JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

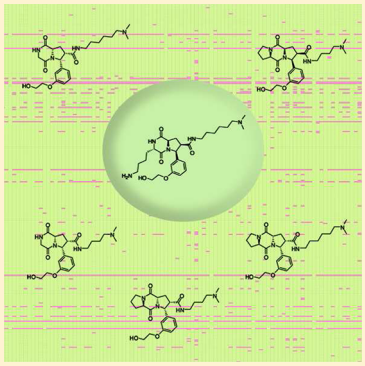
Data Mining of Protein-Binding Profiling Data Identifies Structural Modifications that Distinguish Selective and Promiscuous Compounds

Austin B. Yongye and José L. Medina-Franco*

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

Supporting Information

ABSTRACT: Activity profiling of compound collections across multiple targets is increasingly being used in probe and drug discovery. Herein, we discuss an approach to systematically analyzing the structure−activity relationships of a large screening profile data with emphasis on identifying structural changes that have a significant impact on the number of proteins to which a compound binds. As a case study, we analyzed a recently released public data set of more than 15 000 compounds screened across 100 sequence-unrelated proteins. The screened compounds have different origins and include natural products, synthetic molecules from academic groups, and commercial compounds. Similar synthetic structures from academic groups showed, overall, greater promiscuity differences than do natural products and commercial compounds. The method implemented in this work readily identified structural changes that differentiated highly specific from promiscuous compounds. This approach is general and can be applied to analyze any other large-scale protein-binding profile data.



1. INTRODUCTION

The advent of chemogenomics with sets of compounds screened across multiple biological end points has motivated the development of computational strategies to capture structure−multiple-activity relationships.[1] Recent examples include a systematic analysis of kinase profiling data by Milletti and Hermann.[2] In that work, the authors developed a program that identifies chemical transformations that resulted in selectivity against a specific unwanted kinase, while maintaining activity for the target kinase.[2] A few examples involving the analysis of chemogenomic data sets include the following: Niijima et al. developed a deconvolution approach to dissect kinase profiling data;[3] Steffen et al. evaluated different structure representations with respect to their ability to describe the similarity of biological activity profiles of marketed drugs across a range of proteins relevant for drug design;[4] Medina-Franco and Waddell explored the structure−activity relationships of a set of compounds screened in different bioassays and stored in PubChem.[5] Several of these studies can be considered part of "multitarget activity landscapes", a concept pioneered by Bajorath et al.[6] and largely developed by his and other groups.[7−10] These methods, reviewed in Bajorath et al.,[11,12] can be conveniently adapted to identify structural changes associated with the different number of proteins to which a compound binds from a large set of compounds screened under the same assay conditions across several different molecular targets.

As part of an effort to identify structural features associated with specific and promiscuous compounds, Clemons et al. conducted a large scale protein-binding activity profile of more than 15 000 compounds across 100 diverse, sequence-unrelated proteins.[13] In that work, binding profiles relating the shape and stereochemical complexities of compounds from different sources were compared including commercial compounds (CC), natural products (NP), and diverse synthetic compounds (DC) from academic groups. The binding specificity of compounds from each source was determined in terms of the relative proportion of highly specific compounds (defined in that work as hits binding exactly one protein) and promiscuous (hits binding six or more proteins). It was concluded that compounds from different sources have distinct protein-binding profiles. It was also concluded that increased molecular complexity in natural products and diverse compounds, as measured by the content of sp^3-hydridized and stereogenic carbon atoms, relative to compounds from commercial sources, is associated with improved selectivity and frequency of binding.[13] These results support the hypothesis to develop libraries of complex compounds with balanced physicochemical properties.[14,15] In a follow up study, Clemons et al. conducted a comprehensive analysis of the physicochemical properties, three-dimensional shape, and diversity of biological performance of the screened collections, and concluded that molecular shape could be associated with binding specificity.[16] These published studies are, of course, of great significance but do not explore the relationship between changes in structure with changes in the number of proteins to which a compound binds, which is the focus of this work. Other properties such as molecular weight, lipophilicity, pK_a, molecular shape, and structural features such as the presence of basic moieties,

Received: June 7, 2012
Published: August 2, 2012

dx.doi.org/10.1021/ci3002606 | J. Chem. Inf. Model. 2012, 52, 2454−2461

carboxylic acids, and charge are examples of molecular descriptors that have been associated with promiscuity and selectivity from large-scale profiling data, as reviewed by Leach and Hann.[17]

In this work, we report a chemoinformatic analysis of the 15252 compounds screened across 100 diverse proteins reported by Clemons et al.[13] Despite the fact that the data set of Clemons et al. represents preliminary indication of protein binding; it is one of the very few chemogenomics data sets in the public domain that reports activity data for a large set of diverse proteins using the same assay conditions. The focus of this work was to systematically identify specific and, in particular, small structural changes that distinguish a highly specific compound from a promiscuous one in three different types of screening collections (CC, NP, and DC) assembled in a large scale protein-binding profile screening data. The approach implemented in this work is based on combining the systematic pairwise comparison of structure-similarity using a fingerprint-based structure representation with the difference in the number of proteins to which two compounds bind. This work is part of an ongoing effort in our group to perform chemoinformatic characterizations of compound databases from different sources.[15,18,19] The results of the test case considered in this work further complement the previous analysis of Clemons et al.[13,16] We want to emphasize that the strategy presented here is general and can be applied to other chemogenomics data sets of compounds tested across several different biological end points. As such, we expect to encourage other research groups to analyze their own chemogenomics data using the principles presented in this analysis.

