JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Consensus Models of Activity Landscapes with Multiple Chemical, Conformer, and Property Representations

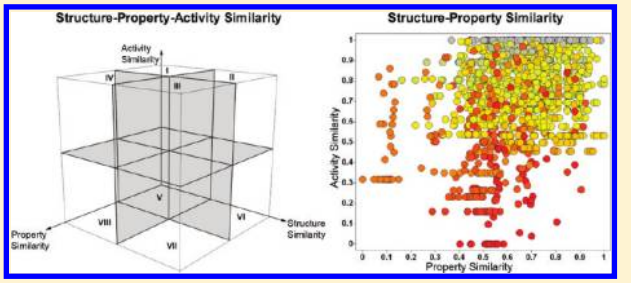
Austin B. Yongye,[†] Kendall Byler,[†] Radleigh Santos,[†] Karina Martínez-Mayorga,[†] Gerald M. Maggiora,[‡] and José L. Medina-Franco*,[†]

[†]Torrey Pines Institute for Molecular Studies, 11350 SW Village Parkway, Port St. Lucie, Florida 34987, United States

[‡]Department of Pharmacology & Toxicology, University of Arizona College of Pharmacy, 1703 E. Mabel Street, Tucson, Arizona 85721, United States and Translational Genomics Research Institute, 445 N. Fifth Street, Phoenix, Arizona 85004, United States

Supporting Information

ABSTRACT: We report consensus Structure−Activity Similarity (SAS) maps that address the dependence of activity landscapes on molecular representation. As a case study, we characterized the activity landscape of 54 compounds with activities against human cathepsin B (hCatB), human cathepsin L (hCatL), and *Trypanosoma brucei* cathepsin B (TbCatB). Starting from an initial set of 28 descriptors we selected ten representations that capture different aspects of the chemical structures. These included four 2D (MACCS keys, *GpiDAPH3*, *pairwise*, and *radial* fingerprints) and six 3D (*4p* and *piDAPH4* fingerprints with each including three conformers) representations. Multiple conformers are used for the first time in consensus activity landscape modeling. The results emphasize the feasibility of identifying consensus data points that are consistently formed in different reference spaces generated with several fingerprint models, including multiple 3D conformers. Consensus data points are not meant to eliminate data, disregarding, for example, "true" activity cliffs that are not identified by some molecular representations. Instead, consensus models are designed to prioritize the SAR analysis of activity cliffs and other consistent regions in the activity landscape that are captured by several molecular representations. Systematic description of the SARs of two targets give rise to the identification of pairs of compounds located in the same region of the activity landscape of hCatL and TbCatB suggesting similar mechanisms of action for the pairs involved. We also explored the relationship between property similarity and activity similarity and found that property similarities are suitable to characterize SARs. We also introduce the concept of structure−property-activity (SPA) similarity in SAR studies.



INTRODUCTION

Activity landscape modeling is an emerging concept to systematically characterize the structure−activity relationships (SAR) of compound data sets.[1,2] The activity landscape can be conceptualized as the chemical space with the addition of biological activity as another dimension.[3] Understanding the activity landscape and the early detection of *activity cliffs* (chemical compounds with highly similar structures but significantly different biological activities)[4] can be crucial to the success of predictive computational models.[5,6] Activity landscape modeling is, however, a challenging task due to the strong influence of molecular representation on chemical space.[7,8] Since different molecular descriptors capture different aspects of molecular structures,[9,10] each set of descriptors can give a different landscape. As a consequence, the discontinuous or continuous regions of the SAR may involve different groups of structures.[11,12] In order to reduce the dependence of activity landscape on chemical representation, the use of multiple molecular representations has been proposed.[12−14] The main goal is to identify regions in the landscape that are consistently formed in different reference spaces. Consensus models are designed to "prioritize" activity cliffs and other consistent regions in the landscape that are captured by several molecular representations.

Several methods have been developed to characterize activity landscapes.[1,3,15] One of the first methods proposed is the Structure−Activity Similarity (SAS) map.[16,17] SAS maps describe the relationship between structural similarity and activity similarity for all pairs of compounds in a data set. Figure 1 shows a prototype SAS map where the molecular and activity similarities are represented on the *X*- and *Y*-axes, respectively. Four major regions can be roughly distinguished in the SAS maps. Pairs of compounds that fall in region I have low structure similarity but high activity similarity and are indicative of scaffold or side chain (R-group) hopping.[18,19] Region II denotes pairs of compounds with both high structure and activity similarity and represents compounds in smooth or continuous SARs. Compounds in region IV have high structure similarity but low activity similarity and therefore correspond to activity cliffs or discontinuous SARs.

Received: February 16, 2011
Published: May 24, 2011

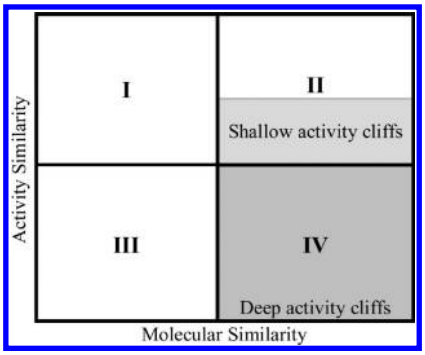


Figure 1. Prototype structure−activity similarity (SAS) map depicting four major segments. The regions provide the following information for pairs of compounds: Region I, scaffold or side chain hopping; Region II, smooth or continuous SAR; Region IV, discontinuous SAR and activity cliffs. Subsegments denoting deep and shallow activity cliffs are also shown (see text for details).

This quadrant is a region of *high SAR information content*[16,20] because it reveals structural patterns that are crucial for activity. Activity cliffs can be classified further as either shallow or deep.[12] Shallow activity cliffs are comprised of pairs of compounds displaying large (but not extreme) differences in activity as a result of slight changes in structure, and they occupy the borderline area between regions II and IV. Conversely, a deep activity cliff is formed by pairs of compounds with very high structure similarities but remarkably low activity similarities. Region III is perhaps the least interesting containing pairs of molecules with low molecular similarity and low activity similarity.

Herein, we explored the activity landscapes of 54 compounds with reported activities against three cysteine proteases: human cathepsin B (hCatB), human cathepsin L (hCatL), and *Trypanosoma brucei* (*T. brucei*) cathepsin B (TbCatB). Analysis of these compounds is part of our ongoing efforts to develop compounds to modulate the biological activity of cathepsins.[21] The overexpression of human cathepsins has been observed in pathological conditions such as cancer, atherosclerosis, arthritis, and osteoporosis (cathepsins B, F, K, and L). The protozoan parasite *T. brucei* is the pathogen that causes human African trypanosomiasis (HAT).[22] Presently few drugs are approved for the treatment of HAT and issues of toxicity limit their use.[23] Recent studies have shown that even a moderate inhibition of TbCatB is lethal for *T. brucei*.[24,25] Analysis of the activity landscape of the 54 compounds was based on 1431 pairwise relationships obtained with selected 2D and 3D structural representations as well as six physicochemical properties. The different landscapes for each target correspond to the different representations. For each target, pairwise SAR was characterized using SAS maps and quantitative analyses of the SAS maps were used to develop consensus models of the activity landscape. In addition, we compared the landscapes of two of the three targets with interesting activity profiles. The approaches described here are not restricted to the data set employed, molecular representations, or molecular targets. The emphasis in this work is on the comprehensive activity landscape modeling of compounds tested across different targets using multiple molecular representations, including multiple conformers.

METHODS

Data Set. The structures of the 54-compound data set along with their activities (pIC_{50}) against hCatB, hCatL, and TbCatB are presented in Table 1. The IC_{50} values were collected from ChEMBL[26] and were reported by the same group[27] under similar conditions.

