Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection

Jens Auer and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

A concept termed *Emerging Chemical Patterns* (ECPs) is introduced as a novel approach to molecular classification. The methodology makes it possible to extract key molecular features from very few known active compounds and classify molecules according to different potency levels. The approach was developed in light of the situation often faced during the early stages of lead optimization efforts: too few active reference molecules are available to build computational models for the prediction of potent compounds. The ECP method generates high-resolution signatures of active compounds. Predictive ECP models can be built based on the information provided by sets of only three molecules with potency in the nanomolar and micromolar range. In addition to individual compound predictions, an iterative ECP scheme has been designed. When applied to different sets of active molecules, iterative ECP classification produced compound selection sets with increases in average potency of up to 3 orders of magnitude.

1. INTRODUCTION

Different types of methodologies have been developed or adopted for molecular classification analysis and chemical database mining including, among others, cell-based or statistical partitioning methods,[1] advanced clustering tools,[2,3] neural network techniques,[4,5] Bayesian models,[6] decision trees,[7,8] or kernel-based methods.[9,10] Machine learning techniques for chemical classification generally depend on the availability of training sets, and their performance is often much influenced by the choice of such data. Limited quality of training data often presents a considerable problem for machine learning approaches, and only a few methods have been devised to handle noisy data.[6] Furthermore, in many instances, available data sets are too small for accurate learning or have an unbalanced composition of class labels (e.g. active/inactive).

Our intention has been to develop a novel classification approach that could also be applied in situations where only very limited training data are available, for example, when attempting to guide compound selection or design on the basis of only a few reference molecules. In medicinal chemistry, this is often the case during the early stages of a hit-to-lead transition or lead optimization program. Traditionally, such efforts have been supported by QSAR-type methods[11] to quantitatively model structure–activity relationships and predict analogues having improved potency. However, QSAR methods also require high-quality training data sets containing as many compounds as possible, which are often not available when analyzing novel hits or leads.

To design a classification methodology that could successfully operate on the basis of differently sized training sets, including very small ones, we have evaluated a concept from computer science termed *emerging patterns*[12] to systematically generate molecular feature patterns and identify those that occur with high frequency in one class of molecules (e.g., highly active ones) but not in another (e.g., weakly active or inactive compounds). In bioinformatics, emerging patterns have previously been analyzed to study gene expression profiles and identify genes that distinguish normal and disease cell lines.[13]

We introduce and adapt the concept of emerging patterns for chemical classifications and derive a special form of property descriptor-based patterns termed *emerging chemical patterns* (ECPs). These patterns are applied to predict potency levels of test compounds and classify molecular data sets according to potency criteria. In test calculations on different sets of active compounds, ECP classification displayed high prediction accuracy comparable to state-of-the-art classifiers. When trained on very small sets consisting of only five or 10 compounds, ECP calculations were still capable of predicting potent compounds. The design and evaluation of the ECP methodology is reported herein.

2. METHODOLOGY

2.1. Emerging Patterns. We illustrate the most important aspect of emerging patterns with the help of an example using hypothetical data, shown in Table 1. The data D_1 consists of descriptor values for 15 compounds that are separated into two classes of seven and eight compounds, respectively, based on their activity. Since emerging patterns are constituted of properties, continuous descriptor values ranges need to be discretized into suitable intervals. In this example, three intervals per descriptor are used (with cutoff values derived from database statistics of value distributions found in the ZINC database[14]).

On the basis of descriptor calculations each test compound generates a set of attribute-value pairs, where each pair (*item*) states the name and value of one of the descriptors. A subset

* Corresponding author phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.

EMERGING CHEMICAL PATTERNS

J. Chem. Inf. Model., Vol. 46, No. 6, 2006 **2503**

Table 1. Sample Data for the Calculation of Emerging Patterns[a]

name	molecular weight			logP(octanol/water)			no. of HB-acceptors			no. of HB-donors		
	[0,310)	[310,412)	[412, ∞)	(-∞,2.44)	[2.44, 4.24)	[4.24, ∞)	[0,2)	[2,4)	[4, ∞)	[0,1)	[1,2)	[2, ∞)
active1			x	x				x			x	
active2	x					x		x			x	
active3	x			x							x	
active4	x			x				x				x
active5	x			x			x			x		
active6	x			x				x			x	
active7		x		x				x			x	
inactive1			x			x			x		x	
inactive2	x					x			x			x
inactive3			x	x				x				x
inactive4		x				x			x		x	
inactive5			x	x					x		x	
inactive6			x	x			x				x	
inactive7	x					x		x				x
inactive8	x					x			x			x
test1		x			x			x			x	
test2			x			x	x			x		

[a] Fifteen hypothetical compounds are separated into two classes and represented by four descriptors, each of which is divided into three discrete ranges. In addition, two test compounds are shown. "HB" stands for hydrogen bond.

of all possible attribute-value pairs is called an *item set* or a *pattern*. The set p_1 = {molecular weight: [310,412), logP-(octanol/water): [4.24, ∞)} is an example of a *pattern*. For clarity, we denote a set of attribute-value pairs {(A_1,V_1), (A_2, V_2)} as {$A_1:V_1$, $A_2:V_2$}. When the attribute is a discretized descriptor, the value V_1 describes the numeric range into which the attribute falls. The frequency of a pattern p in a set of compounds (i.e. the frequency of compounds that share a pattern p) is called its *support* in the data D, abbreviated $supp_D(p)$. From Table 1, we compute the support of the pattern p_1 = {molecular weight: [314,012), (logP(octanol/water): [4.24, ∞)} as $supp_{D_1}(p_1) = 0.07$ since there is exactly one compound in the entire data set that contains this pattern as a subset. By contrast, the support for the pattern p_2 = {molecular weight: [0,310)} is $supp(p_2) = 0.53$, since it is produced by eight of 15 compounds. The pattern p_3 = {logP-(octanol/water): [2.44,4.24)} is not a subset of any instance D_1 and hence has zero support.

The data set in Table 1 consists of two classes, "active" and "inactive". It is evident that certain patterns are more frequent in one class than the other, for example, the pattern {("no. of HB-donors", [0,1))}. This pattern is found only in one instance, one active compound, and thus represents a unique feature of the active class. Patterns with significant support in one class or data set and small support in another are called *emerging patterns* (EPs).[12,15] To estimate the discriminating power of EPs, the coefficient of both support rates is used to calculate the *growth* of an EP with respect to two classes of data D_1 and D_2:

$$\text{growth}_{D_1,D_2}(p) = \frac{\text{supp}(D_1)}{\text{supp}(D_2)}$$

A special case of EPs is presented when the support in the background data set D_2 is zero but not in the other. These patterns are called *jumping emerging patterns* (JEPs)[16] for which the *growth* rate is not defined, but that are expected to become highly discriminatory. However, even in low-dimensional data sets, an abundance of JEPs might occur because any set that includes a JEP as a subset is also a JEP. Considering again the pattern {("no. of HB-donors", [0,1))} having zero support in the inactive subset, adding a

feature to this pattern can only change its support in the active class. Since we add a feature, we decrease its support and thus reduce its predictive ability. This motivated the introduction of *most expressive* JEPs.[16] A JEP is *most expressive* if none of its subsets itself is a JEP and if no superset has a larger support in the data set. Thus, a most expressive JEP becomes a highly discriminatory feature for a compound class. However, within this class, it occurs in as many compounds as possible. The set of most expressive JEPs is computed by selecting those JEPs that are minimal with respect to a subset relationship because the support decreases when patterns contain an increasing number of elements. As implemented, the algorithms guarantee that the computed JEPs are minimal and no further processing is required.

