

# Searching for Target-Selective Compounds Using Different Combinations of Multiclass Support Vector Machine Ranking Methods, Kernel Functions, and Fingerprint Descriptors

Anne Mai Wassermann, Hanna Geppert, and Jürgen Bajorath\*

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

Received December 4, 2008

The identification of small chemical compounds that are selective for a target protein over one or more closely related members of the same family is of high relevance for applications in chemical biology. Conventional 2D similarity searching using known selective molecules as templates has recently been found to preferentially detect selective over non-selective and inactive database compounds. To improve the initially observed search performance, we have attempted to use 2D fingerprints as descriptors for support vector machine (SVM)-based selectivity searching. Different from typically applied binary SVM compound classification, SVM analysis has been adapted here for multiclass predictions and compound ranking to distinguish between selective, active but non-selective, and inactive compounds. In systematic database search calculations, we tested combinations of four alternative SVM ranking schemes, four different kernel functions, and four fingerprints and were able to further improve selectivity search performance by effectively removing non-selective molecules from high ranking positions while retaining high recall of selective compounds.

## 1. INTRODUCTION

Traditionally, the identification of target-selective and potent small molecules has been a major focal point in pharmaceutical research to provide attractive leads for further development into clinical candidates. Despite the increasing notion of polypharmacological drug behavior<sup>1</sup> and the anticipated paradigm shift in evaluating drug candidates, the assessment of compound selectivity continues to be of high relevance for drug discovery. Moreover, with the advent of chemical biology<sup>2</sup> and chemogenomics,<sup>3</sup> target-selective compounds are not only of interest as leads for discovery but also as small molecular probes for elucidating cellular functions of gene products.<sup>4</sup> For example, collections of small molecules that show differential selectivity patterns for individual targets within a protein family can serve as probes for exploring multiple functions. Because the identification of compounds with target selectivity or differential selectivity patterns typically requires significant experimental efforts, the interest in computational approaches to support this process is currently increasing.<sup>5</sup> Recently, similarity searching using 2D molecular fingerprints<sup>6</sup> has been evaluated for its potential to preferentially detect selective over non-selective molecules.<sup>7–9</sup> For example, utilizing an especially designed compound benchmark system consisting of 432 molecules with differential selectivity against 13 targets,<sup>9</sup> systematic test calculations using nearest neighbor fingerprint searching<sup>10</sup> were found to enrich selective compounds in database selection sets.<sup>9</sup> Surprisingly, however, adding non-selective and inactive compounds as additional training classes for naïve Bayesian classification<sup>11</sup> and single- and

dual-step compound selection schemes could not further improve search performance.<sup>9</sup> Thus, we asked the question whether or not machine learning techniques other than naïve Bayes classifiers might be capable of doing so and evaluated support vector machines (SVM) for selectivity analysis.

Support vector machines<sup>12,13</sup> are supervised machine learning algorithms that have become increasingly popular for applications in the chemoinformatics field.<sup>14–16</sup> The SVM approach was originally developed for binary classification problems. Its central idea is to project training objects belonging to two different classes into high-dimensional feature spaces where a linear separation of objects by means of a hyperplane might become feasible. An attractive aspect of SVM classification is that the approach does not only try to reduce the classification error on the training data, but also employs so-called structural risk minimization methods to avoid overfitting effects and enhance generalization performance. Given a derived hyperplane, test objects are classified based on which side of the hyperplane they fall. As the signed distance of an object to the hyperplane provides an intuitive ranking scheme, SVM can in principle also be utilized to provide rankings of test objects. Recently, such ranking functions were successfully applied to virtual screening by ranking of compound libraries in the order of decreasing potential to be biologically active.<sup>16,17</sup> Given the high predictive performance of SVM classification in the search for active compounds,<sup>17</sup> we were encouraged to apply SVM also to selectivity analysis. However, for selectivity searching, SVM analysis had to be further modified. Instead of two training classes, three classes containing selective, non-selective, and inactive molecules are available, thus giving rise to a three-class ranking problem.

\* To whom correspondence should be addressed. Phone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

**Table 1.** Compound Selectivity Sets<sup>a</sup>

targets	selectivity set	potency (pK <sub>i</sub> or pIC <sub>50</sub> )	selectivity ratio (K <sub>i</sub> or IC <sub>50</sub> )	selectives	non-selectives
carbonic anhydrases(CA)	CA IV/I	6–7	0.9–618	10	20
	CA IV/I	7–8	0.2–2308	20	33
	CA IX/I	8–9	0.4–1441	14	11
matrix metalloproteinases(MMP)	MMP 2/I	8–9	0.2–2538	19	14
	MMP 8/I	8–9	3.5–4811	12	16
	MMP 9/I	8–9	2.2–5773	24	16
cathepsins(Cat)	Cat K/B	6–7	0.2–2000	12	10
	Cat S/B	6–7	0.2–860	12	11
	Cat K/L	7–8	0.1–10714	24	15
	Cat K/L	8–9	0.3–32173	31	10
	Cat K/S	7–8	0.1–7857	20	16
	Cat K/S	8–9	0.8–12641	25	21
	Cat S/K	7–8	0.1–26316	24	22
	Cat S/K	8–9	0.3–19697	15	10
	Cat S/L	7–8	0.1–7667	19	27
	Cat S/L	8–9	0.2–5556	14	13
trypsin(Try)	Try/Thr	6–7	0.1–500	13	35
thrombin(Thr)	Try/Thr	7–8	0.1–11429	25	10

<sup>a</sup> Each selectivity set consists of a number of molecules that are selective for one target over a closely related one and a number of non-selective compounds for these two targets (i.e., compounds that are comparably potent against both targets). The size of each subset is reported as “selectives” and “non-selectives”, respectively. For example, the first entry of the table specifies a selectivity set for carbonic anhydrases (CA). The designation “CA IV/I” means that the set consists of molecules that are selective for CA IV over CA I and non-selective molecules for these targets. The column “potency” reports the potency range of all selective and non-selective inhibitors of CA IV and “selectivity ratio” the observed selectivity ratios that result from lower or comparable potency against CA I.

In this study, we investigate three different multiclass SVM ranking strategies, that is, preference, one-versus-all, and two-step ranking, for their ability to enrich small compound selection sets with selective over non-selective test compounds. Hence we aim at high recall of selective compounds but low recall of non-selective ones, which we call “purity” of the final selection set. The performance of these multiclass methods is compared to standard binary SVM-based ranking, termed “simple ranking”, and conventional nearest neighbor similarity searching. Furthermore, different kernel functions are evaluated for SVM-based ranking and preferred combinations of ranking methods and kernel functions are identified. In the following, we provide a detailed description of the different SVM ranking strategies and report the results of our systematic search calculations.