2. METHODS

2.1. Data Set. The data set was comprised of a publicly available group of 15252 compounds assembled by Clemons et al. from 6152 (40.3%) CC, 2477 (16.3%) NP, and 6623 (43.4%) DC.[13] Previously, it was shown that a collection of 660 DC compounds with a spiroxindole scaffold was responsible for the promiscuity of the DC subset.[13] Furthermore, the structures of these compounds are quite complex, hindering a reliable extension of their SAR (structure−activity relationship) analyses. Consequently, the DC′ subset was assembled by excluding the spiroxindole-containing compounds from the DC subgroup. The binding profiles of these compounds determined from a small-molecule microarray assay against 100 sequence-unrelated proteins[13] were also downloaded. The binding activity of each compound with the proteins was described utilizing binary notations: 0 and 1 for inactive and active, respectively.

2.2. Relationship between Structure Similarity and Protein-Binding Difference. The database in SMILES string format was converted to a structure data file using *Molecular Operating Environment* (MOE), version 2008.10.[20] Initially, 10 random samples of binding profiles and the structures of their corresponding compounds were selected proportionately from the CC (403), NP (163), DC (434), and DC′ (434) subsets. The structure files were imported into MOE and defined using the dictionary-based Molecular ACCess System (MACCS) keys (166-bits). The MACCS keys were converted to bit format, retaining only bits that were represented in 5−95% of the compounds. It should be noted that MACCS keys do not take stereochemistry into consideration. Pairwise Tanimoto structure similarities were computed as follows:[21,22]

$$T_n(X_a, X_b) = \frac{\sum_j \min(X_a(j), X_b(j))}{\sum_j \max(X_a(j), X_b(j))} \quad (1)$$

for two compounds X_a and X_b, and feature (bit position) j.

Promiscuity index, P_X, was defined as the number of proteins to which a compound binds and the compounds were classified following the categorization of Clemons et al.: inactive (IN), no protein; highly specific (HSP), 1 protein; partially specific (PSP), 2−5 proteins; and promiscuous (PRM), ≥ 6 proteins.[13] In this study, we did not investigate the potential reasons of promiscuity (e.g., binding promiscuity or aggregation-induced promiscuity)[23] and focused on the annotations reported.[13]

For each pair of compounds, the relationship between structure similarity and the different number of proteins to which each compound in the pair binds was computed using the following equation:

$$\mathrm{SPID}(X_a, X_b) = \frac{|P_{xa} - P_{xb}|}{1 - Tn(X_a, X_b)} \quad (2)$$

where SPID is the Structure-Promiscuity Index Difference, P_{Xa} and P_{Xb} denote the number of proteins to which compounds X_a and X_b are bound, that is, promiscuity indices respectively, and $T_n(X_a, X_b)$ is computed with eq 1. The SPID measure is inspired by the Structure−Activity Landscape Index (SALI) metric[24] commonly used in promiscuity landscape modeling to rapidly identify structural changes that have a large impact on biological activity.[9,10,25,26] Of note, SPID does not account for the specific proteins involved, such as the measures proposed "binding profile similarity"[4] or "multiple-activity similarity".[10] Instead, SPID focuses on the change in the number of proteins bound associated with a change in the molecular structure. As discussed in the following, to address the identities of the proteins, we also computed pairwise binding profile similarities employing the binary profile of each compound as a 100-dimensional vector, effectively a pairwise Tanimoto similarity using eq 1. Thus, we also explored the multiple-assay profile SAR of the data set using the structure-multiple activity landscape index (SmALI) measure.[10] SmALI is a modified version of eq 2 in which the numerator is replaced with the biological profile similarity of the compound pair computed with the Tanimoto coefficient.

Finally, for comparison with the random samples selected proportionately, fixed-sized random samples of the promiscuity indices of 1000 compounds, along with their structures, were selected from the CC, NP, DC, and DC′ subsets, and pairwise SPID values were computed for each group. SPID values were also calculated separately for the 660 spiroxindole compounds.

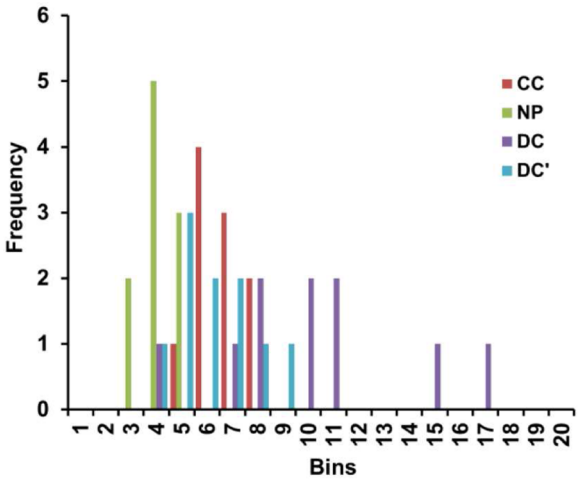
2.3. Systematic Structure Comparison of Promiscuous and Highly Specific Compounds. To identify structural patterns that distinguished selective from promiscuous compounds in the entire collection, the promiscuity indices and corresponding structures of all the 348 PRM and 2054 HSP compounds were extracted from the initial data set of 14592 molecules, that is, the entire set of 15252 structures excluding the 660 spiroxindoles. MACCS keys were computed for each set and a fingerprint model was generated for the PRM set using MOE. The maximum Tanimoto structure similarity between each HSP compound and all the 348 PRM compounds was determined. A heuristic similarity threshold of 0.80 was considered in order to select HSP compounds that were considered "structurally similar" to at least one PRM compound. A total of 403 HSP entries passed this criterion.

The compounds and promiscuity indices of the selected 403 HSP compounds and the 348 from the PRM set were merged, and the SPID values were computed for all 751 structures using eq 2.