2.2. Molecular Classification via Emerging Patterns. Emerging patterns can be used to build various classifiers.[15−17] For example, if test compounds display EPs having strong support in one class and very little or none in others, they are assigned to the class with highest support. However, mining EPs and JEPs is computationally challenging.[18] As an efficient approximation, we apply a hypergraph-based algorithm[19] to mine *most expressive* JEPs from two classes of data. In our study of molecular patterns, we define *emerging chemical patterns* (ECPs) as the *most expressive* JEPs arising from descriptor-dependent feature analysis.

The classification process examines a test instance (molecule) to find all previously computed ECPs from one class that are a subset of the test instance. For these ECPs, support values in the training data are summed and define the final score for the instance to belong to this class. Ultimately, the predicted class is the class with highest score, i.e., the class which has most cumulative support from stored ECPs. This simple scoring scheme is problematic when the training set is unequally distributed among the classes, since the number of ECPs computed from one class might be very large compared to the other class. This problem is overcome by dividing the accumulated support for all ECPs of a test compound by the sum of all ECPs computed for the training set such that the score lies within the interval [0,1]. This procedure normalizes significant differences in the numbers of ECPs when one of the compared classes is much larger.

Table 2. Emerging Chemical Patterns[a]

ECPs: active	support	ECPs: inactive	support
{H-don:[0,1)}	0.14	{MW:[412,∞), logP:[4.24, ∞)}	0.13
{H-acc:[2,4), H-don:[1,2)}	0.57	{MW:[412,∞), H-acc:[4,∞)}	0.25
{MW:[0,310), H-don:[1,2)}	0.43	{MW:[412,∞), H-don:[2,∞)}	0.13
{MW:[0,310), logP:(-∞,2.44)}	0.57	{logP:[4.24, ∞), H-acc:[4,∞)}	0.50
{MW:[0,310), H-acc:[0,2)}	0.43	{logP:[4.24, ∞), H-don:[2,∞)}	0.38
{MW:[310,412), logP(-∞, 2.44)}	0.14	{H-acc:[4,∞), H-don:[2,∞)}	0.25
		{MW:[310,412), logP:[4.24, ∞)}	0.13
		{MW:[412,∞), H-acc:[0,2)}	0.13

[a] Listed are ECPs computed from the data in Table 1.

Table 3. Compound Classes and Potency Levels[a]

class	n	<1 μM	>1 μM	max	min	av
BZR	321	283	38	0.00034	250.00	2.02
DHFR	586	249	337	0.0023	929.00	16.67
GSK3	464	281	183	0.0008	1000	12.84
HIVPROT	967	821	146	0.000015	200	2.68

[a] Compound potencies are reported as IC50 values [μM]; "n" is the number of compounds per class, and the next two columns report how many of these compounds have IC50 values below or above 1 μM; "max", "min", and "av" give the highest, lowest, and average compound potency, respectively.

To illustrate the classification process, we again consider the data provided in Table 1. After dividing the data set into two classes according to hypothetical activity (active/inactive label), we derive the ECPs shown in Table 2. These patterns are reminiscent of Lipinski's rule-of-five[20] and reflect its content in a concise way: since ECPs represent a minimal set, each pattern including more than one descriptor value falling into the range of largest values is a likely signature of an inactive compound.

Given these ECPs, we can now classify the two hypothetical test compounds reported in Table 1. Compound test1 contains only the ECP {H-acc:[2,4), H-don:[1,2)} from the active class so that the accumulated support for this class is 0.57. For the inactive class, its support is zero since it contains none of the ECPs of this training class. Thus, test1 would be predicted to be active. By contrast, compound test2 contains the ECP {H-don:[0,1)} of the active class and ECPs {MW:[412,∞), H-acc:[0,2)}, {MW:[412,∞), logP:[4.24, ∞)} of the inactive class. The resulting support is 0.14 for the active and 0.28 for the inactive class, to which test2 would then be assigned.

2.3. Discretization of Continuous Descriptor Values. The ECP classification depends on the availability of discrete attribute-value pairs. Since many chemical descriptors are continuous (or pseudocontinuous) in nature, their value ranges must thus be divided into discrete intervals. Therefore, we have investigated two discretization techniques implemented in the WEKA library for machine learning.[21] The first approach uses a simple binning scheme where the value range of each descriptor is divided into 10 bins of equal width, not taking into account any information concerning the overall distribution of descriptor values. The second method is based on an attribute splitting criterion often used in the construction of decision trees.[22] The splitting points $\{T_i | i \in \{1,...,n\}\}$ are recursively determined for each descriptor $D_i \in \{ D_1,...,D_n\}$ as those points between two instances that minimize the *class information entropy* (representing a measure for the randomness of the induced partition). For a

compound set S with binary class labels, the information entropy E of a subset $S_j \subseteq S$ is defined as

$$E(S_j) = -P_1(S_j)*\log (P_1(S_j)) - P_2(S_j)*\log (P_2(S_j))$$

$P_1(S_j)$ and $P_2(S_j)$ represent the proportion of compounds in class 1 and 2, respectively.

If the data set S is divided into two partitions S_1,S_2 by assigning compounds with a value for descriptor D_i smaller than T_i to the partition S_1 and the remaining compounds to partition S_2, then the *class information entropy* is defined as

$$E(D_i,T_i;S) = \frac{|S_1|}{|S|}E(S_1) + \frac{|S_2|}{|S|}E(S_2)$$

Here, the expression $|S|$ is defined as the number of compounds in a data set S.

The class information entropy estimates the degree of randomness in both partitions for a given split-point T_i of descriptor D_i. Minimizing the class information entropy for split-points minimizes the randomness of the resulting data intervals. Following each split, the procedure is recursively applied to both partitions until a convergence criterion is reached that balances the number of splits and the information content of the resulting partitions.

Besides the fact that this method takes the class value distribution into account and derives data-dependent split-points, it also maps values of descriptors having no information content into one single interval, which makes it readily possible to eliminate them.

In our classification calculations, class information entropy-dependent discretization of descriptors produced consistently better results than the simple binning scheme. An exemplary calculation is provided in the Results section.