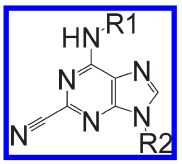
Fingerprint Representations and Structure Similarity. Twelve 2D and four 3D fingerprints were computed (Figure S1 in the Supporting Information). The 3D fingerprints included three- and four-point pharmacophores from Canvas[28] (*3p* and *4p*) and MOE[29] (*piDAPH3* and *piDAPH4*). Recently we showed that despite the inherent issues of employing a single conformer to represent 3D structures, the 3D fingerprints were valuable in characterizing activity landscapes.[12,13] In this work, the effects of 3D conformers on activity landscapes were investigated employing different input conformations for the 3D fingerprints. Initially, four sets of conformations were utilized, which were generated as follows: 1) *BO*, the conformer with best overlay with a compound in the database, which displayed the best activity toward hCatB, hCatL, and TbCatB; 2) *BO xtal*, the conformer with best overlay on a purine nitrile analogue cocrystallized with a related enzyme, cruzain;[30] 3) *LE*, the lowest energy conformer; and 4) *IP*: a SMILES-generated input conformer, i.e, the Daylight SMILES strings downloaded from ChEMBL were converted to 3D structures employing the Merck Molecular Force Field 94x in MOE. Of note, other approaches to generating 3D conformers can also be utilized.[31,32] For the first three approaches, the conformers were generated, and the overlays were performed using OMEGA[33] and the Rapid Overlay of Chemical Structures (ROCS) software,[34] respectively. These additional considerations gave rise to a total of twenty-eight structural representations (12 2D representations plus 16 (4 × 4) 3D representations, Figure S1) that were the starting point in this work. Structure similarities were computed with the Tanimoto coefficient,[35] which metric has been successfully applied in activity landscape modeling. However, other measures such as Euclidean distance can be used. For example, in a previous work, both Euclidean distance and the Tanimoto coefficient were used to model activity landscapes using three 2D fingerprint representations.[11] Authors of that work noted that the landscapes generated with Euclidean distance and Tanimoto similarity were often similar.[11] It is worth pointing out that the 54 compounds in this work have in common a purine nitrile scaffold, hence similarity determinants could be restricted to side chain (R-group) differences.[36,37] Despite the fact that whole-molecule similarity measures were used several fingerprint models could capture small structural differences and were suitable to model the activity landscape of these compounds. It remains to explore the activity landscape of the 54 compounds with other measures such as Euclidean distance to evaluate the further differentiation of the molecular structures.

Activity Similarity. The pairwise activity similarities were measured separately for the three targets employing the following equation used in earlier studies[12,13]

$$AS_{i,j} = 1 - \frac{|A_i - A_j|}{\max - \min} \quad (1)$$

where A_i and A_j are the activities (pIC_{50}) of the *i*th and *j*th compounds, respectively, and max-min is the activity range of the data set for each target. The activity similarity values obtained with eq 1 are in the same scale of the similarity values. Of note, the activity range across the targets of the data set considered in this work is similar (Table 1). Activities can also be compared by computing the absolute difference between the pIC_{50} values of the pair. This approach, independent of the activity range of the

Table 1. Structures and pIC_{50} Values of the Compounds Employed in This Study[a]



compound	R1	R2	pIC_{50} (hCatB)	pIC_{50} (hCatL)	pIC_{50} (TbCatB)
1	2-methylbutyl	benzyl	4.52	5.23	5.30
2	cyclohexyl	propyl	4.72	5.60	5.08
3	cyclohexyl	cyclohexyl	4.63	5.48	5.08
4	cyclohexyl	benzyl	4.75	5.42	5.64
5	hexyl	butyl	4.52	4.63	4.52
6	hexyl	cyclopentyl	4.52	4.55	4.52
7	hexyl	cyclohexyl	4.52	4.52	4.52
8	hexyl	benzyl	4.52	4.52	4.52
9	α-methylbenzyl	butyl	4.52	5.03	4.52
10	α-methylbenzyl	cyclopentyl	4.52	4.93	4.52
11	α-methylbenzyl	benzyl	4.52	4.52	4.52
12	4-MeO-benzyl	cyclopentyl	4.52	5.40	5.15
13	3-Me-benzyl	butyl	4.52	5.41	4.96
14	3-Me-benzyl	cyclopentyl	4.52	4.98	4.52
15	3-Me-benzyl	cyclohexyl	4.52	4.52	4.52
16	3-Me-benzyl	benzyl	4.52	4.98	5.48
17	benzyl	butyl	4.52	5.36	4.85
18	benzyl	cyclopentyl	4.52	4.87	4.52
19	benzyl	benzyl	4.52	4.91	5.22
20	4-Cl-benzyl	butyl	4.52	5.30	5.00
21	4-Cl-benzyl	cyclopentyl	4.52	4.87	4.72
22	4-Cl-benzyl	cyclohexylmethyl	4.52	4.52	4.52
23	4-Cl-benzyl	benzyl	4.52	4.52	4.52
24	phenylethyl	pyridine-3-ylmethyl	4.52	4.77	5.00
25	phenylethyl	butyl	4.52	4.77	4.80
26	phenylethyl	benzyl	4.52	4.52	4.52
27	4-MeO-benzyl	butyl	4.52	5.64	5.22
28	4-MeO-benzyl	benzyl	4.52	5.09	5.35
29	4-Me-benzyl	3-hydroxypropyl	5.38	5.11	5.38
30	4-Me-benzyl	butyl	4.52	5.23	4.52
31	4-Me-benzyl	benzyl	4.52	4.89	4.52
32	benzyl	3-hydroxypropyl	4.52	5.64	4.92
33	benzyl	tetrahydropyran-2-ylmethyl	4.52	4.97	4.96
34	benzyl	4-MeO-benzyl	4.52	4.57	4.96
35	benzyl	3-CF_3-benzyl	4.52	4.52	5.55
36	4-F-benzyl	4-hydroxybutyl	4.52	5.18	4.74
37	4-F-benzyl	butyl	4.52	5.16	5.15
38	4-F-benzyl	benzyl	4.52	4.87	5.48
39	4-Cl-benzyl	pyridine-3-ylmethyl	4.52	4.52	4.60
40	4-Cl-benzyl	3-hydroxypropyl	5.98	6.21	6.24
41	4-Cl-benzyl	4-hydroxybutyl	4.52	5.38	5.32
42	4-Cl-benzyl	4-MeO-benzyl	4.52	4.52	4.64
43	4-Cl-benzyl	3-CF_3-benzyl	4.52	4.52	4.52
44	4-Br-benzyl	butyl	4.52	5.02	5.08
45	4-Br-benzyl	benzyl	4.52	4.54	4.52
46	4-Br-benzyl	3-phenylpropyl	4.52	4.52	4.52
47	3,4-Cl_2-benzyl	2-hydroxyethyl	4.52	6.52	5.85
48	3,4-Cl_2-benzyl	butyl	4.52	5.74	5.39

dx.doi.org/10.1021/ci200081k |*J. Chem. Inf. Model.* 2011, 51, 1259–1270

Table 1. Continued

compound	R1	R2	pIC_{50} (hCatB)	pIC_{50} (hCatL)	pIC_{50} (TbCatB)
49	3,4-Cl_2-benzyl	3-hydroxypropyl	5.27	6.52	6.57
50	3,4-Cl_2-benzyl	2,3-dihydroxypropyl	4.52	6.38	5.92
51	3,4-Cl_2-benzyl	5-hydroxypentyl	4.52	6.10	6.03
52	3,4-Cl_2-benzyl	2-(methoxymethoxy)ethyl	4.52	6.06	6.10
53	3,4-Cl_2-benzyl	3-methoxypropyl	4.52	6.11	5.60
54	3,4-Cl2-benzyl	3,3-dimethoxypropyl	4.52	6.30	5.47

[a] hCatB: human cathepsin B; hCatL: human cathepsin L; TbCatB: *Trypanosoma brucei* cathepsin B.

Table 2. Correlation Matrix between Selected Molecular Representations

	4p BO	4p LE	4p IP	piDAPH4 BO	piDAPH4 LE	piDAPH4 IP	Radial	MACCS moe	Pairwise	GpiDAPH3	PropSim
4p BO	1.00										
4p LE	0.75	1.00									
4p IP	0.66	0.66	1.00								
piDAPH4 BO	0.08	0.05	0.04	1.00							
piDAPH4 LE	0.07	0.09	0.06	0.74	1.00						
piDAPH4 IP	0.05	0.05	0.05	0.67	0.67	1.00					
Radial	0.42	0.66	0.32	0.05	0.10	0.03	1.00				
MACCS moe	0.16	0.16	0.18	0.44	0.55	0.42	0.01	1.00			
Pairwise	0.53	0.62	0.42	0.16	0.16	0.11	0.73	0.13	1.00		
GpiDAPH3	0.10	0.08	0.08	0.75	0.78	0.68	0.01	0.64	0.16	1.00	
PropSim	0.52	0.51	0.54	0.01	0.05	0.03	0.23	0.22	0.33	0.08	1.00

data set, gives rise to a modified version of the SAS map that has been used previously to model activity landscapes.[13]

Property Similarity. The following six drug-like physicochemical properties were computed in MOE: octanol/water partition coefficient (SlogP), topological polar surface area (TPSA), molecular weight (MW), counts of hydrogen bond acceptors (HBA) and hydrogen bond donors (HBD), and the number of rotatable bonds (RB). Properties were first autoscaled with mean centering over all the molecules being considered using the equation

$$p_{ki} = \frac{P_{ki} - \overline{P_k}}{\sigma_{P_k}} \quad (2)$$

where p_{ki} denotes the scaled version of the kth property for the ith molecule, P_{ki} denotes the unscaled value, and $\overline{P}_k$ and σ_{P_k} denote, respectively, the mean and standard deviation of the kth property over all molecules in the study.

