Ligand-Based Virtual Screening by Novelty Detection with Self-Organizing Maps

Dimitar Hristozov,[†] Tudor I. Oprea,[‡] and Johann Gasteiger*[,†]

Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany, and Division of Biocomputing, University of New Mexico School of Medicine, MSC 11 6145, 1 University of New Mexico, Albuquerque, New Mexico 87131-0001

Received January 30, 2007

We describe a novel method for ligand-based virtual screening, based on utilizing Self-Organizing Maps (SOM) as a novelty detection device. Novelty detection (or one-class classification) refers to the attempt of identifying patterns that do not belong to the space covered by a given data set. In ligand-based virtual screening, chemical structures perceived as novel lie outside the known activity space and can therefore be discarded from further investigation. In this context, the concept of "novel structure" refers to a compound, which is unlikely to share the activity of the query structures. Compounds not perceived as "novel" are suspected to share the activity of the query structures. Nowadays, various databases contain active structures but access to compounds which have been found to be inactive in a biological assay is limited. This work addresses this problem via novelty detection, which does not require proven inactive compounds. The structures are described by spatial autocorrelation functions weighted by atomic physicochemical properties. Different methods for selecting a subset of targets from a larger set are discussed. A comparison with similarity search based on Daylight fingerprints followed by data fusion is presented. The two methods complement each other to a large extent. In a retrospective screening of the WOMBAT database novelty detection with SOM gave enrichment factors between 105 and 462—an improvement over the similarity search based on Daylight fingerprints between 25% and 100%, when the 100 top ranked structures were considered. Novelty detection with SOM is applicable (1) to improve the retrieval of potentially active compounds also in concert with other virtual screening methods; (2) as a library design tool for discarding a large number of compounds, which are unlikely to possess a given biological activity; and (3) for selecting a small number of potentially active compounds from a large data set.

1. INTRODUCTION

Virtual screening of databases is a popular method for selecting available compounds for biological assays. It involves the scoring of molecules in a database of chemical compounds in order of decreasing probability of biological activity. The result of this process is a ranked list in which the highest ranked compounds are assumed most likely to share the target's activity. In such a manner, potential hits may be acquired or synthesized and then tested in the early stages of a lead-discovery program. When the 3D structure of the biological target protein is available, structure-based approaches to virtual screening may be used.[1,2] However, when such information is not available, ligand-based approaches can be used, with similarity search[3] being the most common one. Similarity searching requires a known active structure, such as a hit from high-throughput screening, used as query. This query is compared with a set of potentially active compounds by means of a similarity measure.

The use of known active compounds as template to rank a set of compounds with unknown activity has been explored in a variety of studies.[4−6] Once the compounds from the screened set are compared to the query, the set is sorted using the similarity coefficients; the top-ranked compounds are expected to share the biological properties of the query molecule. Most commonly 2D fingerprints (as descriptors) and Tanimoto coefficients[3] (as similarity metric) are utilized, although other types of similarity measures[7,8] exist. Similarity measures are often continuous values and can be adapted to real-value representation. A number of studies have applied autocorrelation vectors,[5] reduced graphs,[9] or other real valued descriptors[10] to this problem.

The availability of published active compounds often presents an opportunity to use a set of active structures rather than a single query. Several reports[11−13] utilizing this information for similarity search exist. A common way for incorporating such multitarget data is via data fusion.[12] The fusion process usually improves the results, and Hert at el.[14] recently reported on additional improvements when the nearest nonactive neighbors are used as queries. Another possibility for utilizing data from multiple targets is to apply novelty detection techniques.

Novelty detection refers to the attempt of identifying patterns that do not belong to the space covered by a given set.[15−17] The set usually contains patterns for only one of the available classes, hence novelty detection techniques are sometimes called one-class classification. This approach is valuable in many cases where data about a given state of the system can be acquired easily, while the collection of data for the other states is difficult. Novelty detection is commonly applied for fault detection, network intrusion

* Corresponding author phone: (49)-9131-815668. E-mail: Gasteiger@molecular-networks.com. Present address: Molecular-Networks GmbH, Henkestrasse 91, 91052 Erlangen, Germany.
[†] Universität Erlangen-Nürnberg.
[‡] 1 University of New Mexico.

10.1021/ci700040r CCC: $37.00
Published on Web 09/14/2007

detection, hand written digit recognition, Internet and e-commerce transactions, statistical process control, etc.[16,17] Prior to virtual screening, one can often access different databases of active structures; however, access to compounds that have been found to be inactive in a biological assay is relatively difficult unless such results are reported or stored in online repositories (such as PubChem). Therefore, building two-class classification models that might distinguish actives from inactives is hindered, and a method that does not require proven inactive compounds is preferable.

In a typical novelty detection application, say, fault detection, there is usually much more data for the "normal" case. In such a scenario one is interested in the patterns predicted as novel (signaling fault conditions). However, putting the novelty detection into a virtual screening context the "normal" case is described using only known active compounds. Since the system has a knowledge of the active class alone it predicts as novel those compounds, which lie far enough from the chemical space set of the active compounds. Thus, the process is inverted in the sense that the accent is now on the compounds, predicted as known, that is, active. In addition, while the discovery of new active scaffolds and chemotypes is a desirable feature of any VS method, the term "novelty" throughout this text refers only to compounds lying sufficiently far from the chemical space of the given activity.

Novelty detection can be performed by a big variety of statistical and machine-learning methods. A comprehensive review covering both approaches has been presented by Marcou et al.[16,17] From the large variety of novelty detection techniques we selected one based on Self-Organizing Maps (SOM). SOM is a neural network model well-known for its applications to high-dimensional data analysis in many fields, including chemistry.[18,19] In spite of its well-documented use as a novelty detector in the field of machine learning,[20−22] and the continuous use of SOM for solving chemistry-related problems[23] including virtual screening,[24,25] to the best of our knowledge novelty detection with SOM has not been applied to chemical problems so far.

There are several reasons for the application of SOM as a novelty detector: (a) The method allows the ranking of the accepted structures according to their proximity to the chemical space covered by the trained SOM, besides the ability to detect novel patterns; this feature allows us to compare the results of this method with the results obtained using similarity search. (b) The size of the trained SOM is usually smaller than the size of the entire training set, which offers speed improvements. (c) The SOM is an unsupervised learning method, thus its application to such "one-class" problems is intuitive.

To build a good description of the active space a representative set of actives is needed. Clearly, the success of a retrospective virtual screening with SOM novelty detection cannot be measured on the same compounds, which were used to build the novelty detector. Thus, different methods for splitting the available sets of actives into a training and a test set were studied. Although not strictly needed for a similarity search with Daylight fingerprints, such a split might be useful when a representative set has to be selected in order to decrease the run time of the search.

The aims of this work were to (a) examine the applicability of SOM novelty detection to ligand-based virtual screening; (b) compare SOM novelty detection with the most commonly used similarity search with Daylight fingerprints; and (c) study the effect of different methods for subset selection on the results of a virtual screening experiment.

Table 1. Subsets of Inhibitors Used in This Study

target	abbreviation	N	μ^a	σ^a
acetylcholinesterase	AChE	568	0.433	0.168
cyclooxygenase 2	COX-2	999	0.345	0.142
3′,5′-cyclic-nucleotide phospho-diesterase	PDE4	457	0.430	0.161
thrombin	thrombin	2105	0.377	0.133
u-plasminogen activator	uPA	199	0.397	0.172

[a] μ and σ give the mean and SD of the intraclass similarities, obtained with Daylight fingerprints and Tanimoto coefficient

2. MATERIALS AND METHODS

2.1. Data Sets. *2.1.1. Overview.* Five sets of known inhibitors—568 acetylcholinesterase (AChE), 999 cyclooxygenase 2 (COX-2), 457 3′,5′-cyclic-nucleotide phosphodiesterase IV (PDE4), 2105 thrombin, and 199 u-plasminogen activator (uPA) inhibitors—were extracted from the WOMBAT (World Of Molecular BioAcTivity) database,[26] version 2006.1. WOMBAT is a target-annotated database available from Sunset Molecular Discovery (Santa Fe, NM). Release 2006.1 contains 154 236 compounds, collected from articles in medicinal chemistry journals published between 1975 and 2006. A reduced set of 135 877 unique chemical structures resulted after removing duplicate structures as well as those, for which some of the used physicochemical properties could not be calculated.

In addition to the structural information, WOMBAT also contains the reported activity values, expressed as pK_i value, information about the species in which the tests were performed, and the biological role of the structure (inhibitor, antagonist, etc.) as well as additional properties of interest. For all sets of actives used in this study only inhibitors for the corresponding enzymes tested in human and with reported activity less than 30 μm were selected. An overview of the selected active subsets is given in Table 1.

In addition to the number of compounds in each subset, the average self-similarity, μ, of the whole subset is given. This value was calculated by taking the pairwise Tanimoto similarity between all compounds in the set using the 1024-bit Daylight fingerprints and calculating the average value of all similarity scores. As can be seen, the inhibitors of AChE are the most self-similar group, having the highest value of μ, while the COX-2 set is the least self-similar, having the lowest μ value.

2.1.2. Training Set Selection. Three methods for splitting the available actives into a training and a test set were compared—random, semirandom, and Taylor—Butina[27,28] clustering. In all cases, half of the known actives were used as a training set and the other half as a test set. The test set was merged with the rest of the WOMBAT structures (up to 135 877), and the performance of the different methods was measured by their ability to retrieve the known active compounds from this set.

The semirandom method was based on the activity values, assigned to each structure. Each subset was sorted according to the reported activity and three groups of high (pK_i > 8),

medium ($6 < pK_i <= 8$), and low ($4.5 < pK_i <= 6$) activity were distinguished. Half of the structures from each of these groups were randomly picked as training, and the remaining were kept aside as a test set.

The Taylor−Butina clustering was first described by Taylor[27] and later by Butina.[28] It is an unsupervised nonhierarchical clustering method which guarantees that every cluster contains molecules which are within a distance cutoff of the central molecule. The similarity matrix obtained with Daylight fingerprints and Tanimoto coefficients for each of the active subsets was used as input, and a cutoff value of 0.8 was used. Half of the structures in each cluster were selected as a training set. Half of the singletons were randomly assigned to the training set as well. The clustering as well as data set splitting was carried out using the R environment.[29]

2.2. Chemical Structure Representation. *2.2.1. Binary Fingerprints (BFP).* 1024-dimensional unfolded Daylight fingerprints[30] were generated with the Chemical Descriptors Library.[31]

2.2.2. Topological Autocorrelation (AC2D). Introduced by Moreau et al.[32] the topological autocorrelation descriptors have since then been applied in a number of studies.[5,24,33] The descriptors are calculated according to eq 1

$$A(d) = \sum_{i=1}^{k} \sum_{j=i}^{k} p_i p_j \delta(d - d_{ij}) \qquad (1)$$

where k is the number of atoms in the molecule, p_i is some atomic property of atom i, d_{ij} is the topological distance (i.e., the number of bonds) between atoms i and j, and $\delta(x) = 1 : x = 0$; $\delta(x) = 0$; $x \neq 0$ is the binning function. In the present study, the autocorrelation function was evaluated from 0 to 10 topological distances. Thus, the chemical structures were represented as 11 dimensional vectors with regard to the used atomic property.

Three atomic properties were calculated with the software package PETRA[34] by previously published empirical methods for all atoms in a molecule: sigma electronegativity (χ_σ),[35] effective atom polarizability (α),[36] and partial atomic charge (q_{tot}).[37] In addition, the identity function, i.e., each atom was represented by 1, was used. The atomic properties for each molecule, with exception of the identity, were autoscaled to zero mean and unit variance before applying eq 1. The scaling has been shown[38] to diminish the correlations between the bins of autocorrelation and to better preserve the physicochemical information. The resulting autocorrelation vectors were calculated with ADRIANA-.Code[39] and were additionally autoscaled to ensure that the values are comparable when a distance measure (such as the Euclidian distance, used by SOM) is calculated.

2.3. Virtual Screening Methods. A schematic overview of the three virtual screening methods used in this work is presented in Figure 1. A detailed overview of each of these methods is given in the following sections.

2.3.1. Similarity Search with Subsequent Data Fusion. A schematic workflow for this type of virtual screening is presented in Figure 2.

In the first step, both the query and the screened data set are described by the same structure representation. In the second step, the representations of each of the n structures

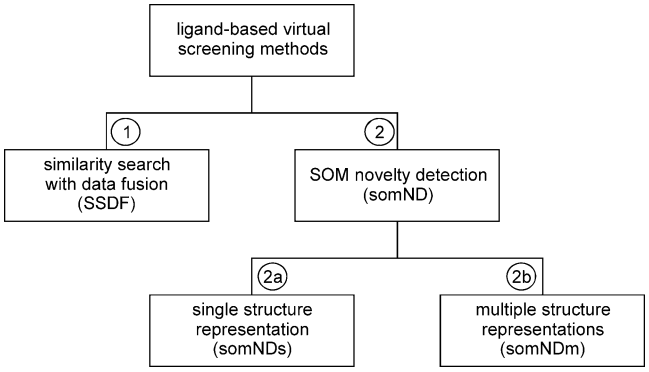


Figure 1. Overview and abbreviations of the three methods for ligand-based virtual screening used in this work. See sections 2.3.1, 2.3.2.a, and 2.3.2.b for details on each method.

