J Comput Aided Mol Des (2007) 21:617–640
DOI 10.1007/s10822-007-9145-8

Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios

Dimitar P. Hristozov · Tudor I. Oprea · Johann Gasteiger

Received: 30 August 2007 / Accepted: 25 October 2007 / Published online: 16 November 2007
© Springer Science+Business Media B.V. 2007

Abstract Four different ligand-based virtual screening scenarios are studied: (1) prioritizing compounds for subsequent high-throughput screening (HTS); (2) selecting a predefined (small) number of potentially active compounds from a large chemical database; (3) assessing the probability that a given structure will exhibit a given activity; (4) selecting the most active structure(s) for a biological assay. Each of the four scenarios is exemplified by performing retrospective ligand-based virtual screening for eight different biological targets using two large databases—MDDR and WOMBAT. A comparison between the chemical spaces covered by these two databases is presented. The performance of two techniques for ligand-based virtual screening—similarity search with subsequent data fusion (SSDF) and novelty detection with Self-Organizing Maps (ndSOM) is investigated. Three different structure representations—2,048-dimensional Daylight fingerprints, topological autocorrelation weighted by atomic physicochemical properties (sigma electronegativity, polarizability, partial charge, and identity) and radial distribution functions weighted by the same atomic physicochemical properties—are compared. Both methods were found applicable in scenario one. The similarity search was found to perform slightly better in scenario two while the SOM novelty detection is preferred in scenario three. No method/descriptor combination achieved significant success in scenario four.

Keywords Ligand-based virtual screening · Novelty detection · Similarity search · Data fusion · ROC curve · BEDROC evaluation · Structure representation

D. P. Hristozov · J. Gasteiger
Computer-Chemie-Centrum, Universität Erlangen-Nürnberg, Nägelsbachstr. 25, 91052 Erlangen, Germany

T. I. Oprea
Division of Biocomputing, University of New Mexico School of Medicine, MSC 11 6145, 1 University of New Mexico, Albuquerque, NM 87131-0001, USA

J. Gasteiger (✉)
Molecular Networks GmbH - Computerchemie, Henkestraße 91, 91052 Erlangen, Germany
e-mail: Gasteiger@molecular-networks.com

Introduction

Virtual screening of compound libraries is often employed for the selection of subsets of chemical structures, which are enriched in active compounds [1–4]. Ligand-based methods are frequently applied when the 3D structure of the biological target of interest is unknown. The ligand-based methods rely on a representative set of reference structures, molecular descriptors, and an appropriate similarity measure [2, 5]. Usually the result of these virtual screening methods is a ranked list of the screened compounds. Highly ranked compounds in such a list are assumed to share the activity of the reference structures.

The retrieval of relevant structures on the top of the ranked list is the broad aim of virtual screening experiments. This aim can be further split into more concrete scenarios, depending on the available resources and on the concrete goal of the experiments. In this work, we have identified and will discuss four such scenarios.

Scenario 1: prioritizing compounds for a subsequent HTS (SC.1)

In this scenario a large database of potentially active compounds is screened and the compounds are ordered in

descending order according to the score assigned by the virtual screening method. Afterwards a certain percentage of the top-ranked compounds are selected and evaluated in a high-throughput screening (HTS) campaign. The assumption here is, as the high-throughput word implies, that a large amount of the compounds in the original database will be screened. Proprietary compound libraries containing millions of structures are nowadays common in the pharmaceutical industry [6]. Therefore, eliminating even a small percentage may decrease the costs in an HTS campaign significantly. In such cases it is important that a virtual screening method is able to guarantee that no potential active structures will be missed, i.e., a small false-negative ratio is required. Since still a relatively large number of compounds will be selected, it is important to determine the size of the ranked list, which provides the best trade-off between recovered active structures (true-positives) and structures of unknown activity (false-positives). A useful virtual screening method in this scenario has to perform better than a random picking of compounds.

Scenario 2: selecting compounds for a subsequent lead-optimization (SC.2)

In this scenario even halving the number of available compounds is not likely to be sufficient. In contrast to a HTS campaign the lead-optimization process requires a decent amount of human intervention. Asking a medicinal chemist to work with half a million compounds is not reasonable. Therefore, a small amount of possibly active compounds is required. In such a scenario a virtual screening method should not only guarantee a performance better than random selection but also that as many actives as possible are retrieved in the beginning of the ranked list—the so called "early recognition" problem. This problem is even more pronounced in a university research laboratory where the resources are usually limited. In this scenario it is irrelevant if all actives have been retrieved in the remaining part of the list (after the predefined number of compounds has been reached) since this part will not be examined. The (relatively) low number of considered structures usually makes an examination of the retrieved compounds by a chemist possible. This scenario—using a different predefined number of the considered compounds, usually between 1 and 10% of the screened database—has been by far the most common application of different virtual screening studies [7–12].

Scenario 3: is a given compound active? (SC.3)

Another possible use of a virtual screening method is to assess the probability that a given structure is active. That is, instead of screening a large dataset, can we provide an answer for a single, already available structure? This question arises often in library design problems. When deciding which structures to include into the chemical library being designed it is beneficial to assess the probability that these structures will exhibit a certain activity [13]. Then, depending on the goal of the structure library which is being designed either structures close to the chemical space of the compounds known to be active or rather far from this space may be purchased.

Usually the above question is answered with the help of a qualitative or quantitative structure-activity relationship (QSAR) model, dedicated to the activity in question. It should be stressed that any QSAR model can be used as a virtual screening device. However, such models are usually built using a limited amount of training data and are unable to extrapolate too far from the chemical space covered by these training data. Therefore, it will be beneficial if a virtual screening method—which is usually faster and does not attempt to make quantitative predictions—can be applied in such a context. A common approach in this direction is to threshold the used similarity metric at a given (perhaps arbitrary) value [14, 15]. When more than one known actives are used (through data fusion) the problem of determining the value of the threshold may become more pronounced. An attractive approach, which handles SC.3 implicitly, is based on so-called novelty detection techniques [16, 17] or one-class classifiers [18].

Scenario 4: identification of the most active compound (SC.4)

A possible application of a virtual screening method is the selection of a "best" structure out of a set of potentially active compounds. The "best" in this context is defined as the structure which will have the highest activity. This question is usually answered with the help of a quantitative structure-activity model. However, provided that a virtual screening experiment has been performed, it may be possible to determine such a "best" structure using only the returned ranked list. In this scenario, the actual ranking of the potentially active structures is important with the assumption that the higher a compound is ranked the higher its activity is. This kind of experiment is common in a structure-based virtual screening experiments based on docking. In these experiments the docking scores are usually correlated with the activity values [19] and a number of references [20, 21] providing overview of the performance and comparison between different docking algorithms exist. However, docking experiments require that the 3D structure of the target is known.

The aim of the present work was to test the applicability of different ligand-based virtual screening methods and chemical structure representations in each of the above scenarios. Thus, in the rest of this paper we first shortly describe the used chemical databases and two different methods for ligand-based virtual screening—similarity search with subsequent data fusion and novelty detection with self-organizing maps. Next, we introduce the measures which will be used to assess the success of our retrospective virtual screening experiments. Then, the results of the application of each of these methods for retrospective virtual screening for eight different biological targets in two large databases—MDL Drug Data Report (MDDR) [22] and WOrld of Molecular BioAcTivity (WOMBAT) [23] using different types of structural descriptors are presented. Different aspects of the results: the optimal size of the training set, the difference in the chemical spaces covered by MDDR and WOMBAT, the bias introduced by the training set selection, the differences in the compounds recovered by different methods or/and descriptors are discussed and the best method–descriptor combination is identified for each scenario.

Materials and methods

Chemical databases

The two databases—MDL Drug Data Report (MDDR) version 2006.1 and World Of Molecular BioAcTivity (WOMBAT), version 2006.01, were used.

MDDR, version 2006.1, contains 159,662 structures together with the associated activity classes. A total of 149,414 molecules were used after removal of duplicates and molecules that could not be processed by some of the used computer programs.

WOMBAT, version 2006.01, contains 154,236 chemical compounds, collected from articles in medicinal chemistry journals published between 1975 and 2006. In addition to the structural information, WOMBAT contains also the reported activity values, expressed as pK_i value, information about the species in which the tests were performed, the biological role of the structure (inhibitor, antagonist, etc.) as well as additional properties of interest. A total of 118,346 chemical structures were used after removing duplicates, the structures which were also found in MDDR, the structures with reported activity less than 30 μmolar, and molecules that could not be processed by some of the used computer programs.

Biological targets

Eight different biological targets were subjected to retrospective virtual screening and are summarized in Table 1. These targets are quite diverse and are common subjects to ligand-based virtual screening experiments [24–26]. The activity classes are referred to by using their WOMBAT activity IDs through the rest of this article.

Virtual screening protocol

The set of known active structures (referred to as "training set" from here on) was always selected from MDDR. The compounds selected as training set were removed from MDDR and were not considered when evaluating the performance. The training set selection consisted of (1) clustering the known actives in MDDR using the Taylor–Butina [27, 28] clustering algorithm; (2) randomly selecting a percentage of known actives from each cluster as such as the total size of the training set equals a predefined number; (3) merging the structures from MDDR and WOMBAT and subjecting the resulting database (which contained 267,760 unique structures) to a similarity search with subsequent data fusion (SSDF) and to novelty detection with SOM (ndSOM) using the selected training set; (4) measuring separately the ability of each method/descriptor combination to recover the active structures from MDDR and from WOMBAT. The results of a single experiment following the described

Table 1 Subsets of active structures used in this study

Activity name	MDDR		WOMBAT	
	Activity ID	# Actives	Activity ID	# Actives
5HT3 antagonists	06233	775	5-HT3	635
5HT1A agonists	06235	953	5-HT1A	2,524
D2 antagonists	07701	487	D2	2,877
Renin inhibitors	31420	1,188	Renin	583
Angiotensin II AT1 antagonists	31432	2,158	AT1	1,361
Thrombin inhibitors	37110	1,122	Thrombin	1,841
HIV protease inhibitors	71523	971	HIV-1 P	2,422
Protein kinase C inhibitors	78374	545	PKC	152

procedure are dependent on the particular training set and may result in too optimistic (or too pessimistic) performance estimates. To eliminate this effect, Step 2 and 3 were repeated 10 times and the means and standard deviations of the corresponding performance metrics were calculated. The actives from MDDR and from WOMBAT were separated in step 4 and the performance was assessed separately in order to investigate the bias introduced by selection of the training set from a particular database (MDDR in this case).

The above protocol was augmented with an additional bootstrapping step when evaluating the "early recognition" capabilities of the studied virtual screening methods. After a ranked list was obtained a number of known active structures were repeatedly removed at random 10 times in a way which leaves only 150 known actives for the subsequent evaluation. This procedure was suggested by Truchon and Bayly [29] and ensures that the BEDROC evaluation (see below) is not prone to saturation effects. The saturation effect appears when there are too many known actives in the beginning of the ranked list, which leads to an early recognition metric with low discriminative power between methods. In addition, the bootstrapped evaluation gives a better estimate of the standard deviation associated with the BEDROC metric.

Assessing the performance

The performance of a virtual screening method in each of the discussed scenarios cannot be assessed by the same metric. A survey of the available performance measures focused mainly on SC.1 and SC.2 can be found in Ref. [30]. Recently, a detailed investigation of the performance metrics, focused on the "early recognition" problem (SC.2 in the present work) was reported [29].

ROC curves

The area under the Receiver Operating Characteristic (ROC) curve [31, 32] is an adequate scalar measure of the performance in SC.1. At the same time, the ROC curve itself can be used to find the best trade-off between false and true positives. The use of ROC curves for assessing the performance of virtual screening experiments is well-established [33, 34]. A ROC curve represents a plot of the number of true active compounds (true positives) included in the sample on the vertical axis, expressed as a percentage of the total number of known actives, against the number of structures with unknown activity (false positives) included in the sample, expressed as percentage of the total number of structures with unknown activity, on the horizontal axis. Sample ROC curves and their corresponding areas are presented on Fig. 1.