## 2. MATERIALS, METHODS, AND CALCULATIONS

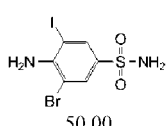
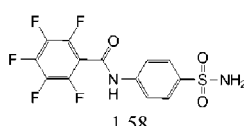
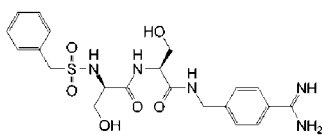
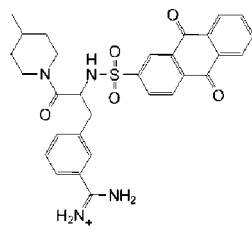
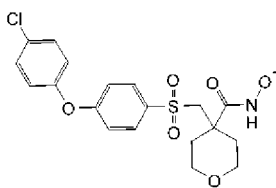
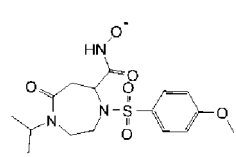
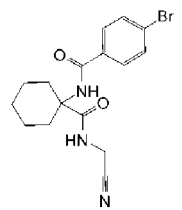
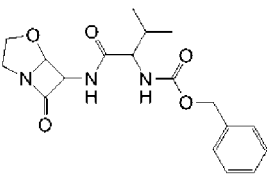
**2.1. Selectivity Sets and Background Database.** The 18 selectivity sets used in this study have previously been described in detail,<sup>9</sup> and their composition is summarized in Table 1. Selectivity set compounds target 13 proteins belonging to four families (carbonic anhydrases, matrix metalloproteases, papain-like proteases, and chymotrypsin-like proteases). Each set contains between 10 and 31 compounds that are selective for one target over a closely related family member as well as between 10 and 35 compounds that are non-selective for these two targets. Molecules are organized into sets according to their potency against the first target of the 18 target pairs that were considered, with all the molecules within a given set having a pK<sub>i</sub> (or pIC<sub>50</sub>) value of 6–7, 7–8, or 8–9. Thus, selective and non-selective compounds within each set only differ in their potency against the second target, which is reflected by their compound selectivity ratios: molecules are considered selective if they have an at least 50-fold higher K<sub>i</sub> (or IC<sub>50</sub>) value for the second than for the first target (for which they are selective), whereas molecules are considered non-selective if the K<sub>i</sub> (or IC<sub>50</sub>) ratio lies between 0.1–10. Representative structures of selective and non-selective

compounds are shown in Table 2, together with their selectivity ratios. For each compound pair, the Tanimoto coefficient (Tc) value<sup>6</sup> based on MACCS keys<sup>18</sup> is reported. The comparison shows that selective and non-selective molecules of a target pair can have substantially varying degrees of structural similarity, ranging from very similar to structurally distinct compounds. As described in detail in their original publication,<sup>9</sup> these compound selectivity sets represent very complex structure–selectivity relationships where highly similar compounds including compounds with identical scaffolds have different selectivity, and on the other hand, many different scaffolds represent both selective and non-selective compounds. Thus, for computational selectivity analysis, these compound sets present challenging test cases. Because there are no simple structural rules or features that can be recognized to correlate with compound selectivity, these compound data sets are also particularly suitable test cases for machine learning. In Figure 1, exemplary distributions of pairwise MACCS Tc similarity values within the subset of selective (intra-class) and between selective and non-selective (inter-class) molecules are reported for data sets CA IX/I (8–9) and Cat S/L (8–9). Especially Figure 1B (set CA IX/I (8–9)) emphasizes that pairs of selective and non-selective but also pairs of selective compounds can be structurally diverse.

As background database for selectivity searching, the MDL Drug Data Report<sup>19</sup> (MDDR) was used because MDDR compounds are in general drug-like and hence provide a more challenging search scenario than, for example, randomly collected synthetic molecules. Known inhibitors of the 13 investigated target proteins as well as inhibitors of closely related targets were removed, resulting in a reduced MDDR version of 152 337 molecules.

**2.2. Support Vector Machines.** The standard SVM approach<sup>12,13</sup> applied to binary classification problems utilizes a set of training data {**x**<sub>*i*</sub>, *y*<sub>*i*</sub>} (*i* = 1, ..., *n*) with **x**<sub>*i*</sub> ∈ ℝ<sup>*d*</sup> being the fingerprint representation (bit vector) of a molecule *i* and *y*<sub>*i*</sub> ∈ {−1, +1} being its class label (positive or negative) to derive a hyperplane *H* that best separates positive from

**Table 2.** Exemplary Compounds from Selectivity Sets)<sup>a</sup>

data set (potency range)	selective molecule: structure + selectivity ratio	non-selective molecule: structure + selectivity ratio	similarity (MACCS Tc)
CA IV/I (6-7)	 50.00	 1.58	0.82
Try/Thr (7-8)	 636.36	 0.35	0.64
MMP 8/1 (8-9)	 444.44	 5.95	0.57
Cat K/S (7-8)	 555.56	 2.60	0.40

<sup>a</sup> Exemplary pairs of selective and non-selective compounds with decreasing (pairwise MACCS Tc) structural similarity are shown with their selectivity ratios.

negative training examples. The hyperplane  $H$  is defined by a normal vector  $\mathbf{w}$  (with Euclidean norm  $\|\mathbf{w}\|$ ) and a scalar  $b$  (called bias) so that

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

where  $\langle \mathbf{x}, \mathbf{w} \rangle$  defines a scalar product. If the training data are linearly separable, an infinite number of such hyperplanes exist and the principal goal of SVM is to determine the one that maximizes the distance (called *margin*) from the nearest training data points to optimize the generalization performance of the classifier. Without loss of generality, the condition for correct classification can be formulated as the following set of inequalities

$$y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 \quad \forall i \quad (2)$$

These imply that the distance from  $H$  to the closest positive and negative training examples is  $1/\|\mathbf{w}\|$  in each case. Hence the maximum-margin hyperplane  $H$  is the one that maximizes  $1/\|\mathbf{w}\|$  or, correspondingly, minimizes  $\|\mathbf{w}\|$  subject to conditions 2.

If the training examples are not linearly separable, the minimization problem has no solution. To overcome this problem, positive *slack variables*  $\xi_i$  ( $i = 1, \dots, n$ ) are

introduced into eq 2, thereby relaxing the condition to permit some training examples to lie either within the margin or even on the incorrect side of  $H$ . However, these classification errors are penalized by their distance from the margin resulting in the following minimization problem, which utilizes a parameter  $C > 0$  to trade off training errors against margin size

$$\text{minimize: } V(\mathbf{w}, \xi) = \|\mathbf{w}\|^2 + C \sum_i \xi_i \quad (3)$$

$$\text{subject to } y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 - \xi_i \text{ with } \xi_i \geq 0 \quad \forall i \quad (4)$$

The minimization problem (3 and 4) can be reformulated into a convex quadratic programming problem applying the technique of Lagrange multipliers<sup>12</sup>

$$\text{maximize } L_D = \sum_i \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \quad (5)$$

$$\text{subject to } \sum_i \alpha_i y_i = 0 \text{ with } 0 \leq \alpha_i \leq C \quad \forall i \quad (6)$$

The solution vector  $\alpha$  uniquely determines the normal vector  $\mathbf{w}$  of  $H$  as a (usually sparse) linear combination  $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_i$ . Here, training examples contributing a nonzero  $\alpha_i$  value to

$\mathbf{w}$  are called *support vectors*. The decision function for the classification of a test molecule given its fingerprint representation  $\mathbf{x}$  then becomes

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

that is, compounds with  $f(\mathbf{x}) = 1$  are assigned to the positive class, those with  $f(\mathbf{x}) = -1$  to the negative class.