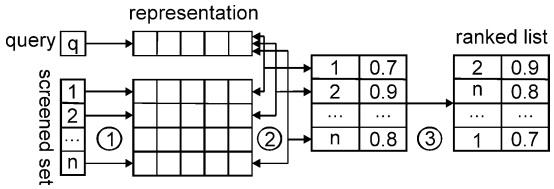


Figure 2. Workflow of a similarity search with a single query: (1) represent the chemical structures; (2) measure the similarity; and (3) sort the list. See text for details.

in the screened data set are compared to the query representation by means of a similarity measure, and a list with the similarity scores is obtained. In step 3 this list is sorted in descending order with regards to the similarity score, and the final ranked list is produced. In the present study for this similarity ranking, the structures were described by 1024-dimensional binary Daylight fingerprints. The Tanimoto coefficient was used as a similarity measure and was calculated according to eq 2

$$S_T = \frac{c}{a + b - c} \qquad (2)$$

where a is the number of bits "on" in the representation of the query structure, b is the number of bits "on" in the screened structure, and c is the number of bits "on" in both, i.e., the union of the two representations.

The workflow presented in Figure 2 applies to only one query structure. However, a set of known high-quality actives is often available. It has been shown by Whittle et al.[11] that in such cases the fusion of the ranked lists obtained from each query structure enhances the results of the virtual screen. Thus, to obtain the highest performance of the similarity search after the m ranked lists were obtained (m being the number of query structures, i.e., the number of actives in the training set) these lists were then subject to data fusion. The data fusion algorithm calculates a new score for each structure from the screened data set by combining the scores, which the structure has obtained in any of the m similarity lists. There are different methods to combine these scores, called "fusion rules".[12] In the present work the MAX rule was used, meaning that each screened structure j obtained a final score, equal to the maximum value of its individual scores, collected from each of the lists, according to eq 3

$$S_{FUS}(j) = \max_i [S^*(i,j)] \qquad (3)$$

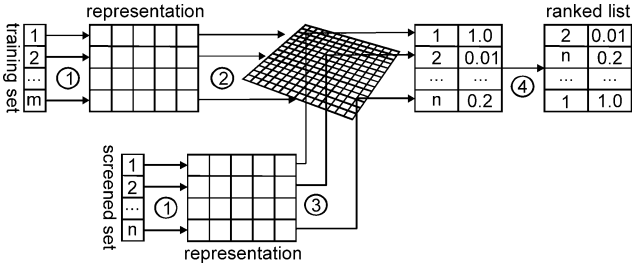


Figure 3. Workflow of a virtual screening with a SOM novelty detector: (1) represent the chemical structures; (2) train a SOM; (3) project the screened data set; and (4) sort the list. See text for details.

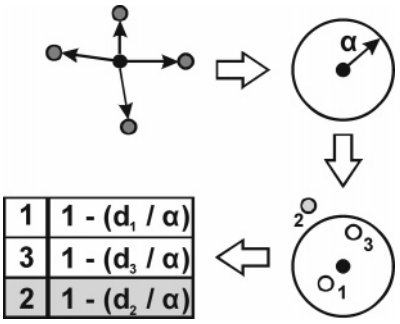


Figure 4. Assignment of local accuracy to a SOM with rectangular topology and lattice and scoring of three hypothetical structures.

where S^* denotes the calculated similarity score between the query structure i and the screened structure j.

2.3.2. Novelty Detection with Self-Organizing Maps (somND). 2.3.2.a. Single Structure Representation (somNDs). The process of using a SOM novelty detector for virtual screening with a single structure representation is shown in Figure 3.

In the first step, both the training set and the screened data set are described by the same structure representation. In the second step, a SOM is trained using only the known active compounds in the training set. Once the trained SOM is obtained, a local accuracy is assigned to each neuron by using the average distance between this neuron and its first-sphere neighbors.[20,22] In step 3, each structure from the screened data set is projected onto the trained SOM, and its best matching neuron (BMN) is found. The (Euclidian) distance between the screened structure and its BMN is then compared to the local accuracy at the BMN, and the screened structure j obtains a score $S_{SOM}(j)$ according to eq 4

$$S_{SOM}(j) = 1 - \frac{d(j,\mathrm{BMN}_j)}{a_{\mathrm{BMN}_j}} \qquad (4)$$

where $d(j, \mathrm{BMN}_j)$ denotes the distance between the structure j from the screened set and its BMN, BMN_j is the best matching neuron for structure j on the trained SOM, and $\alpha_{\mathrm{BMN}j}$ is the local accuracy at BMN_j.

The process of assigning the local accuracy and the scoring of the screened structures is visualized in Figure 4 for a SOM with rectangular topology and lattice.

The local accuracy α is determined as the average distance between the neuron and its first sphere neighbors. All screened structures, falling into the neuron, are scored as shown, d denoting the distance between the neuron weight vector and structure representation vector. The distance d_2 between the screened structure 2 and the neuron weight vector is larger than the local accuracy (see Figure 4), thus its score becomes negative, and consequently it is classified as novel (unlikely to share the biological activity).

2.3.2.b. Two or More Structure Representation (somNDm). The above workflow makes the screening of a data set using a single representation of the chemical structures straightforward. Usually more than one representation is available—in this study autocorrelation vectors, weighted by different atomic properties, were used. Thus, the easiest way to combine such multiple structure representations is by concatenating all vectors together, producing a higher dimensional descriptor and applying the workflow described above.

However, when more than one representation of the chemical structures is available, another method for utilizing this information may be more suited to the task at hand. We adopted here the generalized method for a multisensor environment described by Wong et al.[22] The method relies on training a SOM for each individual chemical structure representation and projecting the training set through these networks. The outputs of each network, i.e., the distance between a pattern and its winning neuron—for the training set are collected in a matrix, which constitutes the system's knowledge of the "normal" (active) case. When a structure is presented for screening, the output of all SOMs gives a vector in the output space, which may be compared to the positions of the training data. This comparison is performed using the Mahalanobis distance (MD) between the screened structure and the training data in the output space. The MD measures the distance between a vector, x, and its mean vector, m_x, scaled by its covariance matrix, C_{xx}. Using these notations, the (squared) MD is given as shown in eq 5

$$\mathrm{MD}^2 = (x - m_x)^T C_{xx}^{-1} (x - m_x) \qquad (5)$$

In the context of this work m_x is a vector containing the column means of the "active" space matrix obtained when building the novelty detector (see the left part of area C in Figure 5) and C_{xx} is the covariance matrix of this "active" space matrix; x denotes a vector containing the output of each individual SOM for the compound being tested (the right part of area C in Figure 5).

De Maesschealk et al.[40] present a comprehensive tutorial about MD and its applications. An important property of the MD is that they are χ^2 distributed. Thus, one can apply a statistically sound threshold when declaring patterns as novel (less likely to possess biological activity, given the previously known active). The threshold may be calculated using the confidence intervals of the χ^2 distribution, thus $\chi^2(\alpha{=}0.99, n)$ will give a value that encapsulates 99% of the data in the active cluster when n different structure representations are used. Figure 5 visualizes the four steps of the described procedure.

The χ^2 distribution of the Mahalanobis distance is based on the assumption that the data used to calculate it are multivariate normal distributed. Figure 6a shows a quantile−quantile (Q−Q) plot of the calculated Mahalanobis distances against the theoretical quantiles of χ^2 distribution with 4 degrees of freedom for the PDE 4 training set, while Figure 6b shows the same plot for the uPA set. The value of the threshold $\chi^2(0.99, 4)$ is presented as a dotted line. As can be seen from Figure 6 the distribution of the Mahalnobis

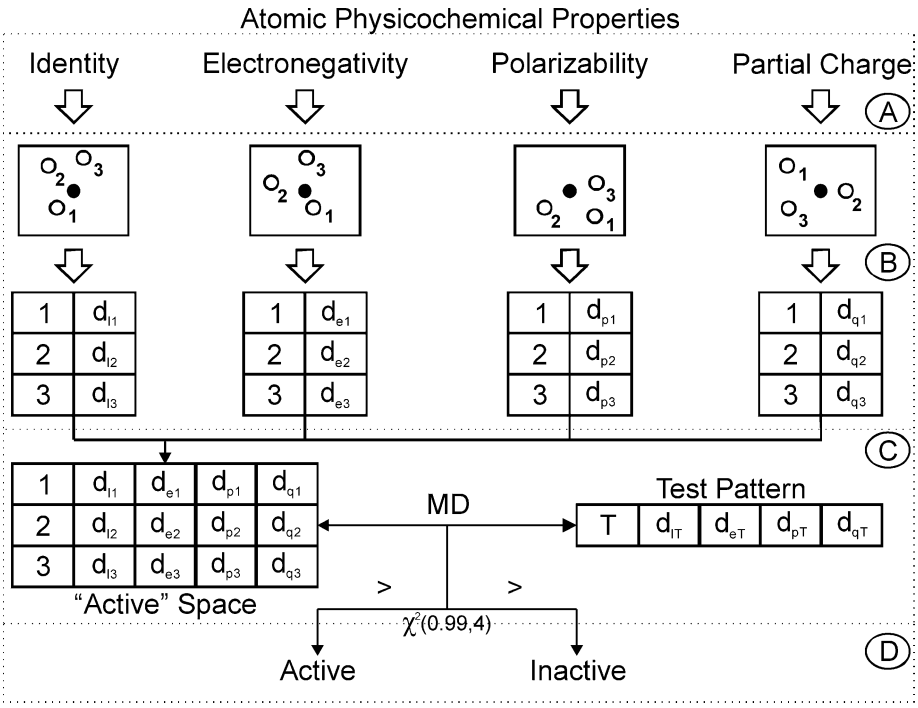


Figure 5. Novelty detection with four different structure representations. After the selected structure representations have been obtained (area A), a SOM is trained for each structure representation (area B). The training set is projected through the trained networks, and the knowledge of the "active" space is collected in a matrix (area C, left). The screened structure is then projected through the same networks, collecting their output in a vector (area C, right). This vector is in turn compared to the "active" space by means of Mahalanobis distance. Based on a statistically chosen threshold, the screened structure is classified as active or inactive (area D).

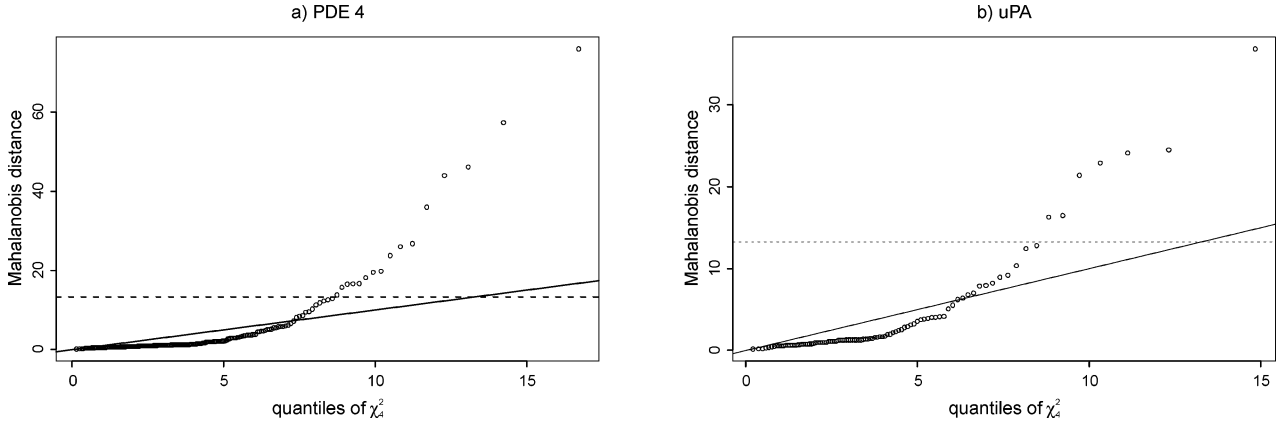


Figure 6. Q−Q plot of the calculated Mahalanobis distance of the training set of PDE 4 inhibitors (a) and for the training set of uPA inhibitors and (b) against the quantiles of χ^2 distribution with 4 degrees of freedom.

distances is deviating from the χ^2 distribution. The observed distribution is heavy tailed in comparison to the χ^2. The Mahalanobis distance can still be used. However, due to the heavy tailing toward larger distances the selected threshold is likely to be conservative, that is, describing the active space very tight. Such a behavior is useful when a low false-positive ratio is needed. False-positive in the context of this paper refers to inactive compounds, predicted as active.