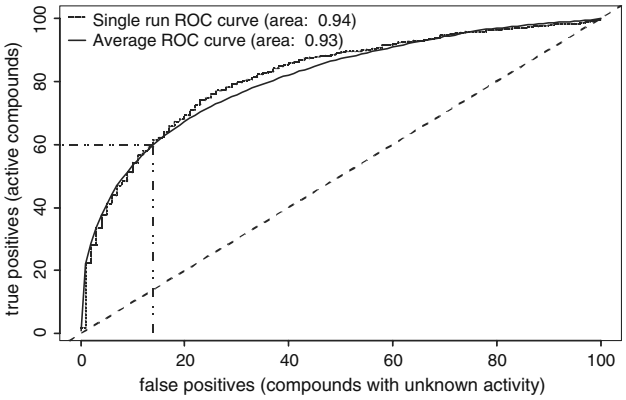


Fig. 1 Sample receiver operating characteristic (ROC) curves. The jagged ROC curve is obtained in a single run, while the smoothed ROC curve results from averaging 10 runs with different training sets. The diagonal line shows the expected ROC curve for random picking. The selection of the desired number of compounds is illustrated assuming that 60% of the known actives should be recovered

BEDROC evaluation

In a recent survey on evaluating virtual screening experiments focused on the early recognition problem Truchon and Bayly [29] proposed the use of Boltzmann-Enhanced Discrimination of ROC (BEDROC) metric. The BEDROC metric is calculated according to Eq. 1

$$BEDROC = \frac{\sum_{i=1}^{n} e^{-\alpha r_i/N}}{\frac{n}{N}\left(\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}\right)} \times \frac{R_a \sinh(\alpha/2)}{\cosh(\alpha/2) - \cosh(\alpha/2 - \alpha R_a)} + \frac{1}{1 - e^{\alpha(1-R_a)}} \quad (1)$$

where n is the number of known active structures, N is the number of inactive (or with unknown activity) structures, r_i is the rank of the ith active structure, R_a is the ratio of active to inactive structures n/N, and α is a weighting factor, which controls the "early recognition" element—higher α values move the region of importance towards the beginning of the ranked list.

The main advantage of BEDROC over the ROC area is that higher weight is given to actives recovered early in the list compared to actives recovered towards the end of the list. This fact makes it very suitable for performance estimation in SC.2.

For a detailed description of this metric together with a mathematical derivation and proofs the reader is referred to the original work of Truchon and Bayly [29]. Following the guidance in Ref. [29] in the present work an α value of 32.2 was used. This value means that 80% of the corresponding BEDROC score was based on the top-ranked five per cents of the compounds in the original database.

Recall and precision

The standard metrics for assessing the performance of a classification task—recall and precision [35], can be applied for evaluating the performance in SC.3. These metrics can be calculated from a 2 × 2 contingency table, such as the one shown in Table 2.

The recall measures the percent of active structures retrieved at a given size of the ranked list (r) and is calculated according to Eq. 2

$$recall_r = \frac{N_{ar}}{N_a} \tag{2}$$

The recall as calculated with Eq. 2 is bound between zero and one. The higher the recall value is the better the performance.

The precision measures how many of the retrieved structures are actually active and is calculated according to Eq. 3:

$$precision_r = \frac{N_{ar}}{N_r} \tag{3}$$

Similarly to the recall, the precision is bound between zero and one and higher values signify better performance.

Kendall's rank correlation coefficient (τ)

There are a couple of so-called rank agreement measures, which can be utilized to measure the agreement between two ranked lists and, therefore, the performance in SC.4. These include rank correlations, such as Spearman's ρ or Kendall's τ, the ndpm measure [36], etc. We have selected the Kendall's τ for the evaluation of SC.4 mainly because it handles weak ordering slightly better than Spearman's ρ. It is calculated according to Eq. 4

$$\tau = \frac{C - D}{\sqrt{(C + D + TR) \times (C + D + TP)}} \tag{4}$$

In this equation, C stands for the number of concordant pairs—pairs of structures that the virtual screening method predicts in the properly ranked order. D stands for the number of discordant pairs—pairs that the virtual screening method predicts in a wrong order. TR is the number of pairs of structures in the true ordering (the ranking determined by the activity values) that have tied ranks (i.e., the same activity) while TP is the number of pairs of structures in the predicted ordering that have tied ranks (the same similarity coefficient).

Kendall's τ calculated with Eq. 4 has the following properties: (1) If the agreement between the two rankings is perfect (i.e., the two rankings are the same) the coefficient has the value of one; (2) If the disagreement between the two rankings is complete (i.e., one ranking is the reverse of the other) the coefficient has a value of minus one; (3) For all other arrangements the value lies between -1 and 1, and increasing values imply increasing agreement between the rankings. If the rankings are completely independent, the coefficient has a value of zero.

Table 2 A 2 by 2 contingency table used to calculate different performance metrics

	Retrieved	Not retrieved	Total
Active	N_{ar}	N_{an}	N_a
Inactive	N_{ir}	N_{in}	N_i
Total	N_r	N_n	N

Virtual screening methods

Similarity search with subsequent data fusion (SSDF)

The similarity search starts with a known active structure, usually called "target" or "query". After this structure has been described by a given representation all compounds from the screened database are compared to it by means of a similarity coefficient. The screened database is then sorted in descending order according to the values of the similarity coefficient. The compounds most similar to the query end up on top of this list.

The procedure described so far requires a single known active. Data fusion is usually applied [37, 38] to adapt it to a case when more than one known active structure is available. Starting with n known actives (training set) a similarity search is carried out with each of them in turn. This results in n separate ranked lists. There are different methods to combine the similarity scores from these lists, called "fusion rules". 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 ranked lists, according to Eq. 5

$$S_{FUS}(j) = \max_i [S^*(i,j)] \tag{5}$$

where S^* denotes the calculated similarity score between the query structure i and the screened structure j. The similarity score used in this work was the Tanimoto coefficient. The Tanimoto coefficient for binary structure representation assumes values between zero and one and is calculated according to Eq. 6:

$$S_T = \frac{c}{a + b - c} \tag{6}$$

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 above formula can be adapted to real-valued structure representations as shown in Eq. 7

$$S(x;y) = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} (x_i)^2 + \sum_{i=1}^{N} (y_i)^2 + \sum_{i=1}^{N} x_i y_i} \quad (7)$$

where x and y are the corresponding real valued vectors of size N describing structures X and Y. The coefficient thus calculated assumes values between -0.333 and one.

Novelty detection with Self-Organizing Maps (ndSOM)

A detailed description of two variants of this virtual screening method has been reported recently [18]. In this work, ndSOM with a single structure representation was used. Briefly, the method requires a set of known actives (training set) from which a Self-Organizing Map (SOM) is built. The SOM belongs to the class of unsupervised learning methods. The typical uses of SOM are for dimensionality reduction and visualization. However, a SOM can also serve as a description of the space covered by the dataset used to build it. Once the SOM has been obtained a local accuracy is determined from the map and the screened database is projected onto this SOM. If the distance between a screened structure and its best matching neuron (BMN) is larger than the local accuracy this structure is deemed novel, that is, unlikely to share the biological activity of the training set.

In our previous work [18], the local accuracy was determined using the average distance between a neuron and its neighbors. However, further experiments have shown that this is prone to rather large fluctuations from a map to map due to spots of low data density, which create "cliffs" in the map. Therefore, in the current work a global accuracy (*ga*) rather than a local accuracy was used. To obtain the global accuracy once the SOM has been trained the training set is projected onto it. The distances between each training structure and its best-matching neuron are calculated. The largest such distance can be used as a global accuracy. However, if an outlier is present in the training set, the maximum distance may become very large. To avoid this, a global accuracy, which classifies approximately 5% of the training structures as novel, i.e., inactive, was used. After the global accuracy was obtained, the screened structures were scored according to Eq. 8

$$S_{SOM}(j) = \frac{ga}{ga + d(j, BMN_j)} \quad (8)$$

where ga is the global accuracy of the map and $d(j, BMN_j)$ denotes the distance between the structure j from the screened set and its best-matching neuron. The score obtained with Eq. 8 is bound between zero and one. Any screened structure for which the calculated score has a value below 0.5 is classified as inactive. This modification eliminates the fluctuations from a map to map significantly and allowed us to use a single rather than 10 SOMs, as described previously. All used SOMs were with hexagonal lattice and were trained using the batch algorithm [39].

Self-organizing map for binary structure descriptors The classic SOM algorithm has been developed for working with real valued vectors. To provide a fair comparison with SSDF with binary fingerprints and to study the applicability of ndSOM with such kind of descriptors we have implemented a version of the Self-Organizing Map algorithm capable of handling binary data.

The easiest way to train a SOM using binary fingerprints is to regard them as real valued descriptor, utilize the standard SOM algorithm and possibly transform the adapted weights back to binary strings at the end of the training. However, this algorithm can be very slow, provided the high dimensionality of the binary fingerprints. Another possibility is to adapt a version of a batch SOM algorithm, initially designed for building SOM on string data [39]. The main idea is to update the weights based on the generalized median of the set of binary fingerprints, S, which form the Voronoi region of the winning neuron and its neighborhood.

The generalized median is a binary string which has the minimum sum of distances to all other binary strings in the set. That is, it minimizes

$$\bar{p} = \arg\min_{p \in U} \sum_{q \in S} d(p, q) \quad (9)$$

where U is the set of all possible binary strings of the given length and $d(p,q)$ is the distance between two binary patterns p and q. By using this algorithm, a ndSOM can be easily applied to binary structure representations. We have used the Tanimoto similarity coefficient as the distance measure, d, in Eq. 9 after subtracting it from 1. Alternatively, one can replace *argmin* in the above equation with *argmax*.

Structure representation

Binary fingerprints

2,048-dimensional Daylight fingerprints (DFP) were generated with the Chemical Descriptors Library [40].

Topological autocorrelation (AC2D)

Introduced by Moreau and Broto [41] the topological autocorrelation descriptors have since then been applied in

a number of studies [42–46]. The descriptors are calculated according to Eq. 10

$$A(d) = \sum_{i=1}^{k} \sum_{j=i}^{k} p_i p_j \delta(d - d_{ij}) \quad (10)$$

Here 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 by previously published empirical methods for all atoms in a molecule: sigma electronegativity (χ_σ) [47], effective atom polarizability (α_d) [48], and partial atomic charge (q_{tot}) [49]. 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. 10. The scaling has been shown [50] to diminish the correlations between the bins of autocorrelation and to better preserve the physicochemical information. The resulting autocorrelation vectors were additionally autoscaled to ensure that the values are comparable when a distance measure (such as the Euclidian distance, used by SOM) is calculated. The final, 44-dimensional description of the chemical structures was achieved by concatenation of the autocorrelations vectors thus calculated. The topological autocorrelation vectors were calculated with the descriptor calculation package ADRIANA.Code [51].

Radial distribution function (RDF)

A radial distribution function is a transform of the 3-dimensional structure of a molecule [52]. All structures were represented by their RDF codes weighted by the same physicochemical properties as the topological autocorrelation—sigma electronegativity (χ_σ), effective atom polarizability (α), partial atomic charge (q_{tot}), and identity (A1). Single, low energy 3D conformations were generated from the 2D constitution using the 3D structure generator CORINA [53, 54]. The RDF codes were calculated according to Eq. 11

$$g(r) = \sum_{i=1}^{N-1} \sum_{j>i}^{N} p_i p_j e^{-B(r-r_{ij})^2} \quad (11)$$

where N is the number of atoms in a molecule, p_i and p_j are properties associated with the atoms i and j, respectively, r_{ij} represents the distance between atoms i and j, and B is a smoothing factor. The above formula was applied with the property p set to each of the four physicochemical properties in turn and 64 dimensional RDF codes were calculated. Analogously to the autocorrelation vectors, the atomic properties for each molecule, with exception of the identity, were auto scaled to zero mean and unit variance before applying Eq. 11. The function $g(r)$ was defined in the interval 0.8–12 Å. The resulting RDF codes were additionally auto-scaled. The final, 256-dimensional description of the chemical structures was achieved by concatenation of the four so RDF codes thus calculated. The RDF codes were calculated with the descriptor calculation package ADRIANA.Code [51].

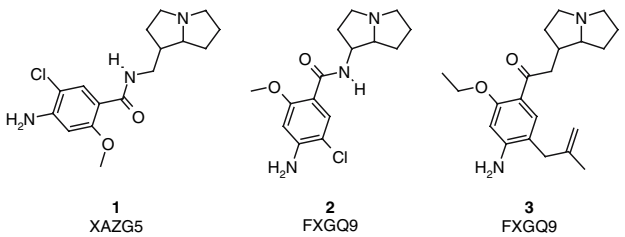
Chemotype analysis

The computer program MeqiLite [55]—a freeware version of the full MeqiSuite program—was used to generate a variety of graph-based indices with different degrees of sophistication. In this work, the unextended cyclic-system skeleton (*UnSkCycMqn*) index was used. This index provides a basic description of the underlying chemotype considering only the connectivity of the original structure. It does not distinguish atom and bond types and thus produces relatively broad categories, which encompass a large amount of chemical structures. Scheme 1 shows an example of three chemical structures, their corresponding *UnSkCycMqn* indices, and the reduced graphs used to calculate the indices.

Structures **2** and **3** from Scheme 1 obtain the same *UnSkCycMqn* because this particular index does not take atom and bond types into account. For a detailed description of all available graph-based indices inside the MeqiSuite the reader is referred to the extensive technical report [56] describing the MeqiSuite software.

