JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

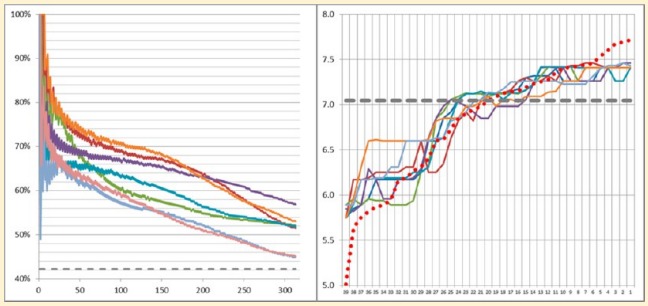
Conditional Probabilistic Analysis for Prediction of the Activity Landscape and Relative Compound Activities

Radleigh G. Santos,* Marc A. Giulianotti, Richard A. Houghten, 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: Structure–property relationships and structure–activity relationships play an important role in many research areas, such as medicinal chemistry and drug discovery. Such methods, however, have focused on providing post-hoc descriptions of such relationships based on known data. The ability for these descriptions to remain relevant when considering compounds of unknown activity, and thus the prediction of activity and property landscapes using existing data, remains little explored. In this study, we present a novel method of evaluating the ability of a compound comparison methodology to provide accurate information about a set of unknown compounds and also explore the ability of these predicted activity landscapes to prioritize active compounds over inactive. These methods are applied to three distinct and diverse sets of compounds, each with activity data for multiple targets, for a total of eight target-compound set pairs. Six methodologically distinct compound comparison methods were evaluated. We show that overall, all compound comparison methods provided an improvement in structure-activity relationship prediction over random and were able to prioritize compounds in a superior manner to random sampling, but the degree of success and therefore applicability varied markedly.



INTRODUCTION

One of the definitions of activity landscape is "any representation that integrates the analysis of the structural similarity and potency differences between compounds sharing the same biological activity.".[1] The activity landscape has also been conceptualized as the multidimensional space resulting from the combination of biological activity and the chemical space of a given compound data set.[2] Activity landscape methods, which can be generalized to property landscape methods, are increasingly being used to systematically describe the structure–property relationships (SPR) or the structure–activity relationships (SAR) of large data sets.[1,3−5] Different graphical and numerical methods are emerging to characterize systematically activity landscapes and are reviewed in several excellent publications.[1,2,6] A number of methods, such as Structure–Activity Similarity (SAS) maps, compare structural similarity and activity similarity for all pairs of compounds in a data set. SAS and related maps, recently reviewed in references 7 and 8 have been applied for a variety of data sets not only to model activity landscapes but also other property landscapes such as the flavor landscape for a comprehensive flavor database.[9,10]

Structure–property relationships and structure–activity relationships play a central role in medicinal chemistry, drug discovery, and several other research areas. The application of these methods has been mainly focused on the systematic description of the SPR/SARs of the compound data sets. A number of approaches are available to develop quantitative models of SPR/SAR including quantitative structure–activity relationships (QSAR), rule-based methods, neural networks, or pharmacophore modeling, to mention a few examples.[11−13] However, some methodologies, like conventional QSAR, make assumptions that are not necessarily valid and may lead to nonpredictive models or other misleading results.[14] Furthermore, descriptive SAR is often treated in a post-hoc manner, with little evidence that the information contained therein offers any predictive capacity on future data sets. Recently, two methods using random forest and support vector machine models, respectively, were proposed to predict activity cliffs.[15,16] In this study we propose a methodology for assessing the predictive capacity of activity and property landscapes and for directly applying that capacity to a set of compounds of unknown activity. Using three distinct and diverse sets of compounds and eight different target-compound set pairs, we gauge the ability of six 2D and 3D structural fingerprint and property similarity methods to predict aspects of the structural activity landscape of a randomly chosen test set using conditional probabilities derived from a disjoint training set. Then, using the same conditional probabilities, we rank compounds in the test set for likely activity and absence of activity. We demonstrate that this method differentiates strongly between compound comparison methodologies, but that in all cases the method demonstrated increased accuracy

Received: April 25, 2013
Published: August 23, 2013

Table 1. Summary of the Data Sets Used in This Study

compd set	# compds	# comparisons in training and test set	average compd comparison method scores						target	% comparisons w/ *ΔA ≤ 0.5 pKi	average ΔA (pKi)
			MACCS	Pharm4Pts	PiDAPH3	radial	TGD	properties			
OR	98	1176	0.74	0.08	0.61	0.17	0.84	0.64	DOR	42.6%	0.68
									KOR	39.3%	0.73
									MOR	33.8%	0.93
									NOC	23.4%	1.20
TPIOR	160	3160	0.69	0.06	0.53	0.12	0.74	0.62	KOR	43.7%	0.65
									MOR	41.8%	0.57
TPIP	78	741	0.55	0.10	0.57	0.16	0.71	0.53	GI	42.2%	0.76
									TV	31.8%	0.99

over random in assessing the activity landscape and in the prioritization of compounds.

METHODS

Data Sets. Three previously published data sets were used in this study. The first, published in reference 17, is a set of 98 compounds tested against four targets: the Delta Opioid Receptor (DOR), the Kappa Opioid Receptor (KOR), the Mu Opioid receptor (MOR), and the Nociception receptor (NOC). We will refer to this as the OR data set in this study. The second set is an in-house collection of small molecules[18] that contains 160 compounds against KOR and MOR. We will refer to this as the TPIOR data set. Finally, the third set, published in reference 19, contains 78 compounds with activities against the parasitic targets *G. intestinalis* (GI) and *T. vaginalis* (TV). We will refer to this as the TPIP data set. Summary information about these data sets is in Table 1.

Compound Comparison Methods. Pairwise structural similarities were computed using the five fingerprint methods previously used in reference 19 having the weakest linear correlation (in this case, $R \leq 0.78$ in all cases) when similarity values are directly compared to one another: MACCS (166-bits), Pharmacophore Atom Quadruplet (Pharm4Pts), Pharmacophore Atom Triangle (PiDAPH3), Typed Graph Distance (TGD), and Radial. Additionally, Property Similarity scores, as described in references 19−25, were also computed between pairs of compounds using the same six drug-like properties previously reported. All values were rounded to two decimals for simplicity. A summary of these compound comparison methods in regards to each of the above data sets is also in Table 1.

Randomized Two-Fold Cross-Validation. All subsequent procedures described below make use of a training set of compounds, which we will denote $\{Tr_i\}$, and a testing set of compounds, which we will denote $\{Ts_j\}$, for each trial. The determination of the training and test sets for a given trial was randomized; for the N^{th} trial, the set of compounds $\{C_k\}$ being analyzed was divided at random into two equal-sized disjoint sets $\{C^{(N)}{}_{k_i}\}$ and $\{C^{(N)}{}_{k_j}\}$. Cross-validation was also performed; first, the assignments $\{C^{(N)}{}_{k_i}\} = \{Tr_i\}$ and $\{C^{(N)}{}_{k_j}\} = \{Ts_j\}$ were made, and then these assignments were reversed, in order to ensure that the randomization procedure played no role in the predictive results. In total, 1,000 cross-validated trials were performed for each compound comparison method on each compound set-receptor pair.

Generating Conditional Probabilities. The activity landscape within the training set was described in terms of conditional probability using the following process. Within the training set associated with the N^{th} trial, the activity differences $\{\Delta A^{(N)}{}_{Tr_i,Tr_j}\}_{j>i}$ (in pIC_{50}) were determined for all unique pairs of compounds. For the K^{th} compound comparison method, the compound similarity scores for all unique pairs of compound in the training set, $\{Sim^{(K,N)}{}_{Tr_i,Tr_j}\}_{j>i}$, was also determined. The conditional probability estimation function was then computed using the standard definition of conditional probability:

$$F^{(K,N)}(x) = P(\Delta A^{(N)} \leq A^* \mid Sim^{(K,N)} \geq x) = \frac{\Sigma_{j>i} I_{\{\Delta A^{(N)}Tr_i,Tr_j \leq A^*\}} I_{\{Sim^{(K,N)}Tr_i,Tr_j \geq x\}}}{\Sigma_{j>i} I_{\{Sim^{(K,N)}Tr_i,Tr_j \geq x\}}} \quad (1)$$

