Ligand Prediction from Protein Sequence and Small Molecule Information Using Support Vector Machines and Fingerprint Descriptors

Hanna Geppert,[†] Jens Humrich,[‡] Dagmar Stumpfe,[†] Thomas Gärtner,[‡] 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, and Fraunhofer Institute Intelligent Analysis and Information Systems IAIS, Schloss Birlinghoven, D-53754 Sankt Augustin, Germany

Received January 8, 2009

Support vector machine (SVM) database search strategies are presented that aim at the identification of small molecule ligands for targets for which no ligand information is currently available. In pharmaceutical research and chemical biology, this situation is faced, for example, when studying orphan targets or newly identified members of protein families. To investigate methods for de novo ligand identification in the absence of known three-dimensional target structures or active molecules, we have focused on combining sequence and ligand information for closely and distantly related proteins. To provide a basis for these investigations, a set of 11 protease targets from different families was assembled together with more than 2000 inhibitors directed against individual proteases. We have compared SVM approaches that combine protein sequence and ligand information in different ways and utilize 2D fingerprints as ligand descriptors. These methodologies were applied to search for inhibitors of individual proteases not taken into account during learning. A target sequence-ligand kernel and, in particular, a linear combination of multiple target-directed SVMs consistently identified inhibitors with high accuracy including test cases where homology-based similarity searching using data fusion and conventional SVM ranking nearly or completely failed. The SVM linear combination and target-ligand kernel methods described herein are intuitive and straightforward to adopt for ligand prediction against other targets.

1. INTRODUCTION

The design of target-specific small molecules for therapeutic intervention has dominated drug discovery research over the past decades.[1] However, the conventional view of specifically active small molecules as potential drug candidates is beginning to change as a consequence of emerging trends including polypharmacological drug behavior[1,2] and chemical genetics interference with cellular functions.[3] Polypharmacology refers to increasing insights that therapeutic effects of many drugs are mediated through interaction with multiple, rather than single targets.[2] In chemical genetics, small molecules are utilized to modulate targets and study functional consequences and resulting cellular phenotypes.[3] In this context, biologically active compounds are not prioritized on the basis of therapeutic potential but rather their ability to interfere with biological functions and elicit certain phenotypes. However, both polypharmacology and chemical genetics suggest evaluating compound specificity in a way different from the conventional one in drug discovery research. Simply put, specific interactions against individual but also multiple targets should be considered when trying to understand biological consequences of small molecule treatment or to identify compounds for chemical genetics or pharmaceutical applications.

Chemical genetics strategies are central to research in chemical biology, which rapidly evolves as an intrinsically interdisciplinary field at interfaces between synthetic and medicinal chemistry and the life sciences.[4] Chemical biology generally focuses on the use of small molecules to probe and elucidate biological functions of target proteins.[4] Such functional analysis can proceed in different ways. For example, small molecules might be evaluated in cell-based assays to generate interesting phenotypes and responsible target proteins would subsequently be identified or, alternatively, selected proteins might be directly targeted using small molecular probes.[4,5] For the study of protein functions using small molecules, a number of requirements must be met. First and foremost, small molecules must be identified that bind to selected targets or elicit a desired cellular phenotype. To these ends, if one would like to not exclusively rely on brute-force compound screening, or might not be able to do so, methods need to be applied that help to focus candidate compounds on specific targets or target families.[6,7] For this purpose, computational design often provides a promising approach.[7,8] Furthermore, in addition to finding suitable molecules for chemical biology applications, their selectivity profiles must be analyzed and confirmed, as mentioned above.[8] For example, if test compounds bind to multiple related target proteins, functions of individual targets might be difficult to discern. On the other hand, an understanding of functional consequences of addressing protein families, subfamilies, or target networks[9] might also be desired and would require the availability of promiscuous (yet specific)

* To whom correspondence should be addressed. Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.
[†] Rheinische Friedrich-Wilhelms-Universität Bonn.
[‡] Fraunhofer Institute Intelligent Analysis and Information Systems IAIS.

10.1021/ci900004a CCC: $40.75
Published on Web 03/23/2009

Table 1. Target and Ligand Data Sets[a]

target	abbreviation	PDB entry	number of ligands	nearest neighbor target	
				name	sequence identity (%)
calpain 1	cal1	1tlo	46	cal2	66.2
calpain 2	cal2	1kfu	49	cal1	66.2
caspase 1	cas1	1ice	21	cas3	27.4
caspase 3	cas3	1gfw	264	cas1	27.4
cathepsin B	catB	1gmy	17	catS	30.8
cathepsin K	catK	1yk7	223	catS	57.8
cathepsin L	catL	1mhw	78	catK	48.8
cathepsin S	catS	1ms6	221	catK	57.8
factor Xa	faXa	1mq5	783	thro	39.6
thrombin	thro	1ppb	281	faXa	39.6
trypsin	tryp	1trn	58	faXa	35.1

[a] For each target protein, a corresponding Protein Data Bank (PDB) entry, the number of ligands, and its nearest neighbor target (including sequence identity) are reported.

compounds or small molecule libraries encoding different selectivity profiles. Moreover, finding small molecule ligands for novel genomic targets with little information available other than their sequences presents additional challenges, both from an experimental and computational point of view.

Given the complexity of the chemical biology applications described above and the efforts and potential difficulties associated with the identification of suitable small molecules, it is understandable that one increasingly attempts to integrate computational concepts into this process.[8] Although the development of such concepts and methods is currently still in its infancy, several recent studies have addressed aspects of chemical biology through computational analysis and design.[8,10] These studies include, for example, systematic mapping of ligand-target interactions,[11] profiling of compounds against arrays of Bayesian classifiers representing different target activities,[12] or focusing on target-selective compounds in chemical database mining.[13−15]

Here, we address one of the chemical biology-related tasks discussed above, namely, the prediction of small molecule ligands for novel members of protein families or proteins for which no ligand information is available. In molecular modeling, ligand design in the absence of known active molecules is typically attempted using docking algorithms[16] or de novo design methodologies[17] that require target structure or pharmacophore information. By contrast, our intention is to ultimately identify small molecule ligands for "orphan" targets by only utilizing their amino acid sequences and ligands of closely or distantly related proteins. A few previous investigations have addressed conceptually similar questions from a computational perspective. For example, early studies have attempted to correlate sequence identity of proteins with chemical similarity of their ligands.[18−20] Building upon these studies, the situation in ligand-based virtual screening was addressed when no active small molecules were available for a target of interest that could serve as screening templates. In such cases, known ligands of homologous targets were utilized as a starting point for what has been termed homology-based similarity searching.[21]

In addition to these investigations, other studies have aimed to apply machine learning techniques, in particular support vector machines (SVMs), to predict active compounds by combining ligand and target sequence information. SVMs[22−24] are algorithms for supervised machine learning that were originally developed for binary object classification (i.e., class label prediction).

The basic idea of SVM learning is to construct a hyperplane in feature space that best separates objects of two given classes, thereby trying to minimize the classification error while maximizing generalization performance to avoid overfitting. Objects are then classified depending on which side of the hyperplane they fall. SVMs have become especially popular with the advent of kernel functions,[25−27] which allow generalization to cases where the classification function is not a hyperplane but any hypersurface. Recently, kernel functions have been introduced that combine small molecule descriptor and target sequence information[28,29] including a sequence homology-based enzyme classification kernel to predict inhibitors of enzymes not used for learning.[29] Furthermore, descriptor vectors were generated that combined equivalent physicochemical descriptors for pairs of ligands and receptor sequences (calculated from ligand structures and amino acids) and were used to distinguish true ligand−receptor pairs from false combinations.[30]

Here we report a different approach that combines ligand and target sequence information through a linear combination of individual target-directed SVM classifiers and enables the prediction of ligands for targets for which no ligands are available during learning. When applied to a collection of proteases that were assembled and used as a benchmark system, our methodology correctly predicted inhibitors for specific targets with high accuracy including cases where reference methods failed.