2.3.3. Selection of SOM Size and Time Complexity. The complexity of projecting a data set of size m where each pattern has a dimensionality d to a SOM with n neurons is linear in all these sizes—$O(mdn)$. Preliminary experiments have shown that increasing the number of neurons beyond 80% of the size of the training set does not lead to significant improvement of the results. Thus, a SOM with a number of neurons roughly equal to $0.8 \times s$, where s is the number of patterns in the training set, was used. The complexity of the similarity search with subsequent data fusion, on the other hand, also scales linearly with m, d, and the number of query structures. Thus, the SOM-based method is about 20% faster than the similarity search for an equal descriptors dimensionality. Provided that the autocorrelation vectors used in this study have a 23 times lower dimensionality than the fingerprints, the execution is faster when a single SOM is used. All run times were measured on a 64 bit Athlon 3700+ PC with 1GB of RAM, running SuSE Linux and using an in-house written program.

2.3.4. Selection of the Best SOM. The training of a SOM is a stochastic approximation process, thus usually several maps need to be built and the best one is selected. A big variety of measures for the goodness of mapping has been proposed,[41,42] but there is still no uniformly recognized method to evaluate the quality of a particular SOM. Thus, in the following experiments with single structure representation ten SOMs were trained, and the obtained ranked lists were merged using eq 3. When multiple representations in

Table 2. Recall in Percent at Rank 1000 for the Five Activity Classes[e]

			autocorrelation of				
activity class	data set split	fingerprints[a]	identity[b]	χ_σ[b]	α[b]	q_{tot}[b]	concatenated[c]
AChE (284)[d]	random	65.1 (182)	29.3	28.2	36.3	36.4	70.4 (199)
	semirandom	67.3 (191)	23.9	22.6	33.3	38.2	66.3 (189)
	clustering	73.1 (208)	24.1	27.4	33.2	37.1	66.5 (189)
COX-2 (500)[d]	random	69.4 (344)	44.1	26.3	35.1	42.7	68.4 (339)
	semirandom	79.8 (400)	48.3	26.4	32.9	40.3	66.8 (335)
	clustering	61.6 (312)	47.8	25.2	34.8	39.8	68.3 (341)
PDE 4 (229)[d]	random	92.8 (213)	36.9	34.2	44.1	38.1	72.4 (166)
	semirandom	88.5 (203)	38.3	35.1	36.4	31.4	73.9 (169)
	clustering	94.3 (216)	33.1	30.3	40.7	33.8	74.1 (170)
thrombin (1053)[d]	random	41.4 (430)	28.4	20.1	27.2	34.3	46.4 (481)
	semirandom	39.2 (409)	32.7	19.9	29.4	35.2	50.2 (529)
	clustering	41.7 (440)	30.3	21.3	29.3	36.4	53.4 (560)
uPA (100)[d]	random	69.4 (69)	34.2	15.9	33.1	32.5	52.3 (52)
	semirandom	74.4 (74)	32.4	19.2	36.6	46.2	70.2 (70)
	clustering	79.7 (80)	24.2	18.1	35.7	46.7	56.4 (56)

[a] Similarity search with subsequent data fusion (method 1 of Figure 1, cf. section 2.3.1). [b] SOM novelty detection with single structure representation (method 2a of Figure 1, cf. section 2.3.2.a), fused results of ten networks, descriptor: topological autocorrelation vectors, weighted by the corresponding atomic property. [c] Same method as footnote *b*, descriptor: the four autocorrelation vectors concatenated. [d] Total number of active compounds.[e] The numbers in parentheses give the number of recovered active compounds.

concert with a Mahalanobis distance were used, the SOM with the lowest quantization error was used for each representation. For a data set containing N patterns, the average quantization error is calculated according to eq 6

$$\epsilon_q = \frac{1}{N}\sum_{i=1}^{N} d(x_i, \mathrm{BMN}_{x_i}) \tag{6}$$

where $d(x_i, \mathrm{BMN}_{x_i})$ denotes the distance between pattern x_i and its best-matching neuron.

2.4. Performance Measures. *2.4.1. Recall.* Virtual screening involves the sorting of a data set of chemical compounds in order of decreasing probability of activity. Once the whole data set has been sorted a subset of the top ranked compounds is considered. The size of this subset is called *rank*.

The results of virtual screening experiments are usually reported by using a *cumulative recall* (most commonly referred to as *recall*).[43] The *recall* at a given *rank* is calculated according to eq 7

$$\mathrm{recall}_r = \frac{F_{\mathrm{act}}}{D_{\mathrm{act}}} \tag{7}$$

where D_{act} is the number of compounds in the screened data set, which exhibit a given biological activity (i.e., the size of the test set), r is the rank, and F_{act} is the number of experimentally validated actives recovered at this rank. The recall gives that fraction of the known actives, which was recovered in the similarity search at a given rank. It is bound between zero and one, with one indicating perfect performance.

2.4.2. Enrichment Factor (ef). The enrichment factor (ef) gives the improvement in the retrieval of active structures at a given rank compared to a random selection with the same rank. It is calculated according to eq 8

$$\mathrm{ef} = \frac{F_{\mathrm{act}}}{r} \times \frac{D_{\mathrm{all}}}{D_{\mathrm{act}}} \tag{8}$$

where D_{all} is the size of the data set, D_{act} is the number of compounds in the data set which exhibit a given biological activity, r is the rank, and F_{act} is the number of experimentally validated actives recovered at this rank. Any method which performs superior to a random selection of r compounds has an enrichment factor greater than one.

2.5. Validation. Two cross-validation-like approaches have been used to validate the performance of the studied virtual screening methods.

In the first approach, the known active structures were randomly split into ten parts. Each one of these parts was kept aside. The training set was selected from the remaining known active structures, and a virtual screening of the whole WOMBAT database was performed. The retrieval of the actives which were kept aside among the 1000 top-ranked compounds (after removing all other known actives from the ranked list) was considered. The whole procedure was repeated five times, and the average recall over the 50 repetitions is reported. This approach ensures a realistic evaluation of the performance of the corresponding virtual screening method.

The second approach was designed in a more aggressive way. Approximately 10% of the known actives were kept aside. However, instead of a random selection, these test sets contained complete clusters, as identified by the Taylor−Butina clustering algorithm. Therefore, the training sets selected from the remaining active structures contained compounds structurally dissimilar to the compounds in the test sets. Since there is a limited variability in the test set selection procedure due to the clustering, the above procedure was executed one time. In this manner the lower bound of the expected performance can be evaluated.

3. RESULTS AND DISCUSSION

Table 2 gives the recall in percent and in parentheses the absolute number of retrieved actives at rank 1000, i.e., after the thousand top-ranked structures, have been retrieved from the screened set of 135 877 compounds.

The five activity classes are shown with regards to the descriptors, analysis method, and data set splitting method. The size of the test set, that is, the actives which were the aim of retrieval, is indicated in parentheses for each activity

Table 3. Cross-Validated Recall in Percent at Rank 1000 Using Random Test Set Selection[d]

	fingerprints[a]		autocorrelation, concatenated[b]		autocorrelation, χ^2 [c]	
activity class	recall	SD	recall	SD	recall	SD
AChE	77.0	6.0	72.3	6.0	61.7	6.2
COX-2	58.9	7.1	84.6	2.9	78.0	6.2
PDE 4	90.3	4.8	79.7	6.5	71.9	6.7
thrombin	48.6	4.2	53.1	4.1	39.0	6.4
uPA	69.8	9.8	60.1	10	57.9	10.1

[a] Similarity search with subsequent data fusion (method 1 of Figure 1, cf. section 2.3.1). [b] SOM novelty detection with single structure representation (method 2a of Figure 1, cf. section 2.3.2.a), fused results of ten networks, descriptor: descriptor: the four autocorrelation vectors concatenated. [c] SOM novelty detection with multiple structure representation (method 2b of Figure 1, cf. section 2.3.2.b). [d] Cf. section 2.5. Mean and SD over 50 repetitions (5 time 10-fold cross-validation).

Table 4. Cross-Validated Recall in Percent at Rank 1000 Using Complete Clusters as Test Set[d]

	fingerprints[a]		autocorrelation, concatenated[b]		autocorrelation, χ^2 [c]	
activity class	recall	SD	recall	SD	recall	SD
AChE	14.0	12.6	31.2	14.2	21.2	10.2
COX-2	9.5	3.8	75.8	13.7	65.0	15.6
PDE 4	30.6	26.1	38.9	20.7	24.5	19.6
thrombin	3.8	3.6	21	5.5	12.1	2.7
uPA	27.5	23.0	30.8	18.2	31.8	23.7

[a] Similarity search with subsequent data fusion (method 1 of Figure 1, cf. section 2.3.1). [b] SOM novelty detection with single structure representation (method 2a of Figure 1, cf. section 2.3.2.a), fused results of ten networks, descriptor: descriptor: the four autocorrelation vectors concatenated. [c] SOM novelty detection with multiple structure representation (method 2b of Figure 1, cf. section 2.3.2.b). [d] Cf. section 2.5. Mean and SD over 10 repetitions (10-fold cross-validation).

class (approximately half of the total number of all known active structures, cf. Table 1).

3.1. Training Set Selection. The utilization of the Taylor−Butina clustering algorithm for selecting the training set gives the best results when a similarity search with Daylight fingerprints and subsequent data fusion is carried out. This tendency is most obvious for the AChE and uPA classes, while in the COX-2 case the results with a clustering based splitting are the worst. The semirandom selection offers significant improvements only in the COX-2 case, while the random selection is comparable with the clustering for PDE 4 and thrombin. The SOM novelty detection was not much affected by the choice of training set. This hints that the compounds in the selected activity classes are relatively evenly distributed among the chemical space, defined by the autocorrelation vectors, thus any subset is able to retrieve a similar number of actives. Due to the fact that quantitative activity information is not always available, or there may not be enough activity spread in the actives set, the semirandom split may not always be applied. Except for the COX-2 case, it did not lead to significant improvements. A Taylor−Butina clustering is generally accepted[44] as the clustering method of choice for the selection of an optimal diverse set of compounds and can easily be integrated into the virtual screening process. However, as can be seen from Table 2, one can do reasonably well with random selection, provided that the selected training set is large enough. For splits that preserve only a minor part for training, the clustering method is preferred over random splitting. For example, using only 20% of the available COX-2 inhibitors as a training set, a recall (at rank 1000) of 38% and 57% was obtained with random and clustering based splits, using Daylight fingerprints with data fusion.

The split based on Taylor−Butina clustering is used for the rest of this work since this method decreases the variability due to different randomly selected training sets and guarantees an even distribution of the compounds in the training set through the activity space.

3.2. Validation. Table 3 shows the recall values considering the top-ranked 1000 structures obtained with 5 times 10-fold cross-validation-like experiment based on random splitting.

Values similar to those reported in Table 2 were observed. The performance of the similarity search method decreases slightly, while the averaged recall obtained with SOM novelty increases. Thus, the gap between the two methods is lower than suggested solely by Table 2. In the case of uPA inhibitors the advantage of the similarity search is much less pronounced in the cross-validated results. For the other activity classes similar performances as in Table 2 are observed. The results obtained with multiple structure representations and Mahalanobis distance in the output space (not included in Table 2) are in general lower than those obtained with concatenated autocorrelation vectors. The deviation of the SOM output from the expected χ^2 distribution, as shown in section 2.3.2.b, might be the reason for this behavior.

Table 4 shows the recall values considering the top-ranked 1000 structures obtained with a 10-fold cross-validation-like experiment, based on leaving out complete clusters.

The first thing to note from Table 4 is that the obtained recall values are 2−10 times lower compared to the values in Tables 2 and 3. Another observation is that the SOM novelty detection performed better for all activity classes. The decrease in the performance for the similarity search is hardly surprising. As the name implies, this methods relies mainly on similarities between the structure in the training set and those in the screened database. By excluding whole clusters an artificial situation in which a complete group of highly self-similar compounds is not presented at all in the training set is created. This has led to a more than a 10-fold loss of performance in the case of thrombin inhibitors (recall of 41.7, cf. Table 2, against 3.8, cf. Table 4). The SOM novelty detection with topological autocorrelation was affected less, leading only to a 2-fold loss of performance on average. This observation confirms that by using topological autocorrelation vectors weighted by physicochemical properties additional aspects of the similarity between the structures are covered. Thus, while complete clusters of structurally self-similar molecules were not present in the training set, between 20 and 70% of these excluded structures were retrieved. However, ultimately most machine learning methods depend on at least some similarities between the structures in the training and in the test set. Thus, the results presented in Table 4 have to be read as a pessimistic measure of the performance since large groups of self-similar compounds were intentionally not represented in the training set.