Here A^* represents a chosen activity difference level of interest, $I_{\{Statment\}}$ is an indicator function with a value of one if *Statment* is true and zero otherwise, and x is the variable compound similarity threshold. Thus, equation 1 measures the probability that a given compound pair with a similarity value greater than or equal to x also has an activity difference less than or equal to A^* pKi. For high values of similarity x, $F^{(K,N)}(x)$ essentially determines the proportion of compound pairs with a similarity of at least x for which there are not activity cliffs; thus a high probability value would reflect an ability for a given comparison method to be directly relevant to activity similarity. $F^{(K,N)}(x)$ was calculated for each $x = \{0.00, 0.01, 0.02, ..., 0.99, 1.00\}$ over all comparison methods. This function may then be used to convert a given similarity score Sim_{C_1,C_2} (rounded to two decimal places) into the conditional probability associated with a threshold of that score:

$$CP^{(K,N)}{}_{C_1,C_2} = F^{(K,N)}(Sim_{C_1,C_2}) \quad (2)$$

In order to validate using the testing set, the above quantification of the activity landscape was then applied to the testing set. Using equation 2, each comparison score over all unique compound pairs, both among the test set compounds and between the test and training set compounds, was converted into conditional probability values:

$$\{CP^{(K,N)}{}_{Ts_i,Ts_j} = F^{(K,N)}(Sim_{Ts_i,Ts_j})\}_{j>i} \;\&\{CP^{(K,N)}{}_{Tr_i,Ts_j} = F^{(K,N)}(Sim_{Tr_i,Ts_j})\}_{\forall i,j} \quad \forall K \quad (3)$$

Additionally, for each unique compound pair, again both among the test set compounds and between the test and training set compounds, the product of the conditional probabilities of all six methods studied was calculated:

dx.doi.org/10.1021/ci400243e | J. Chem. Inf. Model. 2013, 53, 2613−2625

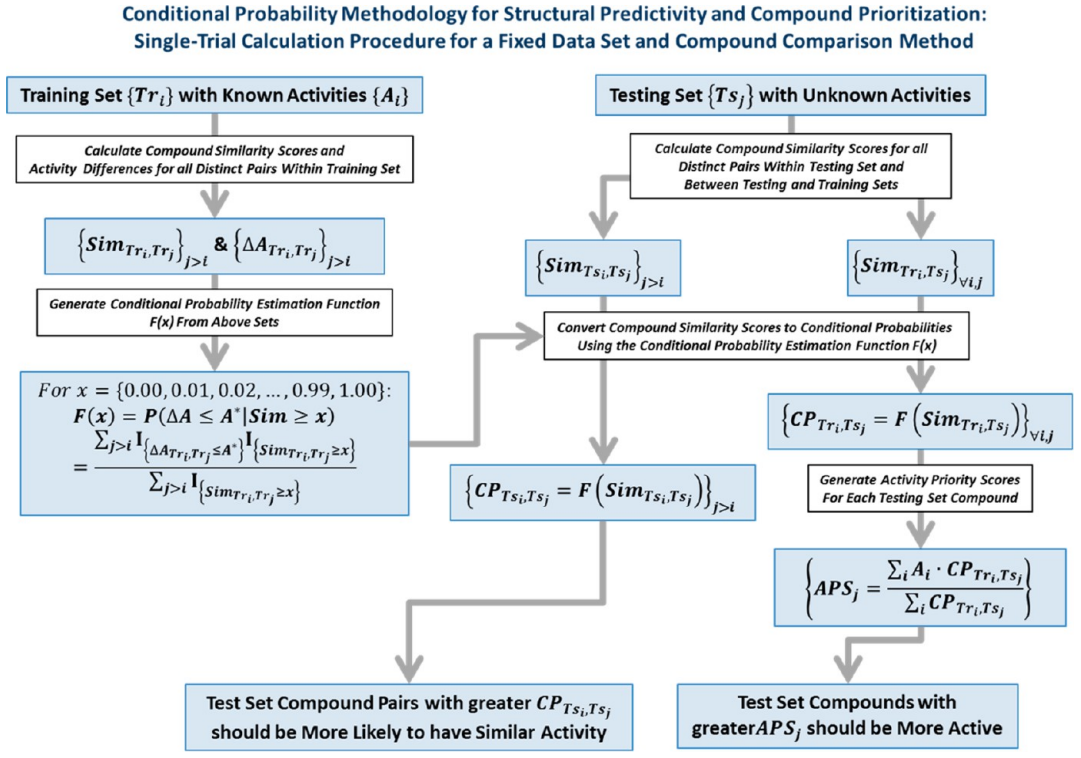
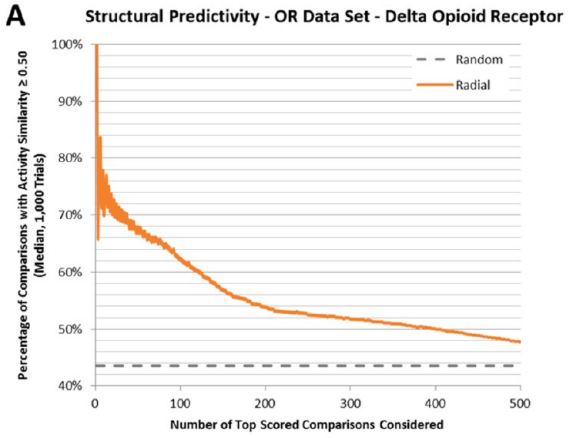
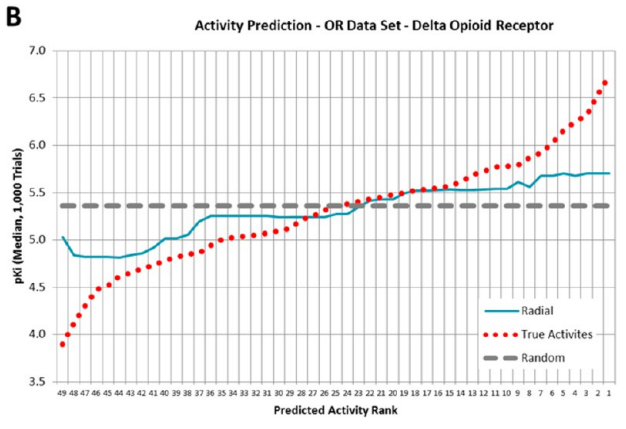


Figure 1. A schematic of a single-trial and single-comparison method implementation of the structural predictivity and activity prioritization method described in this study.





C

Compound Set	Target	Comparison Method	Median Success Rates Over Random in Predicting Activity Similarity				
			Portion of Top-Ranked Comparisons Considered, as % of Theoretical Maximum				
			1%	5%	10%	25%	50%
OR	DOR	Radial	83.8%	65.4%	56.2%	37.8%	20.8%

D

Compound Set	Target	Comparison Method	Active Compound Prediction			
			Median % of Maximum Possible Change (Over Random)		Median Activity Increase/Decrease (pKi) from Random	
			Average of Predicted Bottom 10%	Average of Predicted Top 10%	Average of Predicted Bottom 10%	Average of Predicted Top 10%
OR	DOR	Radial	-48.0%	33.7%	-0.49	0.34

Figure 2. An example of a structural predictivity results plotted (A) and tabulated (C), along with activity prediction results plotted (B) and tabulated (D).

dx.doi.org/10.1021/ci400243e | J. Chem. Inf. Model. 2013, 53, 2613−2625

$$\{CP^{(\Pi,N)}{}_{Ts_i,Ts_j} = \prod_K CP^{(K,N)}{}_{Ts_i,Ts_j}\}_{j>i}$$
$$\&\{CP^{(\Pi,N)}{}_{Tr_i,Ts_j} = \prod_K CP^{(K,N)}{}_{Tr_i,Ts_j}\}_{\forall i,j} \quad (4)$$

The product conditional probability values, then, correspond (up to a constant multiple) to the assumption of compound comparison independence underlying the implementation of the Naïve Bayes algorithm heretofore used in other QSAR contexts[26] but not previously applied in the context of the activity landscape. Thus, for a given trial, each unique compound pair in the test set and between the training and test set was assigned seven values based on conditional probability: one for each of the six methods studied and a "consensus"[27,28] product value. These values will be used to validate both the ability of a given method to describe the activity landscape of the testing set and to predict relative activities within the testing set. A schematic of a single trial is illustrated in Figure 1. The entire cross-validation process is illustrated in Figure S1. To ensure the program was performing correctly, comparison methods representing perfect correlation with activity difference, and random values representing no relationship to activity difference, were run as well.