The Euclidean distance (d_{ij}) between a pair of compounds (i and j) was computed using the following equation[38]

$$d_{ij} = \left[\sum_{k=1}^{K} (p_{ki} - p_{kj})^2\right]^{1/2} \quad (3)$$

where p_{ki} and p_{kj} represent the value of property k. Here, $K = 6$ for the six physicochemical properties. Next, the Euclidean distances were scaled to values between 0 and 1 employing the following expression[12]

$$sd_{ij} = \frac{d_{ij} - \min d_{ij}}{\max d_{ij} - \min d_{ij}} \quad (4)$$

where sd_{ij} is the scaled distance, while min d_{ij} and max d_{ij} specify the range of the computed distances. Lastly, the pairwise property similarities (PS_{ij}) were determined from the scaled distances using the equation[12]

$$PS_{ij} = 1 - sd_{ij} \quad (5)$$

Activity Landscape. For each target, SAS maps were generated employing 1431 pairwise activity and molecular similarities computed from selected structural representations (*vide infra*). To perform quantitative comparisons the SAS maps were partitioned into four regions using thresholds for structure and activity similarity (Figure 1). It should be noted that the molecular similarity thresholds will vary across representations (*vide infra*). In this study the thresholds were defined employing the mean structure similarity for each fingerprint and the mean activity similarity for each target. However, other thresholds could be applied. For quantitative comparisons, the Degree of Consensus (DoC)[13] between two methods was computed employing the expression

$$DoC^{R}_{m,n} = \frac{Cp_{m,n}}{p_m + p_n - Cp_{m,n}} \quad (6)$$

where $Cp_{m,n}$ is the number of pairs placed in the same region by both methods. p_m and p_n are the pairs allocated to the region under consideration by methods m and n, respectively. *DoC* expresses the extent to which two methods, m and n, consistently place identical compound pairs in the same region (R = I − IV) of a SAS map.

Consensus SAS Maps. These maps were generated independently for each protease by combining the pairwise similarities of the orthogonal fingerprints through mean fusion, i.e., computing the mean similarity of the selected fingerprints. Other measures of combining pairwise similarities may also be explored.

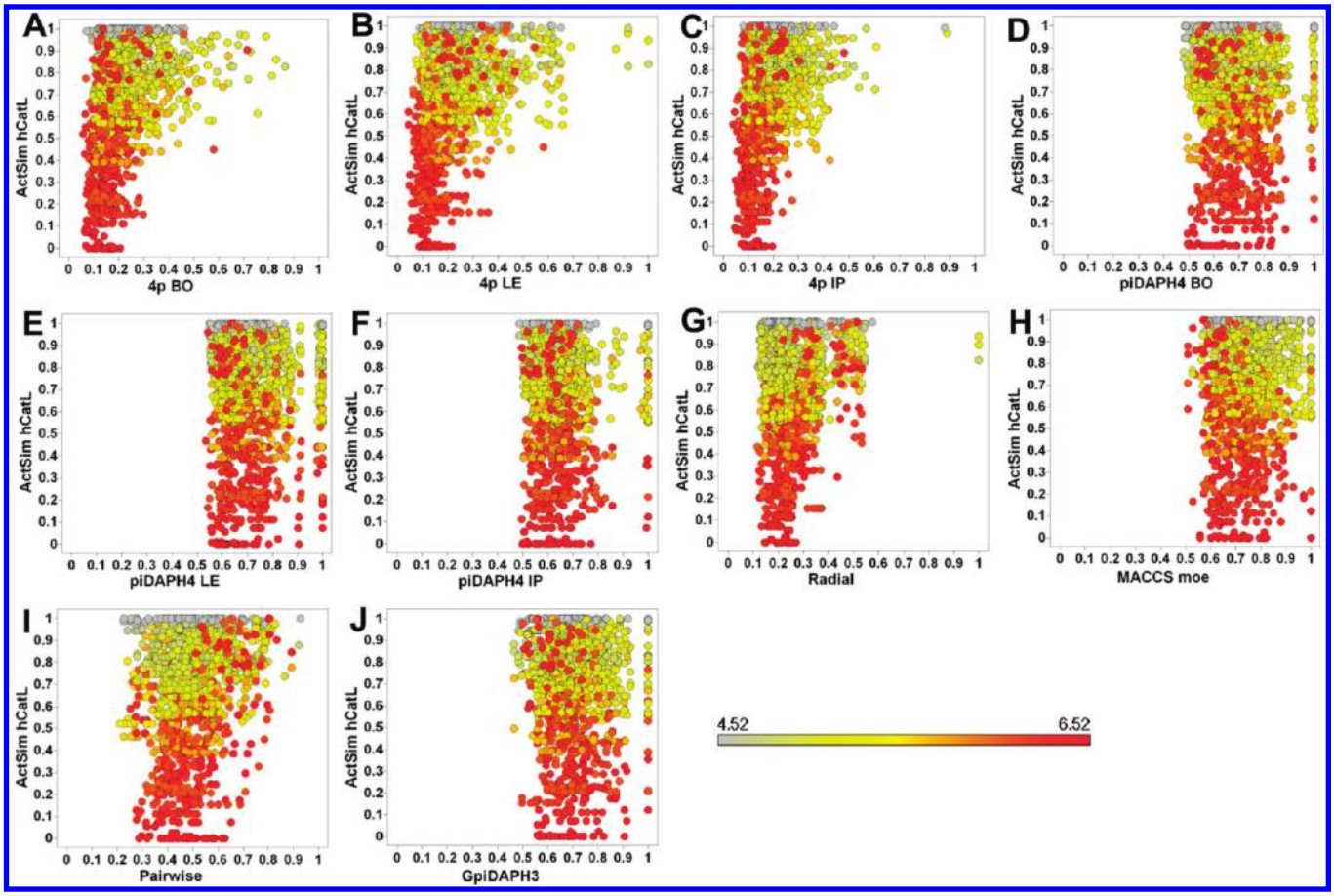


Figure 2. SAS maps for hCatL depicting 1431 pairwise comparisons for different representations of the 54 compounds. Data points are colored based on the activity of the more active member in the pair. Gray: inactive, yellow: moderately active, red: active. Each panel corresponds to a different structural representation: *4p* fingerprints using (A) best overlay (BO) to a compound with best activity against hCatB, hCatL, and TbCatB, (B) the lowest energy (LE) conformer of each ligand and (C) the input conformer (IP); *piDAPH4* using (D) best overlay (BO) to a compound with best activity against all hCatB, hCatL, and TbCatB, (E) the lowest energy (LE) conformer of each ligand and (F) the input conformer; (G) *Radial* fingerprints; (H) *MACCS* keys (166-bits); (I) *Pairwise*; (J) *GpiDAPH3*.

RESULTS AND DISCUSSION

Distribution of Pairwise Similarities for Different Fingerprints. The cumulative distribution functions (CDF) and their corresponding statistics for 2D and 3D representations derived from the 1431 pairwise similarities are shown in Figure S1 (Supporting Information). *TGD* and *TGT* fingerprints were associated with the highest mean and median values displaying the lowest resolution.[12,13] The lowest mean and median similarity values were obtained with the *4p* fingerprint, which also displayed the largest range. Despite the fact that the 54 compounds have a common scaffold, several fingerprint representations could differentiate the structures, indicating they can be employed in evaluating the SARs of analogous series.

