Journal of
CHEMICAL INFORMATION
AND MODELING

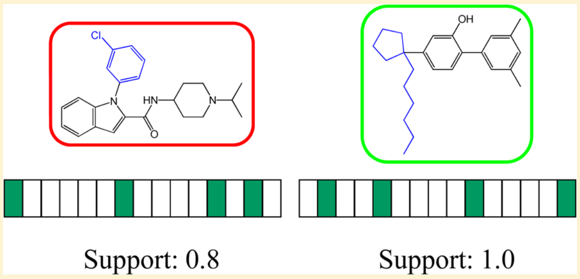
Prediction of Individual Compounds Forming Activity Cliffs Using Emerging Chemical Patterns

Vigneshwaran Namasivayam, Preeti Iyer, and Jürgen Bajorath*

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

S *Supporting Information*

ABSTRACT: Activity cliffs are formed by structurally similar or analogous compounds having large potency differences. In medicinal chemistry, pairs or groups of compounds forming activity cliffs are of interest for structure−activity relationship (SAR) analysis and compound optimization. Thus far, activity cliff assessment has mostly been descriptive, i.e., compound data sets and activity landscape representations have been searched for activity cliffs in the context of SAR analysis. Only recently, first attempts have also been made to depart from descriptive analysis and predict activity cliffs. This has been done by building computational models that distinguish compound pairs forming activity cliffs from non-cliff pairs. However, it is principally more challenging to predict single compounds that participate in activity cliffs. Here, we show that individual compounds having high or low potency can be accurately predicted to form activity cliffs on the basis of emerging chemical patterns.



INTRODUCTION

Activity cliffs are generally defined as pairs of active compounds having a large difference in potency,[1] and they represent the cardinal features of activity landscapes[2] of compound data sets, i.e., representations that integrate compound similarity and potency relationships. Activity landscape and cliff analysis has a priori been descriptive in nature,[2] i.e., landscape representations have been generated for large compound data sets and analyzed to, for example, identify compounds that reveal structure−activity relationship (SAR) information and SAR determinants. Recently, first attempts have been made to predict activity cliffs[3−5] and derive conditional probabilities of different activity landscape features.[6] In addition, compound activities have been predicted on the basis of activity landscape models.[7] Activity cliff predictions reported thus far have utilized different machine learning approaches. Using particle swarm optimization, data sets were searched for groups of compounds forming coordinated activity cliffs, i.e., higher-order activity cliff arrangements involving multiple cliffs.[3] Other studies have attempted to predict compound pairs forming activity cliffs. Specifically, using compound pair descriptor combinations, activity cliff scores were derived using random forests,[4] and in addition, support vector machines were applied to predict activity cliffs using specialized kernel functions.[5]

Although attempts have been made to predict activity cliffs, it has thus far not been possible to predict individual compounds that form activity cliffs, which is the topic of our current study. Here, the difficulty is that chemical characteristics need to be assigned to single compounds that account for their ability to form activity cliffs rather than build computational models to distinguish activity cliffs from non-cliff compound pairs. In order to capture such chemical characteristics of individual compounds, pattern recognition approaches might be considered. The emerging pattern (EP) approach was introduced in computer science to systematically generate class-specific feature patterns for objects.[8−13] It has subsequently been applied in bioinformatics to predict gene expression patterns[14] and adopted in chemoinformatics as *emerging chemical patterns* (ECP) for compound classification,[15] modeling of screening experiments,[16] and analysis of molecular conformations.[17] A hallmark of the ECP approach that sets it apart from many other machine learning methods utilized in chemoinformatics is its ability to operate on the basis of small training sets.[15,16] Recently, the approach has been applied to analyze compound features associated with toxicity[18] and classify compounds with multi-target activities.[19]

In this study, we have used ECP to predict individual compounds forming activity cliffs. The ECP approach has made it possible to identify characteristic patterns for compounds meeting structural and potency criteria to form activity cliffs and distinguish them from others.

METHODS AND MATERIALS

Emerging Chemical Patterns. For deriving ECP classifiers, chemical descriptors must be preselected, and their computed value ranges for compounds must be discretized into defined intervals.[20,21] On the basis of chosen descriptors, each compound yields a set of *attribute-value pairs*. The *attribute* is a descriptor, and the *value* is the numerical interval into which

Received: October 16, 2013

Table 1. Compound Data Sets[a]

			MACCS-based activity cliffs		MMP cliffs	
CHEMBL ID	target name	size	*low ActCliff*	*high ActCliff*	*low ActCliff*	*high ActCliff*
CHEMBL205	carbonic anhydrase II	1519	59	58	53	58
CHEMBL218	cannabinoid CB1 receptor	1532	53	67	42	29
CHEMBL228	serotonin transporter	1382	41	70	35	44
CHEMBL237	kappa opioid receptor	1392	77	91	31	25
CHEMBL244	coagulation factor X	1199	52	138	37	72
CHEMBL253	cannabinoid CB2 receptor	1576	66	87	44	52
CHEMBL256	adenosine A3 receptor	1896	165	278	41	84
CHEMBL259	melanocortin receptor 4	1289	136	175	30	25
CHEMBL261	carbonic anhydrase I	1458	147	65	110	40

[a] For each compound data set, the CHEMBL id, target name, total number of compounds, and number of *low/high ActCliff* for MACCS-based cliffs and MMP cliffs are reported.

