Journal of
Chemical Information
and Modeling

Article

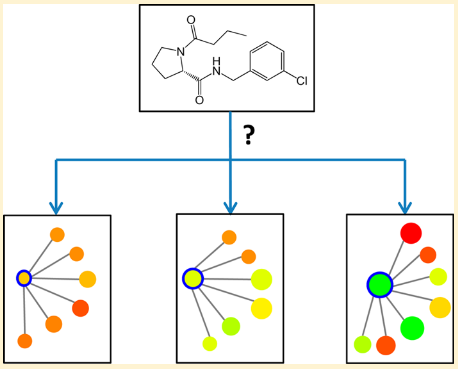
pubs.acs.org/jcim

Prediction of Compounds in Different Local Structure–Activity Relationship Environments Using Emerging Chemical Patterns

Vigneshwaran Namasivayam, Disha Gupta-Ostermann, Jenny Balfer, Kathrin Heikamp, and Jürgen Bajorath*

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

S Supporting Information



ABSTRACT: Active compounds can participate in different local structure–activity relationship (SAR) environments and introduce different degrees of local SAR discontinuity, depending on their structural and potency relationships in data sets. Such SAR features have thus far mostly been analyzed using descriptive approaches, in particular, on the basis of activity landscape modeling. However, compounds in different local SAR environments have not yet been predicted. Herein, we adapt the emerging chemical patterns (ECP) method, a machine learning approach for compound classification, to systematically predict compounds with different local SAR characteristics. ECP analysis is shown to accurately assign many compounds to different local SAR environments across a variety of activity classes covering the entire range of observed local SARs. Control calculations using random forests and multiclass support vector machines were carried out and a variety of statistical performance measures were applied. In all instances, ECP calculations yielded comparable or better performance than controls. The approach presented herein can be applied to predict compounds that complement local SARs or prioritize compounds with different SAR characteristics.

INTRODUCTION

Emerging patterns is a machine learning methodology developed in computer science to identify class-specific feature patterns for objects and utilize these patterns for class label prediction.[1−6] In 2006, the methodology was adapted in chemoinformatics as emerging chemical patterns (ECP) for classification of active compounds.[7] It was found that the ECP approach yielded accurate predictions on the basis of much smaller training sets of active compounds than other machine learning methods,[7] which made ECP attractive for classification tasks when only little information was available for model building. Accordingly, ECP was subsequently applied to simulate iterative screening experiments[8] and compare crystallographic and theoretically derived bioactive compound conformations.[9] More recently, the methodology has also been applied to identify toxic compounds,[10] classify compounds with multitarget activities,[11] and predict individual compounds forming activity cliffs.[12] However, with fewer than 10 ECP applications reported thus far, the approach continues to be underexplored compared to other machine learning methods that are popular in chemoinformatics such as Bayesian classifiers[13] or support vector machines (SVMs).[14] A reason for this might be that ECP calculations are computationally expensive for large data sets. This is the case because pattern mining is an NP-hard problem,[3,4] i.e., the computational time required for mining of typically large numbers of ECP increases exponentially with the number of compounds and descriptors that are used. This is the computational price to pay for one of ECP's two most attractive features, i.e., its ability to effectively utilize very small training sets, as mentioned above. In addition, ECP are often chemically interpretable, depending on the descriptors used, which sets the approach apart from black box machine learning approaches. Hence, at least for modeling data sets of limited size and applications constrained by sparse training data, ECP should have significant potential and merit further investigation.

The most recent ECP application demonstrated that single compounds having either high or low potency could be predicted to form activity cliffs on the basis of characteristic chemical descriptor patterns,[12] which offered an alternative to the canonical compound pair-based approach of assessing activity cliffs.[15] This has encouraged us to ask the question whether local structure–activity relationship (SAR) characteristics could be predicted in a more general manner for individual compounds, which is of considerable interest, for example, for activity landscape modeling and analysis[16] or the selection of SAR-informative compounds from large data sets. Generalized predictions of SAR features would require assigning individual compounds to different local SAR environments, from continuous local SARs, which are characterized by the presence of structurally similar compounds

Received: March 7, 2014
Published: May 6, 2014

with similar potencies, to discontinuous SARs that are formed by structurally similar compounds with significant difference in potency. Because activity cliffs represent the pinnacle of SAR discontinuity, their assessment can be rationalized as a special case of local SAR analysis.

Herein, we generalize the prediction of local SAR environments for single compounds based upon an approach that conceptually differs from the assessment of activity cliffs. ECP models are built to systematically predict compounds exhibiting different degrees of SAR discontinuity covering the entire spectrum of SARs across different data sets. We show that many compounds in different local SAR environments can be correctly predicted, regardless of their biological targets and potency distributions.

■ METHODS AND MATERIALS

Emerging Chemical Patterns. For derivation of ECP, value ranges of chemical descriptors must be discretized into defined intervals.[17,18] For a set of descriptors, a compound produces a set of attribute−value pairs. The attribute is a descriptor and the value the numerical interval into which the descriptor value falls. A subset of all attribute-value pairs represents a pattern. The relative frequency of a pattern p in a learning set D is defined as the support of p in D, i.e., $\text{supp}_D(p)$:

$$\text{supp}_D(p) = \frac{\text{count}_D(p)}{|D|}$$

Here, $\text{count}_D(p)$ reports the number of instances of p in D. A pattern with statistically significant support for positive compared to negative training examples is called an emerging pattern (EP).[1,2] The ratio of support rates of an EP in positive (D_1) and negative (D_2) training data represents its $\text{growth}_{D_1,D_2}(p)$:

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

If the support is >0 in D_1 but 0 in D_2, the EP qualifies as a jumping emerging pattern (JEP)[3] the growth of which remains undefined. A JEP is classified as a most expressive JEP if none of its descriptor subsets is a JEP and if no superset has larger support.[3] Most expressive JEPs from molecular descriptors of training set compounds have been defined as emerging chemical patterns (ECP).[15] For compound classification, descriptor values are calculated, patterns are derived, and matching ECPs from negative and positive training set are identified. For mining of ECPs, the computationally most demanding step, a hypergraph-based algorithm is applied.[4,7] A test set compound is assigned to the class for which matching ECPs yield the largest cumulative support (normalized to the value range [0,1]).

