JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

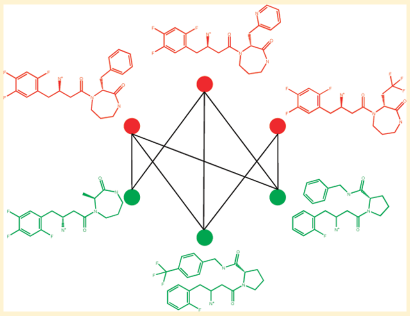
pubs.acs.org/jcim

Searching for Coordinated Activity Cliffs Using Particle Swarm Optimization

Vigneshwaran Namasivayam and Jürgen Bajorath*

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

ABSTRACT: Activity cliffs are formed by structurally similar compounds having large potency differences. Coordinated activity cliffs evolve when compounds within groups of structural neighbors form multiple cliffs with different partners, giving rise to local networks of cliffs in a data set. Using particle swarm optimization, a machine learning approach, we systematically searched for coordinated activity cliffs in different compound sets. Regardless of the global SAR characteristics of these data sets, coordinated activity cliffs introducing strong local SAR discontinuity were identified in most cases. Compound subsets forming coordinated activity cliffs represent centers of SAR discontinuity and have high SAR information content. Through particle swarm optimization guided by subset discontinuity scoring, compounds forming the largest coordinated activity cliffs can automatically be extracted from large compound data sets.



■ INTRODUCTION

Activity cliffs are defined as pairs of structurally similar active compounds having large differences in potency.[1−3] Hence, activity cliffs reveal small structural modifications that have a significant effect on a specific biological activity. This explains the attractiveness of the activity concept for medicinal chemistry.[2,3] In SAR analysis, this "small change−large effect" phenomenon is referred to as SAR discontinuity.[2] By contrast, gradual structural changes in compound series that are accompanied by small to moderate changes in potency are rationalized as SAR continuity.[2] Activity cliffs can be explored for individual compound series or on a large scale through systematic compound data mining.[3,4]

For analysis of activity cliffs, the "similarity" and "potency difference" parameters must be carefully considered. Without defining the degree of similarity two compounds must share to qualify as potential cliff partners, activity cliff analysis remains ambiguous.[3] In addition, the minimal potency difference applied as a cliff criterion must also be specified.[3] Furthermore, it must be taken into account that the assessment of molecular similarity depends on molecular representations (descriptors) and similarity metrics that are used, which presents another variable for activity cliff analysis.[2,3] Simply put, an activity cliff is not a compound data structure written in stone but subject to definition and influenced by the choice of molecular representations, similarity, and potency criteria. Nevertheless, if clearly defined in the context of a given analysis, activity cliffs are of high relevance for SAR analysis.

Thus far, activity cliffs have been mostly considered on the basis of compound pairs, which is the most straightforward way to conceptualize activity cliffs and evaluate structural modifications leading to large potency effects. However, an observation we made in evaluating activity landscape models[2] of many different compound data sets is that activity cliffs are not always formed by individual pairs of compounds. Rather, one can also observe formation of multiple and overlapping cliffs involving groups of structural neighbors. Given these observations, we previously designed a hypothetical data structure termed an "activity ridge" that captured multiple activity cliffs.[5] This data structure was defined to consist of a set of at least five structurally similar compounds with nanomolar potency and another set of five or more structural neighbors of these potent compounds with at least 100-fold higher or lower potency. All high- and low-potency compounds were required to form activity cliffs with all of their counterparts. We searched for this hypothetical data structure in 242 compound sets and detected one or more activity ridges in 71 of them.[5] A characteristic feature of ridge-like structures is that they have high SAR information content due to their high activity cliff density.

Here, we report an approach to systematically search compound data sets for "coordinated" activity cliffs that represent centers of SAR discontinuity. Coordinated activity cliffs, as defined herein, are conceptually similar to activity ridges but represent more flexible data structures. In order to automatically extract the most prominent arrays of coordinated activity cliffs (i.e., combinations of largest cliffs) from compound sets, a particle swarm optimization (PSO)[6] scheme was designed. Compound subsets forming coordinated activity cliffs were detected and differentiated by their contributions to SAR discontinuity. In most of the compound data sets we analyzed, coordinated activity cliffs were identified that introduced strong SAR discontinuity, irrespective of the global SAR phenotype of these sets.

Received: January 27, 2012
Published: March 11, 2012

dx.doi.org/10.1021/ci3000503 | J. Chem. Inf. Model. 2012, 52, 927−934

METHODS AND CALCULATIONS

Definition of Coordinated Activity Cliffs. We searched for compounds forming coordinated activity cliffs meeting the following criteria.

(1) All pairs of compounds within an activity cliff forming subsets had to share more than 80% Tanimoto similarity[7] using MACCS structural keys[8] as a molecular representation (i.e., a MACCS Tanimoto coefficient (Tc) > 0.8).

(2) The potency values of all highly potent compounds had to fall within 1 order of magnitude, and the same criterion was applied to the low-potency molecules.

(3) For formation of an activity cliff, the more potent compound was required to have a potency of at least 100 nM.

(4) Compound subsets containing a minimum of three or six high- and three or six low-potency compounds were considered.