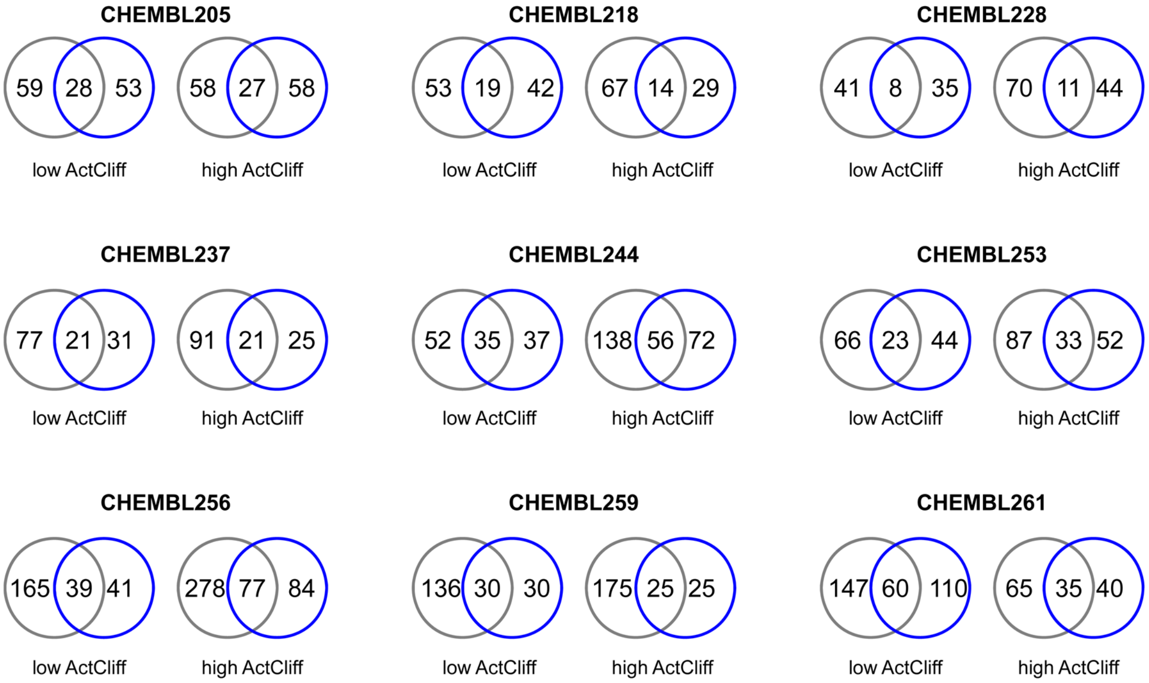


Figure 1. Compound distribution. For the nine compound data sets, Venn diagrams report the number of *high/low ActCliffs* for the two alternative activity cliff definitions (gray, MACCS-based cliffs; blue, MMP cliffs) and the number of conserved compounds.

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 the *support* of p in D, i.e. $\mathrm{supp}_D(p)$:

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

In this equation, $\mathrm{count}_D(p)$ is the number of instances of p in set D. A pattern with statistically significant support for positive relative to negative training examples is called an EP.[8,9] The ratio of support rates of an EP in positive (D_1) and negative (D_2) training data represents its $\mathrm{growth}_{D_1,D_2}(p)$

$$\mathrm{growth}_{D1,D2}(p) = \frac{\mathrm{supp}_{D1}(p)}{\mathrm{supp}_{D2}(p)}$$

If the support is greater than zero in D_1 but zero in D_2, the EP is classified as a *jumping emerging pattern* (JEP).[10] For JEPs, the growth remains undefined. A JEP is further classified as a *most expressive jumping emerging pattern* if none of its descriptor subsets is a JEP and if no superset has larger support.[10] *Most expressive JEPs* obtained for chemical descriptors of training set compounds have been defined as ECPs.[15]

ECP-Based Classification. For classification, descriptor values are calculated for test set compounds, and matching ECPs from negative and positive training set compounds are identified. For systematic mining of ECPs, a hypergraph-based algorithm is utilized.[11,15] A test set compound is then assigned to the class for which matching ECPs yield the largest cumulative support (normalized to the value range [0,1]).

Descriptors. A total of 62 numerical descriptors calculated from molecular graphs and implemented in the Molecular Operating Environment (MOE)[22] were used to generate ECPs. This descriptor set was selected because the descriptors displayed low pairwise correlation but had high information entropy in a large compound database.[23] In addition, the descriptor set was successfully applied in a previous ECP analysis of compounds with multi-target activities.[19] Hence, these descriptors were sensitive to ECP calculations. It is described in detail in Table S1 of the Supporting Information. Most numerical descriptors can adopt many different values or continuous value

Table 2. Qualifying Descriptors[a]

		no. of discretized descriptors	
CHEMBL ID	target name	MACCS	MMP
CHEMBL205	carbonic anhydrase II	37	26
CHEMBL218	cannabinoid CB1 receptor	22	23
CHEMBL228	serotonin transporter	17	42
CHEMBL237	kappa opioid receptor	27	29
CHEMBL244	coagulation factor X	45	36
CHEMBL253	cannabinoid CB2 receptor	26	7
CHEMBL256	adenosine A3 receptor	47	38
CHEMBL259	melanocortin receptor 4	50	6
CHEMBL261	carbonic anhydrase I	46	43

[a] For each data set, the number of descriptors qualifying for ECP analysis following discretization is reported for MACCS-based cliffs and MMP cliffs.

ranges. Thus, for pattern generation, descriptors were discretized with an entropy-based discretization method utilizing an attribute splitting criterion to divide value ranges into suitable intervals.[20,21] If values of a descriptor mapped to a single interval, it was eliminated due to lack of compound-specific information.

Activity Cliff Criteria. The assessment of activity cliffs requires specific consideration of a similarity criterion and potency difference criterion.[1] For our analysis, two alternative similarity criteria were applied to define activity cliffs including Tanimoto similarity[24] calculated on the basis of MACCS structural keys[25] and the formation of a transformation size-restricted matched molecular pair (MMP),[26,27] leading to the definition of MMP cliffs.[27] An MMP is defined as a pair of compounds that are only distinguished by a structural change at a single site, which corresponds to the exchange of a pair of substructures, a so-called chemical transformation.[26] The first similarity criterion relies on calculated whole-molecule similarity, and the second is substructure-based. For the latter, transformation size restrictions were introduced such that compound pairs forming MMP cliffs were confined to structural analogs.[27] Specifically, the size of an exchanged substructure was limited to 13 non-hydrogen atoms and the size difference between exchanged substructures to maximally eight non-hydrogen atoms.[27] Applying these two alternative similarity criteria, activity cliffs were defined as follows: (1.1) Similarity criterion: MACCS Tanimoto coefficient (Tc) value, 0.80; (1.2) Potency difference criterion: At least 2 orders of magnitude;[1] (2.1) Similarity criterion: MMP formation; and (2.2) Potency difference criterion: At least 2 orders of magnitude.

Thus, in both cases, an at least 100-fold difference in potency was required (i.e., at least two pK_i units). In our analysis, high potency activity cliff compounds (*high ActCliff*) and low potency activity cliff compounds (*low ActCliff*) were separately predicted. In addition to the similarity and potency difference criteria, the following conditions were applied to limit the analysis to well-defined cliffs involving high potency compounds and account for compounds forming multiple cliffs: (i) To qualify as *high ActCliff*, a compound was required to have a pK_i value of at least 8.0 and at least two low potency cliff partners with a pK_i value not larger than 6.0. (ii) To qualify as *low ActCliff*, a compound was required to have a pK_i value not larger than 6.0 and at least two high potency cliff partners with a pK_i value of at least 8.0.