Correlations between Fingerprints. A correlation matrix was constructed from the 1431 pairwise similarities to determine the degree of association between the twenty-eight 2D and 3D representations (data not shown). In general, the correlations were weak between the 2D and 3D fingerprints, highlighting the different structural information that are encoded in 2D and 3D representations.[9,10] Exceptions were *GpiDAPH3* (2D) and *piDAPH3* (3D) that exhibited correlations between 0.80 and 0.86 for any of the four different 3D conformations. This correlation indicates that these specific three point pharmacophores derived from the molecular graph or the 3D conformations, respectively, captures essentially the same information. Tables S1 and S2 (Supporting Information) show the correlation matrices of the Pearson's correlation coefficient between the 1431 pairwise similarities for each pair of 2D and 3D representations. In both categories the fingerprints displayed high intracorrelations but weak or no intercorrelations. We also examined the degree of correlation between the four different conformers represented by each 3D fingerprint. For several 3D fingerprints *LE* and *BO xtal* conformers showed high correlation indicating that the lowest energy conformer is similar to the conformer obtained through the best overlay with a cocrystallized compound (see Methods).

Selection of Fingerprints. We selected structural representations as different as possible with correlations *lower than* 0.80 (considered in this and previous works[12] as an arbitrary "high" correlation). Also, we seek to choose representations that capture different aspects of the molecular structures.[9] In this work, we selected a total of ten structural representations including four 2D fingerprints (MACCS keys (166-bits) from MOE, *GpiDAPH3*, *pairwise*, *radial*) and two 3D fingerprints (*4p*, *piDAPH4*) with each including three conformers, *BO*, *LE*, and *IP*. The correlation matrix between the ten representations is

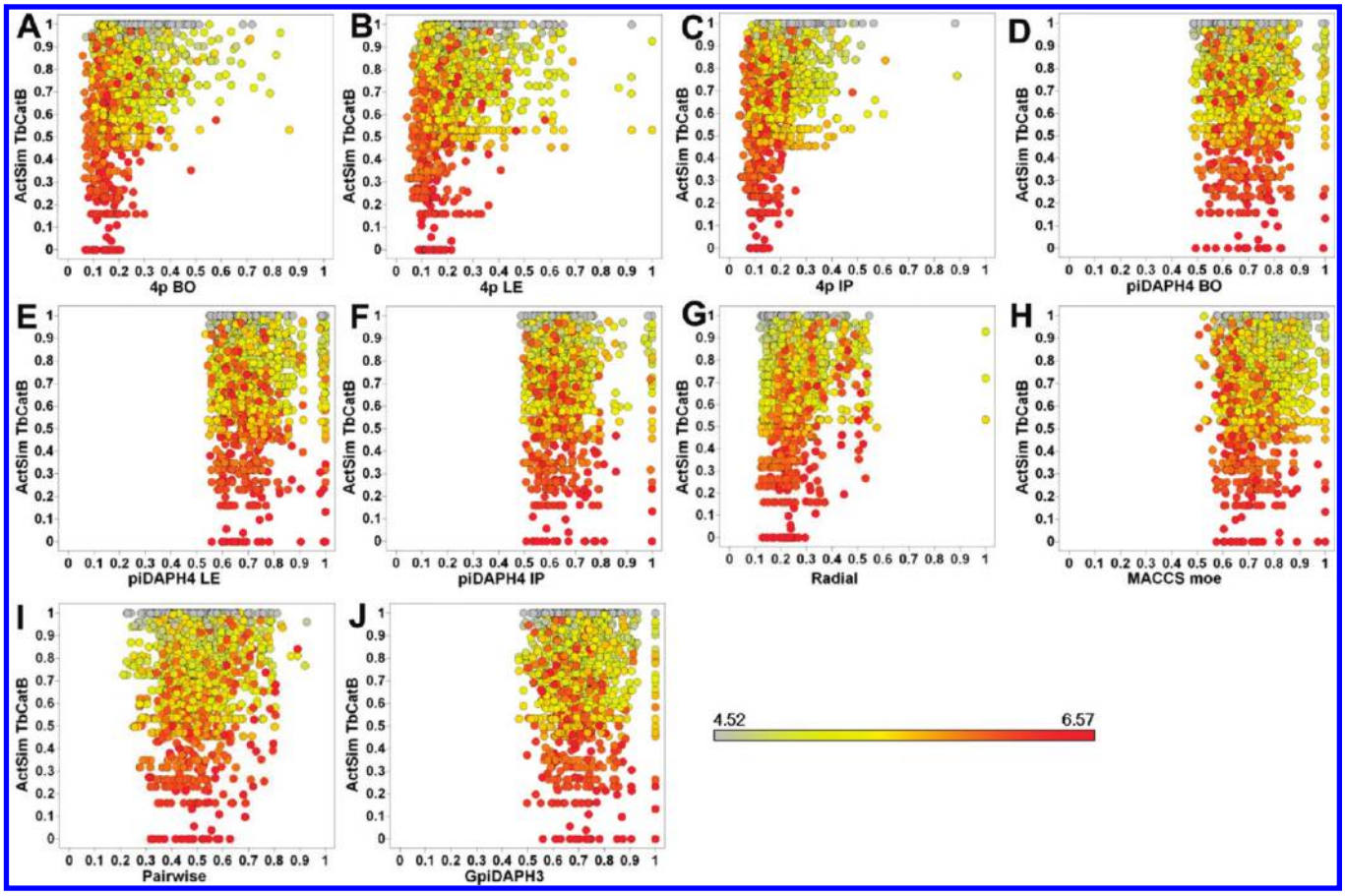


Figure 3. SAS maps for TbCatB depicting 1431 pairwise comparisons for different representations of the 54 compounds. Data points are colored based on the activity of the more active member in the pair. Gray: inactive, yellow: moderately active, red: active. Each panel corresponds to a different structural representation: *4p* fingerprints using (A) best overlay (BO) to a compound with best activity against hCatB, hCatL and TbCatB, (B) the lowest energy (LE) conformer of each ligand and (C) the input conformer (IP); *piDAPH4* using (D) best overlay (BO) to a compound with best activity against all hCatB, hCatL and TbCatB, (E) the lowest energy (LE) conformer of each ligand and (F) the input conformer; (G) *Radial* fingerprints; (H) *MACCS* keys (166-bits); (I) *Pairwise*; (J) *GpiDAPH3*.

presented in Table 2 which shows that the *maximum correlation between any of the representations is 0.78*. In addition, selected fingerprints have conceptually different designs capturing dissimilar aspects of the chemical structures. For example, MACCS keys used in this work are a predefined set of 166 structural keys; *GpiDAPH3* fingerprints are graph-based three-point pharmacophores employing any set of three possible atom types (pi system, donor, acceptor); *pairwise* are fingerprints based on the concept of atom pairs;[39] *radial* fingerprints entail growing a set of fragments radially from each heavy atom over a series of iterations.[39,40] *Radial* fingerprints are equivalent to the extended connectivity fingerprints (ECFPs).[41] Among the 3D fingerprints, *4p* and *piDAPH4* were selected over *3p* and *piDAPH3*, respectively because of their lower correlation with other fingerprints. Since high correlations were observed between the crystal and low energy structures (*vide supra*), the low energy structure representation was selected given that crystal structures may not always be available (considering this method can be generalized to other biological end points).

Structure−Activity Landscapes. Representative SAS maps (for hCatL and TbCatB) generated with the ten selected structural representations are shown in Figures 2 and 3, respectively. The SAS maps for hCatB are in Figure S2 (Supporting Information). In each map, there are 1431 data points color-coded by the activity of the more active compound in the pair: gray (inactive), yellow (moderately active), and red (active). The mean pIC_{50} for each target was used as the threshold for activity. Overall, data points have different distributions for the SAS maps obtained with different representations. There are few pairs of compounds (<30) with identical structure similarities (1.0) in the SAS maps obtained with *piDAPH4 BO*, *piDAPH4 LE*, *piDAPH4 IP*, *MACCS*, and *GpiDAPH3* (Figures 2/3 D, E, F, H, and J, respectively). The maps for the *4p LE* (Figures 2/3 B) and *radial* fingerprints (Figures 2/3 G) also exhibited pairs of compounds with similarity equal to one, albeit to a much lesser extent. These suggested a higher resolution of structural data by the *4p* (3D) and *radial* (2D) fingerprints.

Quantitative Analyses of the SAS Maps. To perform quantitative characterizations, the maps were divided into four regions (Figure 1). The mean molecular similarity for each representation was used to define the similarity thresholds, while the mean activity similarity of the 54 compounds was employed to delineate the activity boundary for each target. It is worth noting that other measures can be employed.[12,13] Statistical measures for the entire database as well as only for active compounds per target are shown in Table S3 (Supporting Information).

