JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

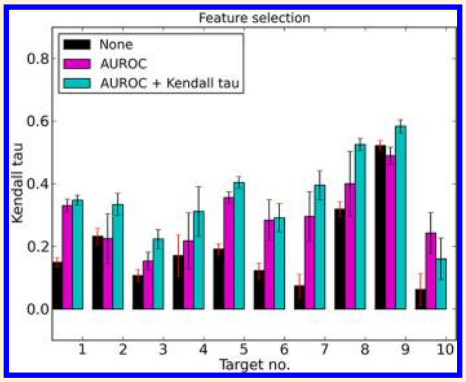
Similarity Searching for Potent Compounds Using Feature Selection

Martin Vogt and Jürgen Bajorath*

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

Supporting Information

ABSTRACT: In similarity searching, compound potency is usually not taken into account. Given a set of active reference compounds, similarity to database molecules is calculated using different metrics without considering compound potency as a search parameter. Herein, we introduce a feature selection method for fingerprint similarity searching to maximize compound recall and preferentially detect potent compounds. On the basis of training examples, fingerprint features are selected that identify potent compounds and produce high recall. Using the reduced fingerprint representations, potent hits are preferentially detected, even if reference compounds have only moderate or low potency. Small sets of simple chemical features are found to yield high search performance.



INTRODUCTION

Similarity searching is one of the major approaches for ligand-based virtual screening.[1,2] For similarity searching, fingerprints, i.e., bit representations of molecular structure and properties, are among the most popular descriptors.[2] Irrespective of the molecular representations and similarity measures that are used, similarity searching typically does not take compound potency as a search parameter into account. An exception is provided by QSAR models adapted for virtual screening,[3] which fall outside the conventional spectrum of whole-molecule similarity-based search methods.[2] Because compound potency represents the dependent variable in QSAR, models can be used to predict the potency of database compounds (if they are structurally related to training set molecules). Apart from a limited number of QSAR-based database search applications, only very few attempts have thus far been made to explicitly consider compound potency in virtual screening. Specifically, two different methods for compound classification have been introduced that take compound potency into account including a potency-directed compound mapping technique[4] and a support vector machine (SVM) classifier.[5] The mapping algorithm was designed to assign database compounds to consensus positions of potent compounds in numerical descriptor spaces.[4] Thus, this methodology takes compound potency indirectly into account. In addition, the SVM method was based upon a so-called structure−activity kernel considering potency directly and involved a linear combination of different models.[5]

For similarity searching using fingerprints, potency-based methods are currently not available. It is often attempted to utilize highly potent molecules as reference compounds hoping that similar database molecules might include potent hits. An open question is how one might utilize potency information more specifically in fingerprint searching to preferentially detect potent hits. Therefore, we have developed a dual-purpose approach for potency-directed similarity searching and optimization of compound recall that is based upon fingerprint feature selection. In fingerprint engineering, defined as the specific modification of fingerprint structures to improve search performance and/or balance molecular complexity effects,[6] statistical feature selection methods have previously been applied to generate compound class-specific fingerprint representations or merge different types of fingerprints.[7,8] These engineered fingerprints often produce higher compound recall than the original fingerprint versions.[7,8]

In our current study, feature selection is applied to identify fingerprint components that detect potent compounds and assign them to high database rank positions. Features are also selected to maximize compound recall. This combined feature selection approach leads to a significant enrichment of highly potent compounds among correctly identified hits. Herein, the development and initial application of the dual-purpose feature selection methodology is reported.

METHODS AND DATA SETS

Feature Selection. The similarity search approach introduced herein relies on compound class-specific feature selection. Different from previous studies[7,8] where statistical methods were applied to assess the significance of individual fingerprint features, we utilize a "wrapper" approach[9] for feature selection, which is conceptually straightforward and can be summarized as follows:

1. Select a subset of features.
2. Evaluate the similarity search performance for this subset using an objective function *f*.

Received: May 29, 2013
Published: June 28, 2013

3. Repeat steps 1 and 2 until a predefined termination condition is met and return the subset that yields the best search performance.

Despite its conceptual simplicity, the wrapper approach has computational requirements that need to be carefully addressed. First, the number of all possible subsets of features grows exponentially with the total number of features and it is computationally intractable to evaluate the performance of all possible subsets. Therefore, heuristic optimization strategies need to be considered. Herein, we apply a backward elimination stategy[9,10] to select fingerprint features for similarity searching. Backward elimination begins with all features and in each step an individual feature is removed, as illustrated in Figure 1a. If the removal results in an improvement of search performance the feature is permanently removed and the process is repeated. If the performance decreases the feature is added back to the subset and another feature is removed. This process continues until no feature remains in the subset whose removal increases the performance (Figure 1a). The backward elimination procedure has a computational complexity that is directly proportional to the square of the number of features.