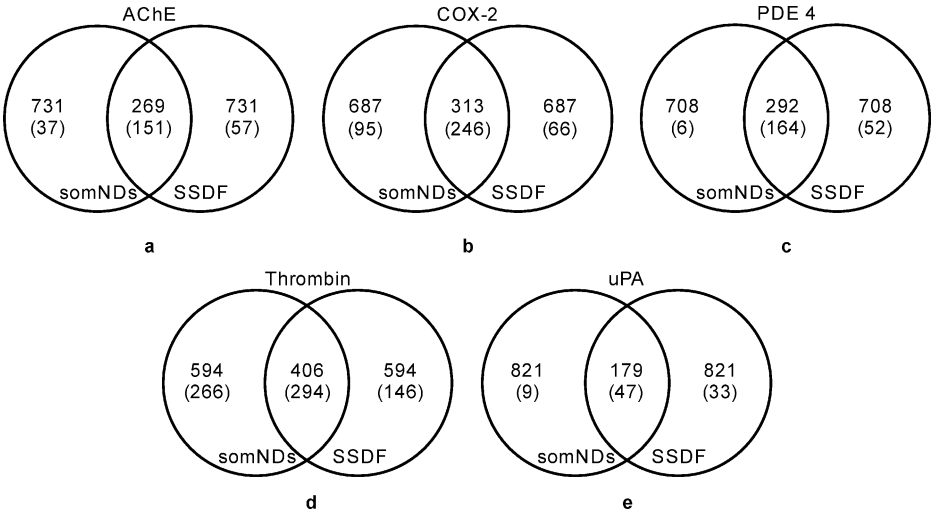


Figure 7. Venn diagrams of the intersection between the 1000 top ranked compounds retrieved by similarity search with subsequent data fusion and Daylight fingerprints (method 1 of Figure 1) and SOM novelty detection with concatenated autocorrelation vector (method 2a of Figure 1). The number of actives in the intersection and in the unique part of the lists is given in parentheses. Inside each circle the numbers on top add up to the rank value of 1000, and the numbers in parentheses add up to the values given in Table 2.

Finally, the very high standard deviations shown in Table 4 deserve a note. A closer expectation of the performance in the individual runs has shown that the main source of these high standard deviations is the same for all activity classes. The run in which the largest homogeneous cluster was omitted from the training set always resulted in low recall values, thus creating a high variability in the ten runs. This observation once again demonstrates the dependency of the both studied methods on at least some structural similarity between the structures in the training and in the test set.

We conclude this section noting that in a cross-validation-like experiment based on random splitting the obtained recall values shown in Table 3 match closely those in Table 2. Thus, the expected performance of the methods presented in this study is somewhere around the recall values in Table 3. On the other hand, the lower-bound of the performance is shown in Table 4. The novelty detection with autocorrelation vectors is expected to give better results in cases where compounds structurally dissimilar from those in the training set are to be retrieved.

3.3. Method Comparison. Comparing the recall values obtained with the different methods (using clustering based split) shows that the SOM novelty technique performs better or comparable to the Daylight fingerprints when the concatenated autocorrelation vectors are used to describe the structures. The SOM based novelty detection performs better for COX-2 and thrombin and comparable for the AChE, while for the PDE 4 and uPA sets, the fingerprint based method gives better results. When an autocorrelation vector, weighted by a single property is used, the results are generally of lower quality. However, taking into account the much lower dimensionality (11 compared to 1024), the fact that the SOM novelty detection is capable in most of the cases to recover around 50% of the actives that are recovered with fingerprints is remarkable.

Using different physicochemical properties enables one to describe different aspects of the active set. Thus, while the identity accounts strictly for structural similarity, i.e., it is equivalent to a histogram of the bond distances in a molecule, the partial charge accounts for the electrostatic properties of the compound, while the polarizability descriptors are related to size and hydrophobic effects. The electronegativity values take into account the hydrogen-bonding properties of the compound. Although all these are rather crude descriptions of the underlying effects, based on the connectivity matrix alone, they provide a good basis for this type of virtual screening experiments in which no quantitative structure−activity correlations are made. Therefore, the different autocorrelation vectors are expected to recover different sets of actives, which explains the significant improvement when the concatenated vector is used. This 44-dimensional concatenated autocorrelation vector has an approximately 23 times lower dimensionality than the 1024-bit Daylight fingerprints. This, together with the smaller size of the SOM, offers a significant run time improvement. Thus, for the thrombin subset, which is the largest one used in this study, the fingerprints based method runs in approximately 130 s, while projecting the whole test set onto a single network is twice as fast, approximately 65 s. This advantage is compensated for by the fact that the recall is hardly related to any of the SOM quality measures, which leads to the requirement of using several networks as a better description of the "active" space.

3.4. Methods Complimentary. Looking at the recall values in Table 2 one can see that neither of the methods was able to recover the full set of known actives. This observation leads us to investigate how different are the ranked lists, obtained by the two methods. Figure 7 shows Venn[45] diagrams of the intersection between the fingerprints based similarity search (method 1 of Figure 1) and SOM novelty detection with concatenated autocorrelation vectors (method 2a of Figure 1).

3.4.1. Intersection. There is a clear difference between the two sets of recovered compounds, the highest intersection in the case of the thrombin subset being equal to 40% of the total ranked list size. As expected, the majority of the recovered actives are found in the intersection. A look at Figure 7 reveals that the intersection between the ranked lists obtained with the two methods (the area shared by both circles) is highly enriched in active structures. The percentage of active structures among these compounds found in the

Table 5. Percent Active Compounds and Enrichment Factors Obtained When Only the Compounds Found in the Intersection between the Ranked Lists Returned by Similarity Search with Data Fusion (Method 1 in Figure 1) and SOM Novelty Detection with Concatenated Autocorrelation Vectors (Method 2a in Figure 1) Are Considered[b]

	intersection				actives recovered by[a]	
activity	compds	active	% actives	enrichment	SSDF	somNDs
AChE	269	151	56	268	86	126
COX-2	313	246	79	213	166	204
PDE 4	292	164	56	333	159	135
thrombin	406	294	72	93	238	276
uPA	179	47	26	357	44	39

[a] Number of active compounds recovered by the corresponding method considering the top-ranked compounds inside a list with the same length as the intersection (column 2). [b] The number of actives recovered in a ranked list of the same length (cf. column two) by each method alone is shown as well.

Table 6. Percent Active Compounds and Enrichment Factors Obtained When the Compounds Found in the Union between the Ranked Lists Returned by Similarity Search with Data Fusion (Method 1 in Figure 1) and SOM Novelty Detection with Concatenated Autocorrelation Vectors (Method 2a in Figure 1) Are Considered[b]

	union				actives recovered by[a]	
activity	compds	active	% actives	enrichment	SSDF	somNDs
AChE	1731	245	14	68	238	205
COX-2	1687	407	24	65	366	378
PDE 4	1708	222	13	77	220	186
thrombin	1594	706	44	57	571	672
uPA	1821	89	5	66	88	64

[a] Number of active compounds recovered by the corresponding method considering the top-ranked compounds inside a list with the same length as the union (column 2). [b] The number of actives recovered in a ranked list of the same length (cf. column two) by each method alone is shown as well.

intersection of the two lists as well as the corresponding enrichment factors are summarized in Table 5. From the values presented in Table 5 it is clear that starting with two ranked lists returned by each of the methods (each list containing 1000 structures) by taking their intersection a very short list which is highly enriched in active structures can be obtained. In addition, the list obtained by considering the intersection contains from 3 to 75% more active structures than the ranked lists of the same size obtained with each of the methods alone.

3.4.2. Union. Each of the methods, however, is capable of recovering structures, which the other method has missed. Thus, by concatenating the unique compounds in the lists a new list is obtained, which is, of course, longer but still more enriched than the corresponding individual lists of the same length. Considering the thrombin subset, by summing up the numbers on top in both circles (594 + 406 + 594) the size of the combined ranked list—1594—is obtained. The number of known actives recovered in this combined ranked list—706—is obtained by summing up the numbers in parentheses (266 + 294 + 146). To compare the result of the above union to the results obtained by each of the methods alone, a ranked list of the same size—1594—was obtained with each method. The number of recovered actives in these lists (not shown in Figure 7) was 571 and 672 for the similarity search with Daylight fingerprints and SOM novelty detection with concatenated autocorrelation vector, respectively. Thus, a union of the two lists at rank 1000 gives a list, which is more enriched with actives. The same holds true for the other four sets as summarized in Table 6. Even when the fingerprints similarity search seems to recover almost all actives, as in the case with the PDE 4 subset where the recall is 94% (cf. Table 2, clustering split) the SOM novelty recovers 6 of the 13 (6%, cf. Table 2) missed actives.

3.5. Scaffold Analysis. Ultimately, one of the highly desired properties of any virtual screening method is the ability to recover new active scaffolds. We want to remind that the term "novelty" as used throughout this text does not refer to the discovery of novel classes of active compounds. The SOM novelty detection like any machine-learning method relies on some commonalities between the active compounds. While the autocorrelation descriptor attempts to take the chemistry as well as structural features into account, it is still dependent on the underlying chemical structure. Therefore, it is unrealistic to expect the discovery of completely new active chemotypes.

Nevertheless, we were interested in the difference in terms of chemotypes between the structures obtained with SOM novelty detection and with similarity search with subsequent data fusion. To achieve this we performed a graph-based scaffold analysis with the help of MeqiSuite.[46]

MeqiSuite calculates 66 different graph-based indices. Detailed description of these indices and the software itself can be found on the MeqiSuite Web site (http://www.pannanugget.com) and in the technical report "An Introduction to the MeqiSuite Indices".[47] In our work, we considered the ordering index *CyclicSystemOrd*. This is a composite hierarchical index meant to facilitate the browsing of diverse compound collections. The *CyclicSystemOrd* index is formed by the concatenation of 11 other MeqiSuite indices. The concatenation is done in a way which groups the similar ring systems together. All compounds which do not contain a cyclic system are grouped together as well. However, almost all of the active compounds used in this study have at least one cycle thus this particular index was considered. For all classes of active compounds a comparison between the unique active structures recovered by each method (the numbers in parentheses in the left- and right-hand circles in Figure 7) was performed. These compounds were also compared to the training set. An outline of the scaffold analysis procedure is presented in Figure 8 with AChE inhibitors as an example, and the results for all five activity classes are summarized in Table 7.

3.5.1. AChE Inhibitors. A look at the third column of Table 7 shows that there were 37 AChE inhibitors recovered only by SOM novelty detection and 57 recovered only by similarity search (these are the same figures as given in parentheses in Figure 7a). The *CyclicSystemOrd* MequSuite index distinguished 28 different chemotypes for SOM novelty and 41 for the similarity search—the fourth column in Table 7. Eight chemotypes were shared by both sets of recovered structures thus leaving 20 unique chemotypes recovered by SOM novelty detection alone and 33 unique chemotypes recovered by similarity search alone—the fifth column in Table 7.

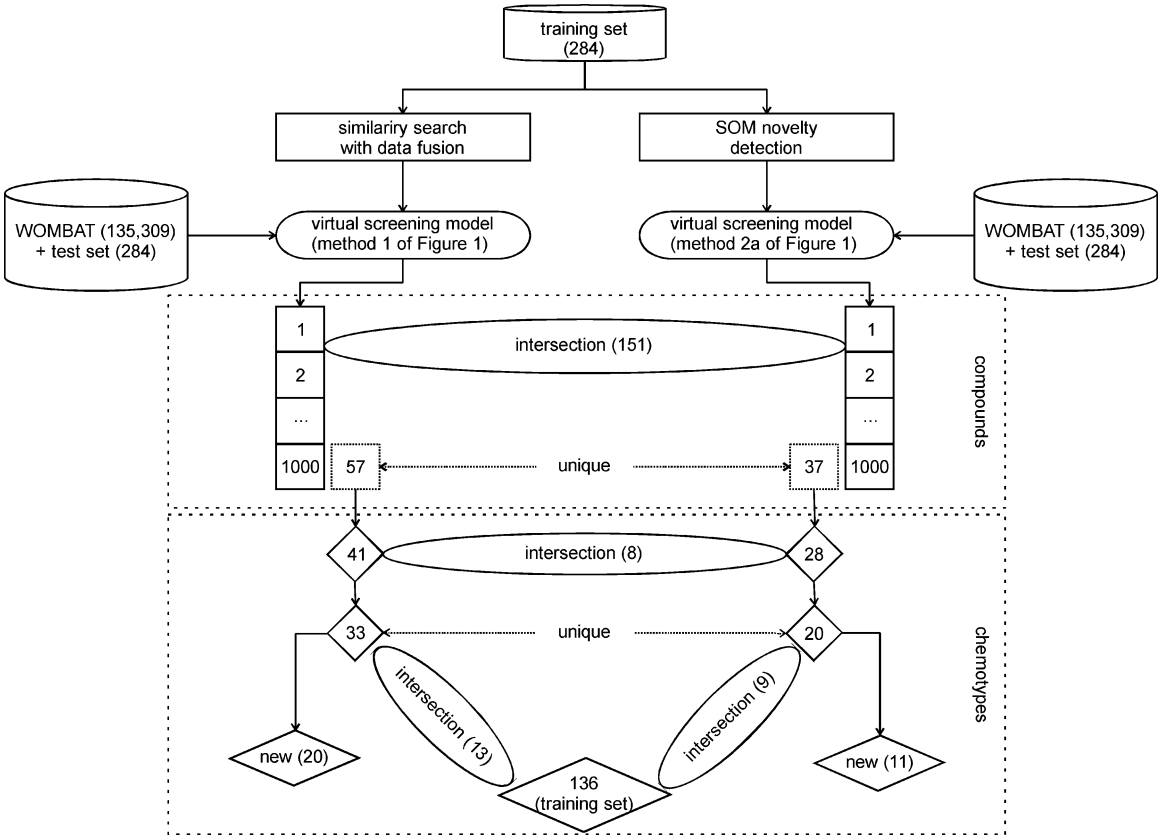


Figure 8. Outline of the scaffold analysis workflow with AChE inhibitors as an example. Using the training set a virtual screening model is developed, and the whole WOMBAT database plus the test set is screened. The top-ranked 1000 structures are considered. Among them the active structures, recovered exclusively by each method, are subject to scaffold analysis by means of the *CyclicSystemOrd* MequSuite index. A comparison between the active chemotypes recovered by each method and between the recovered chemotypes and the chemotypes contained in the training set is made. The results for all activity classes are summarized in Table 7.

