Assessing How Well a Modeling Protocol Captures a Structure−Activity Landscape

Rajarshi Guha

School of Informatics, Indiana University, Bloomington, Indiana 47406

John H. Van Drie*

Van Drie Research LLC, Andover, Massachusetts 01810

Received April 23, 2008

We introduce the notion of structure−activity landscape index (SALI) curves as a way to assess a model and a modeling protocol, applied to structure−activity relationships. We start from our earlier work [*J. Chem. Inf. Model.*, **2008**, *48*, 646−658], where we show how to study a structure−activity relationship *pairwise*, based on the notion of "activity cliffs"—pairs of molecules that are structurally similar but have large differences in activity. There, we also introduced the SALI parameter, which allows one to identify cliffs easily, and which allows one to represent a structure−activity relationship as a graph. This graph orders every pair of molecules by their activity. Here, we introduce the new idea of a SALI curve, which tallies how many of these orderings a model is able to predict. Empirically, testing these SALI curves against a variety of models, ranging over two-dimensional quantitative structure−activity relationship (2D-QSAR), three-dimensional quantitative structure−activity relationship (3D-QSAR), and structure-based design models, the utility of a model seems to correspond to characteristics of these curves. In particular, the integral of these curves, denoted as SCI and being a number ranging from −1.0 to 1.0, approaches a value of 1.0 for two literature models, which are both known to be prospectively useful.

1. INTRODUCTION

Modeling a structure−activity relationship is a triumph of hope over experience.

The "gold standard" for any model is its prospective utility. By the term "prospective", we mean applying a model to molecules that neither the model nor the modeler has seen previously. Finding models that seem to work well retrospectively is a simple matter, whether one uses two-dimensional quantitative structure−activity relationship (2D-QSAR), three-dimensional quantitative structure−activity relationship (3D-QSAR), or structure-based methods. But far too often, building a prospectively useful model is a hit-or-miss affair. Too frequently, models that seem to work well retrospectively fail miserably in prospective tests in their application to real-world drug discovery,[1] causing even leaders in the field to question whether researchers in modeling, such as ourselves, are playing a "glass bead game", divorced from reality.[2] At the other extreme, regulatory agencies are considering using QSAR models in their evaluation of drug candidates.[3,4] If the field of modeling is to fulfill its potential, we must find ways to build prospectively useful models more reliably.

In modeling a structure−activity relationship (SAR), are we being led astray by the statistical methods upon which we rely to protect us from fitting chance correlations? Are these statistical methods simply inappropriate for modeling an SAR? Bootstrapping, jackknifing, and cross-validation all have the intent to ensure that a high-quality model results; their utility in other fields is well-vindicated. However, in modeling an SAR, these methods fall short of their aim.[5] Furthermore, it has been observed by multiple authors that the models which perform best, retrospectively, are often the worst prospectively;[6,7] this observation has been labeled the "Kubinyi paradox".[8] When examining the models that have been published in the literature, many do not describe any prospective tests of the model, and they rely solely on these retrospective statistical measures; hence, it is difficult to judge their utility. A recent rigorous test of structure-based design approaches shows that even they rarely achieve good marks, even in retrospective tests,[9] although there are many examples where structure-based design indisputably has played a key role prospectively in advancing molecules toward drug-candidate status.[10]

We propose to take a new look at the issue of modeling an SAR, whether one approaches the problem with QSAR, pharmacophores, or structure-based design. We begin with the observation that models of an SAR usually aspire to predict the activity of a new molecule. However, in real-world drug discovery, it is not necessary that a model accurately predict the actual measured potency: a model that would consistently predict the correct ordering of any proposed pair of molecules would be tremendously valuable. For example, if the most potent molecule has a K_i value of 10 nM, and a chemist devises 10 ideas for new molecules to make, a model that accurately and consistently predicts which of those 10 ideas would represent a step forward in potency would be tremendously useful. Whether the model predicts 1 nM or 3 nM is less important; more important is the model's ability to determine reliably which of those 10 ideas would be more potent than 10 nM. In actual real-world modeling practice, however, it is not uncommon to have

* Author to whom correspondence should be addressed. E-mail: johnvandrie@mindspring.com.

10.1021/ci8001414 CCC: $40.75 © 2008 American Chemical Society
Published on Web 08/08/2008

models whose prospective performance is worse than random: "we'd have done better flipping a coin".

We propose to focus on a model's ability to predict the rank ordering of potency correctly. In the world of statistics, the ability of a model to rank-order the data correctly is called concordance. Kendall's τ statistic is one measure of concordance, and it is sometimes viewed as preferable to the familiar Pearson's correlation coefficient, r^2. While we are not using Kendall's τ statistic here, the ideas that we will present shortly are inspired by this notion of concordance. In some sense, this is "lowering the bar" for a model: we only expect that it rank-order the data properly; we do not concern ourselves with how closely it predicts any single activity number. Statistically, this decreases the number of degrees of freedom of the space of possible models, which should decrease the chance of discovering a model by relying on chance correlation.

In addition to focusing on a model's ability to rank-order the data, the inherent structure of the data in an SAR must be acknowledged. Recently, Maggiora[11] questioned whether an SAR can have a "landscape" that intrinsically makes it unsuited for typical QSAR modeling protocols. In other words, we know that many receptor-mediated SARs have sharp "activity cliffs", molecules which are very similar but have large changes in activity (e.g., changing a methyl to an ethyl kills activity). In Maggiora's metaphor, any modeling protocol will struggle if it assumes that the structure–activity landscape is relatively smooth, like the hills of Kansas, when it is, in fact, full of sharp cliffs like Bryce Canyon.

Stimulated by Maggiora's article,[11] we recently proposed a way to assess the smoothness of the landscape of a given data set, using the "structure–activity landscape index" (SALI).[12] We showed how this index can quickly highlight the activity cliffs of a dataset (the pairs of molecules most similar with the largest change in activity) and allows one to visualize an SAR in terms of a graph. Our intuition suggests that a model that is incapable of capturing the significant activity cliffs is a model that is destined for poor prospective utility. Motivated by this intuition, and the notion of intending to predict rank ordering rather than actual activity values, we have now extended the application of the SALI to formulate "SALI curves", which may be helpful in assessing the quality of a modeling protocol, as applied to a given dataset. These SALI curves provide a measure for how well a given model is capable of capturing the activity cliffs that are inherent in an SAR, and intriguingly, from our limited experience, we observe that models of demonstrated prospective utility seem to have a characteristic shape for their SALI curves.

This SALI curve approach for assessing modeling protocols is specifically targeted toward SARs where the biological activity is receptor-mediated (i.e., where one is looking at a recognition event). We do not anticipate that this would be applicable to chemometric settings, or for pharmacokinetic parameters, where the measured activity is dominated by physical properties and not molecular recognition events. In those cases, the structure–activity landscape is quite different.

1.1. Outline.

In this paper, we describe an approach to measuring the quality of a model in terms of its ability to predict the order of edges in a series of SALI graphs correctly. We first motivate the construction of these SALI curves, and then we apply them to a variety of 2D-QSAR models on published datasets, one published pharmacophore/3D-QSAR model, and one published structure-based design modeling protocol, and finally we show that SALI curves behave robustly under a variety of computational control experiments.

The paper is organized as follows. Section 2 describes the terminology used in the paper and the SALI curve algorithm. Section 3 describes the datasets used in this study. Section 4 describes the results of the computational experiments, and section 5 describes a set of computational controls for the model quality metric. We end with a discussion of the implications of these results for predictive modeling in general and future lines of investigation.

2. METHODOLOGY

Following our earlier work,[12] we define the structure–activity landscape index (SALI) as

$$\mathrm{SALI}_{ij} = \frac{|A_i - A_j|}{1 - \mathrm{sim}(i,j)} \qquad (1)$$

where A_i and A_j are the activities of the ith and jth molecules, and $\mathrm{sim}(i, j)$ is the similarity coefficient between the two molecules. This definition gives rise to a symmetric matrix, which is called the *SALI matrix* of that dataset; note that this matrix is dependent on both the dataset and the similarity measure used.

This index also allows one to represent an SAR as a fully connected graph, where the nodes are molecules, and the edges are weighted by the SALI_{ij} values. We call this the *SALI graph* of the dataset. These graphs can be stupendously complex, but they can be easily visualized by an interactive graphical tool and by limiting the display to only those pairs for which the SALI_{ij} value exceeds some threshold (i.e., one need only look at the sharpest cliffs). The set of activity cliffs exceeding some threshold X (expressed as a fraction of the maximum SALI value for the dataset) creates another graph, which is a subgraph of the SALI graph; we call this the SALI subgraph for X. An example is shown in Figure 1.

Given an SAR dataset, a model of that data, and a SALI graph, one can ask, for each edge in the SALI graph, whether the model correctly predicts the ordering of activity for the pair of molecules that comprises that edge. That is, if molecule 1 has activity $\mathrm{p}K_i = 1.0$ and molecule 2 has activity $\mathrm{p}K_i = 2.0$ (both in units of μM), then if the model predicts molecule 2 is more active than molecule 1, one says that the model has correctly predicted that edge of the SALI graph.

For any SALI subgraph with a threshold X, one can derive a quantity $S(X)$ that captures the model's ability to predict edges (similar in spirit to the ideas described by Pearlman et al.[13] for protein-based methods). Initializing a sum to 0.0 and looping over all edges in this subgraph, if the model correctly predicts that edge, the sum is increased incrementally by 1 and if it mispredicts that edge, the sum is decreased incrementally by 1. For cases where the model predicts the same activity for a pair of compounds, the sum is left unchanged. This sum, which is normalized by the number of edges in the subgraph, is denoted $S(X)$. A value of $S(X) = 1.0$ indicates a perfect prediction of all edges, whereas a value of $S(X) = -1.0$ indicates perfect misprediction; $S(X) = 0.0$ corresponds to random predictions.