2. DATA SETS

2.1. Targets. For our analysis, eleven targets from four different protease families were selected and combined into a target set (Table 1). This set was designed to include subsets of closely related targets. These subsets were either distantly related or essentially unrelated to each other (Figure 1). The selected enzymes belonged to the papain C1 (cathepsins B, K, L, and S), calpain C2 (calpains 1 and 2), caspase C14 (caspases 1 and 3), and chymotrypsin S1 family (trypsin, thrombin, and factor Xa).[31] As illustrated in Figure 1, the targets of the papain family include two subfamilies, with cathepsins (cat) K, L, and S belonging to one subfamily and cat B to another.[32] The papain family and the calpain family form the papain superfamily. Caspases and the papain

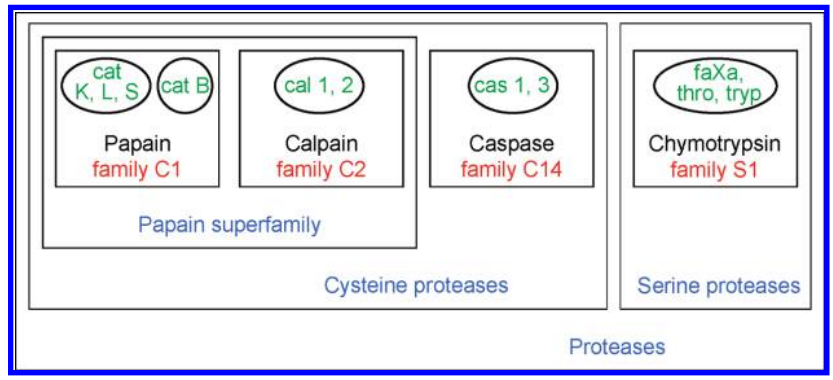


Figure 1. Target relationships. The relationships between the 11 proteases forming the target set are illustrated, as discussed in the text.

Table 2. Pairwise Target Sequence Identity[a]

	cal1	cal2	cas1	cas3	catB	catK	catL	catS	faXa	thro	tryp
cal1	100.0										
cal2	66.2	100.0									
cas1	3.0	8.2	100.0								
cas3	9.8	13.2	27.4	100.0							
catB	14.6	14.4	8.0	1.5	100.0						
catK	13.7	13.1	0.7	2.4	27.4	100.0					
catL	12.9	12.8	6.1	10.1	23.2	48.8	100.0				
catS	5.7	13.1	4.1	2.6	30.8	57.8	45.4	100.0			
faXa	8.7	9.8	14.7	8.6	3.9	16.4	9.9	12.1	100.0		
thro	6.0	7.6	16.8	15.2	9.2	12.0	8.1	7.9	39.6	100.0	
tryp	6.2	8.9	11.9	10.9	4.3	6.8	8.6	4.5	35.1	34.4	100.0

[a] For all possible target pairs, sequence identities (in %) are reported.

superfamily are mammalian cysteine proteases.[33] By contrast, members of the chymotrypsin family are serine proteases.

Pairwise sequence comparisons for the protease set were carried out using EMBOSS[34,35] with default settings. For this purpose, amino acid sequences were extracted from Protein Data Bank[36] (PDB) entries listed in Table 1. Calculated sequence identities ranged from 0.7% to 66.2% and are reported in Table 2. On the basis of the sequence identity matrix, the most similar target, termed "nearest neighbor target" (see Table 1), was identified for each protease as the starting point for homology-based similarity search and simple SVM calculations, as described below.

2.2. Ligands. For the assembly of ligand sets for all proteases, compounds were taken from the Molecular Drug Data Report[37] (MDDR), the BindingDB database,[38−41] and original scientific literature. A total of 2,041 different protease inhibitors were collected for our 11 protein targets and organized in 11 ligand sets, with each individual set containing between 17 and 783 compounds, as reported in Table 1. Given the sources of our compounds, rule-of-five or other drug-likeness filters were not applied. For inclusion into the ligand set of a target, a candidate inhibitor was required to have higher than 1 μM potency (K_i or IC_{50}) against its target. In case of ligand promiscuity, that is, inhibitory activity against multiple proteases, a compound was assigned to the highest potency target. Thus, a candidate inhibitor could only be part of one ligand set. The composition of our ligand sets and exact source information including original references for each compound are reported in Supporting Information Table S1. The molecular weight of ligands ranged from 174 to 844 Da, with a mean of 489.

3. METHODS AND CALCULATIONS

In our study, we have considered six different search strategies that were compared for their ability to predict ligands for targets for which no ligand information was utilized. These strategies were also applied in standard virtual screening where known ligands for the investigated target were included in the training set. These calculations served as a reference point for SVM ligand predictions. As small molecular representations, different 2D fingerprints were calculated, which transform 2D molecular graphs into vectors of binary values.[42] These types of fingerprints are widely used descriptors for the identification of active molecules.[43]

3.1. Similarity Searching Using Data Fusion. Similarity searching[42] is the traditional way to utilize molecular fingerprints in ligand-based virtual screening. The basic idea is to determine fingerprint overlap between a database and active reference compound and use it as a quantitative measure of molecular similarity. Database compounds are then ranked by decreasing similarity scores, and those with high scores are considered to have a high probability to be biologically active. Fingerprint overlap is assessed by means of a similarity metric, the most prominent being the Tanimoto coefficient[42] (Tc), which is also applied in this study. If multiple active reference compounds are available, so-called data fusion methods can be utilized.[44] For example, a database compound is separately compared to each reference molecule and either only the highest similarity value is retained, referred to as "MAX" fusion rule, or the sum of all individual similarity values is calculated, referred to as "SUM" fusion rule, to yield the final similarity score. In "homology-based similarity searching",[21] ligands from the

nearest neighbor target are taken as reference compounds while ligands of the target itself are not utilized or unavailable.

3.2. Simple SVM Ranking. While similarity searching only relies on active reference molecules (or positive training examples), simple SVM ranking also takes inactive molecules (negative training examples) into account. Solving a convex quadratic optimization problem, a hyperplane H is derived that is defined by a normal vector $\mathbf{w}$ and a scalar b

$$H = \{\mathbf{x} | \langle \mathbf{x}, \mathbf{w} \rangle + b = 0\} \quad (1)$$

such that active training molecules are located in the positive half-space and inactive training molecules in the negative half-space. During the optimization procedure, some training compounds are allowed to be placed on the incorrect side of the plane by introducing so-called slack variables that penalize these training errors. After training, test compounds can be classified as active or inactive depending on which side of H they map. However, to obtain a compound ranking instead of two class labels, test compounds $\mathbf{x}$ can be sorted according to the value of

$$g(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle \quad (2)$$

By doing so, compounds located in the positive half-space are ranked by decreasing distances from H, followed by compounds in the negative half-space ranked by increasing distances from H. To permit decision boundaries that do not linearly depend on the training data, scalar products $\langle \cdot, \cdot \rangle$ occurring in the objective function of the SVM optimization problem can be replaced by a kernel function $K(\cdot, \cdot)$.[25−27] However, for simple SVM ranking, a linear kernel was used in this study that is equal to the standard scalar product.

To apply SVM for ligand prediction, we defined "homology-based SVM" in an analogous manner to homology-based similarity searching. That is, ligands of the nearest neighbor target were employed as positive training examples and as negative training examples, a randomly chosen subset of the screening database was utilized, as described in section 3.6.

3.3. SVM Linear Combination. To generate an SVM linear combination (SVM-LC), for each target $t_i \in T$ $i = (1,..., n)$ with a known ligand set L_i, an individual weight vector $\mathbf{w}_i$ is derived. This weight vector results from the hyperplane $H_i = \{\mathbf{x} | \langle \mathbf{x}, \mathbf{w}_i \rangle + b_i = 0\}$ generated by conventional binary SVM learning with linear kernel function when using L_i as positive training examples and randomly selected screening database molecules as negative training examples. For a target t_j of interest, we then build the final weight vector $\mathbf{w}_j^{\text{final}}$ as a linear combination of the individual $\mathbf{w}_i$ using entries $d(t_j,t_i)$ of the sequence identity matrix d as factors

$$\mathbf{w}_j^{\text{final}} = \sum_{t_i \in T} d(t_j, t_i) \mathbf{w}_i$$

Note that in the case of standard virtual screening where ligands for t_j are available ($t_j \in T$), the individual weight vector $\mathbf{w}_j$ contributes to $\mathbf{w}_j^{\text{final}}$ with the factor $d(t_j,t_j) = 100$. In contrast, in the case of ligand prediction where no ligand information for t_j is available ($t_j \notin T$), $\mathbf{w}_j^{\text{final}}$ does not contain the term $\mathbf{w}_j$. To obtain a compound ranking for target t_j, test compounds $\mathbf{x}$ are sorted according to the value of $g(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w}_j^{\text{final}} \rangle$.