Table 7. Number of Chemotypes among the Active Compounds Recovered Exclusively by One of the Two Methods[e]

			chemotypes			
activity class	method	unique active compds[a]	total	unique	not in train set	not in train set, UnSkCyc[b]
AChE	somNDs[c]	37	28	20	11	3
	SSDF[d]	57	41	33	20	6
COX-2	somNDs	95	57	48	26	6
	SSDF	66	35	26	10	2
PDE 4	somNDs	6	6	6	4	2
	SSDF	52	38	38	24	12
thrombin	somNDs	266	157	134	64	18
	SSDF	146	112	89	49	30
uPA	somNDs	9	7	6	2	1
	SSDF	33	19	18	6	3

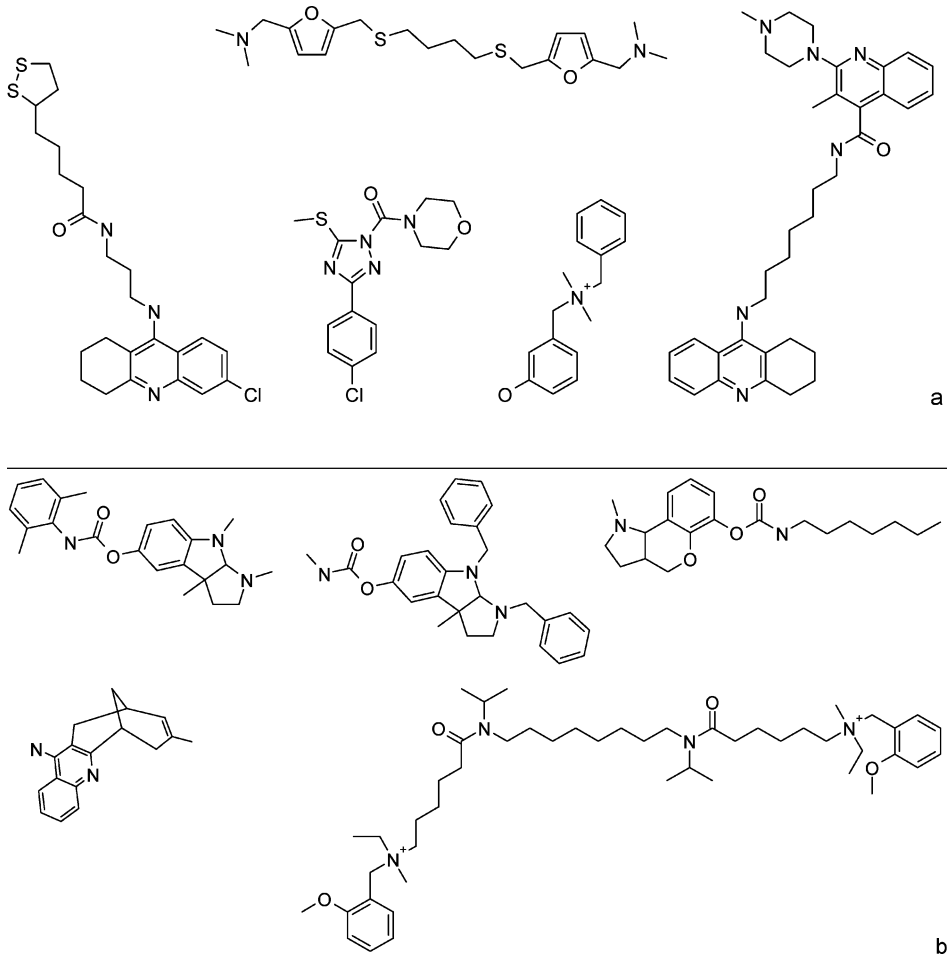
[a] Unique active structures recovered by each method (the numbers in parentheses in the left- and right-hand circles in Figure 7). [b] The UnSkCyc****Mqn index ignores atom and bond types and considers the unextended cyclic system. [c] SOM novelty detection with concatenated autocorrelation vectors (method 2a in Figure 1). [d] Similarity search with Daylight fingerprints and subsequent data fusion (method 1 in Figure 1). [e] Cf. Figure 8. The MeqiSuite *CyclycSystemOrd* index was used for the identification of chemotypes with the exception of the last column.

Five from the 20 unique chemotypes recovered by SOM novelty detection are depicted in Chart 1a, while Chart 1b shows five from the chemotypes, recovered by similarity search alone. As can be seen from Chart 1 the SOM novelty detection was able to recover chemotypes missed by the similarity search and vice versa.

Another interesting question was whether the methods recover chemotypes not present in the set of query structures. As can be seen from the sixth column in Table 7, the *CyclicSystemOrd* MeqiSuite index distinguished 11 chemotypes among the 37 AChE inhibitors recovered exclusively by SOM novelty which were not present in the training set. The number of chemotypes not present in the training set was 20 for the set of 57 AChE inhibitors recovered exclusively by similarity search. These numbers suggest that both methods were able to recover scaffolds not present in the training set. A careful examination of the corresponding structures revealed that although they do have ring systems absent from the training set the differences are mainly due to the position of certain heteroatoms. Considering only the unextended cyclic-system skeleton index *UnSkCyc****Mqn*—which does not distinguish between atom and bond types, the number of chemotypes not present in the training set decreased to 3 for the SOM novelty and to 6 for the similarity search—the seventh column in Table 7. This shows that some structural similarities between the recovered compounds, and the set of known actives used as training set do exist.

However, as already mentioned it is unrealistic to expect the recovery of completely new classes of active compounds based on a general structural descriptor. In the case of similarity search with binary fingerprints such new classes are not recoverable since by definition they are dissimilar to the training set. For the SOM novelty with autocorrelation vectors such compounds will be typically predicted as inactive (novel) since they may lie outside the space covered

Chart 1. AChE Inhibitors Illustrating Some of the Chemotypes[a]



[a] As perceived by the *CyclicSystemOrd* MeqiSuite index and recovered (a) exclusively by SOM novelty detection (method 2a in Figure 1) and (b) exclusively by similarity search with subsequent data fusion (method 1 in Figure 1).

by the training set. From this perspective the recovery of chemotypes not present in the training set is remarkable.

3.5.2. COX-2 Inhibitors. Chart 2a illustrates five of the 57 chemotypes unique to SOM novelty. Chart 2b shows five of the 36 chemotypes unique to similarity search.

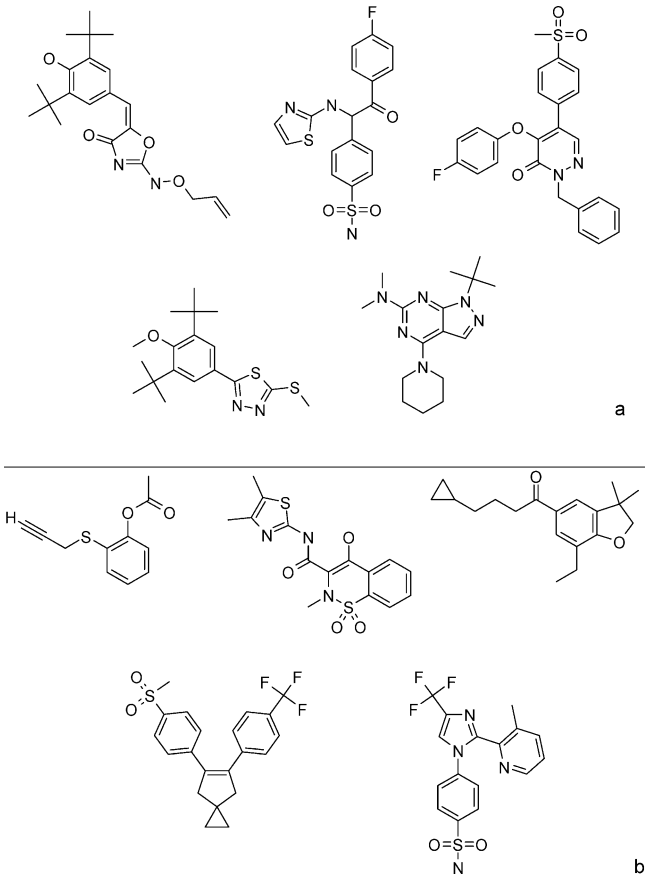
The same analysis as for the AChE inhibitors was performed on the 95 compounds recovered exclusively by SOM novelty detection and on the 66 COX-2 inhibitors recovered exclusively by similarity search with subsequent data fusion. The corresponding numbers for the different chemotypes are shown in the COX-2 row in Table 7. Similar observation as in the case of AChE inhibitors can be made. However, considering that the COX-2 training set was the less diverse one in terms of unique chemotypes in the training set (on average 3 structures per chemotype as determined by the *CyclicSystemOrd* MeqiSuite index) the recovery of some additional ones deserve a note.

3.5.3. PDE 4 Inhibitors. In this particular case, the similarity search with data fusion recovered almost all known actives among the top-ranked 1000 compounds. Thus, it was interesting to examine if the six PDE inhibitors recovered exclusively by SOM novelty detection contain different chemotypes than the rest. As can be seen from Table 7, column five inside the PDE 4 row, this was indeed the case. Four of the chemotypes recovered by SOM novelty were not present in the PDE 4 training set as well. Chart 3a shows all six PDE 4 inhibitors recovered exclusively by SOM novelty detection. Five compounds representing some of the chemotypes recovered only by similarity search are shown in Chart 3b.

A detailed description of the MeqiSuite indices lies out of the scope of this article. However, in the following we provide a short discussion centered on the compounds shown in Chart 3a which should facilitate the understanding of this section. As already mentioned all six compounds in Chart 3a are perceived as representing different chemotypes according to the composite *CyclicSystemOrd* index—a result, which is hardly surprising having in mind the structures shown. Two structures can have the same value for this composite index only when they have the same cyclic system. The *CyclicSystemOrd* index accounts for the atom and bond types as well, thus it creates a relatively large number of different chemotypes, i.e., it has high resolution.

On the other hand, although structures **1** and **2** in Chart 3a are by any means different chemical entities—structure **1** being a xanthine sulfonamide and structure **2** being a pyrazolopyrimidine-2,4-dione sulfonamide—there is high similarity in their skeletons. To achieve a broader grouping, a MeqiSuite index with lower resolution should be used. The composite indices like the *CyclicSystemOrd* used here are formed by concatenating some of the other MeqiSuite indices (When a hierarchical ordering is needed, additional care has to be taken to follow the corresponding relationship between

Chart 2. COX-2 Inhibitors Illustrating Some of the Chemotypes[a]



[a] As perceived by the *CyclicSystemOrd* MeqiSuite index and recovered (a) exclusively by SOM novelty detection (method 2a in Figure 1) and (b) exclusively by similarity search with subsequent data fusion (method 1 in Figure 1).

the individual indices; the reader is referred to ref 47 for details.).

Deleting the individual indices from which such a hierarchical index is built from right to left leads to a decrease in the resolution and subsequently larger groups of compounds are perceived as belonging to the same chemotype. This is exemplified in column seven in Table 7. To obtain the numbers shown in column seven six indices were repeatedly deleted from the *CyclicSystemOrd* index starting with the right-most one. In this manner the resolution was effectively determined by the *UnSkCyc***Mqn* index. The *UnSkCyc***Mqn* index provides a description of the unextended cyclic-system skeleton. It treats all atoms as carbon and all bonds as single. The index is "unextended" because the bridging atoms—like the ether bridge in compound **4** in Chart 3a—are not taken into account. It should be apparent now that based on *UnSkCyc***Mqn* index structures **1** and **2** from Chart 3a are perceived as identical. Of course, one can argue that structures **1** and **2** still differ in their side chains. While this is certainly true and the MeqiSuite does provide the corresponding side chains and functional group indices, we will limit the discussion only to the cyclic skeleton since, as can be seen from all charts shown, the cyclic skeleton can be seen as the main building block for all classes of active compounds.

We conclude this section noting that the chemotypes of structures **2**, **3**, **5**, and **6** were not found in the complete set of known actives which were used to build the novelty detector, while structures **3** and **6** from Chart 3a have been perceived as chemotypes missing from the training set even when the *UnSkCyc***Mqn* index was used.

3.5.4. Thrombin Inhibitors. Five examples for unique chemotypes discovered by SOM novelty and by similarity search with subsequent data fusion are shown at Chart 4a,b.

Similar observations for the AChE, COX-2, and PDE 4 classes can be drawn. The SOM novelty detection has recovered more actives in the top ranked 1000 structures. This has resulted in a higher number of unique chemotypes as can be seen from Table 7. Compared to the similarity search with Daylight fingerprints 134 additional chemotypes were recovered by SOM novelty detection.