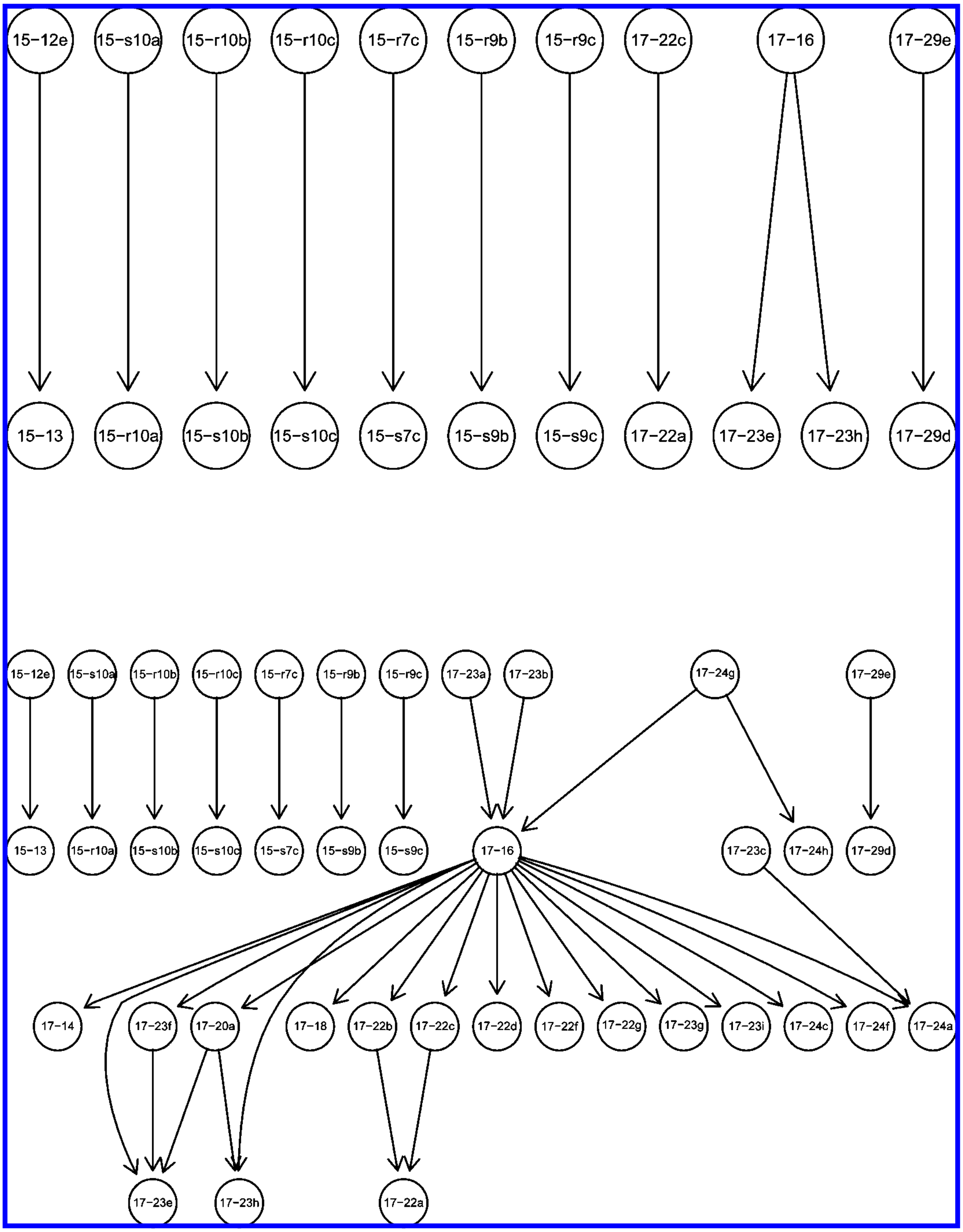

Figure 1. Graph representations of the structure−activity landscape index (SALI) values for a given dataset. Labels identify molecules. An edge occurs between two nodes if the SALI value for that pair is greater than a user-specified threshold. The upper graph was generated using a large threshold and highlights the most significant activity cliffs; the lower graph was generated with a smaller threshold value.

The case where the predicted values for a pair of molecules are the same is rare for regression problems. However, it will be common for classification problems (for example, a pharmacophore might classify a pair of molecules as both being active). The above formulation allows us to handle regression and classification problems in the same framework. At this point, we are not explicitly considering models that do not attempt to predict an edge, although, in principle, this definition of $S(X)$ accommodates such cases.

A SALI curve is simply a plot of $S(X)$ as X ranges from 0.0 to 1.0. The solid and dashed lines in Figure 2 show how one might expect the SALI curve for useful models to look; most of the significant cliffs are correctly predicted by the model, and most of the smaller cliffs also are correctly predicted by the model. It is unreasonable to expect a model to predict all the small cliffs correctly, because the activity differences in such cliffs are usually smaller than the experimental uncertainty (i.e., repeating the experimental measurement might flip the ordering of that edge). Big cliffs are robust to repeat measurements, and we only demand of a model that it correctly predict edges robust to experimental uncertainties. The red dotted line in Figure 2 shows how a model making random guesses would perform; this is repreented as a line oscillating around $S(X) = 0$.

Although the SALI curve in its entirety is informative, it is convenient to summarize the curve numerically. The integral of the curve, denoted as SCI (SALI curve integral), summarizes many features in one number. In practice, we

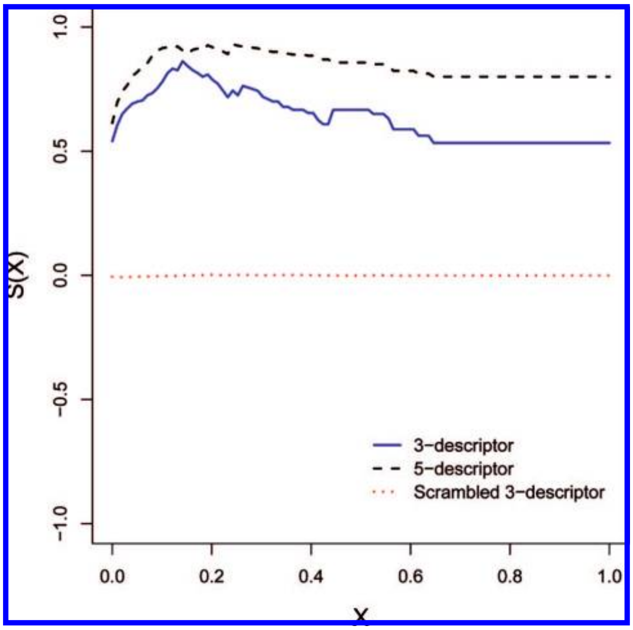

Figure 2. SALI curves for the three-descriptor and five-descriptor models built using the PDGFR dataset. The SALI curve for a Y-scrambled version of the three-descriptor model is also shown. *X* represents the SALI threshold, expressed as a fraction of the maximum SALI value in the dataset.

use a summation of the individual $S(X)$ values, suitably normalized. Specifically, we have,

$$\mathrm{SCI} = \frac{1}{N}\sum_{i=0}^{N} S(X_i) \qquad (2)$$

where X_i represents a fraction of the maximum SALI value for the dataset. We consider fractions ranging from 0.0 to 1.0, such that $X_0 = 0.0$ and $X_N = 1.0$. In this study, we choose an increment of 0.01, so that $N = 100$. Although evaluating $S(X)$ at smaller intervals results in a higher resolution SALI curve and, subsequently, a more accurate value of SCI, values of N greater than 100 did not lead to any appreciable difference in the calculated SCI values. Given the aforementioned definition, the SCI can range from 1.0 (perfect prediction) to −1.0 (perfect misprediction), with 0.0 indicating random predictions. For the SALI curves in Figure 2, the curve for the five-descriptor model has an SCI value of 0.83, the three-descriptor model has an SCI value of 0.63, and the curve for the scrambled three-descriptor model has a SCI value of −0.0008.

Note that a SALI curve is determined by the dataset, the model, and the similarity measure used to calculate the SALI. At the same time, it is important to note that the algorithm does not require the explicit use of a model. All it requires is the predicted values from the model. Furthermore, because the algorithm focuses on ordering, it is possible to replace the predicted values with the ranks of the compounds and obtain the same results.

All the computational experiments were performed using the R 2.6.0 program,[14] running on a MacbookPro (2.16 GHz, 1GB RAM).

3. DATASETS

Given the generality of the SALI curve method, we considered both ligand-based QSAR models as well as structure-based models. For the case of QSAR models, we focused on four datasets taken from the literature, for which we developed linear regression models. For structure-based models, we considered a dataset that was studied using CoMFA and a dataset that was studied using a docking procedure. Details of the datasets are provided below.

3.1. 2D-QSAR Datasets. The first data set consisted of 79 derivatives of 4-piperazinylquinazolines studied for their ability to inhibit PDGFR.[15] A previous study[16] developed both linear regression and neural network models to predict the IC_{50} values of the inhibitors. In this study, we used the original descriptor pool, which consisted of 321 descriptors that were calculated using ADAPT.[17] This pool was reduced using correlation and variance tests to a reduced pool of 43 descriptors.

