Exploring Uncharted Territories - Predicting Activty 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 SAR 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. Our results indicate that the predictive models are able to prioritize molecules in terms of their ability to activity cliffs and can be used as a way to extend an observed activity landscape to identify molecules that could lead to significant improvements in activity. Conversely, the method can be used to deprioritize molecules that are predicted to show an activity cliff with a potent molecule already in the dataset.

1 Introduction

The landscape paradigm for 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 activties) 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

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

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

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

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 multitarget 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 Cavalli dataset was employed in our previous studies consisted of 30 molecules studied by Cavalli et al[15] as possible hErg inhibitors using a pharmacophore modeling approach. The remaining datasets were obtained from ChEMBL. All three assays involved 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 pre-existing 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 from 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 or the CDK 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.[19] This was

motivated by the fact that such models can be proven to not overfit 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 SARs are linear. A random forest model, being an algorithmic approach[20] (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,[21] using the defult settings.

In this work, we employed the CDK to evaluate 109 2D and constitutional descriptors for each individual 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 size of the final descriptor pools is summarized in Table 1.

One feature of this approach is that one can build relatively robust models even 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?

One question that arises from the preceding discussion is the possibility that we are duplicating information between the dependent and independent variables. More specifically, both the dependent variable (i.e., the SALI values) and the aggregated 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 indepdent and dependent variables (the former being based on multiple atom and bond features and the latter derived from purel 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 the problem of including correlated structural information simultaneously in the dependent and independent variables is minimal.

4 Results

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

tions 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, though f_{mean} appears to lead to the best performance.

While the models are certainly not the most accurate models, the RMSE's and R^2 values are reasonable. 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 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.

We observed that the use of the geometric mean did not lead to models that were any different from those generated using the arithmetic mean as the aggregation function. This was also true of the other datasets and so the following discussion ommits the results obtained when using f_{geom}. Fig. 3 summarizes the models built on the ChEMBL datasets using f_{diff} and f_{mean} as the aggregation functions. Table 2 reports the model statistics. While the differences in model performance do not vary significantly with the aggregation function used, we see that for the Costanzo and Kalla datasets, f_{diff} leads to a slightly better model than when using f_{mean}, though for the Dai dataset, the difference is negligable.

Given the relatively small differences between models based on different aggregation functions, we arbitrarily chose f_{mean} as the aggregagtion functin for subsequent analyses. Figs. 4, 5 and 6 summarize the performance of the models built using the reported activity data and the pairwise SALI values for the Costanzo, Kalla and Froimowitz datasets, respectively. 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 the higher end of the SALI values. However in all three cases, the most significant cliffs exhibit significant variance. In comparison, the models developed on the reported activity values are relatively poor. This can be partially ascribed to the small sizes of the dataset.

4.1 Extending a landscape

Clearly, modeling the pairwise SALI values leads to reasonable models the datasets discussed here. However, the primary goal of this approach is to extend the the structure-activity landscape for a set of molecules. In other words, can one use the model to identify compounds that will exhibit an activity cliff with one or more molecules in the dataset. Given a new molecule, one can evaluate the

descriptors for this molecule and then evaluate the pairwise descriptor values with each member of the training set. Thus, given a new molecule, we must obtain predictions for n "pairwise objects", where n is the number of molecules in the training set. Given these n predictions one can proceed to identify training set molecules with which the new molecule might exhibit a cliff.

We first performed this analysis with the Kalla dataset. Since predictions of new molecules involve n predictions per molecule, we removed three molecules from the dataset and rebuilt the random forest model. The hold out molecules, were specifically selected since they displayed significant activity cliffs with various members of the training set. We then evaluted the pairwise descriptor values for these three with the training set molecules and obtained predictions of the log(SALI) values. Fig. 7A 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 he prediction set degrades somewhat (RMSE of 0.51 versus 0.41 for the training set). As noted above, the variance of the prediction is higher at the lower scale of log(SALI) values. However, at the higher end of the scale there a number of pairs that are well predicted. Yet, it is clear that a 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 comletely informative. Specifically, we are interested in the predictions for the actual activity cliffs, where both the difference in activity and the structural similarity is very high. Fig. 7B visualizes this information. Each point coresponds 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.

Fig. 8 displays the structures of the 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 absoute 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, Fig. 7C makes it clear that the predictions for this molecule were, on average, worse than for the other two molecules.

Fig. 9 summarizes the results for the same set of analyzes described above, applied to the Dai dataset. As before we removed three molecules (top row of Fig. 10) and rebuilt the random forest on the paired cases with the remaining 41 molecules. We then predicted the pairwise SALI values for the three hold-out molecules with the remaining 41 molecules (i.e., the prediction set). A plot of the predicted versus observed log(SALI) values for the training and prediction sets are

displayed in Fig. 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). 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 (Fig. 9C) are reatively 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 Fig. 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 the three 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 poredicted 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 significant ranging from 1.1-fold to 13-fold. Two such examples are shown in Fig. 10, where the only difference is in the position of the methyl substituent. 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. Thus while we have selected a few ChEMBL assays for the purposes of highlighting this approach, we have found that for some assays, the model built on pairwise data performs worse than the model built on the original data. Given that most of the datasets are small in size (less than 40 molecules), the poor performance of the initial model is not always surprising. 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 (Fig. 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 selecion

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 highight the modeling procedure and we believe that the random forest allows us to highight the utility of this approach without sacrificing too much by way of predictive performance.

