760	0.016
1000A_100I	0.850	0.005	0.647	0.009	0.644	0.007	0.702	0.024
1000A_200I	0.872	0.006	0.678	0.008	0.654	0.006	0.730	0.020
1000A_500I	0.893	0.003	0.708	0.006	0.670	0.005	0.767	0.019
1000A_1000I	0.905	0.003	0.732	0.003	0.676	0.007	0.795	0.017
(C)								
BAZ2B	P/Z-Z		P/P−P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.835	0.012	0.711	0.013	0.696	0.011	0.759	0.025
100A_200I	0.849	0.009	0.728	0.013	0.708	0.009	0.771	0.020
100A_500I	0.862	0.008	0.739	0.015	0.709	0.008	0.773	0.020
100A_1000I	0.868	0.007	0.747	0.010	0.708	0.008	0.771	0.015
200A_100I	0.852	0.008	0.724	0.011	0.710	0.006	0.771	0.021
200A_200I	0.867	0.009	0.747	0.011	0.728	0.007	0.790	0.016
200A_500I	0.880	0.007	0.765	0.005	0.732	0.007	0.798	0.009
200A_1000I	0.890	0.008	0.778	0.008	0.730	0.008	0.801	0.011
500A_100I	0.868	0.007	0.732	0.012	0.716	0.012	0.785	0.019
500A_200I	0.881	0.006	0.766	0.009	0.731	0.010	0.814	0.013
500A_500I	0.901	0.003	0.794	0.005	0.754	0.003	0.831	0.008
500A_1000I	0.911	0.004	0.810	0.006	0.759	0.005	0.837	0.009
1000A_100I	0.873	0.006	0.751	0.009	0.723	0.012	0.793	0.021
1000A_200I	0.892	0.004	0.782	0.008	0.741	0.008	0.826	0.014
1000A_500I	0.912	0.002	0.810	0.004	0.764	0.005	0.853	0.007
1000A_1000I	0.924	0.001	0.830	0.003	0.775	0.005	0.863	0.008