3.4. SVM with Target-Ligand Kernel. Learning from multiple target classes by using an SVM with target-ligand kernel (SVM-TLK) was recently described by Erhan and L'Heureux[28] and Jacob and Vert.[29] The underlying idea has been to use true/false target-ligand pairs as positive/negative training examples for a conventional SVM, instead of active/inactive molecules. For this purpose, scalar products $\langle \cdot, \cdot \rangle$ between target-ligand pairs that have to be calculated during the SVM optimization and ranking procedure are determined by means of a target-ligand kernel function. Given appropriate individual representations $\mathbf{t}$ and $\mathbf{x}$ for targets and ligands, respectively, a target-ligand kernel relating two different target-ligand pairs to each other can be defined by combining two separate kernels for target pairs and ligand pairs[28,29]

$$K((\mathbf{t}, \mathbf{x}), (\mathbf{t}', \mathbf{x}')) = K(\mathbf{t}, \mathbf{t}') \times K(\mathbf{x}, \mathbf{x}')$$

After derivation of a separating hyperplane H in target-ligand pair feature space, molecules $\mathbf{x}$ of the screening database are combined with the target $\mathbf{t}_j$ (for which new ligands should be identified) to form pairs $(\mathbf{t}_j,\mathbf{x})$ and sorted by the signed distance of $(\mathbf{t}_j,\mathbf{x})$ from H, as described in section 3.2.

Different kernels for targets and ligands have been proposed.[28,29,45−48] However, to make our SVM-TLK approach directly comparable to simple SVM ranking and SVM-LC, the linear kernel was applied as ligand kernel and the protein sequence identity matrix d as target kernel resulting in the combined target-ligand kernel

$$K((\mathbf{t}, \mathbf{x}), (\mathbf{t}', \mathbf{x}')) = d(\mathbf{t}, \mathbf{t}') \times \langle \mathbf{x}, \mathbf{x}' \rangle$$

We were able to use the protein sequence identity matrix d as the target kernel because it was positive semidefinite on our set of proteins, which is a critical requirement during the SVM optimization procedure.

For SVM-TLK training, true and false target-ligand pairs had to be generated in a systematic fashion. In order to obtain a positive training set, each ligand was combined with its corresponding target. For the generation of negative training sets, two different strategies were evaluated, referred to as SVM-TLK1 and SVM-TLK2, respectively. In SVM-TLK1, false target-ligand pairs were obtained by combining randomly selected database compounds with each of our eleven targets. In SVM-TLK2, each ligand that was used to generate a positive training example was also combined with an incorrect target randomly selected from the 10 remaining ones. Thus, in this case, database compounds were not utilized for training. This second strategy corresponded to the generation of training examples reported by Jacob and Vert.[29]

3.5. Computational Complexity of SVM-LC and SVM-TLK. Because the different screening approaches take different chemical information during training into account, it is worth considering their training time complexity. Let $|A|$ and $|I|$ be the number of active and inactive training molecules (with $|A| \ll |I|$), respectively, which are utilized per target, and $|T|$ be the number of reference targets. Compared to the three SVM-based methods, similarity searching does not involve a training step. Instead, each database compound must be compared to $|A|$ reference molecules during ranking, whereas in SVM-based methods, database molecules must only be combined with a single weight vector. When comparing the different SVM-based approaches, it must be considered that deriving the hyperplane generally requires computational time that is ap-

Table 3. Composition of Training or Reference Compound Sets[a]

method	positive training examples or reference compounds	negative training examples
	ligand prediction	
homology-based similarity searching homology-based SVM simple ranking	5 ligands of the nearest neighbor target	1000 DB compounds
SVM-LC	5 ligands of each target, except the investigated one, that is, 50 ligands in total	1000 DB compounds
SVM-TLK1		1000 DB compounds paired with each of the 11 targets, that is, 11 000 target-ligand pairs in total
SVM-TLK2	5 ligands of each target, except the investigated one, paired with its own target, that is, 50 target-ligand pairs in total	5 ligands of each target, except the investigated one, paired with an incorrect target, that is, 50 target-ligand pairs in total
	standard virtual screening	
similarity searching SVM simple ranking	5 ligands of the investigated target	1000 DB compounds
SVM-LC	5 ligands of each of the 11 targets, that is, 55 ligands in total	1000 DB compounds
SVM-TLK1		1000 DB compounds paired with each of the 11 targets, that is, 11 000 target-ligand pairs in total
SVM-TLK2	5 ligands of each of the 11 targets, paired with its own target, that is, 55 target-ligand pairs in total	5 ligands of each of the 11 targets, paired with an incorrect target, that is, 55 target-ligand pairs in total

[a] For each search strategy, the training (SVM) or reference sets (similarity searching) are described. LC and TLK stand for linear combination and target-ligand kernel, respectively; DB means database.

proximately cubic relative to the number of training data points. However, this relationship is critical because the number of training points for the alternative approaches is differently influenced by $|T|$, resulting in the following time complexities:

- simple SVM ranking: $O((|A| + |I|)^3)$
- SVM-LC: $O(|T| \times (|A| + |I|)^3)$
- SVM-TLK1: $O(|T|^3 \times (|A| + |I|)^3)$
- SVM-TLK2: $O(|T|^3 \times |A|^3)$

Thus, when using 10 targets for training, SVM-LC takes 10 times longer than simple SVM ranking but SVM-TLK1 takes 100 times longer than SVM-LC. Hence, SVM-LC is computationally much more efficient than SVM-TLK1.

3.6. Test Calculations. The similarity and SVM search methods discussed above were evaluated in systematic virtual screening calculations on each of our 11 targets using 100 000 randomly selected ZINC7[49] compounds as a background database. As ligand descriptors, three alternative 2D fingerprints were applied: MACCS structural keys,[50] the circular atom environment fingerprint Molprint2D,[51,52] and the two-point pharmacophore-type[53] fingerprint TGD[54] that uses 15 distance ranges to monitor graph distances between atom pairs assigned to seven different pharmacophore feature types.

For ligand prediction on each target, database search trials were carried out with 10 different randomly chosen training and test compound sets, as summarized in Table 3.

For SVM-LC and SVM-TLK learning, five ligands were selected from each of the 10 reference targets and used as active training set compounds. Thus, a total of 50 active training molecules were utilized to predict inhibitors for the remaining target. For homology-based similarity searching and simple SVM ranking, five compounds from the nearest neighbor target were chosen as active reference molecules.

For standard virtual screening that was carried out as reference, 55 active training molecules from all 11 targets were assembled for SVM-LC and SVM-TLK and five compounds from each individual target for conventional similarity searching and simple SVM ranking. Thus, these reference calculations took ligand information for each target directly into account.

For SVM learning, 1000 background database compounds were randomly selected as inactive training molecules, while the remaining 99 000 molecules served as database decoys. For each individual target under investigation, the remaining N-5 compounds were added as active test molecules, that is, potential database hits. As a performance measure, recovery rates (RR: number of detected active ligands of a target divided by the total number of potential database hits) were determined for compound selection sets ranging from 10 to 5000 molecules and averaged over 10 independent trials per target.

All SVM calculations reported in this study were carried out using SVM^{light}, a freely available SVM implementation, with standard parameter settings.[55,56] MAX and SUM similarity search calculations were facilitated using in-house Perl scripts.