These criteria are similar to the ones applied for defining activity ridges.[5] However, for our current analysis and systematic PSO calculations, three changes were made. First, the numbers of required compounds per subsets were varied. Second, whereas the similarity criterion applied for detection of compound ridges was the presence of topologically equivalent scaffolds in all high- and low-potency level compounds, here a Tanimoto similarity threshold was used instead. This was done to permit detection of compounds with similar but not necessarily topologically equivalent scaffolds. Third, compound subsets were selected such that the largest potency margin between low- and high-potency level compounds was at least 2 orders of magnitude. Because potency variations within an order of magnitude were permitted at the high- and low-potency level, this meant that a compound included in a subset was not required to form activity cliffs with a potency difference of at least 2 orders of magnitude with all of its counterparts. Taken together, these modifications introduced for detection of coordinated activity cliffs relaxed the criteria used to define activity ridges. Consequently, it was possible to evaluate larger numbers of compound subsets in the search for the most prominent centers of coordinated activity cliff formation. An exemplary compound subset forming coordinated activity cliffs is shown in Figure 1.

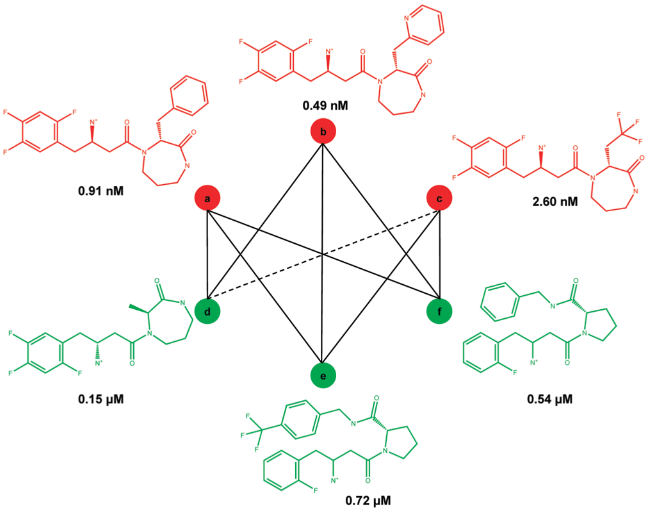


Figure 1. Coordinated activity cliffs. Formation of coordinated activity cliffs is illustrated for dipeptidyl peptidase IV inhibitors. High- and low-potency level compounds are colored in red and green, respectively. All pairwise compound comparisons yield a MACCS Tc > 0.8. Activity cliffs formed between compounds with a potency difference of at least 2 orders of magnitude or a smaller potency difference are connected by solid and dashed lines, respectively.

Particle Swarm Optimization. The PSO methodology was introduced in 1995.[6] PSO is a population-based global search technique that emulates collaborative swarm behavior (such as flocking of birds or schooling of fish). PSO does not utilize evolution operators such as crossover and mutation. Rather, each possible solution is regarded as a "particle" that navigates search space by learning from nearest neighbors.

In a D-dimensional search space, the PSO algorithm is initialized with a population ("swarm") with N random solutions termed particles. Each particle of the swarm is represented by two attributes including its position and velocity. The position of the particle in the swarm is determined by a vector $x_{id} = x_{i1}, x_{i2}, ..., x_{iD}$ with $i = 1, 2, ..., N$ being the number of particles in the swarm and $d = 1, 2, .., D$ the number of dimension of each particle. On the basis of their velocity, represented by a vector $v_{id} = v_{i1}, v_{i2}, ..., v_{iD}$, the particles explore the search space. During each *iteration* (t), a particle evaluates its current position relative to its former best position (*pbest*) and best position of its neighboring particles (*nbest*) and updates its velocity and position according to eqs 1 and 2. The search is guided by optimization of a fitness function until convergence is reached.

$$v_{id}(t+1) = wv_{id}(t) + c_1 rand()[pbest_{id}(t) - x_{id}(t)] + c_2 rand()[nbest_{id}(t) - x_{id}(t)] \quad (1)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad (2)$$

Here, c_1 and c_2 are the cognitive and social confidence coefficients. They represent the acceleration constants that change the velocity of a particle toward *pbest* and *nbest*, respectively. In addition, *rand*() is a random function uniformly distributed between the values of 0 and 1, and w is the inertia weight, a critically important parameter for balancing the global and local search capacities of particles in the swarm.

PSO was originally developed to solve optimization problems in search spaces with continuously valued dimensions. Kennedy and Eberhart also proposed a binary version of PSO for dimensions with binary values.[9] In binary PSO, the particle velocity is also updated according to eq 1. However, due to the binary ("0" or "1") encoding of dimensions, the particle position is updated through a sigmoidal transformation of the velocity according to eqs 3 and 4

$$S(v_{id}(t+1)) = \frac{1}{1 + e^{-v_{id}(t+1)}} \quad (3)$$

$$\text{if } (rand() < S(v_{id}^{t+1})) \text{ then } x_{id}(t+1) = 1$$
$$\text{or else } x_{id}(t+1) = 0 \quad (4)$$

A pseudocode outline of binary PSO that was used in our analysis is given in Figure 2. All PSO calculations were carried out with an in-house program.

The PSO approach has initially been developed in the machine learning field but has in recent years also been applied in the context of molecular design and SAR analysis.[10−15] Our choice of the PSO approach to search for coordinated activity cliffs was motivated by a previous application where PSO was

```
Step 1: Initialize the swarm (S) by randomly generating 'N' particles
Step 2: Evaluate the fitness of each particle in the swarm
Step 3: Compare particle's fitness value to identify and update both personal best ( $pbest$ ) and
neighbors' best ( $nbest$ ) positions
Step 4: Update the velocity $v_{id}$ of all particles using eqn.1
Step 5: Compute sigmoidal transformation $S(v_{id})$ and update position $x_{id}$ using eqn. 3 and 4
Step 6. Repeat step 2 to 5 until the termination criterion (number of evaluations) is reached.
```

Figure 2. Particle swarm optimization. Steps involved in the binary PSO protocol applied to search for coordinated activity cliffs are summarized in a pseudocode format.

