UCP Journal of Science and Technology
Vol. 1, Issue 2 (January - June 2024)
Journal website: http://ojs.ucp.edu.pk/index.php/ucpjst/index

Emerging Perception of Activity Cliffs: A Brief Review

Hafiz Saqib Ali[1*]

Abstract

Activity cliffs (ACs) can be characterized as the collection of structurally similar molecules with significant differences in their potencies. Such molecules are of core importance in medicinal and computational chemistry as any minute change in their structure greatly influences their biological action. They play an important role in the optimization during drug discovery and can be analyzed by structure-activity relationships (SAR), but the factors like the molecular representation, selection of the data sets, and the descriptor used greatly affect the end results. Due to these factors, ACs were thought to be a rarity as the concept is contrary to the similarity property principle (SPP), which forms the basis of quantitative structure-activity relationship (QSAR) modeling and likeness based strategies. Today, the data available on activity cliffs has been refined as well as increased a lot. In this review, we have described literature ranging from 1988 to 2021 and highlighted the factors that are important in analyzing ACs and selecting the data sets for the analysis. Moreover, several strategies including matched molecular pairs (MMP) have been developed. MMP is mostly used for finding similar molecules but having a different group(s) responsible for the change in their potency. Furthermore, the role of ML (machine learning) in ACs has also been discussed as it could further refine the analysis of ACs by developing various logarithms and minimizing the faulty results.

Keywords: Activity cliffs; Computational chemistry; Matched molecular pairs; Quantitative structure-activity relationship; Similarity property principle

1. Introduction

The groups of molecules having high structural similarities but different potencies are called activity cliffs (ACs) (Maggiora, 2006). Not only are they intriguing, but also are of prime importance in medicinal and computational chemistry and have remained under discussion for the last three decades, representing the details of SAR-discontinuity (Silipo & Vittoria, 1991; Stumpfe & Bajorath, 2012; Stumpfe *et al*., 2014). Early determination of activity cliffs and an accurate understanding of the activity landscape (AL) are indispensable for the progression of computational models designed for the prediction of the activity of molecules (Guha & Van Drie, 2008; Guha & Van Drie, 2008). AL is used to characterize SAR (structure-activity relationship) by considering two or three dimensions in which chemical space is predicted as 2D projection and potency of compounds as the third dimension, thus making AL similar to the geographical maps that can be apprehended easily (Wassermann *et al*., 2010; Waver *et al*., 2010; Peltason & Bajorath, 2010). If any insignificant

[1] Chemistry Research Laboratory, Department of Chemistry and the INEOS Oxford Institute for Antimicrobial Research, University of Oxford, 12 Mansfield Road, Oxford OX1 3TA, UK
[*]Corresponding author's E-mail: hafiz.ali@chem.ox.ac.uk
Article History:
Received: 26-02-2023; Received in revised form: 29-11-2023; Accepted: 15-12-2023
Available online: 01-04-2024

