Activity Cliffs As Protein-Related Phenomenon: Investigation Using Machine Learning Against Numerous Protein Kinases

Safa Daoud
Applied Sciences Private University
Mutasem Taha (mutasem@ju.edu.jo)
Faculty of Pharmacy, University of Jordan https://orcid.org/0000-0002-4453-072X

Research Article

Keywords: Activity Cliffs, Machine Learning, Protein Kinases, Protein Descriptors.

Posted Date: January 4th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1120840/v1

Abstract

Activity cliffs (ACs) are analogous compounds of significant affinity discrepancies against certain biotarget. We propose that the ACs phenomenon is protein-related and that the propensity of certain target to have ACs can be predicted by some intrinsic protein properties. We pursued this assumption by collecting the crystallographic structures of 84 protein kinases, each of which has numerous reported inhibitors (hundreds). Following data augmentation using synthetic minority oversampling technique (SMOTE), we attempted to correlate the presence/absence of ACs within the ligand pools of collected protein kinases with their corresponding protein properties using genetic algorithm (GA) coupled with variety of machine learners (MLs). Very good GA-ML models were achieved with accuracies of around 75% against external testing set. The models were further validated by Y-scrambling. Shapely additive explanations highlighted the significance of protein rotatable bonds, hydrophobic and acidic residues in relation to the presence of ACs. These results support the hypothesis that ACs are protein-related.

Introduction

Activity cliffs (ACs) are defined as pairs of closely analogous compounds that exhibit significant affinity differences against certain biotarget [1]. The widespread occurrence of ACs within SAR data [2] makes it imperative for modern computer-aided drug design and discovery to successfully handle this phenomenon [3–9]. Nevertheless, ACs represent significant hurdles for bioactivity-supervised discovery methods that assume smooth and continuous structure-activity relationships [10].

ACs are generally explained to exist due to subtle local differences between active and inactive pairs in their ligand-receptor enthalpic 3D contacts [15]. In this context, machine learning methods (e.g., random forests, support vector machine models and particle swarm optimization) were used to identify and predict activity cliff pairs through ligand descriptor patterns [2, 11–13] or target-specific pharmacophore constraints [14]. Additionally, free energy perturbation (FEP) and molecular dynamics were also used to explain ACs [16–20], nonetheless with affinity prediction errors [21].

Upon careful evaluation of a single class of enzymes, namely, protein kinases, we noticed that some kinases showcase many ACs, while others seem to be resistant to this phenomenon despite numerous (hundreds) reported inhibitors. This observation prompted us to propose that the existence of ACs is target protein-related. However, since all protein kinases share analogous ATP catalytic sites, which are normally targeted by designed inhibitors, we propose that the propensity of having ACs is related to the whole protein matrix rather than the binding site only.

To pursue this assumption, we focused on protein kinases of high-resolution crystallographic structures and numerous reported inhibitors. We then determined the number of ACs within the inhibitors' population of each protein. Subsequently, we scanned numerous machine learners (MLs) to assess the possibility of linking protein properties with the existence and/or absence of ACs within the corresponding ligands population. Moreover, we applied genetic algorithm [22] to identify the most

probable protein descriptors that control the ACs phenomenon. We subsequently evaluated the relative influence of significant protein descriptors using Shapely additive explanations [23].

Methods

Data Collection

The crystallographic structures of protein kinases (kinase domains) were collected from the protein databank. Wild type proteins were given priority over mutants. In case that certain protein is represented by several entries in protein databank, we opted for the one that combines the best possible resolution and longest complete (uncut) peptide chain. Subsequently, ChEMBL database was queried to collect inhibitors for each protein kinase in the list. However, only inhibitors that have their bioactivities reported in Ki format were collected for subsequent processing to minimize errors resulting from bioassay discrepancies among different labs. Duplicate structures and compounds that have their bioactivities reported as being more or less than certain values were discarded. This left us with 84 proteins, each associated with list of inhibitors ranging from 393 to 2514.

AC pairs were identified in each pool of inhibitors using the “Find Activity Cliffs” protocol implemented in Discovery Studio 4.5 (Biovia Inc., USA). This protocol performs Matched Molecular Pairs (MMPs) transformations to find activity cliffs [24, 25]. MMPs are pairs of ligands that differ by a single localized structural change. MMPs with activity difference threshold exceeding 100 folds (i.e., 2 log cycles) were considered ACs provided that the active AC member is of Ki value $\leq$ 100 nM. Table 1 shows the selected proteins, their resolutions, count of collected of inhibitors, and count of AC pairs. The actual data (protein structures and corresponding ligands) is found in the supplementary materials.

Protein Descriptors

Each protein (kinase domain) was downloaded from the protein databank. Hydration water and co-crystallized ligands were removed. Hydrogen atoms were added utilizing the Discovery Studio 4.5 hydrogen atoms template. The following descriptors were determined for each protein using Discovery Studio 4.5 working environment: Count of non-hydrogen atoms, count of intramolecular hydrogen bonds (allowing hydrogen-bonds to exist over a maximum distance of 3.5 Å); count of intramolecular π-π stacking interactions; count of intramolecular bumps (pair of atoms at close proximity such that they violate each other VDW spheres by at least 30% without being covalently bonded); count of disulfide linkages; and count of rotatable bonds. The normalized versions of these descriptors (calculated by dividing each descriptor by the count of non-hydrogen atoms within the particular protein) were also included. Additionally, the counts of each individual amino acid (i.e., Ala, Arg, Asn, Asp, Cys, Gln, Glu, Gly, His, Ile, Leu, Lys, Met, Phe, Pro, Ser, Thr, Trp, Tyr, and Val), count of basic amino acids (summation of Lys, Arg and His residues), count of acidic amino acids (summation of Glu and Asp residues); count of hydrophilic amino acids (summation of Arg, Asn, Asp, Gln, Glu, His and Lys residues); count hydrophobic

amino acids (summation of Ala, Cys, Ile, Leu, Met, Phe, Val) together with their normalized versions (calculated by division by the total count of amino acids of each protein) were also included as protein descriptors.