3. CALCULATIONS

3.1. Compound Classes. Emerging chemical patterns were analyzed for four publicly available compound data sets with greatly varying potency (IC50) distributions: benzodiazepines (BZR),[23] dihydrofolate reductase inhibitors (DHFR),[23] glycogen synthase kinase-3 inhibitors (GSK3),[24] and HIV protease inhibitors (HIVPROT).[24] The composition of these data sets is summarized in Table 3, and their potency distribution is further illustrated in Figure 1. For each of these classes, IC50 values cover at least 3 orders of magnitude. In Figure 2, representative structures spanning the potency range are shown. When each data set is separated into two classes of compounds with potency above or below 1 μM, the relative sizes of these classes substantially differ across the four data sets (Table 3). The 1 μM threshold value was used
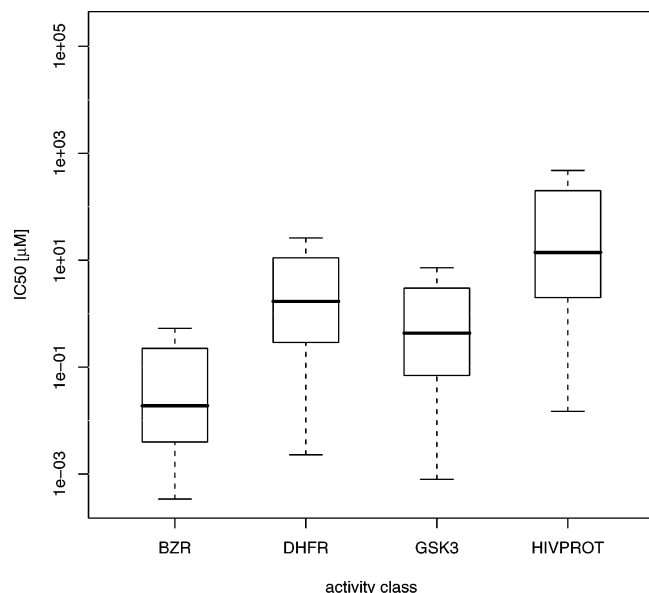
Figure 1. Potency distribution within compound activity classes. Shown are box plots for each activity classes (see Table 3). In these plots, the box shows the 0.75 (top) and 0.25 quartile (bottom) separated by the median (horizontal bar). The lines indicate the largest and smallest values (falling within a distance of maximum 1.5 times the box size from the nearest hinge).

to generate compound classes of higher ("nanomolar") and lower ("micromolar") potency for compound classification analysis.

3.2. Descriptors. We used a set of 61 1D and 2D molecular descriptors implemented in MOE[25] that were previously shown to display only little pairwise correlation but to have high information content in a screening database of 1.34 million compounds.[26] These descriptors were then discretized with both discretization techniques. The binning scheme divided the value range found in each data set into

10 subranges. For each activity class, the entropy-based discretization method eliminated a number of descriptors that were mapped to a single interval and thus not suitable for ECP calculations. This process substantially reduced the number of descriptors for BZR and DHFR (where 18 and 19 remained, respectively), whereas the majority of descriptors passed discretization for GSK3 (47) and HIVPROT (45). All descriptors used for ECP calculations are described in Supporting Information Table 1.

3.3. Classification and Method Comparison. The ECP approach was applied to classify compounds belonging to either the sets with potency above or below an IC50 value of 1 μM and compared to two established and widely used standard classification methods, binary QSAR[6] and decision trees, as implemented in MOE. For each classification, training sets of increasing size (10−50% of compounds belonging to each activity class) were selected. For binary QSAR, we applied MOE contingency analysis to select a subset of most suitable descriptors prior to model building. For classification, a probability threshold value of 0.5 was applied for each compound to belong to the higher potency class. For decision trees, the 61 pre-selected descriptors available in MOE were evaluated during the tree construction process. The constructed tree was postprocessed in a pruning step to remove not required attributes and increase general applicability. We applied a procedure that builds two trees in a 2-fold cross-validation step and selects the best one as the final decision tree structure.

For all three methods, classifiers were trained on 500 randomly selected sets, the resulting models applied to predict the potency-dependent class label of the remaining compounds and average accuracies calculated. Initially, training sets containing 10−50% of compounds per class were used. In a second series of calculations, we focused on training sets of very small size, only consisting of three, five,

Figure 2. Representative structures. For each activity class, the most (max) and least (min) potent compounds are shown together with one having an IC50 value equal or close to the median.

Table 4. ECPs for Activity Class GSK3[a]

ECP IC50 < 1 μM	support	ECP IC50 > 1 μM	support
{peoe_vsa-3:′(7.317−10.975]′}	0.37	{slogp_vsa0:′(-inf-7.169]′, a_nbr:′(-inf-0.3]′, vsa_don:′(-inf-4.356]′,}	0.48
{peoe_vsa+3:′(5.938−11.876]′, a_ns:′(-inf-0.2]′, vsa_don:′(8.712−13.068]′}	0.35	{a_nbr:′(-inf-0.3]′, vsa_don:′(-inf-4.356]′, vsa_pol:′(-inf-5.166]′,}	0.46
{peoe_vsa+2:′(7.943−15.887]′, peoe_vsa-4:′(-inf-4.7228]′,}	0.35	{a_don:′(-inf-0.6]′, a_nbr:′(-inf-0.3]′,}	0.46
{smr_vsa2:′(15.308−22.961]′, slogp_vsa3:′(-inf-11.064]′, vdisteq:′(3.19268−3.439]′,}	0.34	{smr_vsa3:′(-inf-3.346]′, slogp_vsa8:′(-inf-9.434]′, a_nbr:′(-inf-0.3]′,}	0.39
{peoe_vsa+2:′(7.94−15.887]′, peoe_vsa+3:′(5.938−11.876]′,}	0.34	{peoe_vsa+3:′(-inf-5.938]′, a_don:′(-inf-0.6]′,}	0.39
{peoe_vsa+2:′(7.943−15.887]′, vsa_don:′(8.712−13.068]′,}	0.34	{peoe_vsa+3:′(-inf-5.938]′, vsa_don:′(-inf-4.35602]′,}	0.39
{smr_vsa2:′(15.308−22.961]′, slogp_vsa3:′(-inf-11.064]′, slogp_vsa5:′(-inf-12.281]′, a_nf:′(-inf-0.3]′,}	0.32	{peoe_vsa+5:′(-inf-4.06]′, slogp_vsa0:′(-inf-7.169]′, slogp_vsa8:′(-inf-9.434]′, a_nbr:′(-inf-0.3]′, b_triple:′(-inf-0.1]′,}	0.37
{slogp_vsa3:′(-inf-11.068]′, slogp_vsa6:′(-inf-0.441]′, vdisteq:′(3.193−3.439]′, a_ns:′(-inf-0.2]′,}	0.32	{peoe_vsa+4:′(-inf-3.943]′, peoe_vsa+5:′(-inf-4.06]′, slogp_vsa0:′(-inf-7.169]′, a_nbr:′(-inf-0.3]′,}	0.37
{peoe_vsa+6:′(-inf-3.84]′, slogp_vsa3:′(-inf-11.064]′, slogp_vsa5:′(-inf-12.281]′, vdisteq:′(3.193−3.439]′,}	0.32	{smr_vsa3:′(-inf-3.346]′, a_nbr:′(-inf-0.3]′, vsa_don:′(-inf-4.356]′,}	0.37
{peoe_vsa+3:′(5.938−11.876]′, a_nbr:′(-inf-0.3]′, vsa_don:′(8.712−13.068]′,}	0.32	{slogp_vsa0:′(-inf-7.169]′, slogp_vsa4:′(-inf-5.545]′, vsa_don:′(-inf-4.356]′,}	0.37

[a] ECPs with highest support are reported for a single training calculation on a total of 25% GSK3 compounds. In this example, descriptor value intervals were determined using the linear binning scheme. (inf stands for infinity).

or 10 compounds per class from each potency range. This presents a typical lead optimization scenario where only a few active compounds are available as an information source for predictions.

3.4. Simulated Lead Optimization. To simulate a lead optimization process, we applied an iterative classification procedure. During each iteration, we randomly selected training sets of five or 10 compounds from the currently available set of test compounds and divided them into two classes of high and low potency compounds with their mean potency value as the threshold. This compound set was then used to train the ECP classifier to distinguish higher from lower potency compounds. Then, the class label of all remaining test compounds was predicted, assigning each test compound either to the high or low potency class. All compounds predicted to have low potency were then removed from the test set, and only compounds classified as highly potent were retained for the next iteration.