3. RESULTS AND DISCUSSION

First, we aimed to compare the relationships between the structure and the change in promiscuity index for three types of screening collections. We employed the SPID measure that is reminiscent of the SALI metric and adapted it to capture differences in the number of proteins bound related to changes in molecular structure. According to eq 2, large SPID values are indicative that a small structural change (i.e., compounds with high structural similarity) is associated with a large promiscuity index difference. Therefore, compound pairs with large SPID values point to small changes in structure that have a significant impact on the number of proteins bound, in other words, changes in structure that distinguish a promiscuous from a selective compound. According to the concept of "activity cliffs",[27] such pairs represent a special case of "property cliffs".

3.1. SPID Value Distributions of Data Sets from Different Sources. Figure 1 shows a histogram with the distributions of the mean SPID values of the ten random samples selected proportionately from the CC, NP, DC, and DC′ subsets.



	Compounds per sample	Mean of ten SPID values	Lowest mean	Highest mean	Mean structure similarity
CC	403	6.10	4.68	7.42	0.29
NP	163	3.41	2.02	4.53	0.44
DC	434	9.63	3.89	16.73	0.46
DC′	434	5.72	3.21	8.01	0.44

Figure 1. Distributions of the mean SPID values of 10 random samples selected proportionately from the CC, NP, DC, and DC′ sets. Structure similarities were computed with MACCS keys/Tanimoto. The statistics of the mean SPID values are summarized in the table. The full statistics for each random sample are in Table S1 (Supporting Information).

The distributions of the SPID values can also be represented as cumulative distribution functions (CDF). Based on the data, one would expect the CDFs of the DC and NP sets to be the most right-shifted and left-shifted, respectively, because of the high and low distributions of their mean SPID values, respectively. On the other hand, the CDFs of the CC and DC′ sets would be similar to each other, and located between those of the NP and DC sets. The full statistics and the number of pairs employed to compute the SPID values for each random sample are presented in Table S1 of the Supporting Information.

The mean MACCS/Tanimoto structure similarity of the different data sets indicated that the CC set was the most diverse, while the NP, DC, and DC′ sets have comparable diversity. Of note, the similarity values for each data set are very similar across the 10 random samples (Table S1). This result was consistent with the previous notion that random samples of 1000 compounds were large enough to represent the structural diversity of compound collections including combinatorial libraries.[15,18,19,28] The data in Figure 1 also indicated that although the NP, DC, and DC′ sets have equivalent mean structure similarity, the three sets have different SPID values because of differences in their promiscuity indices. With the exception of the DC set, none of the samples displayed a mean SPID value greater than 8.50. Indeed, the highest SPID values were ascribed to the DC set, while SPIDs from the NP set exhibited distributions primarily at the low end of the spectrum, with the highest mean SPID value being less than 4.6 (Figure 1). The low and high SPIDs for the NP and DC sets, respectively, could be attributed to the low and high PRM/HSP ratios, respectively, in the NP (1:33.6) and DC (1:7.3) sets. Thus, there were fewer chances of generating a PRM_HSP category pair in the NP set than in the DC set. On the other hand, the PRM/HSP ratio for the CC set was 1:3.0, suggesting a higher chance of observing a PRM_HSP pair relative to the DC set. However, the SPID values of CC were less than those of the DC set (Figure 1). This discrepancy was because, on average, the pairwise structure similarities and the range of the mean SPID values were lower and smaller, respectively, for the CC set. When the 660 spiroxindole scaffold-containing compounds were excluded from the DC set to derive the DC′ set, the distributions and averaged SPID values with the CC set were comparable. From these distributions, the amount of pairs of compounds with similar chemical structure and large promiscuity index difference was as follows: DC > CC ≈ DC′ > NP. It should be noted that these results are not directly comparable with the analysis of the distribution of molecular complexity and other properties reported by Clemons et al.[13] (see the Introduction). SPID is a separate and complementary analysis. For example, a small structural change between a pair of compounds with the same or similar molecular complexity can be associated with a large difference in the number of proteins bound.

To determine whether this trend depended on the sizes of the samples, 1000 compounds were selected randomly from each set and their SPID values were determined (see the Methods section). In addition, all the 660 spiroxindoles were analyzed separately. Table 1 summarizes the results of the distribution of the SPID values along with the mean structure similarity and number of pairwise comparisons for each set of compounds. It should be noted that while the data in Figure 1 were determined from 10 random samples from each collection source, those in Table 1 are for a single sample.

Table 1 clearly shows that the overall trend of the number of compounds with similar chemical structure and large promiscuity index difference was preserved, namely; DC > CC ≈ DC′ > NP. Moreover, the mean Tanimoto structure similarities were statistically indistinguishable from those observed in the samples that were selected proportionately (Table S1, Supporting Information). Remarkably, the spiroxindoles had the highest SPID values, while the DC′ was

Table 1. SPID Measure Statistics for a Single Random Sample of 1000 Compounds Per Screening Collection[a]

	L95	mean	U95	median	max	min	Q1	Q3	mean structure similarity	no. of pairs[d]
CC[b]	6.31	6.35 ± 9.80	6.39	2.17	214.95	0.00	1.42	6.27	0.30 ± 0.12	193619
NP[b]	2.74	2.76 ± 3.50	2.78	1.91	95.75	0.00	1.54	2.73	0.44 ± 0.16	114701
DC[b]	13.35	13.58 ± 50.95	13.80	2.27	3941.00	0.00	1.64	4.39	0.46 ± 0.16	200444
DC′[b]	5.57	5.63 ± 15.19	5.70	2.16	1556.00	0.00	1.63	3.72	0.45 ± 0.15	195396
spiroxindoles[c]	32.15	32.52 ± 70.99	32.90	4.20	1979.00	0.00	2.07	16.95	0.49 ± 0.18	138193

[a]The 660 spiroxindoles are included. [b]Total number of pairwise comparisons: 499500. [c]Total number of pairwise comparisons: 217470. [d]Number of pairs used to derive these statistics, which excluded any IN_IN and HSP_HSP pairs. The numbers of bits retained from the MACCS keys were 118, 79, 116, 117, and 51 for the CC, NP, DC, DC′, and spiroxindoles sets, respectively.