Machine Learning

For subsequent ML scanning, the protein descriptors were employed as explanatory variables, while the response variable was defined as binary code based on the presence or absence of ACs, such that if the list of inhibitors for a particular protein is devoid of ACs, it is given response label of “Zero”, while the presence of at least one pair of ACs warranted a response label of “One”.

The collected kinase data (Table 1) was divided into training and testing sets by ascending ranking of the collected proteins according to their count of intramolecular hydrogen bonds (arbitrary), then each fifth protein was moved from the list to the testing set (16 proteins, marked with asterisks in table 1 including 6 devoid-of-ACs and 10 showing-ACs). The remaining proteins were retained as training set (68 proteins including 26 devoid-of-ACs and 42 showing-ACs).

However, since the number of collected protein kinases is inadequate for efficient ML modelling we decided to augment the training and testing lists, separately, using Synthetic Minority Oversampling Technique (SMOTE) [26] as implemented in the SMOTE KNIME (version 4.3.3) node. The algorithm works by creating synthetic rows by extrapolating between a real object of a given class and one of its nearest neighbors (of the same class). It then picks a point along the line between these two objects and determines the attributes (cell values) of the new object based on this randomly chosen point. We implemented the following parameters: Select a maximum of 5 neighbors and oversampling by 3 folds. The resulting SMOTE-enhanced training list includes 272 observations (rows), of which 104 are labeled as “devoid of ACs” while 168 are labeled as “showing ACs”. Similarly, the SMOTE-enhanced testing set includes 64 observations, of which 24 are labeled as “devoid of ACs” and 40 as “showing ACs”.

Screening Machine Learners (MLs)

Initially, 10 MLs were scanned using all protein descriptors as explanatory variables, while the response variable was set to the presence or absence of ACs. The screened learners were: K-Star [27], Locally weighted learning (LWL) [28, 29], Logistic Model Trees (LMT) [30, 31], LogitBoost [32], Support Vector Machine (SVM) [33, 34], Naïve Bayes [35], Random Forests [36], Probabilistic Neural Network (PNN) [37, 38], Xgboost [39] and k-nearest neighbors (k-NN) [40]. The MLs were implemented as corresponding KNIME 4.3.3 nodes with default parameters. Each ML was evaluated based on its accuracy (Equation 1) [41–45] and Cohen’s Kappa (κ, Equation 2) [46] values in classifying the training set into “One” (i.e., showing ACs) or “Zero” (i.e., devoid of ACs) implementing the leave-20%-out cross-validation.

$$Accuracy = \frac{TP + TN}{N} \ldots\ldots\ldots\ldots (1)$$

$$\kappa = \frac{P_0 + P_e}{1 - P_e} \ldots\ldots\ldots\ldots (2)$$

where N is the number of all observations in the training list, TP and TN are the numbers of truly identified proteins as “showing ACs” and “devoid of ACs”, respectively, by the particular ML. P_o is the relative observed accuracy (i.e., agreement among raters), and P_e is the probability of chance agreement (hypothetical). This is done by using the observed data to calculate the probabilities of each observer randomly seeing each category. If the raters are in complete agreement, then κ = 1. If there is no agreement among the raters other than what would be expected by chance (as given by P_e), κ = 0. Negative Cohen’s kappa value implies the agreement is worse than random [46].

Genetic algorithm (GA) for descriptor selection

High-ranking MLs were subsequently individually combined with GA to select optimal groups of protein descriptors capable of achieving best Cohen’s Kappa values in classifying the training set implementing leave-20%-out cross-validation. The GA workflow within KNIME Analytics Platform (Version 4.3.3) was implemented for this purpose.

GA operates through a cycle of four stages [22]: (i) Encoding mechanism; (ii) Definition of a fitness function; (iii) Creating a population of chromosomes; (iv) Chromosomes Genetic manipulation. A gene-based encoding system is implemented herein, whereby the presence or absences of a certain descriptor(s) in a particular suggested model is encoded by gene (chromosome) format, i.e., each value in the gene string represents an independent variable (0 = absent, 1 = present). Each chromosome is associated with a fitness value that reflects how good it is compared to other solutions. Genetic manipulation involves mating among successful chromosomes and mutation of some genes within randomly selected chromosomes to help exit local minima. GA control parameters that need to be configured prior to modeling are: the population of initial random chromosomes, the maximum number of generations to exit from a basic cycle and fitness criterion [22]. In the current project, these were configured as follows: population size = 500, maximum number of generations = 10000 and the fitness criterion was set to Cohen’s Kappa value of the ML model resulting from features selected by each genetic chromosome.

Shapley Additive Explanations (SHAP)

Shapley additive explanation (SHAP) of a particular feature (protein descriptor) indicates how much the feature has contributed to the deviation of the prediction of certain observation from the base prediction (i.e., the mean prediction over the full sampling data) [23]. The SHAP node within KNIME Analytics

Platform (Version 4.3.3) was implemented using K-means to summarize the data of the evaluated feature before forming coalition among remaining features.