The second major requirement of the wrapper approach is the calculation of Tanimoto coefficient (Tc)[11] values for reference and screening database compounds for each feature set in order to determine the rank of database hits. From the ranks of available active test compounds, a receiver operating characteristic (ROC)[12] curve is generated and the area under the ROC (AUROC) curve[12] is calculated as a measure of similarity search performance. Benchmark trials need to be carried out for each feature subset and a training set of reference compounds, which also becomes computationally infeasible. Therefore, explicit calculations are circumvented by applying the so-called conditional correlated Bernoulli model[13] (CCBM) for modeling the distribution of Tc values of a screening database with respect to a reference compound. The CCBM approach is summarized in Figure 1b. Given the frequencies and correlation coefficients of all features of a given fingerprint, which are derived from the screening database, the Tc value distribution can be easily modeled for any subset of features by applying the CCBM. As previously demonstrated, ROC curves and AUROC values can be accurately predicted using the CCBM.[13] For training, each highly or weakly potent training set compound (as further specified below) is used once as reference compound while only highly potent training compounds are added as potential hits to the screening database. Then AUROC values are predicted by applying the CCBM. The average AUROC value is then used as an objective function to guide feature selection. Combining backward elimination and the CCBM makes it computationally feasible to apply the wrapper approach for feature selection to optimize similarity search performance. To further improve computational efficiency, features that occur with very similar frequency in the training set and the screening database are removed prior to the backward elimination process because they are not expected to significantly contribute to recall performance. Specifically, all fingerprint features for which the frequency in the training set did not differ from the frequency of the screening database using a significance level of 0.25 as a *p*-value threshold were eliminated. This value was heuristically determined on the basis of initial test calculations comparing fingerprint features in active and database compounds.

Prioritization of Potent Compounds. A major goal of our approach is similarity searching for potent hits. As discussed above, feature selection can specifically focus on highly potent compounds to consider potency implicitly. Moreover, potency criteria can be directly taken into account. Therefore, the objective function, here the AUROC measure, is augmented with an additional term to prioritize highly potent over weakly potent compounds. This can be facilitated by introducing a correlation function that relates the rank of a compound in a similarity search to its potency. For this purpose, we apply the Kendall τ rank correlation coefficient.[14] Considering the Tc values of active compounds relative to a reference as well as the potencies values of these compounds, the Kendall τ coefficient

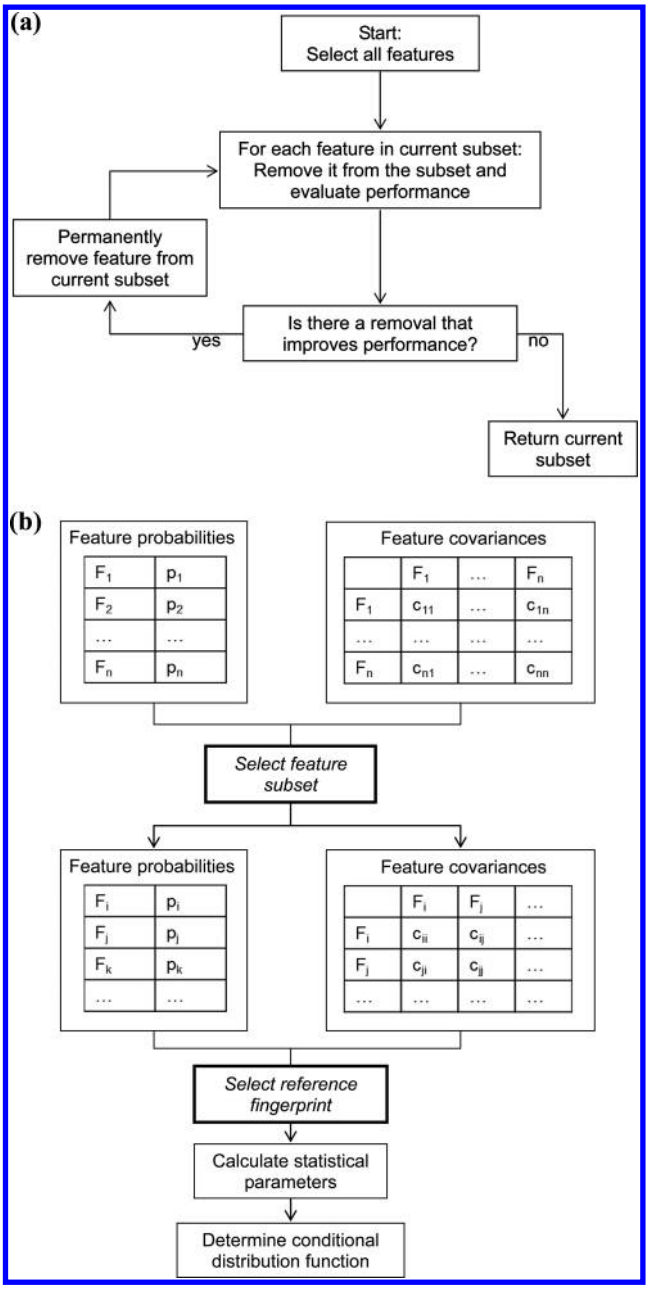


Figure 1. Flowchart representation of key approaches. (a) Feature selection. An outline of the backward elimination feature selection approach is provided. (b) Conditional correlated Bernoulli model (CCBM). The CCBM methodology is summarized. The only time-consuming step in the CCBM workflow is the initial estimation of feature probabilities and covariances. From these, individual CCBM distributions can be directly derived for feature subsets and individual reference fingerprints.

dx.doi.org/10.1021/ci4003206 | J. Chem. Inf. Model. 2013, 53, 1613−1619

quantifies the correlation between a ranking of the compounds based upon the Tc values and a corresponding ranking based upon the potency values.