Table 1. Compound Data Sets and PSO Results[a]

				no. of compounds in the subset forming coordinated activity cliffs							
				n = 6				n = 3			
class	target	global SARI	size	total	high potency	low potency	discont. score	total	high potency	low potency	discont. score
1	arachidonate 5-lipoxygenase	0.79	253	0	0	0	0.00	0	0	0	0.00
2	cyclooxygenase-1	0.79	310	0	0	0	0.00	0	0	0	0.00
3	dopamine transporter	0.75	872	0	0	0	0.00	0	0	0	0.00
4	MAP kinase p38 α	0.58	807	0	0	0	0.00	12	9	3	0.86
5	cathepsin B	0.72	304	0	0	0	0.00	0	0	0	0.00
6	protein kinase C β	0.37	120	0	0	0	0.00	11	8	3	1.00
7	carbonic anhydrase XII	0.51	361	0	0	0	0.00	10	7	3	1.00
8	Tyr phosphatase 1B	0.73	472	0	0	0	0.00	0	0	0	0.00
9	matrix metalloproteinase-3	0.52	433	12	6	6	0.37	15	12	3	0.98
10	caspase-1	0.19	177	18	11	7	0.40	8	5	3	0.74
11	Tyr kinase TIE-2	0.52	299	17	11	6	0.56	17	13	4	1.00
12	protein kinase C α	0.43	274	39	33	6	0.73	20	17	3	1.00
13	c-Jun N-terminal kinase-1	0.72	250	14	7	7	0.79	7	4	3	0.89
14	matrix metalloproteinase-13	0.57	552	29	7	22	0.85	7	4	3	1.00
15	matrix metalloproteinase-9	0.43	400	15	6	9	0.85	11	6	5	0.95
16	β-secretase 1	0.52	650	18	7	11	0.87	8	4	4	1.00
17	renin	0.28	557	17	10	7	0.87	11	8	3	1.00
18	cyclooxygenase-2	0.76	562	13	6	7	0.87	9	5	4	0.95
19	Tyr kinase SRC	0.52	696	19	12	7	0.89	11	7	4	0.98
20	dipeptidyl peptidase IV	0.44	1232	14	7	7	0.90	6	3	3	1.00
21	cytochrome P450 2D6	0.76	460	21	13	8	0.91	21	13	8	0.91
22	Tyr kinase LCK	0.38	419	12	6	6	0.91	8	5	3	0.98
23	Ser/Thr kinase Chk1	0.52	457	15	8	7	0.92	7	4	3	1.00
24	norepinephrine transporter	0.74	1378	16	9	7	0.96	9	6	3	1.00
25	phosphodiesterase 7A	0.54	129	15	9	6	0.98	15	10	5	0.98
26	matrix metalloproteinase-1	0.51	501	13	6	7	0.98	12	4	8	0.98
27	acetylcholinesterase	0.51	681	26	18	8	0.99	13	10	3	1.00
28	Ser/Thr kinase AKT	0.31	458	14	6	8	0.99	12	9	3	1.00
29	matrix metalloproteinase-2	0.51	500	12	6	6	0.99	8	3	5	1.00
30	carbonic anhydrase I	0.55	931	30	12	18	1.00	11	8	3	1.00
31	butyrylcholinesterase	0.59	635	20	9	11	1.00	13	9	4	1.00
32	coagulation factor X	0.26	1553	15	6	9	1.00	38	5	33	1.00

[a] Thirty two compound data sets including various classes of enzyme or transporter inhibitors and receptor antagonists are listed. For each compound set, the target name, its global SARI score, and size are given. Furthermore, results of the PSO calculations are reported. In each case, the composition and discontinuity scores of the best-scoring subsets containing at least three (n = 3) or six (n = 6) high and low potency compounds are reported.

successfully used to extract compounds representing discontinuous local SARs from compound data sets.[14] This problem is conceptually related to yet distinct from the search for specific activity cliff patterns, and we hence reasoned that PSO might also be a promising approach for our current analysis.

Subset Discontinuity Score. To differentiate between the magnitudes of coordinated activity cliffs and their contribution to local SAR discontinuity, a "subset discontinuity score" was used here as a fitness function for particles in the swarm to guide the PSO calculations.

$$\mathrm{raw}_{\mathrm{disc}} = \frac{\sum_{\{i,j|\mathrm{sim}(i,j)>0.65,\mathrm{potdiff}(i,j)>1|i\neq j\}} \mathrm{potdiff}(i,j)\times \mathrm{sim}(i,j)}{|\{i,j\,|\,\mathrm{sim}(i,j)>0.65,\mathrm{potdiff}(i,j)>1\,|\,i\neq j\}|} \qquad (5)$$

Here, *potdiff*(i,j) is the absolute potency difference of compound i and j and sim(i,j) the MACCS Tc value resulting from comparison of i and j. For calculation of raw scores, a MACCS Tc similarity threshold > 0.65 was applied and a potency difference between i and j of at least 1 order of magnitude. These conditions ensured that only compound pairs representing a basic level of discontinuity were included in the score calculation. The subset discontinuity score is a variant of the SAR Index (SARI)[16] that is designed to quantitatively characterize the global SAR phenotype of a compound data set. SARI consists of separate SAR continuity and discontinuity score components that are combined to yield the final SARI value.