Descriptors and Compound Data Sets. A previously reported set[11] of 62 molecular graph-based numerical descriptors implemented in the molecular operating environment (MOE)[19] was used for ECP analysis. The descriptors are reported in detail in Table S1 of the Supporting Information. This descriptor set is characterized by low pairwise correlation and high information content.[20] For discretization of numerical descriptors, an entropy-based discretization method was applied, which employs an attribute splitting criterion to divide value ranges into defined intervals.[17,18] If all values of a descriptor calculated for the data set compounds mapped to a single interval it was eliminated because it did not capture compound-specific information.

From ChEMBL[21] (release 15), 15 compound activity classes were assembled for which equilibrium constants (K_i) below 1 μM for a wide range of human targets at the highest database confidence level (ChEMBL confidence score 9)[21] were available. Only compounds were selected that had at least five structural neighbors in the data set on the basis of similarity calculations using the extended connectivity fingerprint with bond diameter 4 (ECFP4)[22] and applying a Tanimoto coefficient (Tc)[23] threshold value of 0.3, which indicates at least remote structural similarity.[24] These activity classes consisted of 123−1701 compounds, as summarized in Table 1. All data sets are made available without restrictions via the following URL: http://dx.doi.org/10.5281/zenodo.8626.

Table 1. Compound Data Sets[a]

ChEMBL target ID	target	no. cpds (NBRS ≥ 5)	low disc.	intermediate disc.	high disc.
204	thrombin	750	357	216	177
218	cannabinoid CB1 receptor	1452	672	432	348
237	κ opioid receptor	1350	619	445	286
245	muscarinic acetylcholine receptor M3	398	142	127	129
255	adenosine A2b receptor	872	400	288	184
256	adenosine A3 receptor	1701	626	676	399
1821	muscarinic acetylcholine receptor M4	123	28	49	46
1889	vasopressin V1a receptor	397	164	145	88
2001	purinergic receptor P2Y12	509	276	126	107
3155	serotonin 7 (5-HT7) receptor	342	137	125	80
3594	carbonic anhydrase IX	953	429	309	215
4308	bradykinin B1 receptor	474	190	180	104
4608	melanocortin receptor 5	249	133	63	53
4644	melanocortin receptor 3	355	196	83	76
5071	G protein-coupled receptor 44	448	210	132	106

[a] For each data set, the ChEMBL target ID, target name, the number of compounds with at least five structural neighbors (NBRS), and the number of compounds with low, intermediate, or high discontinuity are reported.

SAR Discontinuity. SAR discontinuity was quantified for individual compounds by utilizing the discontinuity score component of the SAR Index[25] for the calculation of per-compound discontinuity scores.[26] The (non-normalized) discontinuity score is defined as

$$\text{raw}_{\text{disc}}(i) = \frac{\sum_{\{j|\text{sim}(i,j)>0.3, i\neq j\}} \text{potdiff}(i,j) \times \text{sim}(i,j)}{|\{j|\text{sim}(i,j) > 0.3,\ i \neq j\}|}$$

Here, $\text{potdiff}(i, j)$ is the absolute potency difference between compounds i and j and $\text{sim}(i,j)$ the calculated fingerprint

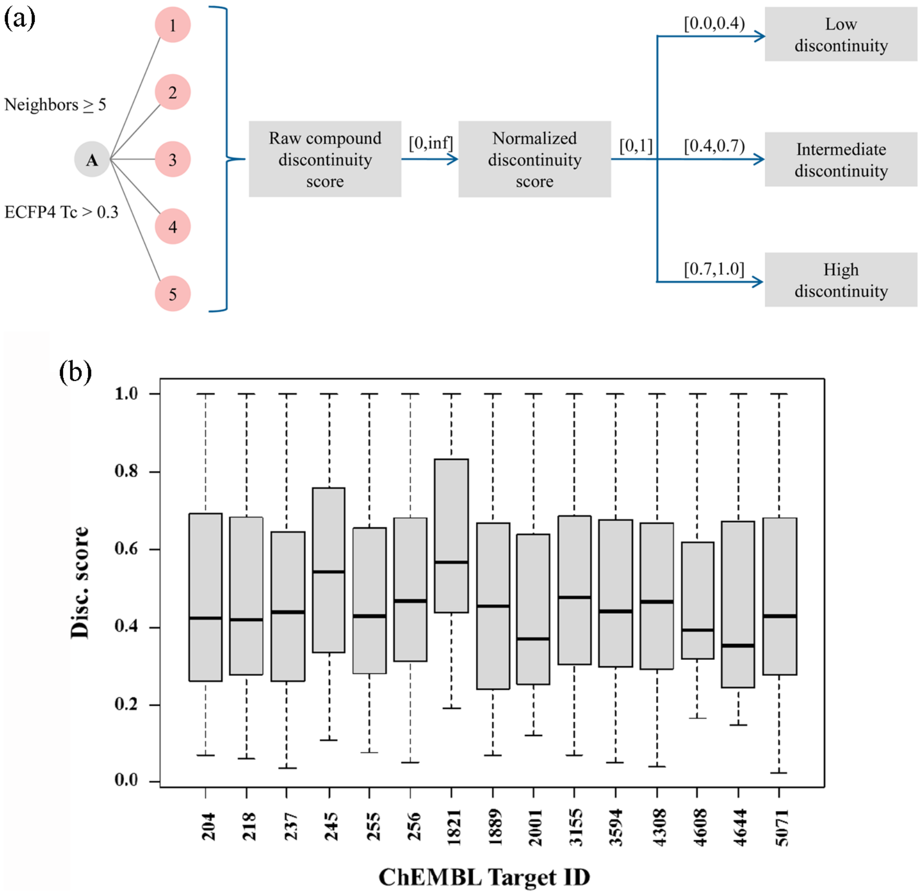


Figure 1. SAR discontinuity-based compound categorization. (a) Outline of the compound categorization approach is provided, as described in the text. (b) Box plot representations of per-compound discontinuity score distributions are reported for all data set, which rationalize the choice of score intervals for low, intermediate, and high discontinuity.

similarity for compounds *i* and *j*. Hence, the discontinuity score is calculated as the average of potency differences of all compound pairs in a data set multiplied by their similarity. As a discontinuity score similarity threshold, an ECFP4 Tc of 0.3 was applied.[24] Raw scores were converted into Z-scores using the sample mean and standard deviation of the score distribution of each individual data set[25] and cumulative probabilities assuming a normal distribution were calculated to map scores onto the value range [0, 1]. Hence, per-compound discontinuity scores of 0 and 1 indicate minimal and maximal SAR discontinuity, respectively. Our goal has been to predict into which local SAR regions of given data sets test compounds fall on the basis of per-compound discontinuity score distributions. Accordingly, normalized scores are preferred over raw scores.