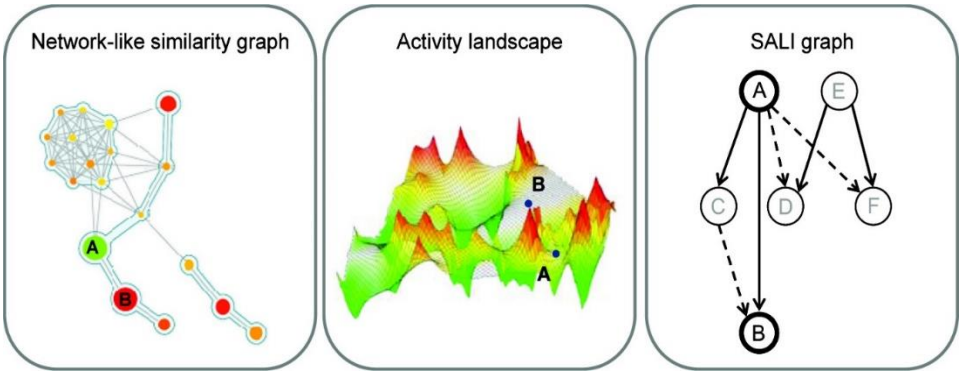

Figure 1 AC can be represented by three methods such as network-like similarity graph, AL and SALI graph. In the first graph, compounds are represented by nodes whereas pairwise similarity relationship is denoted by edges. In the second graph (AL), chemical space between compounds is represented as 2D projection whereas the third dimension is added to represent the potency of the compounds. In the third method (SALI), compounds are denoted by the nodes and edges representing ACs of different magnitudes. Reproduced with permission from Ref. (Stumpfe & Bajorath, 2012). Copyright © 2012, American Chemical Society.

change in the chemical structure of compounds alters biological activity moderately, they are represented by smooth regions, and such compound mapping areas correspond to continuous SARs (Peltason These have proved to be advantageous for QSAR modeling and likeness-based strategies for analytical tools based on SPP (similarity property principle) (Guha & Van Drie, 2008; Bajorath *et al*., 2009; Johnson & maggiora, 1990). Contrarily, discontinuous SARs refer to the canyon like regions formed when a little modification in the chemical structure of the compounds has a drastic impact on their potencies, giving rise to activity cliffs (Waver *et al*., 2010). In other words, ACs are the form of discontinuity in SARs which is the basis of the lead optimization ((Waver *et al*., 2010; Dimova *et al*., 2013).The reliability of the data under consideration is very important before the interpretation of the SARs. Moreover, the size of the data sets is also crucial as data sets with few compounds are harder to study for activity cliffs compared to the ones consisting of a large number of compounds (Medina-Franco, 2013). The activity of these compounds can be studied efficiently against single- or multi-targets leading to single-and multi-target activity landscapes (Wassermann *et al*., 2011). A set criterion has been established and is used for the analysis of ACs quantitatively to determine high structural similarity and difference in activity. For instance, medicinal chemists use empirical rules that are beneficial in converting the qualitative AL data into quantitative one. However, empirical rules are beneficial unless misused as Lipinski's rule of five that overlooked the limitations of the rules (Faller *et al.*, 2011; Ganesan, 2008). Moreover, network-like similarity graph (NSG), structure-activity landscape index (SALI) and SARI (SAR index) can be employed to quantify SARs, and analyze ACs, as represented in the Figure 1 (Stumpfe & Bajorath, 2012; Peltson & Bajorath, 2007; Hu & Bajorath, 2012) SALI have been proposed by Guha and Van Drie to assess the biochemical SAR model and is derived from examining the activities of specific interactions that don't variate linearly with linear property changes (LeDonne *et al*., 2011).

2. Role of molecular structures in ACs

Molecular structures are not only crucial for obtaining actual ACs but also for drawing efficient SAR analysis. If the placement of bonds, protonation or tautomer formation is not proper, the results will be faulty. The type of descriptor can also affect the results. For instance, 2D molecular representation

showing two molecules as the same cliff can show different activity if analyzed by 3D molecular representation, thus making the existence of ACs obscure. This deviation can be explained by using the concept of stereo-isomerization according to medicinal chemists. It is accepted now that actual ACs can be determined by 3D methods compared to 2D approaches (Yongye & Medina-Franco, 2012). Moreover, multiple molecular representation approach is best for minimizing the faulty cliffs in which final results are obtained by considering common conclusions (Yongye & Medina-Franco, 2012; Medina-Franco *et al*., 2009). Additionally, the concept of "data set modelability" has been introduced by determining the effect of ACs on the working of QSAR models (Hu *et al*., 2012).