Results and discussion

For all virtual screening scenarios studied in this work the following questions were considered of interest:



Scheme 1 Example of the MeqiSuite UnSkCycMqn index. The reduced graph used for the calculation is marked with thicker lines

- What is the optimal size of the training set?
- Which method to use?
- Which descriptor to use?
- Is there a difference when screening different chemical space?
- Is there an advantage of using a 3D descriptor?

Scenario 1: prioritizing compounds for a subsequent HTS

In addition to the above questions in this case the following additional questions are of interest:

- Is the virtual screening method able to prioritize compounds in a way, better than random ordering?
- At which size can a ranked list be truncated in a way which provides the best trade-off between number of false-positives and recovered actives? In other words what portion of the ranked list should be examined to guarantee that a given amount of known actives have been recovered?

What is the optimal size of the training set?

Figure 2 shows the obtained area under the ROC curve as a function of the size of the training set for all activity classes under investigation. The results with both methods when retrieving actives from the corresponding database with autocorrelation vectors as a descriptor are shown. Similar tendencies were observed for the other two investigated descriptor types.

As can be seen from Fig. 2, the expected tendency of obtaining better results with a higher amount of training data is confirmed. This tendency is much stronger when the actives from MDDR database are considered (remember that the training set is always selected from MDDR only, cf. "Materials and methods"). On the other hand, when actives from an external database (WOMBAT) are considered increasing the size of the training set beyond 100 active compounds brings only marginal improvement. Thus, increasing the size of the training set as much as six times (from 50 to 300 training structures) brings, on average, less than 6% improvement when the retrieval of MDDR actives is considered and around 3% when a retrieval of active compounds from an external database (WOMBAT) is considered. Therefore, the results obtained with a training set consisting of 100 known active compounds will be discussed in the rest of this section.

Is there a difference when screening different chemical spaces?

Three distinct cases with regards to the amount of active structures in MDDR and in WOMBAT were identified. In the case of 5-HT3 antagonists (Fig. 2a) both MDDR and WOMBAT contain a similar number of active compounds (cf. Table 1). WOMBAT, on the other hand, contains around two times more 5-HTA1 (Fig. 2b) agonists than MDDR. As can be seen from Fig. 2 the obtained results when retrieving structures from WOMBAT were significantly lower in these two cases. In the case of PKC inhibitors (Fig. 2h) MDDR contains almost four times more PKC inhibitors than WOMBAT. MDDR seems to cover the chemical space (as defined by the 44-dimensional autocorrelation vectors) of WOMBAT relatively well in this case and the bias towards retrieving compounds from the same database from which the training set was selected is less pronounced. This has resulted in obtaining a similar performance regardless of the origin of the known actives. Similar tendencies were observed with the other activity classes (Fig. 2c–g).

The ROC areas (average over 10 runs) and the corresponding standard deviations obtained when retrieving actives from MDDR and from WOMBAT, respectively, are shown in Fig. 3. There are two cases which deserve a note. In the first case, illustrated by 5-HT1A, D2, Thrombin, and HIV-1 P activity classes, more active structures are available in WOMBAT than in MDDR (cf. Table 1). A lower performance was obtained when retrieving WOMBAT actives in most of these cases (except HIV-1 P inhibitors, cf. Fig. 3). These results hint that the few actives available in MDDR are unable to cover the full activity space in WOMBAT. In the second case, illustrated by 5-HT3, AT1, Renin, and PKC activity classes, the number of known active structures is higher in MDDR (cf. Table 1). In most of these cases (AT1, Renin, and PKC), similar performances were obtained regardless of the source of the active structures (cf. Fig. 3).

However, lower performances were obtained when retrieving 5-HT3 antagonists originating from WOMBAT with all descriptors (cf. Fig. 3) and when retrieving PKC inhibitors originating from WOMBAT with Daylight fingerprints (cf. Fig. 3a). Thus, a certain amount of bias towards recovering compounds from MDDR is introduced.

The source of this bias is likely due to the fact that the two databases used in this study cover different aspects of the corresponding activity spaces. To investigate this assumption we have mapped the activity spaces covered by the actives in MDDR and WOMBAT on a plane using the Sammon projection algorithm [57]. The distance matrix obtained with the corresponding descriptor using one

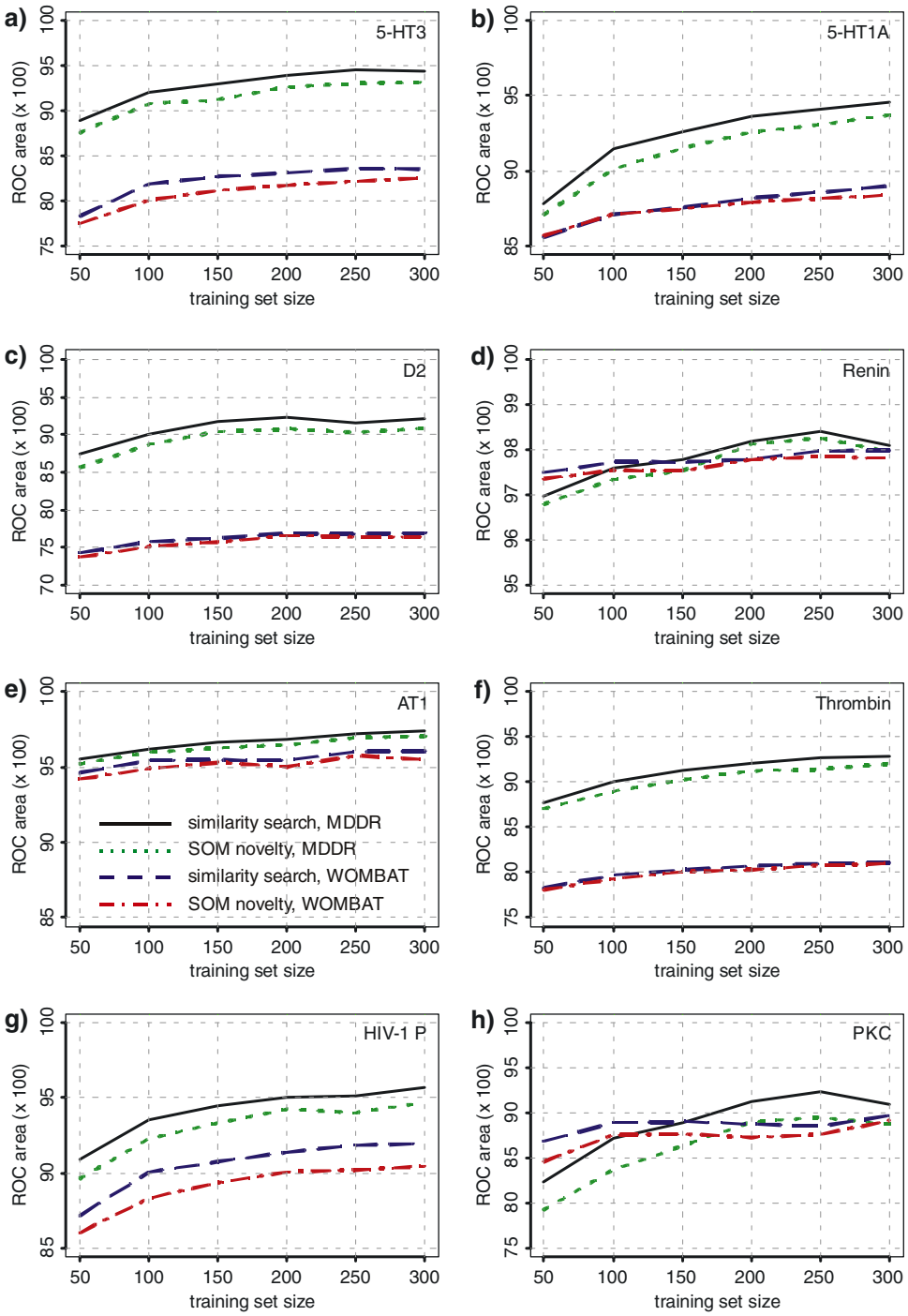


Fig. 2 Area below the ROC curve (×100) as a function of the training set size. The structures were described with 44-dimensional autocorrelation vectors

minus the Tanimoto coefficient was used as an input to the R [58] implementation [59] of the Sammon mapping algorithm. Figure 4 shows the Sammon projections for the 5-HT3 antagonists (Fig. 4a–c) and for the HIV-1 P inhibitors (Fig. 4d–f) with the three structure representations under investigation.

Starting with Fig. 4a more than 35% of the 5-HT3 antagonists in WOMBAT are mapped with *X* axis values higher than 0.25 while less than 10% of the 5-HT3 antagonists in MDDR are found in this region of the projection. This fact confirms the suspected difference between the chemical spaces covered by the 5-HT3 antagonists in MDDR and WOMBAT. The difference exists when using topological autocorrelation and RDF codes as well, as can be seen from Fig. 4b, c. In the case of topological autocorrelation (Fig. 4b) a region below −0.2 on the *Y* axis exists where 15% of the WOMBAT actives are found while only nine MDDR structures have been mapped there. The RDF codes (Fig. 4c) have mapped the WOMBAT

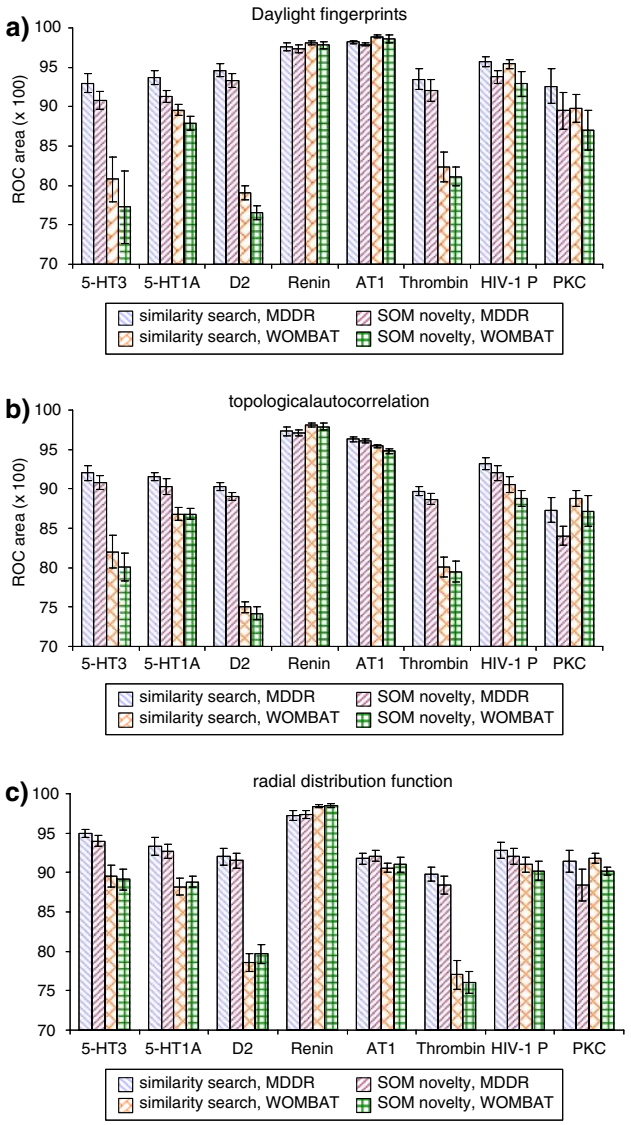


Fig. 3 Area under the ROC curve (×100) obtained when retrieving actives from MDDR or from WOMBAT. A training set of one hundred active compounds selected from MDDR was used. The structures were described with (**a**) 2,048-dimensional Daylight fingerprints; (**b**) 44-dimensional topological autocorrelation vectors; and (**c**) 256-dimensional RDF codes

activity space better than the 2D descriptors when considering 5-HT3 antagonists. There is a region above −0.05 on the *X* and above 0.1 on the *Y* axis, which is occupied mainly by WOMBAT actives. However, still more than sixty MDDR actives are mapped to the same region. This is reflected in the lower difference in the performance when retrieving actives from WOMBAT with RDF codes (cf. Fig. 3). Thus, the use of RDF codes resulted in around 5% lower ROC area when retrieving WOMBAT 5-HT3 antagonists compared to around 13 and 11% when using binary fingerprints and topological autocorrelation, respectively.

In the case of HIV-1 P inhibitors two the actives from WOMBAT form two distinct clusters when using 2-dimensional descriptors (Fig. 4d, e) while no such clustering is observed with RDF codes (Fig. 4f). However, in all cases the actives from MDDR are spread through the WOMBAT space somehow even, therefore the bias towards retrieving MDDR actives is lower (cf. Fig. 3).