The raw subset discontinuity score (raw_{disc}) was calculated by taking into account all qualifying pairwise comparisons for compounds comprising a subset. Raw scores were converted into Z scores utilizing the discontinuity score distribution of an external reference panel of compound sets. Assuming a normal distribution the cumulative probability for each Z score was calculated and mapped to the value range (0,1). Scores of 0 and 1 indicate minimal and maximal subset discontinuity, respectively.

Calculation Parameters. The inertia weight, cognitive and social parameters, and two independent random sequences are the main parameters of the PSO algorithm that determine the search characteristics and convergence behavior.[17] As in our previous PSO-based SAR studies,[14,15] we applied the parameter settings suggested by Clerc[18] with an inertia weight w = 0.721348, confidence coefficients c_1 and c_2 = 1.193147, a swarm size of 35 particles, and a maximum number of 5000 evaluation steps. For each compound data set and subset size, 10 independent PSO runs were carried out and the overall highest scoring compound subset was selected. During each iteration of the optimization process, the Euclidean distance between the particle position $x_{id}(t)$ and the compounds in the data set was determined. Compounds were ranked in the order of increasing Euclidean distance from the particle position. Database molecules falling within the similarity threshold radius (MACCS Tc > 0.8) of the top-ranked compounds were selected and sorted according to their potency values. If the number of database compounds meeting the criteria to form coordinated activity cliffs was large enough (i.e., $n = 3$ or $n = 6$), a particle-based subset was selected and scored.

Data Sets. For our analysis 32 compound data sets were assembled from ChEMBL[19] consisting of different classes of inhibitors and antagonists, as reported in Table 1. These data sets contained between 120 and 1553 compounds. Whenever available, K_i values were used as potency measurements. If not, only IC_{50} values were considered instead. If multiple K_i (or IC_{50}) values were available, the geometric mean was calculated and used as the final potency annotation. On the basis of their global SARI scores, which are also reported in Table 1, these data sets represented different global SAR phenotypes, ranging from mostly continuous (SARI scores > 0.70)[16] to heterogeneous (scores around 0.5) and mostly discontinuous (scores < 0.30) SARs. Thus, these compound selections ensured broad coverage of activity classes and global SAR phenotypes for our analysis.

RESULTS AND DISCUSSION

Search Characteristics. We monitored the score and subset distributions of the PSO calculations and their convergence behavior. According to our definition, the $n = 3$ subset consisted of at least six (three high plus three low potency) and the $n = 6$ subset of at least 12 (six plus six) compounds.

Subset Sizes and Scores. In the course of individual PSO runs, compound subsets of very different size and discontinuity scores were detected including differently sized subsets yielding high discontinuity scores. A representative example is shown in Figure 3. Hence, particles were sensitive to subsets of different composition. Omission of individual compounds from subsets would ultimately thus still yield high discontinuity scores. In the course of the optimization, subset sizes and scores notably narrowed down. In general, more candidate subsets were found for $n = 3$ (Figure 3a) than for $n = 6$ subsets (Figure 3b), as one should expect. For different compound classes, PSO calculations converged at rather different subset sizes, as further discussed below.

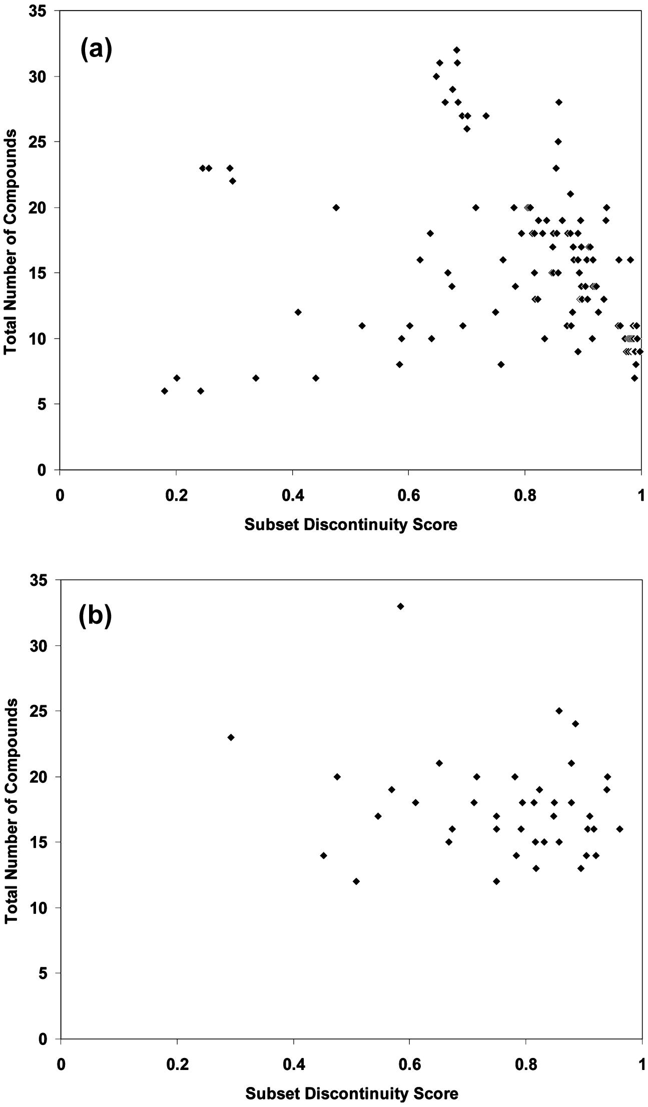


Figure 3. Candidate compound subsets. For all subsets of norepinephrine transporter inhibitors detected by PSO in the course of the optimization run identifying the best-scoring coordinated activity cliffs, the number of compounds and their subset discontinuity scores are reported. Results are shown for subsets containing at least (a) three $n = 3$ or (b) six $n = 6$ high- and low-potency compounds.