On the basis of normalized discontinuity scores (DS), data set compounds were assigned to three different categories (as further discussed below): low discontinuity, DS < 0.4; intermediate discontinuity, 0.4 ≤ DS < 0.7; high discontinuity, DS ≥ 0.7. For each data set, the distribution of compounds over these three different SAR discontinuity categories is reported in Table 1. Alternatively, thresholds for these categories might also be assigned based on global SAR discontinuity score distributions of individual data sets. However, such assignments would be less general.

ECP Calculations. For each classification trial, a set of 3, 5, or 10 reference compounds was randomly selected from each discontinuity category to represent positive training examples and corresponding numbers of compounds were added from other categories as negative training examples. For example, a training set with 10 positive examples from the high discontinuity category also included 10 negative examples from the intermediate and 10 negative examples from the low discontinuity category (i.e., a total of 30 compounds). All remaining data set compounds represented the test set. ECP were derived from positive and negative training examples, respectively, and each test compound was assigned to the SAR category for which matching ECP produced the largest cumulative support. In each case, 100 different trials with randomly assembled training and test sets were carried out to obtain statistically meaningful results.

For each discontinuity category and data set, the performance was assessed by calculating the sensitivity (i.e., true positives/(true positives + false negatives)), specificity (i.e., true negatives/(true negatives + false positives)), and balanced accuracy (i.e., (sensitivity + specificity)/2) averaged over all 100 trials. Furthermore, we have assessed the performance of the overall classification using Matthew's correlation coefficient (MCC) for multiple categories[27] and Cohen's κ coefficient.[28]

Reference Calculations. As control calculations, random forest[29] (RF) and support vector machine[14] (SVM) models were derived for each training set containing 10 positive training examples (smaller training sets are usually not suitable for SVM classification). The same set of 62 numerical descriptors used for ECP was also used for RF and SVM modeling. The value range of each descriptor was normalized to values between 0 and 1. All classifiers were built using the freely available Python implementation scikit-learn[30] using standard parameters to ensure reproducibility. For RF models, 10 trees were used and the splits were determined using the Gini index. For the different SVM models, a Gaussian (radial basis function) kernel[31,32] was used with $\gamma = 1/n$ features, i.e., 0.0161, and uniform class weights. To enable SVM multiclass classification, different strategies were applied including one-vs-all and one-vs-one. In both cases, three classifiers were trained using the training examples of each class as positive instances. In the one-vs-all case, negative instances were provided by the training examples of the other two classes. Test compounds were assigned to the class corresponding to the model that gave the maximum decision value, i.e., the largest distance from the margin. In the one-vs-one cases (occasionally also referred to as all-vs-all), each classifier used only the examples of one class as negative instances. Instead of classifying new compounds using the largest distance from the margin, the outcome of each classifier was used as a vote and the label with the majority votes was selected.

RESULTS AND DISCUSSION

ECP Classification Scheme. The major goal of our analysis has been the prediction of compounds with different local SAR characteristics using ECP. For example, a compound introduces strong SAR discontinuity if it has a potency (high or low) that significantly differs from its structural neighbors. By contrast, local continuity is introduced if a compound has only small potency differences compared to structural analogues. Such local SAR behaviors can be consistently accounted for by calculating normalized per-compound discontinuity scores. Predicting compounds to belong to different SAR categories, defined by discontinuity score intervals over the entire range, generalizes predictions of per-compound SAR features. The formation of local SAR environments depends on the presence of structurally related compounds and their potency relationships. Hence, compounds with no structural neighbors (singletons) cannot participate in local SARs. Therefore, we have assembled a data set in which all active compounds had structural relationships to others, which resulted in the presence of local SAR environments for all compounds (albeit very different ones) and enabled consistent predictions. Table 1 shows that compounds in all data sets were fairly evenly

Table 2. Qualifying Descriptors[a]

ChEMBL target ID	discretized descriptors
204	42
218	35
237	42
245	36
255	26
256	44
1821	17
1889	44
2001	36
3155	12
3594	18
4308	50
4608	45
4644	53
5071	22

[a] For each data set, the number of descriptors qualifying for ECP analysis after information entropy-based discretization is reported.

Table 3. Emerging Chemical Patterns Statistics[a]

(a) positive training examples =3

ChEMBL target ID	low discontinuity	intermediate discontinuity	high discontinuity
204	1067	687	828
218	489	408	425
237	526	499	472
245	580	484	641
255	110	99	98
256	1267	1135	1179
1821	46	52	44
1889	530	599	517
2001	148	119	152
3155	21	22	21
3594	21	20	18
4308	906	654	650
4608	48	45	41
4644	1150	629	744
5071	104	112	99

(b) positive training examples =5

ChEMBL target ID	low discontinuity	intermediate discontinuity	high discontinuity
204	3496	2038	2633
218	1425	1176	1217
237	1370	1346	1320
245	1735	1417	1866
255	233	220	186
256	4095	3565	4043
1821	73	80	68
1889	1386	1566	1192
2001	310	210	308
3155	31	35	29
3594	34	33	30
4308	2470	1898	1889
4608	61	67	62
4644	2942	1714	1942
5071	218	213	205

(c) positive training examples =10

ChEMBL target ID	low discontinuity	intermediate discontinuity	high discontinuity
204	8596	15 063	10 962
218	4287	5426	4691
237	4943	5210	4785
245	4756	6365	6628
255	485	555	456
256	17 672	20 405	19 297
1821	135	135	111
1889	4345	5012	3530
2001	587	840	881
3155	56	47	43
3594	67	65	57
4308	6549	8868	6612
4608	79	84	108
4644	6305	10 008	7080
5071	520	495	474

[a] For each data set, the average number of ECP identified in 100 individual trials with randomly selected training sets of (a) 3, (b) 5, or (c) 10 positive training examples are reported for compounds with low, intermediate, or high discontinuity.