It is important to note that while the modeling approach here focuses on predicting SALI values rather than actual activities, one can obtain an idea of the latter based on the (known) activity of the training set molecules. Thus for a new molecule, predicted to show a cliff with a member of the training set, we can determine whether this molecule might be the potent member of the cliff, by noting whether the corresponding training set molecule is less or more potent. Given that one is usually interested in activity cliffs because of the possibility of increasing potency, this observation implies that for a collection of new molecules, one can prioritize them as being predicted to be more or less potent. From this point of view, the approach described here is more suited for prioritization, rather than prediction of absolute values.

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-exisiting 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 modls 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, G. M. On Outliers and Activity Cliffs–Why QSAR Often Disappoints.. *J. Chem. Inf. Model.* **2006,** *46,* 1535–1535.

[3] Leach, A.; Jones, H.; Cosgrove, D.; Kenny, P.; Ruston, L.; MacFaul, P.; Wood, J.; Colclough, 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.

[4] Shanmugasundaram, V.; Maggiora, G. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. In *CINF-032. 222nd ACS National Meeting, Chicago, IL, United States*; American Chemical Society: Washington, D.C., 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.

[6] Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007,** *50,* 5571–5578.

[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.

[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.

[9] Dimova, D.; Wawer, M.; Wassermann, A. M.; Bajorath, J. Design of Multitarget Activity Landscapes That Capture Hierarchical Activity Cliff Distributions. *J. Chem. Inf. Model.* **2011,** *51,* 258–266.

[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.

[11] Hu, Y.; Bajorath, J. Molecular Scaffolds with High Propensity to Form Multi-Target Activity Cliffs. *J. Chem. Inf. Model.* **2010,** *50,* 500–510.

[12] Wassermann, A. M.; Bajorath, J. Chemical Substitutions that Introduce Activity Cliffs Across Different Compound Classes and Biological Targets. *J. Chem. Inf. Model.* **2010,** *50,* 1248-1256.

[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.

[14] Scheiber, J.; Jenkins, J. L.; 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.

[16] Costanzo, M. J.; Almond, Jr, H. R.; Hecker, L. R.; Schott, M. R.; Yabut, S. C.; Zhang, H.-C.; Andrade-Gordon, P.; Corcoran, T. W.; Giardino, E. C.; Kauffman, J. A.; Lewis, J. M.; de Garavilla, L.; Haertlein, B. J.; Maryanoff, B. E. In-depth study of tripeptide-based alpha-ketoheterocycles as inhibitors of thrombin. Effective utilization of the S1' subsite and its implications to structure-based drug design. *J. Med. Chem.* **2005,** *48,* 1984–2008.

[17] Kalla, R. V.; 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.

[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.

[19] Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees;* Chapman & Hall/CRC: Boca Raton, FL, 1984.

[20] Breiman, L. Statistical Modeling: Two Cultures. *Stat. Sci.* **2001,** *16,* 199–231.

[21] R Development Core Team, "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria, 2008 ISBN 3-900051-07-0.

Table 1: Datasets considered in this study.

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

Table 2: Performance metrics for the random forest models built for the datasets used in this study. RMSE and R^2 are based on out-of-bag data[19]

Dataset	Y-range	Aggregation Function	RMSE	R^2
Cavalli	7.21	f_{diff}	1.11	0.57
		f_{mean}	0.98	0.67
		f_{geom}	1.01	0.65
Costanzo	288.03	f_{diff}	12.88	0.24
		f_{mean}	13.56	0.17
Kalla	88.25	f_{diff}	7.51	0.62
		f_{mean}	8.26	0.53
Dai	32.71	f_{diff}	2.60	0.44
		f_{mean}	2.67	0.41

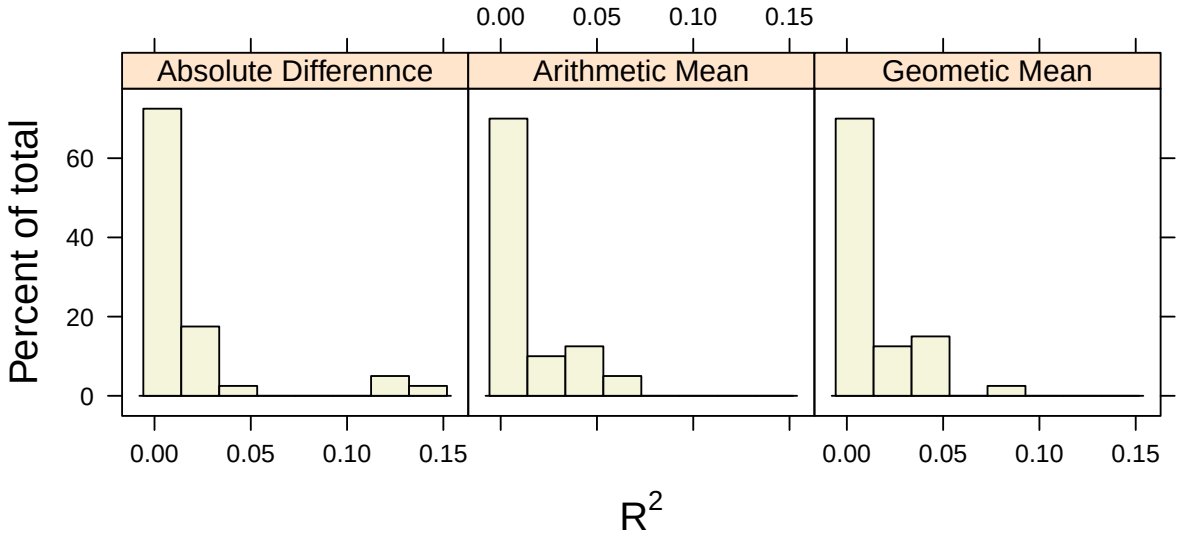


Figure 1: Distribution of Pearson correlations between SALI values and the descriptors for the Cavalli dataset. Three descriptor aggregation functions are considered.