Convergence. Compared to different types of clustering approaches, PSO calculations produce at least comparably accurate results but usually converge much faster.[20] Efficient convergence behavior has also been observed in our current analysis. Most PSO calculations converged well within 5000 evaluation steps at very high subset discontinuity score levels, as illustrated in Figure 4. A characteristic feature of many successful search calculations was that $n = 3$ (Figure 4a) and $n = 6$ (Figure 4b) subsets with high discontinuity scores > 0.9 were already detected during the first ~100 evaluation steps. Then, only small increases in scores were observed until convergence was reached at the highest scoring level (in many instances, close or equal to 1.00). At high scoring levels, a "staggered" convergence behavior was often observed when a number of evaluations within a local minimum did not yield a score improvement until a single step led to a further increase. This is illustrated for the $n = 3$ subset in Figure 4a. The $n = 6$ subset in Figure 4b displayed much quicker convergence. This staggered convergence behavior could be attributed to the fact that most of the search proceeded at already high scoring levels where further improvements were difficult to achieve. Nevertheless, as illustrated in Figure 3, subsets with significantly different scores were detected and evaluated during the initial rounds of evaluation. Taken together, these findings indicated that in successful cases coordinated activity cliffs making large contributions to SAR discontinuity were already detected early on during the PSO calculations, prior to reaching the final convergence level. Thus, in practical applications, short optimizations might often suffice to produce high-quality candidate subsets representing coordinated activity cliffs if available in a data set.

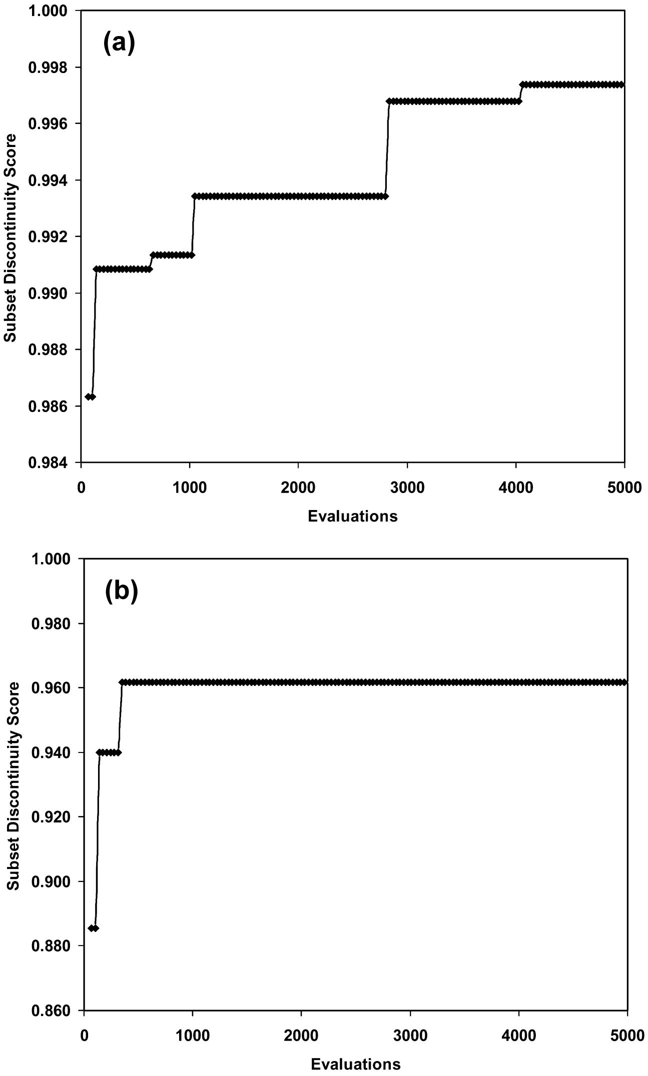


Figure 4. Convergence behavior. For the PSO runs identifying the best-scoring coordinated activity cliffs formed by subsets of norepinephrine transporter inhibitors according to Figure 3, convergence characteristics are reported by monitoring maximal intermediate subset discontinuity scores as a function of the number of evaluation steps: (a) $n = 3$ and (b) $n = 6$.

Subset Distribution. We also monitored the distribution of evaluated subsets over all compound classes. Figure 5 reports the total number of $n = 3$ and $n = 6$ subsets that were detected during the PSO runs, yielding the overall best solutions for each compound class. The figure reveals that depending on the compound data set significantly varying numbers of subsets were evaluated during the simulations. In a number of cases, between 50 and 100 different $n = 6$ subsets or between 100 and 200 $n = 3$ subsets were detected, whereas in others only a few subsets were found. For some classes, many more $n = 3$ than $n = 6$ subsets were detected, indicating that in these cases only small groups of compounds formed coordinated activity cliffs. The compound class-dependent differences in subset distributions observed during PSO analysis also mirrored the different frequencies with which coordinated activity cliffs occurred.