The distribution of molecule pairs among the four regions of the SAS maps for hCatL and TbCatB is shown in Table 3 (the distribution of pairs for hCatB is shown in Table S4 of the Supporting

Table 3. Distribution of Data Points Amongst the Different Segments of the SAS Maps for hCatL and TbCatB

		region I		region II		region III		region IV	
hCatL	threshold	total	active pairs	total	active pairs	total	active pairs	total	active pairs
4p BO	0.22	399	181	438	191	462	462	132	132
4p LE	0.22	402	177	435	195	451	451	143	143
4p IP	0.18	369	173	468	199	437	437	157	157
piDAPH4 BO	0.68	502	210	335	162	281	281	313	313
piDAPH4 LE	0.71	512	241	325	131	337	337	257	257
piDAPH4 IP	0.64	502	209	335	163	314	314	280	280
Radial	0.26	520	199	317	173	456	456	138	138
MACCS moe	0.75	458	233	379	139	364	364	230	230
Pairwise	0.50	399	150	438	222	389	389	205	205
GpiDAPH3	0.70	441	179	396	193	320	320	274	274
PropSim	0.22	267	134	570	238	379	379	215	215
MeanSim	0.50	476	207	361	165	442	442	152	152
		region I		region II		region III		region IV	
TbCatB	threshold	total	active pairs	total	active pairs	total	active pairs	total	active pairs
4p BO	0.22	427	220	395	167	434	434	175	175
4p LE	0.22	431	216	391	171	422	422	187	187
4p IP	0.18	397	209	425	178	409	409	200	200
piDAPH4 BO	0.68	469	195	353	192	314	314	295	295
piDAPH4 LE	0.71	502	241	320	146	347	347	262	262
piDAPH4 IP	0.64	478	209	344	178	338	338	271	271
Radial	0.26	548	245	274	142	428	428	181	181
MACCS moe	0.75	441	246	381	141	381	381	228	228
Pairwise	0.50	437	217	385	170	351	351	258	258
GpiDAPH3	0.70	440	220	382	167	321	321	288	288
PropSim	0.22	300	148	522	239	346	346	263	263
MeanSim	0.50	499	248	323	139	419	419	190	190

Information). The number of *active pairs* in each region is also included. In this work an active pair is defined as a data point in which at least one of the compounds had a pIC_{50} value greater than the mean pIC_{50} for the corresponding target. Categorical representations of the SAS maps are in Figures S3−S5 (Supporting Information). Overall, and across the three targets, there was a tendency of the majority of data pairs to group in region I, followed by region II, while region IV had the fewest number of pairs. The frequencies of the active pairs for hCatL were generally the following: III > II ≈ I > IV. The frequencies of the active pairs in TbCatB had the general trend III > IV ≈ I > II. For hCatB, the trend for the active pairs was the following: III > IV > I > II.

Degree of Consensus. The consensus pairs for the four regions of the landscapes of the three targets are summarized in Tables S5−S7 (Supporting Information). From these distributions a DoC matrix can be computed, which indicates the extent to which the similarity measures put the same pair of compounds in the same region of the SAS map.[12,13] The corresponding DoC matrices for the major regions of the landscapes are shown in Tables S8−S10 (Supporting Information). The DoC was generally low between 2D and 3D models. The DoC followed the order III > I > II > IV, from highest to lowest. For each region, the *4p* conformers generally displayed the highest DoC, except in region IV. Besides the preference for the models to collectively partition compound pairs in region III, these trends also indicated that the molecular representations were more likely to classify the compounds as side chain hops. Since the DoC (eq 6) has the form of the Jaccard index, the statistical significance of DoCs values was assessed based on the calculation of significant values of the Jaccard index studied in detail by Real and Vargas.[41,42] Typically, at 5% significance two methods with a DoC greater than approximately 0.45 would have more agreement than we would expect at random. Also, two methods with a DoC of less than approximately 0.20 would have more disagreement than we would expect at random. The exact critical values will depend on the total number of pairs ($p_n + p_m - Cp_{n,m}$) and can be estimated with equations shown in the Supporting Information. Based on these values at 5% significance, approximately 50% of the DoC values in region I and 31% of the DoC values in region II are higher than random for all three targets. In contrast, about 24% of the DoC values in region IV are higher than random. In region III, about 92% of the DoCs are higher than random for hCatB and hCatL (Tables S8−S10).

Specific All-Fingerprint Consistent Pairs and Their SARs. Despite the overall different distribution of data points in the SAS maps and low DoC, it was possible to identify consensus data pairs in several regions of the landscape for all three targets. Consensus data pairs are particularly valuable in the analysis of SAR because these points are consistently formed in different reference spaces, *reducing* the dependence of the SAR on

dx.doi.org/10.1021/ci200081k |J. Chem. Inf. Model. 2011, 51, 1259–1270

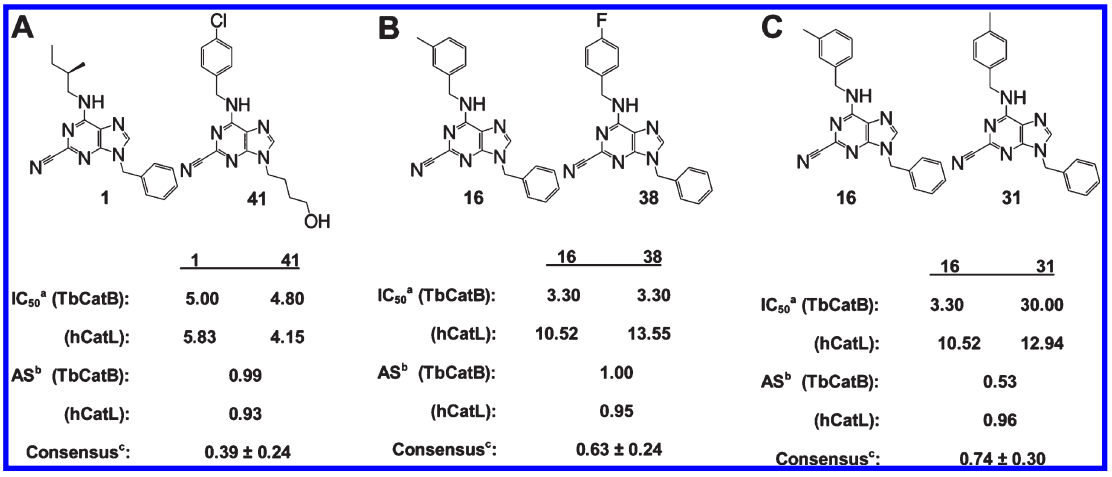


Figure 4. Representative chemical structures of pairs of compounds assigned to the three most informative regions of the SAS maps of TbCatB. The values for hCatL have been included for completeness. (A) Region I: side chain hops, i.e., different R1, R2 groups with *low* structure similarity but high activity similarity; (B) Region II: smooth SAR region displaying both high activity and structure similarities; (C) Region IV: activity cliff region, i.e., high structure similarity but very low activity similarity. [a] IC_{50} values in μM. [b] Activity similarity. [c] Consensus structural similarity computed by mean fusion of ten selected representations.

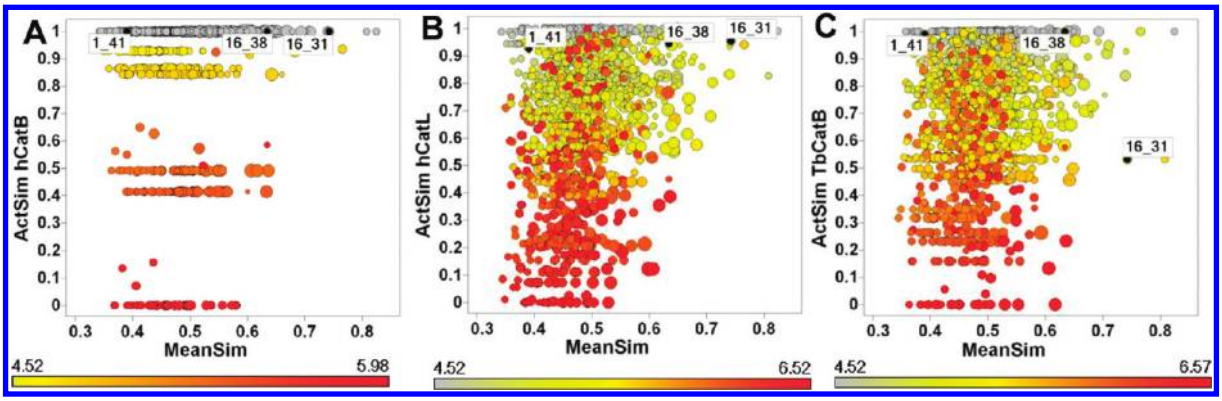


Figure 5. Consensus SAS maps generated with the mean of the similarity values of ten selected representations. Data points are sized by standard deviation and colored based on the activity of the more active member in the pair (gray: inactive, yellow: moderately active, red: active). Each panel corresponds to a different target: (A) hCatB; (B) hCatL; (C) TbCatB (see text for details).

chemical representation. Therefore, one can prioritize the analysis of the SAR based on these consistent pairs. The total number of pairs and active pairs grouped by all the fingerprint models into the three most informative regions of the SAS maps is given in Table S11 (Supporting Information). For each target, several consistent pairs were located in the different regions of the SAS maps. Overall, the frequency of the consistence was I > II > IV, which mirrored the trend observed in the pairwise DoC. Because of the high number of inactive compounds reported for hCatB, the discussions for the remainder of the manuscript will be focused on hCatL and TbCatB. Of note, the activity range of these two targets is the same (Table 1). Therefore, the activity similarity values computed with eq 6 for hCatL and TbCatB are scaled by the same value.