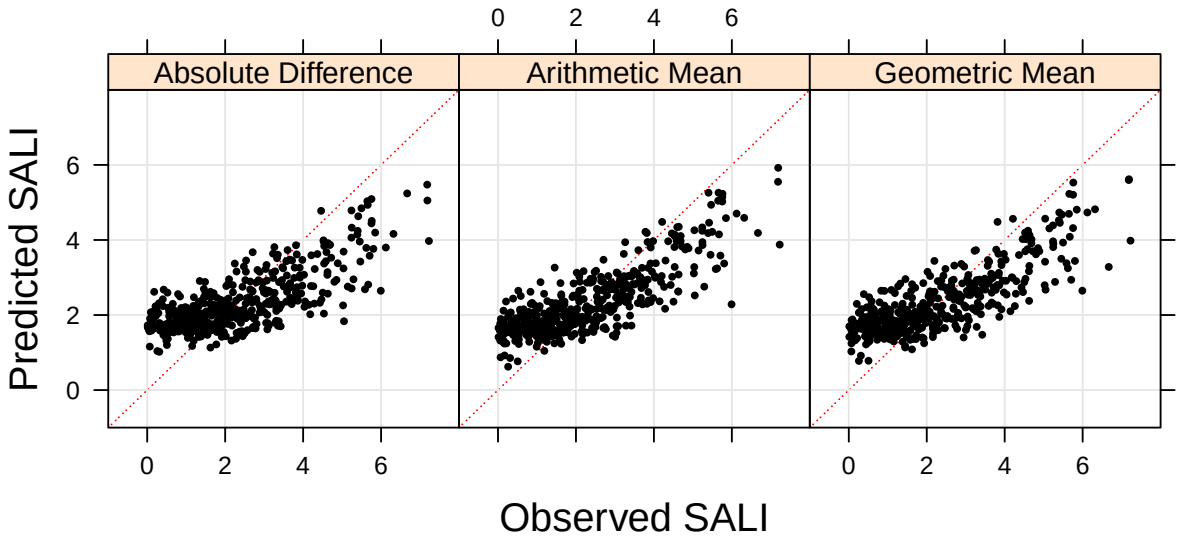
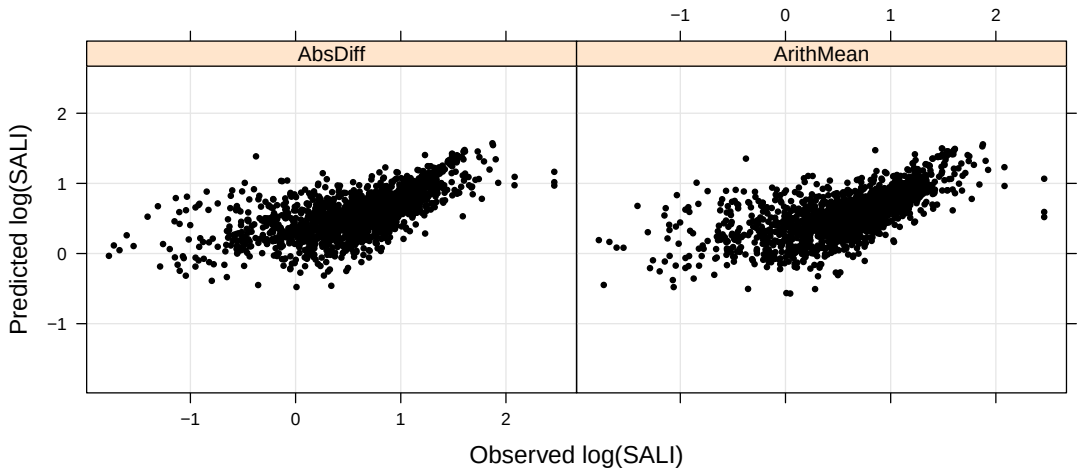
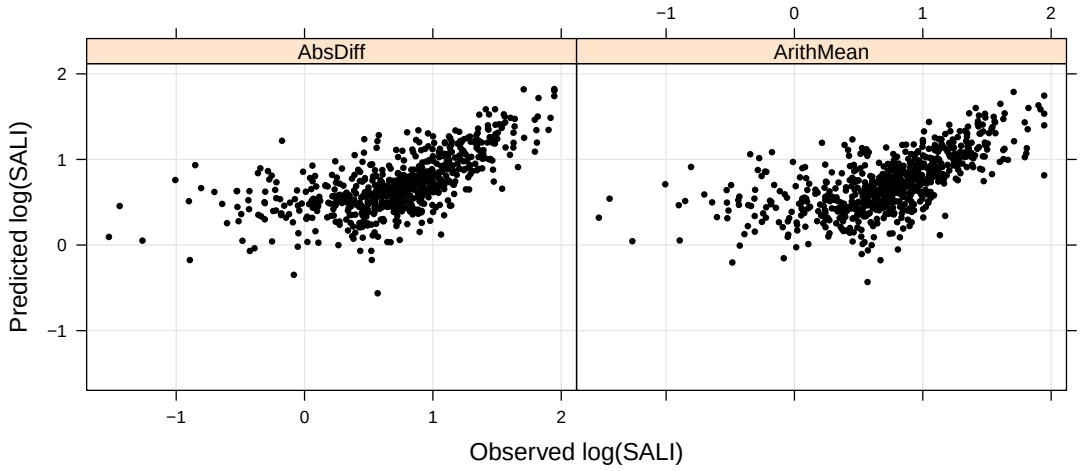


Figure 2: Plots of predicted versus observed SALI values obtained using random forest models, on the Cavalli dataset. Each panel corresponds to the use of a different descriptor aggregation function.