Table 3. ECP Statistics[a]

	MACCS		MMP	
(a) positive training examples = 3				
CHEMBL ID	*low ActCliff*	*high ActCliff*	*low ActCliff*	*high ActCliff*
CHEMBL205	27735	25224	8722	11404
CHEMBL218	10084	10689	10260	9180
CHEMBL228	1754	1721	88557	94090
CHEMBL237	21270	18837	22059	17077
CHEMBL244	131081	153011	70901	59640
CHEMBL253	17787	19568	610	605
CHEMBL256	136289	124566	65513	73945
CHEMBL259	111380	146619	206	236
CHEMBL261	45021	57740	47582	47956
(b) positive training examples = 5				
	MACCS		MMP	
CHEMBL ID	*low ActCliff*	*high ActCliff*	*low ActCliff*	*high ActCliff*
CHEMBL205	65586	66366	16892	24227
CHEMBL218	21103	21418	18486	15658
CHEMBL228	3044	2549	264182	294115
CHEMBL237	51032	46568	39949	31879
CHEMBL244	454146	520138	199357	178723
CHEMBL253	45319	45290	823	690
CHEMBL256	452761	445340	189115	229687
CHEMBL259	372728	508319	211	255
CHEMBL261	124077	140376	139763	137232
(c) positive training examples = 10				
	MACCS		MMP	
CHEMBL ID	*low ActCliff*	*high ActCliff*	*low ActCliff*	*high ActCliff*
CHEMBL205	211229	209550	41912	55562
CHEMBL218	60033	57621	32462	31917
CHEMBL228	6586	3994	1106478	1156193
CHEMBL237	145914	135682	90670	75886
CHEMBL244	2143950	2425959	766904	643773
CHEMBL253	138534	132232	1152	865
CHEMBL256	2081218	2205140	753897	874963
CHEMBL259	1646607	2301299	206	278
CHEMBL261	399508	524176	493716	552953

[a] For each data set, the total number of ECPs identified in 100 individual trials with randomly chosen training sets containing (a) 3, (b) 5, and (c) 10 positive training examples is reported for *low/high ActCliff*.

Compound Data Sets. For our analysis, nine pK_i value-based compound data sets were assembled from ChEMBL[28] that contained at least 25 qualifying *high/low ActCliffs* for both MACCS Tc- and MMP-based activity cliff definitions. Other compounds were considered non-cliff compounds and potential false positives. For compounds with multiple p*K*i values spanning 1 order of magnitude, a geometric mean of these measurements was calculated as the final potency annotation. The data sets consisted of different enzyme inhibitors or receptor ligands and contained between 1199 and 1576 compounds. Their composition is summarized in Table 1.

ECP Calculations and Performance Criteria. For each classification trial, small numbers of 3, 5, or 10 *high ActCliff* or *low ActCliff* were randomly selected as positive training examples. As negative training examples, the same number of *high/low ActCliffs* and non-cliff compounds was randomly selected. For example, a training set with three positive *high ActCliff* training examples consisted of nine compounds, i.e.,

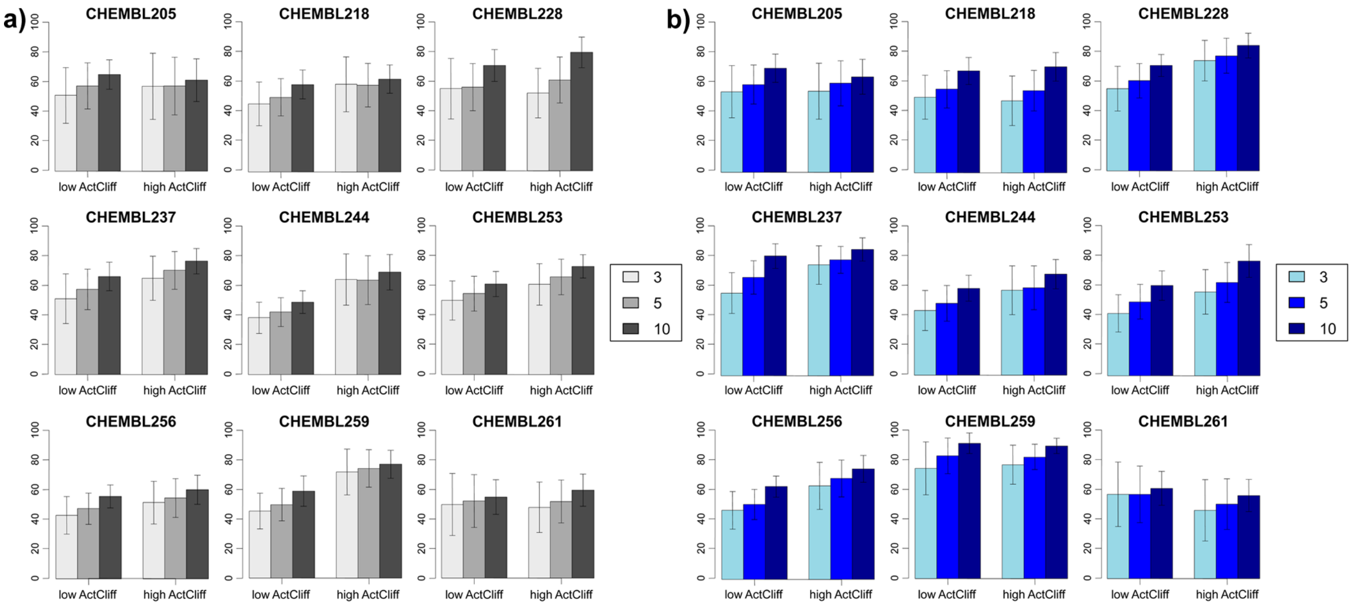


Figure 2. True positive rates. For training sets containing 3, 5, and 10 positive training examples, true positive rates are reported in a bar chart format with standard deviations as error bars: (a) MACCS-based cliffs and (b) MMP cliffs.