Table 4. Exemplary ECP

ChEMBL target ID	discontinuity category	emerging chemical patterns	support
204	low	{SlogP_VSA0:(-inf-45.782856], SlogP_VSA2:(45.710888−58.261097]}	0.9
	intermediate	{SlogP_VSA2:(58.261097−59.430298], SMR_VSA1:(41.477263-inf), VDistEq:(3.630869-inf)}	0.9
	high	{PEOE_VSA+3:(-inf-19.70485], PEOE_VSA+6:(-inf-27.751621], PEOE_VSA_FNEG:(0.174786-inf), SlogP_VSA7:(137.78144−140.07341], TPSA:(-inf-50.004999]}	0.9
255	low	{a_nN:(-inf-6.5], SlogP_VSA2:(-inf-16.724269], SlogP_VSA3:(90.006027-inf), SMR_VSA4:(35.43611-inf)}	0.9
	intermediate	{a_ICM:(1.508357−1.591618], a_nN:(-inf-6.5], PEOE_VSA+5:(24.140093-inf), PEOE_VSA-3:(-inf-7.551808], SlogP_VSA4:(-inf-3.798186]}	0.6
	high	{a_ICM:(1.508357−1.591618], b_1rotN:(-inf-8.5], PEOE_VSA+5:(24.140093-inf), PEOE_VSA-6:(-inf-6.706814], SlogP_VSA4:(-inf-3.798186]}	0.8
1821	low	{PEOE_VSA+6:(-inf-95.614003], SlogP_VSA1:(2.407301-inf), SlogP_VSA5:(-inf-34.140528]}	0.8
	intermediate	{a_ICM:(1.337654-inf), PEOE_RPC-:(0.637844−0.903824], SlogP_VSA1:(2.407301-inf), SlogP_VSA5:(34.140528−86.99664]}	0.8
	high	{a_acc:(1.5-inf), b_1rotR:(-inf-0.080128]}	1.0
3155	low	{a_acc:(-inf-1.5], PEOE_VSA-4:(-inf-30.758184]}	0.9
	intermediate	{PEOE_VSA+6:(64.348719-inf), SMR_VSA2:(15.3383-inf), TPSA:(-inf-49.459999], VDistEq:(-inf-3.97467]}	0.7
	high	{PEOE_VSA+2:(20.274868-inf), PEOE_VSA+6:(64.348719-inf), TPSA:(49.459999-inf), VDistEq:(-inf-3.97467]}	0.7
4644	low	{PEOE_VSA-0:(-inf-122.98079], PEOE_VSA-4:(32.898536-inf), SlogP_VSA5:(55.428746-inf)}	1.0
	intermediate	{SlogP_VSA1:(14.769711−33.492092], SlogP_VSA3:(114.36749−170.929565], SMR_VSA1:(21.171903−59.875202], TPSA:(170.375-inf)}	0.9
	high	{PEOE_VSA_FHYD:(-inf-0.605321], SlogP_VSA9:(-inf-82.576431], SMR_VSA0:(97.410466-inf)}	0.9

For a subset of compound classes, exemplary ECP with strong support for an individual reference set with 10 positive training examples are provided for the low, intermediate, and high discontinuity category. Descriptors are abbreviated according to Table S1 of the Supporting Information; "inf" stands for infinity.

distributed over the three different SAR discontinuity categories representing low, intermediate, and high SAR discontinuity. As a general trend, the number of compounds per category often decreased from low over intermediate to high SAR discontinuity, as would be expected (structurally similar compounds are on average more likely to have comparable rather than significantly different potency).

Figure 1a outlines the SAR discontinuity-based compound categorization approach. For all data set compounds, discontinuity scores were calculated on the basis of all structurally related compounds and their potency values. Then, the scores were normalized. On the basis of normalized scores, compounds were assigned to different SAR categories from which training and test sets were assembled for model building and evaluation. Figure 1b reports the per-compound discontinuity score distribution for all data sets. The box plot representations illustrate that most median values fell within the score interval [0.4, 0.6] and that interquartile distances typically spanned ~40% of the scoring range. These observations rationalized our assignment of score intervals corresponding to low (DS < 0.4), intermediate (0.4 ≤ DS < 0.7), and high (DS ≥ 0.7) SAR discontinuity.

Descriptors and Patterns. Table 2 reports the number of descriptors that were discretized for the activity classes. Depending on the class, the number of qualifying descriptors ranged from only 12 (of 62) to 53, hence reflecting significant class-specific differences in descriptor value distributions. Such differences were further amplified by the numbers of ECP obtained for classification, as reported in Table 3. ECP numbers overall roughly correlated with the number of successfully discretized descriptors. Furthermore, ECP numbers generally depend on training set sizes and there typically was a significant increase in the number of ECP for training sets of increasing size. For example, for thrombin inhibitors (target ID 204), the number of ECP increased from ~700−1000 (3 positive training examples) over ~2000−3500 (5) to ~8500−15 000 (10 examples). In part, extreme differences between ECP numbers per compound class were observed. For instance, in the intermediate discontinuity class (for 10 positive training examples), the ECP numbers ranged from only 47 for serotonin 7 receptor antagonists (target ID 3155) to 20 405 for adenosine A3 receptor antagonists (target ID 256). By contrast, for a given training set size, the number of patterns for a given compound class usually was comparable in magnitude for the different discontinuity categories. Table 4 reports exemplary ECP with strong support for different compound classes and discontinuity categories and illustrates the highly variable composition and the different degrees of complexity such patterns might display. Different value ranges of descriptors such as surface area and partial charge descriptors and/or their combinations were frequently detected as signature patterns with strong support to classify test compounds with respect to different discontinuity categories. In the case of adenosine A2b receptor antagonists (target ID 255), such signatures occurred in combination with characteristic atom or bond counts that differentiated compounds in intermediate and high discontinuity categories.