Costanzo



Kalla



Dai

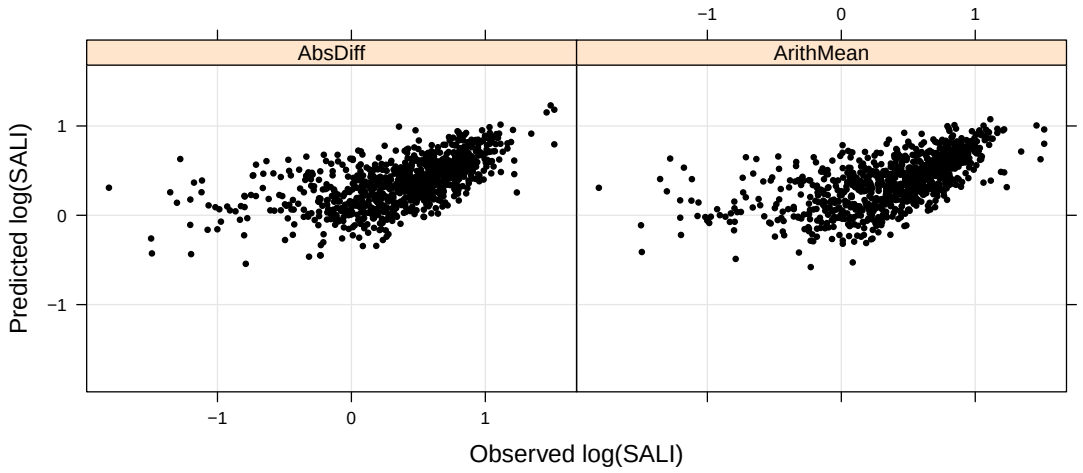


Figure 3: Predicted versus observed SALI values, obtained from random forest models for the three ChEMBL datasets. The three plots correspond to the two different aggregation functions (f_{diff} and f_{mean} respectively).

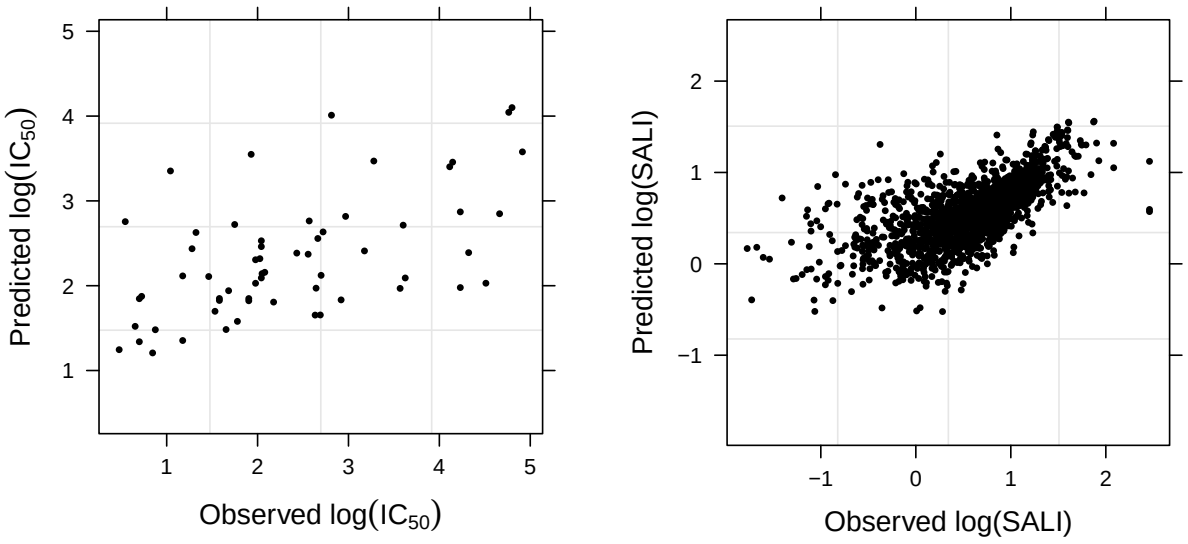


Figure 4: Results of random forest models developed using the Costanzo dataset. The left plot shows the predicted versus observed plot for the model using the original IC_{50} values, and the right hand plot shows the predicted versus observed log(SALI) values.