dx.doi.org/10.1021/ci4002712 | J. Chem. Inf. Model. 2013, 53, 1595−1601

Table 2. continued

(D)								
Nrf2	P/Z-Z		P/P−P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.776	0.005	0.639	0.013	0.614	0.008	0.666	0.029
100A_200I	0.781	0.008	0.653	0.011	0.609	0.011	0.666	0.020
100A_500I	0.795	0.011	0.669	0.009	0.607	0.013	0.674	0.017
100A_1000I	0.801	0.011	0.673	0.007	0.606	0.012	0.671	0.020
200A_100I	0.800	0.007	0.669	0.014	0.631	0.010	0.696	0.017
200A_200I	0.816	0.007	0.679	0.011	0.636	0.010	0.697	0.019
200A_500I	0.826	0.008	0.698	0.009	0.633	0.011	0.699	0.017
200A_1000I	0.837	0.008	0.706	0.007	0.629	0.009	0.703	0.017
500A_100I	0.817	0.008	0.689	0.012	0.650	0.010	0.705	0.013
500A_200I	0.833	0.007	0.714	0.009	0.658	0.011	0.726	0.011
500A_500I	0.856	0.005	0.740	0.007	0.671	0.007	0.744	0.017
500A_1000I	0.869	0.005	0.747	0.004	0.671	0.008	0.738	0.010
1000A_100I	0.824	0.006	0.696	0.007	0.656	0.011	0.708	0.014
1000A_200I	0.846	0.003	0.730	0.007	0.669	0.011	0.742	0.019
1000A_500I	0.868	0.002	0.760	0.008	0.684	0.006	0.770	0.010
1000A_1000I	0.885	0.004	0.771	0.003	0.692	0.006	0.770	0.006
(E)								
SMAD3	P/Z-Z		P/P−P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.824	0.004	0.711	0.014	0.693	0.008	0.739	0.026
100A_200I	0.839	0.007	0.728	0.013	0.700	0.013	0.752	0.029
100A_500I	0.855	0.006	0.745	0.015	0.701	0.013	0.756	0.026
100A_1000I	0.859	0.006	0.750	0.009	0.701	0.012	0.754	0.018
200A_100I	0.834	0.008	0.734	0.013	0.704	0.008	0.767	0.018
200A_200I	0.856	0.004	0.753	0.008	0.723	0.007	0.787	0.017
200A_500I	0.872	0.003	0.773	0.006	0.725	0.004	0.794	0.016
200A_1000I	0.882	0.007	0.783	0.004	0.721	0.009	0.795	0.013
500A_100I	0.852	0.005	0.753	0.010	0.718	0.008	0.783	0.019
500A_200I	0.875	0.006	0.775	0.011	0.738	0.007	0.807	0.014
500A_500I	0.898	0.003	0.800	0.005	0.759	0.007	0.824	0.012
500A_1000I	0.909	0.004	0.815	0.005	0.759	0.007	0.828	0.012
1000A_100I	0.865	0.006	0.762	0.011	0.728	0.009	0.790	0.016
1000A_200I	0.885	0.004	0.790	0.010	0.747	0.005	0.815	0.014
1000A_500I	0.908	0.003	0.820	0.006	0.768	0.007	0.847	0.008
1000A_1000I	0.924	0.005	0.836	0.005	0.780	0.005	0.857	0.009
(F)								
POLH	P/Z-Z		P/P−P		P/Z-P		P/P-Z	
# refs	AUC	SD	AUC	SD	AUC	SD	AUC	SD
100A_100I	0.931	0.005	0.830	0.012	0.824	0.006	0.903	0.011
100A_200I	0.937	0.004	0.837	0.011	0.829	0.008	0.906	0.010
100A_500I	0.942	0.007	0.848	0.005	0.828	0.010	0.911	0.006
100A_1000I	0.944	0.005	0.848	0.007	0.823	0.009	0.905	0.006
200A_100I	0.938	0.005	0.842	0.008	0.837	0.006	0.910	0.013
200A_200I	0.947	0.003	0.858	0.005	0.848	0.005	0.922	0.008
200A_500I	0.954	0.002	0.866	0.004	0.849	0.004	0.924	0.008
200A_1000I	0.955	0.002	0.869	0.005	0.843	0.004	0.924	0.006
500A_100I	0.943	0.004	0.848	0.005	0.839	0.007	0.913	0.008
500A_200I	0.954	0.002	0.865	0.004	0.853	0.003	0.927	0.007
500A_500I	0.962	0.001	0.881	0.003	0.862	0.004	0.938	0.003
500A_1000I	0.966	0.002	0.887	0.003	0.864	0.004	0.940	0.002
1000A_100I	0.947	0.005	0.853	0.006	0.842	0.008	0.922	0.007
1000A_200I	0.959	0.001	0.869	0.006	0.858	0.003	0.934	0.006
1000A_500I	0.968	0.002	0.891	0.003	0.870	0.004	0.947	0.004
1000A_1000I	0.973	0.002	0.900	0.002	0.874	0.003	0.950	0.004

[a]For each data set, average AUC values over 10 independent trials and standard deviations (SD) are reported for different numbers of active and inactive reference compounds (# refs) comprising all possible combinations of 100, 200, 500, and 1000 active (A) and inactive (I) compounds. For example, "100A_100I" refers to a training set consisting of 100 active and 100 inactive compounds. Training/test categories are abbreviated as

Table 2. continued

defined in the Materials and Methods section (P/P−P, P/P-Z, P/Z-P, and P/Z-Z). Virtual screening results are reported for all six data sets, and ECFP4 as the molecular representation: (A) EHMT2, (B) ALDH1A1, (C) BAZ2B, (D) Nrf2, (E) SMAD3, and (F) POLH.