Assessing the Structural Predictivity of a Comparison Method. For the N^{th} trial associated with a given compound-receptor set pair, let M be the total number of unique comparisons among the test set compounds (i.e., those not involving compounds in the training set). Each of the seven lists of M conditional probability values described above, $\{CP^{(K,N)}{}_{Ts_i,Ts_j}\ \forall\ K\ \&\ CP^{(\Pi,N)}{}_{Ts_i,Ts_j}\}_{j>i}$, were rank ordered. Next, the success rate

$$\text{SuccessRate}(m) = \frac{\sum_{j>i} I_{\{\Delta A^{(N)}Ts_i,Ts_j \le A^*\}} I_{\{Rank\ of\ CP^{(K,N)}Ts_i,Ts_j \le m\}}}{m} \quad (5)$$

was calculated for each comparison method and each value of m from 1 to M. One thousand total cross-validated trials were performed, and the median value of the success rate for each value of m from 1 to M was calculated. In order to accurately compare data sets with differing numbers of total comparisons, the theoretical total number of pairs with $\Delta A \le A^*$ in the test set (i.e., the theoretical maximum number of successes within the test set), denoted M^*, was calculated for each data set and comparison method. A single illustrative example is of the above process is shown in Figure 2, for the OR data set against the DOR target using the Radial structural similarity method and $A^* = 0.5$. Success rates are plotted in Figure 2A from $m = 1$ to M^*; as shown in Table 1 there are 1176 total comparisons in the OR test set, of which 42.6% have an activity difference less than or equal to 0.5 for the DOR target, so in this case $M^* = 1176 * 0.426 = 501$. Specific success rate values were tabulated as a percentage over the random baseline, for $m = \lceil 0.01M^* \rceil$, $\lceil 0.05M^* \rceil$, $\lceil 0.10M^* \rceil$, $\lceil 0.25M^* \rceil$, and $\lceil 0.50M^* \rceil$. Figure 2C shows an example of this tabulation; to continue the above example, since $M^* = 501$ for the OR test set against the DOR target, $\lceil 0.10M^* \rceil$ would be 51 comparisons. In this example, over the 1,000 trials run, the median success rate for the top 51 comparisons was 66.5% (as plotted in Figure 2A as an unadjusted success rate). Because the random rate was 42.6%, Figure 2C shows 56.2% under the heading of 10%, since 66.5% is 56.2% greater than 42.6%. Although this method of tabulating success rates is somewhat more complicated than the unadjusted versions plotted in Figure 2A, this method normalizes for differences in both the total number of comparisons in the test set and the baseline random success rate. This normalization allows for meaningful comparisons between receptor-compound set pairs.

Relative Activity Prediction Heuristic. For the N^{th} trial associated with a given compound-receptor set pair, let J be the number of compounds in the testing (and therefore also the training) set. In order to determine the ability for the activity landscape information to prioritize the potential activities of compounds within the testing set, for the K^{th} structural comparison method and the j^{th} compound in the testing set, the Activity Priority Score was calculated as a weighted average of the conditional probabilities associated with the comparisons between that compound and the training set compounds:

$$APS_j^{(K,N)} = \frac{\sum_{i=1,\ldots,J} A_i^{(N)} \cdot CP^{(K,N)}{}_{Tr_i,Ts_j}}{\sum_{i=1,\ldots,J} CP^{(K,N)}{}_{Tr_i,Ts_j}} \quad (6)$$

Here $CP^{(K,N)}{}_{Tr_i,Tr_j}$, defined above, is the conditional probability value associated with the comparison between the j^{th} compound in the testing set and the i^{th} compound in the training set, and $A_i^{(N)}$ is the activity of the i^{th} compound in the training set of the N^{th} trial. In a given trial, the test set compounds are then rank ordered by $APS_j^{(K,N)}$ and sorted along with the actual value of their activity. The median value of the true activity of each compound in the test set at a given rank over all 1,000 trials was then calculated. The results were plotted along with the value associated with random ranking of compounds and the true activity distribution. To continue the above example, see Figure 2B; the rightmost value on each graph represents the median true activity of the top-scored compound in the test set over the 1,000 trials performed for a given comparison method. The average increases and decreases in the top and bottom 10% of these median values were also tabulated in two ways. First, the average percentage of the possible increase or decrease (i.e., as a percentage of the gap between the random value and the actual activity distribution) was calculated with equation 4:

$$\%_{MaxChange} = \frac{TestSetActivity - RandomActivity}{ActualActivity - RandomActivity} \quad (7)$$

Second, the absolute average change in activity (in units of pKi) was calculated. An example of this tabulation is in Figure 2D; for this continuing example, there are 49 compounds in the test set to be ranked for potential activity. Figure 2D indicates that, over the 1,000 total trials, the median true activity of the compounds in the top 10% of these rankings was 33.7% of the maximum possible improvement; this is evident in Figure 2B, with Radial averaging about one-third of the way between the line denoting the true activities (red, dotted) and the line denoting randomness (gray, dashed) over the five right-most points. Figure 2D also indicates that this 33.7% of maximal possible improvement corresponds to an absolute increase in pKi of 0.34. Again, these normalizations are meant to allow for meaningful comparisons between data sets of varying size and activity.

DISCUSSION

Cross-Validation and Parameter Considerations. Over all 1,000 trials, all compound comparison methods, and all data sets, the results of the training-to-testing set analysis described

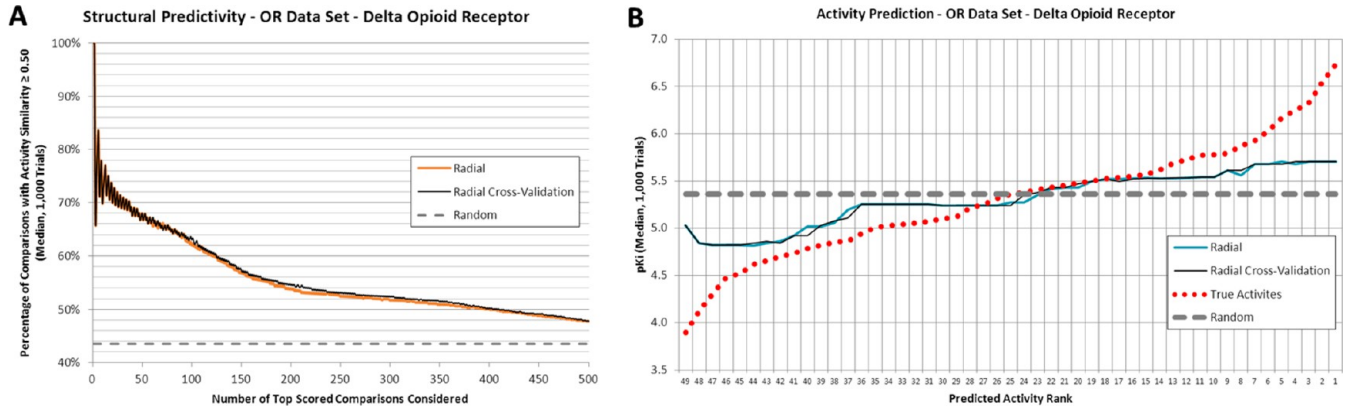


Figure 3. A demonstration of the strong correspondence between the 1,000 validation and cross-validation trials.