A representative subset of consistent pairs identified by all the fingerprint models for hCatL and TbCatB is presented in Table S12 (Supporting Information), including their activity-, structure-, property-, and mean structural similarities (see Methods). For hCatL, four compounds in region I involved **1** and the following compounds: **12**, **29**, **36**, and **41**. In the case of TbCatB, **1** was paired with four other compounds with high activity similarity: **12**, **29**, **41**, and **54**. For both targets, these compounds exemplified the concept of R-/side chain hopping. A striking example is for the **1_41** pair for TbCatB (Figure 4A) wherein the *4p* fingerprint representations had very low structure similarity (between 0.09 and 0.12), while the activity similarity was 0.99. Table S12 includes pairs of compounds that were located in region I of both hCatL and TbCatB: **1_12** and **28_29**, suggesting a similar SAR between the targets for this set of compounds. Additional pairs common to region I of both targets are shown in Table S12.

Exemplary pairs in region II for hCatL are **2_3** and **28_38** against TbCatB (Table S12). Pairs **16_38** and **2_3** were identified in region II of both hCatL and TbCatB. Compounds **16** and **38** (Figure 4B) differ by the presence of a 3-methyl versus a 4-fluoro substituent, respectively, while the difference between compounds **2** and **3** is the presence of a propyl and a cyclohexyl group, respectively. Of note, *MACCS* and *GpiDAPH3* (2D) and *piDAPH4* (3D) fingerprints could not distinguish between this pair, while they could be resolved with *radial* and *pairwise* (2D) and *4p* (3D) fingerprints. Similar observations were reported previously for other biological end points[12] thus underscoring the benefits of employing multiple fingerprint representations to characterize activity landscapes.

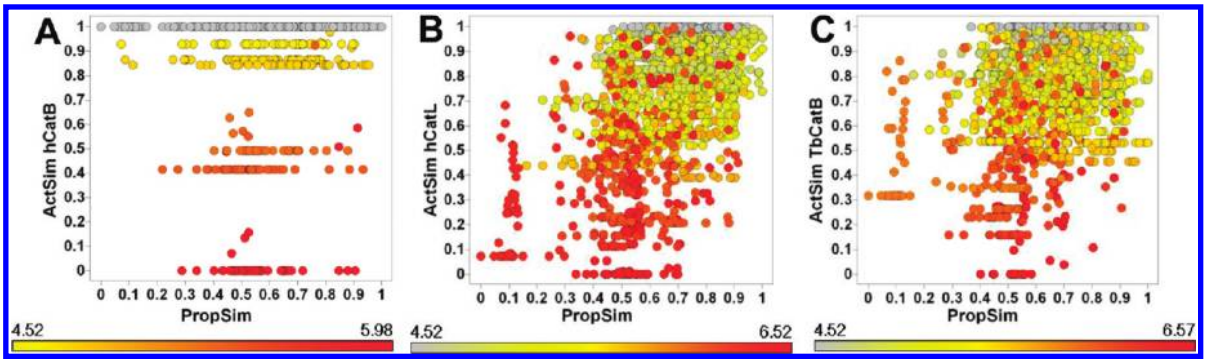


Figure 6. Property-activity similarity maps generated using six physicochemical properties to compute molecular similarity: (HBA, HBD, MW, SlogP, RB, and TPSA). Data points are colored based on the activity of the more active member in the pair (gray: inactive, yellow: moderately active, red: active). Each panel corresponds to a different target: (A) hCatB; (B) hCatL; (C) TbCatB.

Consensus Models of Activity Landscapes. We have proposed that the concept of data fusion[43,44] can be used to generate robust activity landscapes by combining information captured by different molecular representations.[12] Consensus SAS maps for the three targets are shown in Figure 5. Compared to the individual fingerprint SAS maps, the combined structure similarities by mean fusion were located at the center of the maps. The degree of consensus between the individual fingerprints and their mean pairwise similarity is shown in Tables S8−S10 (Supporting Information). These trends echoed those discussed in the "Degree of Consensus" section above for the individual fingerprints. The fused similarity representation successfully captured all compound pairs with at least one active compound in the three most informative regions of the SAS maps identified collectively by all the fingerprints.

Apparent vs Consensus Cliffs and Deep vs Shallow Cliffs. Pairs of compounds that fall in region IV of SAS maps provide pertinent information about molecular features that significantly affect activity.[4] In the course of identifying all-fingerprint consistent activity cliffs, it was customary to observe pairs of compounds that were detected as activity cliffs by one fingerprint representation, but not the others. These compound pairs have been defined previously as *apparent cliffs*,[11,12] the occurrence of which can be deduced from the low DoC for region IV among the targets (*vide supra*) (Tables S8−10). Nonetheless, several activity cliffs were identified for both hCatL and TbCatB by all the fingerprint representations (summarized in Table S12). We want to emphasize that consensus activity cliffs are not meant to disregard cliffs identified by just some representations as these can be "true" cliffs. Consensus cliffs are distinguished in that they are consistently formed in different reference spaces.

The activity cliffs identified in the SAS maps of TbCatB by all representations revealed several patterns. Pairs **1_8** and **1_26** suggest that hexyl and phenylethyl groups at position R1 were undesirable for activity toward TbCatB, consistent with the experimental trends.[27] The cliff between **16_31** (Figure 4C) pointed to the preference of a 3-methylbenzyl over a 4-methylbenzyl group at position R1. The activity cliffs formed by **23_38** and **28_39** indicated that besides fluorine, the other halides on 4-benzyl at position R1 will require less bulky groups at R2. For hCatL compound **1** formed activity cliffs with compounds **8** and **26**. Structurally, compound **1** has a 2-methylbutyl group at position R1 of the purine ring, while in compounds **8** and **26** position R1 contains hexyl and a benzyl group, respectively. Looking further at the activity cliff between **2** and **5** suggested that hexyl moiety will be detrimental to the inhibitory properties of these compounds against hCatL.

Two activity cliffs, **1_8** and **1_26**, were common to both hCatL and TbCatB with compound **1** being the most active in each pair suggesting a similar mechanism of action for these proteases. This is in line with close similarity between the S2 sites of TbCatB and hCatL as well as the general lack of selectivity for the two targets.[27] Nonetheless, the pair **16_31** that was in the smooth SAR region for hCatL was an activity cliff for TbCatB and thus constituted a selectivity cliff.[1,2]

We also explored the presence of shallow cliffs. The shallow cliff region in a SAS map is shown in Figure 1 and includes pairs in which the activity of one of the compounds is close to the activity-defining threshold. The new activity similarity thresholds employed were 0.77 and 0.82 for hCatL and TbCatB, respectively, i.e., the mean+0.1. For hCatL we identified four shallow cliffs listed in Table S13 (Supporting Information). No shallow cliffs were identified for TbCatB.

Rationale for Multiple 3D Conformers. A single conformer is typically selected to represent each compound in activity landscape modeling. However, the conformational landscape of flexible compounds is complex and populated by a large number of conformers.[45] To explore the effect of 3D conformation on activity landscape, the DoC matrices for the activity cliff regions of hCatL and TbCatB were computed utilizing only data pairs with at least one active compound in the pair. The results are summarized in Table S14 (Supporting Information). The off-diagonal elements represent the DoC between two different conformer representations. The low-to-moderate values indicated that *the conformers' representations were identifying activity cliffs that were not present in the landscapes of the other conformers.* Since the bioactive conformers are typically unknown, these results point to potential benefits of including different 3D conformers in the characterization of activity landscapes.