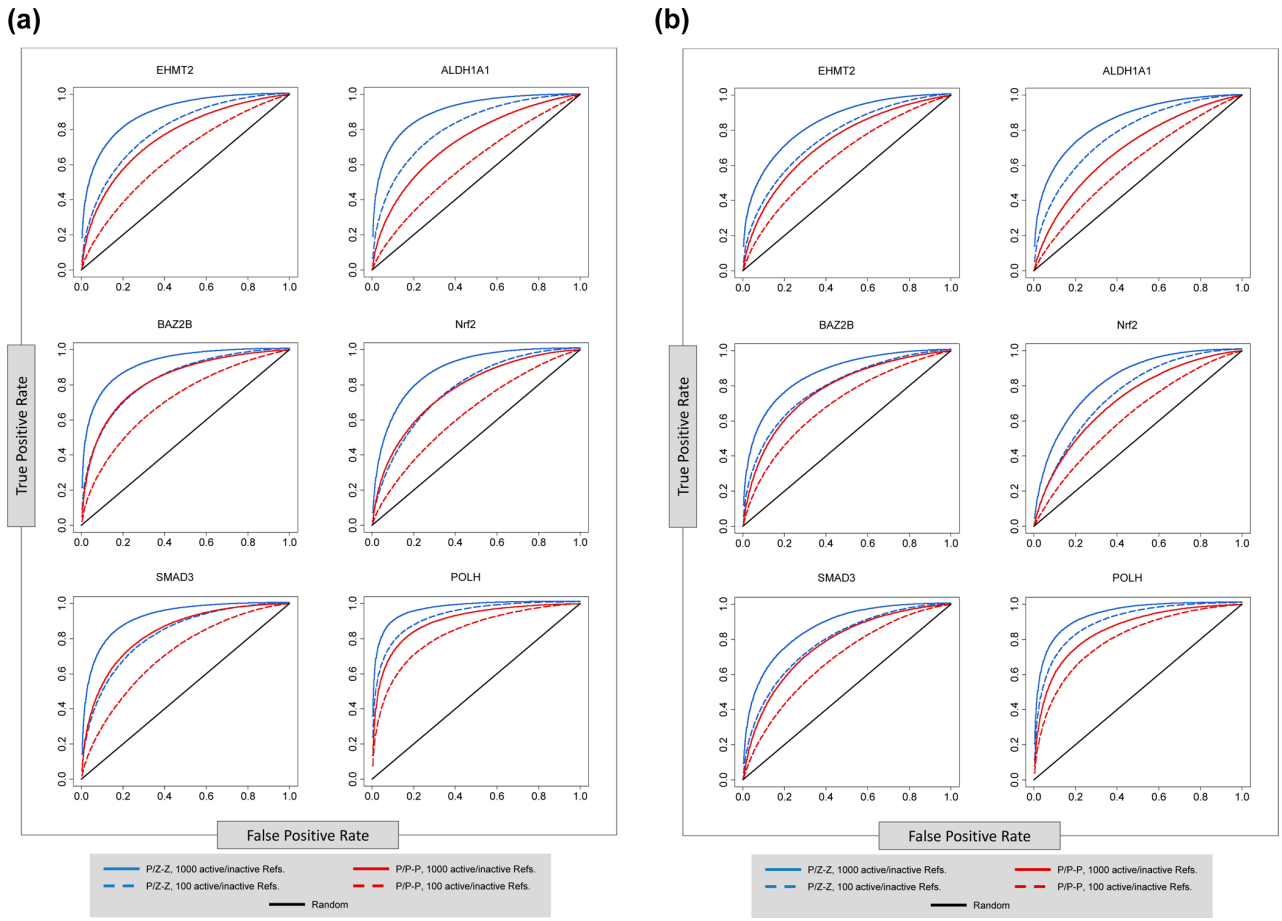


Figure 1. ROC curves. For all data sets, ROC average curves are shown for P/Z-Z (blue) and P/P−P (red) SVM calculations. In both cases, results are reported for 100 active and inactive reference compounds (dashed line) and for 1000 active and inactive training compounds (solid line). The black line represents random search performance. (a) ECFP4, (b) MACCS.

these two molecular representations, a total of 128 virtual screening constellations were obtained.