3. Estimating similarity between molecules

Although the interpretation and detection of activity cliffs are challenging, medicinal chemists can easily extract useful information for analyzing SARs. According to Bajorath *et al.* molecular representation is an essential aspect of AL modeling. Furthermore, practical and interpretable results of the SARs are associated with the correct interpretation of ACs (Golbraikh *et al*., 2014; Agrafiotis *et al*., 2011). The importance of molecular representation has raised many questions like, is there any descriptor that can explain the appearance of ACs? Is there any particular depiction of chemical space that can be beneficial in investigating SAR related to any target? Is there any method for the better representation of AL modeling and for the identification of factual ACs? These questions are not difficult to answer. The response to the first question is that the descriptors used for the interpretation of activity cliffs must provide the information of variables that can be helpful in determining the unknown behavior of compounds regarding activity. For instance, finger-print representation can be used to detect cliffs serving as a process for the structure-based interpretation of activity cliffs (Mendez-Lucio *et al*., 2012). If we talk about the chemical space, with the emergence of new targets and molecular libraries, chemical space can be expanded but many efforts are under process (Nguyen *et al*., 2009; Lopez-Vallejo *et al*., 2012). The last concern can be addressed considering the approach proposed by Hu *et al.* explaining which substructure relationship must be preferred on computed similarity values (Yongye *et al*., 2012). The concept of matched molecular pairs (MMPs) and MMPs-cliffs made the interpretation of results easy from chemical perspective. The main challenge is to establish an interpretable way for the determination of faulty changes in three-dimensions used in ligand-target recognition (Yongye *et al*., 2012; Agrafiotis *et al*., 2011).

4. Methods for the identification of activity cliffs

Many computational protocols have been reported in various publications to apply the concept of AC and their identification. Predominantly, molecular fingerprint and Tanimoto similarity coefficient referred to as molecular graph descriptors are used to calculate similarity values to determine similarity between compounds on the basis of 2D similarity molecular representation (Stumpfe & Bajorath, 2012). Another method for the identification of ACs is matched molecular pair (MPP) formalism which analyze the pair of compounds based on the molecular substructure differing at only a specific site and yield objectively significant chemical explanation (Rabal & Oyarzabal, 2012). Another conscientious approach is based on 3D structures to determine ACs in which the bound ligands represent drastic activity difference despite spatial similarities in the complex structure with a protein of interest (target) as shown in Figure 2 (Rabal & Oyarzabal, 2012).

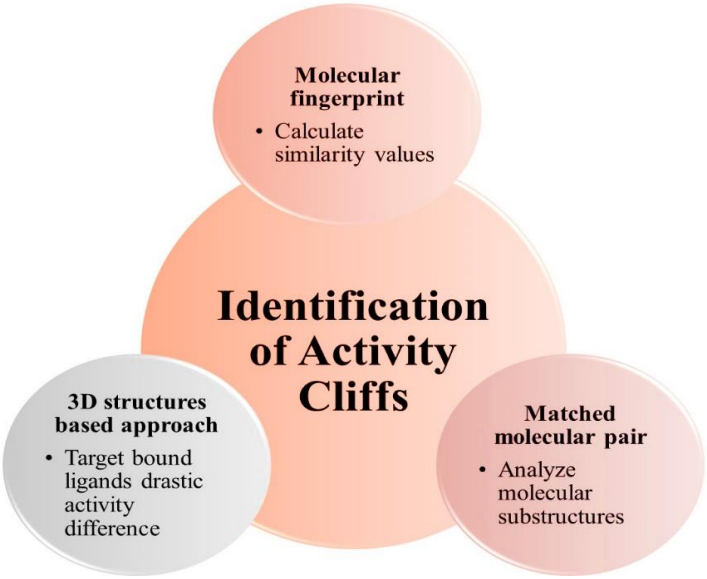

Figure 2 It illustrates the methods for the identification of activity cliffs (ACs). The ACs can be identified by molecular fingerprint, matched molecular pair and 3-dimensional (3D) structure based approach