comparable to the CC set. The PRM/HSP ratio of the spiroxindoles was 1:2 and, along with the mean structure similarity value of 0.49 ± 0.18, explained the high SPID values of these compounds. The statistics of the pairwise structure similarities pointed to comparable structure similarities between the NP, DC, DC′, and spiroxindoles samples, while the CC set had the lowest values. However, among the NP, DC′, and spiroxindoles sets with comparable mean structure similarity values, the latter group had both the widest range and the highest maximum value for the promiscuity index difference (Tables S2 and S3 in the Supporting Information). In accord with the observation of Clemons et al., the origin of the high SPID values for the DC subset was attributed to the 660 spiroxindoles. We want to emphasize that the SPID measure in eq 2 provides additional information to the analysis of Clemons et al. for this large data set[13,16] due to the fact that SPID captures a direct relationship between differences in the number of proteins bound associated with changes in molecular structure.

As discussed in the Methods section, the promiscuity index difference, the numerator in eq 2, does not include information on the particular proteins to which the compounds are binding. To take into consideration the identities of the proteins, pairwise Tanimoto similarities were computed employing the protein-binding profile of each compound as a 100-dimensional vector.[4] However, most of the pairs of compounds showed extremely low (several pairs with zero or close-to-zero) binding profile similarity values (data not shown). This result indicated that the majority of the compounds displayed little overlap in their protein binding profiles. It should be noted that the profile Tanimoto similarities do not provide information about the number of proteins involved (in fact, a pair of highly specific compounds may have exactly the same Tanimoto-based binding profile similarity as a pair of highly promiscuous compounds). On the other hand, two compounds with the same promiscuity index will result in a SPID value of zero whether or not they bind to the same sets of proteins. The Tanimoto-based binding profile similarity measure will separate out these compounds: values of 1.0 and zero will point to the same and entirely different sets of proteins, respectively.

3.2. Comparing Promiscuous versus Highly Specific Compounds. To systematically elucidate the structural changes associated with dramatic changes in promiscuity or specificity, all the PRM and HSP compounds were extracted regardless of the source but excluding the spiroxindoles. The PRM compounds were utilized as the queries. The CDF of the maximum similarities or nearest-neighbor curves between the 348 PRM and 2054 HSP compounds is shown in Figure 2.

Because the aim of this work was to identify changes in structure associated with the most dramatic changes in activity, we focused our analysis on HSP compounds that were more

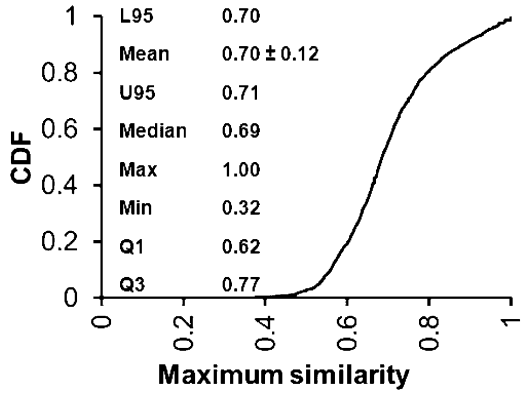


Figure 2. Cumulative distribution function (CDF) of the maximum similarities between promiscuous and highly specific compounds.

structurally similar to PRM, as measured by MACCS keys/Tanimoto. A more comprehensive structural analysis could be performed using several structure representations and generate consensus models as we have extensively described previously.[29−31] However, to illustrate the approach, we used only MACCS keys because this representation provides, overall, reasonable and interpretable results.[32]

The mean of the distribution of the maximum similarity of PRM compounds to any HSP molecule was 0.70 ± 0.12. Out of experience, pairs of compounds with MACCS/Tanimoto similarity values less than 0.80 do not exhibit clearly interpretable structure similarities. Using a heuristic similarity cutoff of 0.80, 403 HSP compounds were selected for pairwise SPID analyses with the 348 PRM compounds. The statistics of the pairwise profile difference, structure similarity and SPID values are presented in Table 2. In principle, SPID values up to 7500 could be observed based on the values of the maximum profile difference (75) and structure similarities (0.99). The actual maximum SPID value was 5000, indicating that a single pair of compounds did not simultaneously display the maximum promiscuity index difference and structure similarity. However, the mean SPID was higher than the expected value of 18.84 (computed from with eq 2 utilizing the mean promiscuity index difference and structure similarity).

The maximum and observed mean SPID values indicated that compounds with similar chemical structure and large promiscuity index difference existed in this data set. Hence, multilevel ranking of the pairwise data was performed employing their SPID and Tanimoto similarity values. Tanimoto thresholds of 0.80, 0.85, and 0.90 were utilized to select pairs of compounds to be analyzed further. The number of unique compounds identified was 109, 91, and 69 at the 0.80, 0.85, and 0.90 limits, respectively. Histograms of the compounds and their frequencies in the selected pairs are

Table 2. Comparison of the 348 PRM and 403 HSP compounds with maximum structure similarities ≥0.80

281625 pairs	L95	mean	U95	median	max	min	Q1	Q3
profile difference	12.20	12.25 ± 10.69	12.30	9.00	75.00	0.00	6.00	16.00
structure similarity	0.35	0.35 ± 0.15	0.35	0.34	0.99	0.00	0.24	0.44
SPID[a]	22.42	22.66 ± 53.83	22.89	14.35	5000.00	0.00	8.80	25.49

[a]Expected SPID value: 18.84.

shown in Figure 3. The order of the top four ranked compounds with the highest number of selected pairs was identical, while six compounds were common across the three cut-offs. The 0.80 and 0.85 limits shared the same top ten compounds. Scaled Shannon entropies[33,34] computed at the 0.80, 0.85, and 0.90 thresholds were 0.863, 0.868, and 0.865, respectively, and indicated that the cut-offs had comparable influences on the distributions.

