Published in final edited form as:
J Chem Inf Model. 2012 August 27; 52(8): 2181–2191. doi:10.1021/ci300047k.

Exploring Uncharted Territories - Predicting Activity Cliffs in Structure-Activity Landscapes

Rajarshi Guha
NIH Center for Advancing Translational Sciences 9800 Medical Center Drive Rockville, MD 20850

Abstract

The notion of activity cliffs is an intuitive approach to characterizing structural features that play a key role in modulating biological activity of a molecule. A variety of methods have been described to quantitatively characterize activity cliffs, such as SALI and SARI. However, these methods are primarily retrospective in nature; highlighting cliffs that are already present in the dataset. The current study focuses on employing a pairwise characterization of a dataset to train a model to predict whether a new molecule will exhibit an activity cliff with one or more members of the dataset. The approach is based on predicting a value for pairs of objects rather than the individual objects themselves (and thus allows for robust models even for small structure-activity relationship datasets). We extracted structure-activity data for several ChEMBL assays and developed random forest models to predict SALI values, from pairwise combinations of molecular descriptors. The models exhibited reasonable RMSE's though, surprisingly, performance on the more significant cliffs tended to be better than on the lesser ones. While the models do not exhibit very high levels of accuracy, our results indicate that they are able to prioritize molecules in terms of their ability to activity cliffs, thus serving as a tool to prospectively identify activity cliffs.

1 Introduction

The landscape paradigm for structure-activity relationship (SAR) data was first proposed 20 years ago[1] and has recently seen a resurgence with a number of studies describing new ways to quantify and visualize activity landscapes. When SAR data is viewed as a landscape, with the X-Y plane representing structural characteristics (which will usually be a 2-dimensional representation of a multi-dimensional descriptor space) and the Z-axis representing the observed activities, one can identify two broad types of regions on the landscape - smooth rolling regions, corresponding to set of molecules exhibiting continuous SAR (i.e., similar structures and similar activities) and rough, gorge-like regions (i.e., very similar structures, but large differences in activity) corresponding to molecules that exhibit SAR discontinuity. The latter have also been term activity cliffs.[2] From a medicinal chemistry point of view, the latter regions of a landscape can be the most interesting as they can provide insight into structural features that are key to improving (or conversely reducing) potency. There is a rich history of methods that have correlated structural differences with corresponding differences in activity – matched molecular pairs,[3] SAS maps[4] and more recently SALI[5] and SARI.[6] Both SALI and SARI focus on numerically characterizing a structure activity landscape. The former is defined for a pair of molecules as

$$S_{i,j} = \frac{|A_i - A_j|}{1 - sim(i, j)} \quad (1)$$

where A_i and A_j represent the observed activities of molecules i and j, and $sim(i, j)$ represents the structural similarity between the two molecules (usually based on some form

of fingerprint similarity). Using the SALI metric, one can take a collection of n molecules and represent them as an $n \times n$ matrix of SALI values - larger values representing more significant activity cliffs. The SARI approach is based on a score defined as

$$\text{SARI} = \frac{1}{2}\left(\text{score}_{cont} + (1 - \text{score}_{disc})\right) \quad (2)$$

where the individual score terms are derived on the basis of potency and pairwise similarities. The reader is referred to Ref. 6 for a detailed discussion of this approach.

In either case, one can take the numerical values and visualize them in a variety of ways ranging from heatmaps of SALI matrices to network representations.[5, 7] These visualizations then allow the user to explore the landscape, quickly identifying a range of activity cliffs, which an then be examined in detail. Apart from identifying individual activity cliffs, a variety of other SAR constructs, such as "activity ridges"[8] and multi-target landscapes[9] can also be identified and characterized.

A feature common to all, recently published work on activity landscapes is that they are primarily retrospective. That is, the methodologies developed are used to analyze SAR datasets for which activities have already been experimentally obtained. For example, using the SALI, one can characterize SAR patterns in a dataset but does not provide insight into whether a new molecule may be part of an activity cliff, with respect to the original dataset. A number of applications have attempted to extract SAR rules based on the landscape (e.g., similarity potency trees[10] and multi-target landscape analyses[11]) or directly identify structural modifications that lead to activity cliffs.[12, 13]

1.1 Motivation

The preceding discussion highlights the utility of retrospective analyses of SAR data using the activity landscape paradigm. But equally, if not more, interesting is determining whether a new, untested molecule might be an activity cliff in the context of the original dataset. For the remainder of this work we focus on the use of SALI to quantify activity landscapes. More specifically, the ability to *predict* SALI values would be useful as it would allow us to both fill in empty regions of an activity landscape as well as extend a structure-activity landscape. Note that this approach to expanding the extent of a SAR dataset does not lend itself to scaffold hopping since the premise of scaffold hopping is that one generates new cores, which differ substantially from the starting structure.

In traditional QSAR modeling approaches, one simply predicts the activity of a new molecule and would then evaluate the SALI (or SARI or some other measure) to determine whether the molecule leads to an activity cliff. However, the fact that an activity cliff represents a SAR discontinuity[2] implies that most statistical and machine learning methods will be unlikely to predict very different activities for two structurally similar molecules. In other words, a new molecule, similar to a subset of the training set, will tend to have a predicted value that is similar to those molecules, rather than a drastically different value.

An alternative approach, that is the focus of this paper, is to directly predict SALI values for pairs of molecules. Thus rather than predict individual activities, we predict SALI values for pairs of molecules. This approach is somewhat similar to the SPREAD method[14] which identified substructures that were predictive of activity differences. However, our solution considers both activity differences and structural similarities. As a result, instead of ranking compounds in terms of their predicted activity, we instead rank a compound in terms of its predicted SALI; i.e., its predicted ability to exhibit an activity cliff when paired with other

molecules in the dataset. This approach could be useful when deciding how far to extend an analog series as well as prioritizing scaffolds for further study.

This does not completely alleviate the problem of discontinuities, since SALI values are infinite when the T_C is 1.0. However, predicting SALI values allows us to work with smaller datasets (since the objects to predict are pairs of molecules), that would ordinarily lead to unreliable models if we were working with activities. Of course, this means that the approach is not practical for very large datasets.

The paper is organized as follows. Section 2 describes the datasets used in this study. Section 3 presents the methodology we employ to predict activity cliffs and Section 4 discusses the results of the predictive models. Finally, Sections 5 and 6 discusses some of the issues underlying this approach and possible extensions of this work.

2 Datasets

For this study we considered a number of datasets, which are summarized in Table 1. The Cav-alli dataset was employed in our previous studies consisted of 30 molecules studied by Cavalli et al[15] as possible hErg inhibitors using a CoMFA modeling approach. The reported endpoint for the molecules was a pIC_{50}. The remaining datasets were obtained from ChEMBL. All three assays in-volved direct binding to a human target and we considered the subset of molecules in each assay that had non-censored experimental values. The Costanzo dataset[16] consisted of 60 α-ketoheterocyclic inhibitors of α-thrombin. The reported IC_{50} values ranged from 3 nM to 82 μM. The Kalla dataset[17] consisted of 38 8-(C-4-pyrazolyl) xanthines, identified as antagonists of the A2B adenosine receptor. The reported K_i values ranged from 0.9 nM to 42 μM. Finally, the Dai dataset[18] consisted of 44 3-aminoindazole derivatives studied for their ability to inhibit the VEGF and PDGF receptor families, with IC_{50} values ranging from 3 nM to 12 μM.