5. Convolutional neural network in forecasting activity cliffs

Recently, convolutional neural networks (CNN) have acquired catbird-seat in chemical informatics and drug designing because specific features from the image data can be obtained by deriving CNN models from 3-dimensional images of ALs and 2-dimensional images of molecular graphs (Lopez-Vallejo *et al*., 2012). Molecular matched pair (MMP) concept and its implementation is an ingenious logic for the well-ordered determination of structurally similar pairs with different potencies that are referred to as AC's (Bajorath, 2017).The use of MPPs for the representation of AC's, leads to the establishment of MPP-cliffs which are developed by the pair of compounds having potency difference but functional against the common target.[22] These MPP-cliffs predict AC's at various levels that is initiated from the utilization of support vector machine, based on compound pair-based kernel function and fingerprint representation. This support vector machine is used to distinguish between MPP-cliffs from MPPs having no or small variation in potency (Thapa *et al.,* 2020). It is a machine learning algorithm that works by constructing hyper-plane H to segregate the training data of two classes via adjusting distance between the classes in space (Iqbal *et al*., 2021). Subsequently, the methodologically simpler prediction of MPP-cliffs is carried out by implementing condensed graph of reaction formalism (Heikamp *et al*., 2012). Moreover, support vector regression is employed for the quantitative prediction of Potency difference generated by MPPs (Blaschke *et al*., 2021). MPP-cliffs are distinguished from MPPs by the fingerprint features that work by tracking back to the original compounds establishing accurate AC predictions by describing the critically crucial structures (Hussain *et al*., 2010).

6. Molecules-in-molecules fragmentation based method

Molecules-in-molecules (MIM) is referred to as the multi-level hybrid energy fragmentation approach based on the supposition that chemical properties are not largely influenced by the groups far off from the area of interest (Thapa *et al*., 2020; Hu *et al*., 2014). The basic principle of MIM is comprised of four steps namely (1) generation of non-overlapping small

fragments of larger molecules by fragmentation (2) devising overlapped primary substructures utilizing local interactions between fragments (3) employing inclusion-exclusion principle for forming derivative subsystem and (4) assessment of the energy of large molecules by adding the independent energies of sole substructures (Hu *et al*., 2014) In 2020, Raghavachari and co-workers have used quantum mechanical (QM) investigation to disclose the formation of ACs by using MIM fragmentation method. The main advantage was the reduced computational cost of QM estimations and calculations. Moreover, MIM method can also identify critical residues by residue specific energy decomposition analysis to distinguish between two ligands. Therefore, MIM is considered as an ideal method to apprehend ACs (Thapa *et al*., 2020).

7. Activity cliffs in drug discovery

ACs are important in drug discovery because a small change in the structure of a particular compound greatly influences its biological activity (Stumpfe *et al*., 2014; Thapa *et al*., 2018). In medicinal chemistry, they play an important role in early phase drug discovery for finding the determinants of high interest in hit-to-lead. Although Acs are crucial in determining SARs information, medicinal chemists encounter many hurdles in the preparation and analysis due to the duality of Acs (Bajorath, 2019). The duality of Acs is similar to the potency difference and similarity parameters for defining Acs and effect their analysis, application and perception, thus making them controversial (Bajorath, 2019). The extent of expertise of medicinal chemist to handle Acs, the computational method for the identification of Acs and the variation in the SAR discontinuity meaning while optimization of lead can result in the duality of Acs (Bajorath, 2017; Bajorath, 2019).

8. Strategies for finding ACs

There exist several strategies for the identification of ACs in databanks (Waver *et al*., 2010). ChEMBL is widely used database for finding the data sets that are formed with time (Stumpfe *et al*., 2013). One of them is the SALI strategy which identifies ACs by comparative scale method because the scale is not definite. SALI has a drawback as it identifies shallow or pseudo cliffs at a particular cutoff value (Waver *et al*., 2010). Some rules were set by professionals to identify the ACs. For example, a molecule must have activity in nanomolar range, an already determined likeness measure is performed, and the activity difference between two molecules must not be less than 100 fold (Waver *et al*., 2010). Matched molecular pairs (MMPs) concept has also been utilized for finding similar molecules that are distinct at some point. For example, the type of ring or an R group which is determined by the implementation of in-house algorithm as proposed by Hussain and Rea (Hussain & Rea, 2010). The characterization of ACs can be performed by various methods like presence of different R-groups and interactions between them, and 3D likeness determined by X-rays during ligand target interactions (Aguayo *et al*., 2014). It must be considered that compound data sets chosen should have reported K_i values (Stumpfe *et al*., 2013).