Property-Activity Similarity. Activity landscapes typically involve describing activity similarity as a function of structural similarity; however, they may not be limited to only these two measures.[1] Recently, we explored property-activity similarity characterizations employing six physicochemical properties: HBA, HBD, MW, SlogP, RB, and TPSA.[12] Here, we investigated further the property-activity similarity (PAS) relationship using other biological end points. The *PAS maps* for these measures are shown in Figure 6. This figure clearly shows that property similarities were more dispersed as compared to the structure similarities.

PAS maps were analyzed to determine the extent to which property similarities could be applied to characterize SARs. The property similarity measure also identified pairs of compounds

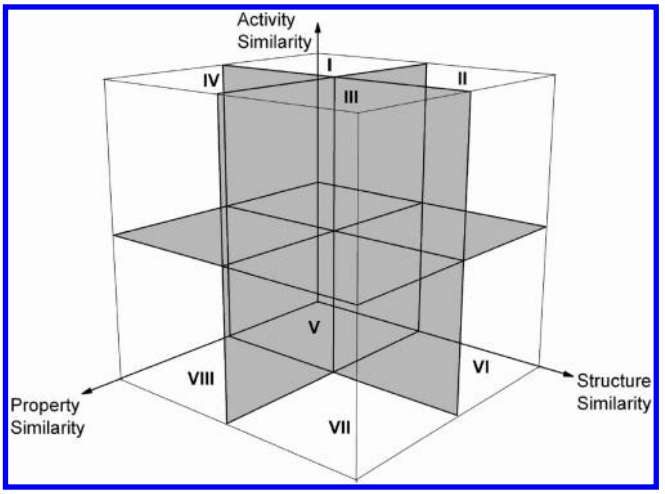


Figure 7. Prototype structure−property-activity (SPA) map showing eight major regions. Region I represents structure/property hops, i.e., compounds with low structure and property similarities, but high activity similarity. Region III contains compounds in the smooth/continuous region of the landscape. Region VII denotes compound pairs involved in activity cliffs. Region V is the least informative region containing pairs of compounds with low structure, property, and activity similarities. The table describes the main regions of the SPA map.

that reflected the definitions of the four defined regions of SAS maps, i.e., the side chain hops, continuous SARs, and activity cliffs. We also examined the extent to which property similarity could detect active pairs that were identified by all the fingerprint models. These comparisons are shown in Table S15 (Supporting Information). Across the three inhibitors, property similarity detected active pairs that were categorized by all the fingerprint models. Taken together, these observations further suggest that it is plausible to employ physicochemical properties as another approach in characterizing activity landscapes.

Structure−Property-Activity (SPA) Maps. An additional dimension could be added to consensus structure−activity maps, e.g. property, to generate structure−property-activity (SPA) landscapes. By including an additional dimension to 2D SAS maps, the compounds are separated further through another measure. A prototypical SPA map is presented in Figure 7, showing eight regions. Regions I, III, V, and VII in Figure 7 are similar to regions I, II, III, and IV, respectively, on SAS maps. In the proposed SPA map, region VII represents a *dual activity cliff* in terms of structure- *and* property-similarities. A full description of the SPA map is provided on the table of Figure 7. Preliminary analysis of the three most informative regions (I, III, and VII) of the SPA map for TbCatB showed promising results. Some examples of compound pairs allocated to these three regions were as follows: **4_50** in region I, **29_41** in region III, and **30_37** in region IV. The SPA maps are general and can be applied to any other data set, using a variety of structure and property representations and similarity measures.

CONCLUSIONS AND PERSPECTIVES

SAS maps were employed to characterize the activity landscapes of 54 compounds with activities against human cathepsin B (hCatB), human cathepsin L (hCatL), and *Trypanosoma brucei* cathepsin B (TbCatB). Starting from 28 structural representations, ten 2D and 3D representations that capture different aspects of the molecular structures were employed. Multiple conformations were utilized to compute 3D similarities. Overall, both the 2D and 3D fingerprint landscapes displayed different distributions of data points across the targets. For quantitative comparisons each SAS map was divided into four regions, defined by the mean structure similarity and the mean activity similarity for each target. Despite the overall different distribution of data point in the SAS maps, for each target we identified consensus data pairs in the different regions of the landscapes. Consensus data points captured by different 2D and 3D representations, including consensus activity cliffs, are particularly valuable in the SAR analysis because these points are consistent in different reference spaces, reducing the dependence of the SAR with molecular representation. These points are intended to prioritize the analysis of the SAR but not eliminate activity cliffs identified by individual molecular representations. We also identified pairs of compounds in the same region of the landscapes across the targets, suggesting similar SARs for the compound pairs involved.

Consensus SAS maps, generated by combining pairwise structure similarities, successfully captured all data pairs with at least one active compound. We also showed that property similarity could be employed to analyze SARs. In the course of this work, we introduced the concept of structure−property-activity (SPA) similarity in analyzing SARs, i.e., by adding property similarity as a new dimension to SAS maps. Further modeling of activity landscapes using SPA maps is warranted. The application of the 3D similarity maps can be extended to represent other properties beyond biological activities for example flavor similarity and "odor cliffs".[46,47]

The strategy presented here is general and can be applied to other data sets with other biological end points. One of the major perspectives of this work is to explore the predictive capabilities of the activity landscape models to anticipate the SAR of new molecules in a prospective manner.

ASSOCIATED CONTENT

Supporting Information. Cumulative distribution functions of the 1431 pairwise similarities using 28 representations (Figure S1); correlation matrix between structural similarities employing 2D fingerprints (Table S1); correlation matrix between structural similarities employing 3D fingerprints (Table S2); SAS maps for hCatB (Figure S2); thresholds employed to define the boundaries of the SAS maps (Table S3); distribution of data points among the different segments of the SAS maps for hCatB (Table S4); categorical SAS maps of hCatB, hCatL, and TbCatB (Figures S3−S5); number of consensus pairs in the SAS maps of hCatB, hCatL, and TbCatB (Tables S5−S7); degree of consensus matrices for the SAS maps of hCatB, hCatL, and TbCatB (Tables S8−S10); statistical significance of the DoC values; number of all-fingerprint consistent pairs in regions I, II, and IV of the SAS maps (Table S11); representative consistent pairs in the most informative regions of the SAS maps (Table S12); active pairs classified as shallows cliffs for hCatL (Table S13); DoC matrices for region IV of the SAS maps of hCatL and TbCatB for the 3D representations calculated using only active pairs (Table S14); the number of active pairs classified into the four major regions of the property-activity similarity maps compared to those identified by all the fingerprints (Table S15). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: (772)345-4685. Fax: (772)345-3649. E-mail: jmedina@tpims.org.

ACKNOWLEDGMENT

We thank the reviewers for their suggestions to improve the manuscript. This work was supported by the State of Florida, Executive Office of the Governor's Office of Tourism, Trade, and Economic Development. Authors are also grateful to OpenEye Scientific Software, for providing ROCS and OMEGA programs. J.L.M.-F. thanks the Menopause & Women's Health Research Center for funding.

REFERENCES

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

(2) Wawer, M.; Lounkine, E.; Wassermann, A. M.; Bajorath, J. Data Structures and Computational Tools for the Extraction of SAR Information from Large Compound Sets. *Drug Discovery Today* **2010**, *15*, 630–639.

(3) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure-Activity Landscapes. *Drug Discovery Today* **2009**, *14*, 698–705.

(4) Maggiora, G. M. On Outliers and Activity Cliffs-Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

(5) Guha, R.; Van Drie, J. H. Assessing How Well a Modeling Protocol Captures a Structure-Activity Landscape. *J. Chem. Inf. Model.* **2008**, *48*, 1716–1728.

(6) Guha, R.; VanDrie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.

(7) Peltason, L.; Bajorath, J. Molecular Similarity Analysis Uncovers Heterogeneous Structure-Activity Relationships and Variable Activity Landscapes. *Chem. Biol.* **2007**, *14*, 489–497.

(8) Medina-Franco, J. L.; Martínez-Mayorga, K.; Giulianotti, M. A.; Houghten, R. A.; Pinilla, C. Visualization of the Chemical Space in Drug Discovery. *Curr. Comput.-Aided Drug Des.* **2008**, *4*, 322–333.

(9) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C. K.; Glick, M.; Davies, J. W. How Similar Are Similarity Searching Methods? A Principal Component Analysis of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2009**, *49*, 108–119.