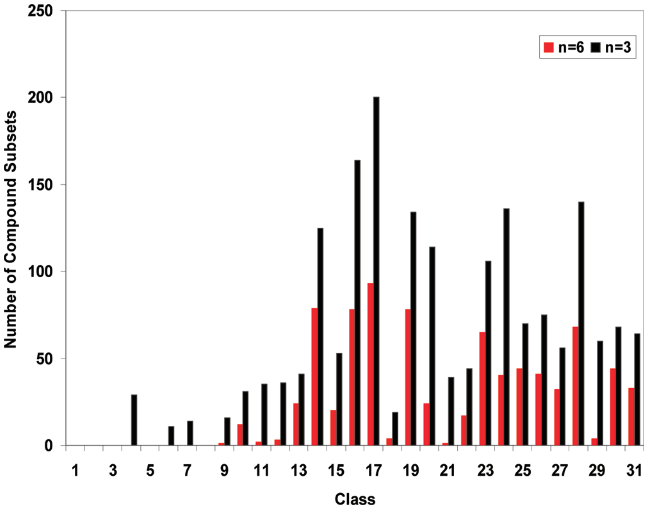


Figure 5. Subset distribution. For 31 activity classes, the total number of compound subsets evaluated by PSO is reported for the runs identifying the best-scoring coordinated activity cliffs (black, $n = 3$; red, $n = 6$). For clarity, class 32 (factor Xa inhibitors) is omitted for which 445 $n = 3$ and 270 $n = 6$ subsets were detected.

Identification of Coordinated Activity Cliffs.

Using our PSO protocol, all 32 data sets were systematically searched for compound subsets forming coordinated activity cliffs. The search results are summarized in Table 1.

Frequency of Detection. We found that most data sets contained coordinated activity cliffs of significant magnitude. For $n = 6$ and $n = 3$ subsets, coordinated activity cliffs were found in 24 and 27 of 32 compound sets, respectively. In only five data sets including inhibitors of arachidonate 5-lipoxygenase, cyclooxygenase-1, dopamine transporter, cathepsin

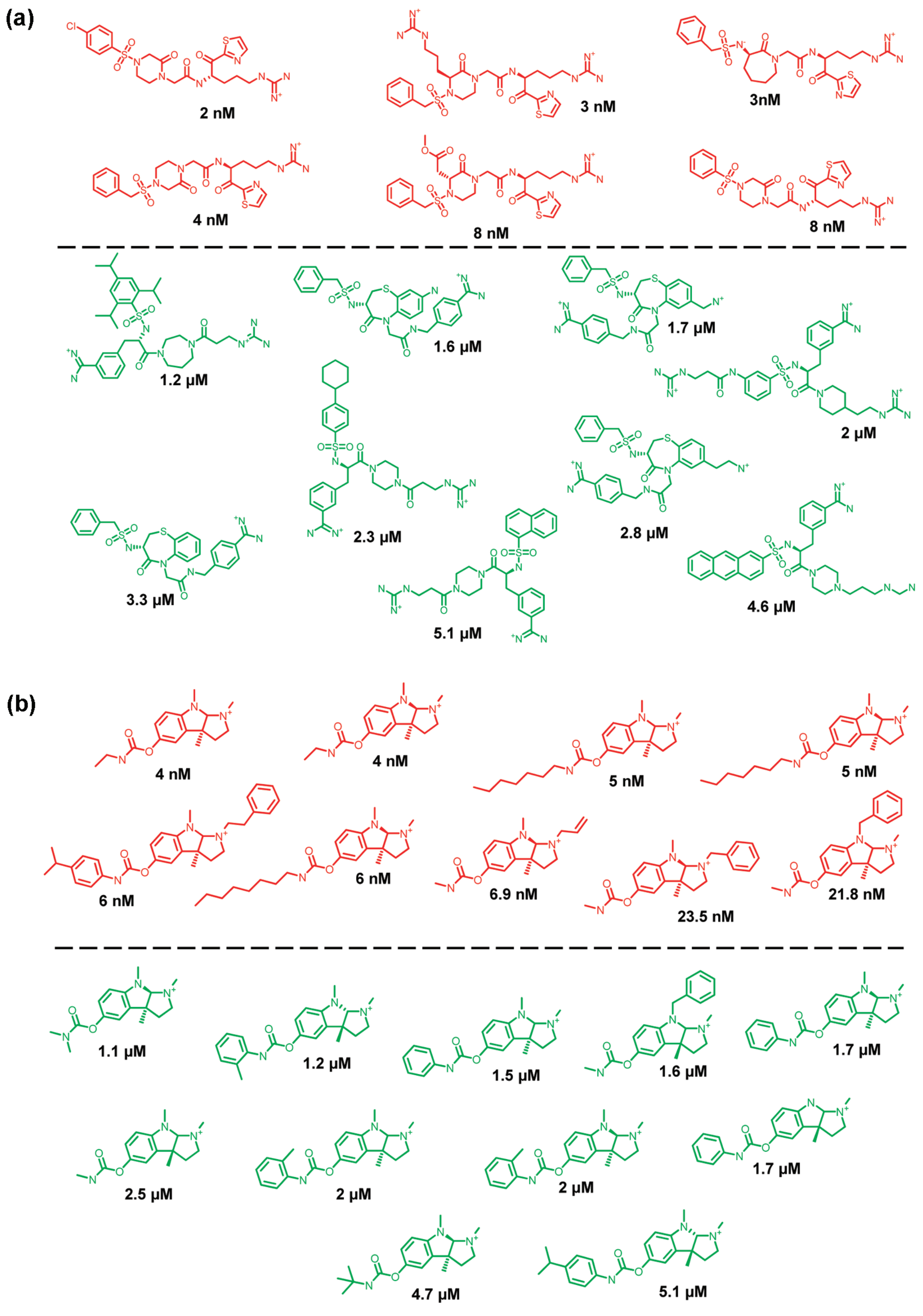


Figure 6. Selected compound subsets. Shown are exemplary high-scoring $n = 6$ compound subsets forming coordinated activity cliffs for inhibitors of (a) factors Xa and (b) butyrylcholinesterase. High and low potency compounds are colored red and green, respectively.