Results

Data Collection

In order to correlate proteins' properties with propensities of having ACs, it was necessary to gather accurate ligand binding data for the collected protein kinases. Duplicate molecules and molecules of inaccurately reported structural information (e.g., undefined stereochemistry) were removed. Additionally, we limited the collected inhibitors to molecules that have their bioactivities expressed as Ki values to minimize the effects of inter-laboratory assay variations commonly seen with other bioactivity indicators (e.g., IC_{50}) [47]. Moreover, ACs were strictly defined as MMPs of bioactivity difference exceeding 100 folds provided that potent ACs members are of Ki values below 100 nM. On the proteins' side; only wide type protein kinases were considered, and if certain protein is represented by several entries in protein data bank, we opted for the one that combines best possible resolution and longest complete (uncut) peptide chain.

However, since the presence/absence of ACs can be a function of the explored chemical space of the particular protein kinase, it can be argued that absence of ACs for certain protein kinase is related to the limited medicinal chemistry efforts against this target rather than due to certain intrinsic factors related to the protein kinase itself. We adopted two measures to remedy this issue. Firstly, we only collected protein kinases of substantial number of reported ligands (Table 1). However, to propose a reasonable minimal number of ligands for a particular kinase to be included in the modeled list (i.e., Table 1) we looked at protein kinases that combine the highest counts of reported inhibitors and least number of ACs. Two kinases met these criteria, namely, LCK (1818 inhibitors and 3 ACs) and FYN (1643 inhibitors and 6 ACs) with ACs-to-ligands ratios of 1 in 606 and 1 in 274, respectively. Their average ratio is 1 in *ca.* 440. Accordingly, based on this ratio, it can be reasonably assumed that if the binding space of a particular protein kinase has been explored by $\geq$ 440 ligands without identifying any ACs then it is likely that the inhibitors space of this target is free from ACs. Based on this plausible assumption we collected protein kinases of at least 440 reported inhibitors. Still, we noticed 5 protein kinases of inhibitors' counts ranging from 393 to 437 (i.e., FES, BRK, ACK1, PYK2, and AXL, table 1), two of which have established ACs (BRK and ACK1) and thus can be safely included for ML. However, the other three are devoid of reported ACs, still we included them for ML because they have inhibitors' counts close to the proposed threshold (i.e., 440 inhibitors), i.e., PYK2 (428 inhibitors), AXL (437 inhibitors) and FES (393 inhibitors). Additionally, these are rather limited in number (just three cases) and therefore should have limited error contribution in ML (if any).

The second measure we took to minimize errors related to limited medicinal chemistry efforts is to squash ACs counts of each protein kinase into a binary response of either "having ACs" (given binary code of one) or "devoid of ACs" (given binary code of zero). This is done to emphasize that the mere

presence of even a single valid AC indicates that protein is vulnerable to other ACs and that counts of reported ACs are irrelevant (in fact all our attempts to correlate the counts of ACs with protein properties were futile). However, although current absence of ACs for certain target does not guarantee that there will be no ACs upon future medicinal chemistry exploration of this target, still, targets labeled as being "devoid of ACs" with ligands counts exceeding 440 can be reasonably considered as being resistant to the ACs phenomenon. Accordingly, the binary label of being "showing ACs" or "devoid of ACs" is really surrogate for "AC-vulnerable" or "AC-resistant" targets. This conclusion is underlined in figure 1.

Figure 1 shows the counts of "devoid of AC" and "showing AC" protein kinases as function of the counts of their reported inhibitors in ChEMBL database. Clearly from the graph the "ACs-devoid" category supersedes the "AC-showing" category in the first two intervals, i.e., 390-700. However, protein kinases of higher ligands' counts tend to incline towards the "AC-showing" class despite the presence of significant "AC-devoid" minority. In the highest ligands' counts interval, i.e., 1301-2514, all 8 collected protein kinases showed ACs among their ligands. Nonetheless, even in this interval, two protein kinases were rather resistant to ACs, i.e., LCK (3 ACs out of 1818 inhibitors reported inhibitors) and FYN (6 ACs out of 1643 inhibitors). Conclusions from figure 1 provide impetus for our proposition that the presence or absence of ACs, within certain protein kinase binders, points to the level of resistance/vulnerability of that target to ACs. However, since it is hard to set certain numerical threshold for the number of ACs to consider a particular protein kinase as being AC-resistant or AC-vulnerable, our use of "devoid of AC" and "showing AC" labels is rather reasonable surrogate of AC resistance/vulnerability.

Machine Learning

To evaluate the relevance of protein descriptors to the propensity of exhibiting ACs, we initially *t*-tested the difference between protein kinases "showing-ACs" versus those "devoid-of-ACs" with respect to calculated protein descriptors. Four descriptors showed significant difference ($p < 0.05$) between the two groups, namely; normalized count of hydrogen bonds (N/H-bonds), normalized count of basic residues (N/basic amino acids), normalized count of hydrophobic residues (N/Hydrophobic amino acids) and normalized count of rotatable bonds (N/Rotatable bonds). Although the number of significantly different descriptors between the two protein kinase categories is rather limited (only four), it should not preclude the possibility of significant differences resulting from interactions between different descriptors warranting subsequent evaluation by MLs.