Based on the mappings shown in Fig. 4 it is clear that for some activity classes there is difference between the chemical space covered by MDDR and WOMBAT. The difference in the performances between retrieving active structures from the same database (MDDR) from which the training set has been selected and from an external database (WOMBAT) can be as high as 14% (5-HT3 antagonists, fingerprints). On average, 6% lower ROC areas were obtained when retrieving actives from an external database with Daylight fingerprints, 5.5% topological autocorrelation, and 5% with RDF codes. Thus, optimistic performance estimates are obtained when active structures from the same database from which the training set has been selected are retrieved.

Which method to use?

To work in a "bias-free" environment and to keep the text concise the discussion from here on will be based on the results obtained when retrieving actives from WOMBAT.

First we look at the performance of the two methods—similarity search with data fusion and SOM novelty detection, illustrated in Fig. 5. Both methods were able to rank the WOMBAT actives in a way superior to a random selection as indicated by the obtained high values of the area under the ROC curve. The ndSOM performed a bit worse when a binary structure representation was used (Fig. 5a). On the other hand, in the case of real-valued descriptors (Fig. 5b, c) there is virtually no difference between the results obtained with SSDF and ndSOM.

Considering the fact that the highest average difference in the ROC areas between the two methods was around 5% (when retrieving 5-HT3 antagonists from WOMBAT with binary fingerprints) we conclude that both methods perform similarly for any practical purposes. Based on this conclusion we favor the SOM novelty detection technique due to the fact that it is twice as fast [18] compared to the similarity search with data fusion.

Which descriptors to use?

Following the conclusions from the previous section, the discussion here is limited to SOM novelty detection for retrieving WOMBAT actives. Figure 6 shows the actual ROC curves and the corresponding areas under them obtained with the three descriptors—Daylight fingerprints, topological autocorrelation vectors, and RDF codes.

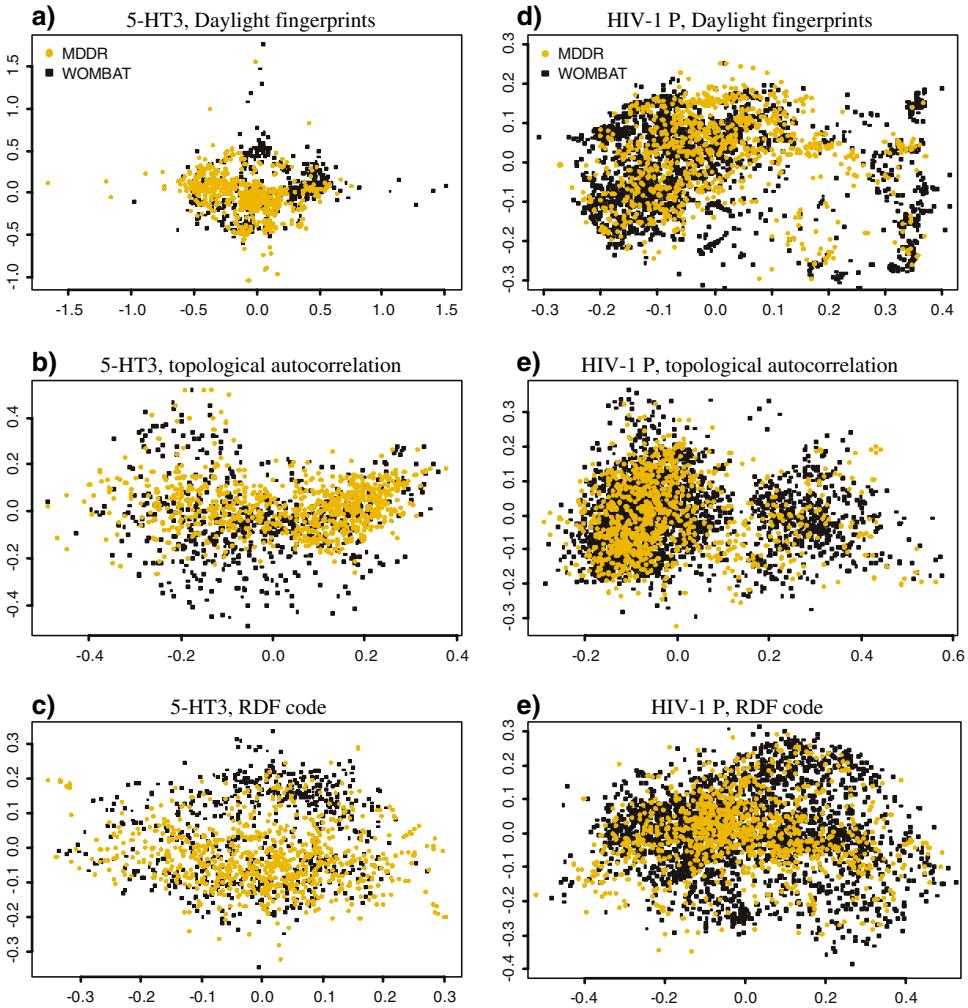


Fig. 4 Sammon projection of the active compounds in MDDR and WOMBAT. (**a**) 5-HT3, Daylight fingerprints. (**b**) 5-HT3, topological autocorrelation. (**c**) 5-HT3, RDF code. (**d**) HIV-1 P, Daylight fingerprints. (**e**) HIV-1 P, topological autocorrelation. (**f**) HIV-1 P, RDF code

According to our expectations, the obtained performance varies with the activity class under investigation. Thus, while the RDF codes gave better performance in the case of 5-HT1, D2, and PKC classes, the Daylight fingerprints performed better in the case of AT1 and Thrombin inhibitors. The topological autocorrelation vectors performed very similar or better than some of the other two descriptors in all cases except D2 antagonists where the autocorrelation vectors have achieved the worst performance.

From the results depicted in Fig. 6 no ultimate "good-for-all" descriptor can be identified. However, the performance obtained with topological autocorrelation vectors—although never the best one—shows that good results can be obtained even when a low-dimensional 2D descriptor is used.

Is there an advantage of using a 3D descriptor?

Looking at Fig. 6, in six out of eight cases no particular advantage of utilizing 3D descriptor can be seen. Considering the fact that the RDF codes used throughout this study were of approximately six times higher dimensionality than the topological autocorrelation functions the results seem to indicate that little, if any, value was added through moving to a 3D-based representation of the chemical structure. This is in agreement with previous studies on the subject [14, 60]. This observation suggests that molecular features, which play important role for the biological activity, can be deduced in most of the cases from the 2D representation of the chemical structure and do not require explicit accounting for the conformational parameters.

The behavior observed in here may be due to the fact that a single low-energy conformation was used. It should be stressed that the aim of this study was not the development of precise quantitative models. Using a single conformation in the manner described in this study is a gross simplification. Even if a conformation minimized by a quantum mechanical method or force -field is used it may be far away from the biologically active conformation. However, the goal here was to study if the inclusion

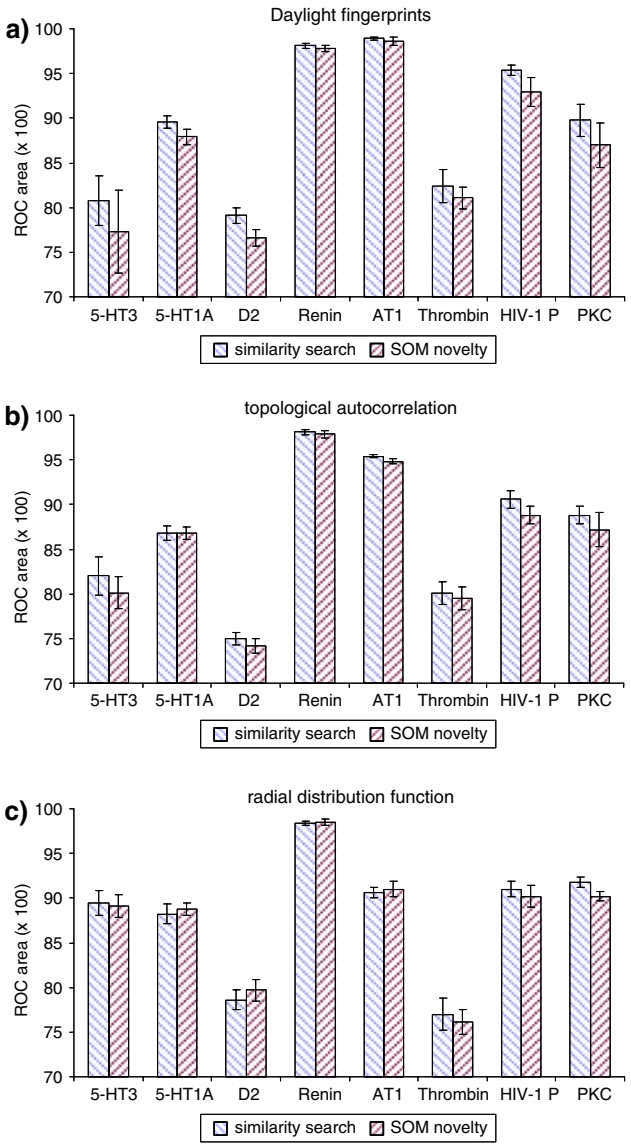


Fig. 5 Area under the ROC curve (×100) obtained when retrieving WOMBAT active structures. The structures were described with (**a**) 2,048-dimensional Daylight fingerprints; (**b**) 44-dimensional topological autocorrelation vectors; and (**c**) 256-dimensional RDF codes

of 3D information, even if oversimplified, can improve the virtual screening results. The 3D structure generator CORINA has been shown [61] to reproduce the PDB conformations of co-crystallized enzyme complexes reasonably well. In addition CORINA is very fast, thus the conversion from topological constitution to a single conformation does not hamper the speed of the virtual screening procedure significantly. On the other hand, with the aim of screening millions of compounds the trade-off between the execution speed and the obtained improvement due to the use of 3D descriptor is not attractive in this scenario.

Scenario 1: specific questions

Based on the already shown results the answer to the first question—is the virtual screening method able to prioritize compounds in a way, better than random ordering—is clearly yes. It was shown—by means of the obtained large areas under the ROC curve—that all method/descriptor combinations are able to prioritize compounds in a way, better than random ordering.

To answer the second question—at which size a ranked list can be truncated in a way which provides the best trade-off between number of recovered false and true positives—one simply needs to define the desired number of actives or the allowed number of structures with unknown activity and to read the corresponding numbers from the ROC curve, as illustrated in Fig. 6 for a list that contains 5% false positives.

Scenario 2: selecting compounds for a subsequent lead-optimization

In addition to the questions posted in the beginning of the discussion in this case the following additional questions were considered of interest:

- How well is a virtual screening method able to retrieve actives in the very beginning of the ranked list ("early recognition")?
- How different are the compounds retrieved by each method at the pre-specified size of the ranked list?
- How different are the compounds retrieved by each descriptor?
- Is the retrieval of new known active chemotypes (which are not present in the training set) possible?
- How diverse are the "false-positives" in terms of chemotypes? Do these "false-positives" look promising as a new leads for the given activity?

In this scenario we assume that a relatively small number of the original compounds is considered—usually the 1–10% top-ranked compounds. However, we want to emphasize that even 1% of structures from a database which contains 1 million compounds still results in 10,000 compounds—a size, which might be acceptable for a pharmaceutical company but is not likely to be of use to a university research laboratory. We concentrate mainly on the "early recognition" problem as perceived in the pharmaceutical industry. The performance of the methods in retrieving known active structures amongst the top-ranked 5% of the original database—corresponding to approximately 13,500 structures—is discussed. However, a short discussion of the one hundred false-positives is presented at