Table 4. Prediction Accuracy for MACCS-Based Activity Cliff-Forming Compounds[a]

		sensitivity		specificity	
CHEMBL ID	positive training examples	*low ActCliff*	*high ActCliff*	*low ActCliff*	*high ActCliff*
CHEMBL205	3	0.49	0.55	0.75	0.69
	5	0.54	0.54	0.77	0.71
	10	0.58	0.53	0.78	0.72
CHEMBL218	3	0.43	0.57	0.71	0.68
	5	0.45	0.55	0.72	0.71
	10	0.49	0.56	0.73	0.75
CHEMBL228	3	0.52	0.50	0.70	0.67
	5	0.51	0.58	0.73	0.69
	10	0.62	0.77	0.75	0.73
CHEMBL237	3	0.49	0.64	0.67	0.69
	5	0.55	0.69	0.66	0.70
	10	0.61	0.74	0.67	0.72
CHEMBL244	3	0.35	0.63	0.71	0.75
	5	0.36	0.62	0.72	0.77
	10	0.37	0.67	0.75	0.79
CHEMBL253	3	0.48	0.60	0.73	0.67
	5	0.51	0.64	0.74	0.69
	10	0.54	0.70	0.76	0.71
CHEMBL256	3	0.42	0.51	0.69	0.69
	5	0.46	0.54	0.69	0.68
	10	0.53	0.59	0.71	0.70
CHEMBL259	3	0.44	0.71	0.73	0.62
	5	0.48	0.73	0.72	0.64
	10	0.55	0.75	0.73	0.68
CHEMBL261	3	0.49	0.45	0.73	0.68
	5	0.50	0.48	0.73	0.68
	10	0.51	0.52	0.76	0.70

[a] Average sensitivity and specificity over 100 individual trials for randomly chosen training sets with 3, 5, and 10 positive training examples are reported for *low/high ActCliffs*.

Table 5. Prediction Accuracy for MMP Cliff-Forming Compounds[a]

		sensitivity		specificity	
CHEMBL ID	positive training examples	*high ActCliff*	*low ActCliff*	*low ActCliff*	*high ActCliff*
CHEMBL205	3	0.52	0.52	0.72	0.64
	5	0.55	0.56	0.75	0.64
	10	0.63	0.57	0.76	0.69
CHEMBL218	3	0.47	0.43	0.79	0.74
	5	0.51	0.46	0.78	0.73
	10	0.59	0.57	0.80	0.74
CHEMBL228	3	0.51	0.73	0.80	0.78
	5	0.54	0.75	0.80	0.80
	10	0.60	0.80	0.82	0.81
CHEMBL237	3	0.51	0.71	0.82	0.78
	5	0.60	0.73	0.84	0.82
	10	0.72	0.76	0.86	0.85
CHEMBL244	3	0.39	0.55	0.73	0.77
	5	0.40	0.56	0.74	0.80
	10	0.43	0.63	0.78	0.83
CHEMBL253	3	0.38	0.54	0.77	0.71
	5	0.44	0.59	0.78	0.76
	10	0.49	0.72	0.80	0.82
CHEMBL256	3	0.42	0.62	0.70	0.72
	5	0.44	0.66	0.70	0.73
	10	0.51	0.71	0.70	0.75
CHEMBL259	3	0.72	0.74	0.82	0.84
	5	0.79	0.78	0.85	0.86
	10	0.87	0.83	0.89	0.88
CHEMBL261	3	0.56	0.42	0.71	0.68
	5	0.55	0.43	0.73	0.69
	10	0.57	0.41	0.77	0.70

[a] Average sensitivity and specificity over 100 individual trials for randomly chosen training sets with 3, 5, and 10 positive training examples are reported for *low/high ActCliffs*.

three additional negative *low ActCliff* and three non-cliff training examples. In each case, 100 different trials with randomly assembled training and test sets were carried out.

As performance measures, true positives, *sensitivity* (true positives/(true positives + false negatives)), and *specificity*

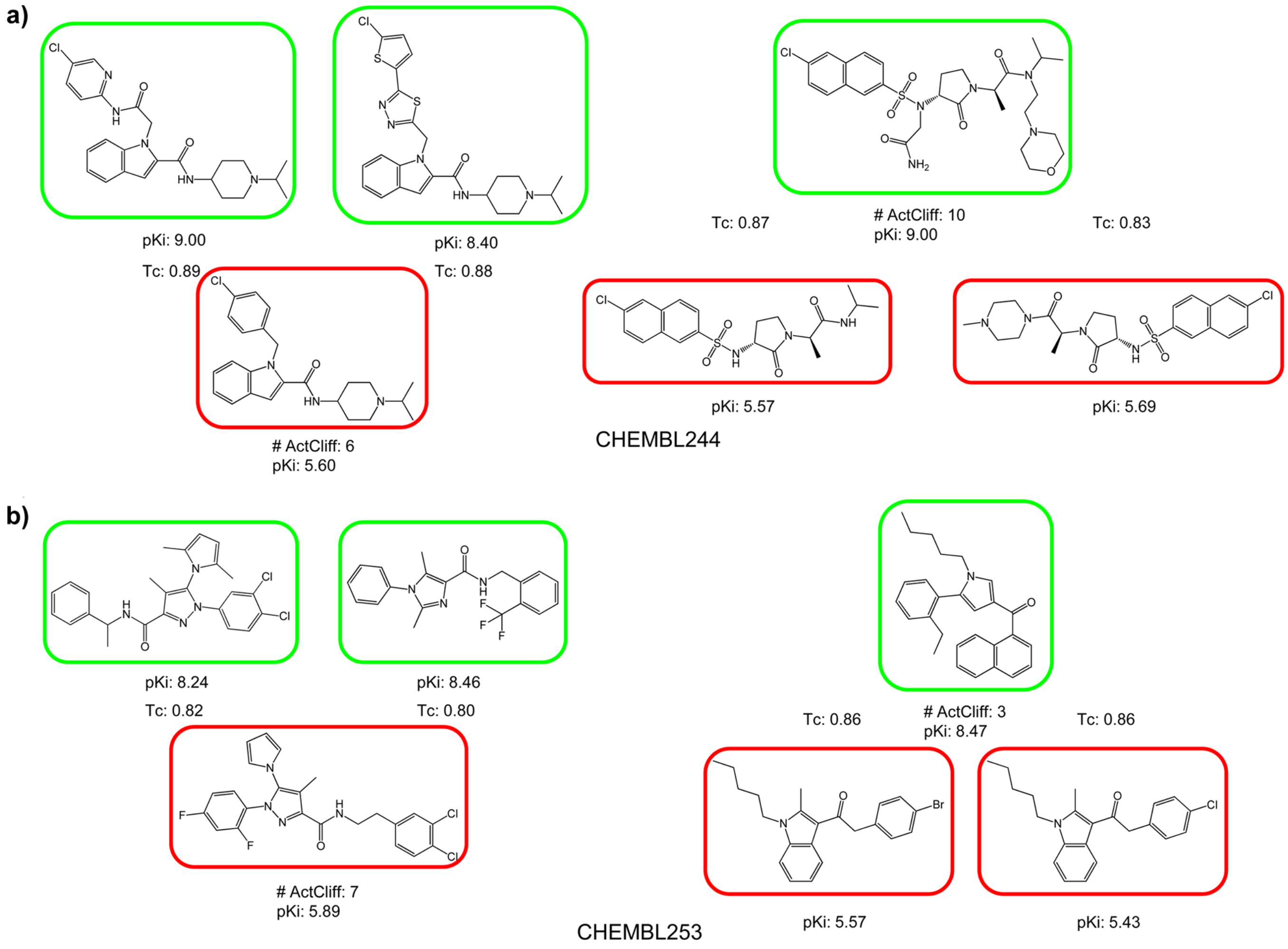


Figure 3. Exemplary MACCS-based activity cliff compounds. Structures of exemplary *high/low ActCliffs* are shown for MACCS-based cliffs originating from (a) CHEMBL244 and (b) CHEMBL253. *high ActCliff* and *low ActCliff* are placed in green and red rectangles, respectively, and the total number of cliff partners are reported. Tc values for activity cliff partners and pKi values are also provided.