ML studies commenced by splitting the collected kinases into training and testing sets (as in table 1, testing compounds are marked with asterisks). However, due to the limited number of protein kinases, i.e., for effective machine learning, we opted to augment the training and testing lists, separately, using Synthetic Minority Oversampling Technique (SMOTE) [26]. However, SMOTE was used to augment both classes "Showing-ACs" or "Devoid-of-ACs" (i.e., not only the minority class) to yield an overall 272 training observations of which 104 are labeled as "devoid of ACs" while 168 are labeled as "showing ACs". The SMOTE-enhanced testing set included 64 observations, of which 24 are labeled as "devoid of ACs" and 40 as "showing ACs".

Subsequently, we scanned ten ML algorithms (see experimental section) implementing “Showing-ACs” or “Devoid-of-ACs” as response labels, while all calculated protein descriptors were used as independent variables. However, seven MLs yielded accuracies exceeding 70% (based on leave-20%-cross validation) warranting subsequent combination with GA to identify the best predictive descriptors. Two MLs succeeded in maintaining significant accuracies and Cohen’s Kappa values despite GA-driven feature reduction, namely; K-star (K*) and XGBoost (XGB). Table 2 summarizes the statistical criteria of the best models.

The K-star algorithm (K*) is an instance-based classifier that uses entropy-based distance function [48]. Classification with K* is made by summing the probabilities from the new instance to all of the members of a category [27]. On the other hand, eXtreme Gradient Boosting (XGBoost, or XGB) is a tree-based standardized ensemble method that relies on an ensemble of weak decision tree (DT)-type models to create new subsequent boosted DT-type models of a reduced loss function. XGB system includes a tree learning algorithm, a theoretically justified weighted quantile sketch procedure with parallel and distributed computing [49].

Accuracies and Cohen’ Kappa values in Table 2 suggest that it is possible to predict the propensity of ACs for certain protein kinase by applying few protein descriptors in certain ML model(s). Cohen’s Kappa values of SMOTE-enhanced observations in table 2 position the corresponding ML models within moderate to substantial interrater reliability.

However, to exclude the possibility of chance correlation, we decided to reevaluate the models using the original unaugmented training and testing sets and to test the validity of the models using Y-scrambling [50]. The results are summarized in table 2. Interestingly, both ML models maintained successful statistical criteria even upon using the original unaugmented data.

In Y-scrambling, 100 random bioactivity data were generated based on either the SMOTE-enhanced training set or the original unaugmented training set. Subsequently, each successful learner and corresponding features were challenged to use these random data to generate ML models of equal or better accuracies compared to the original nonrandomized data, as judged by Leave-20%-Out cross-validations [50]. Table 2 summarizes the results of Y-scrambling while supporting Tables S1 to S4 show the results in details. The results show that in both successful MLs, the non-randomized data yielded models of significantly superior accuracy and Cohen’s Kappa values compared to all randomized trials. The effect is particularly evident in Cohen’s Kappa values. Overall, the results support the validity of our ML models.

Table 2 points to interesting facts: (i) Most of the significant descriptors, from both models, are normalized suggesting that protein size is not of significance with regard to ACs. (ii) The two ML models emphasize the significance of two classes of amino acid residues, namely, acidic and hydrophobic amino acids. Acidic amino acids are represented by N/Asp, N/Glu and N/Acidic amino acids, while hydrophobic groups are represented by N/Hydrophobic amino acids and N/Val. (iii) K* ML model uniquely emphasizes

on the significance of N/Rotatable bonds, while the XGB model underlines the influence of hydrogen bonds count on the propensity of having ACs.

Emergence of N/Hydrophobic, N/Rotatable and N/H-bonds agrees with the *t*-test results mentioned earlier. Strangely though, N/basic failed to emerge in any of the top models despite being significantly different between the two protein kinases (Showing and Devoid of-ACs) according to *t*-test. It appears that the interaction between different descriptors in the optimal ML models in table 2 overshadowed the effects of N/basic.

Significance of Individual Features

Although ML-models in table 2 have reasonable accuracies and Cohen's Kappa values, it is hard to infer the role contributed by each protein feature in predicting the class label of a particular testing protein kinase. For example, it is hard to tell how N/Rotatable bonds contributes to the "Showing ACs" class within the context of GA-K* ML model (table 2). Therefore, we decided to implement SHapley Additive exPlanations (SHAP) values to explain the relative contributions of individual descriptors in predicting class labels [23].

SHAP values originated in game theory, however, within the context of machine learning they are useful for explaining model predictions. The SHAP approach enables prioritization of features that determine the predicted classification of certain observation using any ML model. The SHAP value of certain feature for a certain observation indicates how much this feature (percentage wise) has contributed to the deviation of the prediction from the base prediction (i.e., the mean prediction over the full sampling data) for this observation. SHAP of all features should add up to the difference between the mean prediction and the actual prediction for certain testing observation.

Figure 2 shows the average SHAP values for GA-selected descriptors in the optimal K* and XGB models. Each average SHAP was calculated as the mean of SHAP values of the particular descriptor across "showing ACs" or "devoid of ACs" SMOTE-enhanced testing compounds. Clearly from the figure, the selected descriptors exhibited SHAP probability contributions consistent with the class categories of testing observations, i.e., they yielded positive probabilities towards the "Showing ACs" label for protein kinases that have ACs, while they also gave positive probabilities for "Not showing ACs" label for protein kinases that do not show ACs.