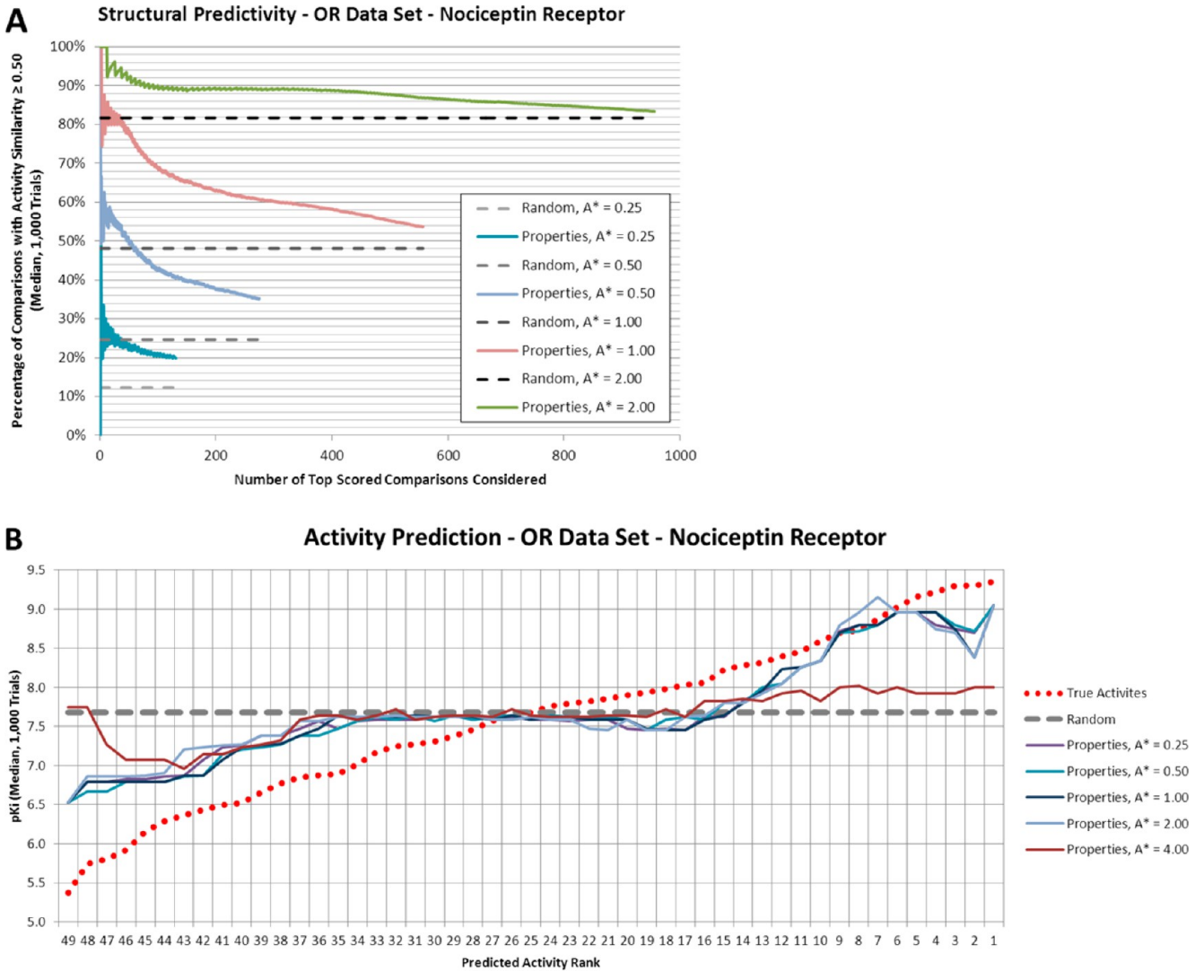


Figure 4. The structural predictivity and activity prioritization methodologies are robust with respect to the choice of activity threshold *A**.

above corresponded to the cross-validation results when the training and testing sets were swapped. This indicates that all results presented herein are not contingent upon the manner in which the training and testing sets were randomly assigned. See Figure 3. Random and perfect control data sets performed as expected in all cases. Furthermore, it was determined that both the structural predictivity and compound prioritization methodologies described above are exceptionally robust to the choice of *A**, as exemplified in Figure 4, which shows the Properties comparison method applied to the OR data set and the NOC target. In Figure 4A, we see that the choice of *A** naturally effects the number of pairs qualifying as a "success" and therefore increases the baseline random level and the theoretical maximum number of success as *A** increases. In

dx.doi.org/10.1021/ci400243e | J. Chem. Inf. Model. 2013, 53, 2613−2625

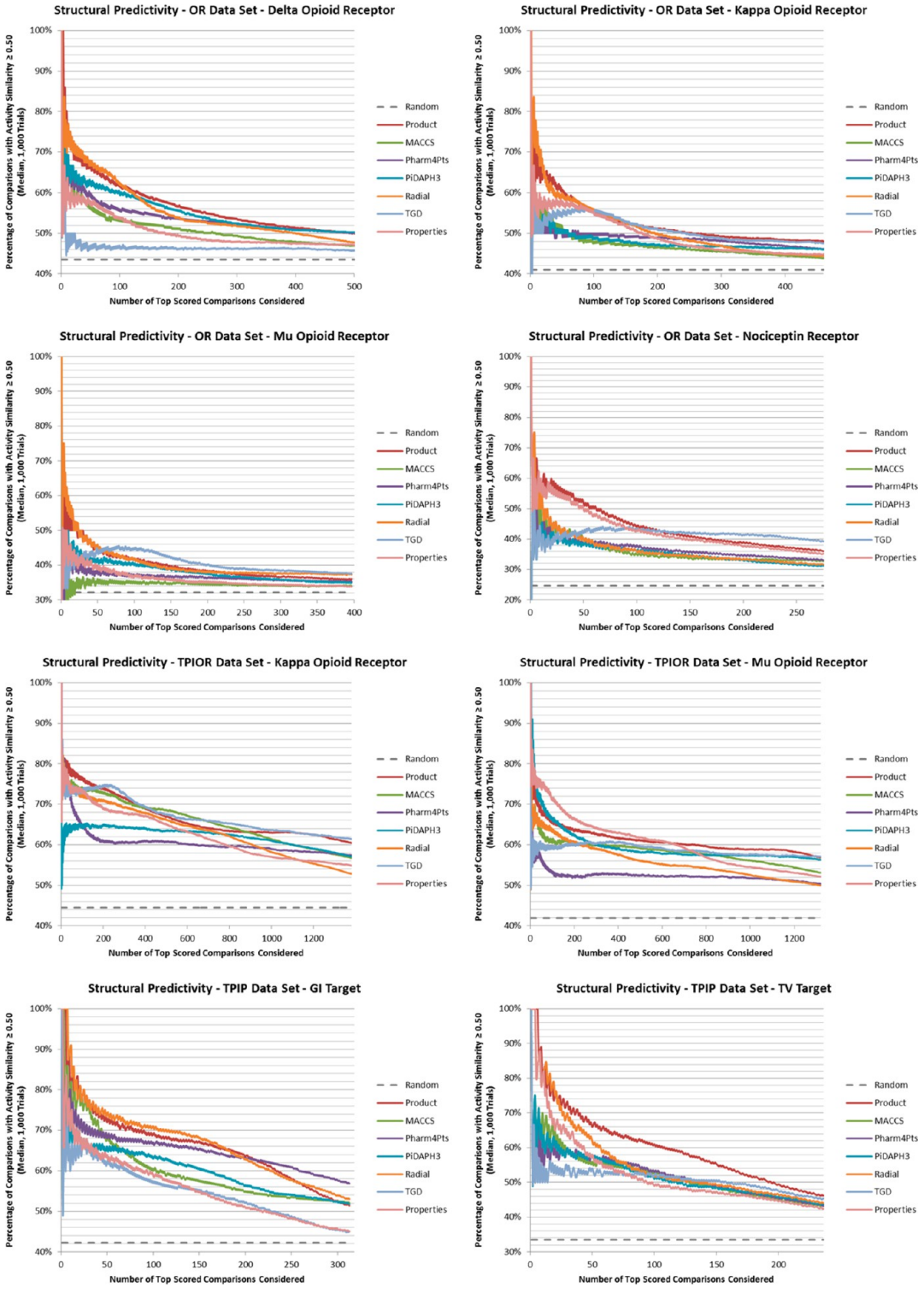


Figure 5. All structural predictivity results plotted, for all data sets, targets, and comparison methods.

all cases, however, success rates are clearly superior to random. In Figure 4B, we see that the activity prioritization results remain fairly consistent for all but the most extreme choices of A^*; this result is unsurprising, since we are focusing on rank-order relative activities rather than absolute values of $APS_j^{(K,N)}$. Because of this, the choice of $A^* = 0.5$ was used for all of the following analysis, so that the activity difference associated with a successfully close pair would correspond to an order of magnitude interval length.