4. RESULTS AND DISCUSSION

The major task of our study has been the identification of active compounds using SVM-based search methods for targets for which ligand information was not taken into account during training. For a set of protease targets representing close, remote, or essentially nonexisting sequence relationships, inhibitors were collected and SVM strategies developed that combined ligand structure and target sequence information in different ways. Applying the resulting models, known ligands hidden in a background database were predicted for protease targets only using their amino acid sequences as input. Ultimately, these methods aim at identifying ligands for orphan targets based on training sets including closely and distantly related proteins and their available ligand information. In order to benchmark the

Table 4. Search Results for Standard Virtual Screening

method	MAX		SUM		SVM		SVM-LC		SVM-TLK1		SVM-TLK2	
set size	100	1000	100	1000	100	1000	100	1000	100	1000	100	1000
						MACCS						
cal1	54.6	67.8	42.2	53.4	62.4	78.5	69.0	80.7	65.4	86.8	6.3	17.3
cal2	87.0	94.1	86.8	92.0	91.1	93.9	93.9	99.5	80.0	97.7	18.6	32.5
cas1	79.4	85.0	63.1	74.4	80.0	89.4	76.9	91.9	83.1	91.9	20.0	36.9
cas3	28.4	49.6	20.8	41.7	26.6	50.4	27.3	56.1	23.3	58.2	1.6	9.8
catB	68.3	79.2	60.8	76.7	70.8	83.3	73.3	95.8	70.0	93.3	20.8	30.0
catK	17.8	33.8	16.1	33.1	27.1	60.5	28.4	71.9	26.8	68.4	2.2	10.2
catL	32.5	41.0	30.7	44.7	38.5	67.9	42.6	89.6	43.0	80.0	3.2	16.7
catS	29.4	56.0	29.8	60.2	32.2	66.3	32.3	77.8	29.1	71.7	2.5	12.5
faXa	11.6	28.7	9.7	31.7	11.7	41.3	11.7	50.2	10.9	47.9	1.2	7.8
Thro	30.3	52.9	31.6	57.4	33.3	76.4	33.8	81.6	29.9	73.0	4.1	16.6
Tryp	36.2	51.3	35.8	49.4	54.2	87.0	57.0	90.8	53.8	86.8	2.5	15.5
average	43.2	58.1	38.9	55.9	48.0	72.3	49.7	80.5	46.8	77.8	7.6	18.7
						Molprint2D						
cal1	79.4	93.4	77.1	94.1	82.7	96.8	80.0	98.3	90.0	98.3	57.3	64.9
cal2	94.5	99.3	93.6	100.0	94.5	100.0	98.2	100.0	99.3	100.0	81.1	88.0
cas1	82.5	92.4	83.1	90.6	91.9	92.5	92.5	94.4	92.5	100.0	65.0	81.3
cas3	35.8	58.3	29.5	51.2	36.6	74.1	36.2	79.6	34.2	82.4	22.0	45.3
catB	84.2	89.2	75.8	83.3	89.2	93.3	92.5	93.3	92.5	93.3	64.2	76.7
catK	32.4	59.4	34.5	59.9	39.6	74.1	36.5	82.9	37.8	84.5	29.0	51.1
catL	44.9	57.6	43.4	61.8	57.7	88.9	66.0	98.6	75.2	98.4	43.7	67.5
catS	40.6	64.1	39.3	61.9	43.2	76.3	37.2	87.3	40.1	87.6	32.5	63.1
faXa	12.8	42.2	12.5	35.4	12.7	63.6	12.5	72.3	12.2	70.6	11.9	43.7
thro	32.0	59.4	32.6	59.3	34.5	79.2	34.2	87.1	33.7	86.1	30.5	55.8
tryp	47.1	58.7	55.3	71.3	77.4	87.4	78.3	92.5	74.5	91.1	44.3	58.3
average	53.3	70.4	52.4	69.9	60.0	84.2	60.4	89.7	62.0	90.2	43.8	63.2
						TGD						
cal1	23.2	35.1	17.8	31.7	32.2	53.2	37.8	63.9	41.0	72.2	10.0	18.0
cal2	62.5	74.5	68.4	77.3	83.0	93.0	87.5	95.2	79.1	95.5	25.0	45.7
cas1	80.6	80.6	59.4	70.6	93.1	99.4	94.4	100.0	80.6	100.0	18.1	31.9
cas3	18.8	28.4	13.9	20.5	30.5	52.6	33.9	57.2	27.8	59.0	2.2	10.8
catB	38.3	47.5	40.0	50.0	79.2	87.5	77.5	87.5	71.7	85.0	23.3	32.5
catK	14.9	26.8	11.2	22.3	19.1	46.2	17.4	50.2	17.5	47.8	2.0	10.7
catL	31.1	45.5	23.8	45.8	23.7	47.8	24.4	57.0	32.1	68.9	8.8	28.2
catS	21.9	36.1	14.1	28.8	22.8	44.3	13.2	50.5	14.3	50.7	1.9	15.6
faXa	9.0	19.5	10.0	27.8	11.6	61.9	10.0	63.4	9.2	48.1	5.5	28.7
thro	30.7	58.4	32.2	59.0	34.1	81.2	33.7	84.9	28.5	77.9	14.6	44.1
tryp	44.5	62.6	39.1	65.1	69.4	97.5	69.1	98.3	54.7	90.9	20.9	47.5
average	34.1	46.8	30.0	45.4	45.3	69.5	45.3	73.5	41.5	72.4	12.0	28.5

performance of our target-ligand SVM methods, different types of reference calculations were carried out.

4.1. Conventional Virtual Screening. The results of conventional virtual screening applying the six different database ranking approaches are reported in Table 4 and Supporting Information Table S2. Similarity searching using data fusion of five known active reference molecules for a target produced overall reasonable search results (with MAX being slightly superior to the SUM fusion rule). For example, for database selection sets of 100 compounds and the MACCS/MAX combination, recovery rates (RR) exceeded 50% for four (cal1, cal2, cas1, catB), fell in the range between 28−37% for five (cas3, catL, catS, thro, tryp), and were lower than 20% for only two targets (catK, faXa).

Compared to similarity searching, simple SVM ranking systematically improved search performance, consistent with previous findings[57] on different compound activity classes. This improvement was a likely consequence of the gain in available chemical information by explicitly taking inactive training compounds into consideration. While simple SVM ranking performance was overall comparable to MAX for small selection sets of up to 50 compounds, RR increased a lot for larger sets in many experiments. For example, for selection sets of 1000 compounds and MACCS, RR increased by about Δ10% for three targets and Δ24−36% for four others. For the TGD fingerprint applied to catB and faXa, RR increased by more than Δ40%.

Compared to similarity searching with multiple reference compounds and simple SVM ranking, SVM-LC takes more information into account because ligands from other targets are also utilized during training. However, in standard virtual screening, SVM-LC search performance was comparable to simple SVM ranking for small selection sets and only slightly better for larger selection sets. For the different fingerprints and selection sets of 1000 compounds, by using SVM-LC RR was increased by on average Δ8.3% for MACCS, Δ5.4% for Molprint2D, and Δ4.0% for TGD, respectively. Furthermore, the performance of the target-ligand kernel method SVM-TLK1 was very similar to SVM-LC, as reported in Table 4. For selection sets of 100 and 1000 compounds, RR fluctuations for individual targets were almost always smaller than ±Δ10%, with an average of Δ1−2% higher SVM-LC RR. By contrast, the performance of SVM-TLK2 was substantially lower. Whereas SVM-TLK1 had on average an RR of 80.1% for selection sets of 1,000 compounds, SVM-TLK2 achieved only 36.8%. Thus, the way negative training examples were generated was critical for the performance of target-ligand kernel methods. When randomly