It can be concluded from Figure 2 that the most significant contributor to probability predictions in the GA-K* model is the normalized number of rotatable bonds (N/Rotatable bonds). On the other hand, the most significant contributor to probability predictions in the GA-XGB model is the normalized number of aspartic acid moieties (N/Asp), while the normalized number of hydrogen bonds and hydrophobic groups (N/H-Bonds and N/Hydrophobic groups, respectively) come second.

Another interesting conclusion from Figure 2 is the orthogonality of the two models, as can be deduced from the differing probability contributions of descriptors among the two models suggesting the

possibility of stacking the two ML models in a meta-learning model (e.g., consensus voting) [51].

Discussion

Emergence of N/Rotatable bonds, N/H-Bonds, H/Hydrophobic and N/acidic amino acids as major contributors in our successful GA-ML models probably encodes for significant role played by protein flexibility in the propensity of having ACs: Presence of extensive hydrogen bonding, hydrophobic and/or electrostatic networks within certain protein matrix is suggestive of conformational rigidity, while on the other hand, numerous rotatable bonds suggest tendency towards molecular flexibility [52–54].

Reports of some protein receptors to undergo drastic conformational rearrangements upon binding to certain ligand(s) while other analogous ligand(s) fail to cause similar response [55–59] prompted us to assume that this trend is related to our findings, namely, the connection between protein flexibility and the propensity of having ACs.

Interestingly, significant conformational modifications in the receptor usually accompany entropy-driven ligand binding while enthalpy-driven binding involves only slight (or even no) associated receptor conformational modifications [55, 56, 60–62]. For example, the close potent analogues in table 3 [60] exhibit significantly discrete conformational influences on their target protein in such a way that the entropic binder causes significant conformational rearrangement in the target protein, while its enthalpic twin causes only subtle protein modifications. In this context, our findings of connecting protein flexibility with the propensity of having ACs, combined with the fact that activity cliffs have nearly identical 3D binding features lead us to theorize that AC twins differ in that one member, i.e., the more potent one, binds to the receptor and induces significant entropically-driven conformational modifications in the host protein causing additional enthalpic binding interactions. The less potent AC member, on the other hand, is enthalpically-driven binder that fails to induce significant conformational modifications in the protein receptor, and therefore fails to recognize additional hidden binding interactions [63]. This assumption is not without precedence. For example, Table 4 shows the thermodynamics and affinities of activity cliff pair against carbonic anhydrase (isoform CA-XIII). Clearly, the potent twin has entropic and enthalpic advantages over the inactive one [62]. Therefore, the existence of ACs should correlate with protein flexibility, while more rigid receptors should me be immune to the ACs phenomena.

Conclusions

This study and related conclusions are of immense significance in the fields of drug design and discovery because it suggests that it is possible to predict whether a drug discovery endeavor will face significant ACs-related issues or not *a priori*.

Declarations

Acknowledgments

The authors thank the Deanships of Academic Research at The University of Jordan and The Applied Sciences Private University, Amman, Jordan, for their generous funds and supporting this project.

Data and Software Availability

All proteins, their properties and collected ligands are available in the Supporting Information. All methods are comprehensively described in the methods section of the article with detailed workflow descriptions. Third party software (e.g., Discovery Studio and KNIME) versions are clearly identified and all related workflows and associated parameters are described and their websites are provided: DiscoveryStudio (Version 4.5), Biovia Inc. https://www.3dsbiovia.com/, USA; KNIME Analytics Platform (Version 4.1.0), https://www.knime.com/

References

1. Stumpfe D, Hu H, Bajorath J (2020) Advances in exploring activity cliffs. J Comput Aided Mol Des. https://doi.org/10.1007/s10822-020-00315-z
2. Namasivayam V, Iyer P, Bajorath (2013) Prediction of individual compounds forming activity cliffs using emerging chemical patterns. J Chem Inf Model. https://doi.org/10.1021/ci400597d
3. Maggiora GM (2006) On outliers and activity cliffs why QSAR often disappoints. J Chem Inf Model. https://doi.org/10.1021/ci060117s
4. Stumpfe D, Bajorath J (2012) Exploring activity cliffs in medicinal chemistry: miniperspective. J Med Chem. https://doi.org/10.1021/jm201706b
5. Bajorath J (2012) Modeling of activity landscapes for drug discovery. Expert Opin Drug Discov. https://doi.org/10.1517/17460441.2012.679616
6. Peltason L, Bajorath J (2007) SAR index: quantifying the nature of structure– activity relationships. J Med Chem. https://doi.org/10.1021/jm0705713
7. Guha R, Van Drie J (2008) Structure– activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model. https://doi.org/10.1021/ci7004093
8. Vogt M, Huang Y, Bajorath J (2011) From activity cliffs to activity ridges: informative data structures for SAR analysis. J Chem Inf Model. https://doi.org/10.1021/ci2002473
9. Hu Y, Bajorath J (2012) Extending the activity cliff concept: structural categorization of activity cliffs and systematic identification of different types of cliffs in the ChEMBL database. J Chem Inf Model. https://doi.org/10.1021/ci300274c
10. Daoud S, Taha MO (2020) Pharmacophore modeling of JAK1: a target infested with activity-cliffs. J Mol Graph Model. https://doi.org/10.1016/j.jmgm.2020.107615
11. Heikamp K, Hu X, Yan A, Bajorath J (2012) Prediction of activity cliffs using support vector machines. J Chem Inf Model. https://doi.org/10.1021/ci300306a
12. Namasivayam V, Bajorath J (2012) Searching for coordinated activity cliffs using particle swarm optimization. J Chem Inf Model. https://doi.org/10.1021/ci3000503