(true negatives/(true negatives + false positives)) were calculated and averaged over 100 independent trials.

RESULTS AND DISCUSSION

Study Concept. In this study, we have aimed to predict individual compounds forming well-defined 2D activity cliffs. It should be noted that activity cliffs have been analyzed in two and also three dimensions.[29,30] 3D activity cliffs have been studied by comparing compound binding modes on the basis of X-ray data, and information provided by 2D and 3D cliffs was often found to be complementary.[29,30] However, for medicinal chemistry application, molecular graph representations are usually preferred.[1] For practical applications, retrospective/descriptive analysis of activity cliffs is typically carried out. The prediction of activity cliffs in compound data sets goes beyond such applications. While first attempts have been made to predict activity cliffs through machine learning,[3−5] individual compounds forming activity cliffs have thus far not been predicted. The ability to identify individual cliff-forming compounds removes the compound pair dependence of cliff assignments and predictions. The prediction of individual compounds that form activity cliffs is central aspect of our study. In our analysis, stringent criteria to define activity cliff-forming compounds and alternative similarity criteria for cliff formation have been applied. For two reasons, we have evaluated the ECP methodology for the prediction of individual cliff-forming compounds. First, because such compounds are generally rare in data sets,[1] it is a hallmark of ECP classification to be capable of operating on the basis of small training sets,[15,16] and second, because ECPs typically distinguish compounds at high resolution.[16,17] The latter aspect was considered particularly relevant because *high ActCliff* and *low ActCliff* are per definition structurally similar or analogous compounds and hence generally difficult to distinguish on the basis of chemical structure. In addition, signature patterns for cliff compounds must be capable in implicitly accounting for SAR determinants (leading to large potency differences). Hence, predicting individual compounds to form activity cliffs was considered a challenging task.

Cliff-Forming Compounds. Figure 1 reports the distribution of *high ActCliff* and *low ActCliff* over all data sets when the alternative MACCS-based cliff and MMP cliff definitions were applied. Two key observations can be made. First, the number of cliff-forming compounds was typically small, with few exceptions (e.g., *low ActCliff* in CHEMBL256 or CHEMBL261). Second, there was only limited overlap between cliff-forming compounds for the alternative cliff assignments. This was a direct consequence of applying alternative similarity criteria. In general, MACCS-based cliffs, which relied on whole-molecule similarity calculations, yielded more cliff-forming compounds

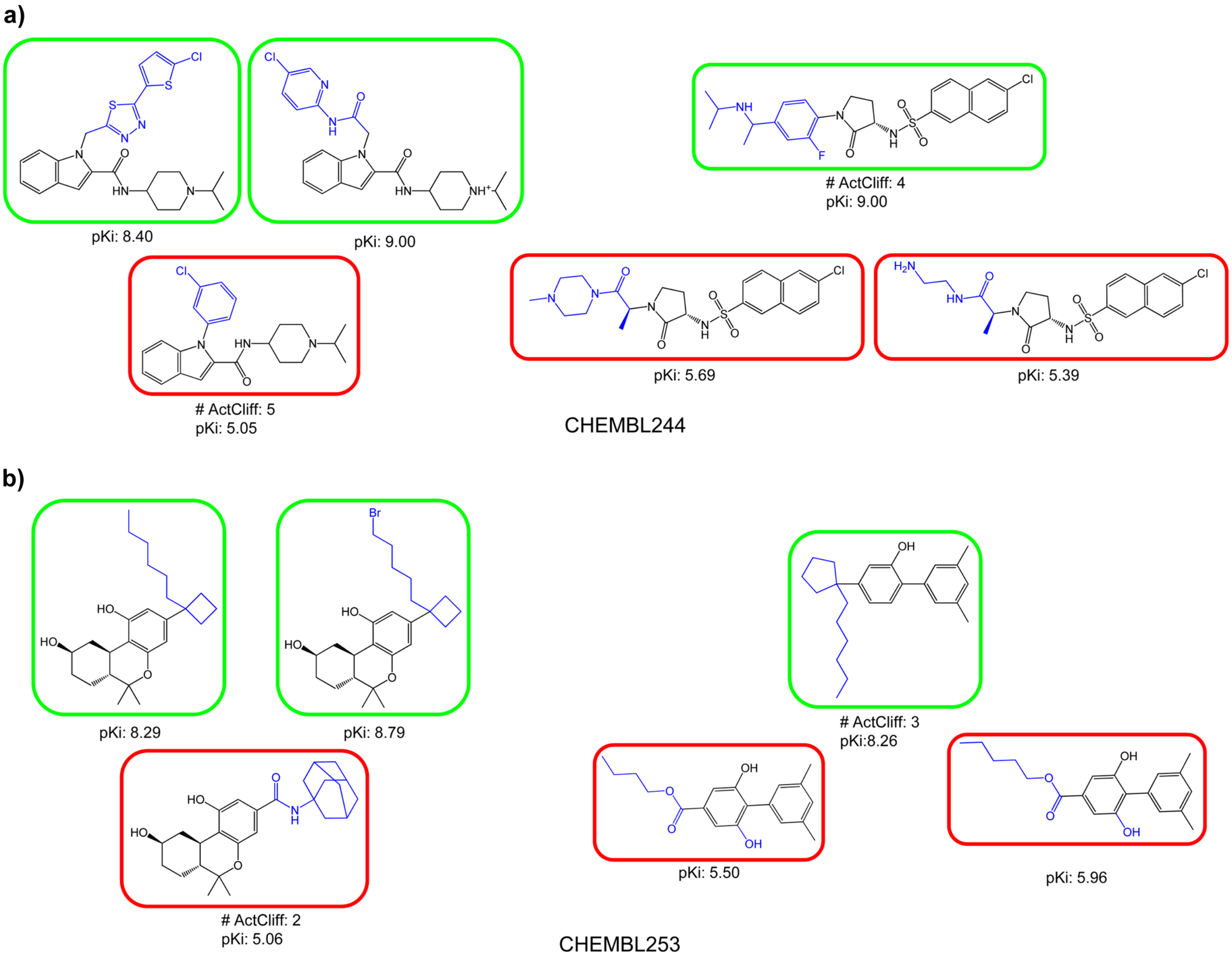


Figure 4. Exemplary MMP cliff compounds. Structures of exemplary *high/low ActCliffs* are shown for MMP cliffs cliffs originating from (a) CHEMBL244 and (b) CHEMBL253. *high ActCliff* and *low ActCliff* are placed in green and red rectangles, respectively, and the total number of cliff partners are reported. Substructures distinguishing cliff partners are colored blue. p*Ki* values are also provided.