3.5.5. uPA Inhibitors. Six uPa inhibitors representing the six unique chemotypes (cf. Table 7, column five) discovered by SOM novelty alone are depicted in Chart 5a. Five structures representing five of the 18 chemotypes discovered by similarity search alone are shown in Chart 5b.

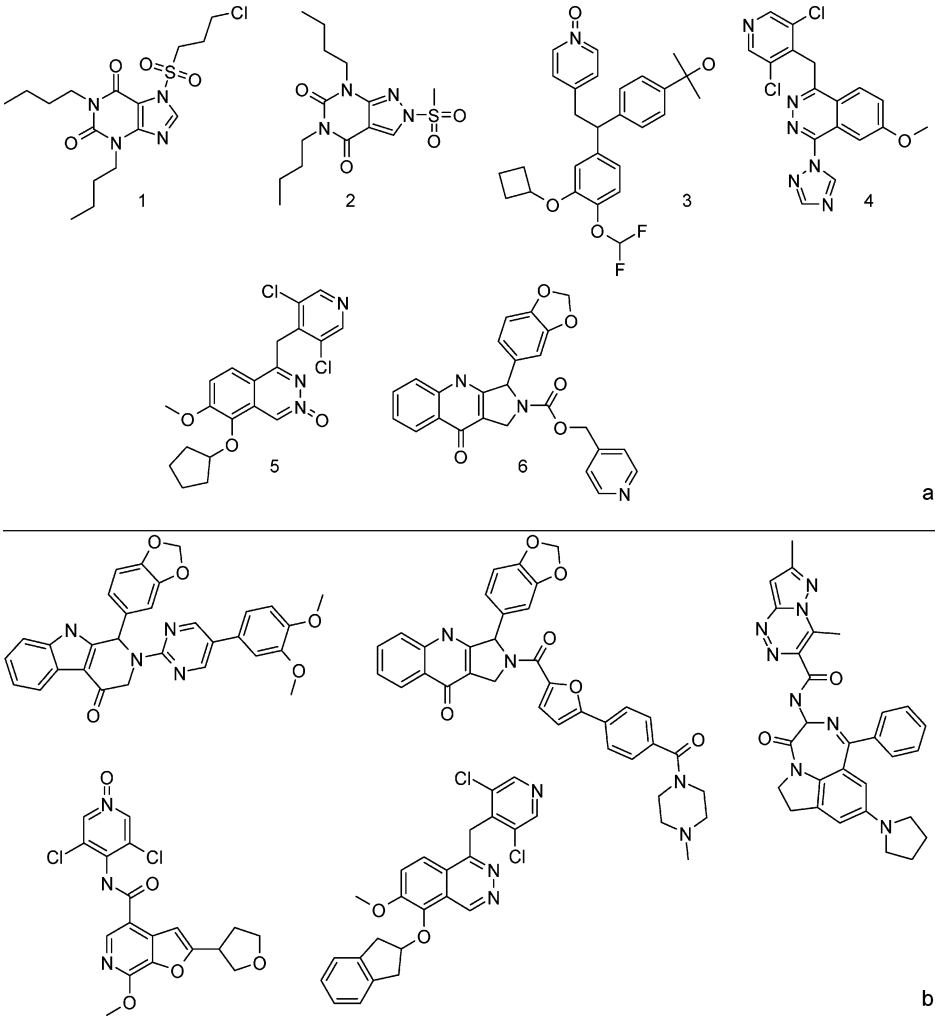
As in the case of PDE 4 inhibitors, although the similarity search recovered almost all known actives, the SOM novelty detection was able to recover additional chemotypes. As can be seen from Chart 5 almost all of the structures shown have a terminal amino or amidine group. This is an example when the use of the side chain and functional groups MeqiSuite indices may be more informative in distinguishing different chemotypes with regards to uPA inhibitory activity.

Based on the above discussion it is clear that the SOM novelty detection method with a 44-dimensional concatenated autocorrelation vector has recovered a significant amount of chemotypes which were missed by the similarity search with Daylight fingerprints. Thus, it is a useful tool for retrieving chemotypes which otherwise would have been missed. On the other hand, the similarity search with Daylight fingerprints has recovered chemotypes missed by the SOM novelty detection. Therefore, as already discussed in section 3.4, the two methods are not by any means orthogonal to each other, and they compliment each other pretty well.

3.6. Rejection Rates. Until now a comparison between the ranked lists obtained with both methods was made. While this is the only kind of results available from a nearest-neighbor based similarity search, novelty detection has the additional capability of immediately classifying as inactive those structures that are outside the space covered by the training set. This one-class classifier offers the advantage of directly rejecting a significant number of patterns without the need of a numerical threshold that always carries some degree of randomness with it. It is commonly accepted[48,49] that a Tanimoto coefficient greater than 0.85, when using binary fingerprints, yields similar compounds. Often, in designing targeted libraries, only one from such pairs of compounds is kept. The validity of this threshold, however, has been subject to criticism.[50] By utilizing novelty detection techniques, no artificial threshold is needed since compounds that are sufficiently far from the chemical space defined by the training set will automatically be classified as novel, i.e., probably inactive. Figure 9 shows the percent of the compounds from the whole WOMBAT data set, which were immediately classified as inactive when using SOM novelty detection with concatenated autocorrelation vector.

The SOM ensemble trained on the PDE 4 training set classified 75% of the remaining WOMBAT structures as inactive. Thus, 101 791 compounds were directly classified

Chart 3. PDE 4 Inhibitors Illustrating Some of the Chemotypes[a]



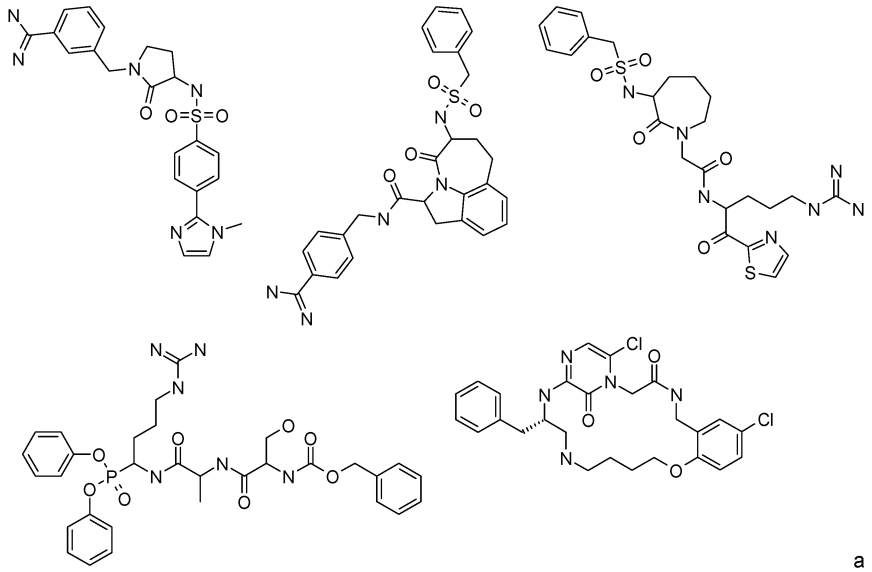
[a] As perceived by the *CyclicSystemOrd* MeqiSuite index and recovered (a) exclusively by SOM novelty detection (method 2a in Figure 1) and (b) exclusively by similarity search with subsequent data fusion (method 1 in Figure 1).

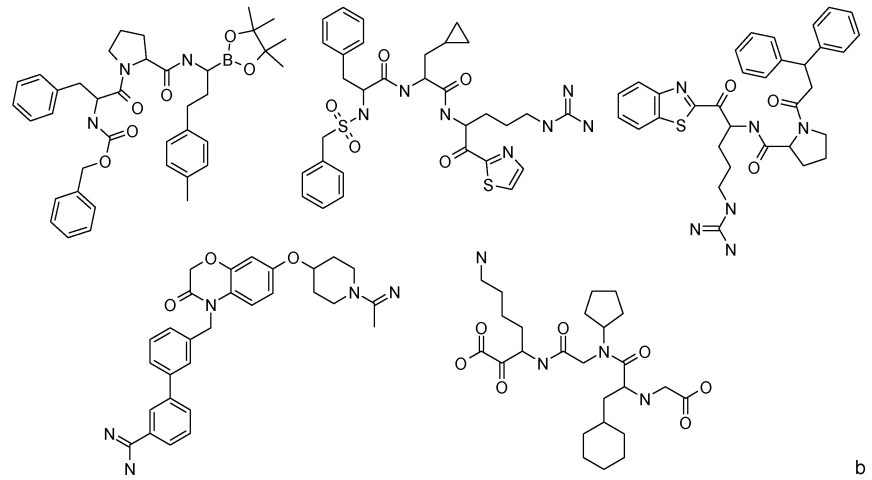
as unlikely to be PDE 4 inhibitors. The smallest number of compounds directly perceived as inactive was obtained for the uPA subclass, which partially explains the lower recall values produced by the SOM novelty detection (cf. Table 2). The number of structures immediately perceived as inactive can be explained well when the structure of the corresponding enzyme is considered. Thus, for the enzymes with a well-defined binding pocket, which can accommodate somehow typical substrates—COX-2, PDE 4, and thrombin—the number of structures immediately perceived as inactive is high. On the other hand, AChE is known to be inhibited by at least two different mechanisms. Here, the number of structures classified as inactive therefore decreases. In the extreme case of uPA, an enzyme with a large binding pocket which can accommodate ligands of different structural types only a small amount (∼22%) of the WOMBAT structures are immediately perceived as inactive.

3.7. Multiple Structure Representations with Mahalanobis Distance. An interesting observation is the fact that this method provides a very tight description of the "active" space when applied to the five activity classes from Table 1. This confirms that indeed the discussed (cf. section 2.3.2.b) deviation of the Mahalanobis distances distribution from the χ^2 has led to a conservative threshold. Figure 10 shows the number of accepted structures for each active subset over the bars and the percent of actives contained in the ranked lists of the given size by the three methods discussed so far. A threshold, equal to $\chi^2(\alpha=0.99, 4)$, which gives a value that encapsulates 99% of the data in the active cluster was used.

The number of accepted structures is very low with any of the active subsets, in the extreme case of uPA only 39 structures of the original 135 877 were classified as belonging to the active space. The percentage of the actives inside the accepted structures is always more than 50%, thus the short lists so obtained are highly enriched in actives—enrichment factors (ef) of 325 for AChE, 236 for COX-2, 489 for PDE4, 92 for thrombin, and 732 for uPA. A comparison with the similarity search with Daylight fingerprints and with the SOM novelty detection with concatenated autocorrelation vectors at these low ranks given on top of the bars in Figure 10 favors SOM novelty detection with multiple structure representations and Mahalanobis distance in the output space method as well. Such a tight description, however, may not be desirable, since a significant number of actives were wrongly rejected and declared as inactive, i.e., the method exhibits a rather high false-negative rate. On the positive side, the high enrichment and the short lists produced may be useful when a limited number of compounds have to be tested, for example in a limited-resource research environ-

Chart 4. Thrombin Inhibitors Illustrating Some of the Chemotypes[a]





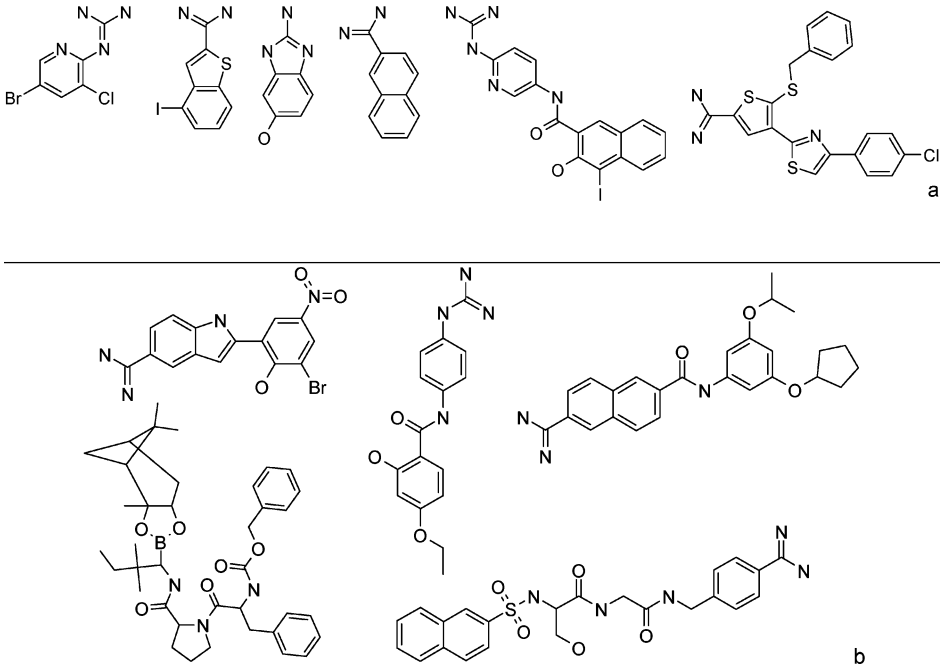
[a] As perceived by the *CyclicSystemOrd* MeqiSuite index and recovered (a) exclusively by SOM novelty detection (method 2a in Figure 1) and (b) exclusively by similarity search with subsequent data fusion (method 1 in Figure 1).

ment. In spite of the small number of accepted structures, these still contain some "false" positives (We put "false" in quotes since the majority of the structures have not been tested against the target enzyme; therefore, we do not factually know if these molecules are inactive.). A close examination of the lists revealed that most of the "false" positives are actually inhibitors with a low activity (above 30 μm) which were left out of consideration when building and testing the novelty detector (cf. Materials and Methods). Furthermore, they contained actives for the same target enzyme but in different or unspecified species. The third and probably most interesting group contained inhibitors of other enzymes. It should be stressed that in general such "false" positives are the main target of virtual screening since they are most likely to become the next lead. To illustrate the above discussion Charts 6 and 7 show some of the "false" positives, i.e., structures which were accepted by the SOM novelty with multiple structure representations although they are not marked as actives in WOMBAT, for the COX-2 and thrombin activity classes. All "false" positive structures were clustered using the *k*-medoids clustering method in R environment[29] with $k = 5$, and the five cluster centers as identified by the algorithm are shown for both activity classes.