The second dataset was a collection of 81 molecules that were studied as agonists[18] and antagonists[19] of the human melanocortin-4 receptor. The original works reported the activity of these compounds in terms of their K_i values. To generated the SALI graphs, we used the reported values directly. For modeling purposes, we used the negative logarithm, because of the order-of-magnitude differences in the reported K_i values. For this dataset, we evaluated 102 topological and constitutional descriptors, using the CDK,[20] which were then reduced to 29 descriptors, using correlation and variance tests.

The third dataset was obtained from the study reported by Sutherland et al.[21] The study focused on ligands for a variety of receptors. We considered the 405 ligands for the benzodiazepine receptor. The activities of the ligands were reported as IC_{50} values ranging from 0.34 nM to greater than 70 μM. Many compounds had indeterminate IC_{50} values and were excluded. Several compounds were also excluded because of errors in the fingerprint calculation, resulting in a final data set of 301 molecules. As with the melanocortin dataset, we used the reported IC_{50} values to generate the SALI graphs and the negative logarithm of the activities to generate the predictive models. The structures obtained from the study had three-dimensional (3D) coordinates that allowed us to evaluate 170 topological, constitutional, and geometric descriptors using the CDK. These were then reduced to a set of 54 descriptors, using correlation and variance testing.

The final dataset was a collection of 179 artemisinin analogs[22] that were studied for their antimalarial activity. Previous work[23] developed linear regression and neural network models, although the linear regression model was developed for interpretative, rather than predictive, purposes. For this dataset, we used the original ADAPT generated descriptor pool (299 descriptors), which was reduced to 55 descriptors, using correlation and variance testing. Because the reported activity was already logarithmic in nature, both SALI graphs and predictive models used the same dependent variable.

For all the datasets, we used the reduced pool in the experiments that has been described in this work. To generate the SALI graphs and curves, we used the BCI 1052-bit fingerprints,[24] and similarities were calculated using the Tanimoto coefficient.

3.2. 3D Datasets. The first dataset was a set of HIV-1 protease inhibitors that was described by Holloway et al.[25] The original work performed energy minimizations of the inhibitors within the protein active site and reported signifi-

Table 1. Summary Statistics for the Three- and Five-Descriptor Models Built for the PDGFR Dataset

descriptor[a]	root-mean-square error, RMSE	r^2	F	p-value
MDEN-33, FNHS-1, RNH	0.49	0.48	23.31	9.18×10^{-11}
MDEC-12, FNHS-1, 1SP3, RNH, MDEN-33	0.45	0.56	18.92	5.08×10^{-12}

[a] Legend of terms: MDEN-33 = molecular distance edge between tertiary nitrogens;[34] MDEC-12 = molecular distance edge between primary and secondary carbons;[34] 1SP3 = count of single bound carbon bound to one other carbon; FNHS-1 = fraction of the hydrophilic surface area;[35] and RNH = relative hydrophilicity.[35]

cant correlations between the intermolecular interaction energy (E_{inter}) and the in vitro enzyme inhibition of the inhibitors. For this dataset (and structure-based design in general), we did not utilize the predicted IC_{50} values, but, instead, used the calculated E_{inter} values to rank the compounds, and, subsequently, we used the rankings to generate SALI curves.

The second dataset consisted of a set of compounds that were active against the hERG channel and was studied by Cavalli et al.,[26] who aligned the molecules using a pharmacophore and then used the pharmacophore-based alignment to develop a CoMFA[27] model. For this dataset, we used the predicted IC_{50} values that were reported by the authors to derive the SALI curve.

4. RESULTS

4.1. SALI Curves. Generally, it makes sense to compare SALI curves generated from different models for a given dataset, rather than multiple datasets. SALI curves generated for two QSAR models for the PDGFR dataset are shown in Figure 2, which plots $S(X)$ versus X (the SALI threshold, which is expressed as a fraction of the maximum SALI value for the dataset) for a three-descriptor regression model and a five-descriptor linear regression model. The model statistics are summarized in Table 1. Although there is no significant difference in the statistical significance of the models, the root-mean-square error (RMSE) and r^2 values indicate that the five-descriptor model exhibits better predictive accuracy.

When considering the SALI curves, we note three important features. First, the value of $S(0.0)$ represents the ability of the model to capture all the SARs in the dataset. Second, the value of $S(1.0)$ represents the models ability to capture the most significant cliffs in the dataset. Because of the fact that the SALI subgraph at this threshold will consist of a small subset of the dataset, the $S(1.0)$ value does not directly correspond to the RMSE of the model. The third important feature of the graph is the variation of the curve between these two values.

For two thresholds—X and X', and $X < X'$—a SALI graph that is generated at a threshold X is always a subgraph of one generated at a threshold X'. As a result, it is possible that the fraction of edges correctly predicted at X' is greater than that at X, simply because of the size of the graphs involved. As a result, we do not believe that the explicit variation in the graph is significant.

Given these observations, we see that the curve for the three-descriptor graph has $S(0.0) = 0.77$ and that for the five-descriptor model has $S(0.0) = 0.87$. These values indicate that the five-descriptor model is able to capture a larger proportion of cliffs overall, as well as a larger portion of the most significant cliffs. These results can be succinctly restated in terms of their SCI values of 0.83 for the five-descriptor model and 0.63 for the three-descriptor model, which clearly indicates the overall, better encoding of the structure−activity landscape by the five-descriptor model. However, this should not be surprising, because by sufficiently increasing the size of the model, one can fit the data perfectly. This suggests that one should not compare SALI curves that have been derived from models of different sizes.

To obtain a more meaningful comparison, we developed linear regression models for all four datasets, using an exhaustive procedure. We developed three-, three-, six- and five-descriptor models for the PDGFR, artemisinin, melanocortin, and benzodiazepine datasets, respectively, by evaluating all possible subsets from their respective descriptor pools. For each dataset, we identified the models with the minimum, maximum, and median RMSE values. The statistics of the models are summarized in Table 2, and Figure 3 displays the SALI curves for the three models for each dataset.

The ordering of the models, for a given dataset, in terms of quality, is mirrored by the $S(0.0)$ values for each SALI curve. Thus, the best model has the highest $S(0.0)$ value and the worst model has the lowest value. In other words, the $S(0.0)$ value tracks the behavior of the RMSE of the models. Although it makes sense intuitively, it is slightly surprising, because the $S(0.0)$ value is derived from a pairwise approach, whereas the RMSE considers the entire dataset. We note that the $S(0.0)$ calculation is similar in nature to the definition of concordance, as measured by Kendall's τ statistic. Indeed, simulation using random data suggests that, within certain limits, Kendall's τ statistic does correlate well with the RMSE for a given dataset.

Next, we focus on the $S(1.0)$ values for the various SALI curves. For the PDGFR dataset, we observe that the ordering of the $S(1.0)$ values mirrors that of the model quality. On the other hand, the artemisinin dataset indicates that the model with the lowest RMSE has the poorest ability to capture the significant cliffs (those beyond $X = 0.5$). Rather, the model with the median RMSE value exhibits the best ability to capture significant activity cliffs. However, an inspection of Table 2 indicates that the median-RMSE and maximum-RMSE models explain little to none of the variance in the dataset and also do not have significant F-values. In the case of the melanocortin dataset, we see that the median-RMSE model seem to correctly predict a larger fraction of the most significant cliffs, although the difference in the $S(1.0)$ values is not very large. Interestingly, we see that the $S(1.0)$ values for the minimum-RMSE and maximum-RMSE models are identical. Given the model statistics in Table 2, it is likely that the $S(1.0)$ of the maximum-RMSE model is simply due to luck. Finally, for the benzodiazepine dataset, we see much of the same behavior. Note that the minimum-RMSE model does not exhibit very good predictive ability on this dataset (although the model itself is statistically significant). However, there is a clear distinction in the ability of the three models to capture significant activity cliffs.

Figure 4 provides a summary of the SCI values for the SALI curves shown in Figure 3. It is evident that the models

Table 2. Summary Statistics for the Models Built for the Four Datasets in This Study[a]

descriptors[b]	RMSE	r^2	F	p-value
PDGFR Dataset				
MDEN-33, FNHS-1, RNH	0.49	0.48	23.31	9.18×10^{-11}
RPH, NAB, 3SP3	0.63	0.14	4.09	9.53×10^{-4}
V5CH, MOMI, SULF	0.67	0.00	0.02	0.99
Artemisinin Dataset				
N7CH, NSB, WTPT-2	0.92	0.66	112.1	2.20×10^{-16}
ALLP-5, GEOM-2, APMIN	1.52	0.05	3.37	1.98×10^{-2}
EDIF-1, FPSA-3, WNHS-2	1.56	0.00	0.002	0.99
Melanocortin Dataset				
AMR, TPSA, ATSc5, SPC-6, VP-6, C2SP3	0.57	0.63	21.1	3.40×10^{-14}
ALogP2, AMR, WTPT-2, WTPT-4, WTPT-5, SC-3	0.69	0.46	10.44	2.46×10^{-8}
ATCs3, Kier3, SC-5, C2SP3, VC-4, VCH-4	0.93	0.04	0.52	0.79
Benzodiazepine Dataset				
Weta2.unity, BCUTc.1, FNSA-3, WTPT-5, ATSc5	0.89	0.40	39.86	2.20×10^{-16}
Wlambda2.unity, Wnu1.unity, Weta2.unity, Weta3.unity	1.06	0.16	11.47	3.94×10^{-10}
MDEC.12, MDEC.22, MDEN.12, MDEN.13, C2SP2	1.15	0.00	0.05	0.99

[a] For each dataset, we exhaustively evaluated all possible models and reported the statistics of the best, median, and worst (in terms of RMSE) model. [b] Legend of terms: MDEN-33 = molecular distance edge between tertiary nitrogens;[34] 3SP3 = count of triply bound carbon bound to three other carbons; FNHS-1 = fraction of the hydrophilic surface area;[35] RNH = relative hydrophilicity;[35] RPH = relative hydrophobicity;[35] NAB = number of aromatic bonds; MOMI = principal component of moment of inertia along the X-axis; V5CH = fifth-order valence-chain χ index;[36] and SULF = number of sulfurs.

built on the artemisinin dataset perform very poorly, in terms of encoding the SAR landscape, and the model with the expected best performance has the lowest SCI value. In the case of the melanocortin dataset, we see that there is no significant difference between the ability of the models to encode the SAR landscape, whereas for the PDGFR and benzodiazepine datasets, we see that there is differentiation between the performance of the models and their ability to encode the SAR landscape. However, it is clear that none of them are able to encode the landscape of the respective datasets to a significant degree.