The SVM method has become particularly popular because it can easily be generalized to cases where the classification function does not linearly depend on the training data, which often proves to be more appropriate. This is accomplished by applying the so-called *kernel trick*<sup>20</sup> that replaces the scalar products in equations 5 and 7 by a kernel function  $K(\mathbf{x}_1, \mathbf{x}_2)$ . Conceptually, kernel functions correspond to a mapping  $\Phi: \mathbf{R}^d \rightarrow \mathcal{H}$  of the original vectors into a high-dimensional space  $\mathcal{H}$  (such that  $K(\mathbf{x}_1, \mathbf{x}_2) = \langle \Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2) \rangle$ ) in which a linear separation of the training examples might

be feasible. In this study, four different kernel functions were considered: the *linear kernel* corresponding to the standard scalar product (eq 8), the *Gaussian kernel*, also known as radial basis function kernel (eq 9), the *polynomial kernel* (eq 10), and the *Tanimoto kernel* (eq 11).

$$K_{\text{linear}}(\mathbf{x}_1, \mathbf{x}_2) = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle \quad (8)$$

$$K_{\text{Gaussian}}(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\gamma \|\mathbf{x}_1 - \mathbf{x}_2\|^2) \quad (9)$$

$$K_{\text{polynomial}}(\mathbf{x}_1, \mathbf{x}_2) = (\langle \mathbf{x}_1, \mathbf{x}_2 \rangle + 1)^d \quad (10)$$

$$K_{\text{Tanimoto}}(\mathbf{x}_1, \mathbf{x}_2) = \frac{\langle \mathbf{x}_1, \mathbf{x}_2 \rangle}{\langle \mathbf{x}_1, \mathbf{x}_1 \rangle + \langle \mathbf{x}_2, \mathbf{x}_2 \rangle - \langle \mathbf{x}_1, \mathbf{x}_2 \rangle} \quad (11)$$

### 2.3. SVM-Based Ranking Strategies for Selectivity Searching.

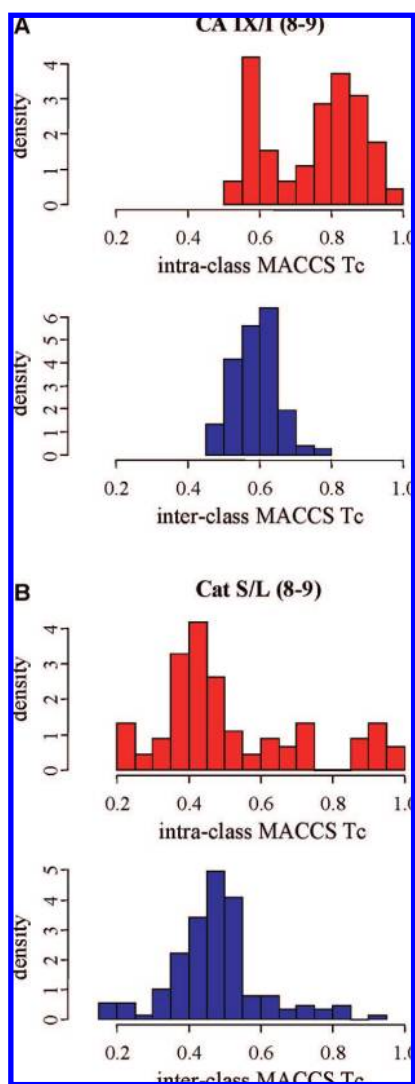
**2.3.1. Simple Ranking.** Originally, the SVM method was introduced for binary classification problems, that is, test data are assigned to a positive or negative class. However, an intuitive way of transforming SVM classification into a ranking approach is defining the rank of an object  $\mathbf{x}$  via the (signed) distance of its embedding  $\Phi(\mathbf{x})$  from the hyperplane determined in  $\mathcal{H}$ . Since the mapping  $\Phi$  is usually not explicitly known, which is actually the key aspect of the kernel trick, the function  $g: \mathbf{R}^d \rightarrow \mathbf{R}$  can be used producing an equivalent ranking

$$g(\mathbf{x}) = \sum_i \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) \quad (12)$$

The first ranking strategy considered in this work and referred to as *simple ranking* (see Figure 2A) sorts test molecules according to the value of function (12) after SVM training with selective (class label +1) versus inactive (class label -1) template molecules. By doing so, compounds located in the positive half-space in  $\mathcal{H}$  are ranked by decreasing distances from  $H$ , followed by compounds in the negative half-space ranked by increasing distances from  $H$ . Thus, we assume compounds in the positive half-space that are most distant from  $H$  to have the highest potential of selectivity.

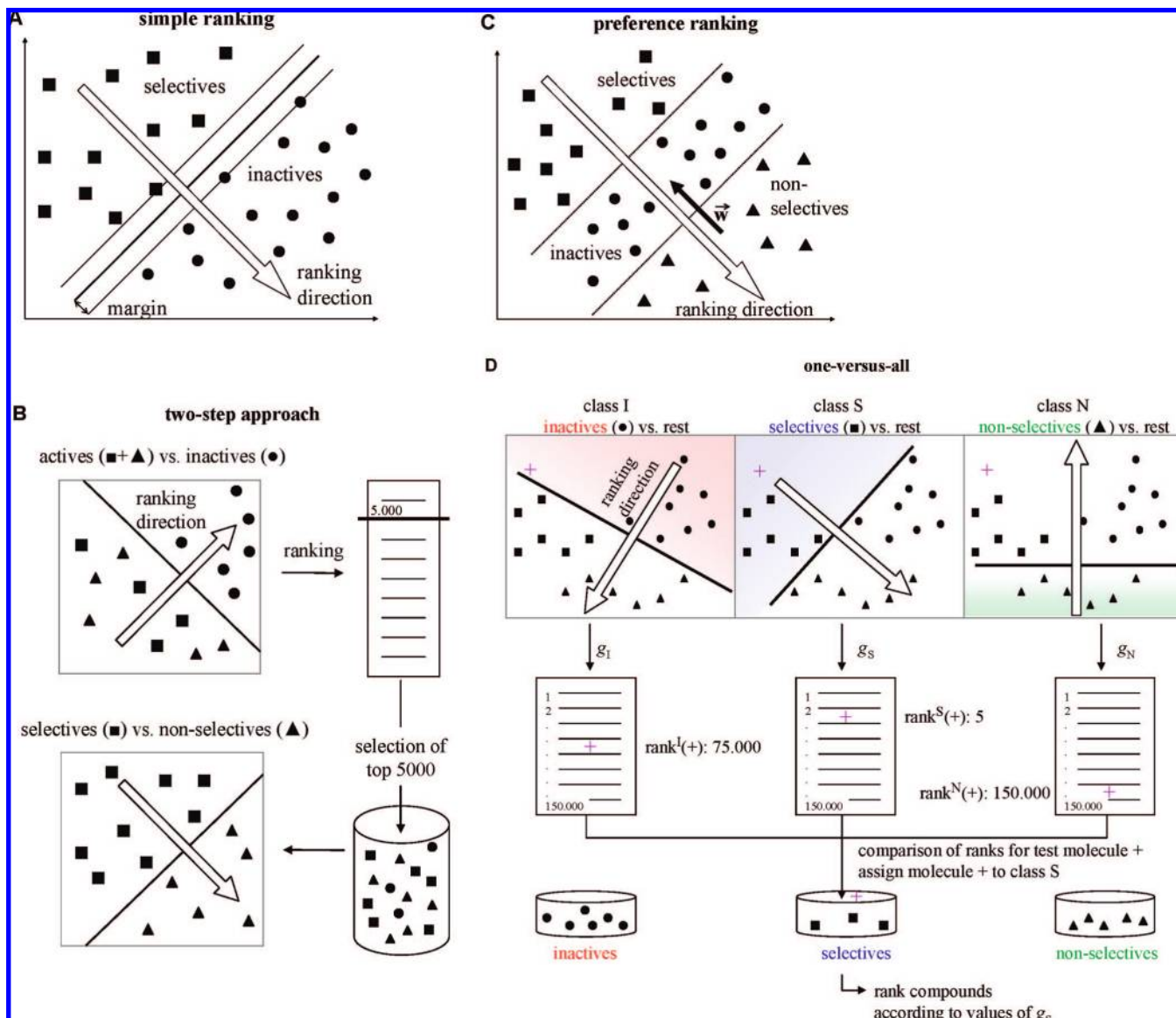
**2.3.2. Two-Step Approach.** The two-step approach was introduced by Stumpfe et al.<sup>9</sup> and is adapted for SVM herein, as illustrated in Figure 2B. In the first step, an SVM is trained using the union of selective and non-selective (i.e., active) reference molecules as positive training set and the inactive reference compounds as negative set to obtain a ranking function  $g_1$  (similar to eq 12). Test compounds  $i$  are ranked according to  $g_1(\mathbf{x}_i)$  and the top 5000 molecules are selected. In the second step, SVM training is repeated to derive a ranking function  $g_2$  by utilizing the selective molecules as positive training examples and assigning the non-selective compounds to the negative class (i.e., inactive reference molecules are not included in this step). Then, the preselected 5000 test molecules are reranked by their values of function  $g_2$ .