Incorporation of the rank correlation coefficient modifies function f such that two objectives are taken into consideration:

1. Improve similarity search performance (AUROC).
2. Prioritize highly over weakly potent compounds (Kendall τ).

Therefore, two individual score components are combined into a single objective function using a weighted average:

$$f = (1 - w)\text{AUROC} + w\tau$$

Depending on the weighting scheme, recall performance of active compounds and prioritization of highly potent compounds are differently balanced. The training protocol described above is carried out analogously using the dual-purpose objective function to optimize AUROC values and the Kendall τ rank correlation coefficient simultaneously.

Test calculations revealed notable differences in search characteristics when extreme weights of $w = 0$ and $w = 1$ were applied. However, stable search results were obtained for most weight variations of $0 < w < 1$. Therefore, a constant weight of $w = 0.5$ was ultimately applied in all subsequently reported calculations resulting in an equal weight for both the potency and the recall criterion.

Data Sets and Calculations. For our calculations, compound data sets covering different potency levels were extracted from ChEMBL (version 15).[15] Compounds active against human targets with available K_i values were considered. For a given target, as a minimal requirement, at least 20 compounds with a pK_i value of 7 or greater and 20 compounds with a pK_i of 5 or smaller had to be available. A total of 49 compound activity classes comprising 111−2153 compounds were selected, as reported in Table S1 of the Supporting Information. As a screening database and basis for CCBM application, a random sample of 250 000 compounds was taken from ZINC (version 12).[16] For all feature selection and search calculations, the MACCS fingerprint[17] was used, given its intuitive composition and the interpretability of structural keys.

The compound classes were randomly divided into training and test sets of equal size. Subsequently, each set was subdivided into compound subsets falling into four potency ranges: p$K_i \geq 7$, $7 >$ p$K_i \geq 6$, $6 >$ p$K_i \geq 5$, and $5 >$ pK_i. Table S2 of the Supporting Information reports the composition of all subsets for each activity class.

During training, only compounds falling into the highest potency range (p$K_i \geq 7$) were considered as potential hits for the AUROC calculations. As references for search calculations, compounds having high or low potency were used, i.e., different trials were carried out in which reference compounds were exclusively taken from the set of highly potent compounds (p$K_i \geq 7$) or from weakly potent compounds falling into the range $6 >$ p$K_i \geq 5$. Accordingly, it was also possible to determine whether selected fingerprint features were capable of directing similarity search calculations toward the detection of potent hits, even if weakly potent reference compounds were used.

In summary, given a training set and a subset of features, the determination of the dual-purpose objective function can be described as follows:

1. For the current subset of features derive the CCBM.
2. For each compound falling into potency range p$K_i \geq 7$ (or $6 >$ p$K_i \geq 5$),
   a. Select this compound as reference compound.
   b. Consider all compounds from potency range p$K_i \geq 7$ except the reference compound as potential hits and predict their Tc-based rank using the CCBM.
   c. Determine AUROC values from these ranks.
   d. For all compounds of the training set except the reference compound, determine the Tc values for the reference using the current subset of features.
   e. Determine the Kendall τ correlation coefficient for the Tc- and potency-based rankings.
3. Determine the average AUROC and the average Kendall τ and return the weighted average of these as value of the objective function.

All trials were repeated 25 times for activity classes randomly divided into training and test sets to obtain statistically meaningful results. During feature selection, feature order was randomized at each individual step to avoid order-dependent local minima effects.

For the evaluation of search performance, each highly or weakly potent test set compound was once used as a reference compound for an individual search calculation using the characteristic fingerprint feature set identified during training. These single-reference compound calculations can be easily adapted to calculations using multiple reference compounds using nearest neighbor calculations or other data fusion strategies. For our proof-of-concept investigation, we have given preference to calculations using individual reference compounds to assess the performance of the approach without introducing an additional variable through data fusion.

For highly or weakly potent reference compounds, average AUROC values were determined for test set compounds and search results were separately monitored for the different potency ranges specified above. As a second performance measure, the average Kendall τ coefficient was calculated for the final similarity search rankings.

RESULTS AND DISCUSSION

Potency-Directed Similarity Searching. For potency-directed similarity searching, we have designed a dual-purpose strategy that is based upon feature selection and balances optimization of global compound recall and the preferential detection of potent hits. Dual-purpose training was facilitated by optimizing AUROC values in combination with Kendall τ rank correlation. Selected fingerprint features were then used for similarity searching. As references, either highly or weakly potent compounds were used during training. Hence, feature selection was applied to identify features that improved recall of highly potent compounds when highly potent compounds were used as references or features that improved recall of highly potent compounds when weakly potent compounds were used as references.