Ten iterations were carried out per activity class, and averages were calculated over 500 independent runs. Over these iterations, the enrichment of potent compounds in the test sets of decreasing size was monitored.
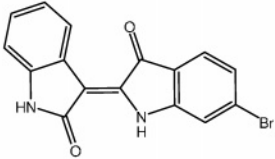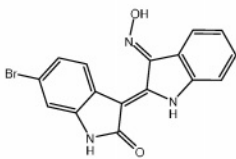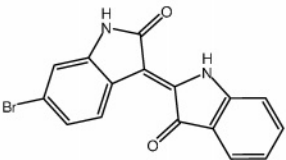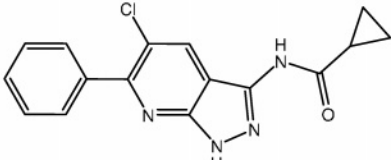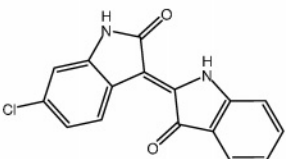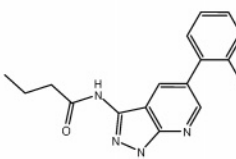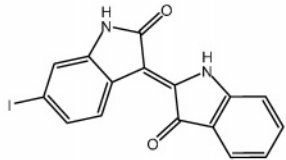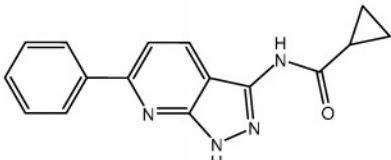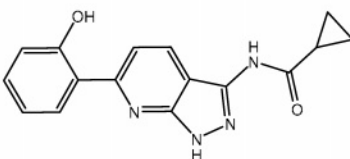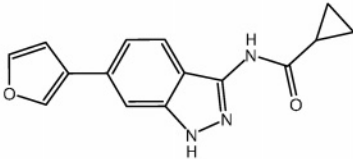
4. RESULTS

4.1. Emerging Chemical Patterns. To illustrate the basis for ECP classification, Table 4 reports 10 patterns with highest support obtained by training with randomly selected 25% of the GSK3. For the remaining 75%, the class label was predicted. ECPs are often chemically intuitive and can be easily interpreted. In this test calculation, the simple descriptor binning scheme was applied for demonstration purposes. For example, the top ECP for the <1 μM class consists of a single descriptor range and occurs in 37% of the compounds of this class but none of the other. Another interesting ECP is the pattern {a_don:′(3.6−4.2]′}, accounting for hydrogen bond donors, having a support of 11% in the <1 μM class. In the >1 μM class, patterns involving descriptors counting the number of bromine atoms and hydrogen bond donors with small value ranges are prominent and occur in direct combination in the pattern {a_don:′(-
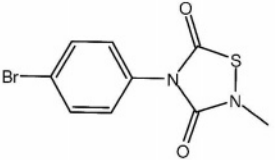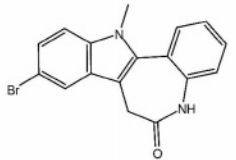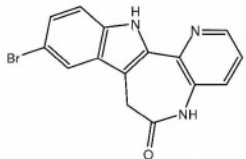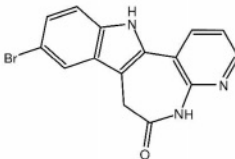
inf-0.6]′,a_nbr:′(-inf-0.3]′,} with 46% support. In addition to the top 10 patterns reported in Table 4, we also identified patterns pointing at sulfur atoms as a discriminating feature ({a_ns:′(1.8-inf)′} with 17% support) and limited polar surface area within the range 28.81−41.83, captured by ECP {tpsa:′(28.81−41.83]′} having a support of 22%. Figure 3a shows the top GSK3 compounds predicted to belong to the <1 μM class and reveals accurate predictions; only one compound was incorrectly classified. This false-positive prediction can easily be explained when considering the training data. The numerical descriptors were divided into 10 equal-size intervals, which represents a low resolution encoding assigning exactly the same intervals to compounds 1 and 3 in Figure 3a. Thus, the classifier considered them identical molecules. The top 10 GSK3 compounds predicted to belong to the >1 μM class are shown in Figure 3b. Here three of 10 compounds were misclassified. The ECP classifier correctly inferred that thiadiazolidinone (TDZD) derivatives have potency >1 μM, although the GSK3 data set only contained 29 thiadiazolidinone derivatives with IC50 values between 2 and 100 μM. Thus, ECPs were highly discriminatory even for relatively underrepresented chemotypes. These calculations were then repeated applying information entropy-based descriptor discretization. The top 10 compounds for each predicted class are shown in Figure 4a,b. Here the classifier achieved 100% accuracy, both for "nanomolar" and "micromolar" compounds. These findings demonstrate the predictive value of patterns derived from descriptor settings subjected to entropy-based discretization.

4.2. Comparison of Different Classification Methods. We then compared ECP with binary QSAR and decision tree classification for training sets of varying size, predicting test compounds to belong to either the "micromolar" or "nanomolar" potency class. Prior to theses calculations, continuous descriptor values were discretized using the information entropy-based methodology on the complete data set. The results are reported in Table 5. For these training sets consisting of 10−50% of all active compounds, all three classifiers performed almost equally well, achieving overall prediction accuracy at the 80% level. In fact, it was interesting to see that the difference in performance between

EMERGING CHEMICAL PATTERNS

J. Chem. Inf. Model., Vol. 46, No. 6, 2006 **2507**

(a)

1 — potency: 22	2 — potency: 0.005
3 — potency: 0.045	4 — potency: 0.234
5 — potency: 0.14	6 — potency: 0.027
7 — potency: 0.055	8 — potency: 0.425
9 — potency: 0.036	10 — potency: 0.035

(b)