Table 5. Search Results for Ligand Prediction

method	MAX		SUM		SVM		SVM-LC		SVM-TLK1		SVM-TLK2	
set size	100	1000	100	1000	100	1000	100	1000	100	1000	100	1000
						MACCS						
cal1	46.8	56.6	39.0	47.8	45.9	61.7	49.3	71.7	49.8	76.1	14.1	21.2
cal2	52.0	69.0	45.0	56.8	61.8	78.4	68.6	90.7	50.2	84.5	0.0	0.9
cas1	14.3	31.9	5.6	9.4	19.4	51.9	9.4	31.9	9.4	31.3	1.3	3.8
cas3	18.8	29.6	14.3	29.7	21.6	41.9	15.0	46.3	8.6	39.1	0.8	5.5
catB	0.0	7.5	0.8	13.3	9.2	40.0	30.8	83.3	38.3	84.2	0.0	1.7
catK	2.6	10.1	7.2	26.5	14.9	45.2	20.0	59.2	16.5	49.4	2.7	10.6
catL	6.8	13.2	6.3	18.5	16.7	40.5	28.8	74.4	25.1	61.1	0.1	3.2
catS	1.6	10.8	5.9	24.3	15.3	42.1	24.4	62.9	17.7	48.8	1.5	8.9
faXa	3.8	12.0	6.3	17.9	9.4	34.9	9.7	34.9	5.5	21.8	0.4	3.3
thro	1.8	9.9	4.1	17.4	11.5	36.2	25.3	67.2	12.6	51.2	0.8	7.4
tryp	2.7	10.8	5.8	14.0	14.5	32.3	43.0	60.4	24.3	45.3	0.8	4.5
average	13.7	23.8	12.8	25.0	21.8	45.9	29.5	62.1	23.5	53.9	2.0	6.4
						Molprint2D						
cal1	58.3	79.4	57.3	85.6	60.5	86.6	60.0	87.6	68.8	87.3	52.9	66.3
cal2	80.8	94.5	81.1	98.4	86.8	99.8	98.4	100.0	98.9	100.0	67.0	85.9
cas1	28.8	48.1	11.3	43.1	57.5	74.4	38.8	93.8	34.4	92.5	0.6	7.5
cas3	21.6	35.6	21.9	42.8	32.1	52.9	20.1	51.2	18.2	51.3	5.7	13.7
catB	1.7	8.3	5.0	12.5	18.3	54.2	43.3	90.8	45.0	86.7	10.0	36.7
catK	8.3	21.1	12.6	33.3	21.1	55.4	26.5	73.3	26.1	72.9	6.7	19.8
catL	14.5	43.4	26.8	65.8	34.5	62.2	51.8	94.2	54.1	93.0	22.9	45.9
catS	7.6	22.9	11.9	29.9	19.1	42.6	23.3	64.8	21.5	63.1	14.2	30.7
faXa	5.9	21.6	6.4	24.4	11.4	45.0	9.8	46.4	6.1	44.7	4.7	16.7
thro	9.0	20.8	8.0	16.6	20.9	45.7	31.7	82.9	28.6	80.9	12.3	28.6
tryp	11.5	23.9	10.0	18.1	43.6	63.2	57.2	86.0	43.4	77.9	14.2	27.2
average	22.5	38.1	22.9	42.8	36.9	62.0	41.9	79.2	40.5	77.3	19.2	34.5
						TGD						
cal1	32.7	41.5	24.4	32.9	28.3	40.7	31.7	45.1	24.9	45.1	8.8	16.8
cal2	24.6	37.0	17.5	32.3	49.3	65.9	48.9	86.6	18.0	73.2	3.4	10.0
cas1	10.0	11.9	10.6	18.1	41.9	73.1	20.0	76.3	15.6	72.5	1.3	8.8
cas3	21.2	32.9	17.2	28.6	35.4	54.4	27.5	52.7	19.9	51.5	3.6	10.6
catB	0.0	2.5	0.8	2.5	12.5	32.5	37.5	59.2	35.0	60.8	1.7	3.3
catK	2.3	8.4	2.9	9.4	7.6	21.3	12.1	41.3	9.2	41.7	2.0	10.1
catL	7.5	14.9	7.9	16.6	7.0	25.6	11.2	32.7	9.5	21.4	1.1	9.2
catS	1.9	7.8	2.0	10.0	3.2	17.5	3.3	27.5	3.3	19.1	1.4	8.1
faXa	4.5	22.3	5.4	32.8	8.0	48.4	8.1	55.2	7.5	47.9	1.2	8.7
thro	2.2	9.9	8.3	25.4	22.0	73.8	30.1	81.8	28.8	79.7	10.1	24.7
tryp	1.1	9.4	11.1	23.4	34.9	78.9	52.6	87.7	48.9	80.4	9.2	24.2
average	9.8	18.1	9.8	21.1	22.7	48.4	25.7	58.7	20.0	53.9	4.0	12.2

chosen database decoys were combined with target sequences to form false target-ligand pairs (TLK1), virtual screening performance was high, but when molecules active against one target were combined with another target (TLK2) the performance was only low.

Thus, taken together, SVM-based methods performed clearly better than fingerprint similarity searching in a standard virtual screening situation. However, the benefit of using multiple active training classes in SVM-LC and SVM-TLK1 compared to simple SVM ranking was only small when ligands for the target under investigation were included during training.

4.2. Ligand Prediction Using Homology-Based Similarity Searching. In contrast to conventional similarity search calculations, similarity searching using five active reference compounds of the nearest neighbor target, that is, homology-based similarity searching, produced overall low RR and completely failed in a number of instances, as reported in Table 5. High similarity search performance was only observed for two targets, cal1 and cal2, and acceptable results were obtained for cas3. However, the ability to predict ligands was dramatically reduced for all other targets. Here, the combination of MACCS or TGD with MAX or SUM hardly reached 5% RR for selection sets of 100 molecules, and RR values were always much smaller than 25% for selection sets of 1000 molecules. For catB, essentially no ligand was detected in selection sets smaller than 1000 compounds.

An explanation for the success or failure of homology-based similarity searching is provided in Figure 2 that shows distributions of pairwise MACCS Tc compound similarities that were of relevance for the detection of cal1 (Figure 2A), catB (2B), and catS inhibitors (2C); corresponding similarity distributions for the remaining sets are provided in Supporting Information Figure S1. The diagrams on the left in Figure 2 compare similarity distributions obtained for pairs of cal1-cal1 (2A), catB-catB (2B), and catS-catS (2C) inhibitors with distributions obtained for pairs of cal1-ZINC, catB-ZINC, and catS-ZINC compounds, respectively. As can be seen, for these three targets, pairwise similarities between ligands are clearly shifted to higher similarity values than similarities between a ligand and a database compound. Thus, in these cases, ligands were easily discriminated from database compounds, consistent with the observed high RR of 67.8%, 79.2%, and 56.0%, respectively, for the MACCS/MAX strategy and selection sets of 1000 compounds. The diagrams

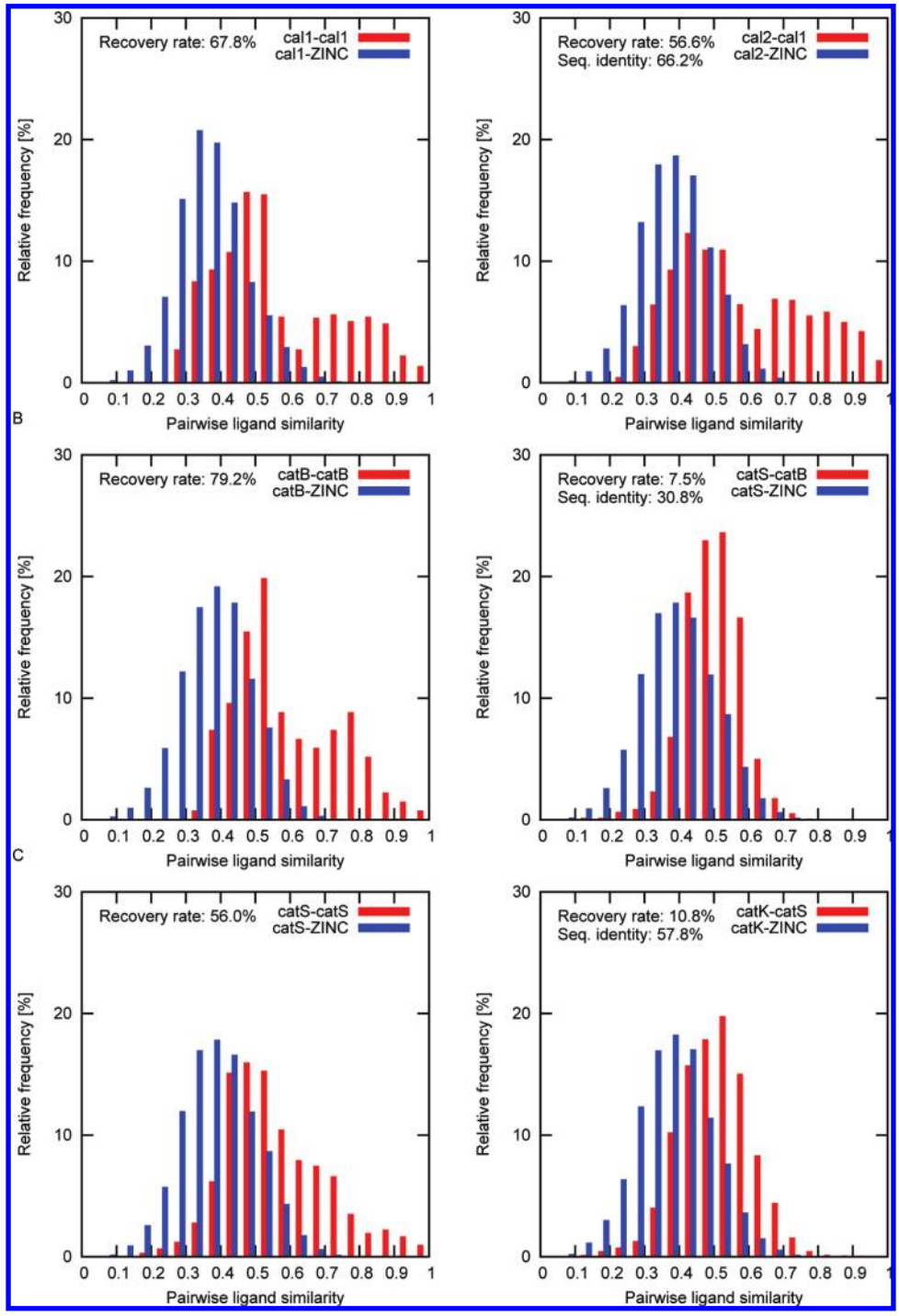


Figure 2. Compound similarity. Distributions of MACCS Tc compound similarity values between pairs of active (red) and between active and ZINC molecules (blue) are shown. Overlap of similarity distributions strongly influences the detection of (A) calpain 1, (B) cathepsin B, and (C) cathepsin S inhibitors in conventional similarity searching (graphs on the left) or homology-based similarity searching (right). The sequence identity to the nearest neighbor target is reported (right). Recovery rates are reported for selection sets of 1000 compounds and the MAX search strategy.