9. Synthetic relevance

In chemoinformatics, molecular fingerprint descriptors form the basis for the calculation of Tanimoto similarity values for defining ACs (Stumpfe *et al*., 2014; Thapa *et al*., 2018). It is hard to understand Tanimoto similarity values as they don't depend on synthetic or analog relationship, thus MMPs (matched molecular pairs) are used to represent ACs as MMPs cliffs but with size-restricted chemical modifications (Hussain and Rea, 2010). However, there is a drawback of MMP cliffs that they are unable to

determine synthetic relationships; therefore, a modified version retrosynthetic MMP cliffs (RMMP cliffs) was proposed in accordance with retrosynthetic rules by which systematic computational framework consisting of exocyclic single bonds in MMPs was substituted with fragmentation. RMMPs have been formed systemically from currently available compounds having activity data of high confidence (Hu and Bajorath, 2018). The main advantage of retrosynthetic bond fragmentation method is the small fragment space occupancy giving rise to smaller number of RMMP cliffs compared to MMP cliffs (Hussain & Rea, 2010; Hu & Bajorath, 2019).

10. Target set-dependent differences in activity and optimization of compounds

It is widely considered that potency difference is already set for the determination of ACs but actually potency difference vary depending on the target set (particular pharmacological target and compound's activity class) (Hu & Bajorath, 2019). In order to avoid the ignorance of target set- dependent activity difference and enhancing SARs exploration, ACs need to be redefined. Accordingly, the threshold for AC formation is obtained by measuring the pair of compounds in a particular target set following the AC similarity criterion (Hu & Bajorath, 2019; Vogt *et al.*, 2011). ACs duality may be due to the fact that how compounds are dealt by medicinal chemists or how they are analyzed computationally. As different target sets consist of different compounds from different sources, they require different optimization procedures and efforts to obtain factual ACs (Thapa *et al.*, 2018).

11. Coordination and frequency of ACs

When analyzing a data set, number of ACs are obtained in a coordinated manner with huge variation in activities of structurally related compounds. Moreover, a single compound has the ability to form a large number of ACs with different analogs. In AC network, compounds present in a data set are represented as nodes, and pair-wise edges serve as activity cliffs and coordinated cliffs, produced by subsets of compounds, leading to disjoint cluster formation (Stumpfe *et al.*, 2014; Demova *et al.*, 2015). AC clusters are more favorable for the analysis of SARs compared to isolated cliffs because >95% of the cliffs of various data sets are produced in a coordinated manner (Stumpfe & Bajorath, 2015; Stumpfe & Bajorath, 2012). Clusters of ACs often consist of "hubs" with numerous partner compounds forming center of local ACs representing molecules as nodes and such molecules are referred as "activity cliff generators".(Stumpfe & Bajorath, 2012). Frequency of the occurrence of ACs for different data sets has also been found along with the coordination of ACs, and the information has been increased tremendously over time as total number of activity data was doubled from 2011 to 2015 with more than 17,000 MMP cliffs in 2015 (Stumpfe & Bajorath, 2012).

12. Determination of potency differences

The measurement of potency differences related to AC production depends on the comparison of experimental values. The validity of AC assignments is certified by assessing the potency difference (in theory) using K_D (dissociation constant) or K_i (assay-independent equilibrium) values. The evaluation of ACs can also be completed formally with increasing potency difference as continuation of pairs of compounds (Hu *et al*., 2012). The use of the constant potency difference is preferred not only in the analysis of ACs, but also for finding ACs in databases (Stumpfe *et al*., 2012; Stumpfe *et al*., 2014). Pair-wise potency difference is lower compared to constant potency difference threshold in analysis of analogs, and is significant statistically. Moreover, potency difference of nearly 100-fold is

usually utilized in AC analysis (Stumpfe *et al.,* 2019). The role of constant potency difference threshold has made computational search for ACs easier, but it doesn't consider class dependent activity difference in the distribution of compound potency which vary greatly in activity class due to similarity relationship between compounds (Stumpfe *et al.,* 2020). AC analysis can further be refined by the class dependent deviation of potency difference threshold. For the determination of class dependent threshold, average of the compound pair-based potency difference distribution plus two standard deviations are statistically performed (Vogt, 2011).