**2.3.3. Preference Ranking.** Preference ranking was originally introduced to optimize the retrieval quality of search engines.<sup>21</sup> The underlying idea is to derive a ranking function from pairwise preference constraints (e.g., “document  $d_i$  should rank higher than document  $d_j$ ”) that can be derived from the query-log of search engines. Joachims demonstrated that preference ranking results in an optimization problem that is equivalent to that of an SVM.<sup>21</sup> To adapt preference



**Figure 1.** Exemplary MACCS Tc similarity analysis. MACCS Tc value distributions are shown for selectivity sets (A) CA IX/I (8–9) and (B) Cat S/L (8–9). Intra-class means that the Tc distribution was determined from all pairwise compound comparisons within the selective compound subset, whereas inter-class refers to the distribution obtained by pairwise comparison of each molecule of the selective subset with each molecule of the non-selective subset.





**Figure 2.** SVM-based ranking strategies. (A) In the simple ranking strategy, selective and inactive training molecules are separated by the maximum margin hyperplane and test compounds are ranked by their signed distance from that hyperplane (indicated by the arrow). (B) In the two-step approach, first simple SVM-based ranking is applied after learning with the union of selective and non-selective (i.e., active) versus inactive training molecules. Test compounds are ranked in the direction of the “active” to the “inactive” half-space (indicated by the arrow) and only the 5000 top-ranked molecules are retained. In the second step, these preselected compounds are resorted, this time using simple SVM-based ranking after training with selective versus non-selective molecules. (C) SVM preference ranking utilizes the three training classes simultaneously to derive a weighting vector  $\mathbf{w}$  perpendicular to the two parallel hyperplanes that best separate selective from inactive and inactive from non-selective training molecules. Then, test compounds are ranked along the opposite direction of  $\mathbf{w}$ . (D) Following the one-versus-all strategy, the multiclass problem is reduced to three binary problems, each of which derives the hyperplane best separating one class from the union of the two others. Test compounds are subjected to each binary classification task and three separate rankings are produced. Then, for each test compound (+), its three positions in the individual rankings are compared and the compound is assigned to the class producing the best ranking position. Finally, only test molecules assigned to the selective class S are further considered and sorted by their initial ranking values for S.

ranking for selectivity searching, we formulate two preference constraints: (1) selective molecules should obtain a higher rank than non-selective and inactive molecules; (2) inactive compounds should rank higher than non-selective compounds to maximize purity. These conditions are formalized in the binary relation

$$R = (S \times N) \cup (S \times I) \cup (I \times N) \quad (13)$$

with S, N, and I representing the class of selective, non-selective, and inactive training molecules, respectively. Now we aim to derive a linear ranking function  $g$  represented by

a weight vector  $\mathbf{w}$  such that all pairwise preference constraints defined by  $R$  are met

$$\forall (i, j) \in R: \langle \mathbf{x}_i, \mathbf{w} \rangle > \langle \mathbf{x}_j, \mathbf{w} \rangle \quad (14)$$

As visualized in Figure 2C, this corresponds to constructing a normal vector  $\mathbf{w}$  for a set of parallel hyperplanes dividing the vector space into ordered layers of selective, inactive, and non-selective molecules. If a perfect separation of the training classes is not possible, slack variables  $\xi_{i,j}$  are introduced in analogy to the binary classification problem described above. Adding SVM regularization for

margin maximization leads to the following convex optimization problem (corresponding to eqs 3 and 4) that can be solved similarly to the one of the SVM classification problem:<sup>21</sup>

$$\text{minimize } V(\mathbf{w}, \xi) = \|\mathbf{w}\|^2 + C \sum_{i,j} \xi_{i,j} \quad (15)$$

$$\text{subject to } \langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{w} \rangle \geq 1 - \xi_{i,j} \text{ with } \xi_{i,j} \geq 0 \\ \forall (i,j) \in R \quad (16)$$

The resulting weight vector  $\mathbf{w}$  can be represented as a linear combination  $\mathbf{w} = \sum_i \alpha_i \mathbf{x}_i$  of the training vectors  $\mathbf{x}_i$  (with  $\alpha_i$  being derived from the solution of the Lagrangian reformulation of the problem), which leads to a ranking function of the form

$$g(\mathbf{x}) = \sum_i \alpha_i \langle \mathbf{x}_i, \mathbf{x} \rangle \quad (17)$$

By replacing scalar products by kernels, nonlinear ranking functions can also be obtained (as discussed in section 2.2).

**2.3.4. One-Versus-All.** The *one-versus-all* (or one-against-all) strategy is a common strategy to solve a multiclass problem with a binary classification method. In this study, for the selective (S), non-selective (N), and inactive (I) training classes, three individual “one-versus-rest” SVM-based ranking functions  $g_S$ ,  $g_N$ ,  $g_I$  are derived by training with classes S, N, or I versus the union  $N \cup I$ ,  $S \cup I$ , or  $S \cup N$ , respectively, as illustrated in Figure 2D. Test molecules are ranked three times according to their values of each individual function  $g_S$ ,  $g_N$ , and  $g_I$ . Let  $\text{rank}^X(j)$  be the ranking position obtained for a molecule  $j$  in the ranking for class  $X$  ( $X \in \{S, N, I\}$ ), then  $j$  is assigned to class

$$\arg \min_{X \in \{S, N, I\}} \text{rank}^X(j) \quad (18)$$

To obtain a final (partial) ranking, only test molecules assigned to class S (selective) are further considered. These molecules are again sorted by their values of  $g_S$ .

**2.4. Calculations.** The performance of the four SVM-based ranking strategies and alternative kernel functions (linear, Gaussian, polynomial, and Tanimoto) was evaluated in systematic selectivity search calculations on 18 selectivity sets using four different 2D fingerprint types that vary in their design and complexity: MACCS structural keys,<sup>18</sup> TGD,<sup>22,23</sup> GpiDAPH3,<sup>23</sup> and Molprint2D.<sup>11,24</sup> TGD and GpiDAPH3 represent two- and three-point pharmacophore-type fingerprints,<sup>25</sup> respectively, which are calculated, however, from the 2D connectivity table of a molecule. Molprint2D is a circular atom environment fingerprint that stores compounds as sets of strings only implicitly defining bit vectors, whereas MACCS, TGD, and GpiDAPH3 produce explicit bit vector representations in which molecular features correspond to predefined fingerprint bit position. To transform a Molprint2D string set into a bit vector, all different atom environments occurring in our screening database and selectivity sets were determined, and the resulting ~50 000 features were assigned to unique bit positions.