than MMP cliffs, which represent a structurally more conservative activity cliff definition.[27]

Qualifying Descriptors. These differences in key compound numbers also affected the number of qualifying descriptors for ECP analysis following discretization, as reported in Table 2. For MACCS-based cliff compounds, more qualifying descriptors were typically (but not always) obtained. For example, for CHEMBL253 or CHEMBL259, MACCS-based cliff and MMP cliff compounds yielded 26 and 7 or 50 and 6 descriptors, respectively. By contrast, for CHEMBL228, the opposite was observed. In this case, MACCS-based cliff and MMP cliff compounds selected 17 and 42 descriptors, respectively.

ECP Statistics. On the basis of qualifying descriptors, ECPs were systematically computed for all training sets containing 3, 5, and 10 positive training examples. The resulting ECP statistics are reported in Table 3. Very large numbers of ECPs were obtained in many cases, frequently more than 100,000 per training class, for limited numbers of qualifying descriptors (hence, reflecting the high-resolution characteristics of ECP ensembles). There was a general trend of increasing ECP numbers for training sets of increasing size. In addition, in part strikingly large differences in ECP numbers between MACCS-based cliff and MMP cliff-forming compounds were observed. For example, for CHEMBL253 in Table 3a, MACCS-based *low ActCliff* and *high ActCliff* produced ~111,000 and ~146,000 ECPs, respectively, whereas MMP cliff *low ActCliff* and *high ActCliff* generated only 206 and 236 ECPs, respectively. By contrast, for CHEMBL228 in Table 3b, 3044 and 2549 ECPs were obtained for MACCS-based *low ActCliff* and *high ActCliff*, respectively, but ~264,000 and ~294,000 ECPs for MMP cliff *low ActCliff* and *high ActCliff*, respectively. Thus, overall, there was a strong dependence of ECP statistics on the activity cliff representations as well as the compound data sets.

Compound prediction. We then attempted to systematically predict individual *low ActCliff* and *high ActCliff* in test sets for MACCS-based cliffs and MMP cliffs on the basis of the identified ECP ensembles. True positive detections are an initial indicator of prediction performance as well as the sensitivity and specificity of the calculations.

True Positives Rates. In Figure 2a and b, true positive rates are reported for *low ActCliff* and *high ActCliff* from MACCS-based cliffs and MMP cliffs, respectively. A key observation has been that *individual activity cliff-forming compounds were correctly predicted at significant rates across all data sets and for the alternative activity cliff definitions.* Three general trends were observed. First, true positive rates slightly increased with the number of positive training examples. Second, true positive rates were overall higher for *high ActCliff* than *low ActCliff*. Third, the rates were overall higher for MMP cliff compounds than MACCS-based cliff compounds. This was the case although fewer qualifying descriptors and much smaller numbers of ECPs were often available for MMP cliff compounds. These findings indicated that predictive performance was not determined by mere ECP frequency but rather by signature patterns resulting from structural changes leading to cliff formation, as further discussed below. In several cases, true positive rates of 80% or more were observed, especially for *high ActCliff*.

Prediction Accuracy. Tables 4 and 5 report the sensitivity and specificity of systematic predictions of *low ActCliff* and *high ActCliff* for MACCS-based cliffs and MMP cliffs, respectively. Sensitivity values between 0.37 and 0.87 were obtained. The sensitivity was overall higher for *high ActCliff* than for *low ActCliff* and for MMP cliff compounds than MACCS-based cliff compounds. However, differences in sensitivity between *high ActCliff* and *low ActCliff* were generally small as well as differences between MACCS-based cliff compounds and MMP cliff compounds. For example, for learning sets with 10 positive training examples, the average sensitivity for *high ActCliff* was ~0.65 and ~0.67 for MACCS-based cliff and MMP cliff compounds, respectively, indicating that more than 60% of the activity cliff-forming compounds were correctly predicted. While achieving this level of sensitivity is considered a success, it leaves room for further improvement and control of false negative rates. However, the results in Tables 4 and 5 also reveal that the specificity of the calculations was generally high. For example, for learning sets with 10 positive training examples, the average specificity for *high ActCliff* was ~0.72 and ~0.79 for MACCS-based cliff and MMP cliff compounds, respectively. Thus, false positive rates were generally low. Overall, both *low ActCliff* and *high ActCliff* were predicted with significant accuracy.

Exemplary Compounds. In Figures 3 and 4, examples of *high/low ActCliffs* are shown from MACCS-based cliffs and MMP cliffs that were consistently predicted with high accuracy. For each individual *high ActCliff* and *low ActCliff*, two cliff partners are also shown. Well-predicted compounds were not confined to specific chemotypes in different data sets but were structurally rather diverse. These examples also illustrate that MMP cliffs were often easier to rationalize in structural terms than activity cliffs defined on the basis of calculated Tanimoto similarity, consistent with overall higher prediction accuracy observed for compounds forming MMP cliffs.

Signature ECPs. Tables 6 and 7 report signature patterns that had strong support and yielded accurate predictions of *high ActCliff* and *low ActCliff*. These patterns varied greatly in their descriptor and value range composition and complexity across different data sets, ranging from single-descriptor patterns to patterns involving six or more descriptors with variable value ranges (descriptor definitions are provided in Table S1, Supporting Information). Complex descriptors combining different components such as surface representations and charge distributions or other surface properties were frequently found in signature patterns. However, no conserved or recurrent patterns were detected that determined multiple predictions. Given the often large numbers of patterns obtained for positive training examples, as discussed above, highly specific ECPs were generally responsible for the correct prediction of *high ActCliff* and *low ActCliff* across different

Table 6. Exemplary ECPs for MACCs-Based Activity Cliff-Forming Compounds[a]