3 Methodology

As noted above, the problem of identifying activity cliffs involving new molecules can be reduced to predicting the SALI value for the new molecule and a preexisting molecule. Note that this does not result in an activity prediction for the new molecule; rather, it allows us to rank a set of new molecules in terms of their predicted ability to exhibit a significant activity difference with respect to one or more of the molecules in the training set.

Given a training set of N molecules, we generate a new training set of $\frac{N(N-1)}{2}$ objects, where each object is a pair of molecules from the original training set. The dependent variable for each pair i, j, is the SALI value, $S_{i,j}$. SALI values were evaluated using the 1051-bit BCI keyed fingerprints (Digital Chemistry, UK) or the CDK[19, 20] 1024-bit path fingerprints. Based on the definition of SALI, it is possible that a pair of molecules have a $T_C = 1.0$ resulting in infinite values. For such cases, we replaced the infinite value with the highest non-infinite SALI value for that dataset.

The next step is to generate a set of independent variables. Since the new dataset consists of pairs of molecules, we consider the descriptors for the resultant objects as a function of the descriptor values of the individual molecules. For an object, representing the i'th and j'th molecules, its descriptor vector can be taken as the arithmetic mean of the descriptor vectors of the individual molecules. We denote this aggregation function as f_{mean}. Alternative functions that were investigated included the absolute difference of the individual descriptors, denoted by f_{diff} and the geometric mean of the individual descriptor vectors, denoted by f_{geom}.

Given the independent and dependent variables for the pairwise dataset, we can now proceed to model development. For this study we focused on the use of random forest models.[21] This was motivated by the fact that such models tend not to overfit the training data (given a sufficiently large number of trees) and the fact that the algorithm implicitly performs feature selection. As a result, this allows us to forgo an explicit feature selection step and work directly with the descriptor pool (after removal of correlated and constant descriptors). Furthermore, there is no reason, *a priori* to assume that the underlying SAR's are linear. A random forest model, being an algorithmic approach[22] (as opposed to a distributional one such as linear regression) makes no such assumptions. We employed the implementation of random forest from the randomForest package, in R 2.11.0.[23] The models were built using 500 trees, and randomly sampled the descriptor pool for a given dataset, using $\sqrt{(N)}$ descriptors at a time, where N is the number of descriptors in the pool. By default, the method builds individual trees using 63% of the dataset and tested on the remainder (the so called out-of-bag data). The final predicted values are obtained by averaging the predictions for each observation from from all 500 trees. We initially employed a training/test set split (80% of the dataset was randomly assigned to the training set) and these results are reported in Table 2. We also developed models (Section 4.1) using the entire dataset and observed similar RMSE and R^2 values. Given the fact that the random forest algorithm builds a model based on out-of-bag data, the similarity in performance of models built using the entire dataset and models built using a training/test split is not surprising.

We employed the CDK to evaluate 109 2D and constitutional descriptors for each molecule. For each dataset, we performed objective feature selection by removing descriptors with constant or near-constant values followed by removal of descriptors that are highly correlated with others (using an R^2 cutoff of 0.8). The sizes (N_{desc}) of the final descriptor pool for each dataset are summarized in Table 1.

A side effect of the proposed approach is that one can build relatively robust models for datasets of small size – say, 20 molecules. Of course, a large training set is just one aspect of a reliable model and other considerations such as diversity, descriptor selection and so on still play an important role.

3.1 Is structural information duplicated?

The preceding discussion raises the issue of explicitly duplicating structural information in the independent and dependent variables. More specifically, both the dependent variable (i.e., the SALI values) and the aggregated (topological) descriptors characterize the molecules' structure. One might therefore ask whether models built using such data perform over-optimistically. Given the nature of the descriptors used in the independent and dependent variables (the former being based on multiple atom and bond features and the latter derived from purely topological paths), we believe that such a phenomenon is unlikely. Fig. 1 displays a histogram of the pairwise Pearson correlations between the SALI values and each of the descriptor values for the Cavalli[15] dataset, using the different aggregation functions described above. It is evident, that the highest R^2, between any of the descriptors and the SALI values is less than 0.15. Similar behavior was observed for all the datasets used in this study. These observations suggest that problems arising from the inclusion of structural information simultaneously in the dependent and independent variables is minimal.

4 Results

We first consider the application of the SALI prediction methodology to the Cavalli dataset. We developed three random forest models, using the three aggregation descriptor

aggregation functions described in Section 3. Fig. 2 presents the predicted versus observed SALI values from the three models and Table 2 summarizes the performance metrics for the three models. Overall the different aggregation functions do not differ dramatically in terms of the final model performance as root mean square error (RMSE) on the test set ranges from 1.01 to 1.07 log units, which is lower compared to the standard deviation of the dependent variable. However, f_{mean} appears to lead to the best performance (RMSE = 1.01 and R^2 = 0.58). In contrast, the RMSE and R^2 values for Y-scrambled[24] models using the different aggregation functions ranged from 1.58 to 1.61 log units and 0.003 to 0.01, respectively.

Note that the predictions on the low end of the SALI values are not as important as those at the high end - simply because low SALI values correspond to small activity cliffs, which are likely not very interesting. Given that observation, it is encouraging to observe that all three models perform relatively well at the higher end of the SALI spectrum. For this example, the three most significant activity cliffs are in fact not very significant activity cliffs in an absolute sense - the T_C for the three pairs are just 0.2, 0.30 and 0.29, though the activity differences were 4.93, 5.0 and 5.11 log units respectively. But more importantly, it is clear from all three panels that the there appears to be two different systematic prediction errors - smaller SALI values are overestimated whereas larger values are underestimated. A possible explanation for this behavior is that the descriptors do not sufficiently differentiate observations with low and high SALI values. In other words, the descriptor profile for observations with low SALI values is sufficiently similar to those with high SALI values. To test this, we split the Cavalli dataset into two groups - low and high - using the median SALI value as the cutoff. We then evaluated the pairwise descriptor distance distributions for each each of the groups, using the Euclidean distance between all descriptors in the pool. Figure 3 indicates that the distance distributions are essentially identical. Though random forest models employ a subset of the descriptor pool for each tree, the same reasoning can be applied to these subsets. These trends appear in the other datasets (Figure 4) but to varying degrees and not always consistently. For example, in the Costanzo dataset, we see that the higher SALI values are somewhat underestimated. The concept of descriptor similarity could also be applied to these cases. But a more fundamental explanation of this behavior is not available at this time.

We observed that the performance of models based on geometric mean aggregation function did not differ substantially from those built with the arithmetic mean. This was also true of the other datasets and so the following discussion omits the results obtained when using f_{geom}. In fact, a paired *t*-test between the predicted SALI values (for the training set) using models built with the different aggregation functions (i.e., f_{mean} vs f_{diff}, etc.) indicated that the differences between the sets of predictions was statistically insignificant (all *p*-values > 0.1). Fig. 4 summarizes the models built on the ChEMBL datasets using f_{diff} and f_{mean} as the aggregation functions. Table 2 reports the model statistics Of the three, the Costanzo stands out as being rather poor given the training and test set RMSE values being quite close to (or greater than) the standard deviation of the dependent variable. While the differences in model performance do not vary significantly with the aggregation function used, we see that for all the ChEMBL datasets, f_{diff} leads to a slightly better model in terms of RMSE and R^2 than when using f_{mean}. As with the Cavalli dataset, the SALI models exhibit more variance for small SALI values (i.e., less significant cliffs) but appear to perform better at the higher end of the SALI values. However in all three cases, systematic prediction errors are apparent. Based on these observations, we arbitrarily chose f_{diff} as the aggregation function for subsequent analyses. Table 3 and Figure 5 summarizes the performance of the models built using the reported activity data (original, single molecule activity) for the Costanzo, Kalla and Dai datasets. Each of these models employed the entire dataset (rather than splitting into a training and testing set) and exhibit relatively poor performance as indicated

by the RMSE (ranging from 0.76 to 1.00 log units) and R^2 (0.18 to 0.34). The model for the Kalla dataset is especially poor, given the RMSE is greater than the standard deviation of the dependent variable. The poor performance of these models can be partially ascribed to the small sizes of the dataset. It should be note that this methodology is not aimed at modeling the original, usually small, SAR dataset. It is well known (and evident from the current analyses) that small datasets do not necessarily lead to very robust or reliable machine learning models. Instead, the methodology focuses on *pairwise* datasets –which are usually large enough for a statistically significant model. Based on the poor performance observed for the Costanzo dataset, we excluded it from our subsequent discussions.