13. Guha R (2012) Exploring Uncharted Territories: Predicting Activity Cliffs in Structure–Activity Landscapes. J Chem Inf Model. https://doi.org/10.1021/ci300047k
14. Mackey M, Cheeseright T (2020) Identification and analysis of activity cliffs using 3D similarity techniques. ChemRxiv. https://doi.org/10.26434/chemrxiv.12974714.v1
15. Hu Y, Stumpfe D, Bajorath J (2013) Advancing the activity cliff concept. F1000Res. https://doi.org/10.12688/f1000research.2-199.v1
16. Rami Reddy M, Ravikumar Reddy C, S Rathore R *et al* (2014) Free energy calculations to estimate ligand-binding affinities in structure-based drug design. Curr Pharm Des. https://doi.org/10.2174/13816128113199990604
17. Gkeka P, Eleftheratos S, Kolocouris A, Cournia Z (2013) Free energy calculations reveal the origin of binding preference for aminoadamantane blockers of influenza A/M2TM pore. J Chem Theory Comput. https://doi.org/10.1021/ct300899n
18. Christ CD, Fox T (2014) Accuracy assessment and automation of free energy calculations for drug design. J Chem Inf Model. https://doi.org/10.1021/ci4004199
19. Mobley DL, Graves AP, Chodera JD, et al (2007) Predicting absolute ligand binding free energies to a simple model site. J Mol Biol. https://doi.org/10.1016/j.jmb.2007.06.002
20. Medina-Franco JL, Méndez-Lucio O, Martinez-Mayorga K (2014) The interplay between molecular modeling and chemoinformatics to characterize protein–ligand and protein–protein interactions landscapes for drug discovery. Adv Protein Chem Struct Biol. https://doi.org/10.1016/bs.apcsb.2014.06.001
21. Pérez-Benito L, Casajuana-Martin N, Jiménez-Rosés M, van Vlijmen H, Tresadern G (2019) Predicting activity cliffs with free-energy perturbation. J Chem Theory Comput. https://doi.org/10.1021/acs.jctc.8b01290
22. Rogers D, Hopfinger A (1994) Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. J Chem Inform Comput Sci. https://doi.org/10.1021/ci00020a020
23. Rodríguez-Pérez R, Bajorath J (2020) Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. J. Comput Aided Mol Des. https://doi.org/10.1007/s10822-020-00314-0
24. Hussain J, Rea C (2010) Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. J Chem Inf Model. https://doi.org/10.1021/ci900450m
25. Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J (2012) MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. J Chem Inf Model. https://doi.org/10.1021/ci3001138
26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WPr (2002) SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. https://doi.org/10.1613/jair.953
27. Ravikumar S, Kanagasabapathy H, Muralidharan V (2019) Fault diagnosis of self-aligning troughing rollers in belt conveyor system using k-star algorithm. Measurement.

https://doi.org/10.1016/j.measurement.2018.10.001
28. Atkeson CG, Moore AW, Schaal S (1997) Locally weighted learning. Artif Intell Rev.
29. Englert P (2012) Locally weighted learning. In: *Seminar Class on Autonomous Learning Systems.* Citeseer
30. Niels L, Hall M, Frank E (2005) Logistic model trees. Seminar Class on Autonomous Learning Systems. Citeseer
31. Colkesen I, Kavzoglu T (2017) The use of logistic model tree (LMT) for pixel-and object-based classifications using high-resolution WorldView-2 imagery. Geocarto Int. https://doi.org/10.1080/10106049.2015.1128486
32. Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann Stat. https://doi.org/10.1080/10.1214/aos/1016218223
33. Hearst MA, Dumais ST, Osuna E, Platt J, Scholkopf B (1998) Support vector machines. IEEE Intelligent Systems and their applications. https://doi.org/10.1109/5254.708428
34. Zeng Z-Q, Yu H-B, Xu H-R, Xie Y-Q, Gao J (2008) Fast training support vector machines using parallel sequential minimal optimization. In: *2008 3rd international conference on intelligent system and knowledge engineering.* 997-1001
35. John GH, Langley P (2013) Estimating continuous distributions in Bayesian classifiers. arXiv preprint arXiv 1302: 338–345
36. Cano G, Garcia-Rodriguez J, Garcia-Garcia A *et al* (2017) Automatic selection of molecular descriptors using random forest: Application to drug discovery. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2016.12.008
37. Specht DF (1990) Probabilistic neural networks. Neural networks. https://doi.org/10.1016/0893-6080(90)90049-Q
38. Mao KZ, Tan K-C, Ser W (2000) Probabilistic neural-network structure determination for pattern classification. IEEE trans neural netw. https://doi.org/10.1109/72.857781
39. Chen T, Guestrin C (2016) Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining.* https://doi.org/10.1145/2939672.2939785
40. Kozma L (2008) Nearest Neighbors algorithm (kNN). Helsinki University of Technology
41. Kondeti PK, Ravi K, Mutheneni SR *et al* (2019) Applications of machine learning techniques to predict filariasis using socio-economic factors. Epidemiol Infect. https://doi.org/10.1017/S0950268819001481
42. Vehtari A, Gelman A, Gabry J (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. https://doi.org/10.1007/s11222-016-9696-4
43. Triballeau N, Acher F, Brabet I, Pin J-P, Bertrand H-O (2005) Virtual screening workflow development guided by the “receiver operating characteristic” curve approach. Application to high-throughput

docking on metabotropic glutamate receptor subtype 4. Stat Comput. https://doi.org/10.1021/jm049092j

44. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T (2008) Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—what can we learn from earlier mistakes? J Comput Aided Mol Des. https://doi.org/10.1007/s10822-007-9163-6
45. Wang X, Han W, Yan X, Zhang J, Yang M, Jiang P (2020) Pharmacophore features for machine learning in pharmaceutical virtual screening. Mol Divers. https://doi.org/10.1007/s11030-019-09961-4
46. McHugh ML (2012) Interrater reliability: the kappa statistic. Biochem Med. https://doi.org/10.11613/BM.2012.031
47. Burlingham BT, Widlanski T (2003) An intuitive look at the relationship of Ki and IC_{50}: a more general use for the Dixon plot. J Chem Educ. https://doi.org/10.1021/ed080p214
48. Cleary JG, Trigg LE (1995) K*: An instance-based learner using an entropic distance measure. Machine Learning Proceedings. https://doi.org/10.1016/B978-1-55860-377-6.50022-0
49. Carpenter KA, Huang X (2018) Machine learning-based virtual screening and its applications to Alzheimer's drug discovery: a review. Curr Pharm Des. https://doi.org/10.2174/1381612824666180607124038
50. Lipiński PF, Szurmak P (2017) SCRAMBLE'N'GAMBLE: a tool for fast and facile generation of random data for statistical evaluation of QSAR models. Chem Zvesti. https://doi.org/10.1007/s11696-017-0215-7
51. Vilalta R, Giraud-Carrier CG, Brazdil P, Soares C (2004) Using Meta-Learning to Support Data Mining. Int J Comput Sci Appl
52. DuBay KH, Geissler PL (2009) Calculation of proteins' total side-chain torsional entropy and its influence on protein–ligand interactions. J Mol Biol. https://doi.org/10.1016/j.jmb.2009.05.068
53. Zhang J, Liu JS (2006) On side-chain conformational entropy of proteins. PLoS Comput Biol. https://doi.org/10.1371/journal.pcbi.0020168
54. Doig AJ, Sternberg MJ (1995) Side-chain conformational entropy in protein folding. Protein Sci. https://doi.org/10.1002/pro.5560041101
55. Klebe G (2019) Broad-scale analysis of thermodynamic signatures in medicinal chemistry: are enthalpy-favored binders the better development option? Drug Discov Today. https://doi.org/10.1016/j.drudis.2019.01.014
56. Amaral M, Kokh D, Bomke J *et al* (2017) Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. Nat Commun. https://doi.org/10.1038/s41467-017-02258-w
57. Koch C, Heine A, Klebe G (2011) Ligand-induced fit affects binding modes and provokes changes in crystal packing of aldose reductase. Biochim Biophys Acta. https://doi.org/10.1016/j.bbagen.2011.06.001

58. Steuber H, Czodrowski P, Sotriffer CA, Klebe G (2007) Tracing changes in protonation: a prerequisite to factorize thermodynamic data of inhibitor binding to aldose reductase. J Mol Biol. https://doi.org/10.1016/J.JMB.2007.08.063
59. Klebe G (2015) Applying thermodynamic profiling in lead finding and optimization. Nat Rev Drug Discov. https://doi.org/10.1038/nrd4486
60. Ehrmann FR, Stojko J, Metz A *et al* (2017) Soaking suggests “alternative facts”: Only co-crystallization discloses major ligand-induced interface rearrangements of a homodimeric tRNA-binding protein indicating a novel mode-of-inhibition. PloS one. https://doi.org/10.1371/journal.pone.0175723
61. Zubrienė A, Smirnov A, Dudutienė V et al (2017) Intrinsic Thermodynamics and Structures of 2, 4-and 3, 4-Substituted Fluorinated Benzenesulfonamides Binding to Carbonic Anhydrases. ChemMedChem. https://doi.org/10.1002/cmdc.201600509
62. Kišonaitė M, Zubrienė A, Čapkauskaitė E *et al* (2014) Intrinsic thermodynamics and structure correlation of benzenesulfonamides with a pyrimidine moiety binding to carbonic anhydrases I, II, VII, XII, and XIII. PLoS One. https://doi.org/10.1371/journal.pone.0114106
63. Mousa L, Hatmal M, Taha M (2021) Exploiting Activity Cliffs and Machine Learning for Building Pharmacophore Models and Comparison with other Pharmacophore Generation Methods: Sphingosine Kinase 1 as Case Study. J Comput Aided Mol Des *In press*

Tables

Due to technical limitations, tables 1-4 are only available as a download in the Supplemental Files section.