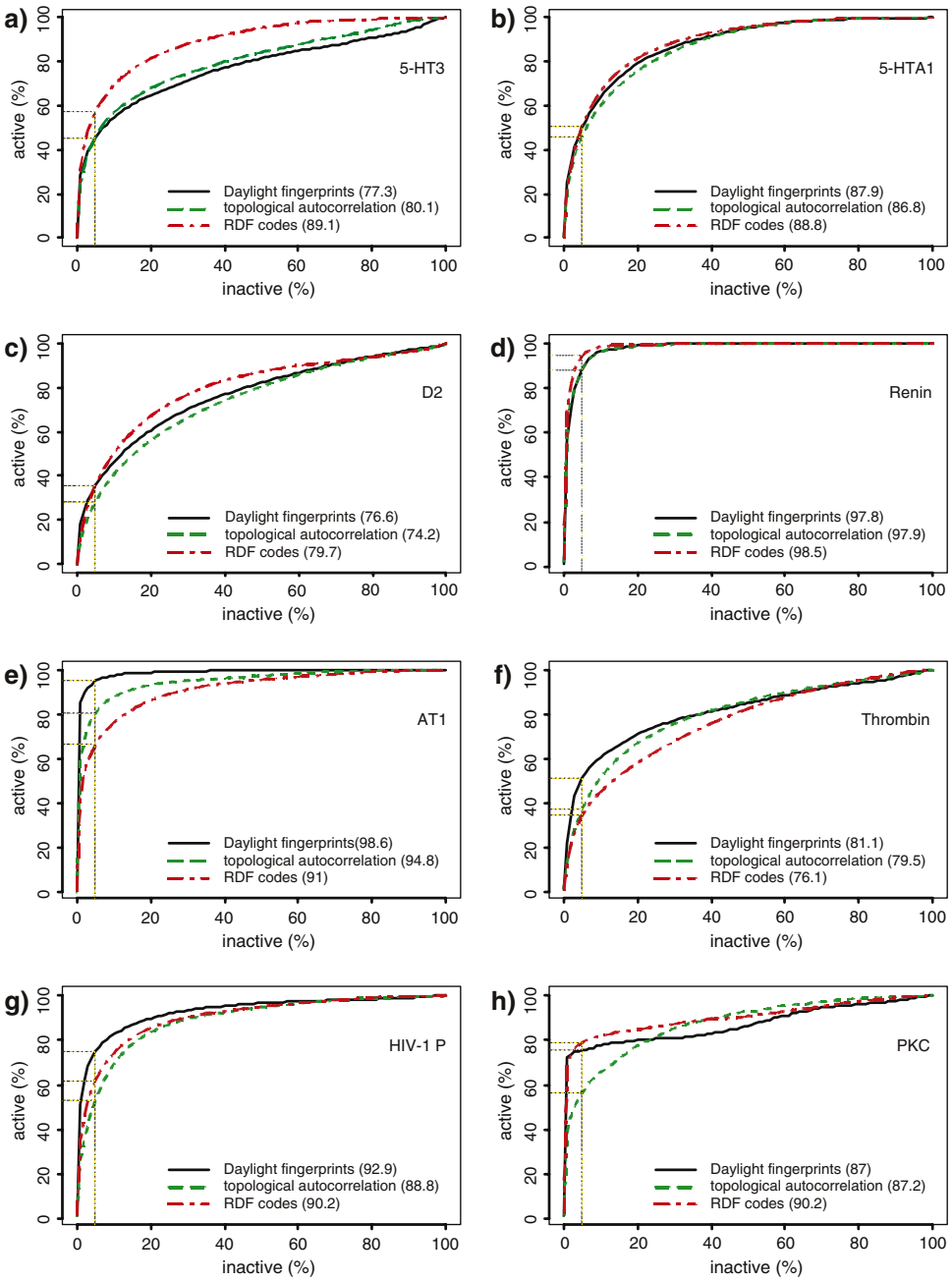


Fig. 6 ROC curves obtained with SOM novelty detection when retrieving WOMBAT active compounds. The area below the corresponding curve is shown in parentheses. A hypothetical cut-off of the ranked list compromising 5% false negatives and the expected per cent of true positives is shown by dotted lines

the end of the section in an attempt to show the applicability of the proposed methods when limited resources are available.

Similarly to the area under the ROC curve, the BEDROC score can be interpreted as the probability that a given active compound will be ranked higher than a random compound, drawn by an exponential distribution with given α parameter. In the present study $\alpha = 32.2$ was used. This value ensures that 80% of the BEDROC score will be based on the 5% top-ranked compounds. The theoretical expectation of the BEDROC score is 0.5. We want to stress that the BEDROC evaluation as, in fact, any evaluation of early recognition problems, is ultimately dependent on the selected parameters. That is, a given score has always to be interpreted taking into account the definition of "usefulness". A method, which retrieves, say, 70% of the actives amongst the top-ranked five per cents may result in a BEDROC score lower than 0.5, although in most of the cases such a performance will be considered rather good. We have provided the full ranking data used in this study as a part of the supporting information, which accompanies this article. These data allow the interested reader to evaluate the performance of each method/descriptor combination under her/his definition of "usefulness".

What is the optimal size of the training set?

The same tendency as in Scenario 1 was observed. Similarly to the area under the ROC curve the obtained BEDROC scores ($\alpha = 32.2$) did not increase significantly after one hundred active structures were used as training set. Therefore, the following discussion is still based on the results obtained with a training set consisting of one hundred known active compounds.

Is there a difference when screening different chemical spaces?

The bias towards retrieving actives from MDDR as already discussed in Scenario 1 was more pronounced when the early retrieval is considered. Thus, in what follows only the retrieval of actives from an external database—WOMBAT—will be discussed. The obtained BEDROC scores when retrieving WOMBAT actives with a training set consisting of 100 active compounds are summarized in Table 3. Mean value and standard deviation over 100 bootstrapped repetitions (cf. "Materials and Methods") are reported.

Which method to use?

As can be seen from Table 3 the similarity search with subsequent data fusion produced better BEDROC ($\alpha = 32.2$) scores than the SOM novelty detection method in almost all cases regardless of the descriptor. In three of the eight activity classes—D2, Renin, and HIV-1 P—the SOM novelty detection was slightly better when RDF code was used to describe the structures.

In contrast to Scenario 1 in which the difference in the ROC areas was deemed irrelevant from a practical point of view, in the early recognition scenario the similarity search method produced up to 16% better BEDROC scores (5-HT3 activity class, for example, cf. Table 3). A paired *t*-test was performed to compare the obtained BEDROC scores between the two methods by statistical means. A statistically significant difference was found in almost all cases except with the 5-HT1A and D2 activity classes based on RDF codes. Based on the results of the *t*-test we conclude that the similarity search method performs better in this scenario. However, we want to point out that a statistically significant difference does not necessarily equal to a practically meaningful difference. Thus, for example, considering the Renin class with binary fingerprints (cf. Table 3) the mean of the difference between the BEDROC scores ($\times 100$) obtained by the two methods is 0.4—a figure which hardly bears any practical consequences but nevertheless the two means differ from a statistical point of view (paired *t*-test).

Which descriptors to use?

A comparison between the BEDROC scores obtained with similarity search is presented in Fig. 7. Similar observations as for Scenario 1 can be made. The performance of the descriptors is once again activity dependent. The Daylight fingerprints performed best for the AT1, Thrombin, and HIV-1 P activity classes while for the rest of the activity classes, except PKC inhibitors, all three descriptors show similar performance. In the case of the PKC inhibitors the topological autocorrelation gave inferior performance compared to the other two structure representations. Thus, based on the obtained BEDROC scores no "best" descriptor can be identified as a general rule.

Table 3 BEDROC scores ($\times 100$, $\alpha = 32.2$) obtained when retrieving WOMBAT active structures—mean and standard deviation over one hundred bootstrapped runs with 10 different training sets of 100 active compounds

Activity	DFP				AC2D				RDF			
	SSDF		ndSOM		SSDF		ndSOM		SSDF		ndSOM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
5-HT3	42.9	8.2	36.0	8.8	40.0	5.4	34.9	4.8	48.0	6.9	43.3	6.5
5-HT1A	41.5	3.9	35.9	4.4	34.5	3.2	32.5	3.3	35.3	3.9	34.9	3.5
D2	27.6	3.1	25.1	2.9	20.9	2.5	19.0	2.7	23.6	2.7	23.8	2.9
Renin	73.4	4.4	69.6	6.4	74.9	4.7	74.2	4.3	73.5	4.1	77.7	4.4
AT1	86.1	3.0	85.7	3.2	68.7	3.4	66.2	3.4	53.3	4.5	52.6	4.0
Thrombin	36.5	4.4	34.6	4.4	24.8	2.9	24.2	3.1	24.8	3.4	23.0	3.4
HIV-1 P	62.5	4.6	59.7	5.7	41.8	4.5	38.1	4.2	45.3	4.7	46.0	5.3
PKC	73.8	2.5	72.1	2.1	51.8	5.7	46.0	7.4	73.2	2.7	71.8	3.0

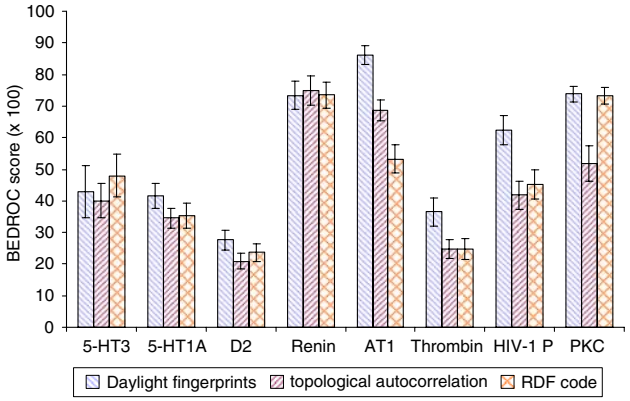


Fig. 7 BEDROC scores (×100) obtained when retrieving WOMBAT actives with similarity search using different descriptors

Is there an advantage of using a 3D descriptor?

As can be seen from Fig. 7 and Table 3 the observations from Scenario 1 with regards to this question are confirmed—no gain in the early retrieval ability of each of the studied virtual screening methods was achieved by utilizing a 3D descriptor. For most of the investigated activity classes—5-HT3, 5-HT1A, D2, Renin, AT1, Thrombin, and HIV-1 P the use of RDF codes gave similar performance to this achieved with topological autocorrelation vectors. Thus, based on the obtained BEDROC scores no convincing reason to use RDF codes based on a single low-energy conformation was found.

How well performed the studied virtual screening methods in the "early recognition" scenario?

From the obtained BEDROC scores—Table 3—both similarity search and SOM novelty detection were able to group significant amounts of the known WOMBAT actives amongst the top 5% of the ranked list. However, the actual results were activity dependent. In the case of Renin, AT1, HIV-1 P, and PKC a good performance—BEDROC scores higher than 0.5—was obtained, while for the rest of the activity classes the grouping of actives amongst the top ranked 5% was not that successful—BEDROC scores lower than 0.5. This is in clear contrast to the area under the ROC curve. As was shown in the discussion of Scenario 1 all methods and descriptors produced ROC area values much superior to the randomly expected area of 0.5. However, when the early recognition is important (which is pretty much the case in any virtual screening application) there are cases where the performance is not optimal. Thus, while an analysis based on the area under the ROC curve is useful for providing an overview of the virtual screening method ability to perform better than random (or, precisely speaking, uniform) picking it is not directly indicative of the early recognition capabilities of a given virtual screening method. Of course, the actual ROC curve gives a rather accurate picture—the highest BEDROC scores obtained (Renin, AT1, HIV-1 P, PKC) correspond to the steepest ROC curves, cf. Fig. 6.

Chemotype analysis

In this section, a comparison between the different virtual screening methods and descriptors with regards to the recovered active structures in terms of chemotypes—as perceived by MeqiLite *UnSkCycMqn* index (cf. "Materials and methods", Section "Chemotype analysis")—is presented.

How different are the active compounds retrieved by each descriptor? The ability of each descriptor to recover different structures and chemotypes is summarized in Table 4. A few observations from Table 4 deserve attention.

Table 4 WOMBAT active structures and chemotypes recovered by different structure representations

Activity	DFP				AC2D				RDF			
	Active structures		Chemotypes		Active structures		Chemotypes		Active structures		Chemotypes	
	Total	Unique	Total	Unique	Total	Unique	Total	Unique	Total	Unique	Total	Unique
5-HT3	332	34	107	7	320	28	115	10	380	50	114	9
5-HT1A	1,393	320	316	58	1,178	252	313	57	1,216	179	287	30
D2	1,087	291	298	54	856	202	266	47	962	235	285	52
Renin	528	6	131	1	518	7	136	3	551	9	139	1
AT1	1,307	134	233	26	1,110	11	206	3	891	2	156	1
Thrombin	962	209	230	41	689	80	204	29	662	61	154	13
HIV-1 P	1,874	330	493	72	1,341	72	396	16	1,431	110	400	30
PKC	120	7	22	2	94	3	20	2	124	9	24	1

Mean values over 10 similarity search runs with different training sets

First, the correspondence between the BEDROC scores and the total number of retrieved actives amongst the top-ranked 5% can be seen. The higher the BEDROC score the higher is the number of retrieved actives. However, the relationship is not strictly linear since, as already discussed, the BEDROC score gives higher weight to compounds, which are found early in the ranked list. Thus, considering the Thrombin inhibitors as an example, the retrieval with autocorrelation vectors and RDF codes produced equal BEDROC scores (cf. Table 3) while the average number of recovered compounds differs slightly (cf. Table 4).

Another observation from Table 4 is that, as expected, different descriptors cover different aspects of the activity space and consequently retrieve different compounds. This is illustrated in the "unique compounds" columns. These numbers were produced by taking the difference between the sets of active compounds recovered by each descriptor and averaging the size of the obtained sets over the 10 repetitions with different training sets.

Thus, considering 5-HT3 as an example, one can see from Table 4 that on average 34 compounds were recovered exclusively by Daylight fingerprints, 28 exclusively by autocorrelation vectors, and 50 exclusively by RDF codes. Depending on the activity class, each descriptor recovered between 1 and 25% active structures, which have been missed by the other two. In addition, as can be seen from the "unique chemotypes" columns in Table 4, the active structures recovered uniquely by each descriptor more often than not contain additional chemotypes. Figure 8 shows the percentage of different chemotypes amongst the recovered active structures for each descriptor.