4.1 An alternative validation approach

Since the model is built on pairs of observations, performance measures characterizing the model are based on the accuracy of pairwise SALI values. Thus the preceding section validated models based on holding out pairs of observations and evaluating the RMSE and R^2 for the predicted values of these held out pairs.

The model can also be tested in a slightly different approach by checking whether it can correctly identify cliffs between a new molecule and the members of the training set. In fact, such a scenario would be the anticipated, prospective use case for such a modeling methodology. To test this, we consider the original set of n individual molecules in a dataset and hold out m molecules (as opposed to holding out a set of observations from the pairwise dataset). We then evaluate the SALI values $(n - m)(n-m-1) / 2$ pairs of molecules and develop a random forest model. In contrast to the models previously described, we rebuilt these models using the entire set of n molecules. The model is then used to predict the SALI values between the m holdout molecules and the $(n - m)$ remaining molecules. As a result, instead of a just m predicted values, we obtain $m \times (n - m)$ predicted values for the test set.

We first performed this analysis with the Kalla dataset. Since predictions of new molecules involve n predictions per molecule, we removed ten molecules from the original dataset and rebuilt the random forest model. The hold out molecules included cases that displayed significant activity cliffs with various members of the training set, as well as other cases that did not really exhibit cliffs with any member of the training set. We then evaluated the pairwise descriptor values for these ten molecules with the training set molecules and obtained predictions of the log(SALI) values. Figure 6A summarizes the performance of the model on the training set as well as for the prediction set. From a numerical point of view, the performance of the model on the prediction set degrades somewhat (RMSE of 0.41 versus 0.63 for the training set). As noted above, the variance of the prediction is higher at the lower scale of log(SALI) values. This is highlighted in Figure 7, where we divided the observed log SALI values into three groups, *low*, *medium* and *high* (based on the quartiles of the values) and aggregated the prediction residuals for the observations belonging to each group. Clearly, for lower SALI values (i.e., less significant cliffs), the prediction errors are larger and more widely distributed. While 6A indicates that a number of the more significant activity cliffs are quite well predicted, Figure 7 indicates that as a group the observations with higher log SALI values tend to exhibit a small range of non-zero residual errors. However it is clear from the scatter plot that number of relatively significant activity cliffs have been underestimated.

Given that large SALI values can arise due to high structural similarity, even when the activity difference is small, simply examining log(SALI) values is not completely informative. Specifi-cally, we are interested in the predictions for the actual activity cliffs, where both the difference in activity and the structural similarity is very high. Figure 6B visualizes this information. Each point corresponds to a pair of molecules, which are shaded by the absolute prediction residual for that pair. Thus, the “true” cliffs are represented by

points lying towards the bottom right corner of the plot. In this case we see that there are a number of pairs of molecules with $1 - T_C \leq 0.1$ and 100-fold or better difference in activity. While many of these pairs are associated with a high prediction residual, there are a number with low to medium residuals.

Figure 8 displays the structures of three hold out molecules (top row) and a selection of structures, which which these hold out molecules were predicted to exhibit an activity cliff. For each molecule, the K_i value is listed along with the absolute residual value (log(SALI) units). For molecules **349288** and **349699**, the predicted cliffs are relatively accurate, as evidenced by the low residuals. For **349138** on the other hand, most of the residuals were were relatively high. While we have highlighted a number of cliffs, Figure 6C makes it clear that the predictions for this molecule were, on average, worse than for the other two molecules.

Figure 9 summarizes the results for the same set of analyzes described above, applied to the Dai dataset. As before we removed ten molecules (top row of Figure 10) and rebuilt the random forest on the paired cases with the remaining 34 molecules. We then predicted the pairwise SALI values for the ten hold-out molecules with the remaining 34 molecules. A plot of the predicted versus observed log(SALI) values for the training and prediction sets are displayed in Figure 9A. As before the model performs poorly on the lesser cliffs, but improves for the more significant cliffs (RMSE for the training set and prediction sets were 0.37 and 0.43 log units respectively). Figure 7 also highlights the fact that the model performs poorly at the lower end of the log(SALI) spectrum. However, the 2 most significant cliffs are under-estimated. However, in comparison to the Kalla dataset, the distribution of absolute residuals aggregated by the hold-out molecule (Figure 9C) are similar, with a median absolute residual of less than 0.3 log units. For this dataset, the number of "real" cliffs is relatively low as shown in Figure 9B. Interestingly, the compound pairs that display moderate to significant cliffs are relatively well predicted - in fact, the maximum residuals are observed for the least significant cliffs.

We then considered three of hold-out molecules and some of the members of the training set with which they are predicted to show activity cliffs. For this dataset, there are relatively few "true" activity cliffs. For example, while molecules **371259** and **371307** have a 1000-fold difference in activity, their $T_C = 0.68$, and would not represent a truly significant cliff. However, the pair is well predicted with a residual of 0.04 log units. if we consider the subset of predictions where the $T_C \geq 0.8$ we note that the median absolute residual is 0.24 log units. However, the activity differences are not always very large ranging from 1.1-fold to 13-fold. Two such examples are shown in Figure 10, where the only difference is in the position of the methyl substituent (**371259** and **371224**). Similar behavior is observed with the other hold-out molecules.

5 Discussion

The results described above suggest that predicting pairwise SALI values is a useful way to identify whether a new molecule will form an activity cliff with one or more members of the training set. As with all QSAR models, one is limited by the nature of the underlying data, both in terms of accuracy as well as applicability. We have selected a few ChEMBL assays for the purposes of highlighting this approach. Given that most of the datasets are small in size (30 to 50 molecules), models built on the actual activity data do not perform very well. In contrast the models built on the pairwise data usually exhibit improved performance - which could certainly be ascribed to the larger dataset involved. However, in cases where the pairwise model does not exhibit improved performance, we believe that predictive modeling of activity cliffs in the original dataset will not be fruitful.

Given that many of the models show relatively poor statistics on the pairwise data, it is important to note that much of this is due to the large variance in the predictions for the less significant cliffs. One possible reason for the increase in variance for these portions of the dataset is the fact that the distribution of the log(SALI) values tends to be skewed to the right (Figure 11). In general we see that the smaller SALI values constitute a relatively small portion of the dataset.

Another factor affecting the models built on pairwise data is the descriptors that are employed. While the choice of initial descriptors is always open to debate, we partially avoid a biased selection by employing a random forest model to perform implicit feature selection. However, given an initial set of descriptors we observe that the models are robust to the different aggregation functions to derive a pairwise descriptor representation from the descriptors for the individual compounds. This is a useful observation given that there appears to be no rule to select one aggregation function over another *a priori*. While random forests do allow us to avoid feature selection, it is clear that other models, such as neural networks or support vector machines, could lead to more predictive solutions. At this stage, our aim is to highlight the modeling procedure and we believe that the random forest allows us to highlight the utility of this approach without sacrificing too much by way of predictive performance.