1 — potency: 3	2 — potency: 0.4
3 — potency: 0.018	4 — potency: 6
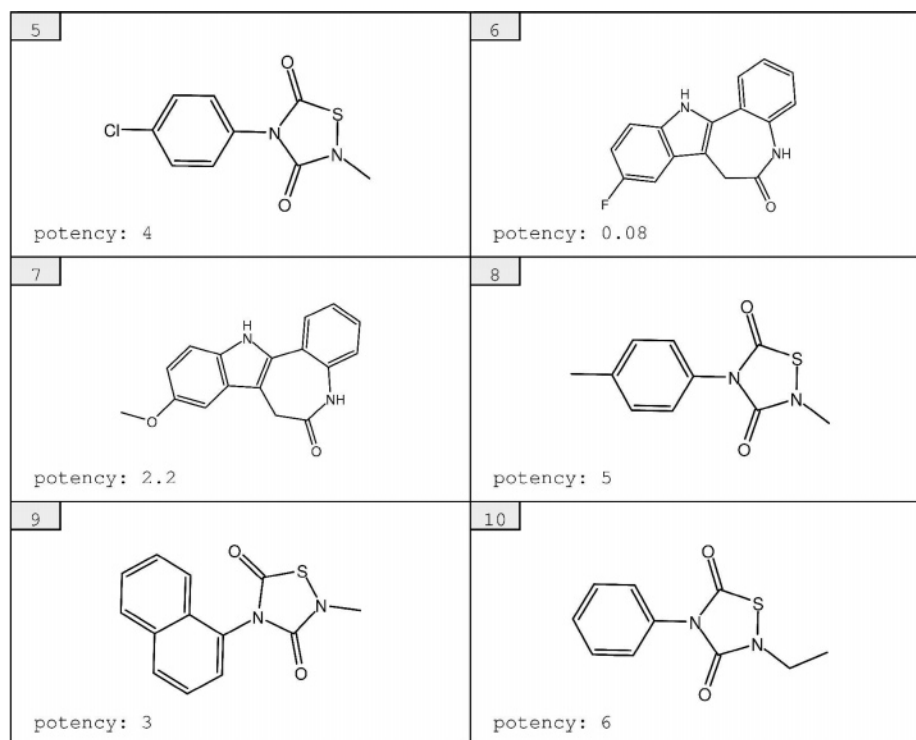
Figure 3. ECP classification of GSK3 compounds based on simple descriptor binning. Shown are the top 10 compounds predicted to belong to the (a) <1 μM class or (b) >1 μM class.
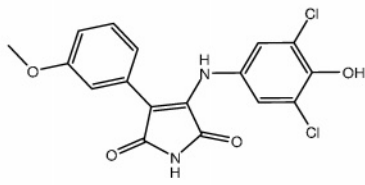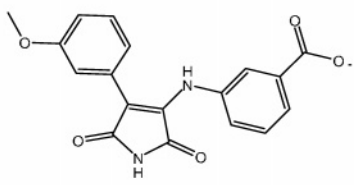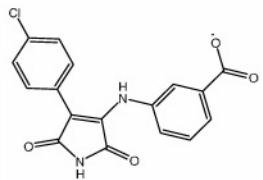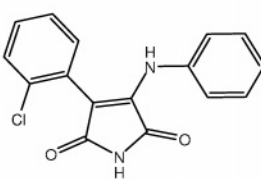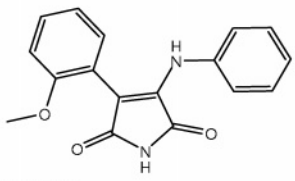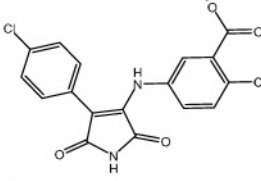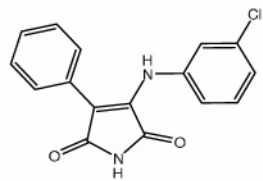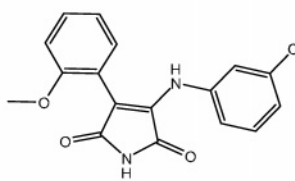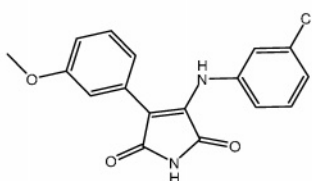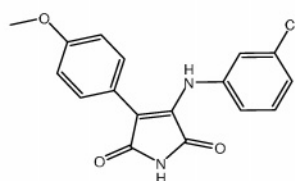
these methods was very small for all four activity classes. Prediction accuracy was consistently lowest for DHFR (approximately 60−70%) and highest for HIVPROT (approximately 80−90%), irrespective of the methodology.

Having confirmed the general predictive ability of ECP classification we focused on predictions based on very small training sets of 10 or fewer compounds. It was thought that ECP should be well suited to operate under such unusual training conditions because of the high level of resolution emerging chemical patterns displayed in compound ranking. Thus, it should be possible to extract discriminatory patterns from only a few compounds. Therefore, for each activity set we assembled training sets of 10, 5, or 3 compounds from each of the <1 μM and >1 μM potency classes by random selection and predicted the class label of the remaining compounds. The results are presented in Table 6. Under these challenging training conditions, ECP performed better than binary QSAR or decision tree classification, in three cases still reaching approximately 80−90% average prediction accuracy for training on only three compounds. Overall ECP performed best on three compound sets and binary QSAR on one. For the smallest learning sets, decision tree classification lost any predictive ability, because it classified all compounds as highly potent, yielding an artificial prediction accuracy of 100% for the < 1 μM class. For three of our classes, binary QSAR showed slightly better results for prediction of compounds from the <1 μM class than ECP. However, for all classes, the prediction accuracy for compounds from the >1 μM class was clearly below random, thus indicating the presence of systematic prediction errors. Thus, binary QSAR could not be applied in a meaningful way to these test cases when only three compounds were used for training. For training sets composed of five compounds, where apparent systematic prediction errors were absent, the prediction accuracy of ECP was significantly
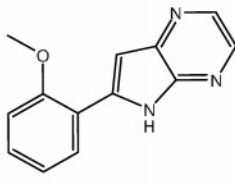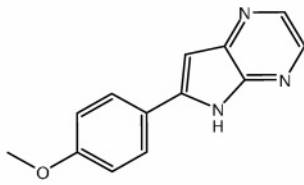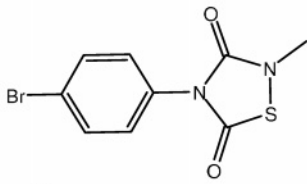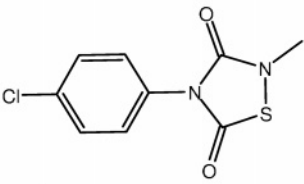
higher than for decision tree and binary QSAR calculations. Thus, overall only ECP calculations displayed consistent predictive ability for very small training sets.

4.3. Simulated Lead Optimization Studies. On the basis of these findings, we analyzed the potential of ECP to systematically enrich potent compounds in selection sets when trained on 10 or fewer "micromolar" and "nanomolar" compounds, a scenario that is fairly typical during the initial stages of a lead optimization effort. The results are shown in Figure 5. For each activity class and training set size, the calculations consistently reached convergence during the first eight or fewer iterations, producing selection sets of fewer than 10 compounds with average potency in the submicromolar range. Training sets of five or 10 compounds produced comparable results. During the first few iterations a sharp decline in compound numbers was observed for all classes accompanied by significant reduction in average activity. Observed potency enrichments ranged from one (GSK3) to three orders of magnitude (HIVPROT). To achieve potency enhancements of this magnitude, random selection of small learning sets and subdivision into classes covering the top half or bottom half of the compounds based on a simple potency ranking was sufficient, consistent with the high-resolution nature and discriminatory power of emerging chemical patterns. For comparison, the same lead optimization simulation procedure was implemented for decision tree and binary QSAR as a classification technique, and the results are presented in Supporting Information Figures 1 and 2. The observed potency enrichments significantly differed from ECP. As expected, the decision tree classifier could not be used in a meaningful way for only five training compounds because it classified the entire test set to belong to one class, either eliminating all compounds, if classified as low potency molecules, or retaining all compounds for the next iteration, if classified as high potency ones.