Figures

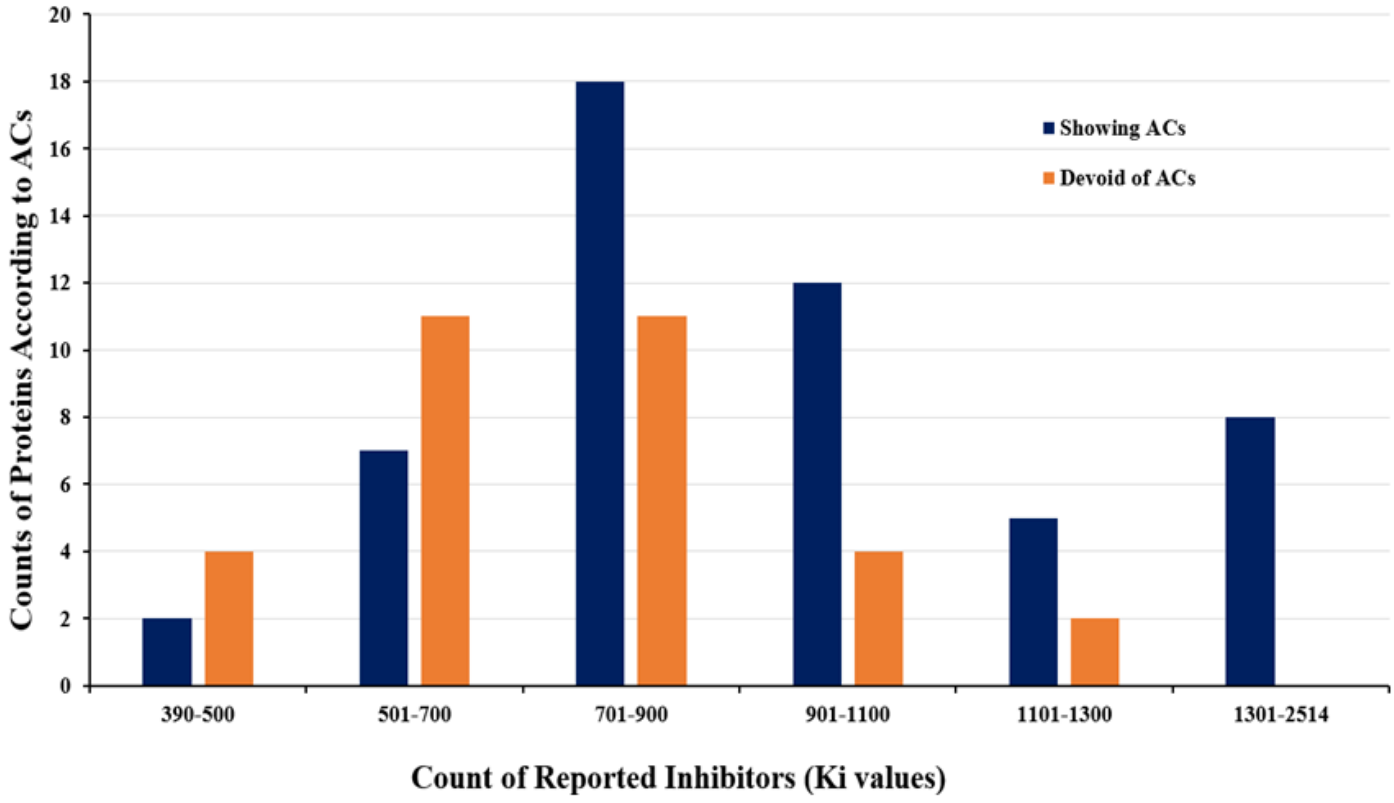

Figure 1

Counts of protein kinases classes (“devoid of AC” and “Showing AC”) as function of number of reported inhibitors in ChEMBL database.

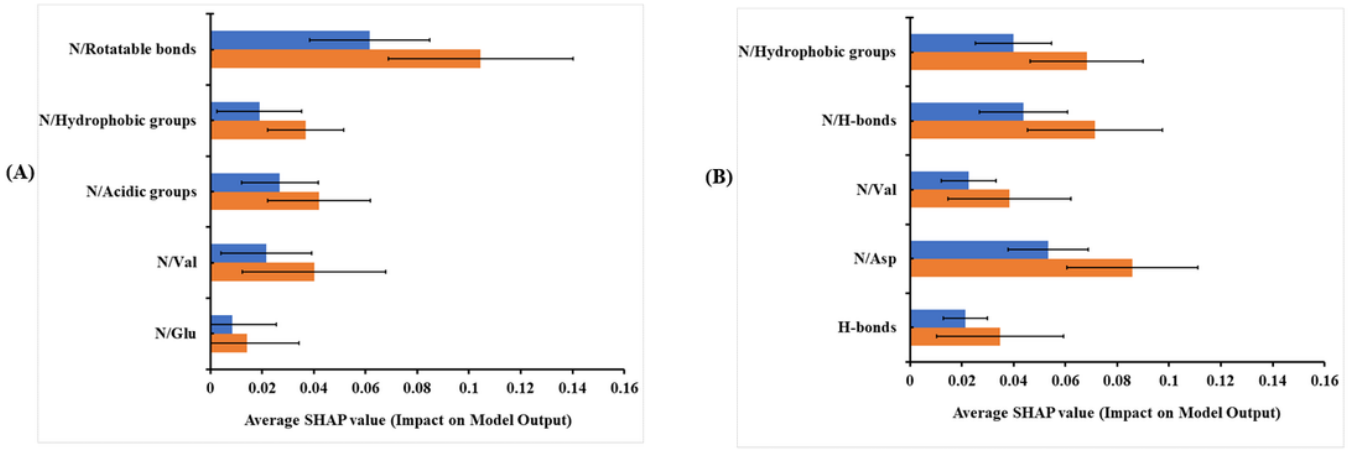

Figure 2: SHAP probability contributions of descriptors emerging in optimal **(A)** GA-K* and **(B)** GA-XGB models. ■ Probability contributions to "Showing Activity Cliffs" label within SMOTE-enhanced testing observations that actually have activity cliffs. ■ Probability contributions to "No Activity Cliffs" label within SMOTE-enhanced testing observations that actually exhibit no activity cliffs". Error bars represent the standard error of probability contributions of certain descriptor.

Figure 2

See image above for figure legend

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Tables.docx
- SupplemntaryMaterials.rar
- SupportingMaterials.docx