ECP-Based Predictions. Table 5 reports the results of SAR discontinuity category predictions over 100 individual trials for test compounds from all data sets. Sensitivity and specificity of ECP classification calculations were determined on a per-category basis. Despite the different composition of the compound data sets under study and the in part significantly different numbers of ECP they produced, compounds were often correctly predicted at comparable rates, on average ~50−60% of the compounds in different SAR discontinuity categories. The predictions were rather consistent across different data sets and there were no true outliers, despite compound class-specific differences. Notably, only small differences in prediction accuracy were generally observed for training sets of different sizes. Even the smallest training sets containing only 3 positive training examples (which is unusually small for machine learning) yielded sensitivity and specificity values that were similar or, on average, only a few percent lower

Table 5. Prediction Accuracy for Compounds in Different Discontinuity Categories[a]

		sensitivity			specificity		
ChEMBL target ID	training set size	low disc.	intermediate disc.	high disc.	low disc.	intermediate disc.	high disc.
204	3	0.50	0.40	0.53	0.80	0.67	0.76
	5	0.55	0.37	0.55	0.78	0.70	0.77
	10	0.61	0.39	0.59	0.81	0.72	0.79
218	3	0.38	0.38	0.36	0.71	0.66	0.69
	5	0.41	0.41	0.36	0.72	0.66	0.72
	10	0.45	0.40	0.39	0.72	0.69	0.73
237	3	0.49	0.35	0.36	0.69	0.71	0.72
	5	0.52	0.35	0.36	0.68	0.72	0.73
	10	0.56	0.37	0.37	0.69	0.72	0.76
245	3	0.53	0.43	0.50	0.80	0.71	0.72
	5	0.56	0.45	0.51	0.82	0.71	0.73
	10	0.62	0.45	0.56	0.82	0.74	0.76
255	3	0.43	0.33	0.41	0.69	0.70	0.70
	5	0.46	0.34	0.43	0.70	0.71	0.72
	10	0.51	0.36	0.46	0.71	0.71	0.75
256	3	0.51	0.36	0.40	0.72	0.72	0.70
	5	0.57	0.34	0.42	0.71	0.74	0.71
	10	0.60	0.36	0.43	0.72	0.74	0.73
1821	3	0.48	0.52	0.69	0.75	0.77	0.84
	5	0.53	0.56	0.74	0.77	0.81	0.87
	10	0.61	0.60	0.79	0.78	0.85	0.90
1889	3	0.71	0.41	0.42	0.76	0.75	0.79
	5	0.72	0.41	0.47	0.77	0.76	0.80
	10	0.75	0.41	0.51	0.79	0.77	0.80
2001	3	0.57	0.32	0.41	0.66	0.74	0.78
	5	0.59	0.36	0.44	0.70	0.74	0.79
	10	0.63	0.41	0.52	0.74	0.77	0.80
3155	3	0.45	0.37	0.47	0.74	0.69	0.71
	5	0.46	0.39	0.51	0.75	0.70	0.72
	10	0.51	0.46	0.59	0.78	0.73	0.76
3594	3	0.46	0.41	0.34	0.67	0.67	0.77
	5	0.50	0.43	0.38	0.70	0.69	0.77
	10	0.57	0.48	0.47	0.75	0.74	0.78
4308	3	0.54	0.39	0.38	0.74	0.70	0.73
	5	0.57	0.40	0.40	0.75	0.71	0.75
	10	0.62	0.40	0.44	0.76	0.73	0.76
4608	3	0.55	0.56	0.69	0.80	0.75	0.82
	5	0.59	0.63	0.75	0.83	0.78	0.84
	10	0.74	0.73	0.84	0.89	0.88	0.87
4644	3	0.81	0.35	0.60	0.80	0.82	0.86
	5	0.81	0.37	0.63	0.81	0.83	0.86
	10	0.82	0.41	0.66	0.82	0.84	0.87
5071	3	0.51	0.43	0.41	0.77	0.70	0.73
	5	0.53	0.45	0.44	0.78	0.72	0.73
	10	0.56	0.49	0.48	0.79	0.75	0.75

[a] Average sensitivity and specificity over 100 individual trials with randomly selected training sets of 3, 5, or 10 positive training examples are reported for compounds with low, intermediate, or high discontinuity.

than those obtained for larger training sets. Overall, best results were obtained for training sets containing 10 positive examples, although the differences were mostly small. The results obtained for training sets with 10 positive examples are reported in Figure 2 and discussed in the following. Results for smaller training sets with 3 and 5 positive examples were similar and are reported in Figures S1 and S2 of the Supporting Information, respectively. As shown in Figure 2a, sensitivity values between ~0.4–0.5 were most frequently obtained, with six individual predictions reaching values >0.7. Standard deviations >0.2 were observed in most cases, hence reflecting an in part strong influence of training set size and composition on the results (which might be expected for relatively small training sets). For 12 of 15 classes, the highest sensitivity of the calculations was detected for compounds with low discontinuity, followed by compounds with high discontinuity (although overall fewest compounds with high discontinuity were available). Differences in sensitivity between individual categories of >20% were observed for the majority of activity classes. With ~50–60% of correctly predicted compounds falling into different SAR discontinuity categories across variety of activity classes, ECP calculations reached reasonable