EMERGING CHEMICAL PATTERNS

J. Chem. Inf. Model., Vol. 46, No. 6, 2006 **2509**

(a)

1 potency: 0.142	**2** potency: 0.195
3 potency: 0.186	**4** potency: 0.216
5 potency: 0.216	**6** potency: 0.109
7 potency: 0.301	**8** potency: 0.114
9 potency: 0.257	**10** potency: 0.156

(b)

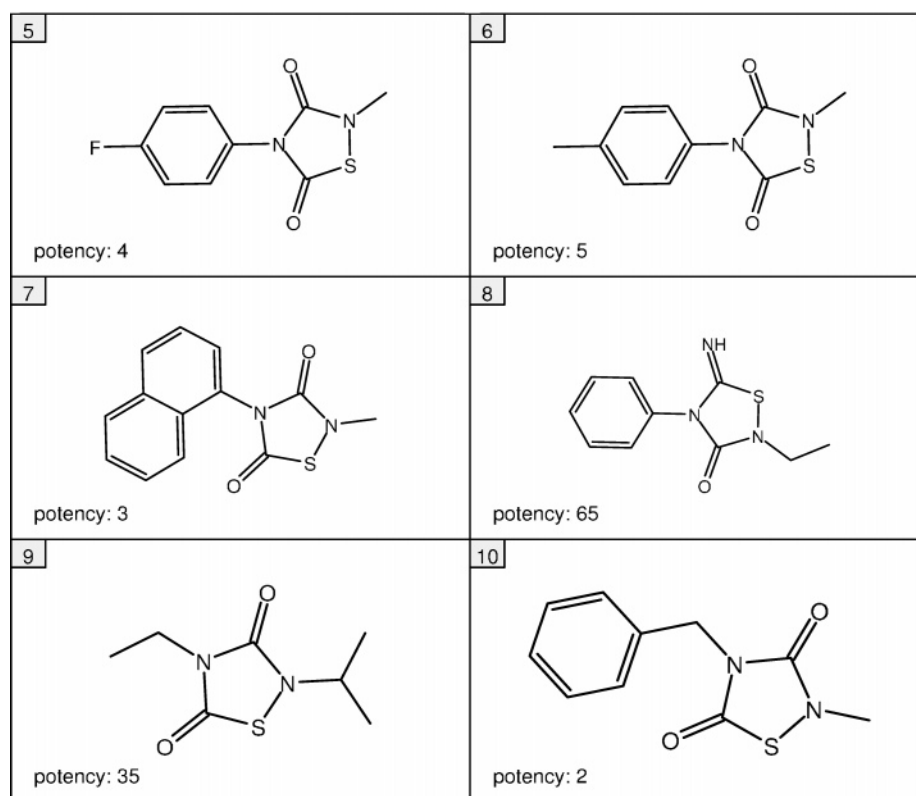|	
1 potency: 3.3	**2** potency: 1.1
3 potency: 3	**4** potency: 4

Figure 4. ECP classification of GSK3 compounds based on entropic descriptor discretization. Shown are the top 10 compounds predicted to belong to the (a) <1 μM class and (b) >1 μM class.

Table 5. Comparison of Classification Methods[a]

| | training set | | | | | | | | | | | | | | |
| | 10% | | | 20% | | | 30% | | | 40% | | | 50% | | |
class	ECP	BIN	DT	ECP	BIN	DT	ECP	BIN	DT	ECP	BIN	DT	ECP	BIN	DT
BZR	0.82	0.78	0.85	0.85	0.67	0.83	0.86	0.73	0.84	0.86	0.77	0.85	0.87	0.80	0.85
DHFR	0.64	0.66	0.62	0.67	0.69	0.65	0.67	0.70	0.67	0.67	0.70	0.68	0.67	0.71	0.68
GSK3	0.76	0.77	0.79	0.78	0.81	0.82	0.79	0.82	0.83	0.79	0.82	0.84	0.80	0.82	0.85
HIVPROT	0.89	0.78	0.84	0.89	0.84	0.85	0.89	0.86	0.86	0.89	0.87	0.87	0.89	0.88	0.86
average	0.78	0.75	0.78	0.80	0.75	0.79	0.80	0.78	0.80	0.80	0.79	0.81	0.81	0.80	0.81

[a] Average prediction accuracies for five training set sizes (10−50% of each activity class) are reported for ECP, binary QSAR (BIN), and decision tree (DT) classification. The last row reports the average accuracy over all classes.

Compared to binary QSAR, ECP produced much better potency enrichments potencies on all classes, by at least one order of magnitude for classes HIVPROT and GSK3. For these classes, binary QSAR calculations failed to produce a median potency value below 1 μM. On training sets with 10 compounds, ECP also performed better than the other methods. Only for GSK3, the decision tree classifier achieved a potency enrichment close to ECP. For HIVPROT, only ECP produced potencies below 1 μM, with a sharp decline in average potency from the third iteration on.

5. DISCUSSION

The concept of emerging patterns has been evaluated to differentiate between chemical features represented through calculation and discretization of conventional molecular descriptors. On the basis of this analysis, emerging chemical patterns have been introduced as a new methodology for compound classification. In our initial test calculations, we found that the ECP method performed as well as binary QSAR and decision tree classifiers in the prediction of

Table 6. Prediction Accuracy for Very Small Training Sets[a]

| | no. of compounds | | | | | | | | |
| | 3 | | | 5 | | | 10 | | |
class	ECP	BIN	DT	ECP	BIN	DT	ECP	BIN	DT
	(a) Potency Class <1 μM								
BZR	0.62	0.72	1.00	0.75	0.58	0.57	0.74	0.57	0.59
DHFR	0.54	0.68	1.00	0.72	0.54	0.58	0.73	0.71	0.59
GSK3	0.57	0.74	1.00	0.80	0.64	0.68	0.82	0.51	0.69
HIVPROT	0.79	0.73	1.00	0.78	0.65	0.63	0.81	0.57	0.66
	(b) Potency Class >1 μM								
BZR	0.88	0.45	0.0	0.75	0.55	0.64	0.79	0.58	0.63
DHFR	0.75	0.39	0.0	0.55	0.55	0.50	0.59	0.44	0.50
GSK3	0.86	0.44	0.0	0.68	0.52	0.65	0.72	0.70	0.62
HIVPROT	0.57	0.45	0.0	0.61	0.52	0.59	0.65	0.62	0.63

[a] Average prediction accuracies are separately reported for both potency classes and very small training sets of three, five, or 10 compounds from each class. Abbreviations are used according to Table 5.