CHEMBL ID	subsets	signature patterns	support
CHEMBL205	*low ActCliff*	{PEOE_VSA-1:(19.947869-inf), PEOE_VSA-4:(-inf-37.541647], PEOE_VSA-5:(-inf-14.335269], SlogP_VSA2:(-inf-6.554696], SMR_VSA7:(98.320915-inf), TPSA:(-inf-101.235]}	0.8
	high ActCliff	{b_1rotN:(1.5-inf), PEOE_RPC-:(-inf-0.223033], SlogP_VSA2:(-inf-6.554696], SMR_VSA2:(0.869983-inf), vsa_base:(16.423053-inf)}	1.0
CHEMBL218	*low ActCliff*	{a_nS:(-inf-0.5]}	0.8
	high ActCliff	{a_nCl:(-inf-1.5], a_nS:(-inf-0.5], chi0v_C:(-inf-17.438794], PEOE_VSA+0:(-inf-189.58638], PEOE_VSA-4:(11.475044-inf), SMR_VSA5:(179.066065-inf)}	0.9
CHEMBL228	*low ActCliff*	{SlogP_VSA9:(-inf-59.803112], SMR_VSA0:(-inf-31.740204], SMR_VSA1:(-inf-21.091469], SMR_VSA3:(-inf-10.419988], SMR_VSA6:(18.041382-inf)}	0.6
	high ActCliff	{PEOE_VSA_FNEG:(0.344313-inf), SlogP_VSA3:(-inf-17.735076], TPSA:(-inf-25.295]}	0.8
CHEMBL237	*low ActCliff*	{PEOE_VSA+2:(-inf-30.514455], PEOE_VSA_FNEG:(-inf-0.396279], SlogP_VSA3:(-inf-73.5439], SMR_VSA0:(141.31085-inf)}	0.9
	high ActCliff	{a_nS:(-inf-0.5], b_1rotR:(-inf-0.194783], PEOE_VSA+2:(30.514455−30.950753]}	0.9
CHEMBL244	*low ActCliff*	{PEOE_VSA+0:(136.86096-inf), PEOE_VSA+5:(-inf-5.442076], PEOE_VSA-4:(43.653418-inf), SMR_VSA2:(-inf-12.373303]}	0.8
	high ActCliff	{SlogP_VSA1:(53.418178−53.456326], SlogP_VSA5:(0.775367−2.05479], SlogP_VSA6:(0.622303−3.955702], SMR_VSA3:(-inf-15.409812]}	1.0
CHEMBL253	*low ActCliff*	{PEOE_VSA+3:(19.813397-inf), SlogP_VSA5:(19.358495-inf)}	0.9
	high ActCliff	{PEOE_VSA-0:(-inf-81.559235], SlogP_VSA2:(4.811791-inf), SlogP_VSA5:(-inf-19.358495]}	1.0
CHEMBL256	*low ActCliff*	{PEOE_VSA-5:(14.818799-inf), petitjean:(-inf-0.458042], SlogP_VSA3:(34.765648-inf), SlogP_VSA5:(33.050338-inf), vsa_other:(-inf-40.701151]}	1.0
	high ActCliff	{PEOE_RPC-:(0.106804−0.139006], PEOE_VSA+4:(-inf-18.716819], PEOE_VSA-3:(23.077686-inf), SMR_VSA0:(-inf-10.367004], SMR_VSA1:(102.92624−104.58829], SMR_VSA4:(24.325969-inf)}	0.9
CHEMBL259	*low ActCliff*	{a_nF:(-inf-2.5], PEOE_VSA+5:(-inf-9.824541], PEOE_VSA-1:(-inf-55.3575], SlogP_VSA9:(90.699826−130.556885], SMR_VSA1:(-inf-47.144615]}	0.8
	high ActCliff	{PEOE_RPC-:(-inf-0.246719], SMR_VSA5:(232.393815-inf), SMR_VSA7:(204.36783-inf), vsa_acid:(-inf-10.745535]}	1.0
CHEMBL261	*low ActCliff*	{PEOE_VSA+1:(-inf-1.10427], petitjean:(0.449495-inf), SlogP_VSA1:(-inf-47.538895], SMR_VSA5:(106.04289−106.441085], vsa_acid:(12.974524−13.726033]}	1.0
	high ActCliff	{a_ICM:(1.514448-inf), PEOE_VSA+1:(-inf-1.10427], PEOE_VSA+2:(31.500751-inf), PEOE_VSA-1:(39.85158-inf), SMR_VSA0:(47.948061−70.021271]}	0.9

[a] For *low/high ActCliff*, exemplary ECPs are reported with their support for an individual trial with 10 positive training examples. Descriptors are abbreviated according to Table S1 of the Supporting Information; "inf" stands for infinity.

Table 7. Exemplary ECPs for MMP Cliff-Forming Compounds[a]

CHEMBL ID	subsets	signature patterns	support
CHEMBL205	*low ActCliff*	{PEOE_RPC-:(0.108325-inf), PEOE_VSA+5:(2.092301−5.172632]}	0.8
	high ActCliff	{PEOE_VSA+5:(5.172632−19.813397], PEOE_VSA-1:(-inf-19.947869], PEOE_VSA-5:(-inf-25.630595], SMR_VSA3:(-inf-5.264666]}	0.9
CHEMBL218	*low ActCliff*	{chi0v_C:(-inf-15.498377], PEOE_VSA+0:(250.58998-inf), SlogP_VSA8:(-inf-150.334635], SMR_VSA5:(-inf-222.530795]}	0.8
	high ActCliff	{a_nS:(0.5-inf), chi0v_C:(-inf-15.498377], PEOE_VSA+0:(250.58998-inf), PEOE_VSA-3:(12.737779−27.148685], PEOE_VSA_FNEG:(-inf-0.164407]}	0.7
CHEMBL228	*low ActCliff*	{SlogP_VSA7:(-inf-54.522341], vsa_pol:(-inf-10.787675]}	1.0
	high ActCliff	{a_don:(-inf-0.5], TPSA:(26.165−26.429999], vsa_other:(-inf-1.10427]}	1.0
CHEMBL237	*low ActCliff*	{a_base:(0.5-inf), SlogP_VSA1:(-inf-20.387569]}	1.0
	high ActCliff	{chi0v_C:(17.619698-inf), PEOE_VSA+1:(8.871125-inf), SMR_VSA0:(-inf-122.20231], vsa_pol:(-inf-42.046135]}	0.9
CHEMBL244	*low ActCliff*	{a_nCl:(1.5-inf), SMR_VSA0:(71.584702-inf), SMR_VSA1:(-inf-23.646768], SMR_VSA3:(-inf-14.216929]}	0.9
	high ActCliff	{a_nCl:(-inf-0.5], chi0v_C:(-inf-16.226208], PEOE_VSA+5:(-inf-2.092301]}	1.0
CHEMBL253	*low ActCliff*	{SlogP_VSA9:(83.584351−136.45901]}	0.8
	high ActCliff	{PEOE_VSA+6:(24.029738-inf), SlogP_VSA7:(130.190405-inf)}	0.9
CHEMBL256	*low ActCliff*	{PEOE_VSA+0:(93.313366-inf), SlogP_VSA3:(-inf-18.811508]}	1.0
	high ActCliff	{SlogP_VSA0:(24.630473−68.424889], SMR_VSA3:(-inf-3.343091], SMR_VSA6:(5.171726−24.433495]}	1.0
CHEMBL259	*low ActCliff*	{PEOE_VSA+3:(-inf-11.173008], SMR_VSA7:(108.71675-inf)}	0.4
	high ActCliff	{PEOE_VSA+3:(-inf-11.173008], SlogP_VSA7:(194.258815-inf)}	0.4
CHEMBL261	*low ActCliff*	{balabanJ:(2.31318−2.556531], b_1rotN:(-inf-0.5], b_1rotR:(-inf-0.080128], SlogP_VSA1: (53.787146−71.176098], vsa_don:(31.724531-inf)}	1.0
	high ActCliff	{PEOE_VSA+5:(2.092301−33.934933], PEOE_VSA+6:(26.904799-inf), SlogP_VSA7:(202.3759-inf), SMR_VSA7:(-inf-6.521481]}	0.9