In each case, 10 different trials with randomly assembled positive and negative training and test sets were carried out. Virtual screening performance was measured using the receiver operating characteristic (ROC), and the area under the ROC curve (AUC)[27] averaged over all 10 trials.

All SVM calculations were carried out using SVM^{light},[28] a freely available SVM implementation, and as calculation parameters, SVM^{light} default settings were used.

RESULTS AND DISCUSSION

Study Goal and Design. The major focal point of our study is the question to what extent the choice of confirmed inactive molecules versus randomly selected database compounds as negative training examples might influence the outcome of SVM-based virtual screening calculations. Typically, the use of negative training data is not much discussed in SVM modeling and virtual screening applications. In benchmark calculations, it is common practice to use randomly selected database compounds, mostly from ZINC, as training compounds assumed to be inactive against a target of interest. This represents an approximation underlying model building, and it would scientifically be more rigorous to utilize confirmed inactive compounds for training. To address this issue, we have carried out systematic SVM calculations for different compound data sets obtained from PubChem Confirmatory Bioassays by applying well-defined training/test categories. These categories represented different combinations of PubChem and/or ZINC compounds for training and as the background database. This made it possible to compare SVM search performance for different training and test settings in detail.

Global Search Performance. Table 2 reports AUC values (and their standard deviations) for all SVM calculations using the ECFP4 fingerprint. The corresponding search results for MACCS structural keys are provided in Table S1 of the Supporting Information. In addition, in Figure 1A and B, ROC curves are shown for the smallest and largest training sets of P/Z-Z and P/P−P calculations for ECFP4 and MACCS, respectively.

Overall, the SVM-based virtual screening results yielded high search performance for the best categories, with AUC values of

dx.doi.org/10.1021/ci4002712 | J. Chem. Inf. Model. 2013, 53, 1595−1601

~0.8−0.9 or even higher for all compound data sets. Highest search performance was generally observed for P/Z-Z calculations, i.e., when ZINC compounds were utilized as negative training examples and as the background database. Over all data sets, these calculations yielded AUC values that were on average ~0.1−0.2 higher than for P/P−P calculations, i.e., when confirmed inactive were used for training and search calculations were carried out in the PubChem screening database. These in part significant differences are reflected in Figure 1. For both fingerprint representations, equivalent trends were observed (Figure 1). AUC values were generally slightly higher for ECFP4 than for MACCS.

Training Set Composition. Increasing numbers of reference compounds generally led to increasing search performance, as illustrated in Figure 1. Average increases in AUC values ranged from ~0.04−0.11. In addition, with increasing numbers of reference compounds, standard deviations of the calculations decreased. However, no notable changes in AUC values were observed when training sets having the same size, but inverted composition were compared, e.g., 200 active and 500 inactive vs 500 active and 200 inactive compounds. Thus, overall increases in the number of reference compounds had a stronger influence on SVM performance than training set permutations.

Category-Dependent Differences in Search Performance. The different training/test categories we defined displayed systematic differences in search performance. As a general trend, search performance over all data sets decreased in the following order: P/Z-Z > P/P-Z > P/P-P > P/Z-P. Highest standard deviations were generally observed for P/P-Z calculations. The observed order indicated that confirmed active compounds were generally easier to identify on a ZINC background than in the screening database from which they originated, regardless of whether the SVM models were trained with PubChem or ZINC compounds as negative training instances. The most likely explanation for this finding is that active and inactive PubChem confirmatory assay compounds are often more similar to each other than active PubChem and random ZINC compounds, which are chemically very diverse. This explanation is also consistent with the observation that P/Z-Z calculations, which best exploited chemical differences between compounds gave the overall best search results. These calculations also produced AUC values that were on average ~0.10−0.21 higher than P/Z-P calculations. Furthermore, when PubChem compounds were used as negative training examples, the search performance was even slightly higher on a ZINC than a PubChem background, with average increases in AUC values of ~0.01−0.06. Thus, active compounds were easier to distinguish from ZINC compounds. These findings are also consistent with a major influence of the background database on the search results, irrespective of training conditions, with ZINC compounds yielding consistently best results.