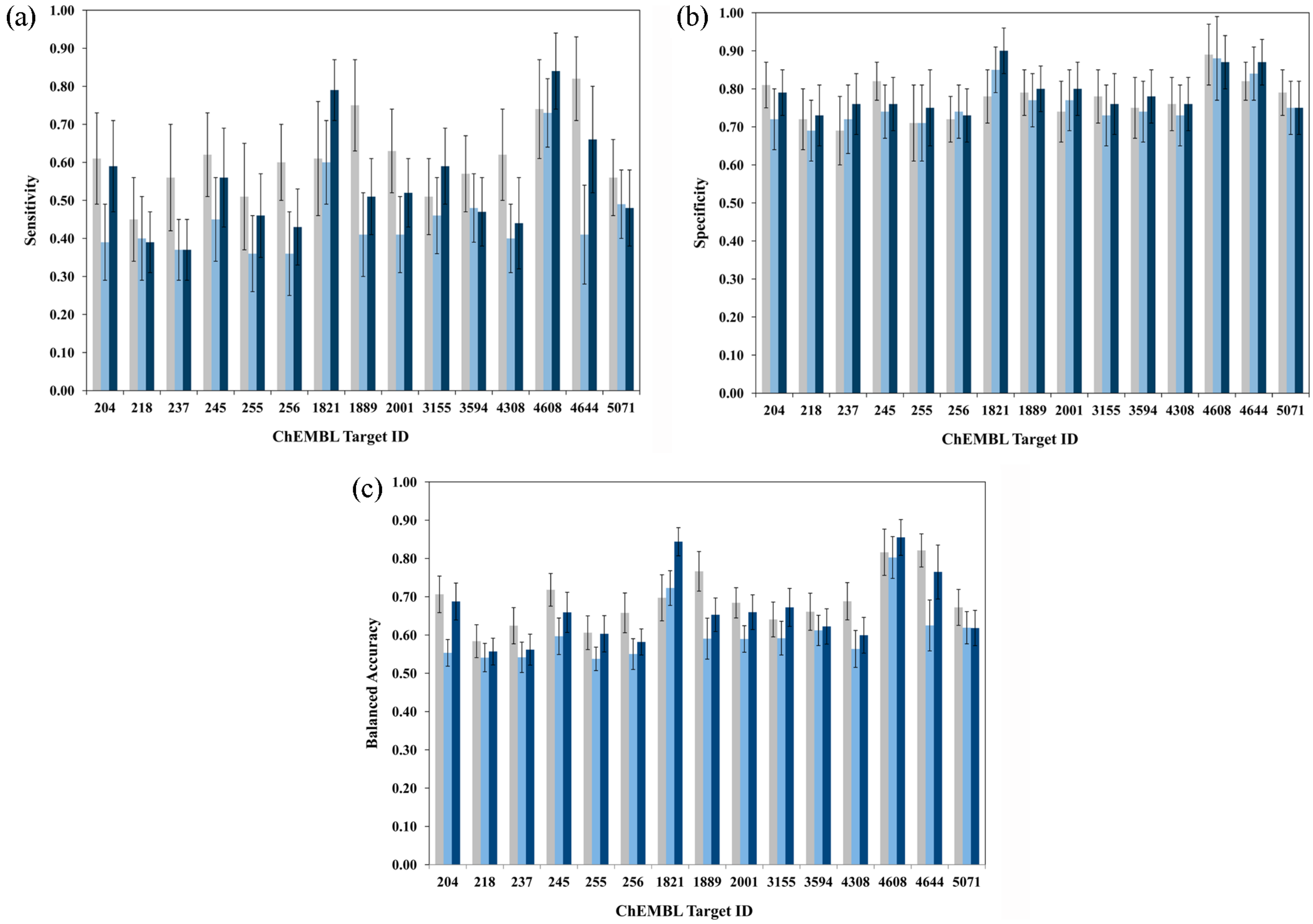


Figure 2. Prediction accuracy. Average (a) sensitivity, (b) specificity, and (c) balanced accuracy are reported for compounds with low (gray), intermediate (light blue), and high SAR discontinuity (dark blue). Results are reported for training sets with 10 positive training examples. Vertical lines report standard deviations. For compound data sets, target IDs are provided according to Table 1.

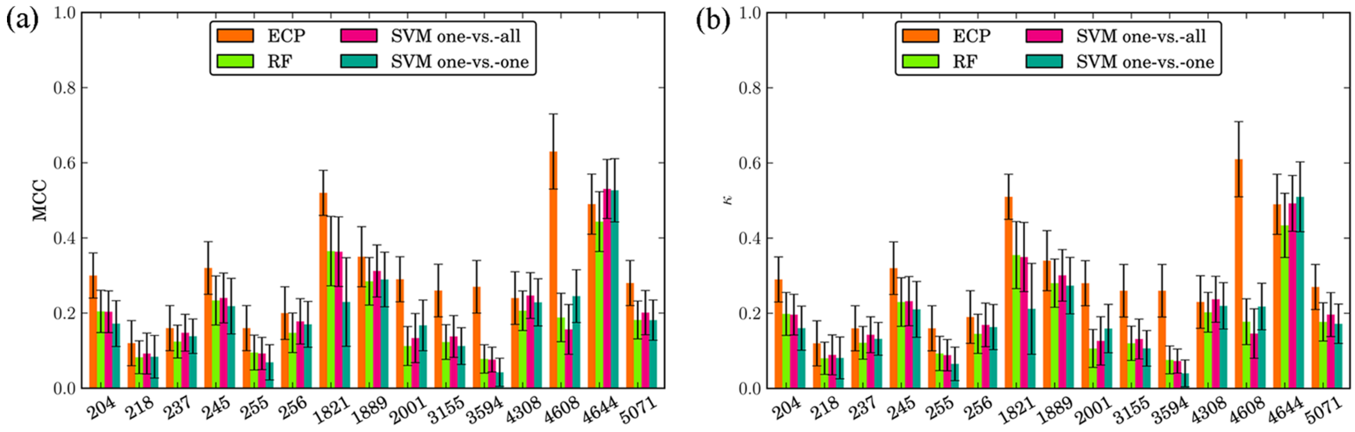


Figure 3. Alternative predictions. Average (a) MCC and (b) κ values compared for ECP, RF, and SVM calculations using training sets with 10 positive training examples. Vertical lines report standard deviations.

sensitivity, which nonetheless leaves some room for further potential improvements. By contrast, the specificity of the calculations was generally high, as shown in Figure 2b. Specificity values of ~0.7−0.9 were consistently observed for different SAR discontinuity categories and data sets. In addition, the differences between categories were small (and standard deviations were lower than for sensitivity calculations). The observed high specificity of ECP calculations meant that false positive rates were generally low. This represents an important aspect for prediction of compounds in local SAR environments, because the characteristics of these local environments are particularly vulnerable to false positive assignments. These results are also mirrored by the calculation of balanced accuracy yielding values of ~0.5−0.8, as reported in Figure 2c. Often, the addition of only one compound with different potency relationships to its neighbors might be