on the right in Figure 2 compare similarity distributions obtained for pairs of cal2-cal1, catS-catB, and catK-catS inhibitors with distributions for pairs of cal2-ZINC, catS-ZINC, and catK-ZINC compounds, respectively. These distributions are relevant for homology-based similarity searching where reference compounds of the nearest neighbor target were used. The high protein sequence identity of 66.2% between the cal1 and cal2 protein is in accordance with a high similarity of ligands of both targets. The search for cal1 ligands using cal2 ligands as reference compounds was successful, yielding an RR of 56.6% (Figure 2A). By contrast, the similarity of catB and catS inhibitors is not notably higher than the similarity between catS and ZINC compounds. Thus, using catS ligands as a reference, catB ligands could not be effectively detected, producing an RR of only 7.5% (2B). Furthermore, the similarity between catK and catS inhibitors is only slightly higher than between catK and ZINC compounds, resulting in RR of 10.8%. Thus, the success of homology-based similarity searching was determined by the shape of compound similarity distributions, which corresponded to the degree of sequence identity between nearest neighbor targets in the case of cal1 (66.2%

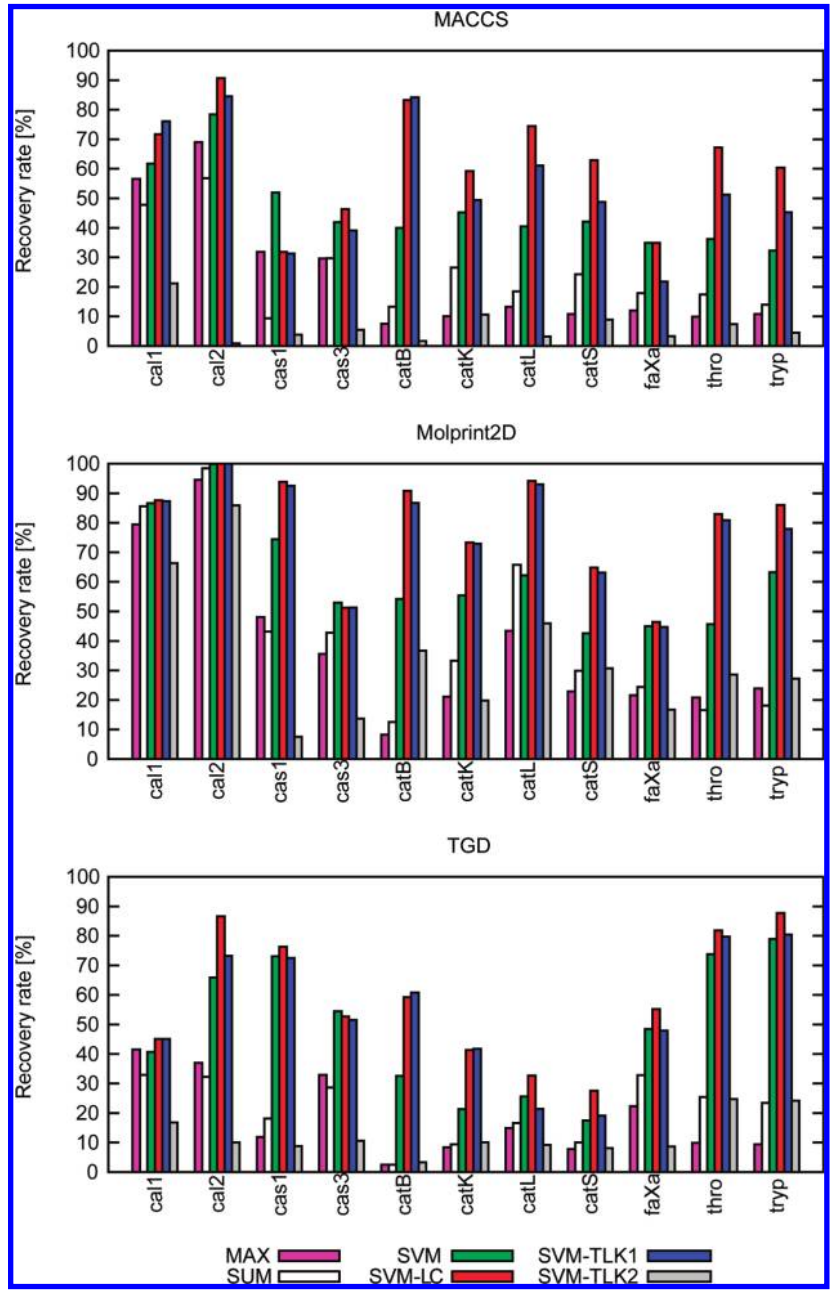
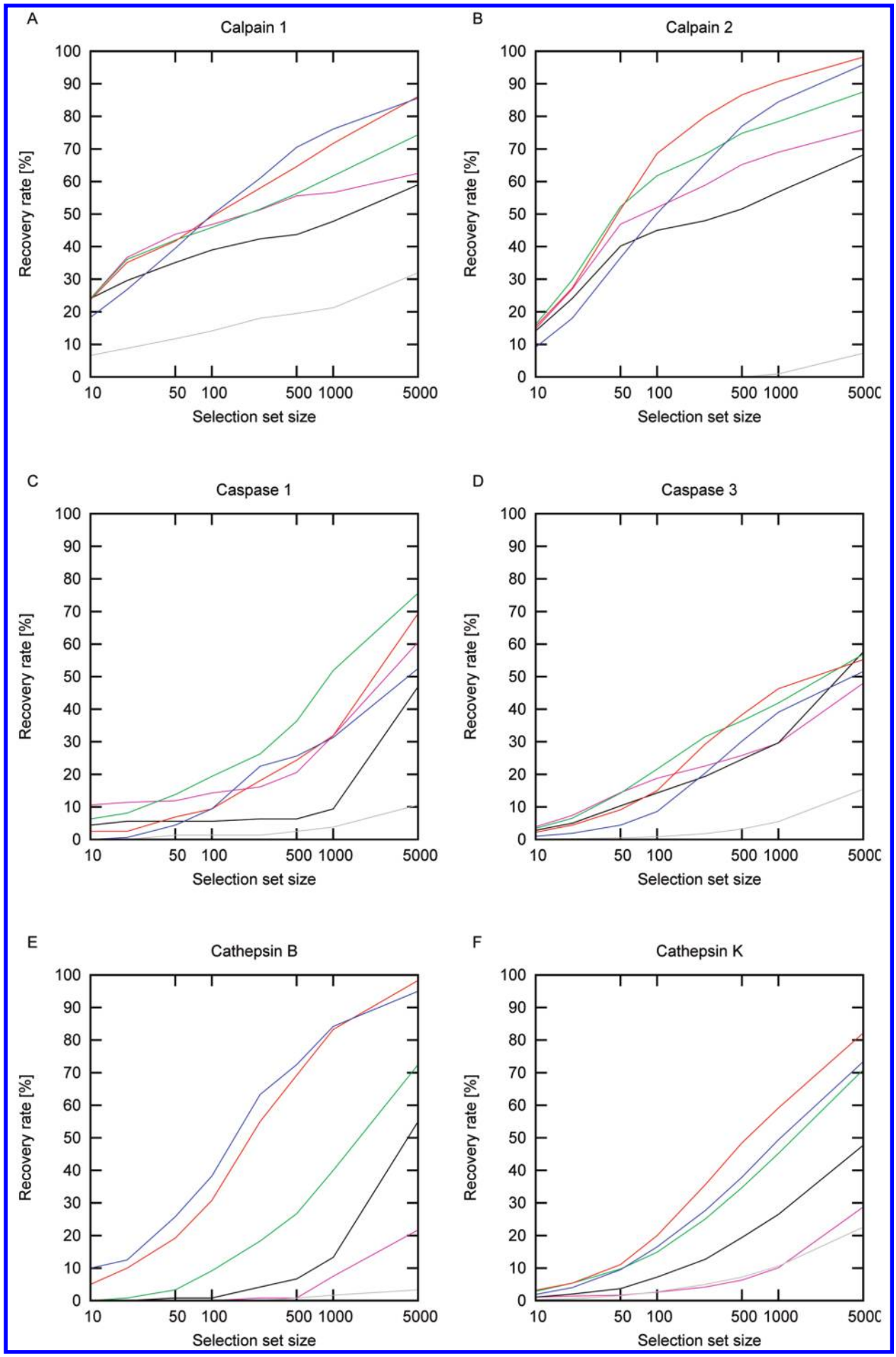