[a] For *low/high ActCliff*, exemplary ECPs are reported with their support for an individual trial with 10 positive training examples. Descriptors are abbreviated according to Table S1 of the Supporting Information; "inf" stands for infinity.

compound classes. ECP calculations were capable of indirectly accounting for small structural differences leading to significant changes in compound potency.

■ CONCLUSIONS

In this study, we have attempted to predict individual compounds that form activity cliffs on the basis of emerging chemical patterns, which have previously been utilized for compound classification. Activity cliff-forming compounds are generally rare in data sets. Predicting such compounds principally requires the availability of characteristic features that implicitly account for small structural changes leading to large potency differences. Therefore, we have focused on the emerging chemical patterns approach that typically generates large numbers of descriptor patterns on the basis of small training sets, which can be mined for signature patterns of activity cliff compounds.

The prediction of individual compounds participating in the formation of activity cliffs represents a rather complex task. First, small numbers of activity cliff compounds with either high or low potency must be distinguished from each other. Second, activity cliff compounds must also be differentiated from much larger numbers of non-cliff compounds present in data sets.

Despite these challenges, single-compound predictions have utility in activity cliff assessment. In practical applications, individual compounds can be identified in large data sets that have a high propensity to form activity cliffs. These compounds are likely to be SAR-informative and might form multiple cliffs with multiple partners. This can be easily determined once individual compounds have been prioritized. Compared to compound pair predictions, which require combined similarity and potency measures, ECP-based single-compound predictions have the advantage that only little compound information is required for training and that activity cliffs do not need to be explicitly considered.

On the basis of ECP calculations, we have been able to predict individual high potency and low potency cliff compounds with reasonable sensitivity and high specificity in different compound data sets. On average, ~60−70% of all compounds participating in the formation of differently defined activity cliffs were correctly detected and effectively distinguished from non-cliff compounds, as indicated by generally low false positive rates. Hence, the approach presented herein should be promising to further expand activity cliff predictions by focusing on individual compounds.

■ ASSOCIATED CONTENT

Supporting Information

Table S1 provides descriptor details. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Stumpfe, D.; Bajorath, J. Exploring activity cliffs in medicinal chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.

(2) Wassermann, A. M.; Wawer, M.; Bajorath, J. Activity landscape representations for structure−activity relationship analysis. *J. Med. Chem.* **2010**, *53*, 8209−8223.

(3) Namasivayam, V.; Bajorath, J. Searching for coordinated activity cliffs using particle swarm optimization. *J. Chem. Inf. Model.* **2012**, *52*, 927−934.

(4) Guha, R. Exploring uncharted territories: Predicting activity cliffs in structure−activity landscapes. *J. Chem. Inf. Model.* **2012**, *52*, 2181−2191.

(5) Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of activity cliffs using support vector machines. *J. Chem. Inf. Model.* **2012**, *52*, 2354−2365.

(6) Vogt, M.; Iyer, P.; Maggiora, G. M.; Bajorath, J. Conditional probabilities of activity landscape features for individual compounds. *J. Chem. Inf. Model.* **2013**, *53*, 1602−1612.

(7) Santos, R.; Giulianotti, M. A.; Houghten, R. A.; Medina-Franco, J. L. Conditional probabilistic analysis for prediction of the activity landscape and relative compound activities. *J. Chem. Inf. Model.* **2013**, *53*, 2613−2625.

(8) 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, 1999, pp 30−42.

(9) 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.

(10) Li, J.; Dong, G.; Ramamohanarao, K. Making use of the most expressive jumping emerging patterns for classification. *Knowl. Inf. Syst.* **2001**, *3*, 131−145.

(11) 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, Florida, 2003; IEEE Computer Society: Los Alamitos, CA, 2003, p 485.

(12) Li, J.; Dong, G.; Ramamohanarao, K.; Wong, L. DeEPs: A new instance-based lazy discovery and classification system. *Mach. Learn.* **2004**, *54*, 99−124.

(13) Wang, L.; Zhao, H.; Dong, G.; Li, J. On the complexity of finding emerging patterns. *Theor. Comput. Sci.* **2005**, *335*, 15−27.

(14) Li, J.; Wong, L. Identifying good diagnostic gene groups from gene expression profiles using the concept of emerging patterns. *Bioinformatics* **2002**, *18*, 725−734.

(15) Auer, J.; Bajorath, J. Emerging chemical patterns: A new methodology for molecular classification and compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 2502−2514.

(16) Auer, J.; Bajorath, J. Simulation of sequential screening experiments using emerging chemical patterns. *Med. Chem.* **2008**, *4*, 80−90.

(17) 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.

(18) 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.

(19) 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.

(20) 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.

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

(23) 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.

(24) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

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

(26) Hussain, J.; Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.* **2010**, *50*, 339−348.

(27) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic identification of activity cliffs on the basis of matched molecular pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.

(28) 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.

(29) Hu, Y.; Bajorath, J. Exploration of 3D activity cliffs on the basis of compound binding modes and comparison of 2D and 3D cliffs. *J. Chem. Inf. Model.* **2012**, *52*, 670−677.

(30) Hu, Y.; Furtmann, N.; Gütschow, M.; Bajorath, J. Systematic identification and classification of three-dimensional activity cliffs. *J. Chem. Inf. Model.* **2012**, *52*, 1490−1498.