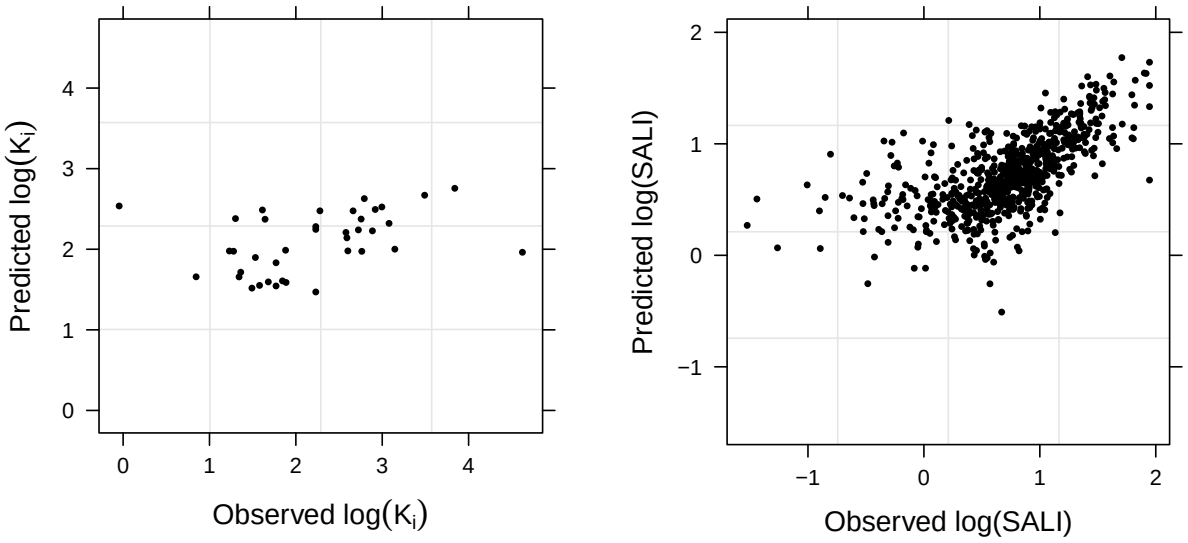


Figure 5: Results of random forest models developed using the Kalla dataset. The left plot shows the predicted versus observed plot for the model using the original K_i values, and the right hand plot shows the predicted versus observed log(SALI) values.

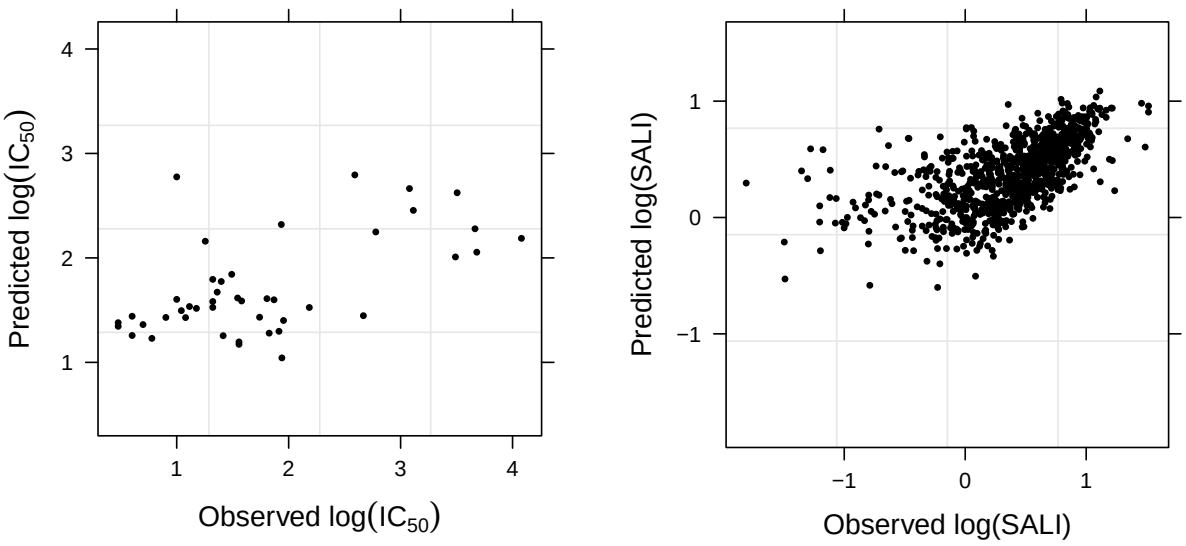


Figure 6: Results of random forest models developed using the Dai dataset. The left plot shows the predicted versus observed plot for the model using the original K_i values, and the right hand plot shows the predicted versus observed log(SALI) values.