As can be seen from Fig. 8, the autocorrelation vectors recovered the most diverse—in terms of *UnSkCycMqn* Meqi index—active structures regardless of the activity. Although the Daylight fingerprints recovered more actives, these structures were usually less diverse. This is not surprising having in mind that the Daylight fingerprints encode exclusively structural information while the autocorrelation and RDF codes make an attempt to account for various physico-chemical properties as well. The RDF codes usually recovered less diverse sets of actives compared to the topological autocorrelation. This may be attributed to the use of only a single low-energy conformation.

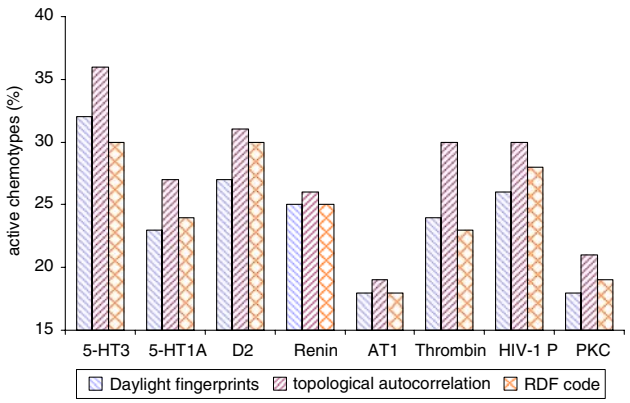


Fig. 8 Number of chemotypes expressed as percentage from the total number of recovered WOMBAT actives. Virtual screening method: similarity search with data fusion

The results shown in Table 4 raise the question is it possible to further improve the retrieval of active structures by combining the ranked lists obtained with different structure representations. In an attempt to answer this question we have combined the ranked lists obtained with Daylight fingerprints, topological autocorrelation, and RDF codes using data fusion. A fusion algorithm, based on the sum of the ranks from the individual ranked lists was used, as described by Ginn [62]. Thus, each structure from the screened database obtained a new rank equal to the sum of its ranks in the ranked lists obtained with the particular structure representation. The results of these experiments are summarized in Table 5.

The results in Table 5 and the following discussion are based on the results obtained with the similarity search virtual screening method. This allows the estimation of the improvement compared to the common use of this method in which only binary fingerprints are used. Similar tendencies were observed for the SOM novelty detection—the other virtual screening method under investigation.

As a first attempt, the lists obtained with similarity search using Daylight fingerprints and topological autocorrelation were fused. As can be seen from Table 5, 1–16% more actives were recovered for the 5-HT3, 5-HT1A, D2, Renin, and PKC activity classes. A slight decrease in the number of recovered actives compared to Daylight fingerprints was observed for the AT1, Thrombin, and HIV-1 P activity classes. However, the decrease, when it occurs, was never more than 3% (Thrombin inhibitors) while an improvement as high as 21% (5-HT3 antagonists) was observed. Therefore, the use of autocorrelation vectors based on atomic physico-chemical properties adds value to the commonly used similarity search with binary fingerprints.

As a second attempt, the lists obtained with topological autocorrelation and RDF codes were fused. In three cases (AT1, Thrombin, and HIV-1 P) 5–8% less actives were recovered compared to Daylight fingerprints (cf. Table 5). However, for the rest of the activity classes the fusion of the lists obtained with topological autocorrelation and RDF codes resulted in the recovery of 6–30% more active structures compared to Daylight fingerprints. This result shows that the use of RDF codes, even though based on single low-energy conformation, leads to better coverage of the activity space.

Table 5 WOMBAT active structures recovered by fusion of the ranked list obtained with different structure representations and improvement over similarity search using only Daylight fingerprints

Activity	AC2D + DFP		AC2D + RDF		AC2D + DFP + RDF	
	# Actives	Improvement over DFP (%)	# Actives	Improvement over DFP (%)	# Actives	Improvement over DFP (%)
5-HT3	403	21.4	399	20.2	431	29.8
5-HT1A	1,612	15.7	1,496	7.4	1,715	23.1
D2	1,243	14.4	1,151	5.9	1,348	24.0
Renin	551	4.4	566	7.2	568	7.6
AT1	1,300	−0.5	1,244	−4.8	1,362	4.2
Thrombin	930	−3.3	883	−8.2	988	2.7
HIV-1 P	1,830	−2.3	1,783	−4.9	1,954	4.3
PKC	121	0.8	128	6.7	130	8.3

Mean values over 10 similarity search runs with different training sets

Finally, the results obtained with all three structure representations under investigation were fused. As can be seen from Table 5, this resulted in the recovery of 3–30% more active structures compared to the case in which only Daylight fingerprints were used.

To summarize, the fusion of the results obtained with different descriptors is most effective when the individual structure representations show similar performance, as demonstrated by the 5-HT3 and 5-HT1A activity classes (cf. Tables 4, 5). The fusion of the lists obtained with two real-valued descriptors (topological autocorrelation and RDF codes weighted by atomic physico-chemical properties) recovered similar or higher number of actives compared to the use of Daylight fingerprints alone. This shows the utility of the real-valued descriptors and the corresponding physico-chemical properties in a ligand-based virtual screening experiment, especially when the much lower dimensionality of the real-valued descriptors is considered. In addition, the application of RDF codes even when based on a single low-energy conformation allows better coverage of the corresponding activity space.

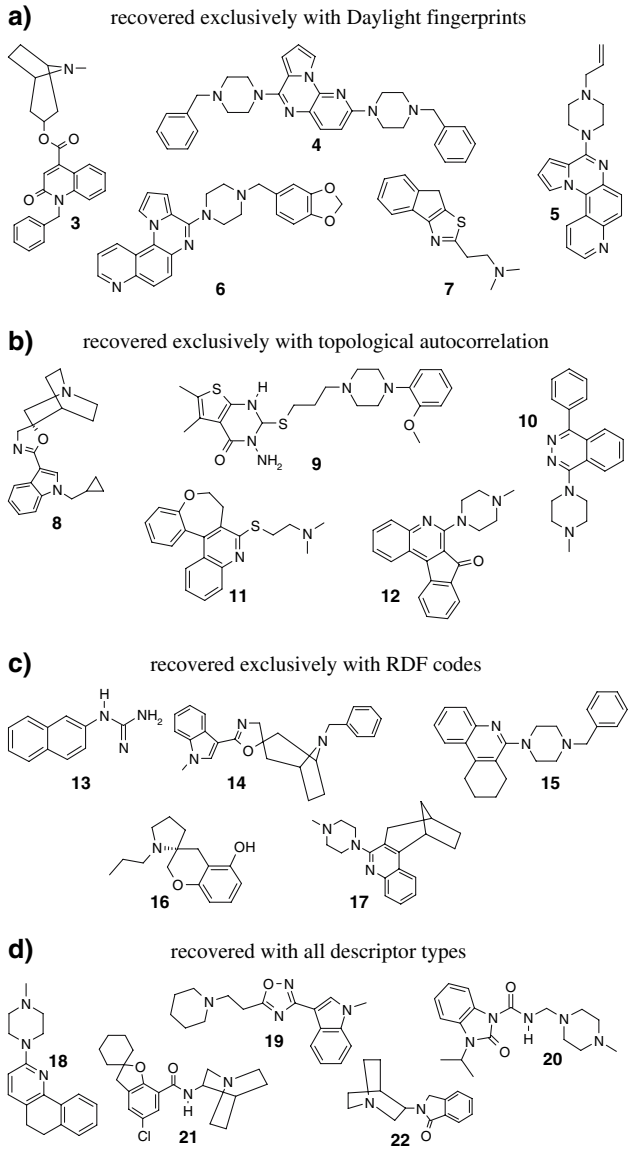
The discussion so far has been based on the average number of active structures and chemotypes recovered from the 10 runs with different training sets. In an attempt to translate some of the numbers discussed above in a chemical language we have randomly selected a single run out of the 10 repetitions for each activity class. This allows us to analyze the actual chemical structures. The results for two of the eight activity classes—5-HT3 and D2 antagonists—will be exemplified. The virtual screening for these activity classes resulted in average (BEDROC scores between 40 and 48, cf. Table 3) and low (BEDROC scores between 20 and 28, cf. Table 3) early recognition performance, respectively.

The structures shown in Scheme 2 illustrate the ability of each individual descriptor to recover structures from different parts of the activity space. Amongst the structures recovered exclusively by Daylight fingerprints a common (piperazinopyridopyrroloquinoxaline) ring system can be seen in structures **5** and **6** and a similar one (piperazinopyridopyrrolopyrazine) in structure **4**. For the compounds discovered by some of the vectorial descriptors no such common structural building block can be distinguished. Thus, the compounds obtained by real-valued vectorial descriptor are somehow more diverse, as already discussed.

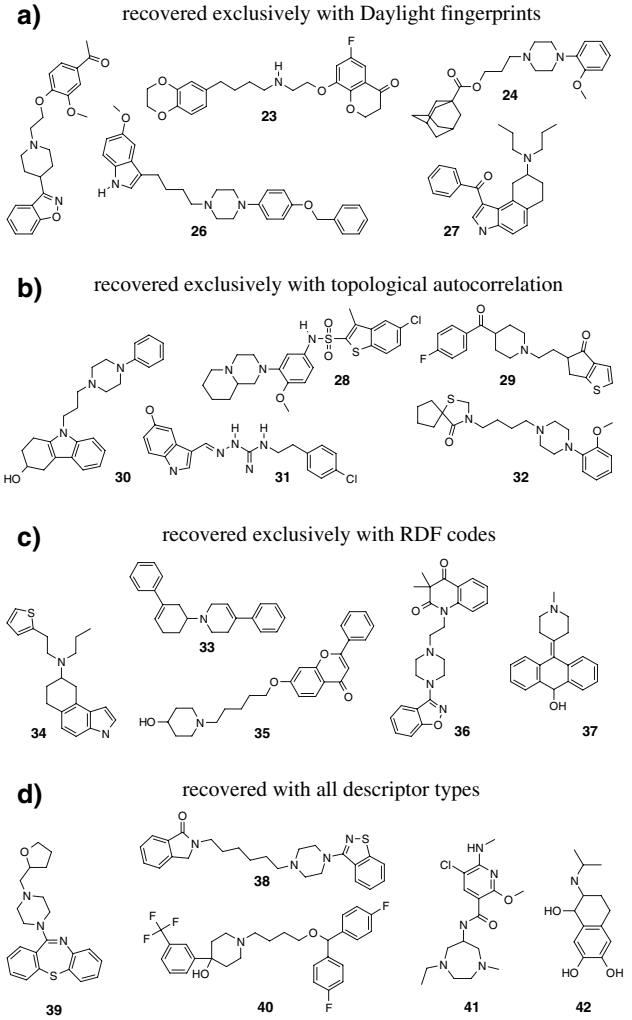
A certain preference for recovering relatively small ligands is observed with RDF vectors—structures **13** and **16**. Considering that structure **13** is not very flexible it is surprising that it was overlooked by the topological autocorrelation. This artifact can be attributed to the lower dimensionality of the autocorrelation vectors which may lead to a slight preference for larger structures.

This preference results from the summation in Eq. 10 which inevitably gives more weight to structures with more atoms. While the RDF is prone to the same problem its higher dimensionality can alleviate the size effect to some extent. The conformation dependence of the RDF code, on the other hand, makes the result dependent on the particular 3D conformation—a fact which significantly complicates the interpretation of the results. However, the use of a single low-energy CORINA conformation leads to reasonable results as can be seen from the structures recovered by all descriptors.

Scheme 3 shows some of the D2 antagonists recovered only by one of the used structure representations. In contrast to Scheme 2, no common fragment can be discovered amongst the structures recovered by Daylight fingerprints. The chemotypes discovered exclusively by some of the used descriptors are quite diverse in all cases. This is not surprising having in mind the relatively low early recognition performance in this case, BEDROC scores between 0.2 and 0.28, cf. Table 3. Thus, it clearly demonstrates that each descriptor covers different parts of the activity space.



Scheme 2 5-HT3 antagonists from WOMBAT. Each structure represents a different chemotype recovered only by the corresponding descriptor. Five structures representing some of the chemotypes recovered by all descriptor types are shown as well. (**a**) Recovered exclusively with Daylight fingerprints. (**b**) Recovered exclusively with topological autocorrelation. (**c**) Recovered exclusively with RDF codes. (**d**) Recovered with all descriptor types



Scheme 3 D2 antagonists from WOMBAT. Each structure represents a different chemotype recovered only by the corresponding descriptor. Five structures representing some of the chemotypes recovered by all descriptor types are shown as well. (**a**) Recovered exclusively with Daylight fingerprints. (**b**) Recovered exclusively with topological autocorrelation. (**c**) Recovered exclusively with RDF codes. (**d**) Recovered with all descriptor types

How different are the compounds retrieved by each method? In this section we compare both virtual screening methods in terms of recovered chemotypes. Since similar tendencies were observed with all structure representations, the following discussion concentrates on the results obtained with autocorrelation vectors, which recovered the most diverse set of actives (see the previous section). The results are summarized in Table 6.