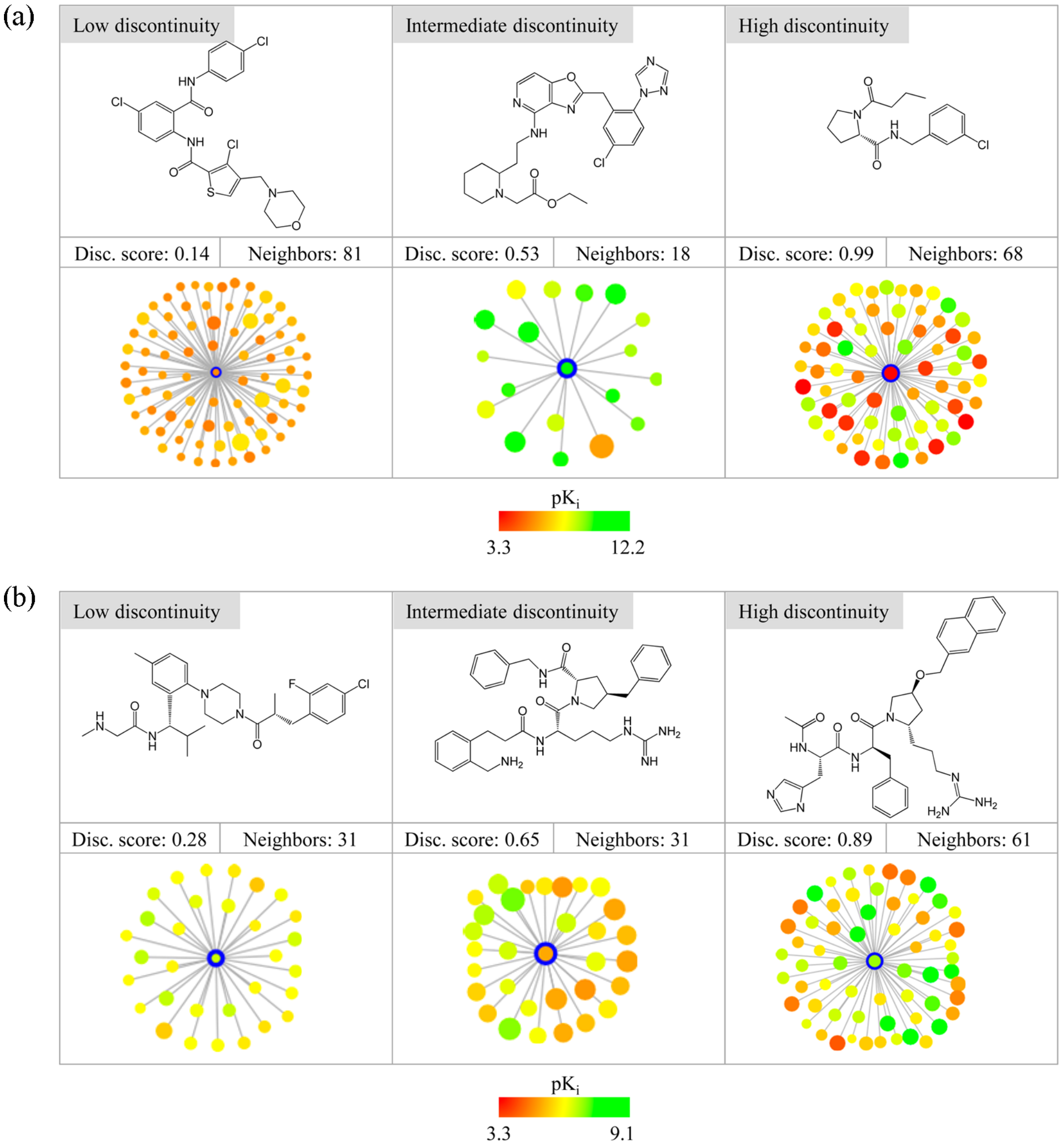


Figure 4. Exemplary compounds. At the top, examples of correctly predicted (a) thrombin inhibitors and (b) melanocortin receptor 3 antagonists with low, intermediate, or high discontinuity are shown. At the bottom, local SAR networks[25] are shown for these compounds and their structural neighbors in which compounds are represented as nodes and similarity relationships as edges. Nodes are color-coded according to compound potency using a continuous color spectrum from red (lowest potency in the data set) over yellow (intermediate) to green (highest compound potency). In addition, nodes are scaled in size according to per-compound discontinuity scores.[26] The exemplary compounds shown at the top are placed in the center of each graph and are encircled.

sufficient to substantially change a local SAR (e.g., by introducing a shift toward increasing discontinuity), consistent with observed standard deviations for sensitivity calculations. Furthermore, it is important to note that compounds in both continuous and discontinuous SAR environments were predicted with high specificity, indicating that ECP implicitly captured very different SAR characteristics of individual compounds.

Control Calculations. ECP is a pattern matching approach that assigns a test compound to one of multiple classes on the basis of cumulative pattern support. This sets the ECP approach conceptually apart from other multiclass machine learning methods. To put ECP performance into perspective, we have also generated different RF and SVM classifiers to predict compounds falling into each discontinuity category. Average balanced accuracy values for RF and SVM calculations are reported in Figure S3 of the Supporting Information. The individually derived classifiers displayed reasonable to high specificity and sensitivity for the discontinuity categories they were trained on. As robust classification metrics, we have also calculated assessed MCC and κ values to assess and compare the overall performance of multiclass predictions using ECP, RF, and SVM models. The results of MCC and κ calculations are reported in are reported in Figure 3a,b, respectively.

With 10 compounds as positive training sets, ECP calculations over all classes in the data set produced average MCC and κ values of 0.31 and 0.30, respectively. The RF classifiers yielded average values of 0.19 for both these measures. Average balanced accuracy values for the low, intermediate, and high discontinuity categories were 0.55, 0.62, and 0.61. The one-vs-all SVM strategy resulted in average MCC and κ values of 0.21 and 0.20, and the one-vs-one SVM strategy produced average MCC and κ values of 0.19 and 0.18 over all classes. The balanced accuracy for the different categories was 0.54, 0.64, and 0.61 for one-vs-all and 0.55, 0.63, and 0.59 for one-vs-one classification. MCC and κ values produced by the different RF and SVM classifiers were compared to those obtained with ECP enabling an at least approximate comparison for the assignment of compounds to individual SAR discontinuity categories (Figure 3). In most cases, ECP produced higher κ values than the other three classifiers. For target sets 1821, 2001, 3155, 3594, and 4608, ECP performed significantly better than RF and SVM classifiers. Specifically, for smaller target sets 1821 and 4608 with fewer than 250 compounds, values greater than 0.5 were obtained. In one case, target set 4644, SVM classifiers reached higher performance than ECP.