The combination of a ranking strategy, kernel function, and fingerprint type resulted in 64 different settings for selectivity searching. These 64 calculations were repeated

**Table 3.** Summary of the SVM Selectivity Search Set-Up

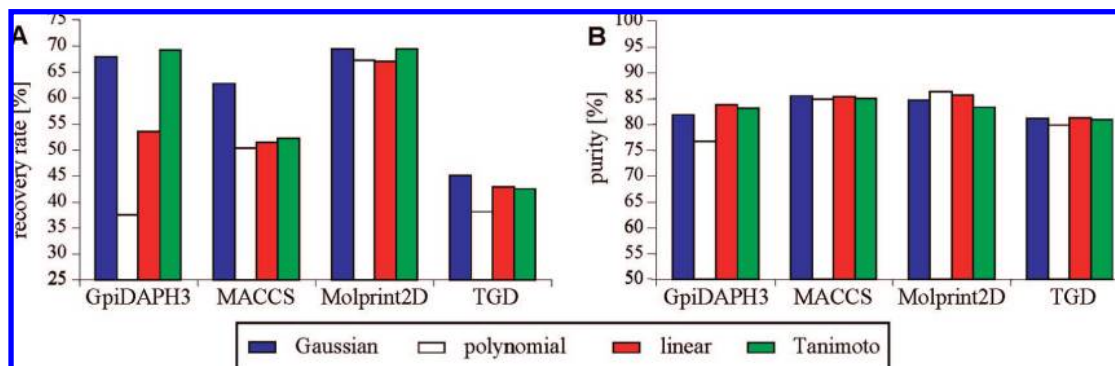
	description
selectivity sets	18 pairs of selective and non-selective compound sets
screening database	MDDR, 152 337 compounds
training molecules	100 inactive compounds, 5–17 selective or non-selective compounds (half of each set)
search strategies	simple ranking, 2-step approach, preference ranking, 1-versus-all
SVM kernels	Gaussian, linear, polynomial, Tanimoto
fingerprints	GpiDAPH3, MACCS, Molprint2D, TGD
trials	10 with different randomly selected compound reference sets

for all 18 selectivity sets using 10 different randomly selected training and test compound sets (i.e., in total 180 trials per combination). Each time, half of the molecules from a selective and non-selective subset served as training molecules, while the other half was added to the MDDR as potential database hits. Furthermore, 100 compounds randomly taken from the MDDR were used as (assumed) inactive training compounds, while the remainder of the MDDR served as the inactive test set. As performance measures in our calculations, recovery rates (RR, number of recovered selective molecules divided by the number of hidden selective test molecules), purity (P, number of recovered selective molecules divided by the number of recovered active (i.e., selective and non-selective) molecules), and the geometric mean of both ( $\text{RRP} = \sqrt{\text{RR} \times \text{P}}$ ) were determined for selection sets of increasing sizes and averaged over all trials. Average values were compared for the 64 different calculation settings and also to the performance of  $k$ -nearest neighbor searching ( $k$ -NN),<sup>10</sup> the so far best strategy reported for selectivity searching.<sup>9</sup> Following the  $k$ -NN search strategy, pairwise Tc similarity values are determined between a database and all selective reference compounds and the average over the  $k$  highest values is used as final score.

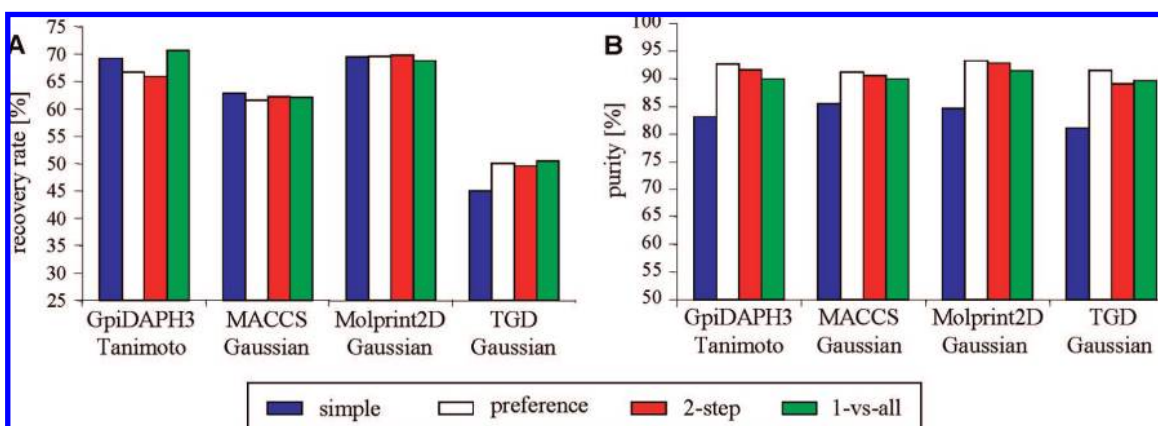
The SVM-based ranking strategies presented in this study were conducted using SVM<sup>light</sup>, a freely available SVM implementation,<sup>26,27</sup> to determine the solution of the convex quadratic programming problem and in-house Perl scripts were used to carry out the two-step and one-versus-all approaches. With the exception of the kernel parameters  $d$  and  $\gamma$  (see equations 9 and 10) that were obtained from test calculations, all other parameters were SVM<sup>light</sup> default settings to make our calculations easily reproducible. The calculation setup of our study is summarized in Table 3.

### 3. RESULTS AND DISCUSSION

In this analysis, we have focused on comparing different SVM ranking methods and kernel functions for their ability to distinguish selective from non-selective and inactive compounds in database mining. Support vector machine classifications can not be interpreted at the level of chemical modifications that render compounds selective. Their major goal is the optimization of recall rates of selective compounds and the purity of database selection sets. In previous investigations, we have also studied Bayesian models for selectivity searching using fingerprint descriptors.<sup>9</sup> Bayes classification assigns statistical weights to individual bit positions. As to whether highly weighted bit positions can



**Figure 3.** Comparison of different kernel functions. For each fingerprint, the (A) recovery rate and (B) purity obtained for the simple ranking strategy in combination with the Gaussian, polynomial, linear, or Tanimoto kernel are reported for a selection set size of 100 compounds. Results are averaged over all selectivity sets and search trials.



**Figure 4.** Comparison of different ranking strategies. For each fingerprint, the (A) recovery rate and (B) purity for the simple, preference, two-step, or one-versus-all ranking strategy in combination with the best-performing kernel are shown for a selection set size of 100 compounds. Results are averaged over all selectivity sets and search trials.

potentially be associated with chemical features conferring compound selectivity depends on the chosen fingerprint descriptors and the nature of highly weighted features.

In our current study, fingerprints of different design were applied to make the results of the calculations independent of the specifics of a chosen descriptor set. Specifically, we have aimed to identify a search strategy that significantly enriches database selection sets with selective compounds (i.e., achieving high recall) and deprioritizes non-selective molecules (i.e., high purity). For selectivity searching, compound recall and purity were considered equally important criteria.