Practical Implications. The results we obtained indicate that building predictive SVM models on the basis of actual screening libraries is more difficult than using combinations of known active and randomly chosen database compounds, although the use of confirmed inactive screening compounds as negative training examples is scientifically more accurate. By comparing the different screening setups explored, the background database was found to play a major role for the success of the virtual screening calculations. Simply put, confirmed active compounds were easier to distinguish from random ZINC selections than from the screening database from which they originated, indicating that many ZINC compounds might contain chemical characteristics that distinguish them from compounds selected for screening libraries. It should be noted that the P/Z-Z category represents a typical benchmark setting. In this case, models were derived from active and random training examples and used to screen a ZINC background database to with known actives were added. This setup produced consistently best results. For practical applications, a more realistic scenario would be to train SVM models on confirmed active and inactive compounds obtained from an initial experimental screen and apply these models to search a larger screening collection. This case exactly corresponds to our P/P−P category. However, under these conditions, search performance was consistently lower than for the P/Z-Z category. These findings indicate that SVM search performance under typical benchmark conditions is likely overestimated. In fact, active compounds utilized here were confirmed screening hits that were not chemically optimized. Hence, such screening hits differ from many optimized active compounds that are usually obtained from medicinal chemistry databases such as ChEMBL and used for benchmarking. Given their generally higher chemical complexity compared to screening hits, such active compounds are even easier to differentiate from random ZINC compounds, which can be expected to further increase search performance in benchmark calculations.

■ CONCLUSIONS

Herein, we have investigated the question to what extent the choice of negative training examples influences the outcome of SVM-based virtual screening. This question is usually only little considered in SVM modeling. For this purpose, we have compared a variety of SVM models that were generated on the basis of confirmed active and inactive compounds from different PubChem Confirmatory Bioassays with corresponding models generated from confirmed actives and randomly selected ZINC compounds. In these calculations, a clear influence of negative training examples on SVM search performance was detected. The results of search calculations using the same numbers of confirmed inactive training compounds and ZINC molecules assumed to be inactive systematically differed. In addition, we also observed that generally highest search performance was achieved when SVM models were screened on a ZINC background, regardless of whether the models were trained using inactive PubChem compounds or ZINC molecules. These findings revealed a strong influence of the background database on the virtual screening results. The best SVM models we obtained were derived from known active compounds and random ZINC collections and applied on a ZINC background, which corresponded to typical benchmark conditions.

■ ASSOCIATED CONTENT

Supporting Information

Table S1 reports average AUC values and standard deviations for MACCS structural keys. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50*, 205−216.

(2) Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 53−62.

(3) Duda, R. O.; Hart, P. E.; Stork, D. G. *Pattern Classification*,2nd ed.; Wiley-Interscience: New York, 2000, pp 20−83.

(4) Bender, 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. Comput. Sci.* **2004**, *44*, 170−178.

(5) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of biological targets for compounds using multiple-category bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124−1133.

(6) Watson, P. Naive Bayes classification using 2D pharmacophore feature triplet vectors. *J. Chem. Inf. Model.* **2008**, *48*, 166−178.

(7) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd Ed.; Springer: New York, 2000.

(8) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discovery* **1998**, *2*, 121−167.

(9) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.

(10) 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.

(11) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−561.

(12) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(13) Liu, T.; Lin, Y.; Wen, X.; Jorisson, R. N.; Gilson, M. K. BindingDB: A Web-accessible database of experimentally determined protein−ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.