Structural Predictivity. *Comparison of Data Sets.* As shown in Figure 5 and Table 2, the ability to predict that two compounds in the test set would have similar activities varies widely, both with regards to receptor-compound set pairing and compound comparison method. Overall, all compound comparison methods against all receptor-compound set pairs exhibited median success rates above random when considering a sufficiently large number of comparisons; in almost all cases, random success rates were outperformed for any number of

dx.doi.org/10.1021/ci400243e | J. Chem. Inf. Model. 2013, 53, 2613−2625

Table 2. All Structural Predictivity and Activity Prediction Results Tabulated for All Data Sets, Targets, and Comparison Methods

Compound Set	Target	Comparison Method	Median Success Rates Over Random in Predicting Activity Similarity					Active Compound Prediction			
			Portion of Top-Ranked Comparisons Considered, as % of Theoretical Maximum					Median % of Maximum Possible Change (Over Random)		Median Activity Increase/Decrease (pKi) from Random	
			1%	5%	10%	25%	50%	Average of Predicted Bottom 10%	Average of Predicted Top 10%	Average of Predicted Bottom 10%	Average of Predicted Top 10%
OR	DOR	Product	83.8%	56.2%	51.6%	37.8%	26.3%	-61.4%	47.0%	-0.68	0.49
		MACCS	83.8%	37.8%	28.6%	21.3%	15.3%	-40.1%	23.2%	-0.43	0.25
		Pharm4Pts	37.8%	47.0%	37.8%	26.8%	21.7%	-53.3%	24.5%	-0.57	0.24
		PiDAPH3	37.8%	47.0%	42.4%	36.0%	23.5%	-50.1%	49.5%	-0.54	0.46
		Radial	83.8%	65.4%	56.2%	37.8%	20.8%	-48.0%	33.7%	-0.49	0.34
		TGD	37.8%	10.3%	5.7%	6.6%	6.2%	-30.6%	13.4%	-0.32	0.15
		Properties	37.8%	37.8%	33.2%	19.4%	10.7%	-64.7%	24.7%	-0.68	0.28
	KOR	Product	46.4%	59.1%	48.5%	32.5%	23.6%	-59.3%	71.2%	-0.66	0.76
		MACCS	46.4%	37.9%	27.3%	15.7%	13.0%	-37.1%	58.0%	-0.40	0.64
		Pharm4Pts	46.4%	27.3%	22.0%	22.0%	18.3%	-13.4%	57.3%	-0.13	0.57
		PiDAPH3	46.4%	27.3%	22.0%	17.8%	15.1%	-20.6%	64.4%	-0.19	0.65
		Radial	95.2%	59.1%	43.2%	32.5%	19.4%	-47.5%	71.1%	-0.51	0.76
		TGD	46.4%	27.3%	32.6%	34.6%	22.5%	-8.7%	37.4%	-0.06	0.44
		Properties	46.4%	37.9%	37.9%	32.5%	15.1%	-43.7%	57.2%	-0.48	0.63
	MOR	Product	55.6%	55.6%	47.8%	28.8%	18.8%	-60.6%	30.2%	-0.90	0.39
		MACCS	-22.2%	8.9%	8.9%	10.0%	7.9%	-18.3%	26.5%	-0.27	0.35
		Pharm4Pts	-22.2%	24.4%	16.7%	16.3%	12.6%	-29.1%	22.9%	-0.46	0.27
		PiDAPH3	55.6%	40.0%	32.2%	25.7%	17.3%	-27.0%	26.4%	-0.42	0.30
		Radial	133.3%	55.6%	47.8%	28.8%	18.8%	-58.3%	46.6%	-0.87	0.68
		TGD	55.6%	24.4%	32.2%	38.3%	25.1%	-49.5%	10.7%	-0.75	0.02
		Properties	55.6%	24.4%	24.4%	13.1%	9.4%	-33.6%	28.3%	-0.48	0.39
	NOC	Product	170.3%	131.7%	131.7%	99.8%	67.5%	-57.1%	83.5%	-1.05	1.33
		MACCS	35.2%	73.8%	73.8%	52.8%	38.1%	-45.3%	53.3%	-0.84	0.84
		Pharm4Pts	35.2%	73.8%	59.3%	58.7%	46.9%	-36.6%	63.9%	-0.70	1.01
		PiDAPH3	35.2%	73.8%	59.3%	52.8%	41.0%	-25.6%	51.3%	-0.47	0.81
		Radial	170.3%	102.8%	73.8%	52.8%	41.0%	-46.5%	74.3%	-0.86	1.18
		TGD	35.2%	44.8%	59.3%	76.3%	70.4%	-50.6%	65.8%	-0.93	1.04
		Properties	170.3%	131.7%	117.2%	88.1%	64.6%	-53.0%	76.9%	-0.99	1.22

dx.doi.org/10.1021/ci400243e | J. Chem. Inf. Model. 2013, 53, 2613−2625

Table 2. continued

Compound Set	Target	Comparison Method	Median Success Rates Over Random in Predicting Activity Similarity					Active Compound Prediction			
			Portion of Top-Ranked Comparisons Considered, as % of Theoretical Maximum					Median % of Maximum Possible Change (Over Random)		Median Activity Increase/Decrease (pKi) from Random	
			1%	5%	10%	25%	50%	Average of Predicted Bottom 10%	Average of Predicted Top 10%	Average of Predicted Bottom 10%	Average of Predicted Top 10%
TPIOR	KOR	Product	76.6%	72.6%	69.4%	58.3%	44.4%	-95.1%	36.0%	-0.70	0.43
		MACCS	76.6%	66.1%	66.1%	57.0%	48.0%	-95.1%	35.8%	-0.70	0.35
		Pharm4Pts	76.6%	53.1%	40.1%	36.2%	34.3%	-95.1%	57.8%	-0.70	0.67
		PiDAPH3	44.5%	46.6%	44.9%	44.0%	41.5%	-79.7%	25.2%	-0.60	0.31
		Radial	76.6%	62.9%	61.2%	54.4%	43.1%	-95.1%	25.5%	-0.70	0.25
		TGD	76.6%	62.9%	66.1%	59.0%	48.0%	-95.1%	34.1%	-0.70	0.45
		Properties	76.6%	66.1%	61.2%	51.8%	38.6%	-95.1%	0.7%	-0.70	0.01
	MOR	Product	83.5%	62.6%	55.4%	48.9%	43.8%	-100.0%	44.3%	-0.70	0.44
		MACCS	65.1%	48.2%	44.5%	42.4%	39.1%	-0.2%	33.8%	0.00	0.30
		Pharm4Pts	46.8%	30.1%	24.7%	26.5%	24.7%	0.0%	36.6%	0.00	0.37
		PiDAPH3	101.8%	62.6%	55.4%	42.4%	38.0%	0.0%	28.5%	0.00	0.30
		Radial	65.1%	55.4%	48.2%	40.2%	30.8%	-14.6%	37.7%	-0.10	0.34
		TGD	46.8%	40.9%	40.9%	44.5%	39.5%	-100.0%	6.0%	-0.70	0.07
		Properties	83.5%	77.1%	66.2%	51.8%	43.1%	-100.0%	5.3%	-0.70	0.05
TPIP	GI	Product	136.7%	92.4%	83.3%	66.9%	58.3%	-81.7%	62.3%	-1.18	0.39
		MACCS	136.7%	77.6%	75.6%	51.8%	35.7%	-65.9%	58.5%	-0.98	0.37
		Pharm4Pts	57.8%	77.6%	68.0%	60.9%	53.8%	-78.0%	58.5%	-1.12	0.37
		PiDAPH3	57.8%	62.8%	52.7%	51.8%	41.7%	-74.2%	62.3%	-1.08	0.39
		Radial	136.7%	92.4%	83.3%	70.0%	59.8%	-71.2%	46.8%	-1.05	0.29
		TGD	57.8%	48.0%	52.7%	39.6%	29.7%	-58.5%	58.5%	-0.89	0.37
		Properties	57.8%	62.8%	52.7%	42.7%	28.2%	-69.5%	62.8%	-1.02	0.39
	TV	Product	198.8%	149.0%	124.1%	97.5%	77.2%	-83.6%	78.2%	-1.46	1.16
		MACCS	49.4%	99.2%	86.7%	67.1%	54.5%	-44.8%	60.9%	-0.90	0.89
		Pharm4Pts	49.4%	74.3%	74.3%	72.2%	54.5%	-78.7%	42.4%	-1.39	0.60
		PiDAPH3	49.4%	74.3%	74.3%	67.1%	49.4%	-88.6%	48.3%	-1.57	0.67
		Radial	198.8%	149.0%	111.6%	77.2%	51.9%	-76.9%	64.2%	-1.35	0.92
		TGD	49.4%	49.4%	61.8%	57.0%	54.5%	-21.1%	39.9%	-0.43	0.56
		Properties	198.8%	124.1%	99.2%	62.1%	44.3%	-62.1%	78.2%	-1.07	1.16

comparisons greater than ten. Results were often very specific to the comparison method and data set considered; Radial, when used to predict activity differences in the OR data set and the MOR target, was 133.3% more accurate than random when considering 1% of the maximum possible number of successes. This was 2.5-fold more effective than any other method for this target and compound set. Drop-off of accuracy as more comparisons were considered was also very specific; when considering the OR data set and the NOC target, Radial and Properties declined in effectiveness when increasing the number of considered comparisons from 1% of the theoretical maximum to 5%, while MACC, PiDAPH3, and Pharm4Pts improved markedly and TGD improved only slightly. As evident by the heat-mapping of Table 2, however, there were some trends of overall relative effectiveness. Table 3 shows the structural predictivity (at the same values of m as Table 2) for each receptor-compound set pair averaged over all compound similarity methods (excluding product). As is evident, the TPIP-TV, TPIP-GI, and OR-Nociceptin pairings produced the most accurate predictions; average success rates were almost twice that of random for small numbers of comparisons and were still about 50% better than random when hundreds of

dx.doi.org/10.1021/ci400243e | J. Chem. Inf. Model. 2013, 53, 2613−2625

Table 3. All Structural Predictivity and Activity Prediction Results Tabulated for Each Data Set, Averaged over All Compound Comparison Methods