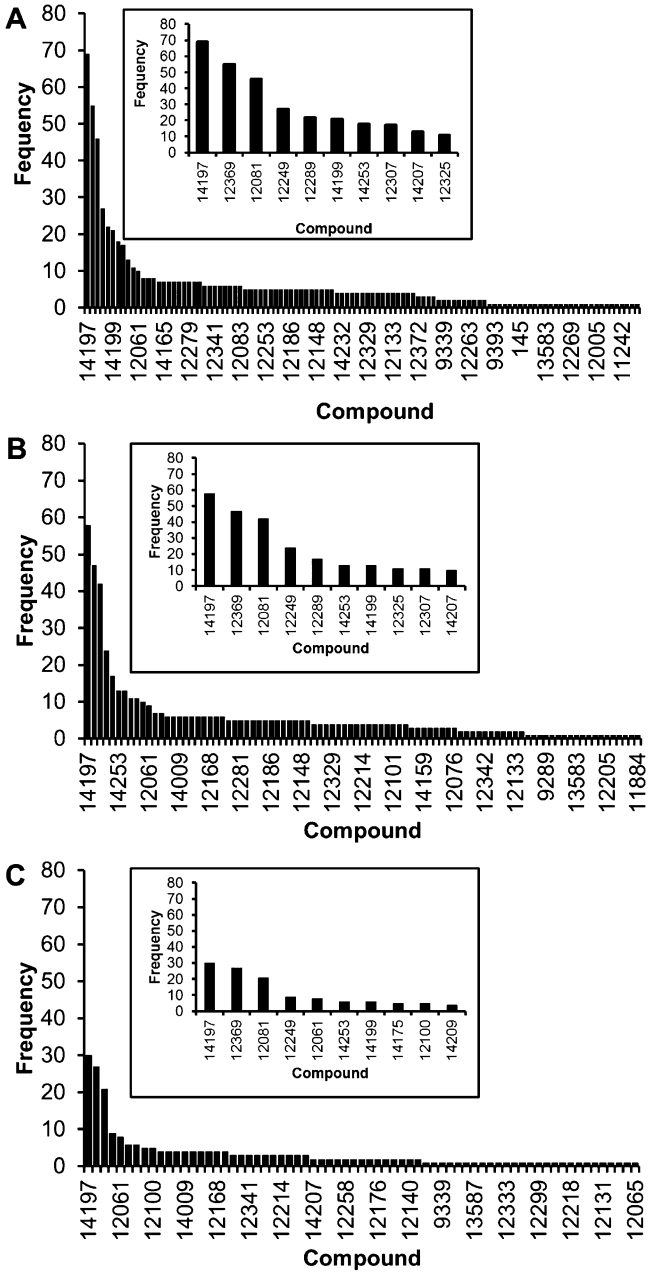


Figure 3. Frequencies of compounds in selected pairs of promiscuous and highly specific compounds with high SPID values for three pairwise Tanimoto similarity cut-offs: (A) 0.80 (B) 0.85, and (C) 0.90. The inserts portray the top 10 compounds with the highest frequencies at each threshold.

Figure 4 shows the three most frequent promiscuous compounds **14197**, **12369**, and **12081**, respectively, in the selected compounds at the 0.90 MACCS/Tanimoto threshold. Each panel also shows the top five highly specific compounds with the highest structure similarity to the corresponding promiscuous, reference molecule. The only exception is in Figure 4C, which includes a compound (**12061**) that binds to 16 proteins and belonged to the PRM category. In Figure 4, MACCS/Tanimoto similarity between the central compound and each of the five surrounding compounds is >0.97. The IDs of the compounds as indicated by Clemons et al.[13] are shown. Only the common structural differences between the central and all five surrounding compounds are marked with blue and red colors (structural differences that occur between the central compound and few but not all surrounding molecules are not colored).

Figure 4A shows the structural relationships between compound **14197**, which binds to 76 proteins, and five highly similar compounds that bind to only one protein. The structural differences between these five molecules and **14197**, highlighted in color, revealed that the presence of a straight-chain primary amine at position three of the pyrrolopyrazine-1,4-dione scaffold, as opposed to an alicyclic system fused to the pyrazine moiety, drastically determined specificity (**14009**), irrespective of stereochemistry (**12145**, **12378**) or the number of carbon atoms linking the tertiary amine to the carboxamide group (**14209**, **12168**) at position seven. Similar conclusions could be drawn from the comparisons with compounds **12369** (Figure 4B) and **12081** (Figure 4C).

Among the five compounds with high structure similarity to **12081** (Figure 4C), compound **12131** contained a tertiary amine at position three and primary amine at position seven. Though **12061** binds to only 16 proteins, the pair **12100_12061** (Figure 4C) possessed identical cyclic systems and stereochemistries. Nonetheless, the presence of two additional carbon atoms between the tertiary amine and the carboxamide group in **12100** rendered it highly specific, while **12061** was promiscuous (structural change marked in green).