6 Conclusions

We have presented an approach to extending a structure-activity landscape in an indirect fashion, by predicting the propensity of a new molecule to exhibit an activity cliff with one or more molecules in a pre-existing SAR dataset. The method is based on building a model on pairwise SALI values (dependent variable) and pairwise aggregated descriptor values (independent variables). For a new molecule, we obtain n pairwise SALI predictions (since we must estimate the SALI value for each training set molecule and the test molecule). The predicted SALI values can then be used to judge whether the new molecule will exhibit a cliff and whether such a cliff is in the desired direction (i.e., improving potency versus worsening potency). To test this strategy we have developed random forest models to predict SALI values for several ChEMBL datasets. While the model performance statistics are not stellar, we observe that this is primarily due to high prediction variance at the lower range of SALI values. In contrast, the more significant cliffs are relatively well predicted, though, unsurprisingly, the most significant cliffs are not always well predicted.

In summary, this approach extends the activity cliff concept, from that of a retrospective analysis tool to a prospective tool that could be used to guide synthetic campaigns in their goals of improved potency.

References

1. Johnson, M.; Maggiora, G. Concepts and Applications of Molecular Similarity;. John Wiley & Sons; New York: 1990.
2. Maggiora GM. On Outliers and Activity Cliffs–Why QSAR Often Disappoints. J Chem Inf Model. 2006; 46:1535–1535. [PubMed: 16859285]
3. Leach A, Jones H, Cosgrove D, Kenny P, Ruston L, MacFaul P, Wood J, Col-clough N, Law B. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. J Med Chem. 2006; 49:6672–6682. [PubMed: 17154498]
4. Shanmugasundaram, V.; Maggiora, G. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. CINF-032. 222nd ACS National Meeting; Chicago, IL, United States. Washington, D.C: American Chemical Society; 2001.

5. Guha R, Van Drie J. The Structure-Activity Landscape Index: Identifying and Quantifying Activity-Cliffs. J Chem Inf Model. 2008; 48:646–658. [PubMed: 18303878]
6. Peltason L, Bajorath J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. J Med Chem. 2007; 50:5571–5578. [PubMed: 17902636]
7. Wawer M, Peltason L, Weskamp N, Teckentrup A, Bajorath J. Structure-Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure-Activity Relationship Indices. J Med Chem. 2008; 51:6075–6084. [PubMed: 18798611]
8. Vogt M, Huang Y, Bajorath J. From Activity Cliffs to Activity Ridges: Informative Data Structures for SAR Analysis. J Chem Inf Model. 2011; 51:1848–1856. [PubMed: 21761918]
9. Dimova D, Wawer M, Wassermann AM, Bajorath J. Design of Multitarget Activity Landscapes That Capture Hierarchical Activity Cliff Distributions. J Chem Inf Model. 2011; 51:258–266. [PubMed: 21275393]
10. Wawer M, Bajorath J. Similarity-potency trees: A method to search for SAR information in compound data sets and derive SAR rules. J Chem Inf Model. 2010; 50:1395–1409. [PubMed: 20726598]
11. Hu Y, Bajorath J. Molecular Scaffolds with High Propensity to Form Multi-Target Activity Cliffs. J Chem Inf Model. 2010; 50:500–510. [PubMed: 20361784]
12. Wassermann AM, Bajorath J. Chemical Substitutions that Introduce Activity Cliffs Across Different Compound Classes and Biological Targets. J Chem Inf Model. 2010; 50:1248–1256. [PubMed: 20608746]
13. Peltason L, Weskamp N, Teckentrup A, Bajorath J. Exploration of Structure-Activity Relationship Determinants in Analogue Series. J Med Chem. 2009; 52:3212–3224. [PubMed: 19397320]
14. Scheiber J, Jenkins JL, Bender A, Milik M, Mikhalov D, Sukuru S, Cornett B, Whitebread S, Laszlo U, Davies J, Glick M. SPREAD - exploiting chemical features that cause differential activity behavior. Stat Anal Data Mining. 2009; 2:115–122.
15. 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. [PubMed: 12190308]
16. Costanzo MJ, Almond HR Jr, Hecker LR, Schott MR, Yabut SC, Zhang HC, Andrade-Gordon P, Corcoran TW, Giardino EC, Kauffman JA, Lewis JM, de Garavilla L, Haertlein BJ, Maryanoff BE. In-depth study of tripeptide-based alpha-ketoheterocycles as inhibitors of thrombin. Effective utilization of the S1′ subsite its implications to structure-based drug design. J Med Chem. 2005; 48:1984–2008. [PubMed: 15771442]
17. Kalla RV, Elzein E, Perry T, Li X, Palle V, Varkhedkar V, Gimbel A, Maa T, Zeng D, Zablocki J. Novel 1,3-disubstituted 8-(1-benzyl-1H-pyrazol-4-yl) xanthines: high affinity and selective A2B adenosine receptor antagonists. J Med Chem. 2006; 49:3682–3692. [PubMed: 16759111]
18. Dai Y, et al. Discovery of N-(4-(3-amino-1H-indazol-4-yl)phenyl)-N′-(2-fluoro-5-methylphenyl)urea (ABT-869), a 3-aminoindazole-based orally active multitargeted receptor tyrosine kinase inhibitor. J Med Chem. 2007; 50:1584–1597. [PubMed: 17343372]
19. 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.
20. Steinbeck C, Han YQ, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinfor-matics. J Chem Inf Comput Sci. 2003; 43:493–500. [PubMed: 12653513]
21. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. Classification and Regression Trees;. Chap-man & Hall/CRC; Boca Raton, FL: 1984.
22. Breiman L. Statistical Modeling: Two Cultures. Stat Sci. 2001; 16:199–231.
23. R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2008.
24. Rucker C, Rucker G, Meringer M. y-Randomization and Its Variants in QSPR/QSAR. J Chem Inf Model. 2007; 47:2345–2357. [PubMed: 17880194]

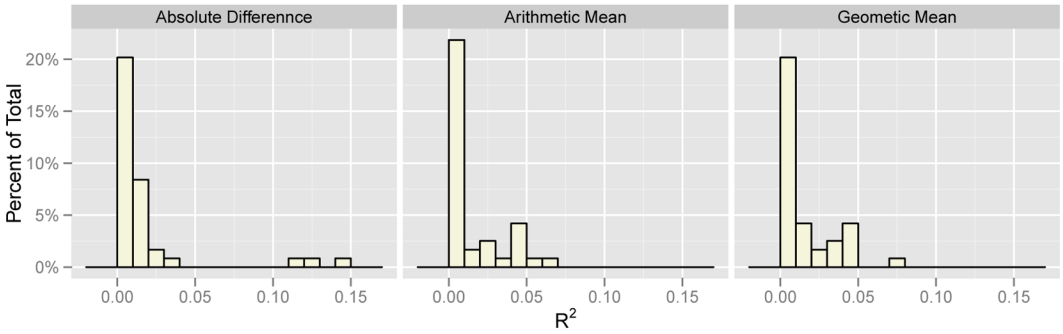

Figure 1.
Distribution of Pearson correlations between SALI values and the descriptors for the Cavalli dataset. Three descriptor aggregation functions are considered.