**3.1. Overall Selectivity Search Performance.** Results of our systematic selectivity search calculations are summarized in Table 4 that reports average recovery rates and purities for each combination of a kernel function and ranking strategy for selection set sizes of 100 compounds. To determine the overall best compromise between compound recall and purity, the geometric mean of both (RRP) was also calculated and included in the comparison. Initially, we examined the search performance of the combination of the linear kernel function and simple SVM ranking (linear/simple), which could be considered as a standard SVM strategy. For different fingerprint designs, the linear/simple combination produced recovery rates of 43–67% and purities of 81–86%, with Molprint2D and TGD leading to highest and lowest values, respectively. Thus, the linear/simple combination already produced on average promising search results, which we considered an encouraging finding.

Next, we studied the effects of using alternative kernel functions and tried to identify the best kernel for each fingerprint type. Figure 3 graphically illustrates changes in selectivity search performance when different kernels were applied in combination with simple ranking. As shown in Figure 3A, the Gaussian kernel produced overall highest recovery rates, with a substantial increase of  $\Delta 14\%$  for GpiDAPH3 and  $\Delta 11\%$  for MACCS when compared to the linear kernel. For GpiDAPH3, the Tanimoto kernel performed slightly better than the Gaussian kernel. By contrast, the polynomial kernel yielded overall lowest recall of selective compounds. Figure 3B reports the effects of different kernel functions on the purity of database selection sets. Different from recovery rates, no significant changes in purity were observed. When using RRP as a quality criterion (Table 4), the Gaussian kernel performed best for MACCS, Molprint2D, and TGD and the Tanimoto kernel for GpiDAPH3.

On the basis of these findings, we next studied the effects of using alternative ranking strategies. Figure 4 reports virtual screening results when combining the fingerprint-dependent preferred kernel with the four different SVM-based ranking approaches. As shown in Figure 4A, no systematic improvement in recovery over simple ranking could be detected when alternative ranking schemes were applied. By contrast, a clear improvement in purity was achieved, as shown in Figure 4B. Replacing simple ranking with any other ranking strategy systematically increased purity to values exceeding 90%. For GpiDAPH3, Molprint2D, and TGD, this corresponded to



**Table 4.** Compound Recall in SVM Selectivity Searching<sup>a</sup>

fingerprint	kernel	strategy	RR [%]	P [%]	RRP
GpiDAPH3	Gaussian	simple	67.9	81.8	74.6
		preference	64.4	91.9	76.9
		2-step	55.5	91.6	71.3
		1-vs-all	68.6	88.7	78.0
	polynomial	simple	37.5	76.7	53.6
		preference	53.6	88.4	68.9
		2-step	38.9	80.7	56.1
		1-vs-all	35.3	81.2	53.5
	linear	simple	53.5	83.7	67.0
		preference	48.8	89.4	66.1
		2-step	52.5	90.6	69.0
		1-vs-all	56.6	90.0	71.4
	Tanimoto	simple	69.2	83.1	75.8
		preference	66.7	92.6	78.6
		2-step	65.9	91.5	77.7
		1-vs-all	70.6	90.0	79.7
MACCS	Gaussian	simple	62.8	85.5	73.3
		preference	61.6	91.2	74.9
		2-step	62.2	90.6	75.1
		1-vs-all	62.0	90.0	74.7
	polynomial	simple	50.4	84.9	65.4
		preference	41.5	88.7	60.7
		2-step	35.8	82.1	54.2
		1-vs-all	48.0	87.6	64.8
	linear	simple	51.5	85.3	66.3
		preference	6.1	29.4	13.4
		2-step	30.8	82.1	50.3
		1-vs-all	47.5	87.1	64.3
	Tanimoto	simple	52.3	85.1	66.7
		preference	9.7	26.9	16.1
		2-step	37.6	85.3	56.6
		1-vs-all	48.6	87.6	65.3
Molprint2D	Gaussian	simple	69.4	84.6	76.7
		preference	69.6	93.3	80.6
		2-step	69.8	92.9	80.5
		1-vs-all	68.8	91.4	79.3
	polynomial	simple	67.3	86.4	76.2
		preference	67.8	93.5	79.6
		2-step	66.7	92.5	78.5
		1-vs-all	68.0	91.2	78.7
	linear	simple	67.1	85.7	75.8
		preference	56.0	92.0	71.8
		2-step	59.0	94.4	74.7
		1-vs-all	67.3	91.0	78.3
	Tanimoto	simple	69.5	83.2	76.1
		preference	65.8	93.5	78.4
		2-step	64.3	93.6	77.6
		1-vs-all	68.5	89.8	78.5
TGD	Gaussian	simple	45.0	81.1	60.4
		preference	50.0	91.5	67.7
		2-step	49.6	89.1	66.5
		1-vs-all	50.5	89.6	67.3
	polynomial	simple	38.1	79.8	55.2
		preference	25.3	65.6	40.7
		2-step	19.1	54.2	32.2
		1-vs-all	37.6	82.2	55.6
	linear	simple	42.9	81.2	59.0
		preference	14.3	40.3	24.0
		2-step	24.4	60.9	38.6
		1-vs-all	40.4	84.4	58.4
	Tanimoto	simple	42.6	80.9	58.7
		preference	13.2	34.6	21.4
		2-step	31.9	74.2	48.7
		1-vs-all	42.4	86.4	60.5

<sup>a</sup> Recovery rates (RR in %), purity (P in %), and combined recovery and purity (RRP) values are reported for systematic selectivity search calculations using a database selection set size of 100 molecules. Values are averaged over all 18 selectivity sets and 10 trials per set. Under “strategy”, “simple” stands for simple ranking approach, “preference” for preference ranking, “2-step” for the two-step approach, and “1-vs-all” for the one-versus-all strategy.

Δ10% improvement over simple ranking. Although observed differences between preference, two-step, and one-versus-all ranking were small, preference ranking consistently produced highest purity. Thus, taken together, our results indicated that an increase in recovery could basically be attributed to a suitable change in the kernel function with the Gaussian kernel generally representing a good choice, whereas purity gained from the application of more sophisticated ranking strategies (especially preference ranking). The second observation can be well rationalized by considering that preference, two-step, and one-versus-all ranking make use of additional structural information by taking non-selective compounds into account during the learning phase, in contrast to simple ranking.

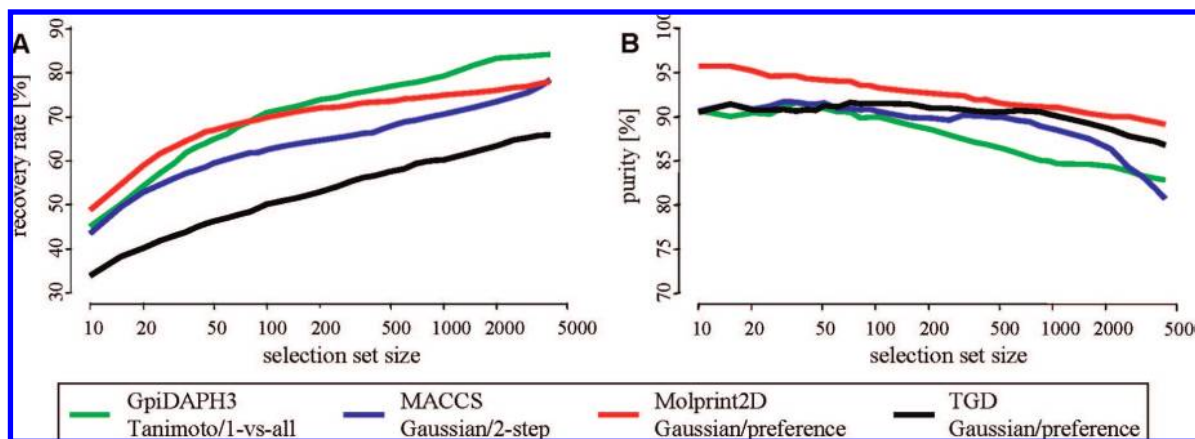
Differences in the relative performance of alternative fingerprint descriptors are reported in Figure 5. For each of the four fingerprint types, the best performing combination of a kernel function and ranking method was selected on the basis of largest RRP (Table 4), and for these combinations, recovery rates and purities are shown for compound selection sets of increasing size. These preferred combinations were Tanimoto/one-versus-all for GpiDAPH3, Gaussian/two-step for MACCS, Gaussian/preference for Molprint2D, and Gaussian/preference for TGD. As shown in Figure 5A, Molprint2D and GpiDAPH3 produced overall highest recall, followed by MACCS and TGD (that consistently had Δ5–10% and Δ15–20% lower recovery rates, respectively). Figure 5B shows that purity values only differed by Δ5% for the alternative fingerprint designs, with Molprint2D performing best.