		Averages Over All Similarity Methods								
		Median Success Rates Over Random in Predicting Activity Similarity					Active Compound Prediction			
Compound Set	Target	Portion of Top-Ranked Comparisons Considered, as % of Theoretical Maximum					Median % of Maximum Possible Change (Over Random)		Median Activity Increase/Decrease (pKi) from Random	
		1%	5%	10%	25%	50%	Average of Predicted Bottom 10%	Average of Predicted Top 10%	Average of Predicted Bottom 10%	Average of Predicted Top 10%
OR	DOR	53.1%	40.9%	34.0%	24.6%	16.4%	-47.8%	28.2%	-0.51	0.29
	KOR	54.5%	36.1%	30.8%	25.8%	17.2%	-28.5%	57.6%	-0.30	0.62
	MOR	42.6%	29.6%	27.0%	22.0%	15.2%	-36.0%	26.9%	-0.54	0.33
	NOC	80.2%	83.4%	73.8%	63.6%	50.4%	-42.9%	64.2%	-0.80	1.01
TPIOR	KOR	71.2%	59.6%	56.6%	50.4%	42.2%	-92.5%	29.9%	-0.68	0.34
	MOR	68.2%	52.4%	46.6%	41.3%	35.9%	-35.8%	24.6%	-0.25	0.24
TPIP	GI	84.1%	70.2%	64.2%	52.8%	41.5%	-69.6%	57.9%	-1.02	0.36
	TV	99.2%	95.0%	84.7%	67.1%	51.5%	-62.0%	55.7%	-1.12	0.80

Table 4. All Structural Predictivity and Activity Prediction Results Tabulated for Each Compound Comparison Method, Averaged over All Compound-Receptor Set Pairs

	Averages Over All Data Sets								
	Median Success Rates Over Random in Predicting Activity Similarity					Active Compound Prediction			
	Portion of Top-Ranked Comparisons Considered, as % of Theoretical Maximum					Median % of Maximum Possible Change (Over Random)		Median Activity Increase/Decrease (pKi) from Random	
Structural Method	1%	5%	10%	25%	50%	Average of Predicted Bottom 10%	Average of Predicted Top 10%	Average of Predicted Bottom 10%	Average of Predicted Top 10%
Product	106.5%	84.9%	76.5%	58.8%	45.0%	-74.8%	56.6%	-0.92	0.67
MACCS	58.9%	56.2%	51.5%	39.7%	31.4%	-43.3%	43.7%	-0.56	0.50
Pharm4Pts	41.0%	50.9%	42.9%	39.9%	33.3%	-48.0%	45.5%	-0.63	0.51
PiDAPH3	53.6%	54.3%	47.9%	42.2%	33.5%	-45.7%	44.5%	-0.61	0.49
Radial	120.0%	80.3%	65.7%	49.2%	35.7%	-57.3%	50.0%	-0.74	0.60
TGD	50.7%	38.5%	43.9%	44.5%	37.0%	-51.8%	33.2%	-0.60	0.39
Properties	90.8%	70.2%	61.5%	45.2%	31.8%	-65.2%	41.8%	-0.76	0.52

comparisons were considered. Other comparisons using the OR data set were not as successful, being about as accurate in their most highly ranked comparisons as the best data sets were over hundreds of theirs. The TPIOR data sets' results lay somewhere in the middle. In particular, these analyses show that the underlying structures involved alone did not determine the successfulness of the method, since the OR data set produced both some of the most accurate comparison predictions with Nociceptin and some of the least accurate with MOR. Nor does the receptor alone necessarily dictate effectiveness, since the TPIOR and OR data sets saw substantial differences in structural predictivity against the KOR and MOR targets. There were also no clear patterns regarding the effect of the overall activity distributions of the data set on effective predictions. OR-Nociceptin had both the lowest percentage of activity comparisons ≤0.5 pKi and the largest average activity difference among its comparison (see Table 1), but TPIP-TV and OR-MOR had similar numbers with drastically different success rates. TPIP-GI showed median values in both percentage of activity comparisons ≤0.5 pKi (42.2% from Table 1) and average activity difference (0.76 from Table 1) but was almost as effective as OR-Nociceptin which had a low value for percentage of activity comparisons (23.4%, Table 1) and high average activity difference (1.20, Table 1). The presented methodology for structural prediction in this study shows at least some median effectiveness over all receptor-compound set pairs, but relative effectiveness across all compound comparison methods is particular to each data set.

Comparison of Similarity Methods. Structural predictivity success rate, as defined in equation 5, essentially measures the ability for a given compound comparison method to describe the structural-activity landscape of a data set in a manner that persists when applied to a different data set. One can view this as a measure of how dependent a given SAR profile is on the compounds contained in that profile. It is therefore useful to examine the compound similarity methods' relative average performance across all receptor-compound set pairs. The results are in Table 4. Using the product conditional probability

dx.doi.org/10.1021/ci400243e | J. Chem. Inf. Model. 2013, 53, 2613−2625

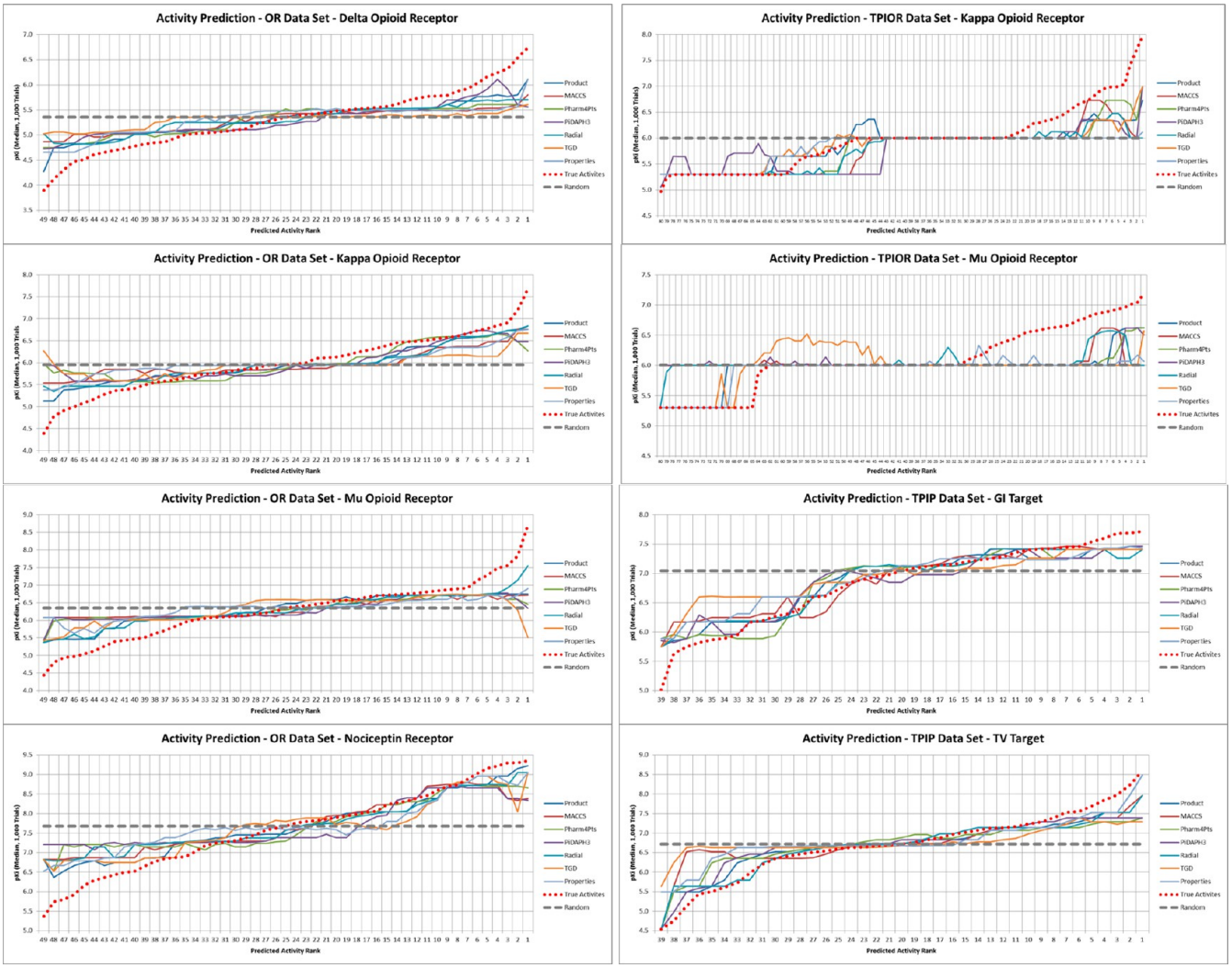


Figure 6. All activity prediction results plotted, for all data sets, targets, and comparison methods.