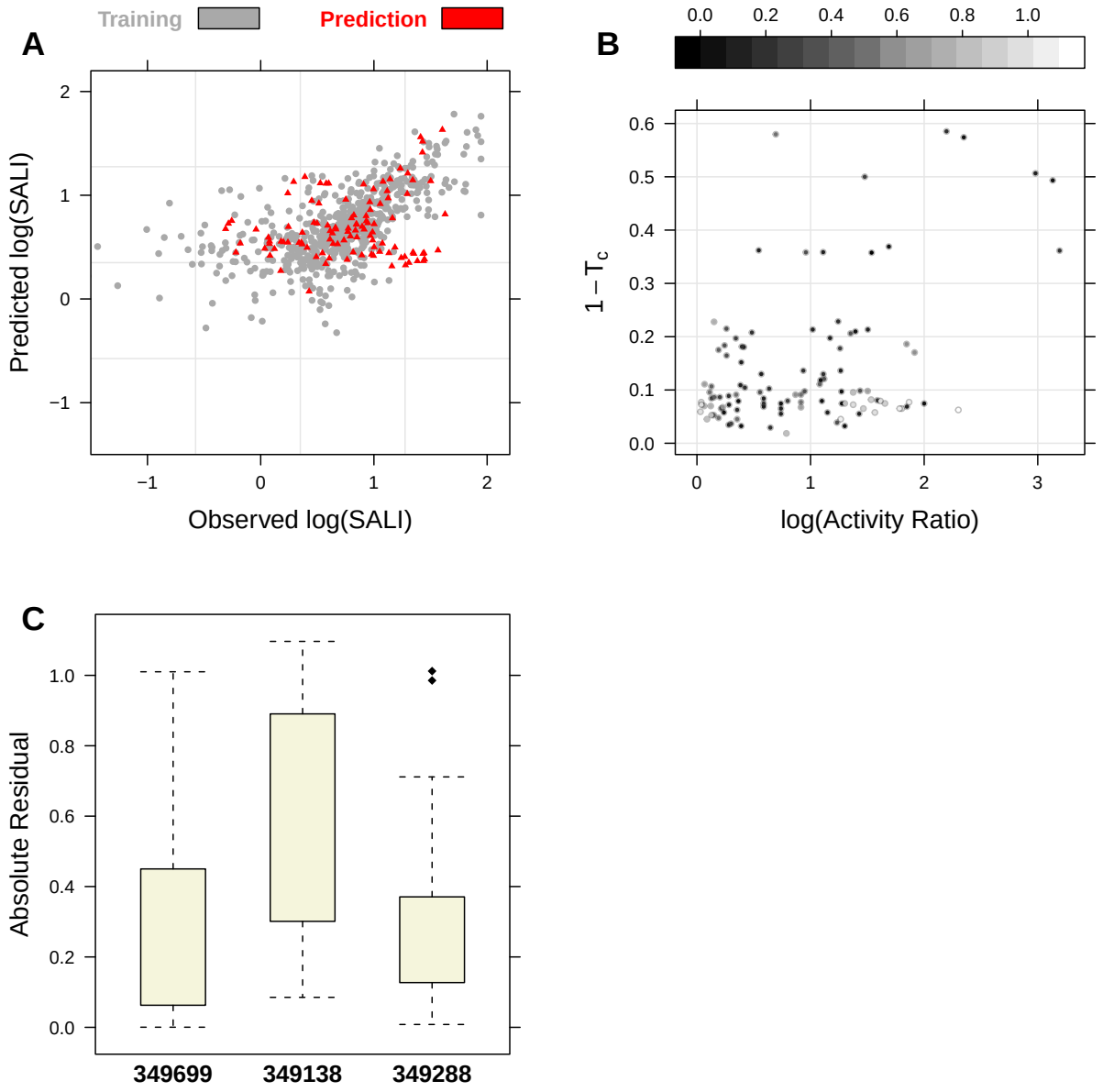


Figure 7: Detailed analysis of SALI predictions for the Kalla dataset. **A** - a plot of predicted versus observed log(SALI) values for the training set and the hold out set. **B** - a summary of the training set, where we plot the structural difference versus the logarithm of the ratio of the activities 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.