Figure 3. Comparison of ligand prediction performance. For each fingerprint descriptor, recovery rates for the six different ligand prediction strategies are compared for selection sets of 1000 compounds. Recovery rates are averaged over 10 independent search trials per target and strategy.

sequence identity versus 56.6% RR) and catB (30.8% vs 7.5%), but not in the case of catS (57.8% vs 10.8%).

4.3. Ligand Prediction Using Support Vector Machines. Compared to homology-based similarity searching, the RR increase in ligand prediction produced by our SVM methods was highly significant. Search results for selection sets of 1000 compounds are compared in Figure 3 and, in addition, cumulative recall curves for all targets and the MACCS fingerprint are shown in Figure 4. Complete search results are reported in Supporting Information Table S3. Relative to homology-based similarity searching using MAX, SVM-LC produced an increase in RR of approximately Δ50% or more for half of the target/fingerprint combinations and selection sets of 1000 compounds; for catB/MACCS, catB/Molprint2D, and tryp/TGD, the improvement was close to Δ80%. Furthermore, for ligand prediction, significant advantages of SVM-LC oversimple SVM ranking were observed. For selection sets of 1,000 compounds, SVM-LC improved the search performance of simple SVM ranking by on average Δ16.1% for MACCS, Δ17.2% for Molprint2D, and Δ10.3% for TGD. As can be seen in Figure 3, this corresponded to RR of more than 60% for two-thirds of the independent calculations and of more than 80% for one-third, although no ligand information for the investigated target was considered during training.

In ligand prediction, SVM-LC also performed better than SVM-TLK1, with an average RR increase over all fingerprints and targets of Δ4.4% and Δ5.0% for selection set sizes of 100 and 1000 compounds, respectively. SVM-LC achieved top RR using Molprint2D with on average 41.9% for 100 and 79.2% RR for 1,000 selected compounds. By contrast, SVM-TLK2 calculations nearly failed to recover active compounds using MACCS and TGD fingerprints and only produced moderate RR with Molprint2D of on average


A
Calpain 1
B
Calpain 2
C
Caspase 1
D
Caspase 3
E
Cathepsin B
F
Cathepsin K
Recovery rate [%]
Selection set size
0
10
20
30
40
50
60
70
80
90
100
10
50
100
500
1000
5000

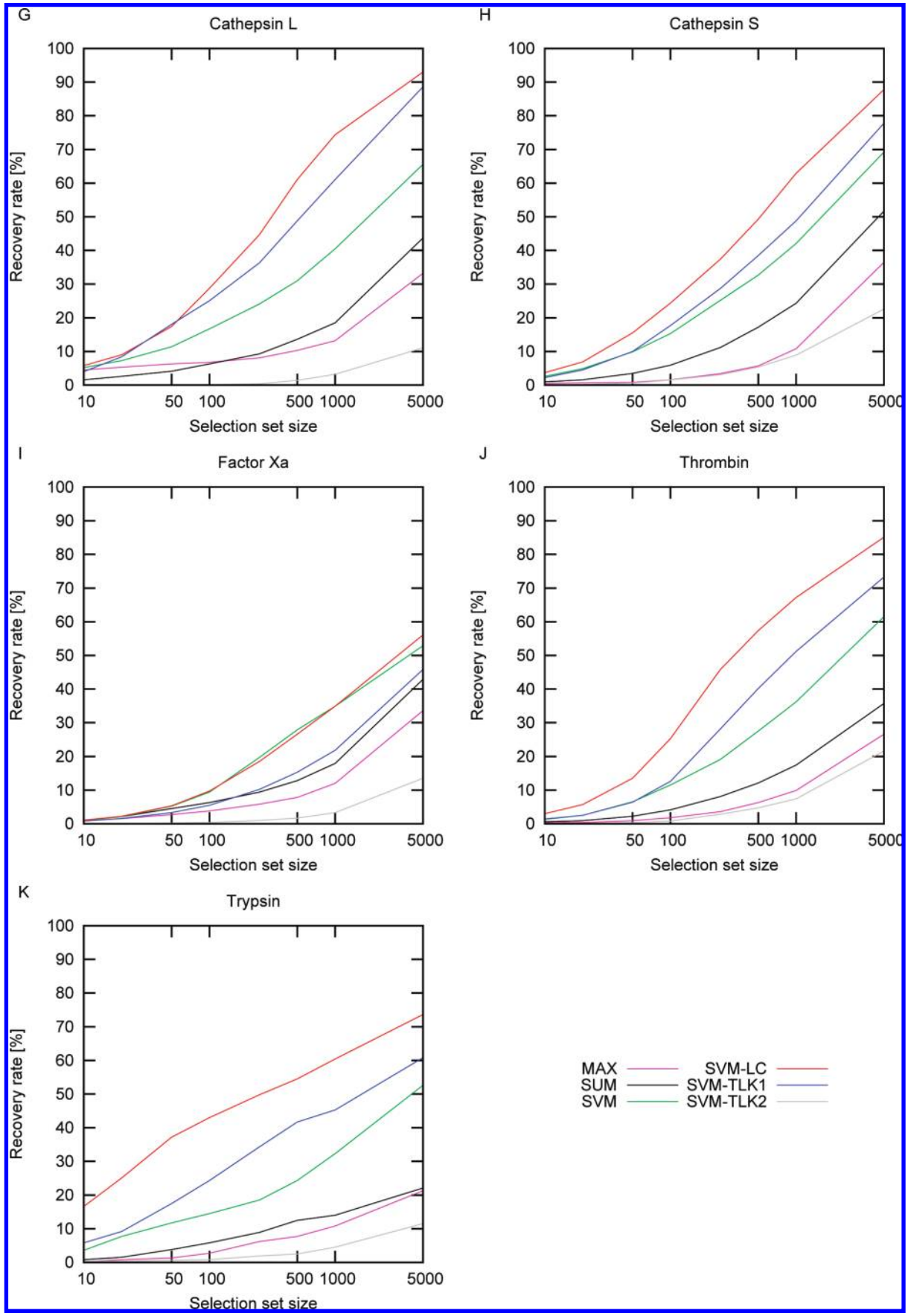


Figure 4. Cumulative recall curves for ligand prediction. Recall curves are shown for all targets using MACCS as a ligand descriptor. Selection set size is given on a logarithmic scale.

19.2% for 100 and 34.5% for 1000 selected database compounds. The dramatic decrease in search performance of SVM-TLK2 further emphasized the crucial importance of utilizing database decoys to generate negative

target-ligand training examples, rather than compounds active against other targets included in SVM model building.

Relative to other methods, SVM-LC was generally most stable when comparing standard virtual screening to ligand prediction. For two-thirds of all trials, RR decreased by less than Δ15% for selection sets of 1000 compounds when omitting the five active compounds of the investigated target during training; for the Molprint2D fingerprint, the decrease in RR was less than Δ5% for half of the targets.

Taken together, SVM-LC produced overall best results in ligand prediction. Hence, for our protease benchmark system, the linear combination of individual target-directed SVMs with sequence-derived weighting factors emerged as a computationally efficient and highly effective method to identify inhibitors for targets when no known ligand information was taken into account.