(10) Bender, A. How Similar Are Those Molecules after All? Use Two Descriptors and You Will Have Three Different Answers. *Expert Opin. Drug Discovery* **2010**, *5*, 1141–1151.

(11) Peltason, L.; Iyer, P.; Bajorath, J. Rationalizing Three-Dimensional Activity Landscapes and the Influence of Molecular Representations on Landscape Topology and the Formation of Activity Cliffs. *J. Chem. Inf. Model.* **2010**, *50*, 1021–1033.

(12) Pérez-Villanueva, J.; Santos, R.; Hernández-Campos, A.; Giulianotti, M. A.; Castillo, R.; Medina-Franco, J. L. Towards a Systematic Characterization of the Antiprotozoal Activity Landscape of Benzimidazole Derivatives. *Bioorg. Med. Chem.* **2010**, *18*, 7380–7391.

(13) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Marín, R. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. Characterization of Activity Landscapes Using 2D and 3D Similarity Methods: Consensus Activity Cliffs. *J. Chem. Inf. Model.* **2009**, *49*, 477–491.

(14) Pérez-Villanueva, J.; Santos, R.; Hernández-Campos, A.; Giulianotti, M. A.; Castillo, R.; Medina-Franco, J. L. Structure-Activity Relationships of Benzimidazole Derivatives as Antiparasitic Agents: Dual Activity-Difference (DAD) Maps. *Med. Chem. Comm.* **2011**, *2*, 44–49.

(15) Peltason, L.; Bajorath, J. Molecular Similarity Analysis in Virtual Screening. In *Chemoinformatics Approaches to Virtual Screening*; Varnek, A., Tropsha, A., Eds.; Royal Society of Chemistry: Cambridge, UK, 2008; pp 120−149.

(16) Maggiora, G. M.; Shanmugasundaram, V. Molecular Similarity Measures. In *Chemoinformatics and Computational Chemical Biology, Methods in Molecular Biology*; Bajorath, J., Ed.; Springer: New York, 2011; Vol. 672, pp 39−100.

(17) Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. CINF-032. In *222nd ACS National Meeting, Chicago, IL, United States*, American Chemical Society: Washington, DC, Chicago, IL, United States, 2001.

(18) Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini-Rev. Med. Chem.* **2006**, *6*, 1217–1229.

(19) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. Scaffold-Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed.* **1999**, *38*, 2894–2896.

(20) Iyer, P.; Wawer, M.; Bajorath, J. Comparison of Two- and Three-Dimensional Activity Landscape Representations for Different Compound Data Sets. *Med. Chem. Comm.* **2011**, *2*, 113–118.

(21) Fricker, S. P. Cysteine Proteases as Targets for Metal-Based Drugs. *Metallomics* **2010**, *2*, 366–377.

(22) O'Brien, T. C.; Mackey, Z. B.; Fetter, R. D.; Choe, Y.; O'Donoghue, A. J.; Zhou, M.; Craik, C. S.; Caffrey, C. R.; McKerrow, J. H. A Parasite Cysteine Protease Is Key to Host Protein Degradation and Iron Acquisition. *J. Biol. Chem.* **2008**, *283*, 28934–28943.

(23) Wilkinson, S. R.; Kelly, J. M. Trypanocidal Drugs: Mechanisms, Resistance and New Targets. *Expert Rev. Mol. Med.* **2009**, *11*, e31.

(24) Enanga, B.; Burchmore, R. J. S.; Stewart, M. L.; Barrett, M. P. Sleeping Sickness and the Brain. *Cell. Mol. Life Sci.* **2002**, *59*, 845–858.

(25) Mackey, Z. B.; O'Brien, T. C.; Greenbaum, D. C.; Blank, R. B.; McKerrow, J. H. A Cathepsin B-Like Protease Is Required for Host Protein Degradation in *Trypanosoma brucei*. *J. Biol. Chem.* **2004**, *279*, 48426–48433.

(26) European Bioinformatics Institute. ChEMBL database, ver. 2. Available at http://www.ebi.ac.uk/chembldb (accessed May 5, 2011).

(27) Mallari, J. P.; Shelat, A. A.; Obrien, T.; Caffrey, C. R.; Kosinski, A.; Connelly, M.; Harbut, M.; Greenbaum, D.; McKerrow, J. H.; Guy, R. K. Development of Potent Purine-Derived Nitrile Inhibitors of the Trypanosomal Protease TbCatb. *J. Med. Chem.* **2008**, *51*, 545–552.

(28) *Canvas, version 1.3*; Schrödinger, L.L.C.: New York, NY, 2010.

(29) Molecular Operating Environment (MOE), *version 2007*; Chemical Computing Group Inc.: Montreal, Quebec, Canada. Available at http://www.chemcomp.com (accessed May 5, 2011).

(30) Mott, B. T.; Ferreira, R. S.; Simeonov, A.; Jadhav, A.; Ang, K. K.-H.; Leister, W.; Shen, M.; Silveira, J. T.; Doyle, P. S.; Arkin, M. R.; McKerrow, J. H.; Inglese, J.; Austin, C. P.; Thomas, C. J.; Shoichet, B. K.; Maloney, D. J. Identification and Optimization of Inhibitors of Trypanosomal Cysteine Proteases: Cruzain, Rhodesain, and TbCatB. *J. Med. Chem.* **2009**, *53*, 52–60.

(31) Musafia, B.; Senderowitz, H. Biasing Conformational Ensembles Towards Bioactive-Like Conformers for Ligand-Based Drug Design. *Expert Opin. Drug Discovery* **2010**, *5*, 943–959.

(32) Yongye, A. B.; Bender, A.; Martínez-Mayorga, K. Dynamic Clustering Threshold Reduces Conformer Ensemble Size While Maintaining a Biologically Relevant Ensemble. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 675–686.

(33) *OMEGA, version 2.3.1*; OpenEye Scientific Software Inc.: Santa Fe, NM. Available at http://www.eyesopen.com (accessed May 5, 2011).

(34) Rapid Overlay of Chemical Structures (ROCS), *version 2.3.1*; OpenEye Scientific Software Inc.: Santa Fe, NM. Available at http://www.eyesopen.com (accessed May 5, 2011).

(35) Jaccard, P. Étude Comparative De La Distribution Florale Dans Une Portion Des Alpes Et Des Jura. *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 547–579.

(36) Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of Structure-Activity Relationship Determinants in Analogue Series. *J. Med. Chem.* **2009**, *52*, 3212–3224.

(37) Agrafiotis, D. K.; Wiener, J. J. M.; Skalkin, A.; Kolpak, J. Single R-Group Polymorphisms (SRPs) and R-Cliffs: An Intuitive Framework for Analyzing and Visualizing Activity Cliffs in a Single Analog Series. *J. Chem. Inf. Model.* **2011**, *51*, 1122−1131.

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

(39) Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W. Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *J. Chem. Inf. Model.* **2010**, *50*, 771–784.

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

(41) Real, R.; Vargas, J. M. The Probabilistic Basis of Jaccard's Index of Similarity. *Syst. Biol.* **1996**, *45*, 380–385.

(42) Real, R. Tables of Significant Values of Jaccard's Index of Similarity. *Misc. Zool.* **1999**, *22.1*, 29–40.

(43) Medina-Franco, J. L.; Maggiora, G. M.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A. A Similarity-Based Data-Fusion Approach to the Visual Characterization and Comparison of Compound Databases. *Chem. Biol. Drug Des.* **2007**, *70*, 393–412.

(44) Willett, P. Similarity-Based Virtual Screening Using 2D Fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.

(45) Borodina, Y. V.; Bolton, E.; Fontaine, F.; Bryant, S. H. Assessment of Conformational Ensemble Sizes Necessary for Specific Resolutions of Coverage of Conformational Space. *J. Chem Inf. Model.* **2007**, *47*, 1428–1437.

(46) Martínez-Mayorga, K.; Medina-Franco, J. L. Chemoinformatics—Applications in Food Chemistry. In *Advances in Food and Nutrition Research*; Taylor, S., Ed.; Academic Press: Burlington, 2009; Vol. 58, pp 33−56.

(47) Martínez-Mayorga, K.; Peppard, T. L.; Yongye, A. B.; Santos, R.; Giulianotti, M. A.; Medina-Franco, J. L. Characterization of a Comprehensive Flavor Database. *J. Chemom.* **2011**, in press. doi: 10.1002/cem.1399.