Overall, ECP reached comparable or better performance in SAR discontinuity-based multiclass predictions than RF and SVM classification, which represent state-of-the-art machine learning approaches in the chemoinformatics field. Multiclass (rather than binary) predictions typically represent a complicated task. From this point of view, the results obtained for predictions of compounds in different categories spanning the entire range of SAR discontinuity in data sets were encouraging.

Exemplary Compounds. In Figure 4, exemplary compounds belonging to different SAR discontinuity categories are shown that were correctly predicted using ECP calculations. Figure 4a shows thrombin inhibitors and Figure 4b melanocortin receptor 3 antagonists. The compounds are displayed together with network-like depictions of their local SAR environments, which illustrate differences between potency distributions of structural neighbors and the different degrees of local SAR discontinuity the test compounds introduce. Consistently accurate predictions were obtained for compound neighborhoods of different composition and SAR information content.

■ CONCLUSIONS

In this study, the ability of the ECP methodology to predict compounds in different local SAR environments was investigated, a previously unconsidered task. Local SAR environments in large compound data sets are often studied on the basis of activity landscape modeling, which represents a descriptive approach. However, it has thus far not been attempted to further extend descriptive to predictive modeling of compounds in different SAR environments. For our analysis, compounds in a variety of activity classes were organized into different SAR discontinuity categories on the basis of per-compound discontinuity scores spanning the entire spectrum of local SARs. This organization scheme has provided a consistent reference frame for machine learning and compound property predictions. Using ECP, signature descriptor patterns were derived for training compounds from different discontinuity categories and used for predictions. ECP was found to accurately predict many compounds in different local SAR environments and performed comparably to or better than RF models and different multiclass SVM strategies. The approach presented herein provides a basis, for example, for the prediction of compounds that complement local SAR environments, i.e., for adding candidate compounds to existing series, or the prioritization of compounds with different SAR characteristics. It also represents another application of the ECP methodology that utilizes one of its strengths, i.e., the assignment of compounds to multiple feature categories (multiclass learning), in this case, focusing on different SAR environments. Another known strength of ECP is its ability to operate on the basis of very small training sets. This enables the application of the methodology in lead optimization projects at early stages when only limited compound information is available. ECP calculations can then be applied to predict compounds that complement local SAR environments in data sets that are only sparsely populated. If predictions are carried out for regions of high local SAR discontinuity, as reported herein, one might obtain highly active compounds to further progress optimization efforts.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Table S1 details the set of descriptor used for ECP and control calculations. Figures S1 and S2 report the results of ECP calculations with three and five training instances, respectively. Figure S3 reports the results of RF and multiclass SVM control calculations. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*J. Bajorath. 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) 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.

(2) 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.

(3) Li, J.; Dong, G.; Ramamohanarao, K. Making Use of the Most Expressive Jumping Emerging Patterns for Classification. *Knowl. Inf. Syst.* **2001**, *3*, 131−145.

(4) 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.

(5) Li, J.; Dong, G.; Ramamohanarao, K.; Wong, L. DeEPs: a New Instance-Based Lazy Discovery and Classification System. *Mach. Learn.* **2004**, *54*, 99−124.

(6) Wang, L.; Zhao, H.; Dong, G.; Li, J. On the Complexity of Finding Emerging Patterns. *Theor. Comput. Sci.* **2005**, *335*, 15−27.

(7) Auer, J.; Bajorath, J. Emerging Chemical Patterns: A New Methodology for Molecular Classification and Compound Selection. *J. Chem. Inf. Model.* **2006**, *46*, 2502−2514.

(8) Auer, J.; Bajorath, J. Simulation of Sequential Screening Experiments Using Emerging Chemical Patterns. *Med. Chem.* **2008**, *4*, 80−90.

(9) Auer, J.; Bajorath, J. Distinguishing between Bioactive and Modeled Compound Conformations through Mining of Emerging Chemical Patterns. *J. Chem. Inf. Model.* **2008**, *48*, 1747−1753.

(10) Sherhod, R.; Gillet, V. J.; Judson, P. N.; Vessey, J. D. Automating Knowledge Discovery for Toxicity Prediction Using

Jumping Emerging Pattern Mining. *J. Chem. Inf. Model.* **2012**, *52*, 3074−3087.

(11) Namasivayam, V.; Hu, Y.; Balfer, J.; Bajorath, J. Classification of Compounds with Distinct or Overlapping Multi-target Activities and Diverse Molecular Mechanisms using Emerging Chemical Patterns. *J. Chem. Inf. Model.* **2013**, *53*, 1272−1281.

(12) Namasivayam, V.; Iyer, P.; Bajorath, J. Prediction of Individual Compounds Forming Activity Cliffs Using Emerging Chemical Patterns. *J. Chem. Inf. Model.* **2013**, *53*, 3131−3139.

(13) Duda, R. O.; Hart, P. E.; Stork, D. G. Pattern Classification, 2nd ed.; Wiley-Interscience, New York, 2000; pp 20−83.

(14) Vapnik, V. N. *The Nature of Statistical Learning Theory*, 2nd ed.; Springer: New York, 2000.

(15) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.

(16) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity Landscape Representations for Structure-Activity Relationship Analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

(17) 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., Ed.; Morgan Kaufmann Publishers: San Francisco, CA, 1993; pp 1022−1027.

(18) 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.

(19) *Molecular Operating Environment (MOE)*, 2013.08; Chemical Computing Group Inc., Montreal, Quebec, Canada, 2013.

(20) 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.

(21) 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.* **2012**, *40*, D1100−D1107.

(22) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

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

(24) Dimova, D.; Stumpfe, D.; Bajorath, J. Quantifying the Fingerprint Descriptor Dependence of Structure-Activity Relationship Information on a Large Scale. *J. Chem. Inf. Model.* **2013**, *53*, 2275−2281.

(25) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571−5578.

(26) Wawer, M.; Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Structure-Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure-Activity Relationship Indices. *J. Med. Chem.* **2008**, *51*, 6075−6084.

(27) Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient. *Comput. Biol. Chem.* **2004**, *28*, 367−374.

(28) Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37−46.

(29) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5−32.

(30) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(31) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−561.

(32) Ben-Hur, A.; Weston, J. A User's Guide to Support Vector Machines. *Methods Mol. Biol.* **2010**, *609*, 223−239.