Compound **12131** (Figure 4C) was highly specific even though a primary amine was present at position seven of the pyrrolopyrazine-1,4-dione scaffold, indicating that promiscuity was more likely to be observed when a primary amine substituent was at position three than at another position. To follow up on this observation, all compounds in the data set containing either a primary amine or a pyrrolidine substructure were extracted and their promiscuity indices were analyzed. Only 81 compounds bore a primary amine, and 68 of them contained the pyrrolopyrazine-1,4-dione scaffold. Table S4 in the Supporting Information summarizes the results of the 81 compounds: 16 and 44 compounds possessed a primary amine

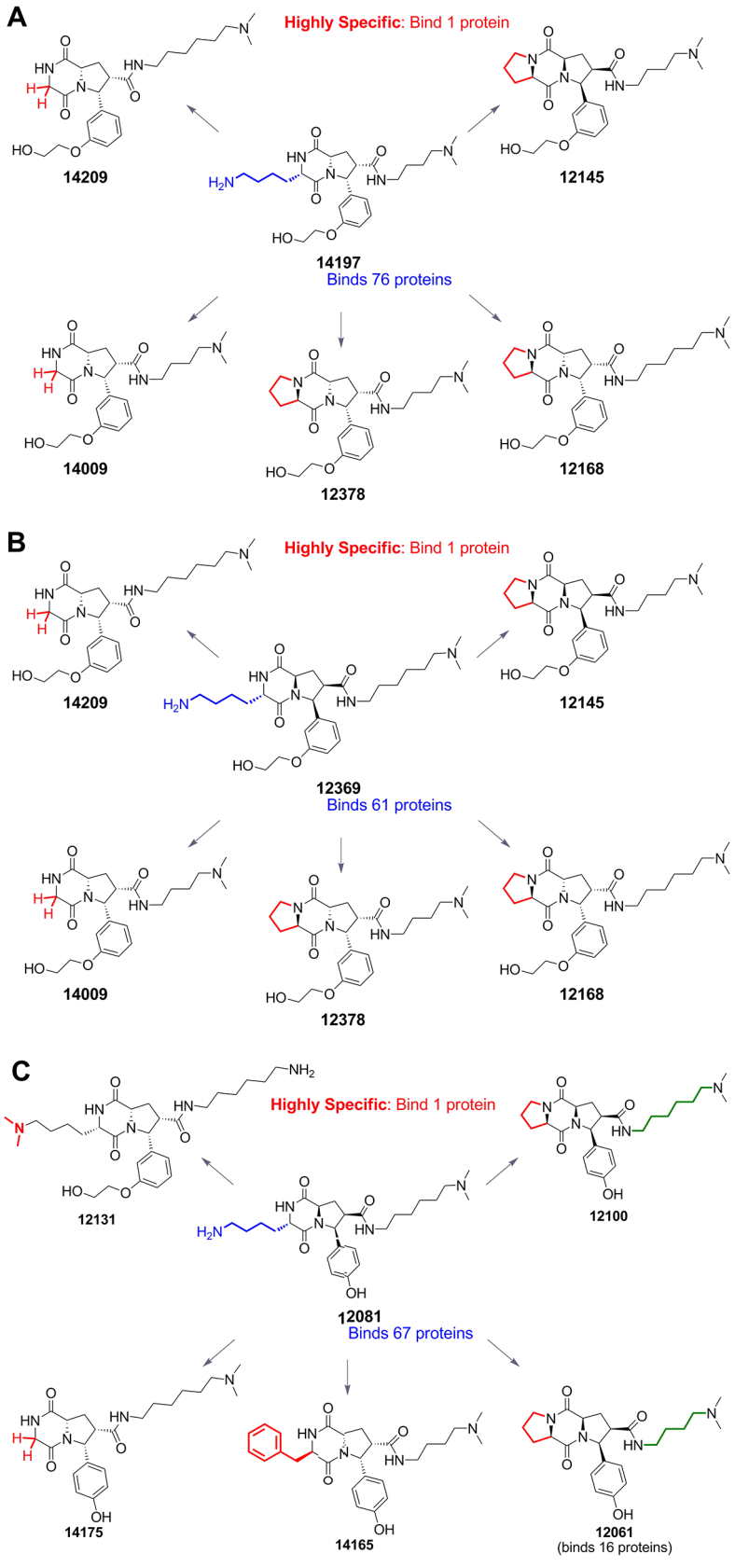


Figure 4. Chemical structures of the top three most frequent promiscuous compounds present in pairs of compounds with high SPID values. (A) **14197**, (B) **12369**, and (C) **12081**. In each panel, the common structural differences between the central promiscuous and five surrounding most similar highly specific compounds (except **12061**), are marked with a blue and red color, respectively. In panel C, the structural difference between **12100** and **12061** is marked with green. See text for details.

dx.doi.org/10.1021/ci3002606 | *J. Chem. Inf. Model.* 2012, 52, 2454−2461

at positions three and seven of the pyrrolopyrazine-1,4-dione scaffold, respectively, while eight had this functionality at both positions. The other 13 did not contain the pyrrolopyrazine-1,4-dione scaffold. On the other hand, 362 compounds contained a pyrrolidine moiety. Consistent with our initial observation that promiscuity is more likely to be conferred by the presence of a primary amine at position three, much higher proportions of compounds bearing a primary amine at position three alone (43.8%) or concurrently at positions three and seven (62.3%) were promiscuous, as opposed to this group present only at position seven (4.5%). The higher probability of promiscuity could also be pointed to the amine functionality at position three by the low percentage of promiscuous compounds with a primary amine not attached to the pyrrolopyrazine-1,4-dione scaffold at position three (15.4%), as well as with the pyrrolidine substructure fused at positions two and three (3.3%) or on a different scaffold (5.3%) (Table S4, Supporting Information). These results are in accord with the trends noticed in the study of Azzaoui et al. with a different data set, where bulky and hydrophobic amines were more likely to be promiscuous.[35] The lower percentage of promiscuous primary amines in the Azzoui et al. study suggests some relevance of the pyrrolopyrazine-1,4-dione scaffold in imparting promiscuity when coupled to a primary amine. These primary amines, more accessible to binding sites because of their attachment to a flexible *n*-butyl linker, may be protonated at the pH of the buffer in which assays were performed. This would corroborate previous observations that bases and positively charged or quaternary bases were more likely to be promiscuous than their negatively charged or neutral counterparts.[36,37]

The observations discussed, obtained with the use of a fingerprint-based representation (MACCS keys) to evaluate molecular similarity, open up the question to systematically explore other substructures and molecular scaffolds in this protein-binding profile data associated with promiscuity/selectivity. A first step in this direction would be a systematic analysis of promiscuous and selective compounds across different substructures and scaffolds using a chemotype-based classification, as previously reported for other compounds.[38] This is the subject of a separate work that will be reported in due course.