relative compound potencies. Different from classical QSAR approaches, ECP does not attempt to predict actual activity
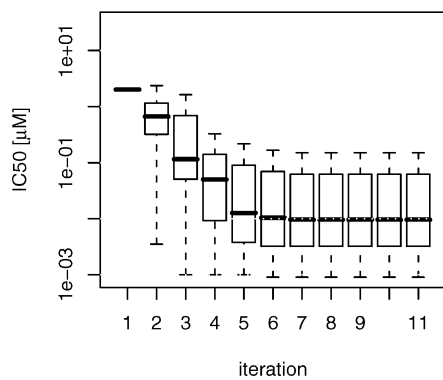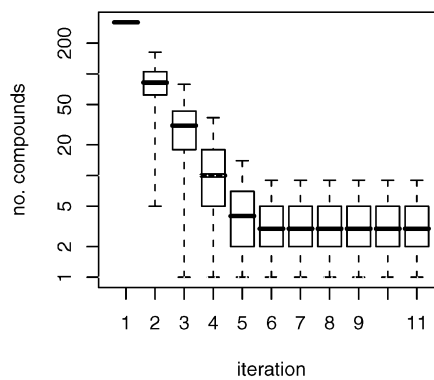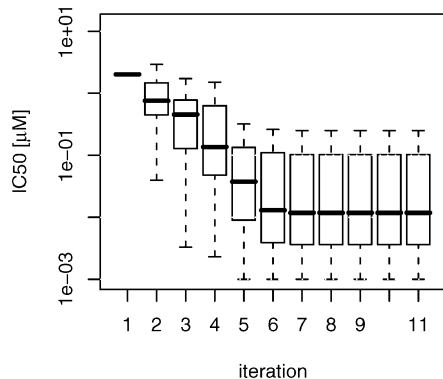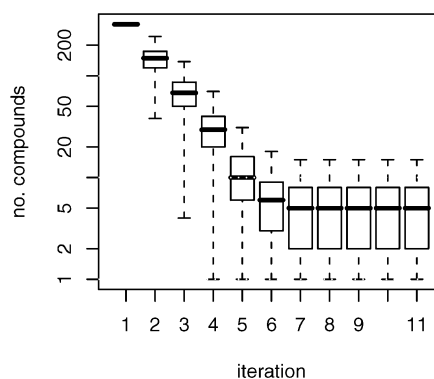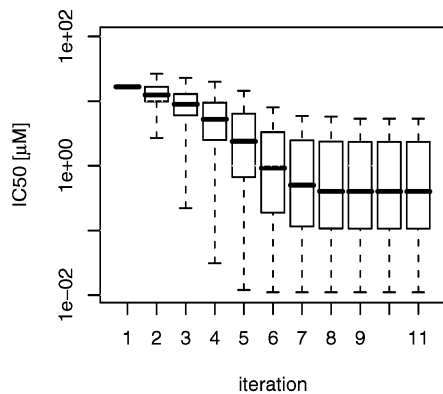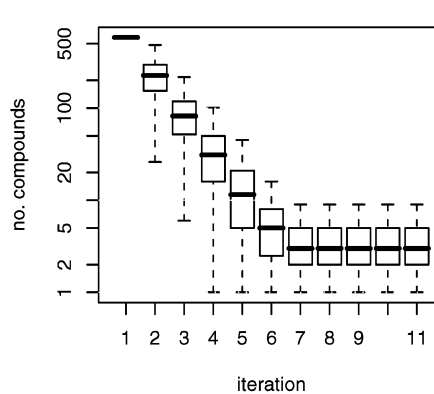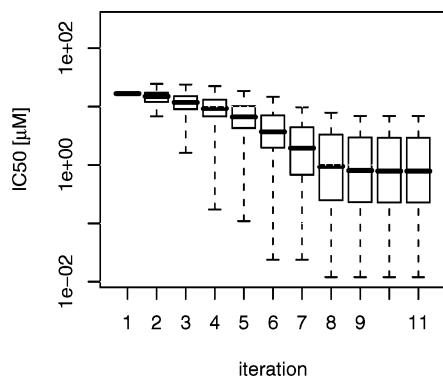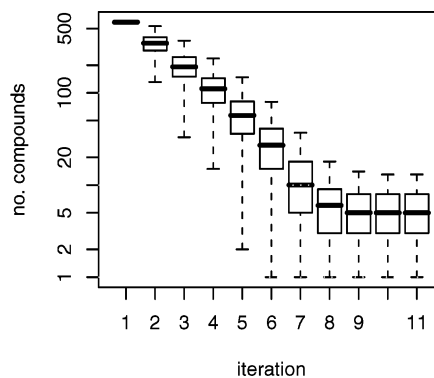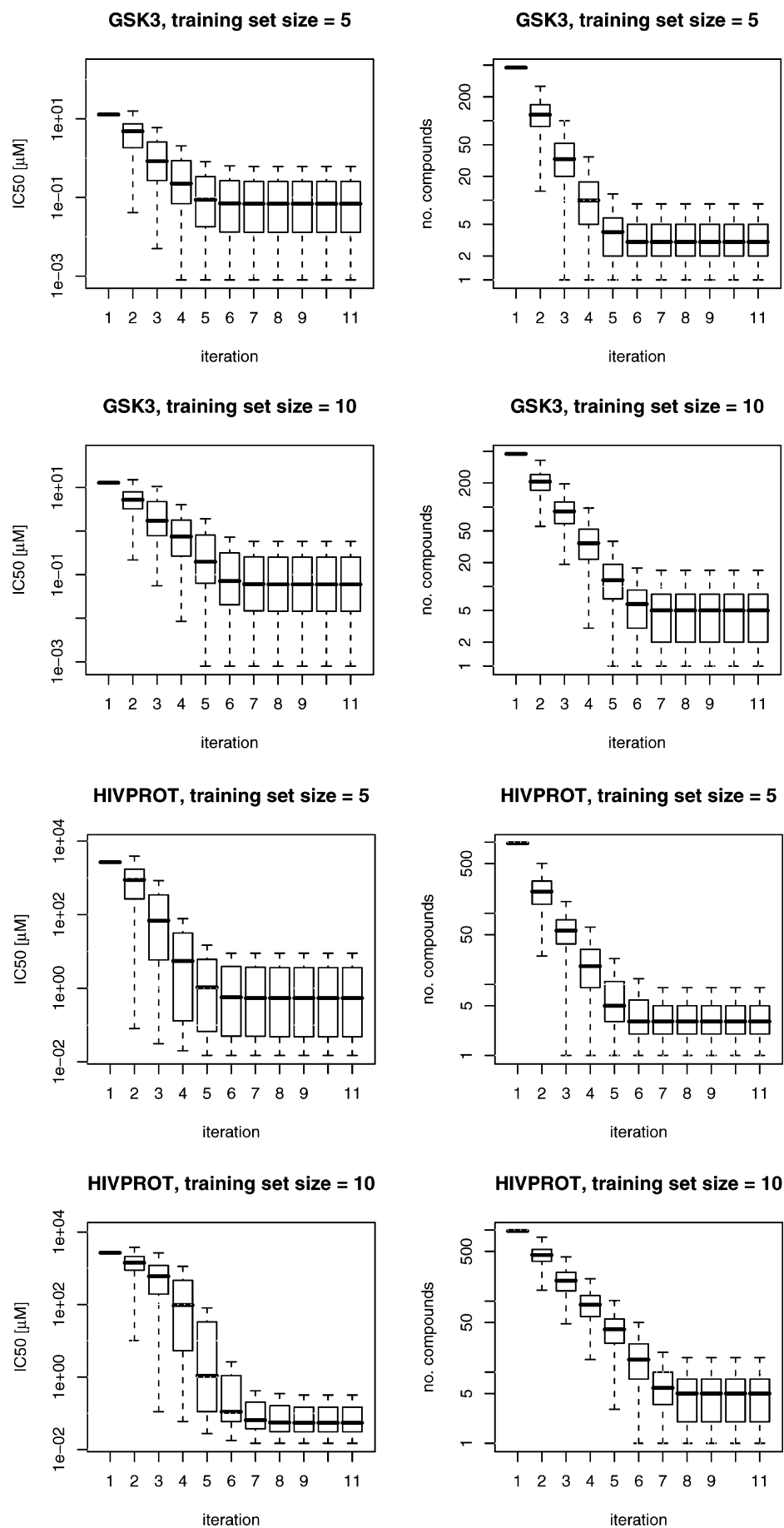
Figure 5. Simulated lead optimization trials. Average potencies and compound numbers over 500 calculations of 10 iterations each are reported for training sets of five and 10 compounds. Box plot representations are according to Figure 1.

values. Rather, it is trained to predict compounds to be active above or below a predefined potency threshold level and is thus conceptually more similar to binary QSAR analysis. However, as we could confirm on the basis of additional test calculations, a major distinguishing feature of ECP is its ability to successfully operate on very small training sets where other classifiers become unreliable and are difficult to apply. The ability to handle a very small training set was a major reason for the design of ECP. Our results suggest that if sets of only three to five highly active and weakly active (or inactive) compounds are available, a predictive ECP model can be built and used to identify other potent molecules or predict the potency level of newly designed compounds. These features make the ECP approach attractive to aid in analogue design during early stages of lead optimization where molecular information is usually rather limited and where often too few molecules with different potencies are available to build conventional QSAR models.