In the following, we first present representative results for 10 activity classes, as specified in Table 1 and, then, discuss the global search performance over all 49 classes.

Similarity Searching Using Highly Potent Reference Compounds. In Figure 2a, search results for the 10 classes using highly potent reference compounds in the absence of feature selection are reported. Search performance is separately monitored for compound subsets (potential database hits) at different potency levels. For 8 of 10 activity classes, recall was highest for highly potent hits, although AUROC value differences between compound subsets with different potency ranges were often small (with the exception of class 11060).

dx.doi.org/10.1021/ci4003206 | J. Chem. Inf. Model. 2013, 53, 1613−1619

Table 1. Compound Data Sets

ChEMBL target id	target name	size
11	thrombin	787
43	beta-2 adrenergic receptor	291
72	dopamine D2 receptor	1765
86	monoamine oxidase A	171
194	coagulation factor X	1191
214	muscarinic acetylcholine receptor M4	211
10498	cathepsin L	206
11003	melanocortin receptor 3	373
11060	carbonic anhydrase VII	247
11627	acyl coenzyme A:cholesterol acyltransferase	137

[a]Exemplary activity classes (enzyme inhibitors or receptor ligands) are listed for which similarity search results are reported in Figures 2−6. Size refers to the number of compounds per class.

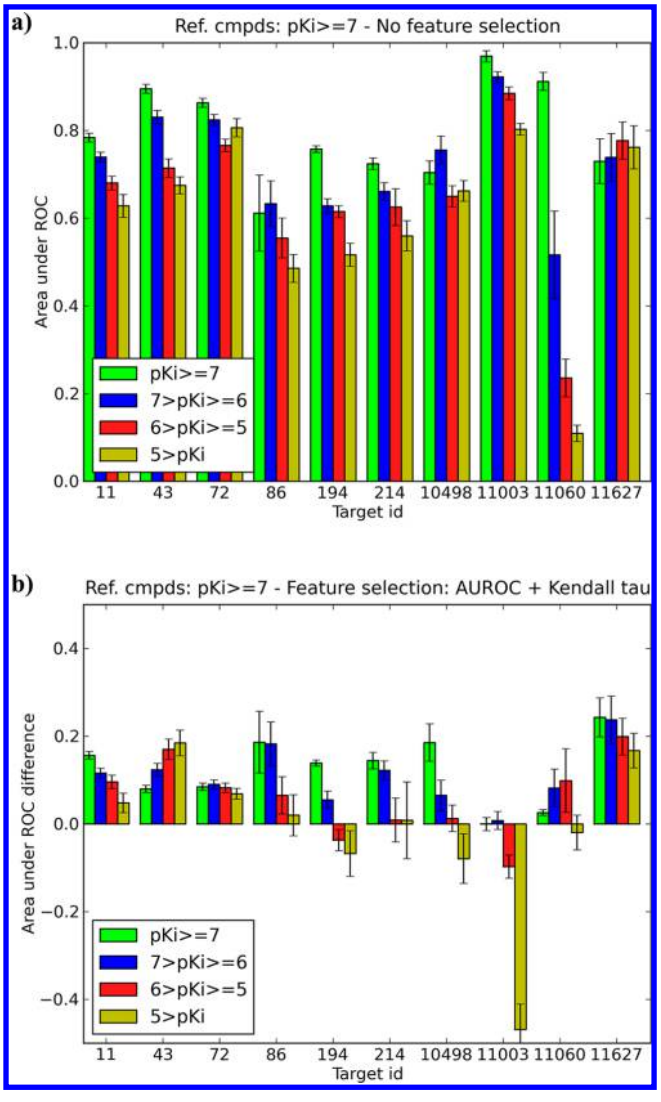


Figure 2. Similarity searching using highly potent reference compounds. (a) Results of similarity searching using MACCS reported in the absence of feature selection. Compound classes are designated using target identifier (id) numbers. The color code indicates search results for compounds at different potency levels. Standard deviations of the search calculations are given at the top of each bar. (b) Differences in search performance reported compared to Figure 2a after applying feature selection in combination with rank correlation (AUROC + Kendall tau).

Recall of the most weakly potent compounds was generally lowest. Thus, standard fingerprint search calculations using highly potent reference compounds displayed a tendency to at least slightly enrich potent hits in database compound rankings, as anticipated.

In Figure 2b, results of search calculations using highly potent reference compounds following dual-purpose feature selection (AUROC plus Kendall τ) are presented. In these cases, differences in recall relative to the standard calculations in Figure 2a are reported. A consistent further increase in search performance was observed, with AUROC difference values of close to or above 0.2 in the majority of cases (leading to nearly ideal search results in several instances). Furthermore, with the exception of three classes (43, 72, and 11060), a notable further enrichment of highly over weakly potent compounds was observed. Moreover, in three cases (194, 10498, and 11003), in part strong deprioritization of weakly potent compounds occurred as a consequence of Kendall τ training, which also represented a desired outcome. Hence, dual-purpose feature selection resulted in further increase of compound recall and enrichment of highly potent compounds.