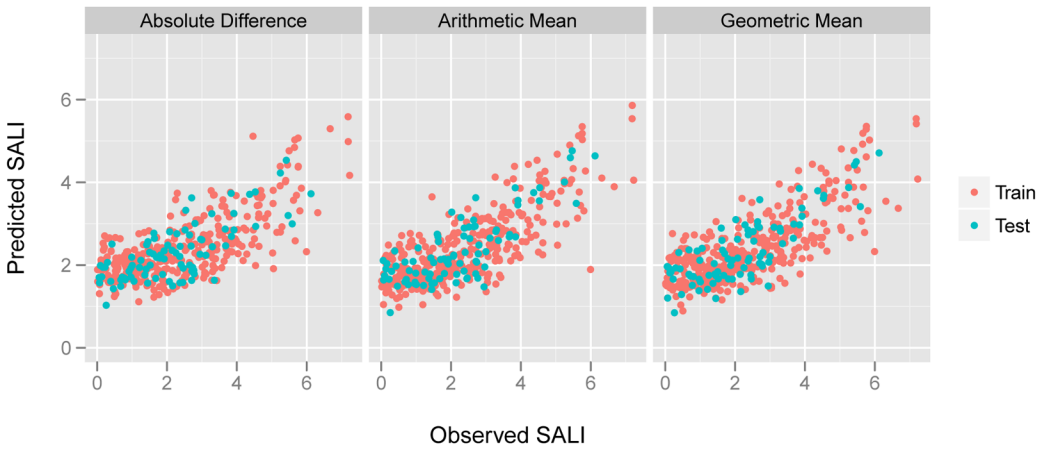

Figure 2.
Plots of predicted versus observed SALI values obtained using random forest models, on the Cavalli dataset. Each panel corresponds to the use of a different descriptor aggregation function.

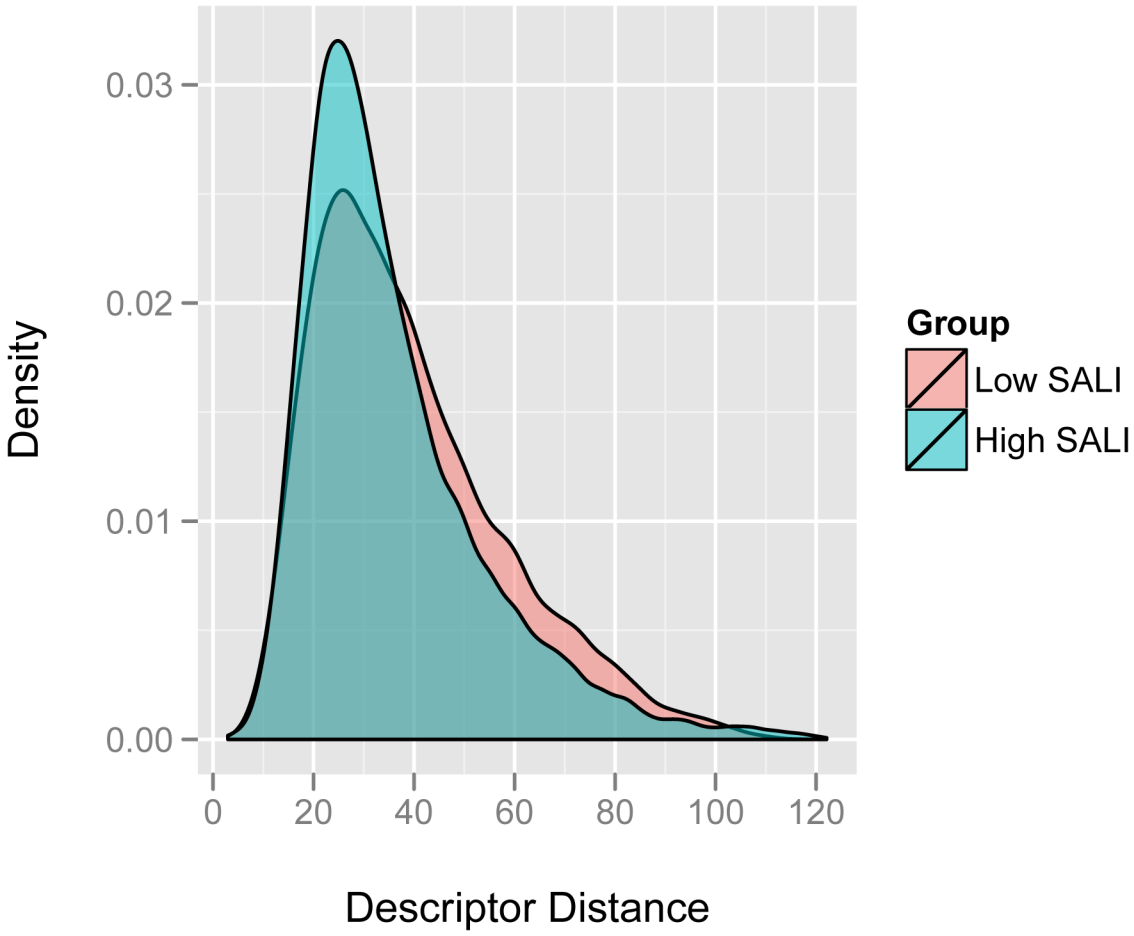

Figure 3.
A comparison of the pairwise descriptor distribution for low and high SALI values in the Cavalli dataset. This figure employed the descriptor values generated using the f_{diff} aggregation function and Euclidean distances were evaluated using all the descriptors in the pool. The *low* group is defined as those observations with SALI less than 2.03 and the remainder are assigned to the *high* group.