13. Activity cliff generations

The evolution of the ACs is contributed to the variation in the manner of addressing similarity and potency difference. This variation played a part in differentiating among the three generations of two dimensional ACs, also known as molecular graph based ACs (Stumpfe et al., 2019; Stumpfe *et al.,* 2020). First generation of ACs is classified as a separate group using constant potency difference threshold in all the activity classes and similarity measures based on sub-structures Stumpfe *et al.,* 2020). In 2015, they (first generation ACs) were reported based on ChEMBL release 20 by extracting 48,244 compounds having K_i values and were active against 746 targets. MACCS structural keys, MPP formalism and extended connectivity fingerprint with bond diameter 4 (ECFP4) were used to identify first generation activity cliffs with $\Delta pK_i \geq 2$, where ΔpK_i represents potential difference threshold (Perez *et al.,* 2015;Stumpfe *et al.,* 2020; Stumpfe *et al.,*2017) Second generation of ACs came into existence because of capturing single substitution site of structural analogs (R), and MMP cliff formalism with varying potency difference threshold depending upon activity class. Their search was initiated in ChEMBL release 23 from which 212 activity classes having potential for AC formation were identified (Hu *et al.,* 2018). These 212 activity classes yielded 16,096 class dependent RMMP-cliffs having ΔpK_i between 1 and 2.5 (Hu *et al.,* 2019). However, 11,773 RMMP-cliffs were obtained in 195 classes when the ΔpK_i was ≥ 2. Moreover, 145 RMMP-cliffs containing inactive compounds that are obtained from screening assays in PubChem were also identified from the eight activity classes with available screening data (Hu *et al.,* 2019). Furthermore, single or multiple substitution analog pairs belonging to same series gave rise to third generation ACs with activity class dependent potency difference threshold (Stumpfe *et al.,* 2020). They were 16,454 analog series-based ACs in ChEMBL release 24.1 having class-dependent potency difference threshold. Only 25.6% of 4204 instances were third generation ACs with multi-site cliffs while others contained a single site for substitution (Stumpfe *et al.,*2019*).*

14. Activity cliffs containing privileged substructures

The concept of privileged substructures (PS) was introduced by Evans et al., representing non-class specific compounds with specific biological activities (Evans *et al.,* 1988). They have been a center of attention in pharmaceutical research for long. It has been found that the activity cliffs comprising selected PS exhibited huge improvements in efficacy of ligands (Hu & Bajorath, 2020).

15. ACs, chemoinformatics and machine learning

QSAR studies are promoted due to continuous SARs and discontinuous SARs have negative impact. Several machine learning (ML) methods have been developed to assist chemoinformatics (Aguiar *et al.,* 2013; Rose, 2013), and are involved in the classification and generalization of data and help in making findings logical (Rose, 2013).However, the ML methods need improvement as they just observe the "rolling hills" as the key signs for the determination of small ACs,