Given the aforementioned discussion, it is interesting to ask why the $S(1.0)$ value of the minimum-RMSE model for the artemisinin model is so low. There are 11 edges in the SALI graph that correspond to a threshold of $X = 1$, and the model is able to predict the ordering of one edge. As shown in Table 3, the Tanimoto coefficient for the compound pairs is generally 1.0. Indeed, many of the compounds in the graph are stereoisomers. Because we used real-valued descriptors, rather than fingerprints, to build the models, it is more reasonable to investigate the similarity in the descriptor space of the model. We evaluated the Euclidean distance between the compound pairs: of 11 pairs, 6 were equal to 0, and the highest value was 3.6 (the maximum distance in the entire dataset was 39.1). It is clear that, even in the descriptor space, the bulk of the pairs are identical, which is a side effect of the fact that the descriptors are not be able to capture the small structural variations. As expected, for those cases, where the molecules are identical in descriptor space, the predicted values are identical and, hence, the model does not order them correctly. It should be stressed that a correct ordering of predictions does not imply predictive accuracy. For example, although compounds 148 and 147 in Table 3 have the correct predicted ordering, compound 147 has a low standardized residual (0.06 log units), whereas compound 148 has a large standardized residual (−0.81 log units). We also note that artemisinin dataset contained 23 molecules for which the reported activity was −4.0 log units, although there were structural differences within the group. No explanation was provided for these compounds, and it is likely that this group of compounds is the cause for the poor performance of the artemisinin models.

4.2. SALI Curves and 3D Modeling Approaches. We now shift our attention away from the statistical models that we have built using molecular descriptors to investigate the behavior of SALI curves for 3D modeling techniques, namely, structure-based design and 3D-QSAR.

To examine a structure-based design approach, we consider the Holloway dataset that has been described in section 3. This dataset is a good one to choose for our purposes, because it was used prospectively in the discovery of Merck's first clinical candidate targeting HIV-1 protease against AIDS. This should provide an example of how a SALI curve might appear for a prospectively useful model.

As noted previously, rather than using the predicted IC_{50} values, we used the rankings of the compounds, derived from calculated E_{inter} values. Thus, if the docking model rank-orders the compounds in a cliff in the same order as their observed IC_{50} values, we consider that cliff to be correctly predicted and evaluate the SALI curve as described previously. Figure 5 shows the SALI curves generated from the reported E_{inter} values.

In contrast to the SALI curves presented for 2D QSAR models, we see that the structure-based design model is able to capture the *a* range of cliffs, ranging from intermediate to the most significant. This may highlight what distinguishes a structure-based approach, or at least a 3D approach, from QSAR models that have been constructed from 2D descriptors. It is interesting to note also that the r^2 value for the Holloway et al. model is mediocre (0.7), yet from the SALI curve perspective, the model seems to be outstanding.

Curious as to whether this type of SALI curve reflected more the nature of structure-based design or the inherent properties of a 3D approach, we next considered a pharmacophore/3D-QSAR model. For this, we chose the dataset that was studied by Cavalli et al.[26] This model has been heavily validated in the literature and may

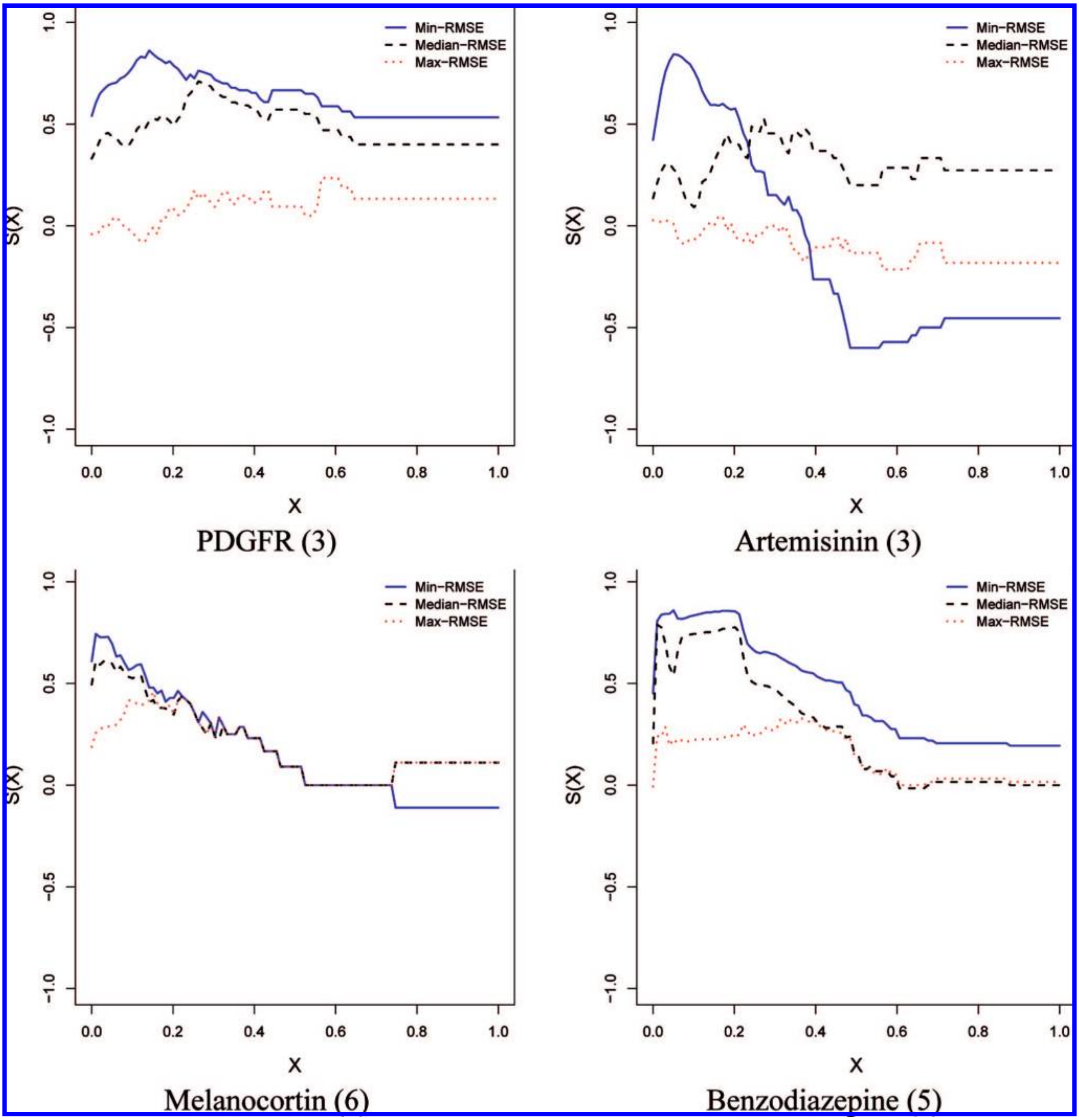

Figure 3. SALI curves for the best, median, and worst OLS models for each dataset. The models were identified by exhaustive evaluation of the descriptor pool, and the numbers given in parentheses indicate the size of the model. *X* represents the SALI threshold expressed as a fraction of the maximum SALI value in the dataset.

represent the gold standard for 3D hERG models that do not rely on a homology model of the hERG channel. Thus, similar to the Holloway dataset, we anticipate that the SALI curve should be of a form indicative of prospective utility. The curve is shown in Figure 6, and the SCI value of 0.99 indicates that the model captures essentially all the details of the structure−activity landscape. We interpret this as a measure of the quality of the entire protocol followed by Cavalli et al.:[26] the selection of compounds for consideration, the pharmacophore used to align the analogs, and the conformational analysis used, as well as a reflection of the inherent nature of CoMFA.

5. COMPUTATIONAL CONTROLS

The preceding discussion highlights the utility of SALI curves to measure the quality of a model in terms of its ability to encode the SAR landscape. In the following sections, we perform a series of computational controls to investigate the behavior of the SALI curves in different scenarios, primarily focusing on the effects of noise.