Similarity Searching Using Weakly Potent Reference Compounds. In Figure 3a, the results of standard search

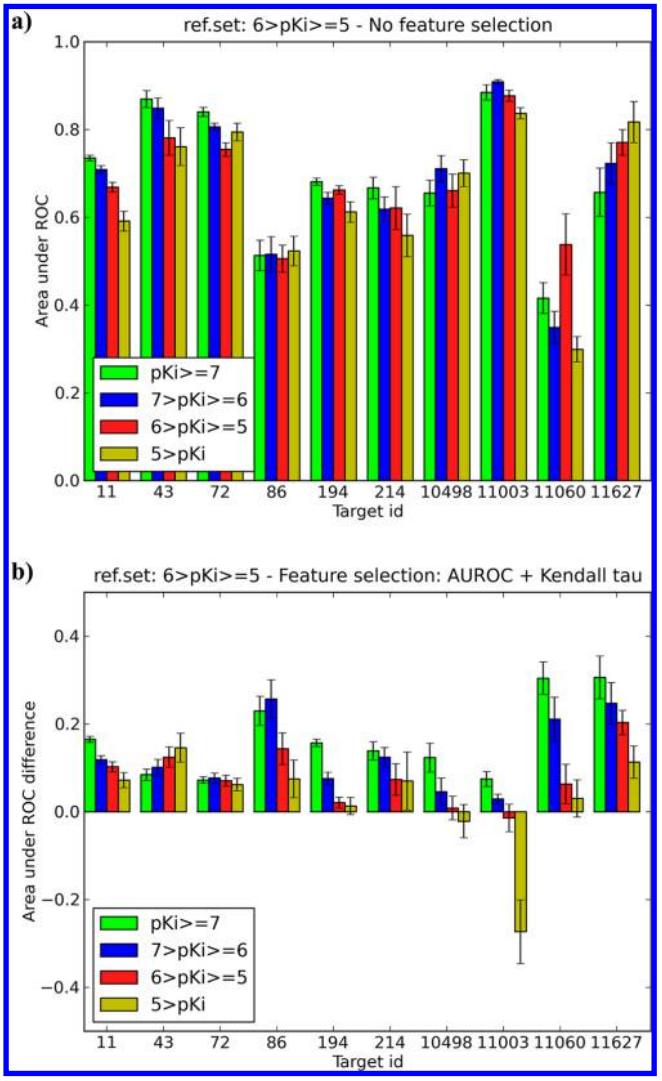


Figure 3. Similarity searching using weakly potent reference compounds. Search results in the absence (a) and presence (b) of feature selection are reported according to Figure 2.