(14) Han, L. Y.; Ma, X. H.; Lin, H. H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z. R.; Cao, Z. W.; Ji, Z. L.; Chen, Y. Z. A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *J. Mol. Graph. Model.* **2008**, *26*, 1276−1286.

(15) Plewczynski, D.; Spieser, S. A. H.; Koch, U. Assessing different classification methods for virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 1098−1106.

(16) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of biological activity profiles using substructural analysis and genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.

(17) Smusz, S.; Kurczab, R.; Bojarski, A. J. The influence of the inactives subset generation on the performance of machine learning methods. *J. Cheminf.* **2013**, *5*, 17 DOI: 10.1186/1758-2946-5-17.

(18) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A free tool to discover chemistry for biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.

(19) *Molecular Drug Data Report (MDDR)*; Accelrys, Inc., San Diego, CA. http://www.accelrys.com (accessed June 28, 2013).

(20) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49*, 6789−6801.

(21) Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B. A.; Bolton, E.; Gindulyte, A.; Bryant, S. H. PubChem's bioassay database. *Nucleic Acids Res.* **2012**, *40*, D400−D412.

(22) Boser, B. E.; Guyon, I. M.; Vapnik, V. N. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*; Pittsburgh, PA, 1992; ACM: New York, 1992; pp 144−152.

(23) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(24) *MACCS Structural Keys*; Accelrys, San Diego, CA.

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

(26) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093−1110.

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

(28) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods − Support Vector Learning*; Schölkopf, B., Burges, C. J. C., Smola, A. J., Eds.; MIT-Press: Cambridge, MA, 1999; pp 169−184.

Summary

SVM modeling was performed using negative training data sets of diverse compositions and applied to different screening databases. SVM models based on confirmed inactive compounds from screening experiments were shown to reduce the prediction performance compared to the use of randomly chosen database compounds. Increasing the number of training compounds, both active and inactive training instances, resulted in improved search performance. Additionally, SVM calculations were strongly influenced by the background database. The recall of active compounds was generally higher when the search was done in ZINC than in the database from the screening experiment. Hence, the typical benchmark setting with database compounds used both as negative training data and as background database was shown to yield the best recall performance.
The supporting information of this publication is available via the following URL: http://dx.doi.org/10.1021/ci4002712.

Conclusion

In this thesis, 2D fingerprint methods for LBVS have been investigated. 2D fingerprints have been analyzed in detail in the context of similarity searching. Furthermore, SVM-based approaches utilizing 2D fingerprint representations have been introduced for applications that cannot be carried out using standard search methods. In addition, benchmark settings in SVM calculations have been investigated. Major results are summarized in this chapter and conclusions are drawn.

Initially, 2D fingerprints were analyzed on a large scale in order to determine their performance range on a wide spectrum of pharmaceutical targets. For this purpose, two conceptually different fingerprints, representing opposite levels of resolution, were used in similarity search calculations. The majority of the calculations yielded good search performances of both fingerprints and were characterized by an early enrichment of active compounds.

Next, the mechanism by which 2D fingerprints recover structurally diverse active compounds was investigated. Two feature selection methods were applied to systematically reduce fingerprint representations. Similarity searching using reduced fingerprints revealed individual features that distinguished subsets of active compounds from database compounds. The assembly of these features led to a cumulative recall of structurally diverse compounds.

In the following, 2D fingerprints were used in SVM calculations. First, SVM linear combination was applied to search for compounds active against related targets. A single multi-class prediction model was obtained by combining individual SVM models for each target through the use of positive and negative linear weighting factors. This model was able to recover compounds with desired activities and deprioritize those having other activity profiles.

We then introduced two SVM-based approaches for potency-directed LBVS. A

potency-oriented SVM linear combination and the structure-activity kernel were designed that incorporated compound potency information in order to direct search calculations to the preferential detection of potent hits. The potency-oriented SVM linear combination and calculations using the structure-activity kernel retained the recall of active compounds and achieved an enrichment of potent compounds at high ranking positions.