5.1. Y Randomization. We first consider the traditional model validation technique of Y-scrambling. In this procedure, we scramble the variable and rebuild the model. We then generate a SALI curve. Now, the result of Y-scrambling is such that, any correlation between the X-variables and

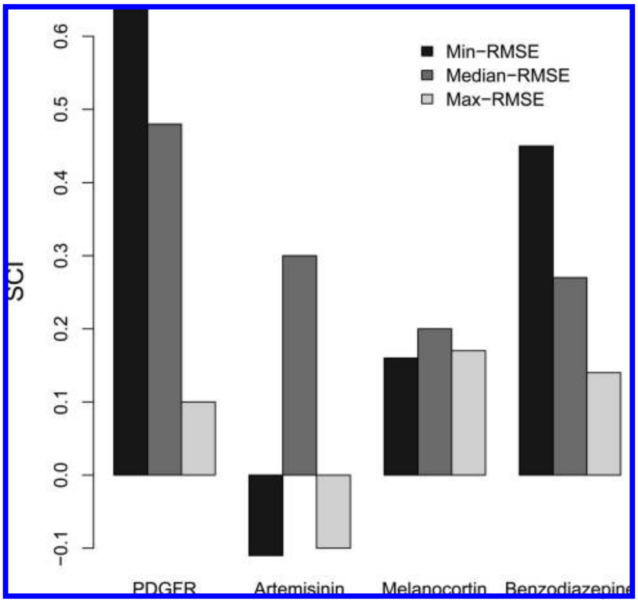

Figure 4. Summary of the SCI values for the SALI curves derived from the three models developed for each QSAR dataset.

Table 3. Summary of the Compounds Comprising the Edges in the SALI Graph for the Artemisinin Dataset Generated Using a Threshold of 100%

ID1	ID2	similarity	Euclidean distance	Δ Obs	Δ Pred	predicted ordering
4	1	1.00	0.00	−0.17	0.00	incorrect
25	8	0.99	3.00	1.65	−0.70	incorrect
31	11	1.00	0.00	−2.34	0.00	incorrect
33	12	1.00	1.00	0.16	−0.21	incorrect
173	54	1.00	0.00	−1.54	0.00	incorrect
65	56	1.00	0.00	−0.02	0.00	incorrect
170	58	1.00	3.61	−1.48	0.71	incorrect
100	99	1.00	1.41	−1.12	0.16	incorrect
138	137	1.00	0.00	0.35	0.00	incorrect
148	147	1.00	1.00	−1.08	−0.21	correct
175	168	1.00	0.00	0.70	0.00	incorrect

the Y-variable is lost. As a result, a scrambled model is expected to predict 50% of the edges correctly, on average, for any SALI subgraph, yielding a SALI curve oscillating around 0.0. We applied this procedure to the three-descriptor model developed for the PDGFR dataset and performed 100 scrambling runs, resulting in 100 SALI curves. We then determined the average of these curves, and the result is shown in Figure 7. In this figure, the red dashed line indicates the theoretical curve, and the black line is the mean curve evaluated from the 100 scrambling runs.

5.2. Addition of Noise. Noise can be added to the X- or the Y-variables, and, in both cases, we expect that the correlation between the dependent and independent variables will be reduced. Figure 7 presents a set of SALI curves derived from the three-descriptor PDGFR model; the black line represents the curve for the original model. The remaining four curves in this figure are derived from models built using increasingly noisy X-data. It is apparent from the plot that, for small amounts of noise, the value of $S(1.0)$ does not differ significantly from that of the original model. We also see the relatively low variation in $S(0.0)$, compared to $S(1.0)$. These observations highlight the fact that the addition of noise to the descriptors can convert pairs of compounds that represent significant activity cliffs to small activity cliffs.

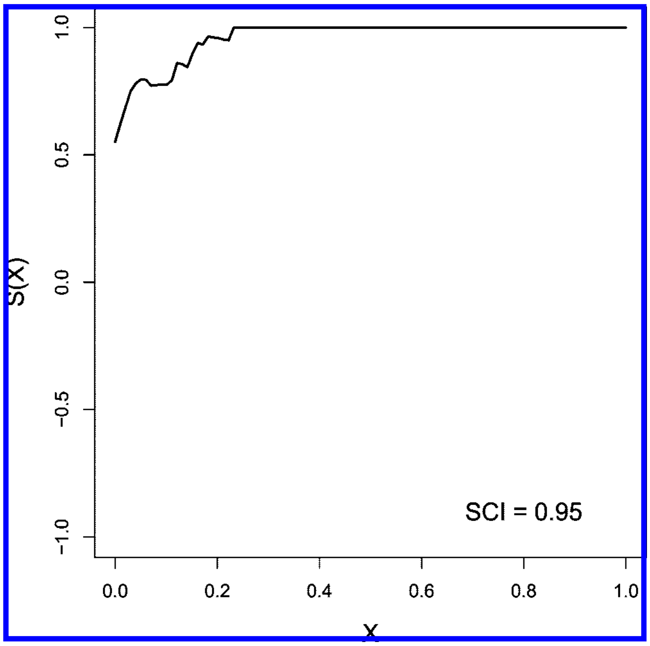

Figure 5. SALI curve for the Holloway dataset, highlighting the ability of a docking model to capture the significant cliffs accurately. Similarity was measured using the Tanimoto coefficient with BCI fingerprints, and predicted activity was replaced with the ranks of the E_{inter} values. X represents the SALI threshold, expressed as a fraction of the maximum SALI value in the dataset.

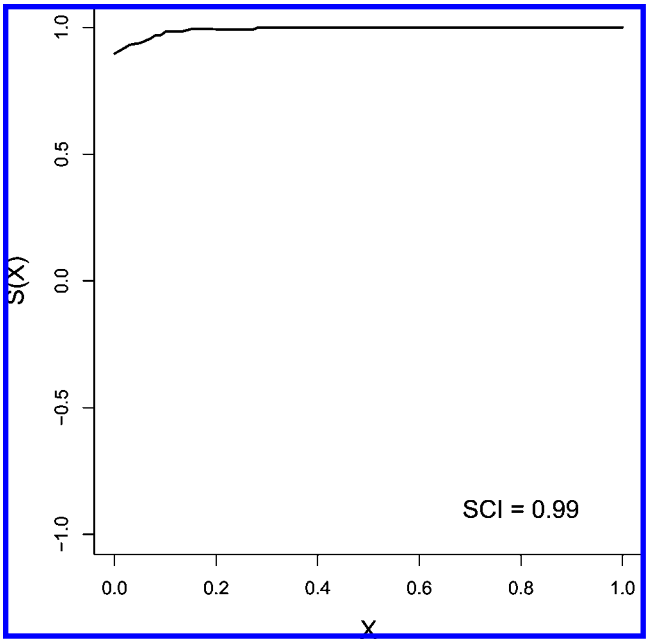

Figure 6. SALI curve for the Cavalli et al.[26] dataset, highlighting the ability of a CoMFA model to capture the significant activity cliffs. Similarity was measured using the Tanimoto coefficient with BCI fingerprints. Predicted activities (IC_{50}) were used to generate the SALI curve. X represents the SALI threshold, expressed as a fraction of the maximum SALI value in the dataset.

We next considered the effect of noise in the dependent variable. This is of practical concern because assay data can be noisy. Figure 8 shows SALI curves for the various datasets, using noisy Y-variables. For a given dataset, we generated a model using the original Y-variable (black curve). We then added increasing amounts of noise (from a

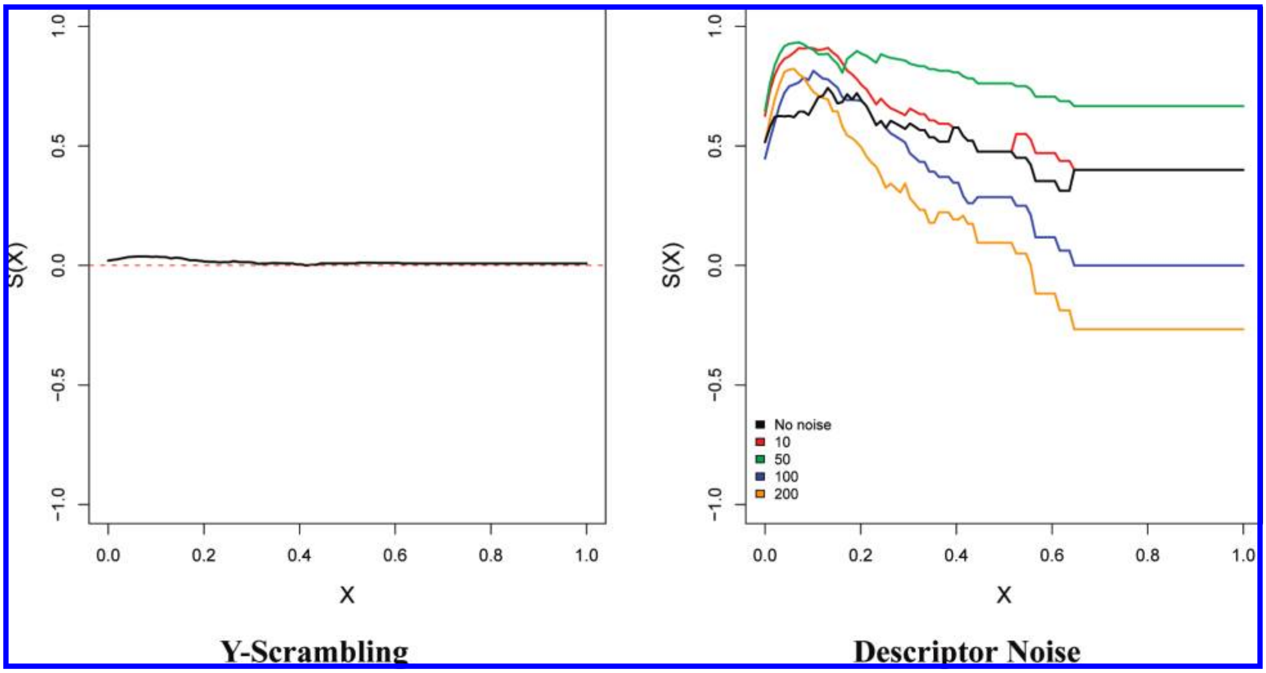

Figure 7. Results of the control experiments performed for the SALI curve. *X* represents the SALI threshold, expressed as a fraction of the maximum SALI value in the dataset. In both cases, we used the three-descriptor model built for the PDGFR dataset. For the scrambling plot, the red line is drawn at the *S*(*X*) = 0 level, and the back line is the mean SALI curve averaged over 100 scrambling runs. For the descriptor noise plot, the numbers indicate the relative amounts of noise added to the descriptor data.

normal distribution). The legend represents noise factors: larger values indicate larger amounts of noise.