dx.doi.org/10.1021/ci4003206 | J. Chem. Inf. Model. 2013, 53, 1613−1619

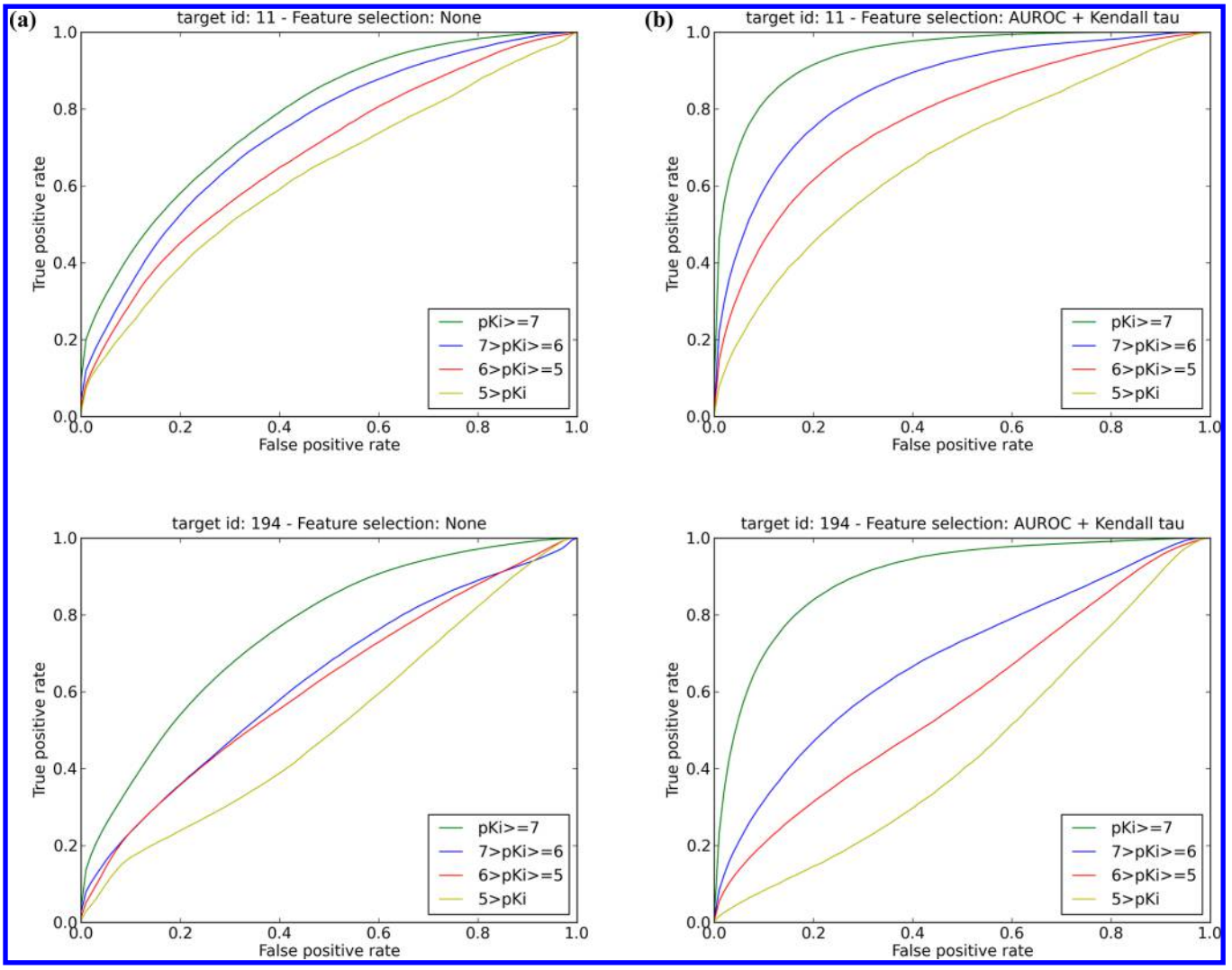


Figure 4. Representative search results. For two compound classes (11 and 194), ROC curves are shown in the absence (a) and presence (b) of feature selection and rank correlation. The color code is according to Figure 2.

calculations for weakly potent reference compounds are shown. Overall, the results were similar to those in Figure 2a, except that recall of highly potent compounds was often comparable to or lower (11060, 11627) than recall of compounds falling into other potency ranges.

In Figure 3b, the results of the corresponding search calculations are shown for feature selection based on weakly potent reference compounds. Again, differences in AUROC values compared to standard calculations are reported. Similar to the results in Figure 2b, consistent improvements in search performance were observed when weakly potent reference compounds were used. Moreover, while deprioritization of weakly potent compounds was only observed in one case (11003), the majority of activity classes displayed a preferential enrichment of highly potent hits. For target 11003, compound recall in the absence of feature selection was already very high (and could hardly be further improved). In the presence of feature selection, recall of highly potent compounds was slightly improved, but weakly potent compounds were strongly deprioritized, which represented a desired result (as stated above). Hence, features extracted from weakly potent reference compounds during training also led to a preferential detection of highly potent hits when weakly potent reference compounds were used.

In Figure 4, exemplary ROC curves are shown in the presence and absence of feature selection that further illustrate the observed effects.

Preferential Detection of Highly Potent Compounds. In Figure 5, average Kendall τ correlation coefficients are reported for search calculations using highly potent reference compounds under different conditions. The correlation coefficients were determined as average Kendall τ values for rankings of the test set using each highly potent compound once as a reference. Search calculations were carried out in the absence of feature selection, the presence of feature selection for improvement of search performance only (AUROC), and the presence of dual-purpose feature selection (AUROC plus Kendall τ). AUROC optimization alone consistently improved rank correlation of potent compounds over standard calculations when highly potent reference compounds were used. Moreover, with one exception (11627), a further increase in rank correlation was observed after dual-purpose training, i.e., highly potent compounds were assigned to higher ranks, leading to a consistent improvement in the detection of potent hits.

Global Search Performance. Table 2 summarizes the search results obtained for all activity classes using highly potent reference compounds. The results of calculations following

dx.doi.org/10.1021/ci4003206 | J. Chem. Inf. Model. 2013, 53, 1613−1619

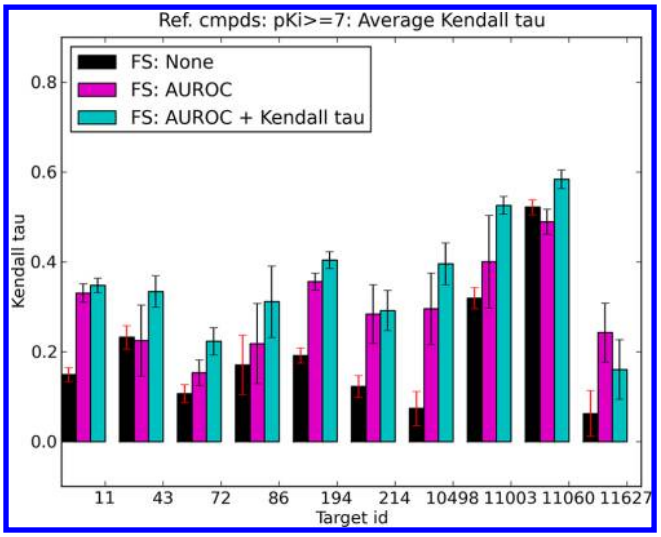


Figure 5. Rank correlation. Average Kendall τ values are reported for search calculations using highly potent reference compounds in the absence of feature selection (FS: None), in the presence of feature selection to improve recall (FS: AUROC), and in the presence of dual-purpose feature selection (FS: AUROC + tau).