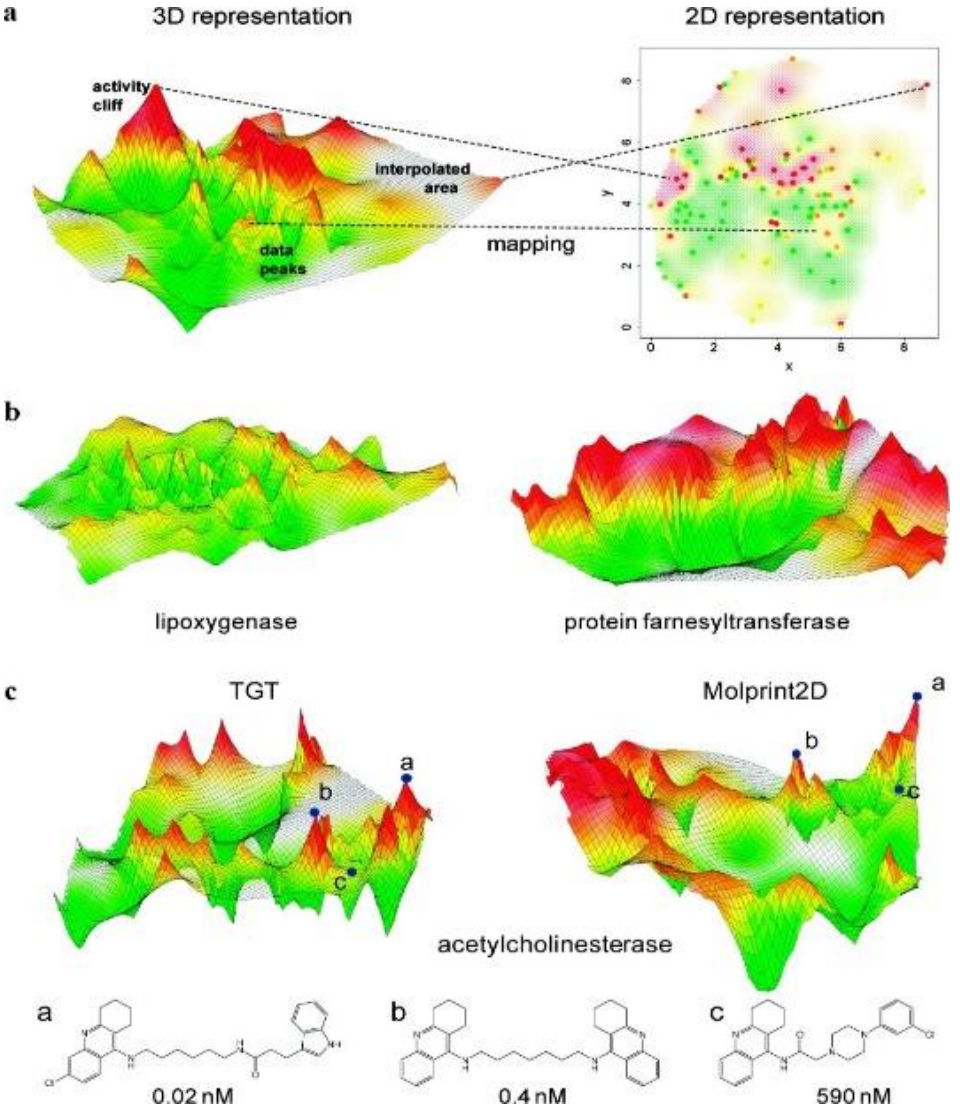

Figure 3 A comprehensive representation of 3D AL. (a) A comparison between 2D and 3D activity landscape formed by dimension reduction for the analysis of a set of 112 acetylcholinesterase inhibitors, represented as MACCS fingerprints. (b) It represents the 3D AL analysis of a set of lipoxygenase inhibitors (252 inhibitors) and protein farnesyltransferase inhibitors (146 inhibitors) based upon Molprint 2D fingerprint representation. AL for lipoxygenase is indicating SARs continuity while AL for protein farnesyltransferase is representing SARs discontinuity. (c) TGT and Molprint 2D fingerprint are used for the analysis of ACs of acetylcholinesterase inhibitors that were used in Figure (a) but here just three compounds are mapped with their reported potencies. It shows that the ACs can be formed or not depending upon the molecular representation. Reproduced with permission from Ref (Wassermann *et al.*, 2010). Copyright © 2010, American Chemical Society.

thus reducing the forth coming forecast steadiness (Guha, 2011). The determination of the occurrence of ACs in an area of ML and explaining the bad effects of ACs is an important task. However, a pre-processing innovative filtering approach (PRISM) was proposed by Smith and Martinez, and according to a study, ISMs represent estimations to an extra precision. This extra precision is not shown by all the ISMs but a huge correlation can be shown between ACs and

Figure 4 It enlightens the reasons behind the poor prediction of activity cliffs and their possible solutions.