B, and Tyr phosphatase 1B, no coordinated activity cliffs were detected. Although the global SAR characteristics of the compound sets we investigated ranged from predominantly discontinuous over heterogeneous to predominantly continuous phenotypes, centers or at least "islands" of SAR discontinuity were detected in the majority of all sets.

Subset Sizes. The sizes of best-scoring subsets varied for different compound classes. For $n = 3$ and $n = 6$ subsets, up to 38 (factor Xa inhibitors) and 39 (protein kinase C α inhibitors) members were detected, respectively. In five cases, $n = 3$ PSO runs identified larger subsets than $n = 6$ calculations that contained more than 12 compounds, including compound sets where $n = 6$ calculations failed. Hence, it has sometimes been easier for particles to approach favorable solutions (and navigate through local minima) when required numbers of qualifying compounds were smaller (which corresponds to less constrained optimization condition). In calculations that succeeded to detect coordinated activity cliffs, the best-scoring

$n = 3$ and $n = 6$ subsets consisted of on average 12.2 and 18.1 compounds, respectively. For 26 of 32 classes, $n = 3$ or $n = 6$ subsets with more than 10 compounds were identified. The proportion of high- vs low-potency compounds per set often varied (and hence the total number of coordinated activity cliffs that were formed). On the basis of our calculation, larger numbers of compounds participated in formation of coordinated activity cliff data than we originally anticipated.

Subset Discontinuity Scores. Table 1 reports for each data set the highest scoring $n = 3$ and $n = 6$ compound subsets forming coordinated activity cliffs. For 21 data sets, both $n = 6$ and $n = 3$ subsets of strong discontinuity were identified (with subset discontinuity scores > 0.70). In three instances including inhibitors of MAP kinase p38 α, protein kinase C β, and carbonic anhydrase XII where no $n = 6$ subsets were found, strongly discontinuous $n = 3$ subsets were identified with scores of 0.9−1.0. Equivalent observations were made for three classes where only $n = 6$ subsets of low to moderate discontinuity were found including inhibitors of matrix metalloproteinase-3, caspase-1, and Tyr kinase TIE-2 (with subsets scores of 0.37, 0.40, and 0.56, respectively). Hence, in these cases, only $n = 3$ simulations identified strongly discontinuous subsets that typically consisted of fewer than 10 compounds. By contrast, for eight compounds sets (numbers 25−32 in Table 1), $n = 3$ and $n = 6$ subsets of nearly maximal or maximal discontinuity were identified (with scores of 0.98−1.00). One would expect that the distribution of compound subsets forming coordinated activity cliffs in different data sets should vary according to their specific SAR characteristics and that not all sets contain coordinated activity cliffs. However, as stated above, coordinated activity cliffs were detected in the majority of data sets, and the corresponding compound subsets were typically strongly discontinuous in nature.

Local vs Global SAR Features. It has also been of interest to relate the occurrence of coordinated activity cliffs to the global SAR phenotypes of compound classes (as indicated by global SARI scoring). The eight compound classes producing subsets of highest discontinuity (25−32 in Table 1) were either heterogeneous or discontinuous in their global SAR character (the latter category including Ser/Thr kinase AKT and factor Xa inhibitors). Hence, it should be expected that such compound classes are rich in activity cliffs. In fact, for globally discontinuous compound data sets, subsets forming coordinated activity cliffs were consistently identified. By contrast, the test cases for which $n = 3$ and $n = 6$ calculations did not identify coordinated activity cliffs were characterized by a high degree of global SAR continuity (numbers 1−3, 5, and 8 in Table 1). However, there also were notable departures from these general trends. For example, for several predominantly continuous classes including inhibitors of c-Jun N-terminal kinase-1, cyclooxygenase-2, and cytochrome P450 2D6, large and highly discontinuous compound subsets were identified that formed coordinated activity cliffs. These examples illustrate the local character of coordinated activity cliff structures; even globally continuous data sets might contain islands of activity cliff formation.

Examples. In Figure 6, representative examples of high-scoring coordinated activity cliffs are shown, as revealed by PSO analysis. The factor Xa inhibitor subset in Figure 6a consists of 6 high- and 9 low-potency compounds with large differences in potency. In this case, each high-potency molecule forms activity cliffs of more than or close to 2 orders of magnitude potency difference with all of its low-potency counterparts. Furthermore, the subset of butyrylcholinesterase inhibitors in Figure 6b contains 9 high- and 11 low-potency compounds. The distribution and magnitudes of coordinated activity cliffs formed among them are comparable to the factor Xa example. These examples illustrate the SAR information gain that is associated with analyzing coordinated activity cliffs rather than individual compound pairs. For example, the coordinated factor Xa activity cliffs in Figure 6a suggest that shape differences and the presence of condensated ring systems are responsible for dramatic reductions in inhibitor potency. This conclusion would be difficult to draw by inspecting individual compound pairs due to the presence of multiple changes that usually differentiate them. In addition, in cliff forming butyrylcholinesterase inhibitors in Figure 6b the position at the phenyl ring where substitutions apparently have major potency effects can be easily identified by inspecting coordinated activity cliffs. Surprisingly, however, many different types of substituents at this site have either positive or negative effects on compound potency. The cliffs also reveal that different ring stereoisomers are permitted in both high- and low-potency inhibitors. Thus, these observations indicate the presence of complex SAR patterns. In fact, considering the coordinated activity cliffs, one might formulate the hypothesis that combinations of certain ring stereoisomers and substitutions of different size at the phenyl moiety might determine the SAR of this compound series. Taken together, these examples illustrate the generally high SAR information content of coordinated activity cliff arrangements and rationalize their attractiveness for SAR analysis. The stereopair observations discussed above also lead to an important methodological point. The choice of different molecular representations (descriptors) is generally expected to change the results of compound similarity evaluation, which is one of two critical parameters for activity cliff analysis. Because structural keys used as descriptors for our analysis are stereoinsensitive, different isomers are considered identical, and hence, stereo-effects do not reduce compound similarity. Consequently, activity cliff formation is not affected, which might or might not be beneficial of SAR analysis, depending on whether stereo-effects are true SAR determinants. For the example in Figure 6b, using stereoinsensitive molecular representations for similarity assessment is suitable because the presence of different stereoisomers changes potency only in a few instances but not in many others. This is also nicely illustrated by analyzing the coordinated activity cliffs where different combinations of compounds can be identified having either the same or dramatically different potency that are identical except for ring stereopairs.