Table 2. Search Results[a]

total no. of data sets	49
recall of active cpds with p$K_i \geq 7$	
no. of data sets with significant gain in AUROC ($\geq 1\sigma$)	46
average gain in AUROC (std dev)	0.11 (0.058)
recall of active cpds with $7 > pK_i \geq 6$	
no. of data sets with significant gain in AUROC ($\geq 1\sigma$)	45
average gain in AUROC (std dev)	0.10 (0.057)
recall of active cpds with $6 > pK_i \geq 5$	
no. of data sets with significant gain in AUROC ($\geq 1\sigma$)	38
average gain in AUROC (std dev)	0.071 (0.078)
recall of active cpds with $5 > pK_i$	
no. of data sets with significant gain in AUROC ($\geq 1\sigma$)	25
average gain in AUROC (std dev)	0.028 (0.12)
Kendall τ rank correlation	
no. of data sets with significant gain in Kendall τ ($\geq 1\sigma$)	48
average Kendall τ without feature selection (std dev)	0.20 (0.12)
average Kendall τ with feature selection (std dev)	0.34 (0.10)
AUROC gain of highly potent cpds ($pK_i \geq 7$) relative to other potency ranges	
no. of data sets with on average larger gain compared to ($7 > pK_i \geq 6$)	32
no. of data sets with on average larger gain compared to ($6 > pK_i \geq 5$)	37
no. of data sets with on average larger gain compared to ($5 > pK_i$)	39

[a]A summary of the similarity search results for the 49 activity classes using highly active reference compounds is reported. "σ" refers to standard deviation (std dev).

dual-purpose training are compared with standard search calculations. Increases in global search performance are reported for compounds at all potency levels.

The average recall of highly potent compounds was found to never decrease for any of the 49 classes. In 46 of 49 cases, a significant improvement in recall of highly potent compounds was observed and in 45 cases, an improvement in recall for compounds having potency within the range $7 > pK_i \geq 6$. Equivalent observations were made for 38 activity classes and compounds falling into potency range $6 > pK_i \geq 5$. Furthermore, in 48 of 49 cases, significant increases in Kendall τ rank correlation coefficients were detected. Moreover, preferential gain in the recall of highly potent over weakly potent compounds was detected for 32 to 39 classes, depending on the potency range (bottom of Table 2). Hence, dual-purpose feature selection resulted in a consistent global improvement in search performance and the preferential detection of highly potent compounds.

Selected Features. Dual-purpose training typically resulted in the selection of on average only ~20 to 40 of the 166 MACCS structural keys, as reported in Figure 6. Hence, rather

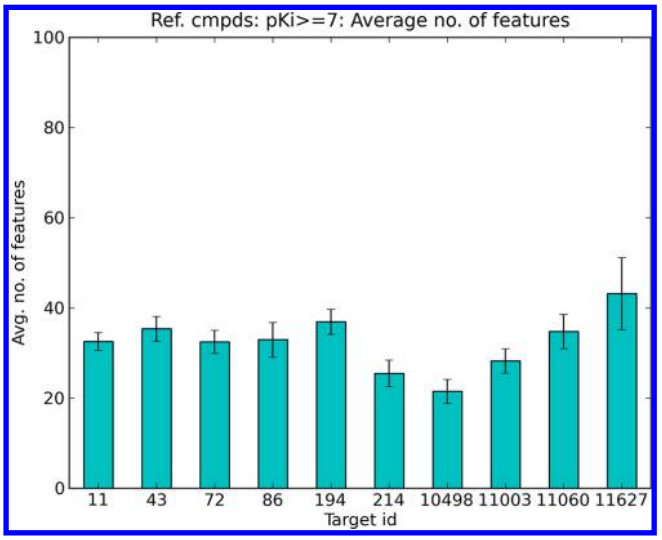


Figure 6. Selected features. Average numbers of features selected using highly potent reference compounds and dual-purpose training are reported.

short fingerprint representations comprising specifically selected features had significant predictive capacity and displayed high search performance, as discussed above. In Table 3, the top 20 most frequently selected MACCS structural keys are reported, which include in part rather simple chemical features. The data set statistics of these features illustrate that frequently selected features might either be over- or underrepresented in highly potent compounds compared to the screening database. For example, MACCS key number 37, NC(O)N, was selected for 30 of 49 activity classes. In three instances, the key had higher frequency in compounds belonging to an activity class than in database compounds, whereas in 27 other cases, it was more frequent in the screening database. By contrast, key 145, 6 M RING > 1, accounting for the presence of more than one six-membered ring in a molecule, which was selected for 27 activity classes, had higher frequency in 24 classes than in database compounds and lower frequency in three cases. Regardless of whether a given feature was over- or underrepresented in an activity class relative to the screening database, the feature was discriminatory because it aided in the prioritization of active compounds or deprioritization of screening database compounds, respectively.