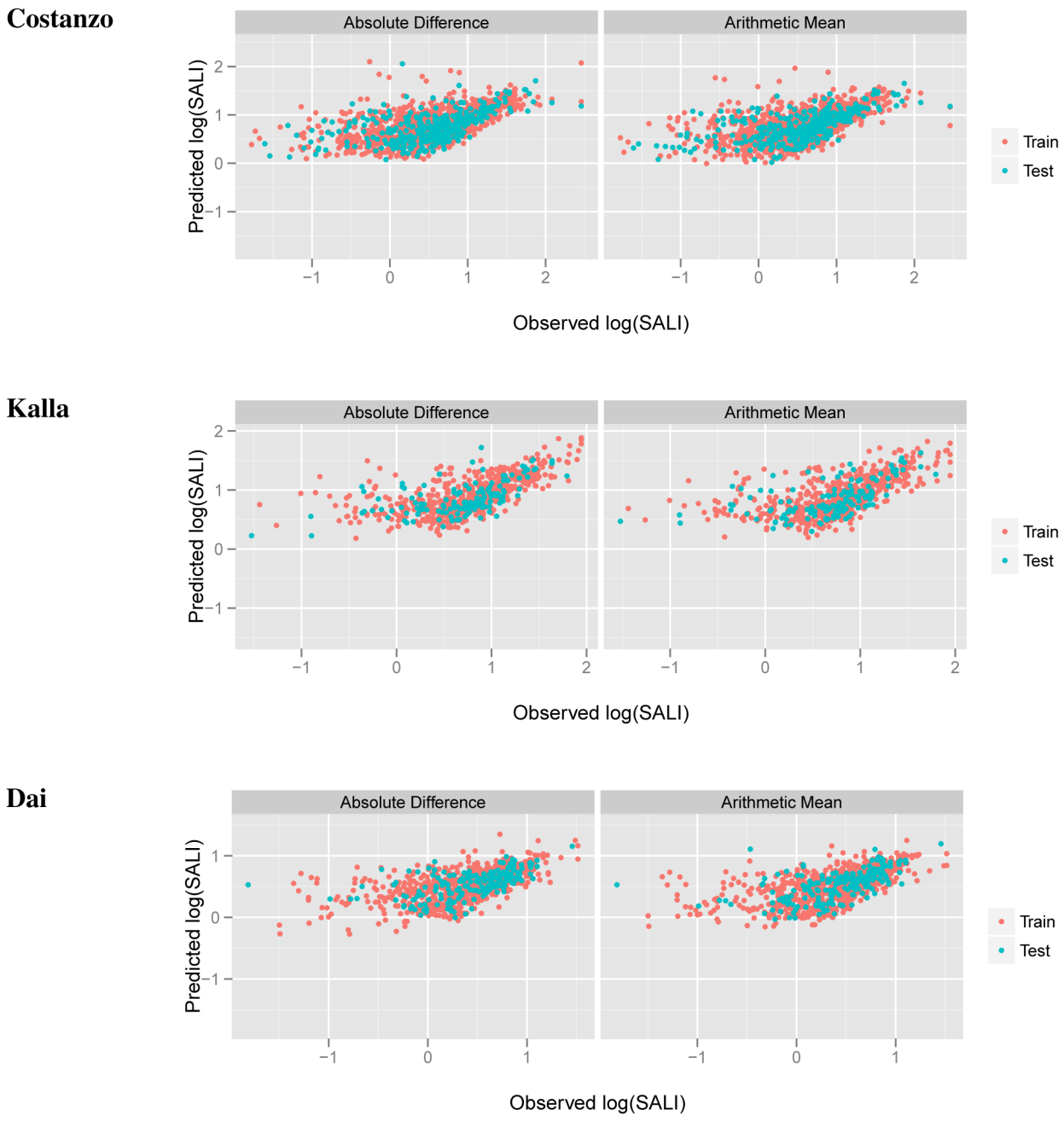

Figure 4.
Predicted versus observed SALI values, obtained from random forest models for the three ChEMBL datasets. The three plots correspond to the two different aggregation functions (f_{diff} and f_{mean} respectively).

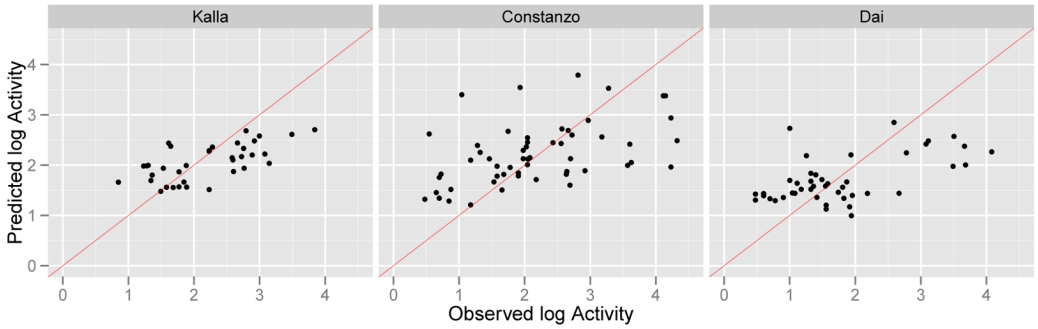

Figure 5.
Results of random forest models developed using the three ChEMBL dataset. In each case the model was built using the $\log_{10}$ of the observed activity.

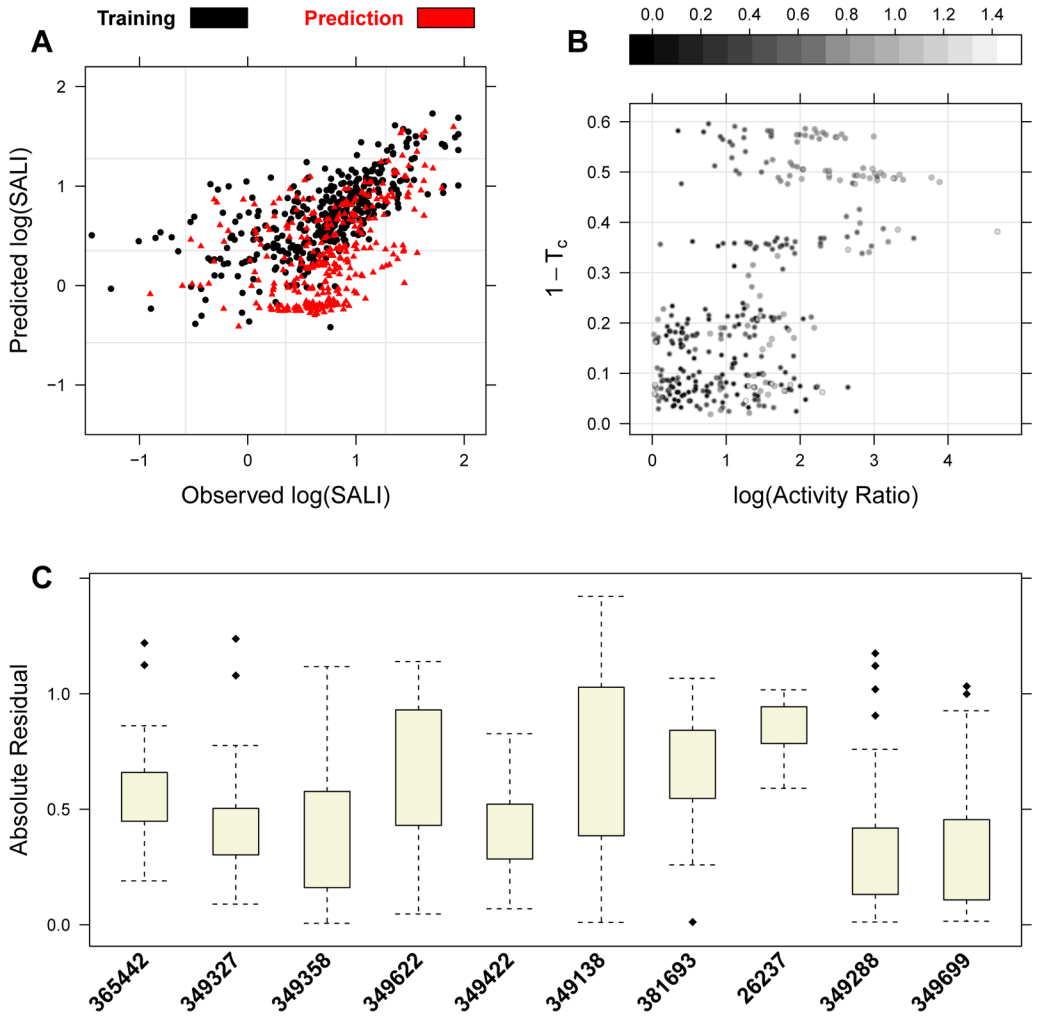

Figure 6.

Detailed analysis of SALI predictions for the Kalla dataset. **A** - a plot of predicted versus observed log(SALI) values for the training set and the hold out set. **B** - a summary of the training set, which plots the structural difference versus the logarithm of the ratio of the activities for each pair of molecules in the prediction set. Points are shaded by their absolute residual. **C** - a box plot summarizing the distribution of residuals associated with predictions from each of the hold out molecules.