**3.2. Set Dependence of SVM Performance.** When considering the search performance for individual selectivity sets, in part significant fluctuations were observed. For instance, using the linear/simple/MACCS combination, recovery rates of 20–100% and purities of 58–100% for selection sets of 100 molecules were obtained for the different selectivity sets. In general, recovery rates correlated with intra-class structural similarity (assessed using MACCS Tc), yielding a Pearson correlation coefficient of 0.76. For example, nearly 83% compound recall was achieved for selectivity set CA IX/I (8–9), shown in Figure 1A, but only 41% for Cat S/L (8–9), shown in Figure 1B, which had higher intra-class structural diversity.

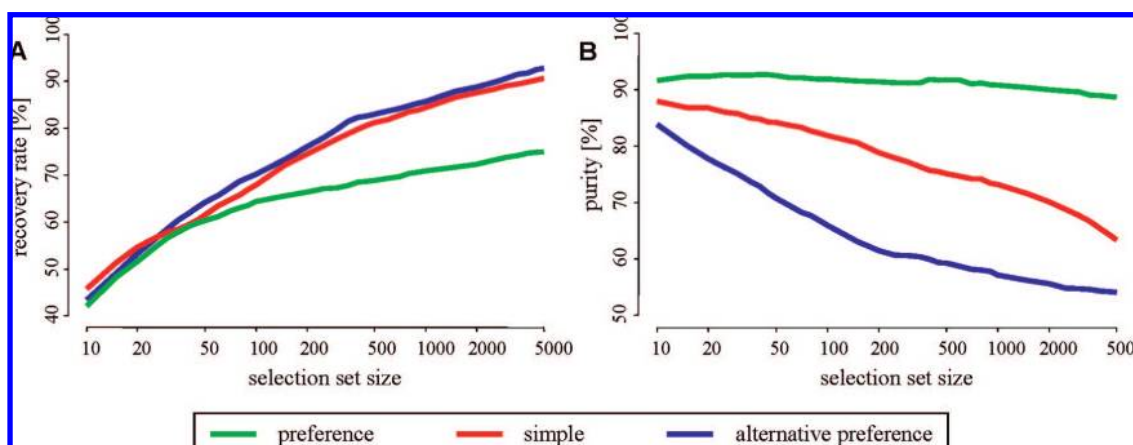
The relationship between purity rates and structural similarity of selective and non-selective compounds was more complex. A possible and intuitive explanation for high purity values would be the presence of significant differences between intra-class (within the selective subset) and inter-class (selective subset vs non-selective subset) structural diversity. This view is consistent with purity of 100% obtained for CA IX/I (8–9) for which the intra-class MACCS Tc distribution was notably shifted to higher similarity values than the inter-class distribution (see Figure 1A). However, for Cat S/L (8–9), a comparable purity of 98% was observed, although the intra- and inter-class similarity distributions strongly overlapped, as shown in Figure 1B. This figure indicates the presence of variable and complex structure-selectivity relationships that were successfully handled by SVM ranking.

**3.3. Characteristic Features of Ranking Methods.** In addition to the high selectivity search performance of the preferred kernel/ranking/fingerprint combinations described





**Figure 5.** Comparison of preferred kernel/ranking approach combinations. The diagram shows the (A) recovery rate and (B) purity achieved using the best combination of a ranking strategy and kernel function for each fingerprint. Rates are reported for selection sets of increasing size. Results are averaged over all selectivity sets and search trials.

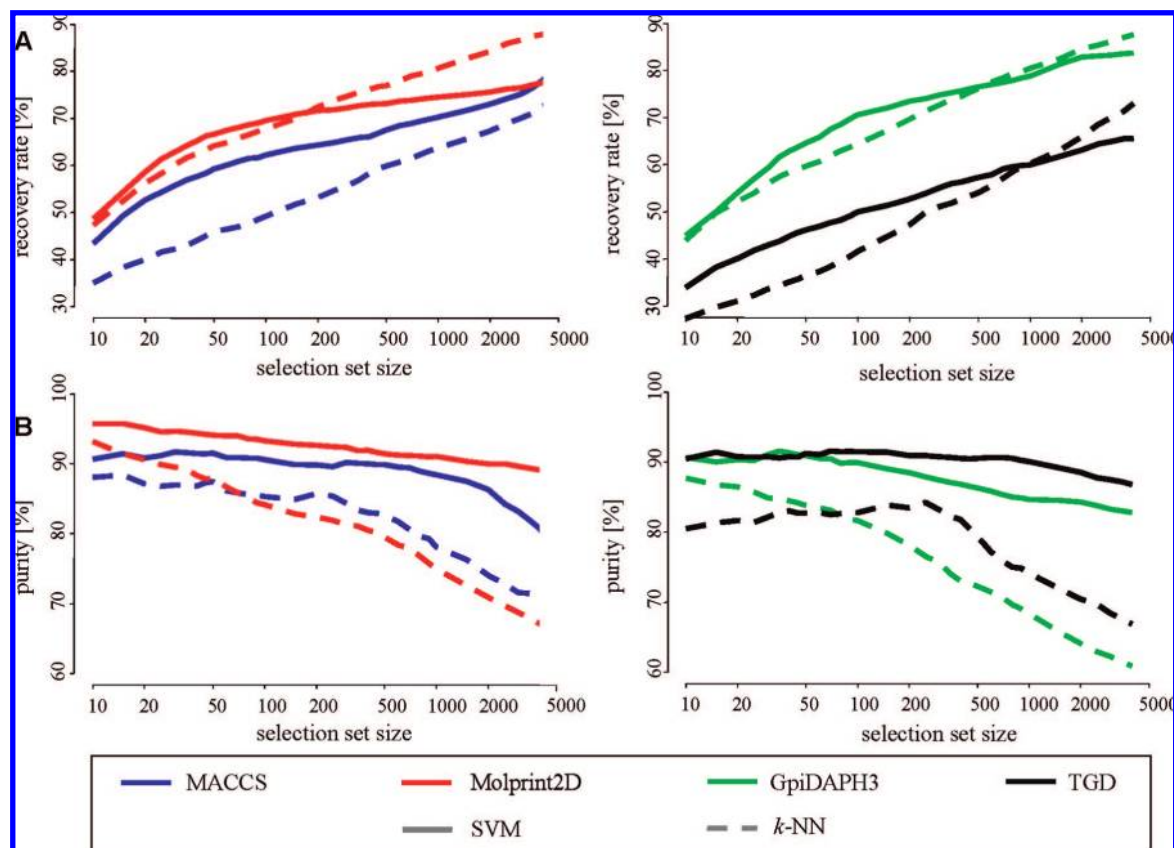


**Figure 6.** Comparison of alternative strategies for preference ranking. Two different strategies for preference ranking are compared to simple ranking. (A) Recovery rates and (B) purities are reported for selection sets of increasing size. “Preference” stands for preference ranking according to Figure 2C, that is, test compounds are sorted in the order *selective–inactive–non-selective*. By contrast, in “alternative preference” ranking, test compounds are sorted in the order *selective–non-selective–inactive*.