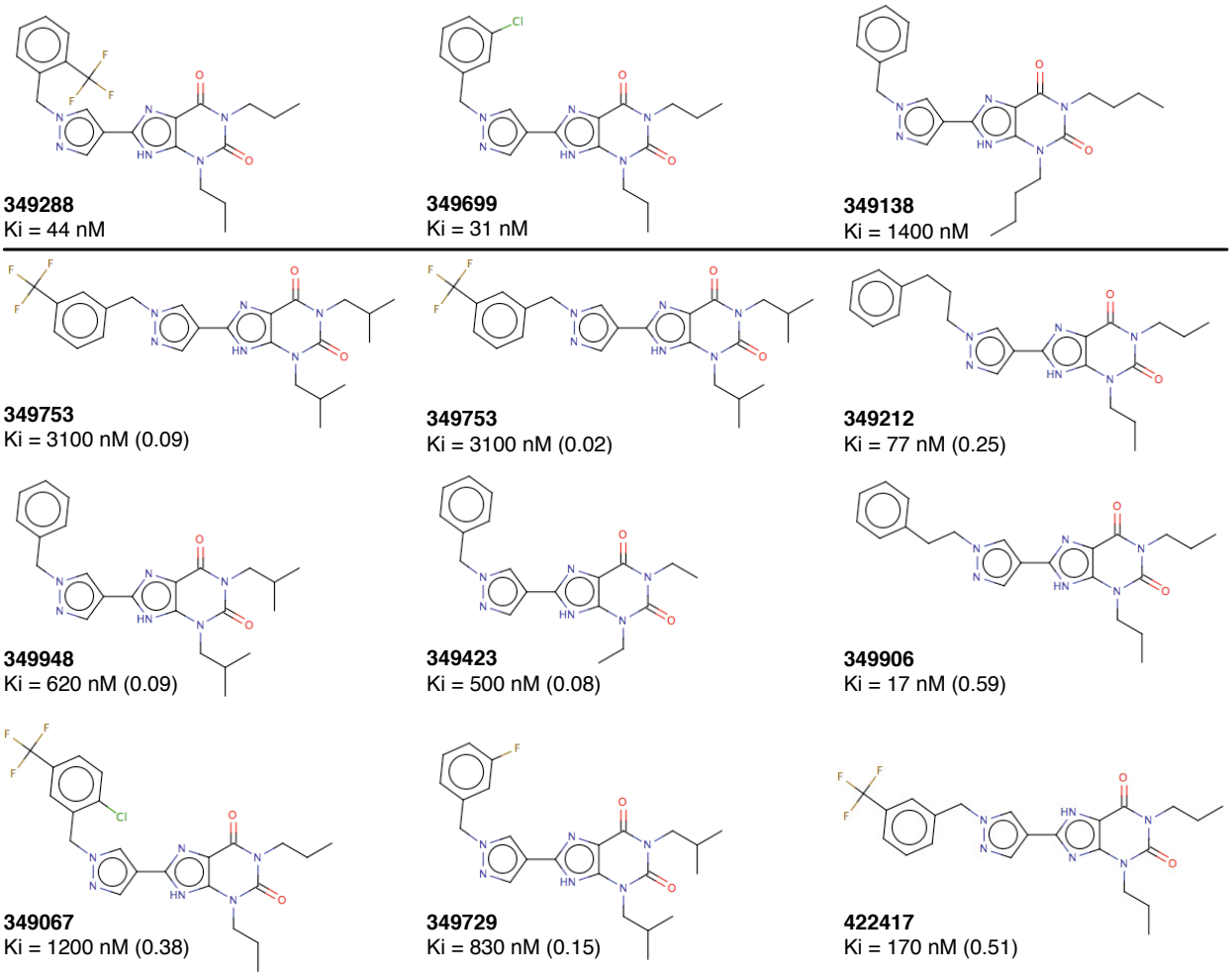


Figure 8: The hold out molecules for the Kalla dataset and training set members with which the hold outs exhibit predicted activity cliffs. Bold numbers are ChEMBL MOLREGNO values and numbers in parentheses are the absolute prediction residual in log(SALI) units.

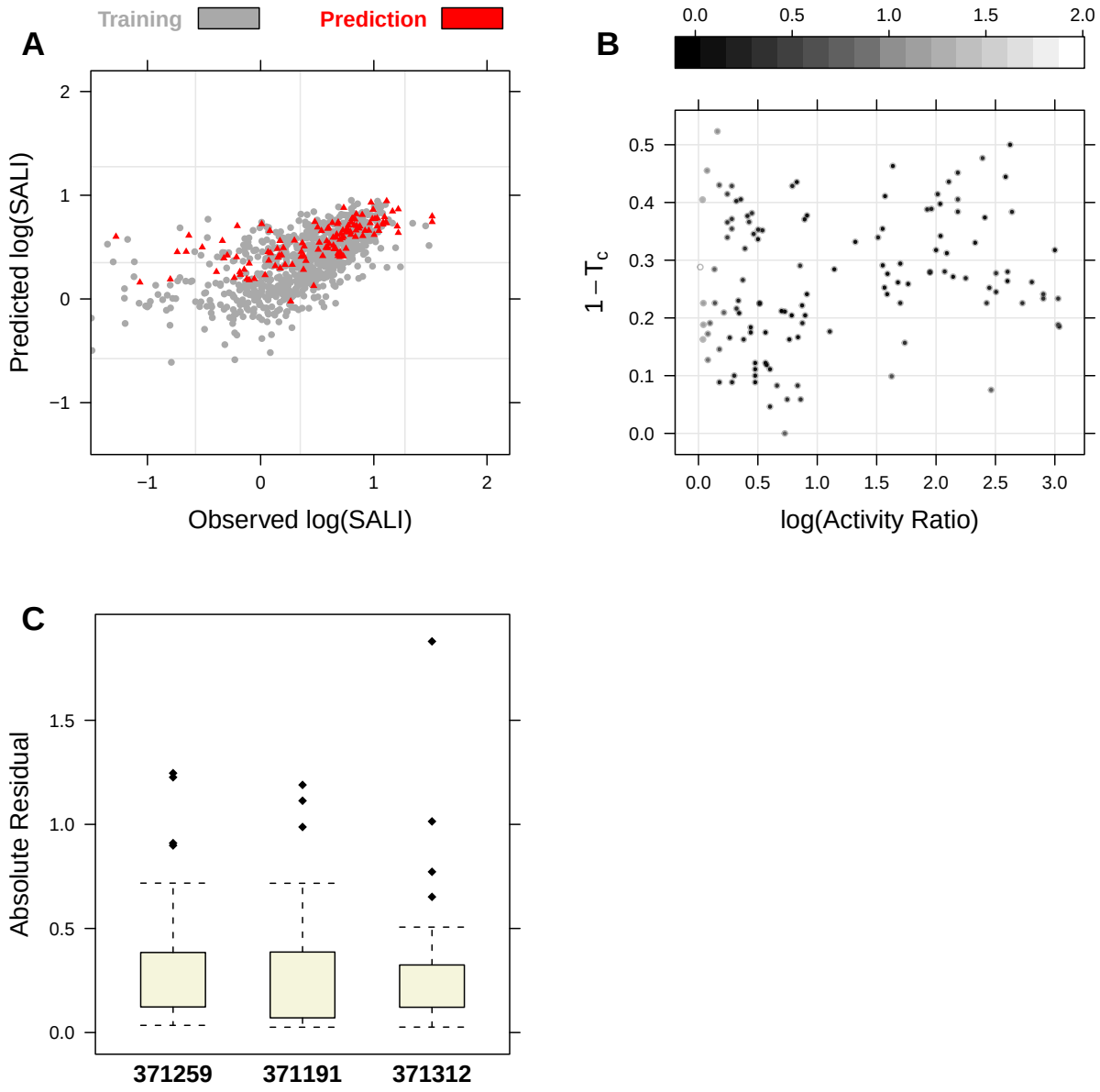


Figure 9: Detailed analysis of SALI predictions for the Dai dataset. **A** - a plot of predicted versus observed log(SALI) values for the training set and the hold out set. **B** - a summary of the training set, where we plot the structural difference versus the logarithm of the ratio of the activities 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.