as a consensus metric consistently performed better than using almost any method individually, regardless of how many comparisons were considered. The only exception is Radial, which was superior to all other methods when looking at only the few most highly ranked comparisons. This is due to Radial's uniquely effective predictivity of the OR-DOR, OR-KOR, and OR-MOR data sets. Of the individual methods considered in this study, Radial showed the strongest overall structural predictivity at multiple numbers of comparisons, followed by Property Similarity (Properties). TGD and Phram4Pts generally displayed the worst overall performance. At the level of 50% of the maximum total number of possible hits, all individual methods essentially performed the same.

Relative Activity Prediction. Both the top and bottom 10% ranking compounds' median true activities were analyzed in this study, as described above; results are shown in Figure 6 and Table 2. High activity relative to random in the top 10% of ranking compounds indicates an ability to find more active compounds using the described methodology than through random chance alone; low activity relative to random in the bottom 10% of ranking compounds indicates an ability to find compounds that are especially unlikely to be active. Both scenarios could be valuable in the prioritization of compounds in large libraries. In both analysis of the top and bottom 10%, no comparison method's median performance was worse than random, indicating that in the long run the use of this method will not negatively impact prioritization of potentially active compounds. Overall effectiveness, both as a percentage of possible effectiveness and in terms of absolute pKi, varied widely from data set to data set and comparison methodology to comparison methodology. Effectiveness was also, in general, asymmetric with respect to the top and bottom 10%, with almost all data sets showing reasonably strong results for at least some comparison methods. Interestingly, effectiveness in predicting relative activities only correlated somewhat with structural predicitivity ability; the highest correlation found was between absolute average pKi difference (Table 2, eighth and ninth numerical column) and the success rate when considering 10% of the maximum number of possible successes (Table 2, third numerical column), with an $R^2 = 0.36$ for the top 10% of ranked compounds and an $R^2 = 0.51$ for the bottom 10%. See Figure 7. That there would be some correlation is not at all surprising, since the weighted average in equation 7 assumes, just as the structural predictivity method does, that the conditional probabilities in the training set will accurately describe relationships in the testing set. It is more surprising that the correlation between methods is not higher; the OR-KOR data set, for example, resulted in predictions that were

dx.doi.org/10.1021/ci400243e | J. Chem. Inf. Model. 2013, 53, 2613−2625

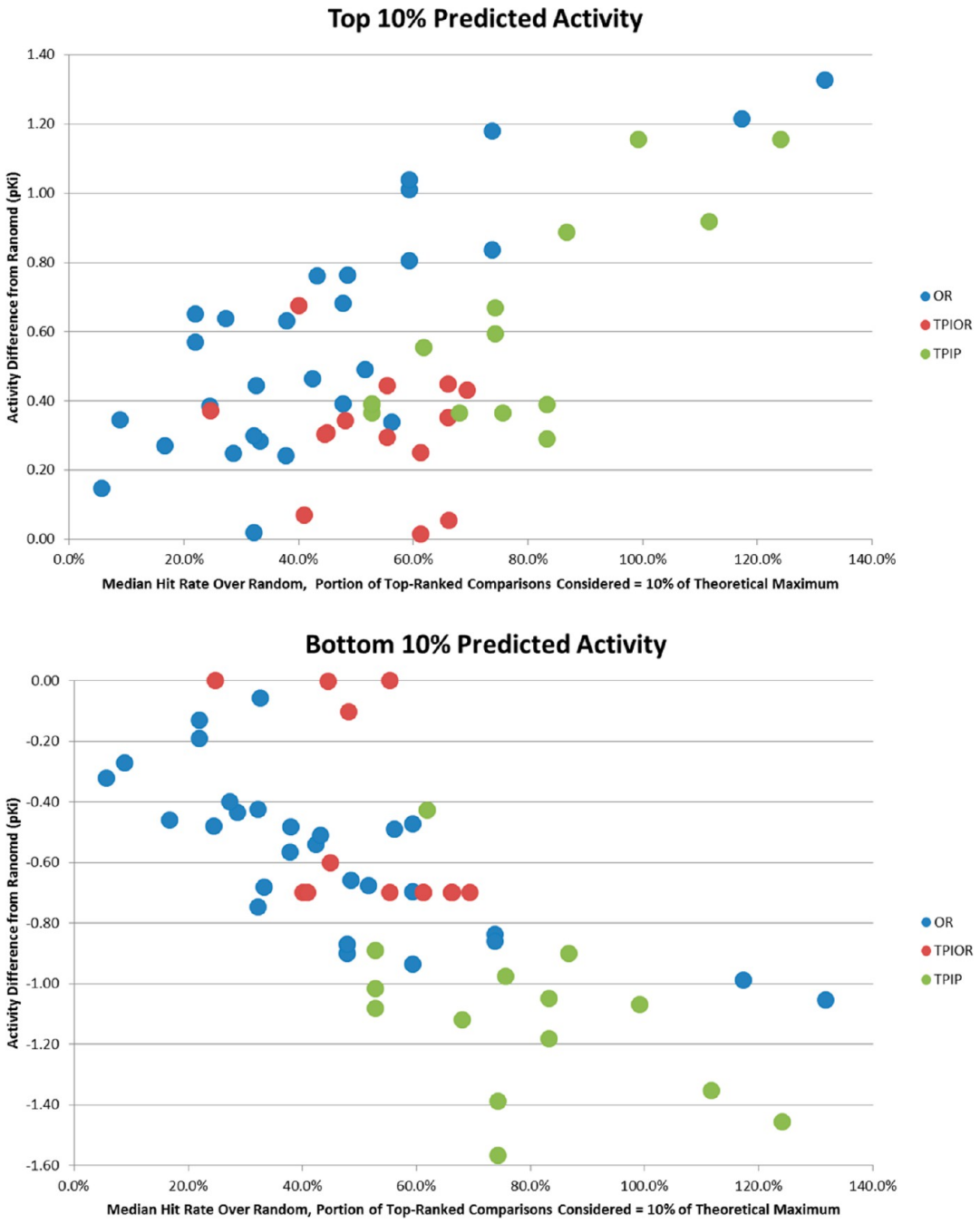


Figure 7. Structural predictivity plotted against activity prediction.

50% or better than the best prediction possible for the top 10% ranked compounds using almost all compound similarity methods, in spite of the fact that it was one of the worst performers overall in structural predictivity (see Table 3). Most likely, this discrepancy is due to specific active or inactive compounds that are more clearly differentiated from the data set as a whole, rather than the median compound.

Comparison of Data Sets. Table 3 summarizes the results averaged by data set (again excluding Product). On average, compounds ranked in the top 10% had a median true activity of a full order of magnitude better than random for the OR-Nociceptin data set (1.01), and at least 3-fold better for the OR-KOR and TPIP-TV data sets as well (0.62 and 0.80, respectively). All data sets averaged at least a 0.24 improvement in median true pKi or a 74% improvement in activity. On average, compounds ranked in the bottom 10% had a median true activity of a full order of magnitude lower for both the TPIP-TV and TPIP-GI data sets (−1.12 and −1.02, respectively) and an over 6-fold reduction in activity for the OR-Nociceptin data set (−0.80). All data sets averaged at least a 0.25 reduction in median true pKi. Because of the ranges of activity in each data set, larger absolute activity differences did not always correspond to greater percentages of the best possible prediction. The TPIOR-KOR data set, for example, had predictions in the bottom 10% that were almost as far below the random value as possible (92.5%); however, it had only a 0.68 reduction in median true pKi. This is due to the narrower range of activities in the TPIOR-KOR data set relative

dx.doi.org/10.1021/ci400243e | J. Chem. Inf. Model. 2013, 53, 2613−2625

to the other data sets. Similarly OR-KOR showed increases in activity relative to random in the top 10% of ranked compounds on par with those of TPIP-TV and TPIP-GI when viewed as a percentage of maximum possible increase.