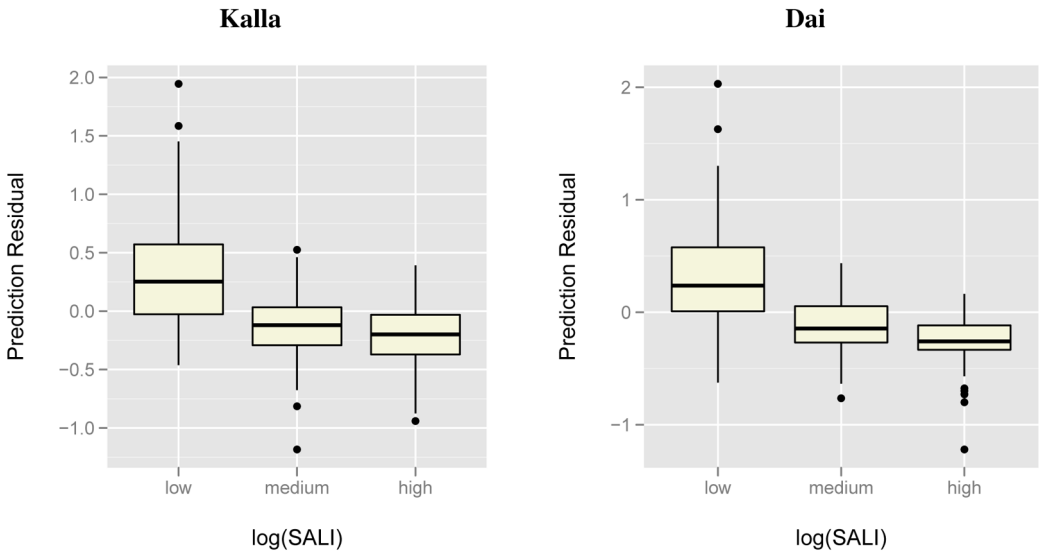

Figure 7.

A summary of the prediction residuals for log(SALI) values, grouped by whether the actual log(SALI) value for that observation was *low, medium* or *high*. The grouping is based on the quartiles of the observed values.

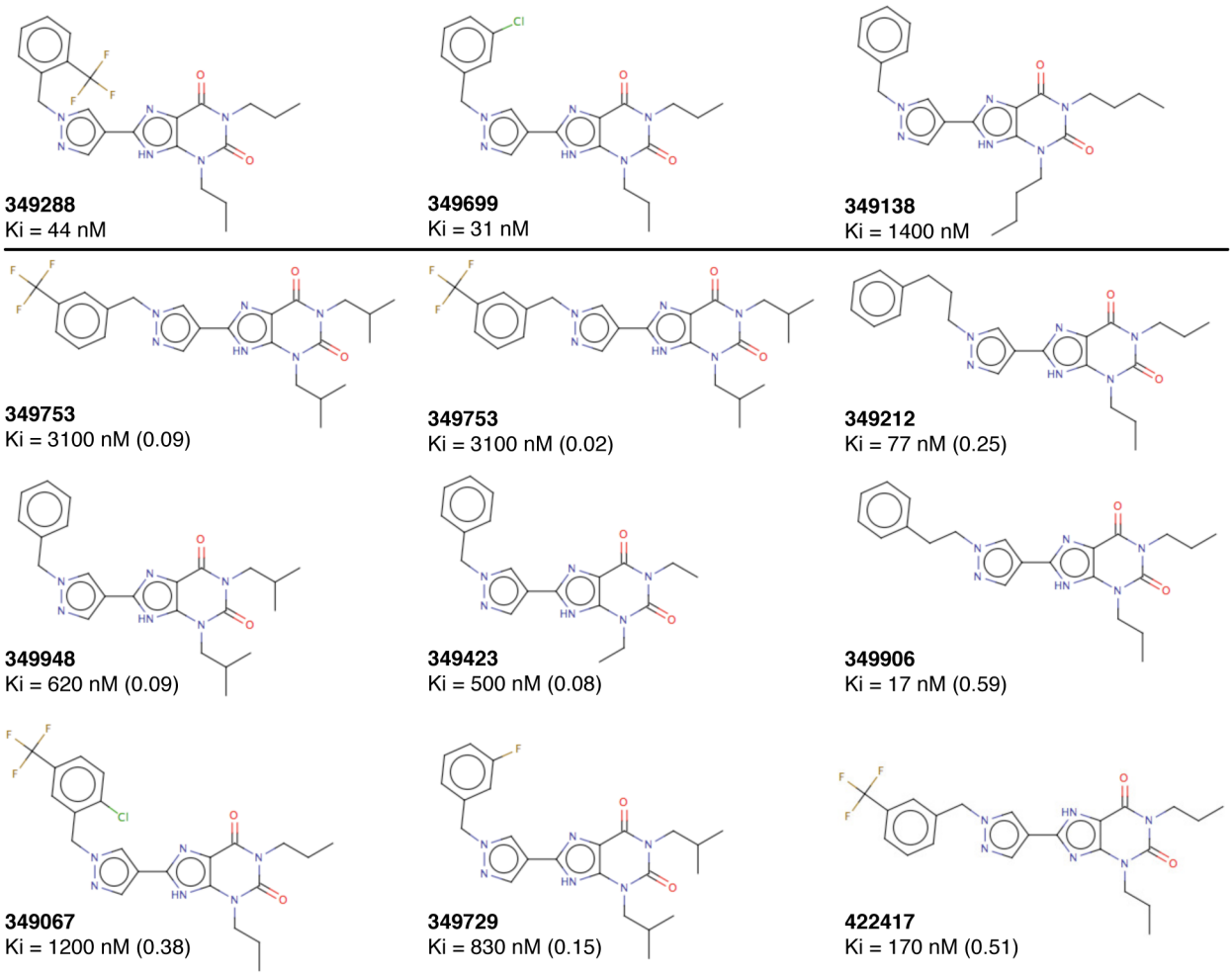

Figure 8.

Three of the hold out molecules for the Kalla dataset and training set members with which the hold outs exhibit predicted activity cliffs. Bold numbers are ChEMBL MOLREGNO values and numbers in parentheses are the absolute prediction residual in log(SALI) units.