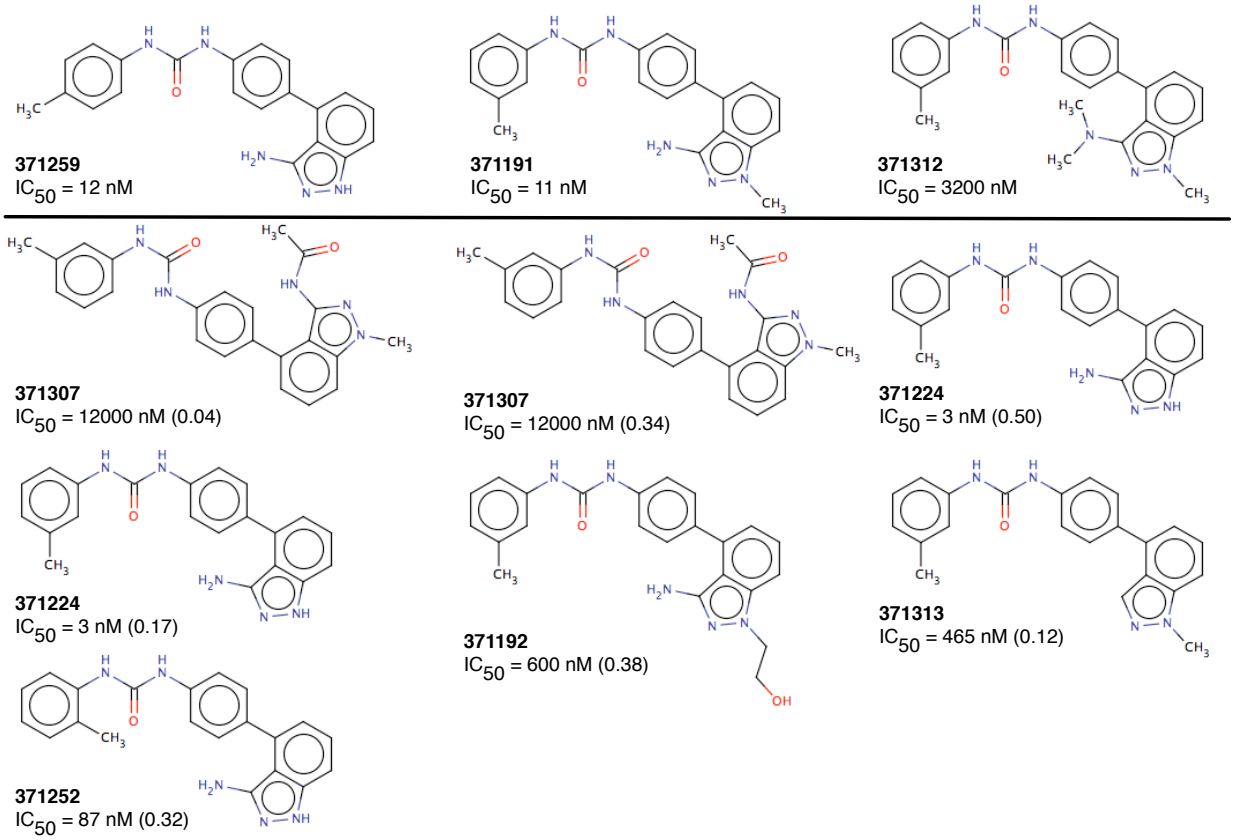


Figure 10: The hold out molecules for the Dai dataset and training set members with which the hold outs exhibit predicted activity cliffs. Bold numbers are ChEMBL MOLREGNO values and numbers in parentheses are the absolute prediction residual in log(SALI) units.

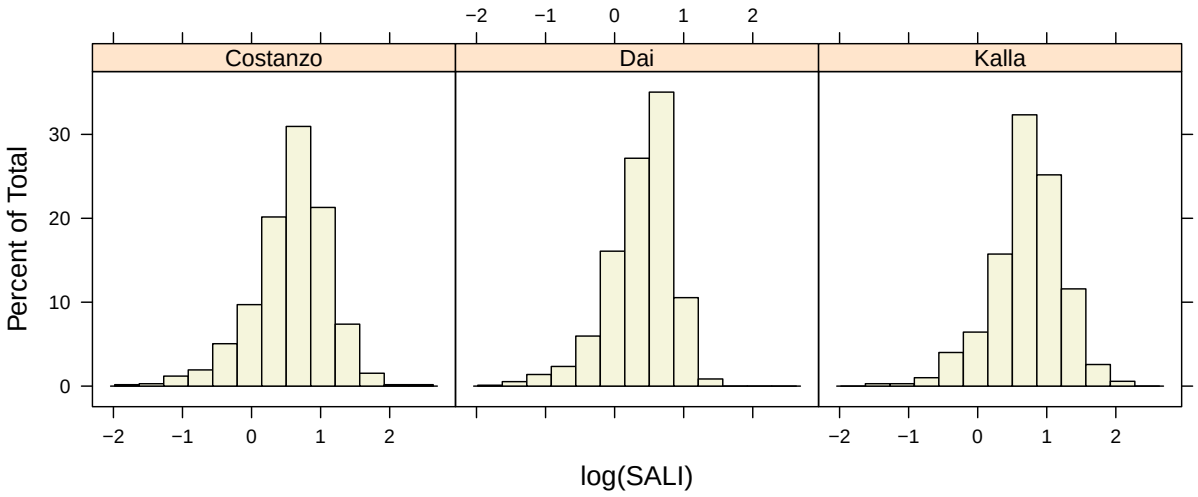


Figure 11: Distribution of log(SALI) values for the three ChEMBL datasets.