The results are qualitatively similar to those in Figure 7, in that more variation is observed at higher thresholds rather than lower thresholds. In both cases, one might expect that the addition of noise would affect the prediction of smaller cliffs rather than larger cliffs. This would lead to greater variation on the left-hand side of the curves, rather than on the right-hand side of the curves. However, our results indicate the opposite. Although this behavior is counterintuitive, it is a side effect of the nature of the scoring function that we use to calculate the value of the SALI curve at any given threshold value. Briefly, the change in the score (due to addition of noise) at any given threshold is inversely proportional to the number of edges in the graph at that threshold. As a result, the change is much less significant at low thresholds, because the graphs have significantly larger numbers of edges than at higher thresholds. The Appendix provides a formal mathematical explanation of this behavior. We also note that the larger variation in the curves in Figure 7, compared to those in Figure 8, is due to the fact that each descriptor had noise added to it. Therefore, cumulatively, much more noise has been added to the X-variables, compared to the case where we add noise to the Y-variable.

However, the counter-intuitive behavior exhibited in Figure 8 can also be ascribed to the fact that the models themselves are probably modeling noise (i.e., experimental error) in the original observed activities. This is probably especially true for the artemisinin dataset, where we see that the SALI curves in the presence of Y-noise have a significantly higher value than the curve from the original model. This can be explained by the fact that there are 23 molecules with the same value in the original dataset. As a result, the addition of noise causes them to differentiate and, as a result, the predicted values may correctly order them. In the original model, the ordering of the predicted values for these molecules can never match the observed ordering (because there is none).

Finally, we also note that, for the QSAR datasets considered here, none of them contained cliffs whose magnitude was due to large differences in activity. Instead, cliffs usually arose because of a high degree of similarity between pairs of molecules. As a result, the activity differences were small enough that sufficient amounts of noise (of the order considered here) were sufficient to “flip” the ordering of many of the significant cliffs. However, this effect is probably specific to datasets with relatively closely spaced activity values.

6. DISCUSSION

The SALI method allows us to characterize the landscape of an SAR, and, in this study, we have shown that it can be used to characterize a model’s ability (or inability) to capture the details of the landscape. Although a numerical characterization of the model, the SALI curve technique does not address numerical accuracy. Instead, it focuses on a model’s ability to capture increasingly significant cliffs. This explicit measure of a model’s ability to capture cliffs provides an alternative view to characterizing predictive models.

More generally, the method provides us with a *framework* within which we can understand what types of models better encode a SAR and what aspects of the SAR they encode successfully (and what aspects they encode unsuccessfully). Note that the SALI framework does not simply focus on the model itself. Instead, it characterizes the modeling protocol (in other words, the nature of descriptors used to build a model as well as the model itself). In some scenarios (such as 2D QSAR) it is possible to access the molecular representation (i.e., descriptors) explicitly. In other cases, such as docking models, the chemical representation is more abstract and we may not have direct access to it. The

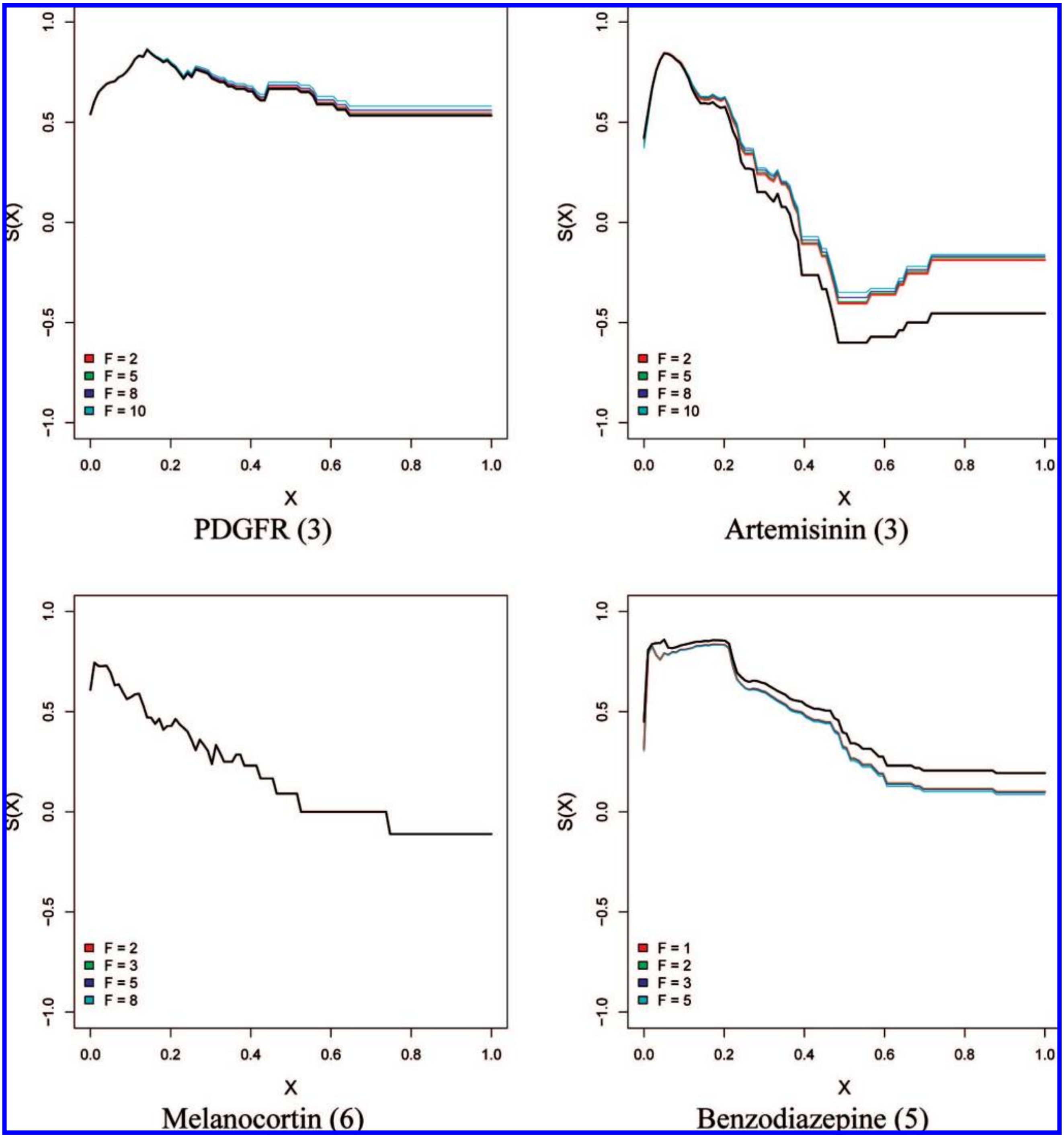

Figure 8. SALI curves for models built using a noise-dependent variable. Multiple curves in a plot correspond to models built using increasing amounts of Gaussian noise. Here, *F* is a factor that indicates the degree of noise that was added to the dependent variable. The black line is the SALI curve for the original model.

advantage of the SALI framework is that we are not dependent on direct access to the underlying molecular representation.

It is important to realize that, from the point of view of machine learning, activity cliffs, especially very large cliffs, represent special cases. By definition, the small structural difference between a pair of molecules, comprising an activity cliff, is specific to the pair. A predictive modeling algorithm faces challenges when trying to generalize the structural differences exhibited by activity cliffs. From the perspective of traditional QSAR, a model that predicts a large number of activity cliffs correctly is usually interpreted to say that the model has memorized the minor structural differences, that is, a model that correctly predicts a large fraction of activity cliffs will probably perform poorly on the bulk of the dataset. When applied to traditional QSAR, one should be cautious about interpreting the significance of a model with a high value of $S(1,0)$, especially if $S(0.0)$ is relatively low.

One result of this approach is that it allows us to investigate certain fundamental problems underlying QSAR modeling that may not be accessible when one considers pure numerical accuracy. For example, the Kubinyi paradox[8] observes that a model with good retrospective performance does not necessarily exhibit good prospective performance. From the point of view of machine learning, this may be due to overfitting. However, it has been shown[5] that characterizing overfitting using metrics such as q^2 are not very reliable. When viewed from the point of view of activity cliffs, one could examine what parts of the dataset may be problematic for the model. When considered prospectively, one could then ask: where, on the structure−activity landscape, is a new molecule located? If it is in the region of a cliff, that model may provide a poor (i.e., unreliable) prediction. It is apparent that the SALI framework could also be used as the basis to understanding issues underlying model applicability[28] (when a model will provide a reliable prediction and when it will not). Traditional approaches[29−32] to this problem have generally considered aspects of the descriptor space of the model. Although a valid view, it is clear that a model encodes an SAR by taking into account both the structural representation of molecules as well as the activities of molecules. Thus, it would appear that, by considering the structure−activity landscape (rather than just the "structure landscape"), we may gain a better understanding of when a model may be used to reliably predict the property for a new molecule.