We want to emphasize that while other studies of chemogenomics data are focused on identifying general trends between physicochemical properties and other descriptors with promiscuity/selectivity, the present work is focused on identifying small and specific structural modifications that have a large impact on the number of proteins to which a compound binds. One of the next logical steps of this work is to further evaluate if the chemical transformations described for the compounds analyzed in this case study are also applicable to newly synthetized compounds or compounds already present in other proprietary screening collections of other groups. We expect that the results of this study will motivate experimental research groups to conduct such validation.

4. CONCLUSIONS

The principles of activity landscape modeling were adapted to conduct a systematic analysis of protein-binding data. Such an approach can be utilized to guide library design or chemical synthesis during lead optimization. As a case study, a recently published data set by Clemons et al. with more than 15000 compounds from different origins tested across 100 diverse proteins was employed. We used a modified form of the SALI measure, Structure Promiscuity Index Difference (SPID), to systematically analyze small structural changes associated with changes in the number of proteins to which a compound binds. This approach is different from the analyses of Clemons et al.[13,16] and represents an additional analysis of the rich chemogenomics data set. For the compounds in the test case analyzed in this work, it was concluded that small structural changes in synthetic compounds from academic groups showed greater promiscuity difference than do synthetic compounds from commercial sources. Natural products showed the lowest drastic changes in promiscuity due to small structural modifications. The specific structural modifications that are associated with the most dramatic change in the number of proteins bound of this data set were identified, that is, small structural changes that differentiated a highly specific (targeting one protein) from a promiscuous compound (targeting more than six proteins). The approach discussed herein is not restricted to the public protein-binding profile data discussed in this work. Indeed, we encourage other research groups to apply this method to other public and proprietary large-scale profiling data. In this work, a fingerprint-based representation was used to assess molecular similarity. Other fingerprint-based or property-based representations can be employed. In addition, alternative representations, such as matched molecular pairs recently utilized to model activity landscapes,[39] can also be implemented.

■ ASSOCIATED CONTENT

Supporting Information

Statistics of 10 random samples selected proportionately from the CC, NP, DC, and DC′ subsets (Table S1); distribution of the promiscuity index differences for a random sample of 1000 compounds. All 660 spiroxindoles were also analyzed separately (Table S2). Distribution of the Tanimoto structure similarities for a random sample of 1000 compounds. All 660 spiroxindoles were also analyzed separately (Table S3). Promiscuity profiles of compounds containing either a primary amine or a pyrrolidine (Table S4). The perl scripts employed to perform the random sampling can be obtained from the authors upon request. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +1-772-345-4685. Fax: +1-772-345-3649. E-mail: jmedina@tpims.org.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are grateful to Karina Martínez-Mayorga for helpful discussions and thank Jacob Waddell for help preparing the figures and proofreading the manuscript. We also thank the anonymous reviewers for their comments and suggestions. This work was funded by the State of Florida, Executive Office of the Governor's Office of Tourism, Trade, and Economic Development.

■ REFERENCES

(1) Lounkine, E.; Kutchukian, P.; Petrone, P.; Davies, J. W.; Glick, M. Chemotography for Multi-Target SAR Analysis in the Context of