above, we also found combinations that led to substantial variations. As reported in Table 4, preference ranking displayed a dramatic loss in recovery rates when combined with kernel functions operating in comparably low-dimensional feature space, that is, the Tanimoto or linear kernel, and low-complexity fingerprints, i.e. TGD or MACCS. A similar but less significant trend was observed for the two-step method, whereas the one-versus-all approach was the overall most “stable” search strategy. This observation might be due to the fact that the one-versus-all method, different from two-step or preference ranking, does not attempt to move non-selective compounds to the end of the final compound list after learning, which could lead to a misclassification of selective database compounds that are structurally very similar to non-selective training molecules. Because the priority ordering of compound classes can easily be adjusted in preference ranking, we tried to evaluate the potential role of the mentioned misclassification effect by investigating an alternative class order (selective–non-selective–inactive) instead of the originally applied order (i.e., selective–inactive–non-selective). Indeed, we observed a consistent increase in recall but at the cost of an in part dramatic loss in purity. Figure 6 shows exemplary results for modified preference ranking in combination with the Gaussian kernel and the GpiDAPH3 fingerprint. As can be

seen, the alternative class order led to a moderate increase in recovery rates of  $\Delta 1$ –10% for selection set sizes up to 200 compounds, whereas purity was reduced by  $\Delta 10$ –30%. Therefore, to achieve a meaningful compromise between recall and purity, the original class order (i.e., selective–inactive–non-selective) was clearly preferred for SVM preference ranking.

**3.4. Comparison of SVM Ranking to Conventional Similarity Searching.** To put the performance of the SVM search strategies evaluated herein into context, the results were compared to previously reported *k*-NN similarity searching using selective reference molecules.<sup>9</sup> Figure 7 shows a cumulative recall and purity comparison of the best performing SVM-based ranking approach and *k*-NN similarity searching for selection sets of 10–5000 molecules. The *k*-NN search results for the MACCS and Molprint2D fingerprints were taken from the original publication, while those for TGD and GpiDAPH3 were calculated using equivalent parameter settings. For MACCS, a consistent increase in recovery rate ( $\sim \Delta 8$ –14%) and purity ( $\sim \Delta 5$ –10%) was observed for SVM-based ranking. For the other three fingerprints, SVM ranking only slightly increased compound recall (by max  $\Delta 10$ %) for selection set sizes of up to about 150 compounds for Molprint2D, 500 for GpiDAPH3, and 1000 for TGD. For larger selection set sizes, recovery rates



**Figure 7.** Comparison of SVM-based ranking and similarity searching. (A) Recovery rates and (B) purities are compared for the overall preferred search strategy identified by Stumpfe et al.<sup>9</sup> (i.e., *k*-NN, dashed lines) and the best combination of a SVM-based ranking approach and kernel function identified in our study (solid lines). Results are averaged over all selectivity sets and search trials.

became higher for *k*-NN (to a comparable extent). However, in accordance with our goal to effectively distinguish between selective and non-selective (but active) compounds, a significant increase in purity was achieved by SVM-based ranking. For small selection set sizes, SVM purity exceeded 90% and remained essentially constant (max decrease  $\Delta 5\%$ ) for selection sets of increasing size. By contrast, *k*-NN fingerprint searching resulted in lower purity values for small selection sets and, in addition, purity systematically decreased by about  $\Delta 25\%$  when proceeding to larger selection sets. Therefore, for selectivity searching, we clearly favor SVM-based ranking over *k*-NN search strategies, which in turn outperformed Bayesian models in our previous studies.<sup>9</sup>

#### 4. CONCLUSIONS

In this study, we have thoroughly investigated different SVM-based approaches to search for target-selective compounds and distinguish them from non-selective molecules and database decoys. Specifically, we have aimed at balancing recall of selective compounds and purity of selection sets of small size. Therefore, we have adapted the original binary classification SVM algorithm to solve a three-class ranking problem. This led to four different SVM-based ranking strategies that we termed *simple*, *preference*, *two-step*, and *one-versus-all* ranking. In exhaustive selectivity search trials involving four alternative kernel functions and four fingerprints of different design, we found preference but also two-step and one-versus-all ranking to consistently improve the ability of conventional (simple) SVM ranking and fingerprint similarity searching to deprioritize non-selective compounds

and thereby increase the purity of selection sets, because of the explicit inclusion of non-selective compounds as third training class. Furthermore, application of the Gaussian kernel function led to substantial increases in the recall of selective compounds over alternative kernels. Thus, taken together, our findings suggest that combining preference ranking and the Gaussian kernel function is of particular value for selectivity searching. Moreover, one-versus-all ranking, which is straightforward to implement, emerged as the overall most stable method for different kernel functions and fingerprints. Given the complexity of the structure–selectivity relationships represented by the compound selectivity sets investigated here, for practical purposes, accurate predictions of compound selectivity using our SVM approaches would be most useful early on during studies of individual targets or target families when little is known about compound selectivity determinants, and no structural rules governing selectivity relationships are available.

#### ACKNOWLEDGMENT

We wish to thank Thomas Gärtner and Martin Vogt for many helpful discussions and review of the manuscript. In addition, we like to thank Dagmar Stumpfe and Jens Auer for help with compound selectivity sets and SVM<sup>light</sup>, respectively.

#### REFERENCES AND NOTES

- (1) 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.

- (2) Spring, D. R. Chemical genetics to chemical genomics: small molecules offer big insights. *Chem. Soc. Rev.* **2005**, *34*, 472–482.
- (3) Bredel, M.; Jacoby, E. Chemogenomics: An emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, *5*, 262–275.
- (4) Stockwell, B. R. Exploring biology with small organic molecules. *Nature* **2004**, *432*, 846–854.
- (5) Bajorath, J. Computational approaches in chemogenomics and chemical biology: current and future impact on drug discovery. *Expert Opin. Drug Discovery* **2008**, *3*, 1371–1376.
- (6) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
- (7) 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.
- (8) 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.
- (9) 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.
- (10) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target protein. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
- (11) 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.
- (12) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowledge Discovery* **1998**, *2*, 121–167.
- (13) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.
- (14) 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.
- (15) 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.
- (16) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.
- (17) 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.
- (18) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.
- (19) *MDL Drug Data Report (MDDR)*; Symyx Software: San Ramon, CA, 2005.
- (20) 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.
- (21) Joachims, T. Optimizing search engines using clickthrough data. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Edmonton, Alberta, Canada, 2002; ACM: New York, 2002; pp 133–142.
- (22) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
- (23) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.
- (24) *MOLPRINT 2D*. URL for the publicly available molecular fingerprint: <http://www.molprint.com> (accessed 15 Jan 2008).
- (25) 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.
- (26) 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.
- (27) *SVM<sup>light</sup>*, version 4.00. <http://svmlight.joachims.org/> (accessed 15 Jan 2008).

CI800441C