The aforementioned discussion leads to the question of whether experimental error present in the dataset or the occurrence of overfitting in the model will be visible in the SALI curve itself. The effect of experimental error is to mask the correlation between the structure and observed activity. Thus, the predictions from the model should be increasingly "scrambled" with larger amounts of experimental error. This implies that, overall, the SALI curve (for a model derived from a dataset with significant experimental error) should correspond to fewer correct edge orderings. In other words, low SCI values (near 0.0) could be indicative of experimental error (although a poor choice of descriptors could also lead to a low SCI value). The issue of overfitting is related to the memorization of cliffs. When we speak of memorization, we refer to the model learning some specific cases. On the other hand, overfitting is a more general problem, in which the model effectively memorizes the entire dataset. The result of overfitting implies that the predicted orderings will be perfect. That is, the SCI will be 1.0. From this point of view, one could suggest that the CoMFA model analyzed in section 4.2 is somewhat overfit. We believe that this is not the case, because the model has been shown to be prospectively useful. Generally, for QSAR models, excessively high SCI values could be indicative of overfitting and, in such cases, additional confirmatory tests should be performed on the model.

One feature of the preceding material is that we have characterized the ability of a model built using a certain molecular representation to capture the cliffs identified in a different representation. Thus, we could imagine that the use of a different representation to compute the molecular similarity in generating the SALI graph could yield different results and alter our SALI curve assessment of the modeling protocol. Ideally, one would identify the cliffs in a given representation and then measure the ability of a model built using that representation to capture those cliffs. However, this approach does not allow one to compare models built using different representations. In a sense, the use of a binary fingerprint to identify cliffs in the dataset represents a global characterization of the activity cliffs in a dataset and is independent of any specific descriptor set used to build models. We do not claim that the BCI fingerprints represent the best global description of a dataset, and it might indeed be useful to identify cliffs using different representations and investigate a model's ability to capture the sets of cliffs so identified.

This observation leads to the possibility of characterizing molecule representations themselves, in terms of their ability to generate a smooth or jagged landscape. One might expect that a representation that is relatively flat (i.e., few major cliffs) would lead to better machine learning models (because there are fewer significant nonlinearities). At the same time, it is true that the descriptors that define such a landscape should have some relation to the property being modeled. In other words, a flat landscape by itself is not a guarantee of well-performing models. Another advantage of this approach is that we can characterize different representations, independent of any models built on them. This is contrary to traditional descriptor selection procedures, which invariably are closely linked to a specific type of model. We will present a more detailed analysis of this approach in a forthcoming paper.

Our use of computational controls to study the effects of scrambling and addition of noise reinforces our intuitive sense for the significance of the SALI curves. Increasing noise has a tendency to drive the curves downward, with the portions of the curve near $X = 1.0$ being more sensitive to noise than those portions near $X = 0.0$.

We also note that, although the results in this paper have focused on linear regression models, we also investigated SALI curves for nonlinear models. Given that activity cliffs represent nonlinearities in the dataset, one might expect that a nonlinear model (for example, a neural network or random forest) might capture more of the significant activity cliffs. This is not always true; it is a function of the modeling method and the molecular representation. Therefore, we developed a computational neural network (CNN) model for the PDGFR dataset using the descriptors selected for the best linear regression model and two hidden neurons. The SALI curve for the CNN model indicated that it performed slightly better on the bulk of the dataset (also confirmed by a slightly lower RMSE than the linear regression model), but captured the same number of significant cliffs as the linear model. Although the descriptor set is not optimal for the CNN model, a nonlinear method clearly does not guarantee better encoding of the SAR. Indeed, this highlights the fact that one may avoid unnecessary increases in model complexity by identifying the degree to which significant cliffs can be identified by a model. Furthermore, our results stress the fact that a given molecular representation may not *allow* improvement in a model's ability to capture significant cliffs, irrespective of the complexity of the model.

7. CONCLUSIONS

Traditional approaches to modeling structure−activity relationships (SARs) have focused on the accuracy of predictions, either in terms of root-mean-square error (RMSE) or rank correlations. There is no doubt that accuracy is important; however, this approach does not usually explain whether a model will perform well prospectively, and if it does not, why it does not. Furthermore, by restricting ourselves to pure numerical measures of accuracy or predictive ability such as r^2 or q^2, we have a tendency to ignore the details of what a model has captured, in terms of an SAR. In other words, how do we understand why a model performs poorly? Is it due to the modeling technique? Is it due to the underlying molecular representation? What are the structural aspects of the landscape that might prevent a model from performing well?

In summary, we have presented an approach to the characterization of arbitrary predictive models, in terms of their ability to capture increasingly significant activity cliffs. The method first uses a two-dimensional (2D) fingerprint (although other representations can be used) to identify activity cliffs of varying degrees. Because each cliff can be considered an ordered edge in the structure–activity landscape index (SALI) graph, we examine how many such edges in a given SALI graph are correctly ordered by the model. This allows us to characterize the model in terms of two values that indicate the model's ability to capture the minor cliffs (constituting the bulk of the dataset) and the most significant cliffs (which represent the "interesting" parts of the SAR landscape). The method was applied to linear regression models that were built on a variety of datasets, as well as a docking model and a CoMFA model. One of the useful features of this approach is that we do not require access to the original models. All that is required are the predicted values (or some form of rankings in the case of docking) and the molecular structures.

We highlight the ability of this approach to differentiate between models of very similar statistical quality, allowing us to focus on models that are able to capture significant parts of the SAR. In this sense, the SALI curve method is not an alternative to traditional numerical characterizations but, instead, provides a complementary approach to measuring model quality. Furthermore, the SALI approach allows us to consider all aspects of the modeling protocol. In this paper, we have focused on characterization of the model itself. As noted previously, future work will focus on the use of the SALI framework to analyze the suitability of molecular representations, independent of the modeling technique applied to perform predictions.

Although we have not analyzed this point in detail, we note that this SALI curve approach is appropriate for receptor-mediated SARs. Many quantitative structure–activity relationship (QSAR) models that have been reported recently in the literature involve studies of pharmacokinetic data, which have a tendency to be dominated by physicochemical properties, with a smaller component of molecular recognition involved. It is an interesting open question whether standard QSAR techniques are more appropriate to pharmacokinetic models rather than models of SARs, possibly because of these aspects of the structure–activity landscape.

Our observations with the SALI curves applied to a variety of models gives insight into some well-known paradoxical behaviors (e.g., the Kubinyi paradox), and the observation that the r^2 value for most structure-based design models rarely exceeds 0.7. We now understand that optimizing solely on the RMSE allows one to obtain models with relatively high $S(0.0)$; however, this says nothing about the SALI curve at larger values. It appears that the plateau values of the SALI curves may have a tendency to say more about how well the model may perform prospectively (i.e., the value of $S(X)$, when $X = 0.5-1.0$). In other words, we hypothesize that a model's ability to capture the largest cliffs may be most indicative of its ability to perform well prospectively.

The magnitude of the SALI curve integral (SCI) seems to be interesting. In the models that have been described in this paper, the SCI values range from below 0.0 for the artemisinin models to greater than 0.90 for the Cavalli hERG model and the Holloway structure-based HIV model, which both are models known to be prospectively useful. More experience is needed to fully appreciate the implications of these values, but we are aware that the SCI is "richer" than the RMSE, because the RMSE corresponds to only one value of the SALI curve: $S(0.0)$. We anticipate that the greatest value of these SALI curves, and the SCI value, will be to assist modelers in "debugging" their modeling protocols—testing the various assumptions that went into the modeling protocol, to identify those assumptions that are flawed.

Of course, the ultimate test will be to steadily test these SALI curves to determine if they allow one to anticipate whether any SAR model will turn out to be prospectively useful. To that end, we have provided a Java application[33] that allows anyone to enter their structure–activity data, as well as the predictions of their model, to construct a SALI curve and to compute the SCI.

ACKNOWLEDGMENT

We would like to thank Gerald Maggiora and Mic Lajiness for much fruitful discussion on structure–activity landscapes.

APPENDIX

According to the definition of activity cliffs, significant cliffs will correspond to pairs of molecules with large activity differences. As a result, if noise is added to a dataset (either to the X- or Y-variables), one would expect it not to affect the significant cliffs. In terms of a SALI curve derived from noisy data, one would thus expect more variation on the left-hand side of the curve, rather than on the right-hand side, compared to curves obtained from noise-free data. However, Figures 7 and 8 indicate the opposite. As noted in section 6, this is partially due to the nature of the manner in which we calculate the value of the SALI curve at a given threshold. We present a mathematical analysis that explains the observed behavior of the SALI curves in the presence of noise.

The value (i.e., score) of the SALI curve at a threshold X can be written as

$$S = \frac{n_{\text{correct}} - n_{\text{wrong}}}{n_{\text{edge}}}$$

where n_{correct} and n_{wrong} represent the number of edges in the SALI graph that were predicted correctly and wrongly, respectively. The parameter n_{edge} is the total number of edges in the graph. Note that the aforementioned form is strictly only correct when there are no ties in the observed activities.