5. CONCLUSIONS

In this study, we have introduced SVM methods for the de novo identification of ligands for simulated orphan targets. Considering the conceptual simplicity of the introduced SVM linear combination and target-ligand kernel, the results of our test calculations were rather promising compared to homology-based similarity searching and simple SVM ranking. Our findings suggest that novel ligands can be effectively identified through SVM methods that utilize sequence information of investigated targets and ligand/sequence information of other proteins. High prediction accuracy was achieved with SVM models trained on a limited number of closely, distantly, and unrelated targets and their ligands. Thus, only little knowledge about closely related targets was required. Furthermore, for representing ligand information, 2D fingerprints were effective tools, consistent with our earlier observations. Although SVM-TLK1 approached the search performance of SVM-LC, the linear combination of individual SVM models was overall preferred for ligand prediction. For practical applications, SVM-LC has the additional advantage of significantly lower computational complexity over target-ligand kernel methods. Furthermore, for SVM-LC any similarity measure defined on proteins can be used to obtain factors for the linear combination of weight vectors whereas for SVM-TLK, the protein similarity matrix needs to be positive semidefinite to be a suitable kernel.

A key finding of our analysis has been the effectiveness of SVM combinations of protein sequence and ligand structure representations in extrapolating from ligand information and identifying active compounds for other targets, without prior ligand knowledge. Thus, the SVM methods presented herein should merit further evaluation in the prospective search for ligands of novel targets and complement currently available tools for computer-aided chemical biology.

ACKNOWLEDGMENT

D.S. is supported by Sonderforschungsbereich (SFB) 704 of the Deutsche Forschungsgemeinschaft (DFG). The authors thank Martin Vogt for many helpful discussions.

Supporting Information Available: Figure S1 shows distributions of pairwise compound similarities and Tables S1−S3 report the source information for all ligands and complete virtual screening and ligand prediction results, respectively. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Sams-Dodd, F. Target-based drug discovery: is something wrong. *Drug Discovery Today* **2005**, *10*, 139–147.
(2) Hopkins, A. L. Network pharmacology: The next paradigm in drug discovery. *Nature Chem. Biol.* **2008**, *4*, 682–690.
(3) Knight, Z. A; Shokat, K. M. Chemical genetics: Where genetics and pharmacology meet. *Cell* **2007**, *128*, 425–430.
(4) Stockwell, B. R. Exploring biology with small organic molecules. *Nature* **2004**, *432*, 846–854.
(5) Spring, D. R. Chemical genetics to chemical genomics: Small molecules offer big insights. *Chem. Soc. Rev.* **2005**, *34*, 472–482.
(6) Schreiber, S. L. Target-oriented and diversity-oriented synthesis in drug discovery. *Science* **2000**, *287*, 1964–1969.
(7) Schnur, D. M. Recent trends in library design: 'Rational design' revisited. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 375–380.
(8) Bajorath, J. Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.
(9) Yildirim, M. A.; Goh, K.-I.; Cusick, M. E.; Barabási, A.-L.; Vidal, M. Drug-target network. *Nat. Biotechnol.* **2007**, *25*, 1119–1126.
(10) Bajorath, J. Computational approaches in chemogenomics and chemical biology: current and future impact on drug discovery. *Expert Opin. Drug Discovery* **2008**, *3*, 1371–1376.
(11) 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.
(12) 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.
(13) Stumpfe, D.; Ahmed, H.; Vogt, I.; Bajorath, J. Methods for computer-aided chemical biology, Part 1: Design of a benchmark system for the evaluation of compound selectivity. *Chem. Biol. Drug. Des.* **2007**, *70*, 182–194.
(14) Vogt, I.; Stumpfe, D.; Ahmed, H.; Bajorath, J. Methods for computer-aided chemical biology, Part 2: Evaluation of compound selectivity using 2D fingerprints. *Chem. Biol. Drug. Des.* **2007**, *70*, 195–205.
(15) Stumpfe, D.; Geppert, H.; Bajorath, J. Methods for computer-aided chemical biology, Part 3: Analysis of structure-selectivity relationships through single- or dual-step selectivity searching and Bayesian classification. *Chem. Biol. Drug. Des.* **2008**, *71*, 518–528.
(16) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.
(17) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nature Rev. Drug Discovery* **2005**, *4*, 649–663.
(18) Frye, S. Structure-activity relationship homology (SARAH): A conceptual framework for drug discovery in the genomic era. *Chem. Biol.* **1999**, *6*, R3−R7.
(19) Mitchell, J. B. O. The relationship between the sequence identities of alpha helical proteins in the PDB and the molecular similarities of their ligands. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1617–1622.
(20) Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J.-J.; Lecchini, S.; Jacoby, E. An ontology for pharmaceutical ligands and its application for in silico screening and library design. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 947–955.
(21) 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.
(22) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discovery* **1998**, *2*, 121–167.
(23) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.
(24) Boser, B. E.; Guyon, I. M.; Vapnik, V. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*; Pittsburgh, Pennsylvania, 1992; ACM: New York, 1992; pp 144−152.
(25) Aronszajn, N. Theory of reproducing kernels. *Trans. Am. Math. Soc.* **1950**, *68*, 337–404.
(26) Schölkopf, B.; Smola, A. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.
(27) Müller, K.-R.; Rätsch, G.; Mika, S.; Tsuda, K.; Schölkopf, B. An introduction to kernel-based learning algorithms. *IEEE Neural Networks* **2001**, *12*, 181–201.
(28) 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.
(29) Jacob, L.; Vert, J.-P. Protein−ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* **2008**, *24*, 2149–2156.

(30) Bock, J. R.; Gough, D. A. Virtual screens for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model.* **2005**, *45*, 1402–1414.
(31) Rawlings, N. D.; Morton, F. R.; Kok, C. Y.; Kong, J.; Barrett, A. J. MEROPS: The peptidase database. *Nucleic Acids Res.* **2008**, *36*, D320−D325.
(32) Guay, J.; Falgueyret, J. P.; Ducret, A.; Percival, M. D.; Mancini, J. A. Potency and selectivity of inhibition of cathepsin K, L and S by their respective propeptides. *Eur. J. Biochem.* **2000**, *267*, 6311–6318.
(33) Rzychon, M.; Chmiel, D.; Stec-Niemczyk, J. Modes of inhibition of cysteine proteases. *Acta Biochim. Pol.* **2004**, *51*, 861–873.
(34) Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277.
(35) EMBOSS. http://www.ebi.ac.uk/Tools/emboss/align/index.html (accessed October 2008).
(36) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(37) *MDL Drug Data Report (MDDR)*; Symyx Software: San Ramon, CA, 2005.
(38) Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R. N.; Gilson, M. K. BindingDB: A web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.
(39) Chen, X.; Lin, Y.; Gilson, M. K. The Binding Database: Overview and user's guide. *Biopolymers Nucleic Acid Sci.* **2002**, *61*, 127–141.
(40) Chen, X.; Lin, Y.; Liu, M.; Gilson, M. K. The Binding Database: Data management and interface design. *Bioinformatics* **2002**, *18*, 130–139.
(41) Chen, X.; Liu, M.; Gilson, M. K. Binding DB: A web-accessible molecular recognition database. *J. Comb. Chem. High-Throughput Screening* **2001**, *4*, 719–725.
(42) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(43) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: Foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
(44) Hert, J.; Willet, 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.
(45) Gärtner, T. A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter* **2003**, *5*, 49–58.
(46) Borgwardt, K. M. Graph Kernels. PhD thesis in Computer Science, Ludwig-Maximilians-University, Munich, Germany, 2007.
(47) Gärtner, T. Kernels for Structured Data. *Series in Machine Perception and Artificial Intelligence*, Vol. 72; World Scientific: Singapore, 2008.
(48) Mahé, P.; Vert, J.-P. Graph kernels based on tree patterns for molecules. *Machine Learning*, in press.
(49) 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.
(50) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.
(51) MOLPRINT 2D. http://www.molprint.com (accessed Jan 2008).
(52) Bender, A.; Mussa, Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT2D): Evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
(53) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.
(54) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.
(55) Joachims, T. Making large-scale SVM learning practical. In *Advances in Kernel Methods—Support Vector Learning*; Schölkopf, B., Burges, C., Smola, A., Eds.; MIT-Press: Cambridge, MA, 1999.
(56) SVM^{light}. http://svmlight.joachims.org/ (accessed Sep 2008).
(57) 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.

CI900004A