Taken together, the results in Figure 6 and Table 3 illustrate that limited numbers of features from intuitive fingerprints of simple design such as MACCS yielded specific molecular representations for similarity searching that produced high compound recall and enriched highly potent compounds in database rankings.

dx.doi.org/10.1021/ci4003206 | J. Chem. Inf. Model. 2013, 53, 1613−1619

Table 3. Most Frequently Selected Features[a]

MACCS feature	description	no. of compound sets		
		total	increased frequency	decreased frequency
163	6 M RING	31	15	16
37	NC(O)N	30	3	27
22	3 M RING	29	2	27
28	QCH2Q	28	5	23
34	CH2=A	27	1	26
145	6 M RING > 1	27	24	3
125	AROMATIC RING > 1	27	14	13
41	CTN	26	1	25
23	NC(O)O	25	5	20
84	NH2	24	19	5
45	C=CN	24	0	24
19	7 M RING	24	1	23
49	CHARGE	23	20	3
46	Br	23	1	22
131	QH > 1	23	20	3
162	AROMATIC	21	13	8
69	QQH	20	14	6
79	NAAN	19	9	10
47	SAN	19	0	19
101	8M+ RING	19	2	17

[a]Reported are MACCS fingerprint features most frequently selected for highly potent referecne compounds compared to the screening database together with their compound set statistics.

CONCLUSIONS

Herein we have introduced a similarity search approach based upon dual-purpose feature selection. In addition to maximizing compound recall, special emphasis has been put on the preferential detection of highly potent hits. The methodology involves training using a two-component objective function and is conceptually simple. We have shown that combined optimization of AUROC values and Kendall τ rank correlation coefficients for highly or weakly potent compounds was sufficient to yield feature sets of limited size with high search performance. Selected features also led to an enrichment of highly potent hits when weakly potent reference compounds were used. Our proof-of-concept investigation was based on the conventional calculation of Tanimoto similarity using MACCS structural keys. MACCS is a low-complexity fragment-based fingerprint of intuitive design, often considered to be too simplistic for virtual screening. We note that on average less than 30% of the 166 MACCS structural keys were sufficient to generate specific molecular representations meeting our dual-purpose optimization requirements. These findings indicate that combinations of limited numbers of simple, yet carefully selected chemical features might often be highly predictive. The underlying dual-purpose feature selection approach is readily transferable to other molecular representations. Potency-directed fingerprint searching, as introduced herein, further extends the current spectrum of similarity search methods.

ASSOCIATED CONTENT

Supporting Information

Supplementary Table S1 reports the composition of all 49 compound classes. Supplementary Table S2 reports the number of compounds falling into different potency ranges for all 49 classes. This information is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

REFERENCES

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

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

(3) Esposito, E. X.; Hopfinger, A. J.; Madura, J. S. Methods for Applying the Quantitative Structure-Activity Relationship Paradigm. *Methods Mol. Biol.* **2004**, *275*, 131−213.

(4) Godden, J. W.; Stahura, F. L.; Bajorath, J. POT-DMC − A Virtual Screening Method for the Identification of Potent Hits. *J. Med. Chem.* **2004**, *47*, 5068−5611.

(5) Wassermann, A. M.; Heikamp, K.; Bajorath, J. Potency-directed Similarity Searching Using Support Vector Machines. *Chem. Biol. Drug Des.* **2011**, *77*, 30−38.

(6) Heikamp, K.; Bajorath, J. Fingerprint Design and Engineering Strategies: Rationalizing and Improving Similarity Search Performance. *Future Med. Chem.* **2012**, *4*, 1945−1959.

(7) Nisius, B.; Vogt, M.; Bajorath, J. Development of a Fingerprint Reduction Approach for Bayesian Similarity Searching Based on Kullback-Leibler Divergence Analysis. *J. Chem. Inf. Model.* **2009**, *49*, 1347−1358.

(8) Nisius, B.; Bajorath, J. Reduction and Recombination of Fingerprints of Different Design Increase Compound Recall and the Structural Diversity of Hits. *Chem. Biol. Drug Des.* **2010**, *75*, 152−160.

(9) Guyon, I.; Elisseeff, A. An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157−1182.

(10) Stracuzzi, D. J.; Utgoff, P. E. Randomized Variable Elimination. *J. Mach. Learn. Res.* **2004**, *5*, 1331−1362.

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

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

(13) Vogt, M.; Bajorath, J. Introduction of the Conditional Correlated Bernoulli Model of Similarity Value Distributions and its Application to the Prospective Prediction of Fingerprint Search Performance. *J. Chem. Inf. Model.* **2011**, *51*, 2496−2506.

(14) Kendall, M. A New Measure of Rank Correlation. *Biometrika* **1938**, *30*, 81−89.

(15) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2011**, *40*, D1100−D1107.

(16) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757−1768.

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