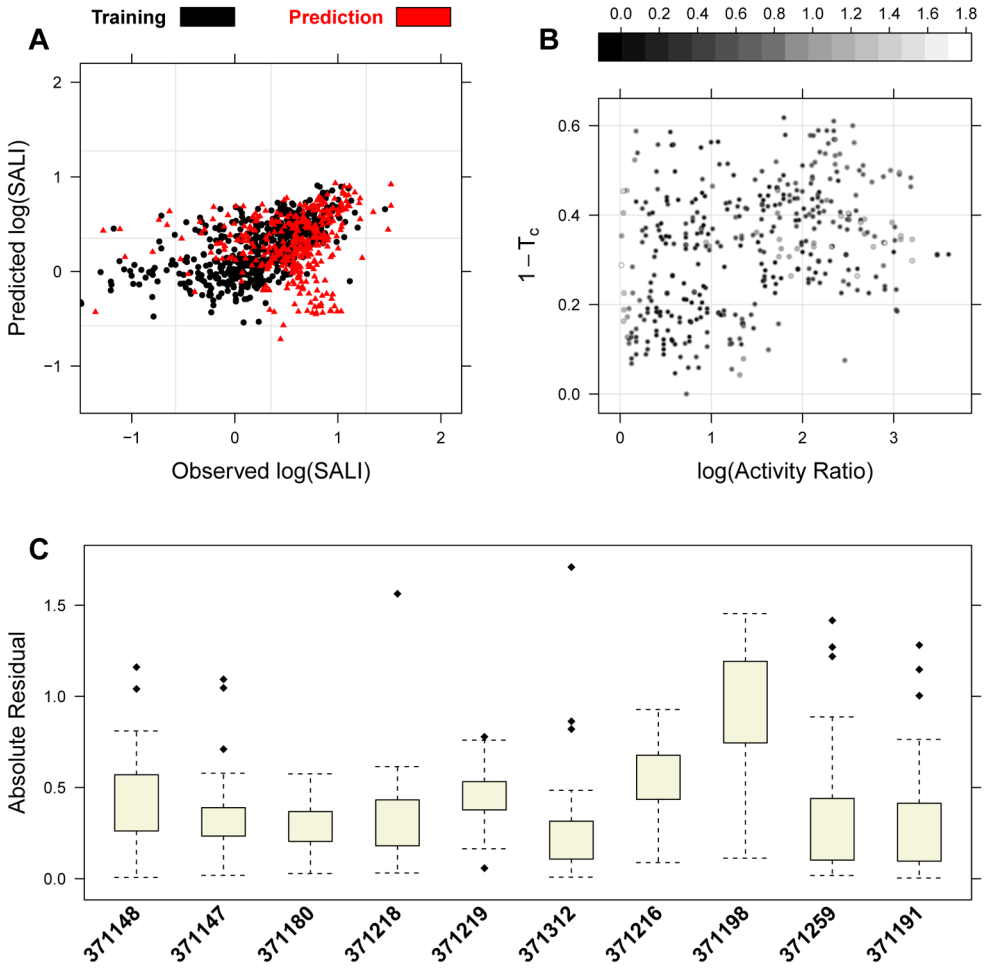

Figure 9.

Detailed analysis of SALI predictions for the Dai dataset. **A** - a plot of predicted versus observed log(SALI) values for the training set and the hold out set. **B** - a summary of the training set, which plots the structural difference versus the logarithm of the ratio of the activities for each pair of molecules in the prediction set. Points are shaded by their absolute residual. **C** - a box plot summarizing the distribution of residuals associated with predictions from each of the hold out molecules.

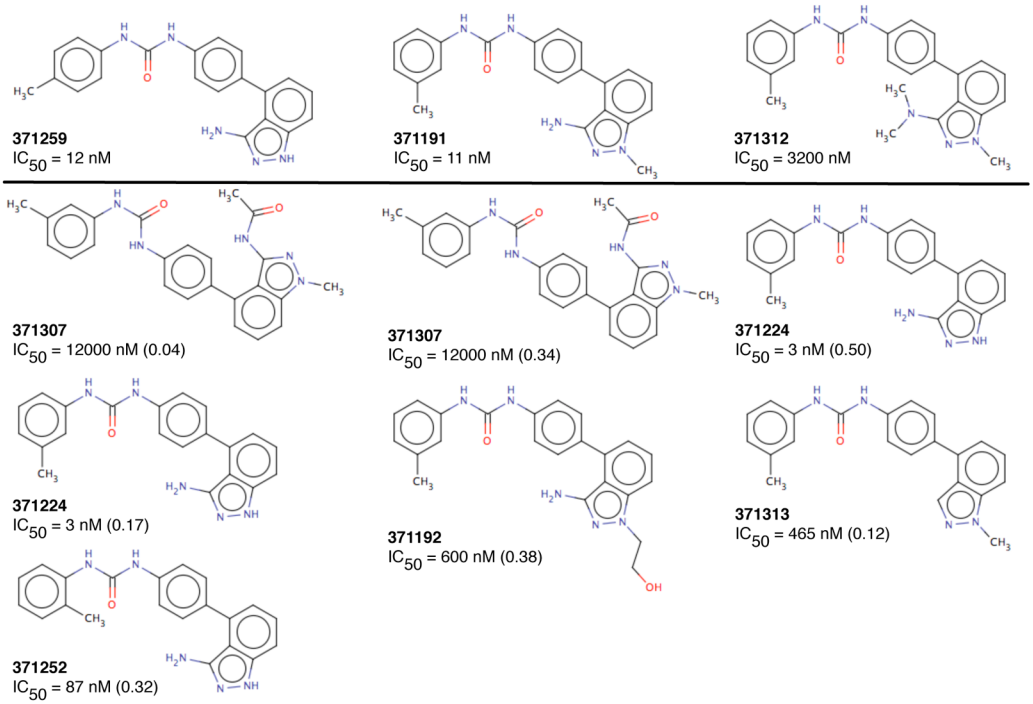

Figure 10.

Three of the hold out molecules for the Dai dataset and training set members with which the hold outs exhibit predicted activity cliffs. Bold numbers are ChEMBL MOLREGNO values and numbers in parentheses are the absolute prediction residual in log(SALI) units.

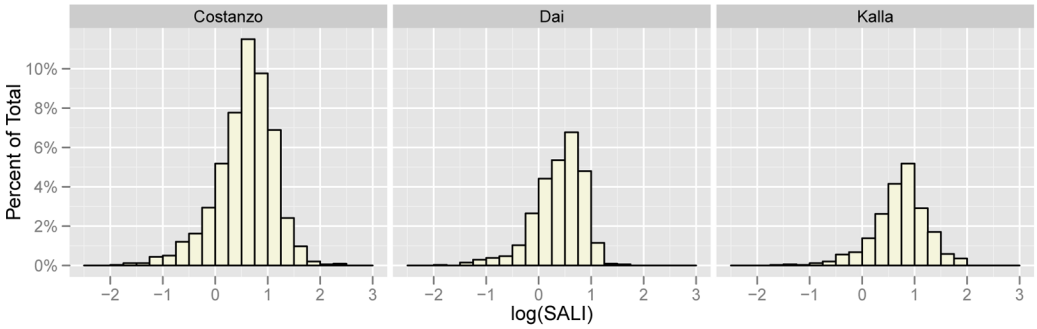

Figure 11.
Distribution of observed log(SALI) values for the three ChEMBL datasets.

Table 1

Datasets considered in this study. N_{desc} is the number of descriptors for the dataset after removal of correlated and low-variance descriptors from the original pool of 109 descriptors.

Name	Description	Endpoint	ChEMBL AID	Size	N_{desc}	Reference
Cavalli	Putative hERG inhibitors (via pharmacophore modeling)	pIC_{50}		30	40	Ref. 15
Costanzo	Thrombin inhibitors	IC_{50}	305283	60	21	Ref. 16
Kalla	A2B adenosine receptor antagonists	K_i	364155	38	10	Ref. 17
Dai	RTK inhibitors	K_i	429056	44	12	Ref. 18

Table 2

Performance metrics for the random forest models built for the datasets used in this study. SD is the standard deviation of the dependent variable.

				Training		Test	
Dataset	**Y-range[a]**	**SD[a]**	**Aggregation Function**	**RMSE**	**R^2**	**RMSE**	**R^2**
Cavalli	7.21	1.57	f_{diff}	1.16	0.54	1.07	0.56
			f_{mean}	1.03	0.64	1.01	0.58
			f_{geom}	1.07	0.61	1.01	0.59
Costanzo	288.02	14.84	f_{diff}	11.99	0.24	17.97	0.10
			f_{mean}	12.88	0.12	16.96	0.20
Kalla	88.25	12.16	f_{diff}	8.12	0.59	6.49	0.57
			f_{mean}	8.79	0.52	8.00	0.47
Dai	32.71	3.48	f_{diff}	2.49	0.48	3.21	0.22
			f_{mean}	2.64	0.41	2.86	0.37

[a]Y-range is calculated using the entire dataset for each study.

J Chem Inf Model. Author manuscript; available in PMC 2012 September 22.

Table 3

Performance metrics for the random forest models built on the single molecule activity data, for each ChEMBL dataset. In each case the model was built using the $\log_{10}$ of the reported activities.

Dataset	Y-range	SD	RMSE	R^2
Costanzo	4.67	1.22	0.80	0.18
Kalla	4.44	0.78	1.00	0.32
Dai	3.60	0.94	0.76	0.34