Comparison of Similarity Methods. Results averaged by compound comparison method are in Table 4. Using the product of the conditional probabilities derived from the training set was, on average, more effective than any singular compound comparison method. For Product, compounds ranked in the top 10% averaged a median improvement of 0.67 pKi (4.7-fold) over random and averaged an improvement of 56.6% of the maximum improvement possible. Compounds ranked in the bottom 10%, on average, showed almost a full order of magnitude median decrease (−0.92 pKi) in activity relative to random; this decrease is equivalent to 74.8% of the maximum possible decrease. For individual compound comparison methods, Radial once again performed the best, having an average median improvement over random in the top 10% ranked compounds of 0.08 pKi more than any other method. Properties performed marginally better than Radial when considering the bottom 10% of ranked compounds but did not distinguish itself when considering the top 10% of ranked compounds. TGD was by far the worst at predicting active compounds using the top 10% of ranked compounds, but MACCS was the worst at predicting inactive compounds using the bottom 10% of ranked compounds. However in both cases these methods were still substantially better than random.

To demonstrate the limitations of the conditional probability prediction method, the OR-KOR and OR-MOR data sets were used, respectively, as training sets in an attempt to rank activities of the TPIOR-KOR and TPIOR-MOR data set compounds. The median average structural difference, over the various compound similarity metrics, between these data sets is 40% higher than the average structural difference within either data set; it was expected that because these sets of compounds were significantly different from one another altogether, that the method would fail, and indeed rankings were indistinguishable from random. To confirm this, the TPIOR-KOR and TPIOR-MOR data sets were divided into two specific subsets that had 50% more structural difference (on average) between them than either subset. Once again, predictions were indistinguishable from random. It is unsurprising that sufficient differences between the internal and external diversities would cause the method to fail. By showing that median values of randomly sampled trials were generally successful, however, it implies that when using this methodology, it is sufficient for the training and testing sets to be about as diverse internally as they are between themselves in order to have a reasonable likelihood of success.

CONCLUSION

Compound comparison methods focusing on structural similarity or property similarity have been widely used both to describe aspects of the activity landscape and to prioritize compounds for testing through comparisons to specific compounds. In this study, we have presented a new, probabilistic approach that systematically both determines the ability of compound comparison methods to capture meaningful information about the activity landscape and uses that information to prioritize compounds based on their likelihood of similarity to other active compounds. We have demonstrated over three separate compound sets and eight distinct compound-target pairs that in all cases at least one compound comparison method (and often many) was able to accurately predict compounds with similar activity in the test set and accurately create compound rankings that trended from more to less active. We have demonstrated the overall superiority of certain methods, such as Radial and Properties, to describe the activity landscape in a relevant manner when later considering novel compounds. Although it has not yet been determined what underlying mechanism drives the relative success rates of different comparison methods, we have shown that no method behaved worse than random, meaning that there is no drawback to using this method as an alternative to random sampling. Improvements by as much as an order of magnitude were seen in the rankings, and it is our belief that we have demonstrated this method's wide applicability over diverse sets of compounds.

ASSOCIATED CONTENT

Supporting Information

A schematic of the entire 2-fold validation process used in this study (Figure S1). This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Phone: +1-772-345-4734. Fax: +1-772-345-3649. E-mail: rsantos@tpims.org.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Discussions with Dr. Rafael Castillo and M.Sc. Jaime Pérez-Villanueva are greatly appreciated. This work was supported by NIH Grant 1R01DA031370 and the State of Florida, Executive Office of the Governor's Department of Economic Opportunity.

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) 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.

(3) Medina-Franco, J. L.; Yongye, A. B.; López-Vallejo, F. Consensus Models of Activity Landscapes. In *Statistical Modeling of Molecular Descriptors in QSAR/QSPR*; Matthias, D., Kurt, V., Danail, B., Eds.; Wiley-VCH: 2012; pp 307−326.

(4) 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.

(5) Bajorath, J. Modeling of Activity Landscapes for Drug Discovery. *Expert Opin. Drug Discovery* **2012**, *7*, 463−473.

(6) Guha, R. Exploring Structure−Activity Data Using the Landscape Paradigm. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 829−841.

(7) 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.

(8) 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.

(9) Medina-Franco, J. L. Scanning Structure-Activity Relationships with SAS and Related Maps: From Consensus Activity Cliffs to Selectivity Switches. *J. Chem. Inf. Model.* **2012**, *52*, 2485−2493.
(10) Martínez-Mayorga, K.; Peppard, T. L.; Yongye, A. B.; Santos, R.; Giulianotti, M.; Medina-Franco, J. L. Characterization of a Comprehensive Flavor Database. *J. Chemom.* **2011**, *25*, 550−560.
(11) Ooms, F. Molecular Modeling and Computer Aided Drug Design. Examples of Their Applications in Medicinal Chemistry. *Curr. Med. Chem.* **2000**, *7*, 141−158.
(12) Tropsha, A. Best Practices for QSAR Model Development, Validation, and Exploitation. *Mol. Inf.* **2010**, *29*, 476−488.
(13) Sanders, M. P. A.; Barbosa, A. J. M.; Zarzycka, B.; Nicolaes, G. A. F.; Klomp, J. P. G.; de Vlieg, J.; Del Rio, A. Comparative Analysis of Pharmacophore Screening Tools. *J. Chem. Inf. Model.* **2012**, *52*, 1607−1620.
(14) Scior, T.; Medina-Franco, J. L.; Do, Q. T.; Martínez-Mayorga, K.; Yunes Rojas, J. A.; Bernard, P. How to Recognize and Workaround Pitfalls in QSAR Studies: A Critical Review. *Curr. Med. Chem.* **2009**, *16*, 4297−4313.
(15) Guha, R. Exploring Uncharted Territories: Predicting Activity Cliffs in Structure−Activity Landscapes. *J. Chem. Inf. Model.* **2012**, *52*, 2181−2191.
(16) Heikamp, K.; Hu, X.; Yan, A.; Bajorath, J. Prediction of Activity Cliffs Using Support Vector Machines. *J. Chem. Inf. Model.* **2012**, *52*, 2354−2365.
(17) 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.
(18) Houghten, R. A.; Pinilla, C.; Giulianotti, M. A.; Appel, J. R.; Dooley, C. T.; Nefzi, A.; Ostresh, J. M.; Yu, Y. P.; Maggiora, G. M.; Medina-Franco, J. L.; Brunner, D.; Schneider, J. Strategies for the Use of Mixture-Based Synthetic Combinatorial Libraries: Scaffold Ranking, Direct Testing, in Vivo, and Enhanced Deconvolution by Computational Methods. *J. Comb. Chem.* **2008**, *10*, 3−19.
(19) 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. *Biorg. Med. Chem.* **2010**, *18*, 7380−7391.
(20) Guha, R.; VanDrie, J. H. Structure-Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646−658.
(21) 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.
(22) 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.
(23) Medina-Franco, J. L.; Waddell, J. Towards the Bioassay Activity Landscape Modeling in Compound Databases. *J. Mex. Chem. Soc.* **2012**, *56*, 163−168.
(24) Yongye, A. B.; Medina-Franco, J. L. Data Mining of Protein-Binding Profiling Data Identifies Structural Modifications That Distinguish Selective and Promiscuous Compounds. *J. Chem Inf. Model.* **2012**, *52*, 2454−2461.
(25) 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. *MedChemComm* **2011**, *2*, 44−49.
(26) Chen, B.; Sheridan, R. P.; Hornak, V.; Voigt, J. H. Comparison of Random Forest and Pipline Pilot Naïve Bayes in Prospective QSAR Predictions. *J. Chem. Inf. Model.* **2012**, *52*, 792−803.
(27) Chu, C.-W.; Holliday, J. D.; Willett, P. Combining Multiple Classifications of Chemical Structures Using Consensus Clustering. *Bioorg. Med. Chem.* **2012**, *20*, 5366−5371.
(28) 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.