The "false" positives and the known actives shown in Charts 2 and 6 show a rather high degree of structural similarity. Thus, it is not surprising that all structures from Chart 6 have been actually tested against COX-2. However, all of them were found to possess activity below 30 μm and therefore were not considered in this study (cf. Materials and Methods).

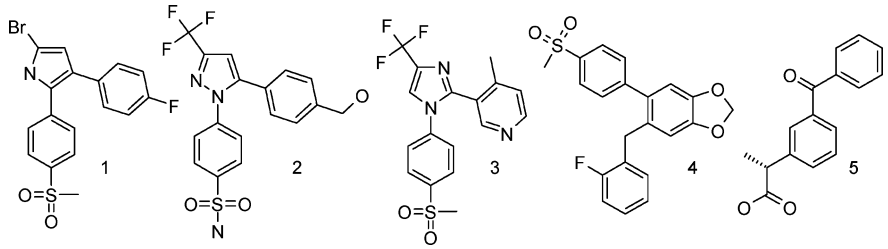
In the case of the thrombin subset, compound **7** from the "false" positives shown in Chart 7 was found highly active against bovine thrombin, while compounds **8** and **9** have been found inactive against human thrombin. Compounds **6** and **10** have not been tested against thrombin, according to the WOMBAT data set. Two conclusions follow from the above observation. First, the method is very good at recovering structurally similar compounds which appear promising in the eyes of a medicinal chemist. However, there is still a lot to be desired in distinguishing pure structural similarity from the cause of a given biological activity. Although the autocorrelation vectors were weighted by different physicochemical properties, the topological nature of these descrip-

Chart 5. uPA Inhibitors Illustrating Some of the Chemotypes[a]



[a] As perceived by the *CyclicSystemOrd* MeqiSuite index and recovered (a) exclusively by SOM novelty detection (method 2a in Figure 1) and (b) exclusively by similarity search with subsequent data fusion (method 1 in Figure 1).

Chart 6. COX-2 "False" Positives (Cluster Centers) as Identified by SOM Novelty Detection with Multiple Structure Representations[a]



[a] Method 2b of Figure 1.

tors together with their low dimensionality limits the approach. Thus, other chemical descriptor sets or better definitions of the active space may prove useful. The proposed method shows promise as a fast and reliable alternative—especially when a short list of putative actives is sought or as a complimentary method to the similarity search with Daylight fingerprints. The SOM novelty with multiple structure representations method also allows the ranking of structures using, e.g., their Mahalanobis distance to the training set. In this manner, structures that are not too distant from the training set could be considered. A comparison with the other two methods at different ranks can be done.

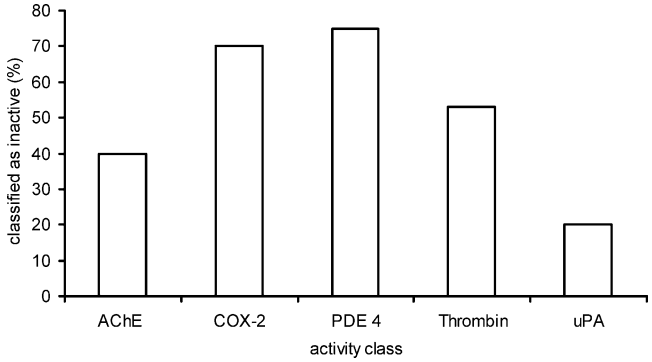


Figure 9. Percent of compounds from the WOMBAT data set immediately classified as inactive by SOM novelty detection with concatenated autocorrelation vector (method 2a in Figure 1).

3.8. Comparison at Different Ranks. Different virtual screening experiments may target a different number of potentially active compounds for a subsequent biological assay. Thus, rather than working at fixed rank, Table 8 shows

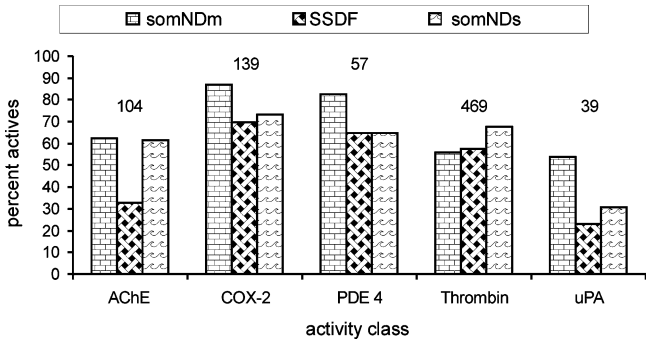
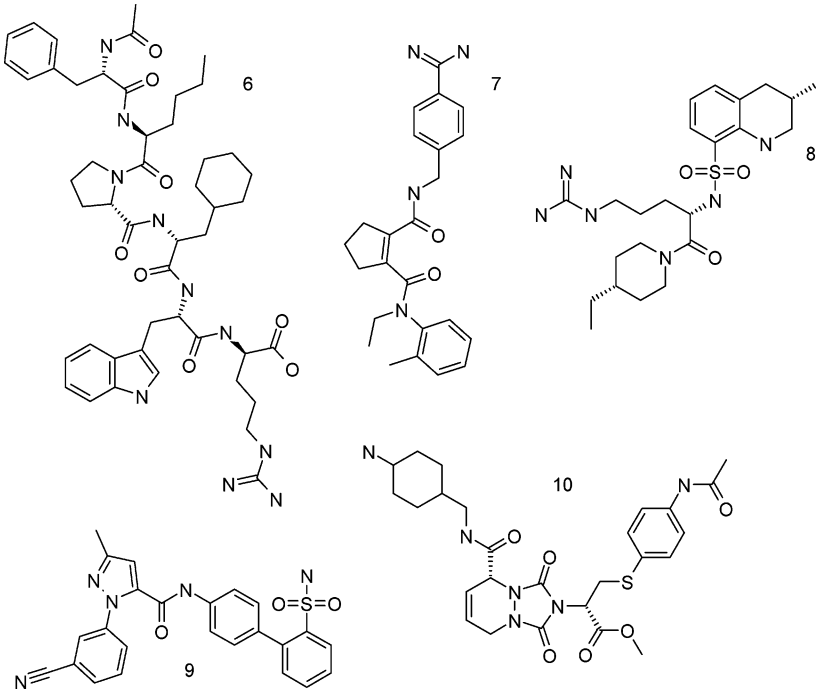


Figure 10. SOM novelty detection with multiple structure representations and Mahalanobis distance (somNDm, method 2b of Figure 1) in the output space as a measure of novelty. The number of structures predicted as active by somNDm is shown above the bars. The height of the bars gives the percent of active compounds among those predicted as active and among the ranked lists of the same size, obtained by the other two methods (SSDF, method 1 of Figure 1, and comNDs, method 2a of Figure 1). The value $\chi^2(\alpha=0.99, 4)$ was used as threshold.

Chart 7. Thrombin "False" Positives (Cluster Centers) as Identified by SOM Novelty Detection with Multiple Structure Representations[a]



[a] Method 2b of Figure 1.

Table 8. Enrichment Factors (ef), Obtained by the Three Methods at Ranks 100 and 1000

		rank 100		rank 1000	
activity class	screening method	no. of actives	ef	no. of actives	ef
AChE (284)	SSDF[a]	32	153	208	100
	somNDs[b]	62	297	188	90
	somNDm[c]	67	321	150	72
COX-2 (500)	SSDF	74	201	312	85
	somNDs	75	204	341	93
	somNDm	93	253	355	96
PDE 4 (229)	SSDF	67	398	216	128
	somNDs	61	362	170	101
	somNDm	79	469	162	96
thrombin (1053)	SSDF	55	71	440	57
	somNDs	76	98	560	72
	somNDm	81	105	391	50
uPA (100)	SSDF	24	326	80	109
	somNDs	24	326	56	76
	somNDm	34	462	55	75

[a] Similarity search with Daylight fingerprints followed by data fusion (method 1 of Figure 1). [b] SOM novelty detection with concatenated autocorrelation vector (method 2a of Figure 1). [c] SOM novelty detection with multiple autocorrelation vectors (method 2b of Figure 1).

the enrichment factors obtained by the three methods at ranks 100 and 1000, while Figure 11 compares the results of the nearest-neighbor similarity search with Daylight fingerprints and the two types of SOM novelty detection with topological autocorrelation descriptors at different ranks.

Both SOM novelty detection methods exhibit similar performance in the case of COX-2, PDE 4, and uPA activity classes, while the SOM novelty detection with concatenated autocorrelation vectors is better in the other two cases, as can be seen from Figure 11. A look at Table 8 confirms the observation (cf. Figure 10) that SOM novelty detection with multiple structure representations gives highly enriched lists at low ranks—it outperforms the other two methods in all cases at rank 100. However, its advantage is lost when increasing the size of the ranked list—at rank 1000 it is outperformed in all cases except of the COX-2 class. Thus, SOM novelty detection with multiple structure representations is preferable when a very small subset of lead candidates is required.

Comparing the SOM novelty detection method using a single concatenated autocorrelation vector with the similarity search with Daylight fingerprints and subsequent data fusion, the SOM novelty detection performs better for COX-2 and thrombin classes, while for the PDE 4 and uPA the similarity search was better. In the case of AChE activity class, the tendency of the SOM novelty detection methods to recover higher number of actives at low ranks can be clearly seen with the fingerprints method catching up at rank 800. The terms "better" and "worse" results are used for comparative purposes. However, these two methods are not mutually exclusive. As has already been shown in the sections 3.4, cf. Figure 7, merging the ranked lists is a valuable way of improving the virtual screening results. On the other hand—as has been demonstrated in section 3.5, the SOM novelty detection succeeded in discovering chemotypes, which are missed by the similarity search with data fusion and vice versa.

4. CONCLUSIONS

Two different methods for novelty detection with Self-Organizing Maps were used for a retrospective ligand-based virtual screening of the WOMBAT database. One method used a single structure representations and data fusion, while another method used multiple representations in concert with a Mahalanobis distance measure. The structures were described by topological autocorrelation functions weighted by atomic physicochemical properties. The results were compared with a traditional similarity search method based on Daylight fingerprints and subsequent data fusion. In addition, different methods for selecting an initial set of targets from