Biological Pathways. *Bioorg. Med. Chem.* **2012**, in press. DOI:10.1016/j.bmc.2012.02.034.
(2) Milletti, F.; Hermann, J. C. Targeted Kinase Selectivity from Kinase Profiling Data. *ACS Med. Chem. Lett.* **2012**, *3*, 383−386.
(3) Niijima, S.; Shiraishi, A.; Okuno, Y. Dissecting Kinase Profiling Data to Predict Activity and Understand Cross-Reactivity of Kinase Inhibitors. *J. Chem. Inf. Model.* **2012**, *52*, 901−912.
(4) Steffen, A.; Kogej, T.; Tyrchan, C.; Engkvist, O. Comparison of Molecular Fingerprint Methods on the Basis of Biological Profile Data. *J. Chem. Inf. Model.* **2009**, *49*, 338−347.
(5) Medina-Franco, J. L.; Waddell, J. Towards the Bioassay Activity Landscape Modeling in Compound Databases. *J. Mex. Chem. Soc.* **2012**, *56*, 179−184.
(6) Peltason, L.; Hu, Y.; Bajorath, J. From Structure-Activity to Structure-Selectivity Relationships: Quantitative Assessment, Selectivity Cliffs, and Key Compounds. *ChemMedChem* **2009**, *4*, 1864−1873.
(7) Wassermann, A. M.; Peltason, L.; Bajorath, J. Computational Analysis of Multi-Target Structure-Activity Relationships to Derive Preference Orders for Chemical Modifications toward Target Selectivity. *ChemMedChem* **2010**, *5*, 847−858.
(8) Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Design of Multitarget Activity Landscapes That Capture Hierarchical Activity Cliff Distributions. *J. Chem. Inf. Model.* **2011**, *51*, 258−266.
(9) Medina-Franco, J. L.; Yongye, A. B.; Pérez-Villanueva, J.; Houghten, R. A.; Martínez-Mayorga, K. Multitarget Structure-Activity Relationships Characterized by Activity-Difference Maps and Consensus Similarity Measure. *J. Chem. Inf. Model.* **2011**, *51*, 2427−2439.
(10) Waddell, J.; Medina-Franco, J. L. Bioactivity Landscape Modeling: Chemoinformatic Characterization of Structure-Activity Relationships of Compounds Tested across Multiple Targets. *Bioorg. Med. Chem.* **2012**, in press. DOI:10.1016/j.bmc.2011.11.051.
(11) 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.
(12) Bajorath, J. Modeling of Activity Landscapes for Drug Discovery. *Expert Opin. Drug Discovery* **2012**, *7*, 463−473.
(13) Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small Molecules of Different Origins Have Distinct Distributions of Structural Complexity That Correlate with Protein-Binding Profiles. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 18787−18792.
(14) Dandapani, S.; Marcaurelle, L. A. Accessing New Chemical Space for 'Undruggable' Targets. *Nat. Chem. Biol.* **2010**, *6*, 861−863.
(15) López-Vallejo, F.; Giulianotti, M. A.; Houghten, R. A.; Medina-Franco, J. L. Expanding the Medicinally Relevant Chemical Space with Compound Libraries. *Drug Discovery Today* **2012**, *17*, 718−726.
(16) Clemons, P. A.; Wilson, J. A.; Dancik, V.; Muller, S.; Carrinski, H. A.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Quantifying Structure and Performance Diversity for Sets of Small Molecules Comprising Small-Molecule Screening Collections. *Proc. Natl. Acad. Sci. U.S.A.* **2011**, *108*, 6817−6822.
(17) Leach, A. R.; Hann, M. M. Molecular Complexity and Fragment-Based Drug Discovery: Ten Years On. *Curr. Opin. Chem. Biol.* **2011**, *15*, 489−496.
(18) Singh, N.; Guha, R.; Giulianotti, M. A.; Pinilla, C.; Houghten, R. A.; Medina-Franco, J. L. Chemoinformatic Analysis of Combinatorial Libraries, Drugs, Natural Products, and Molecular Libraries Small Molecule Repository. *J. Chem. Inf. Model.* **2009**, *49*, 1010−1024.
(19) López-Vallejo, F.; Nefzi, A.; Bender, A.; Owen, J. R.; Nabney, I. T.; Houghten, R. A.; Medina-Franco, J. L. Increased Diversity of Libraries from Libraries: Chemoinformatic Analysis of Bis-Diazacyclic Libraries. *Chem. Biol. Drug Des.* **2011**, *77*, 328−342.
(20) *Molecular Operating Environment (MOE)*, version 2008.10; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2008.
(21) Jaccard, P. Etude Comparative De La Distribution Florale Dans Une Portion Des Alpes Et Des Jura. *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 547−579.
(22) 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.
(23) Seidler, J.; McGovern, S. L.; Doman, T. N.; Shoichet, B. K. Identification and Prediction of Promiscuous Aggregating Inhibitors among Known Drugs. *J. Med. Chem.* **2003**, *46*, 4477−4486.
(24) Guha, R.; VanDrie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.
(25) Hessler, G.; Matter, H.; Schmidt, F.; Giegerich, C.; Wang, L.-h.; Guessregen, S.; Baringhaus, K.-H. Identification and Application of Antitarget Activity Hotspots to Guide Compound Optimization. *Mol. Inf.* **2011**, *30*, 996−1008.
(26) Seebeck, B.; Wagener, M.; Rarey, M. From Activity Cliffs to Target-Specific Scoring Models and Pharmacophore Hypotheses. *ChemMedChem* **2011**, *6*, 1630−1639.
(27) Maggiora, G. M. On Outliers and Activity Cliffs. Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.
(28) Agrafiotis, D. K. A Constant Time Algorithm for Estimating the Diversity of Large Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 159−167.
(29) 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.
(30) Yongye, A.; Byler, K.; Santos, R.; Martínez-Mayorga, K.; Maggiora, G. M.; Medina-Franco, J. L. Consensus Models of Activity Landscapes with Multiple Chemical, Conformer, and Property Representations. *J. Chem. Inf. Model.* **2011**, *51*, 1259−1270.
(31) Medina-Franco, J. L.; Yongye, A. B.; López-Vallejo, F. Consensus Models of Activity Landscapes. In *Statistical Modeling of Molecular Descriptors in QSAR/QSPR*; Dehmer, M.; Varmuza, K.; Bonchev, D., Eds.; Wiley-VCH: Weinheim, Germany, 2012; pp 307−326.
(32) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J. Med. Chem.* **2012**, *55*, 2932−2942.
(33) Godden, J. W.; Bajorath, J. Analysis of Chemical Information Content Using Shannon Entropy. In *Reviews in Computational Chemistry*; Lipkowitz, K. B.; Cundari, T. R., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, 2007; Vol. *23*, pp 263−289.
(34) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR Comb. Sci.* **2009**, *28*, 1551−1560.
(35) Azzaoui, K.; Hamon, J.; Faller, B.; Whitebread, S.; Jacoby, E.; Bender, A.; Jenkins, J. L.; Urban, L. Modeling Promiscuity Based on in Vitro Safety Pharmacology Profiling Data. *ChemMedChem* **2007**, *2*, 874−880.
(36) Peters, J. U.; Schnider, P.; Mattei, P.; Kansy, M. Pharmacological Promiscuity: Dependence on Compound Properties and Target Specificity in a Set of Recent Roche Compounds. *ChemMedChem* **2009**, *4*, 680−686.
(37) Leeson, P. D.; Springthorpe, B. The Influence of Drug-Like Concepts on Decision-Making in Medicinal Chemistry. *Nat. Rev. Drug Discovery* **2007**, *6*, 881−890.
(38) Medina-Franco, J. L.; Petit, J.; Maggiora, G. M. Hierarchical Strategy for Identifying Active Chemotype Classes in Compound Databases. *Chem. Biol. Drug Des.* **2006**, *67*, 395−408.
(39) Hu, X.; Hu, Y.; Vogt, M.; Stumpfe, D.; Bajorath, J. MMP-Cliffs: Systematic Identification of Activity Cliffs on the Basis of Matched Molecular Pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138−1145.