The iterative classification scheme devised for ECP led to significant enrichment of potent compounds in small selection sets, without the need for careful training set assembly. This approach should have considerable potential for the analysis of analogue or target-focused libraries[27] that can be enumerated in silico using different design protocols[28] and the selection of preferred subsets or single molecules. Precomputed libraries focused on known active compounds can be subjected to iterative ECP classification in order to identify library subsets enriched with compounds predicted to be most potent.

Despite its encouraging performance, the ECP approach also has some general limitations. First, ECP classifications are qualitative in nature, and the computed scores do not reflect absolute differences in potencies and are thus not appropriate for potency-based compound ranking. Furthermore, for large compound sets, ECP calculations become computationally expensive, since the computational complexity for mining ECPs is high. Mining ECPs is an NP-hard problem, meaning that the computational time grows exponentially with the number of compounds and descriptors used. Finally, ECP predictive accuracy is influenced by the degree of similarity of training compounds. If the compounds are highly similar, the probability is high that many descriptors will be discretized into only one range and thus be eliminated by the information-based discretization technique, which in turn makes it difficult to identify highly discriminatory patterns. However, this problem also reflects a unique strength of the ECP approach, its ability to derive sound predictive models on the basis of few compounds having different structural features yet similar activity. This suggests that ECP can be successfully applied to predict highly active with significant structural modifications compared to learning set molecules, which also distinguishes the ECP approach from QSAR-type methodologies.

6. CONCLUSIONS

The ECP methodology introduced herein is capable of extracting discriminatory feature patterns from test molecules and using this information for class label predictions. An important feature of the ECP approach is its ability to derive predictive models on the basis of little molecular information. This makes the methodology attractive for applications in, for example, lead optimization programs and screening of analogue libraries. On the basis of the results presented herein, ECP is expected to further add to the spectrum of molecular classification methods and complement currently available computational methodologies to aid in lead optimization.

Supporting Information Available: Molecular descriptors used in this study (Table 1) and the results of simulated virtual screening trials using binary QSAR and decision tree calculations (Figures 1 and 2, respectively). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Stahura, F. L.; Bajorath, J. Partitioning methods for the identification of active molecules. *Curr. Med. Chem.* **2003**, *8*, 707−715.
(2) Tamura, S. Y.; Bacha, P. A.; Gruver, H. S.; Nutt, R. F. Data analysis of high-throughput screening results: application of multi-domain clustering to the NCI anti-HIV data set. *J. Med. Chem.* **2002**, *45*, 3082−3093.
(3) Feher, M.; Schmidt, J. M. Fuzzy clustering as a means of selecting representative conformers and molecular alignments. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 810−818.
(4) Sadowski, J. Optimization of chemical libraries by neural network methods. *Curr. Opin. Chem. Biol.* **2000**, *4*, 280−282.
(5) Keseru, G. M.; Molnar, L.; Greiner, I. A neural network based virtual high throughput screening test for the prediction of CNS activity. *Comb. Chem. High Throughput Screening* **2000**, *3*, 535−540.
(6) Labute, P. Binary QSAR: a new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput.* **1999**, *4*, 444−455.
(7) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(8) Stockfisch, T. P. Partially unified multiple property recursive partitioning (PUMP-RP): a new method for predicting and understanding drug selectivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1608−1613.
(9) Harper, G.; Bradshaw, J.; Gittin, J. C.; Green, D. V. S.; Leach, A. R. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.
(10) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *44*, 549−561.
(11) Esposito, E. X.; Hopfinger, A. J.; Madura, J. D. Methods for applying the quantitative structure−activity relationship paradigm. *Methods Mol. Biol.* **2004**, *275*, 131−214.
(12) Dong, G.; Li, J. Efficient mining of emerging patterns: discovering trends and differences. In *Conference on Knowledge Discovery in Data*, Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, 1999; Chaudhuri, S., Fayyad, U., Madigan, D., Eds.; ACM Press: New York, 1999; pp 43−52.
(13) Li, J.; Wong, L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics* **2002**, *18*, 725−734.
(14) Irwin, J. J.; Shoichet, B. K. ZINC − a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.
(15) Dong G.; Zhang, X.; Wong, L.; Li, J. CAEP: Classification by aggregating emerging patterns. In *Lecture Notes in Computer Science, Vol. 1721*, Proceedings of the Second International Conference on Discovery Science, Tokyo, 1999; Arikawa, S., Furukawa, K., Eds.; Springer-Verlag: London, U.K., 1999; pp 30−42.
(16) Li, J.; Dong, G.; Ramamohanarao, K. Making use of the most expressive jumping emerging patterns for classification. *Knowl. Inf. Syst.* **2001**, *3*, 131−145.
(17) Li, J.; Dong, G.; Ramamohanarao, K.; Wong, L. DeEPs: a new instance-based lazy discovery and classification system. *Machine Learn.* **2004**, *54*, 99−124.
(18) Wang, L.; Zhao, H.; Dong, G.; Li, J. On the complexity of finding emerging patterns. *Theor. Comput. Sci.* **2005**, *335*, 15−27.
(19) Bailey, J.; Manoukian, T.; Ramamohanarao, k. A fast algorithm for computing hypergraph transversals and its application in mining emerging patterns. In *3rd IEEE International Conference on Data Mining*, Proceedings of the 3rd IEEE International Conference on Data Mining, Melbourne, FL, 2003; IEEE Computer Society: Los Alamitos, CA, 2003, p 485.

(20) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(21) Witten, I. H.; Frank, E. Introduction to Weka. In *Data mining: practical machine learning tools and techniques*, 2nd ed.; Morgan Kaufmann Publishers: San Francisco, CA, 2005; pp 365−368.

(22) Fayyad, U. M.; Irani, K. B. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proceedings of the 13th International Joint Conference on Artificial Intelligence, Chambry, France, 1993*; Bajcsy, R., Eds.; Morgan Kaufmann Publishers: San Francisco, 1993; pp 1022−1027.

(23) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. Spline-fitting with a genetic algorithm: a method for developing classification structure−activity relationships. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1906−1915.

(24) Chen, X.; Lin, Y., Gilson, M. K. The binding database: overview and user's guide. *Biopolymers Nucleic Acid Sci.* **2002**, *61*, 127−141.

(25) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: 1255 University Street, Montreal, Quebec, Canada H3B 3X3.

(26) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151−1157.

(27) Schnur, D.; Beno, B. R.; Good, A., Tebben, A. Approaches to target class combinatorial library design. *Methods Mol. Biol.* **2004**, *275*, 355−377.

(28) Rose, S.; Stevens, A. Computational design strategies for combinatorial libraries. *Curr. Opin. Chem. Biol.* **2003**, *7*, 331−339.