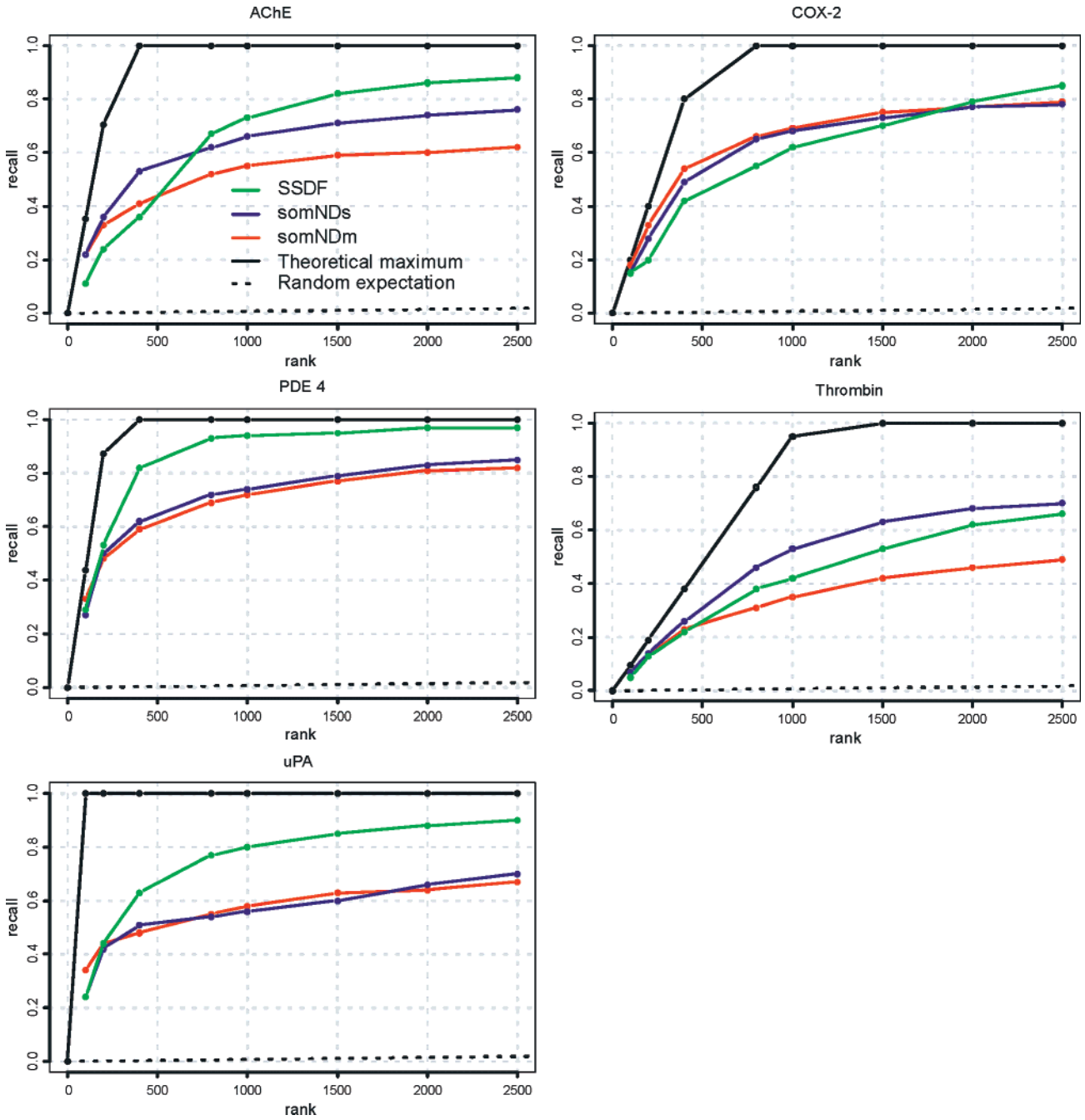


Figure 11. Recall plots for the nearest-neighbor similarity search with Daylight fingerprints (method 1 of Figure 1, green), SOM novelty detection with concatenated topological autocorrelation vectors (method 2a of Figure 1, blue), and SOM novelty detection based on multiple structural descriptors and Mahalanobis distance in the output space (method 2b of Figure 1, red). Also the theoretical maximum (solid black line) and the random expectation (dotted line) are shown.

a larger set were compared. Based on the presented results we conclude the following:

(1) Both SOM novelty detection techniques—with single structure representation (method 2a of Figure 1) and with multiple structure representations (method 2b of Figure 1)—based on topological autocorrelation descriptors are useful for ligand-based virtual screening.

(2) The Taylor−Butina clustering is the method of choice for a subset selection, especially when a small representative set of actives is needed.

(3) Using a 44-dimensional autocorrelation vector and SOM novelty detection gives better (COX-2, thrombin) or comparable results to the similarity search with Daylight fingerprints and data fusion.

(4) Using a 44-dimensional autocorrelation vector and SOM novelty detection is twice as fast compared to the similarity search with Daylight fingerprints and data fusion when a single network is used.

(5) Small sets of compounds highly enriched in active structures can be obtained by considering the intersection between the ranked lists obtained by a combined application of SOM novelty detection with single structure representation and of similarity search with subsequent data fusion.

(6) The SOM novelty detection method with a 44-dimensional concatenated autocorrelation vector complements the Daylight fingerprints based similarity search. Better enriched lists were obtained by merging these results.

(7) The SOM novelty detection method with a 44-dimensional concatenated autocorrelation vector recovered a significant amount of chemotypes which are missed by the similarity search.

(8) The SOM novelty detection method with a 44-dimensional concatenated autocorrelation vector is applicable as a library design tool for discarding a large number of compounds which are unlikely to possess a given biological activity without the need of an artificial threshold.

(9) Using multiple structure representations in concert with a Mahalanobis distance recovers between 34% and 93% of the actives in the top 100 ranked structures. This corresponds to enrichment factors between 105 and 470. Thus, it is the recommended method when a short list of lead candidates is required.

ACKNOWLEDGMENT

Part of this work was funded by National Institutes of Health grant U54 MH074425-01 (National Institutes of Health Molecular Libraries Screening Center Network) and by the New Mexico Tobacco Settlement Fund (T.I.O.).

REFERENCES AND NOTES

(1) Lyne, P. D. Structure-Based Virtual Screening: an Overview. *Drug Discovery Today* **2002**, *7*, 1047−1055.
(2) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A Review of Protein-Small Molecule Docking Methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151−166.
(3) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Model.* **1998**, *38*, 983−996.
(4) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Model.* **1996**, *36*, 118−127.
(5) Fechner, U.; Franke, L.; Renner, S.; Schneider, P.; Schneider, G. Comparison of Correlation Vector Methods for Ligand-Based Similarity Searching. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 687−698.
(6) Evers, A.; Hessler, G.; Matter, H.; Klabunde, T. Virtual Screening of Biogenic Amine-Binding G-Protein Coupled Receptors: Comparative Evaluation of Protein- and Ligand-Based Virtual Screening Protocols. *J. Med. Chem.* **2005**, *48*, 5448−5465.
(7) Nikolova, N.; Jaworska, J. Approaches to Measure Chemical Similarity - a Review. *QSAR Comb. Sci.* **2003**, *22*, 1006−1026.
(8) Whittle, M.; Willett, P.; Klaffke, W.; van Noort, P. Evaluation of Similarity Measures for Searching the Dictionary of Natural Products Database. *J. Chem. Inf. Model.* **2003**, *43*, 449−457.
(9) Gillet, V. J.; Willett, P.; Bradshaw, J. Similarity Searching Using Reduced Graphs. *J. Chem. Inf. Model.* **2003**, *43*, 338−345.
(10) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical Similarity Using Geometric Atom Pair Descriptors. *J. Chem. Inf. Model.* **1996**, *36*, 128−136.
(11) Whittle, M.; Gillet, V. J.; Willett, P.; Alex, A.; Loesel, J. Enhancing the Effectiveness of Virtual Screening by Fusing Nearest Neighbor Lists: A Comparison of Similarity Coefficients. *J. Chem. Inf. Model.* **2004**, *44*, 1840−1848.
(12) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Model.* **2004**, *44*, 1177−1185.
(13) Bender, A.; Jenkins, J. L.; Glick, M.; Deng, Z.; Nettles, J. H.; Davies, J. W. "Bayes Affinity Fingerprints" Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When Are Multitarget Drugs a Feasible Concept? *J. Chem. Inf. Model.* **2006**, *46*, 2445−2456.
(14) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the Effectiveness of Similarity-Based Virtual Screening Using Nearest-Neighbor Information. *J. Med. Chem.* **2005**, *48*, 7049−7054.
(15) Marsland, S. Novelty Detection in Learning Systems. *Neural Comput. Surv.* **2003**, *3*, 157−195.
(16) Markou, M.; Singh, S. Novelty Detection: a Review - Part 1: Statistical Approaches. *Signal Process.* **2003**, *83*, 2481−2497.
(17) Markou, M.; Singh, S. Novelty Detection: a Review - Part 2: Neural Network Based Approaches. *Signal Process.* **2003**, *83*, 2499−2521.
(18) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating Biologically Active Compounds in Medium-Sized Heterogeneous Data Sets by Topological Autocorrelation Vectors: Dopamine and Benzodiazepine Agonists. *J. Chem. Inf. Model.* **1996**, *36*, 1205−1213.
(19) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, 1999.
(20) Vesanto, J. SOM-Based Data Visualization Methods. *Intell. Data Anal.* **1996**, *3*, 111−126.
(21) Zhang, S.; Ganesan, R.; Xistris, G. D. Self-Organising Neural Networks for Automated Machinery Monitoring Systems. *Mech. Syst. Signal. Pr.* **1996**, *10*, 517−532.
(22) Wong, M. L. D.; Jack, L. B.; Nandi, A. K. Modified Self-Organising Map for Automated Novelty Detection Applied to Vibration Signal Monitoring. *Mech. Syst. Signal. Pr.* **2006**, *20*, 593−610.
(23) Noeske, T.; Sasse, B. C.; Stark, H.; Parsons, C. G.; Weil, T.; Schneider, G. Predicting Compound Selectivity by Self-Organizing Maps: Cross-Activities of Metabotropic Glutamate Receptor Antagonists. *ChemMedChem* **2006**, *1*, 1066−1068.
(24) Teckentrup, A.; Briem, H.; Gasteiger, J. Mining High-Throughput Screening Data of Combinatorial Libraries: Development of a Filter to Distinguish Hits From Nonhits. *J. Chem. Inf. Model.* **2004**, *44*, 626−634.
(25) Selzer, P.; Ertl, P. Applications of Self-Organizing Neural Networks in Virtual Screening and Diversity Selection. *J. Chem. Inf. Model.* **2006**, *46*, 2319−2323.
(26) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Cheminformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: New York, 2003; pp 223−241
(27) Taylor, R. Simulation Analysis of Experimental Design Strategies for Screening Random Compounds As Potential New Drugs and Agrochemicals. *J. Chem. Inf. Model.* **1995**, *35*, 59−67.
(28) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Model.* **1999**, *39*, 747−750.
(29) *R Development Core Team. R: A language and environment for statistical computing*, *version 2.2.1*; 2005. http://www.r-project.org (accessed Jan 1, 2006).
(30) Daylight Chemical Information Systems Inc. http://www.daylight.com/dayhtml/doc/theory/theory.finger.html (accessed June 2006).
(31) Sykora, V. *Chemical Descriptors Library*. http://cdelib.sourceforge.net (accessed June 2006)
(32) Moreau, G.; Broto, P. Autocorrelation of a Topological Structure: A New Molecular Descriptor. *New J. Chem.* **1980**, *4*, 359−360.
(33) Spycher, S.; Nendza, M.; Gasteiger, J. Comparison of Different Classification Methods Applied to a Mode of Toxic Action Data Set. *QSAR Comb. Sci.* **2004**, *23*, 779−791.
(34) *PETRA - Parameter Estimation for the Treatment of Reactivity Applications*, *version 4.0*; Molecular Networks GmbH: Erlangen, Germany, 2006. http://www.molecular-networks.com (accessed June 2006).
(35) Hutchings, M. G.; Gasteiger, J. Residual Electronegativity - an Empirical Quantification of Polar Influences and Its Application to the Proton Affinity of Amines. *Tetrahedron Lett.* **1983**, *24*, 2541−2544.
(36) Gasteiger, J.; Hutchings, M. G. New Empirical Models of Substituent Polarisability and Their Application to Stabilisation Effects in Positively Charged Species. *Tetrahedron Lett.* **1983**, *24*, 2537−2540.
(37) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.
(38) Hollas, B. An Analysis of the Autocorrelation Descriptor for Molecules. *J. Math. Chem.* **2003**, *33*, 91−101.
(39) *ADRIANA.Code*, *version 1.0*; Molecular Networks GmbH: Erlangen, Germany, 2006. http://www.molecular-networks.com (accessed June 2006).
(40) De Maesschalck, R.; Jouan-Rimbaud, D.; Massart, D. L. The Mahalanobis Distance. *Chemom. Intell. Lab.* **2000**, *50*, 1−18.
(41) Vesanto, J.; Sulkava, M.; Hollmén, J. On the Decomposition of the Self-Organizing Map Distortion Measure. In *Proceedings of the Workshop on Self-Organizing Maps (WSOM'03)*; Kitakyushu, Japan, 2003; pp 11−16.
(42) Pölzlbauer, G. Survey and Comparison of Quality Measures for Self-Organizing Maps. In *Proceedings of the Fifth Workshop on Data Analysis (WDA'04)*; Sliezsky dom, Vysoké Tatry, Slovakia, 2004, Paralic, J., Pölzlbauer, G., Rauber, A., Eds.; Elfa Academic Press: Košice, 2004.

(43) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *J. Mol. Graphics Modell.* **2000**, *18*, 343−357.

(44) Olah, M.; Bologa, C.; Oprea, T. I. Strategies for Compound Selection. *Curr. Drug Discovery Technol.* **2004**, *1*, 211−220.

(45) Murdoch, D. Venn Diagrams in R. *J. Statist. Soft.* **2004**, *11*, Code Snippet 1.

(46) *MeqiSuite, version 2.30*; Pannanugget Consulting L.L.C.: Kalamazoo, MI, U.S.A.. http://www.pannanugget.com (accessed Jan 20, 2007).

(47) Johnson, M. *An Introduction to the MeqiSuite Indices;* Technical Report 2006/001; Pannanugget Consulting, Inc.: Kalamazoo, MI, U.S.A., 2006.

(48) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(49) Matter, H. Selecting Optimally Diverse Compounds From Structure Databases: A Validation Study of Two-Dimensional and Three-Dimensional Molecular Descriptors. *J. Med. Chem.* **1997**, *40*, 1219−1229.

(50) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

CI700040R