A small advantage of the similarity search is observed by the slightly higher absolute number of recovered actives ("actives" column in Table 6). However, in terms of recovered chemotypes both methods show a similar performance ("chemotypes" column in Table 6). Considering the unique structures and chemotypes recovered by each method it is clear that, even though the same descriptor and training set was used, both methods have recovered different active structures. This observation comes as no surprise considering previous work in this field [63]. Thus, whenever it is possible, the use of more than one virtual screening method is preferable. Combining the lists of compounds returned by each separate method usually leads to an improvement in the obtained results, as has been shown in a number of studies [44, 63], including our previous work on SOM novelty detection [18].

Table 6 WOMBAT active structures and chemotypes recovered by different virtual screening methods

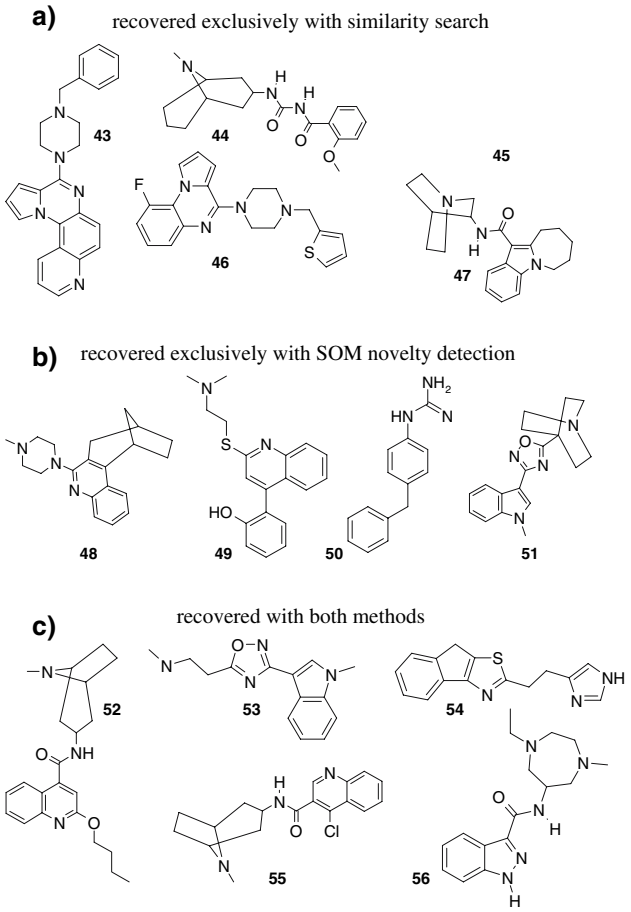
Activity	Similarity search with data fusion				SOM novelty detection			
	Active structures		Chemotypes		Active structures		Chemotypes	
	Total	Unique	Total	Unique	Total	Unique	Total	Unique
5-HT3	320	64	115	12	288	31	110	7
5-HT1A	1,178	222	313	43	1,128	172	308	38
D2	856	201	266	43	795	140	259	36
Renin	518	16	136	3	512	10	134	2
AT1	1,110	69	206	8	1,081	39	203	4
Thrombin	689	119	204	26	655	85	202	24
HIV-1 P	1,341	230	396	50	1,228	118	377	32
PKC	94	15	20	1	85	6	20	2

Mean values over 10 similarity search runs with different training sets and topological autocorrelation vectors as descriptor

When only a single method has to be used, the numbers shown in Table 6 do not show preference for any of the two methods investigated in this study. However, the similarity search recovers slightly larger number of actives, while the SOM novelty detection is twice as fast. Thus, we suggest that SOM novelty detection is used when the execution speed is of concern. Since approximately the same number of different chemotypes is recovered by both methods (even if the particular chemotypes are not 100% the same) the retrieved compounds will still cover significant parts of the activity landscape.

Using the same activity class—5-HT3 antagonists—as in the previous section the ability of both methods—similarity search with data fusion and SOM novelty detection—to discover different chemotypes is illustrated. Scheme 4 shows five of the seven 5-HT3 active structures recovered exclusively by similarity search and the four active structures unique to SOM novelty detection.

Is the retrieval of new active chemotypes possible? The clustering procedure, used for selecting the training set out of the MDDR actives (cf. "Materials and methods", Screening protocol) ensured a diverse training set. Around 70 different chemotypes were found amongst the one hundred MDDR active compounds used as a training set. However, the number of recovered WOMBAT chemotypes exceeded the number of the chemotypes in the training set by a factor as large as six in the case of HIV-1P inhibitors (396 different chemotypes were recovered with topological autocorrelation, cf. Table 4, HIV-1 P row). As demonstrated by structures **4**, **5**, and **6** in Scheme 2 some of the chemotypes, as identified by Meqi *UnSkCycMqn* index, still possess a high degree of structural similarity. However, this observation alone cannot account for the much higher number of chemotypes recovered from WOMBAT. Thus, the retrieval of different chemotypes from the one contained in the training set with the use of the virtual screening methods investigated in this study is possible.



Scheme 4 5-HT3 antagonists from WOMBAT. Each structure represents an active chemotype found only by the corresponding method. Five active structures recovered with both methods are shown as well. (**a**) Recovered exclusively with similarity search. (**b**) Recovered exclusively with SOM novelty detection. (**c**) Recovered with both methods

Do the "false positives" look promising as new leads? In retrospective experiments the term "false positive" is generally used for any structure which does not belong to

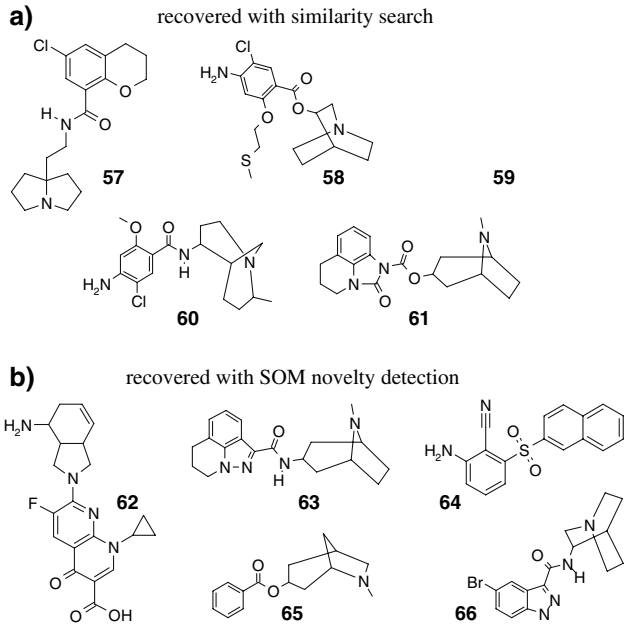
the set of known actives. However, very few, if any, of these structures have been tested against the screened enzyme or receptor. Therefore, these are structures with unknown activity rather than strictly false positives.

In this section, using again the results for 5-HT3 and D2 antagonists, we provide a short analysis of these "false positives". We focus on the one hundred top-ranked false positives in an attempt to show the applicability of the proposed methods in a situation where very limited resources for subsequent biological assays are available, i.e., in a university research laboratory. In this respect, a fact we want to stress is that the actual one hundred top-ranked structures were predominantly known actives—between 50 and 95%, depending on the activity class. However, to simulate a prospective screening situation we concentrate here on the top-ranked false positives. The same particular virtual screening runs as in the previous section were used and the results with topological autocorrelation vectors are discussed.

The one hundred top-ranked structures of unknown activity represented a large number of different chemotypes—64 for similarity search and 70 for SOM novelty detection when screening for 5-HT3 antagonists and 54 for similarity search and 69 for SOM novelty detection when screening for D2 antagonists. Scheme 5 shows the five active structures which represent the most frequently (at least three times) occurring chemotypes amongst the one hundred top-ranked false positives for the similarity search with data fusion—Scheme 5a—and for the SOM novelty detection—Scheme 5b when screening for 5-HT3 antagonists.

Some similarities between the structures on Scheme 5 and some of the known 5-HT3 antagonists depicted on Schemes 2, 4, like the presence of different aza-bicyclic systems can be seen. In addition, most of the false positives shown in Scheme 5 have been tested and found to be active against different receptors, which are closely related to 5-HT3. Compounds **57**, **60**, **61** and **63** are antagonists for different members of the serotonin family of receptors, while compound **65** has been found to act as a dopamine transporter (DAT) antagonist. Of course not all of the shown compounds have been tested against the members of the serotonin or related families of receptors. Compound **59**, for example, belongs to the activity class of bronchodilators in MDDR, while compound **64** is marked as a potent (pK_i of 7.5) HIV-1RT inhibitor in WOMBAT. However, more often than not the top-ranked false-positives are related to the activity under investigation. Thus, considering the top-ranked false-positives is likely to result in the discovery of previously unknown active structures.

Similar observations can be made when examining the one hundred top-ranked false-positives for the seven remaining activity classes. Taking the D2 antagonists



Scheme 5 Active structures which represent the most frequently (at least three times) occurring chemotypes amongst the one hundred top-ranked false-positives. (**a**) Recovered with similarity search. (**b**) Recovered with SOM novelty detection

activity class, most of the false-positive compounds belonged to the antipsychotic MDDR activity class or are marked as antagonists to different alpha adrenergic receptors. The dopamine antagonists, on one hand, are members of the broader antipsychotic MDDR activity class. On other hand, a large percentage of the D2 antagonists in WOMBAT acts as antagonists against different alpha adrenergic receptors as well. Therefore, the top-ranked false-positive compounds recovered by the virtual screening methods investigated in this study are likely to share the target activity.

Scenario 3: is a given compound active?

The problem investigated in this scenario is very similar to a classification task in which only active structures are used as training set. It is well known [35] that for most machine-learning techniques usually the larger the training set is the better are the results obtained. In addition, the bias towards performing better on the same database used to select the training set exists naturally. Therefore, we will not discuss the questions about the optimal size of the training set and about the difference in predicting compounds from different chemical spaces in details, but rather the results with training sets, consisting of one hundred active compounds when classifying actives from an external database (WOMBAT) will be discussed. In addition, for reasons which will become clear in the next section, the result

obtained with SOM novelty detection will be the main focus on our discussion.

Which method to use?

The task of turning a similarity search into a classifier is immediately confronted with the question of determining a threshold value for the used similarity measure. Screened structures which obtain similarity scores lower then this threshold are then considered unlikely to share the activity of the training set.

There are many reports [2, 9, 13, 60] which utilize different kind of binary fingerprints for similarity search or for library design. However, the linking of the Tanimoto coefficient value to the probability of activity has been somewhat difficult. Initially two molecules with a pair-wise Tanimoto coefficient of 0.85 were expected to have an 80% chance of sharing the same activity. However, in later studies this probability was re-evaluated and brought down to 30%. A Tanimoto coefficient as low as 0.55 was found necessary in order to include most examples in a patent [15]. Thus, the value of such a threshold is very hard to determine and the optimal threshold value may vary from one activity class to another. In addition, the problem is even more pronounced when a data fusion is used. This is demonstrated by the fact that in our experiments using a threshold on the binary Tanimoto coefficient of 0.75 predicted, on average, in 823, 768, 809, 1,909, 587, 3,121, 1,835, and 926 structures as belonging to the 5-HT3, 5-HT1A, D2, Renin, AT1, Thrombin, HIV-1 P, and PKC activity classes, respectively.

In the case of real-valued structure representations the problem is even more pronounced since not so much prior studies exist. Thus, we decided to avoid the searching for a "best" threshold value of the real-valued Tanimoto coefficient. Such a value is bound to be activity and descriptor specific and, therefore, with limited practical application. As an example, illustrating the above point, using a threshold of 0.75 with topological autocorrelation almost all screened structures were classified as actives when screening for 5-HT3 antagonists. Using the same threshold with RDF codes classified, on average, 1074 of the WOMBAT structures as 5-HT3 antagonists.

The SOM novelty detection handles the threshold determination implicitly. Based on the difficulties encountered by determining a reasonable value of the threshold, needed to turn a similarity search method into a classifier, the SOM novelty detection was identified as the method of choice in this scenario.

Which descriptor to use?

The precision and recall values obtained with SOM novelty detection and the different structure representations when retrieving WOMBAT actives are summarized in Table 7. A relatively broad description of the active space was achieved by using Daylight fingerprints. As much as 10% of all 267,760 screened structures were classified as actives in some cases (5-HT3, 5-HT1A, HIV-1P, PKC). Consequently, the highest recall values were obtained. However, this comes at the price of many false positives. In other words, the confidence that a structure is correctly predicted as active is low.