The result of this definition is that, for higher thresholds (corresponding to SALI graphs of the biggest activity cliffs), one extra misprediction (or correct prediction) can result in a relatively large change in the final value of the score function. This can be formally shown as follows. We note that

$$n_{\text{edge}} = n_{\text{correct}} + n_{\text{wrong}}$$

Now, let S' denote the score when the model predicts one extra edge correctly (for example, that which is due to the addition of noise to the X- or Y-variable). Thus,

$$S' = \frac{(n_{\text{correct}}+1)-(n_{\text{wrong}}-1)}{(n_{\text{correct}}+1)+(n_{\text{wrong}}-1)}$$

$$= \frac{n_{\text{correct}}-n_{\text{wrong}}}{n_{\text{edge}}} + \frac{2}{n_{\text{edge}}}$$

Similarly, we define S'' to be the score for the scenario where the model mispredicts one extra edge. Therefore, we have

$$S'' = \frac{(n_{\text{correct}}-1)-(n_{\text{wrong}}-1)}{(n_{\text{correct}}-1)+(n_{\text{wrong}}-1)}$$

$$= \frac{n_{\text{correct}}-n_{\text{wrong}}}{n_{\text{edge}}} - \frac{2}{n_{\text{edge}}}$$

Generally, for n extra correct or incorrect edge predictions, we will have

$$S' = \frac{n_{\text{correct}}-n_{\text{wrong}}}{n_{\text{edge}}} + \frac{2n}{n_{\text{edge}}} \quad (3)$$

$$S'' = \frac{n_{\text{correct}}-n_{\text{wrong}}}{n_{\text{edge}}} - \frac{2n}{n_{\text{edge}}} \quad (4)$$

Equations 3 and 4 represent the value of the SALI curve at a threshold X. For a high threshold, the number of edges in the graph is small. However, for lower thresholds, the number of edges can be very large. Thus, in the case of the PDGFR dataset at $X = 0.1$, the number of edges is 299, whereas for $X = 0.9$, the number of edges is 15. Therefore, generally, when $n_{\text{edge}} \gg 2n$, the factor $2n/n_{\text{edge}}$ becomes negligible.

This means that, given a SALI curve for a model and then a SALI curve from a model that uses noise (in X or Y), the effect of extra correct predictions or incorrect predictions of edge orderings at a given threshold value will cause more change at higher thresholds rather than at lower thresholds. As a result, we would expect to see the differences that are due to noise to appear on the right-hand side of the curve, rather than on the left-hand side of the curve.

REFERENCES AND NOTES

(1) Stouch, T.; Kenyon, J.; Johnson, S.; Chen, X.; Doweyko, A.; Li, Y. In silico ADME/Tox: Why Models Fail. *J. Comput. Aid. Mol. Des.* **2003**, *17*, 83–92.

(2) Kubinyi, H. Drug Research: Myths, Hype and Reality. *Nat. Rev. Drug Discov.* **2003**, *2*, 665–668.

(3) Walker, J. QSARs Promote More Efficient Use of Chemical Testing Resources—Carpe Diem. *Environ. Toxicol. Chem.* **2003**, *22*, 1651–1652.

(4) Contrera, J.; Matthews, E.; Kruhlak, N.; Benz, R. Estimating the Safe Starting Dose in Phase I Clinical Trials and No Observed Effect Level Based on QSAR Modeling of the Human Maximum Recommended Daily Dose. *Regul. Toxicol. Pharmacol.* **2004**, *40*, 185–206.

(5) Golbraikh, A.; Tropsha, A. Beware of q^2. *J. Mol. Graph. Model.* **2002**, *20*, 269–276.

(6) Kubinyi, H.; Hamprecht, F.; Mietzner, T. Three-Dimensional Quantitative Similarity−Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553=–2564.

(7) Norinder, U. Single and Domain Made Variable Selection in 3D QSAR Applications. *J. Chemom.* **1996**, *10*, 95–105.

(8) Van Drie, J. Pharmacophore Discovery: A Critical Review. In *Computational Medicinal Chemistry for Drug Discovery*; Bultinck, P., Winter, H., Langenaeker, W., Tollenare, J., Eds.; Marcel Dekker: New York, 2004; pp 437−461.

(9) Warren, G.; Andrews, C.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M.; Lindvall, M.; Nevins, N.; Semus, S.; Senger, S.; Tedesco, G.; Wall, I.; Woolven, J.; Peishoff, C.; Head, M. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(10) Charifson, P.; Kuntz, I. Recent Successes and Continuing Limitations in Computer-Aided Drug Discovery. In *Practical Application of Computer-Aided Drug Design*; Charifson, P., Ed.; Marcel Dekker: New York, 1997; pp 1−39.

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

(12) Guha, R.; Van Drie, J. The Structure−Activity Landscape Index: Identifying and Quantifying Activity−Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.

(13) Pearlman, D.; Charifson, P. Are Free Energy Calculations Useful in Practice? A Comparison with Rapid Scoring Functions for the p38 MAP Kinase Protein System. *J. Med. Chem.* **2001**, *44*, 3417–3423.

(14) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2005 (ISBN 3-900051-07-0).

(15) Pandey, A.; Volkots, D. L.; Seroogy, J. M.; Rose, J. W.; Yu, J.-C.; Lambing, J. L.; Hutchaleelaha, A.; Hollenbach, S. J.; Abe, K.; Giese, N. A.; Scarborough, R. M. Identification of Orally Active, Potent, and Selective 4-Piperazinylquinazolines as Antagonists of the Platelet-Derived Growth Factor Receptor Tyrosine Kinase Family. *J. Med. Chem.* **2002**, *45*, 3772–3793.

(16) Guha, R.; Jurs, P. C. Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179–2189.

(17) Stuper, A.; Brugger, W.; Jurs, P. *Computer Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.

(18) Tran, J. A.; Chen, C. W.; Jiang, W.; Tucci, F. C.; Fleck, B. A.; Marinkovic, D.; Arellano, M.; Chen, C. Pyrrolidines as Potent Functional Agonists of the Human Melanocortin-4 Receptor. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 5165–5170.

(19) Tran, J. A. Pyrrolidinones as Orally Bioavailable Antagonists of the Human Melanocortin-4 Receptor with Anti-Cachectic Activity. *Bioorg. Med. Chem. Lett.* **2007**, *15*, 5166–5176.

(20) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. Recent Developments of the Chemistry Development Kit (CDK)—An Open-Source Java Library for Chemo- and Bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2110–2120.

(21) Sutherland, J.; O'Brien, L.; Weaver, D. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships. *J. Chem. Comput. Sci.* **2003**, *43*, 1906–1915.

(22) Avery, M. A.; Alvim-Gaston, M.; Rodrigues, C. R.; Barreiro, E. J.; Cohen, F. E.; Sabnis, Y. A.; Woolfrey, J. R. Structure Activity Relationships of the Antimalarial Agent Artemisinin. The Development of Predictive in Vitro Potency Models Using CoMFA and HQSAR Methodologies. *J. Med. Chem.* **2002**, *45*, 292–303.

(23) Guha, R.; Jurs, P. The Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1440–1449.

(24) Digital Chemistry, http://www.digitalchemistry.co.uk, last accessed June 2008.

(25) Holloway, M. K.; Wai, J. M.; Halgren, T. A.; Fitzgerald, P. M. D.; Vacca, J. P.; Dorsey, B. D.; Levin, R. B.; Thompson, W. J.; Chen, L. J.; et al. A Priori Prediction of Activity for HIV-1 Protease Inhibitors Employing Energy Minimization in the Active Site. *J. Med. Chem.* **1995**, *38*, 305–317.

(26) Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recanatini, M. Toward a Pharmacophore for Drugs Inducing the Long QT Syndrome: Insights from a CoMFA Study of HERG K+ Channel Blockers. *J. Med. Chem.* **2002**, *45*, 3844–3853.

(27) Cramer, R., III.; Patterson, D.; Bunce, J. Comparative Molecular Field Analysis (CoMFA). I. Effect of Shape on Binding of Steroids to Carrier Protiens. *J. Am. Chem. Soc.* **1998**, *110*, 5959–5967.

(28) Gramatica, P. Principles of QSAR Models Validation: Internal and External. *QSAR Comb. Sci.* **2007**, *26*, 694–701.

(29) Sheridan, R.; Feuston, B.; Maiorov, V.; Kearsley, S. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912–1928.

(30) Stanforth, R.; Kolossov, E.; Mirkin, B. A Measure of Domain of Applicability for QSAR Modelling Based on Intelligent K-Means Clustering. *QSAR. Comb. Sci.* **2007**, *26*, 837–844.

(31) Schroeter, T.; Schwaighofer, A.; Mika, S.; Ter Laak, A.; Suelzle, D.; Ganzer, U.; Heinrich, N.; Muller, K.-R. Estimating the Domain of Applicability for Machine Learning QSAR Models: A Study on Aqueous Solubility of Drug Discovery Molecules. *J. Comput. Aid. Mol. Des.* **2007**, *21*, 485–498.

(32) Guha, R.; Jurs, P. C. Determining the Validity of a QSAR Model—A Classification Approach. *J. Chem. Inf. Model.* **2005**, *45*, 65–73.

(33) Guha, R. SALI Viewer, http://cheminfo.informatics.indiana.edu/~rguha/code/java/salivis, last accessed April 2007.

(34) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction for Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance Edge (MDE) Vector, λ. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387–394.

(35) Stanton, D.; Mattioni, B. E.; Knittel, J.; Jurs, P. Development and Use of Hydrophobic Surface Area (HSA) Descriptors for Computer Assisted Quantitative Structure-Activity and Structure-Property Relationship Studies. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1010–1023.

(36) Kier, L.; Hall, L. *Molecular Connectivity in Structure−Activity Analysis*; John Wiley and Sons: New York, 1986.

CI8001414