■ CONCLUSIONS

In this study, we adapted the PSO approach to systematically search compound data sets of different size, composition, and SAR character for the presence of coordinated activity cliffs. As introduced herein, the coordinated activity cliff data structure consists of multiple and overlapping activity cliffs and is formed by compounds with defined similarity relationships and potency differences. The adaptive machine learning approach applied in our analysis makes it possible to search for coordinated activity cliffs that introduce the strongest local SAR discontinuity in data sets and automatically extract the corresponding compound subsets from large data sets. These are key aspects of our study. In systematic search calculations, we identified strongly discontinuous compound subsets forming coordinated activity cliffs in the majority of test cases. On the basis of our analysis, coordinated activity cliffs frequently occur in compound data sets of different nature, even within largely continuous SAR environments. Taken together, these findings suggest further extending the activity cliff concept beyond compound pairs. As observed in previous investigations, isolated activity cliffs formed by pairs of compounds with no structural neighbors or SAR-insignificant neighbors frequently occur in data sets. However, it is also likely to find coordinated activity cliffs that often involve relatively large numbers of compounds. For SAR analysis, coordinated activity cliffs are of high interest because they are typically rich in SAR information. Hence, providing automated access to highly discontinuous compound subsets that form coordinated activity cliffs in different data sets should be helpful for systematic SAR exploration.

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) Maggiora, G. M. On Outliers and Activity Cliffs—Why QSAR often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

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

(3) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, in press; DOI: 10.1021/jm201706b.

(4) Wassermann, A. M.; Dimova, D.; Bajorath, J. Comprehensive Analysis of Single- and Multi-Target Activity Cliffs Formed by Currently Available Bioactive Compounds. *Chem. Biol. Drug Des.* **2011**, *78*, 224−228.

(5) Vogt, M.; Huang, Y.; Bajorath, J. From Activity Cliffs to Activity Ridges: Informative Data Structures for SAR Analysis. *J. Chem. Inf. Model.* **2011**, *51*, 1848−1856.

(6) Kennedy, J.; Eberhart, R. C. Particle Swarm Optimization. *Proceedings of the IEEE International Conference Neural Networks IV (ICN95)*; 1995; pp 1942−1948.

(7) Willett, P. Searching Techniques for Databases of Two- and Three-dimensional Structures. *J. Med. Chem.* **2005**, *48*, 1−17.

(8) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.

(9) Kennedy, J.; Eberhart, R. C. A Discrete Binary Version of the Particle Swarm Algorithm. *Proceedings of the World Multiconference on Systemics, Cybernetics, and Informatics*; 1997; pp 4104−4109.

(10) Agrafiotis, D. K.; Cedeno, W. Feature Selection for Structure-Activity Correlation Using Binary Particle Swarms. *J. Med. Chem.* **2002**, *45*, 1098−1107.

(11) Lü, J. X.; Shen, Q.; Jiang, J. H.; Shen, G. L.; Yu, R. Q. QSAR Analysis of Cyclooxygenase Inhibitors Using Particle Swarm Optimization and Multiple Linear Regression. *J Pharm. Biomed. Anal.* **2004**, *35*, 679−687.

(12) Namasivayam, V.; Günther, R. PSO@Autodock3: A Fast Flexible Molecular Docking Program Based on Swarm Intelligence. *Chem. Biol. Drug Des.* **2007**, *70*, 475−484.

(13) Hartenfeller, M.; Proschak, E.; Schüller, A.; Schneider, G. Concept of Combinatorial De Novo Design of Drug-like Molecules by Particle Swarm Optimization. *Chem. Biol. Drug Des.* **2008**, *72*, 16−26.

(14) Namasivayam, V.; Iyer, P.; Bajorath, J. Extraction of Discontinuous Structure-Activity Relationships from Compound Data Sets through Particle Swarm Optimization. *J. Chem. Inf. Model.* **2011**, *51*, 1545−1551.

(15) Namasivayam, V.; Iyer, P.; Bajorath, J. Exploring SAR continuity in the vicinity of activity cliffs. *Chem. Biol. Drug Des.* **2012**, *79*, 22−29.

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

(17) Shi, Y.; Eberhart, R. C. Parameter Selection in Particle Swarm Optimization. *Proceedings of the Seventh Annual Conference on Evolutionary Programming*; 1998; pp 591−600.

(18) Clerc, M. Stagnation Analysis in Particle Swarm Optimization or What Happens When Nothing Happens. *Technical Report CSM-460*; Department of Computer Science, University of Essex: 2006; ISSN: 1744-8050

(19) 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.

(20) Omran, M.; Engelbrecht, P.; Salman, A. Particle Swarm Optimization Method for Image Clustering. *Int. J. Pattern Recognit. Artif. Intell.* **2005**, *19*, 297−321.