On the other extreme, a very tight description of the active space is achieved when RDF codes are used to describe the structures. Consequently, the highest precision values are obtained. Therefore, on average, there is around 30% chance that a structure is correctly predicted as active with RDF code.

This high precision comes at the price of a relatively high false negative ratio as indicated by the low recall values. This tight description of the chemical space spanned by the query structures can be attributed to different reasons. One such reason is that the RDF code descriptor is

Table 7 Precision and recall values obtained with SOM novelty detection when retrieving WOMBAT actives

Activity	DFP			AC2D			RDF		
	Predicted as active	Recall	Precision	Predicted as active	Recall	Precision	Predicted as active	Recall	Precision
5-HT3	30,065	53.5	1.9	7,175	35.7	6	645	15.1	19.8
5-HT1A	26,307	58	7.5	10,420	40.7	10.7	1,527	13.2	26.2
D2	10,483	30.3	9.7	7,447	21.2	9.1	1,565	9.1	18.9
Renin	6,737	73.1	6.6	2,561	61.1	18.3	325	26.1	57.3
AT1	1,611	64.3	56.5	3,125	58.5	28.7	369	13.7	58.7
Thrombin	19,592	54.4	5.7	12,661	36	5.5	3,360	14.4	10.5
HIV-1 P	24,784	76.8	12.3	13,542	52.6	13.2	1,142	16.6	47.9
PKC	35,669	77.6	0.8	15,421	61.7	0.8	3,104	67.4	5.9

conformation sensitive. This fact, combined with the aforementioned use of a single low-energy conformation, may lead to narrowing the active space, since in addition to structural and physicochemical features the location of the atoms in the 3D space adds additional constraints. Another possible reason is that the selected dimensionality—four combined 64-dimensional RDF code—is actually too high and the underlying SOM is overfitted to the particular set of query structures.

As can be seen from Table 7 on average around 3.5 % of all screened structures were predicted as actives when using topological autocorrelation. The obtained results were somehow in between SOM novelty detection with binary fingerprints and with RDF codes. A higher false positive ratio than the one obtained with RDF codes and a lower true positive ratio than the one obtained with Daylight fingerprints was observed. Thus, the use of autocorrelation vectors provided the best precision/recall trade-off and is the recommended descriptor in this scenario.

Scenario 4: identification of the most active compound

The objective in this scenario was to investigate to what extent the ranking produced by a virtual screening method correlates with known activity values. To achieve this, the known active structures from the WOMBAT database for which an activity values are available were ranked using similarity search with data fusion and SOM novelty detection. The correspondence between the obtained virtual screening scores and the activity values were measured by Kendall's τ correlation coefficient. The results are summarized in Table 8.

As can be seen from Table 8 no significant difference between both virtual screening methods was found. In all of the screened activity classes no significant correlation—Kendall's τ between -0.1 and 0.2—was found. The best results were obtained for the PKC activity class. Even in this case no Kendall's τ larger than 0.37 was achieved. The low correlation found was not unexpected since both virtual screening methods do not take the activity of the query structures into account. Remember that the query structures were actually drawn from MDDR—a data base in which no measured activity values are available. Therefore, it is not justified to make assumptions regarding the potency of a given structure based on its position in the ranked list, produced by either similarity search or SOM novelty detection.

Conclusions

The applicability of two different virtual screening methods—similarity search with consequent data fusion and novelty detection with Self-Organizing Maps—in four different virtual screening scenarios was investigated. Three different ways of representing chemical structures—binary fingerprints, topological autocorrelation and radial distribution function (RDF) codes—were examined in combination with both virtual screening methods. Both virtual screening methods were found applicable for scenario 1—prioritizing compounds for subsequent high-throughput screening—and scenario 2—selecting a predefined (small) number of potentially active compounds from a large chemical data base. The SOM novelty detection is preferred for scenario 3—assessing the probability that a given structure will exhibit a given activity. Both methods were found inapplicable for scenario 4—selecting the most active structure(s) for a biological assay. The performance of the different descriptors was found to be dependant on the activity class. While no "best-for-all" descriptor was identified, it was found that the topological autocorrelation

Table 8 Kendall's rank correlation between the rank obtained in the virtual screening and the rank according to the activity values when retrieving WOMBAT actives

Activity	DFP				AC2D				RDF			
	SSDF		ndSOM		SSDF		ndSOM		SSDF		ndSOM	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
5-HT3	0.13	0.05	0.16	0.04	0.18	0.02	0.14	0.02	0.15	0.03	0.18	0.04
5-HT1A	0.18	0.03	0.17	0.03	0.18	0.02	0.17	0.02	0.12	0.02	0.11	0.03
D2	0.15	0.02	0.16	0.02	0.14	0.01	0.13	0.01	0.07	0.02	0.09	0.01
Renin	0.14	0.04	0.09	0.06	0.03	0.05	0.02	0.05	0.10	0.05	0.12	0.04
AT1	0.01	0.10	0.01	0.09	0.10	0.03	0.10	0.04	0.12	0.04	0.10	0.05
Thrombin	0.22	0.02	0.22	0.03	0.21	0.03	0.19	0.04	0.23	0.03	0.24	0.04
HIV-1 P	0.15	0.09	0.17	0.08	0.14	0.06	0.11	0.05	0.09	0.06	0.09	0.05
PKC	0.37	0.04	0.34	0.08	0.29	0.04	0.28	0.05	0.29	0.08	0.25	0.06

Mean and standard deviation over 10 runs with different training sets

usually offers the best dimensionality/performance ratio. The use of 3D vectorial descriptor based on a single low-energy conformation (RDF codes) alone gave similar results to the 2D descriptors. However, it covered different parts of the activity spaces under investigation. Consequently, the fusion of the ranked lists obtained with RDF codes and a 2D descriptor improved the results. A difference in the activity space covered by two large databases of biologically relevant compounds—MDDR and WOMBAT—was found. A bias towards retrieving compounds from the same database which was used to select the training set was found. Increasing the size of the training set beyond one hundred compounds did not bring a significant improvement in all scenarios. The studied virtual screening methods were able to recover chemotypes not present in the training set. In addition, an analysis of the top-ranked false positive structures revealed that these structures are likely to share the target activity. Therefore, the proposed methods are likely to work good in prospective virtual screening experiments.

Finally, all ranked lists used throughout this study with the actual similarity scores are freely available at http://www2.chemie.uni-erlangen.de/people/Dimitar_Hristozov/sprt_info.These lists can be used to calculate any performance metric for comparative purposes. The trained self-organizing maps, used as novelty detectors for each activity, are included as well. These can be used to screen any database of chemical compounds, provided that the structures are described with the same descriptor. A simple Python script is provided for this purpose.

A full-featured application allowing the rapid screening of large databases and including the described ligand-based virtual screening methods together with different data fusion techniques is under development and will be available soon.

References

1. Walters WP, Stahl MT, Murcko MA (1998) Drug Discov Today 3:160
2. Bajorath J (2001) J Chem Inf Model 41:233
3. Bajorath J (2002) Nat Rev Drug Discov 1:882
4. Oprea TI, Matter H (2004) Curr Opin Chem Biol 8:349
5. Willett P, Barnard JM, Downs GM (1998) J Chem Inf Model 38:983
6. Bleicher KH, Bohm HJ, Muller K, Alanine A (2003) Nat Rev Drug Discov 2:369
7. Kearsley SK, Sallamack S, Fluder EM, Andose JD, Mosley RT, Sheridan RP (1996) J Chem Inf Model 36:118
8. Bologa C, Revankar CM, Young SM, Edwards BS, Arterburn JB, Kiselyov AS, Parker MA, Tkachenko SE, Savchuck NP, Sklar LA, Oprea TI, Prossnitz ER (2006) Nat Chem Biol 2:207
9. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) J Chem Inf Model 44:1177
10. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) Org Biomol Chem 2:3256
11. Bender A, Jenkins JL, Glick M, Deng Z, Nettles JH, Davies JW (2006) J Chem Inf Model 46:2445
12. Chen B, Harrison RF, Papadatos G, Willett P, Wood DJ, Lewell XQ, Greenidge P, Stiefl N (2007) J Comput Aid Mol Des 21:53
13. Martin YC, Kofron JL, Traphagen LM (2002) J Med Chem 45:4350
14. Matter H (1997) J Med Chem 40:1219
15. Martin YC (2006) QSAR Comb Sci 25:1192
16. Markou M, Singh S (2003) Signal Process 83:2481
17. Markou M, Singh S (2003) Signal Process 83:2499
18. Hristozov D, Oprea TI, Gasteiger J (2007) J Chem Inf Model. http://www.pubs3.acs.org/acs/journals/doilookup?in_doi=10.1021/ci700040r
19. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Nat Rev Drug Discov 3:935
20. Sousa SF, Fernandes PA, Ramos MG (2006) Proteins 65:15
21. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) J Med Chem 49:5912
22. MDL Drug Data Report, version 2006.1
23. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI (2003) In: Oprea TI (ed) Cheminformatics in drug discovery. Wiley-VCH, New York, pp 223–239
24. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) J Chem Inf Model 44:1177
25. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) Org Biomol Chem 2:3256
26. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2005) J Med Chem 48:7049
27. Taylor R (1995) J Chem Inf Model 35:59
28. Butina D (1999) J Chem Inf Model 39:747
29. Truchon JF, Bayly CI (2007) J Chem Inf Model 47:488
30. Edgar SJ, Holliday JD, Willett P (2000) J Mol Graph Model 18:343
31. Hanley JA, McNeil BJ (1982) Radiology 143:29
32. Hanley JA, McNeil BJ (1983) Radiology 148:839
33. Triballeau N, Acher F, Brabet I, Pin JP, Bertrand HO (2005) J Med Chem 48:2534
34. Cleves AE, Jain AN (2006) J Med Chem 49:2921
35. Witten IH, Eibe F (2000) Data mining: practical machine learning tools and techniques. Morgan Kaufmann, San Francisco
36. Yao YY (1995) J Am Soc Inf Sci 46:133
37. Whittle M, Gillet VJ, Willett P, Alex A, Loesel J (2004) J Chem Inf Model 44:1840
38. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2005) J Med Chem 48:7049
39. Kohonen T (2001) Self-organizing maps. Springer, Berlin
40. Sykora V (2007) Chemical descriptors library. Retrieved from cdelib.sourceforge.net 01/2007
41. Moreau G, Broto P (1980) New J Chem 4:359
42. Bauknecht H, Zell A, Bayer H, Levi P, Wagener M, Sadowski J, Gasteiger J (1996) J Chem Inf Model 36:1205
43. Spycher S, Pellegrini E, Gasteiger J (2005) J Chem Inf Model 45:200
44. Fechner U, Franke L, Renner S, Schneider P, Schneider G (2003) J Comput Aid Mol Des 17:687
45. Spycher S, Nendza M, Gasteiger J (2004) QSAR Comb Sci 23:779
46. Teckentrup A, Briem H, Gasteiger J (2004) J Chem Inf Model 44:626
47. Hutchings MG, Gasteiger J (1983) Tetrahedron Lett 24:2541
48. Gasteiger J, Hutchings MG (1983) Tetrahedron Lett 24:2537

49. Gasteiger J, Marsili M (1980) Tetrahedron 36:3219
50. Hollas B (2003) J Math Chem V33:91
51. ADRIANA.Code, version.1.0, 2006, Molecular Networks GmbH, Erlangen, Germany. http://www.molecular-networks.com
52. Hemmer MC, Steinhauer V, Gasteiger J (1999) Vib Spectrosc 19:151
53. Sadowski J, Gasteiger J (1993) Chem Rev 93:2567
54. CORINA, version 3.2. 2003, Molecular Networks GmbH, Erlangen, Germany. http://www.molecular-networks.co
55. Johnson M, MeqiLite, version 2.30, 2007, Pannanugget Consulting L.L.C., Kalamazoo, MI, USA. http://www.pannanugget.com
56. Johnson M (2006) An introduction to the MeqiSuite Indices. Pannanugget Consulting L.L.C. http://www.pannanugget.com/MeqiSuiteIntro.pdf
57. Sammon JR (1969) IEEE T Comput C-18:401
58. R Development Core Team, R: A language and environment for statistical computing, version 2.0, 2005. http://www.r-project.org/
59. Venables WN, Ripley BD (2002) Modern applied statistics with S. Springer, New York, USA
60. Brown RD, Martin YC (1997) J Chem Inf Model 37:1
61. Renner S, Schwab CH, Gasteiger J, Schneider G (2006) J Chem Inf Model 46:2324
62. Ginn C, Willett P, Bradshaw J (2000) Persp Drug Discov Des 20:1
63. Sheridan RP, Kearsley SK (2002) Drug Discov Today 7:903