Kernel functions were also introduced for the comparison of pairs of compounds. These kernel functions captured structural differences and common elements of molecule pairs and allowed the prediction of activity cliffs in compound data sets with high accuracy.

Finally, the impact of differently composed negative training data and background databases on SVM-based VS was analyzed. Both the negative training data and the choice of the screening database were shown to strongly influence the search performance. The calculations revealed that the typical benchmark setting using assumed inactive database compounds for training and as a background database produced higher recall than calculations utilizing confirmed inactives for training.

In conclusion, 2D fingerprints were found to be powerful molecular representations when searching for novel active compounds. They yielded high performance in search calculations and were able to recover structurally diverse compounds despite their simplicity. The SVM methodology utilizing the fingerprint representations further extends the spectrum of available LBVS methods. SVM approaches that incorporate potency information as a search parameter have been introduced. Furthermore, an intrinsic limitation of similarity-based prediction methods, i.e. the presence of activity cliffs, was addressed by the design of appropriate kernel functions. Additionally, the influence of different benchmark settings was analyzed and revealed an overestimation of the search performance by typical benchmark conditions.

Additional publications during Ph.D. period

Original research publications

Balfer, J.; Heikamp, K.; Laufer, S.; Bajorath, J. Modeling of compound profiling experiments using support vector machines. *Chem. Biol. Drug Des.*, in press.

Stumpfe, D.; Dimova, D.; Heikamp, K.; Bajorath, J. Compound pathway model to capture SAR progression: comparison of activity cliff-dependent and -independent pathways. *J. Chem. Inf. Model.* **2013**, *53*, 1067-1072.

Dimova, D.; Heikamp, K.; Stumpfe, D.; Bajorath, J. Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets. *J. Med. Chem.* **2013**, *56*, 3339-3345.

Reviews and perspectives

Heikamp, K.; Bajorath, J. Support vector machines for drug discovery. *Expert Opin. Drug Discov.* **2014**, *9*, 93-104.

Heikamp, K.; Bajorath, J. The future of virtual compound screening. *Chem. Biol. Drug Des.* **2013**, *81*, 33-40.

Heikamp, K.; Bajorath, J. Fingerprint design and engineering strategies: rationalizing and improving similarity search performance. *Future Med. Chem.* **2012**, *4*, 1945-1959.

Eidesstattliche Erklärung

An Eides statt versichere ich hiermit, dass ich die Dissertation "Application and Development of Computational Methods for Ligand-Based Virtual Screening" selbst und ohne jede unerlaubte Hilfe angefertigt habe, dass diese oder eine ähnliche Arbeit noch an keiner anderen Stelle als Dissertation eingereicht worden ist und dass sie an den nachstehend aufgeführten Stellen auszugsweise veröffentlicht worden ist:

Heikamp, K.; Bajorath, J. Large-scale similarity search profiling of ChEMBL compound data sets. *J. Chem. Inf. Model.* **2011**, *51*, 1831-1839.

Heikamp, K.; Bajorath, J. How do 2D fingerprints detect structurally diverse active compounds? Revealing compound subset-specific fingerprint features through systematic selection. *J. Chem. Inf. Model.* **2011**, *51*, 2254-2265.

Heikamp, K.; Bajorath, J. Prediction of compounds with closely related activity profiles using weighted support vector machine linear combinations. *J. Chem. Inf. Model.* **2013**, *53*, 791-801.

Wassermann, A. M.; Heikamp, K.; Bajorath, J. Potency-directed similarity searching using support vector machines. *Chem. Biol. Drug Des.* **2011**, *77*, 30-38.

Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of activity cliffs using support vector machines. *J. Chem. Inf. Model.* **2012**, *52*, 2354-2365.

Heikamp, K.; Bajorath J. Comparison of confirmed inactive and randomly selected compounds as negative training examples in support vector machine-based virtual screening. *J. Chem. Inf. Model.* **2013**, *53*, 1595-1601.

Kathrin Heikamp
Januar 2014
Bonn