ISMs. In AL-modeling, ISM has the ability to perform similar incidences with different names in an area. Moreover, ISM can swing to the EL (elementary landscape) that can be renowned by outliers and noise. Three current outliers were discovered by Smith and Martinez that helped in equating PRISM by mediating discovery principles and noise reduction. Pre-removal of the cases determined by PRISM can help to enhance the precision of classification related to the instrument learning computation formed by initial data groups. Thus, removing the ISMs from ML procedures will make the suitable categorization difficult in advance training certificates (Guha, 2011). Moreover, Weka, a data mining suite, is used for the formation of a computational model and provides maximum flexibility for the frameshifting sites when analyzing a new data set for the analysis of ACs (Frank *et al.*, 2004).

16. Local vs universal molecular likeness

It is necessary that the medicinal chemists use different calculations to analyze ACs in addition to employing different approaches for determining the likeness of compounds. Scalar molecular caption is responsible for regional similarities that involve the determination of molecular design by topological, functional and constitutional attitude and is novel to researchers working on chemoinformatics who are specialists of QSAR and drug-based pathway. On the other hand, medicinal chemists utilize all the molecular illustrations (2D chemical space depiction) based on fingerprints like MACCS keys (Rogers & Hahn, 2010), as represented in Figure 3a, which involves a comparison between 3D and 2D illustrations based on MACCS fingerprint. Furthermore, Figures 3b and 3c show comparison between the ACs of lipoxygenase and protein farnesyltransferase inhibitors and the ACs formed by TGT and Molprint 2D for the same compounds, respectively (Rogers & Hahn, 2010).

17. Reasons for poor prediction

The restriction in the prediction is not closely linked with the methodology but it is retaled with the data. A number of descriptors and statistical approaches are available that are used to validate and predict the data sets probably by cross validation. However, all the data sets can't be predicted due to some reasons. These reasons for poor predictivity involve the absence of activity related features in descriptors, inability of QSAR suitability in complex relation between descriptors and activity, a huge variation in the experimental uncertainty of observed activities among molecules and the formation of more or steeper activity cliffs (Figure 4) (Sheridan *et al.*, 2020).

18. Solutions for poor predictivity

If large uncertainty in observed values in dataset is the issue, routine multiple measurements of molecules to cancel its effect in the average and reduction in the assay experimental error can also

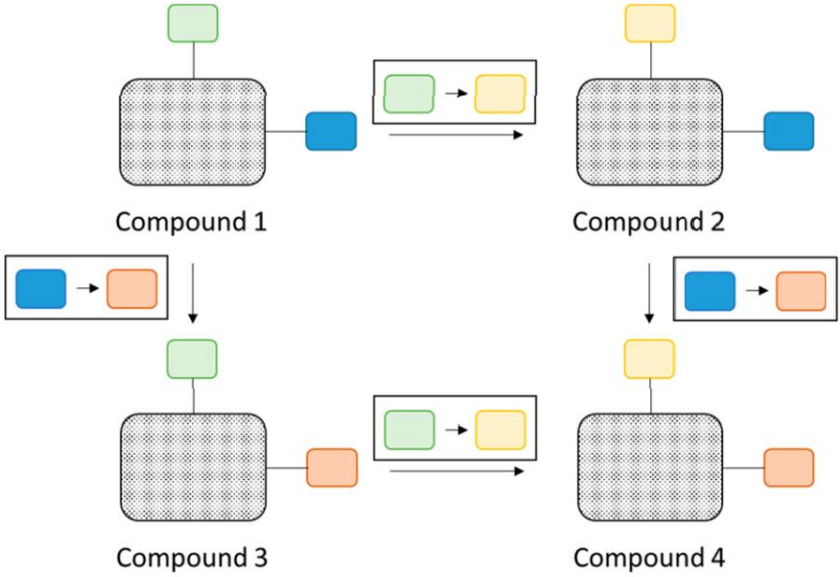

Figure 5 A diagrammatic representation of nonadditivity analysis known as double transformation cycles. It consist of four compounds that are attached by identical transformations; whereas, the colored blocks in the diagram are representing different functional groups. Reproduced with permission from Ref (Kramer *et al*., 2021). Copyright © 2019, American Chemical Society.

ameliorate the issue. The activity cliff metrices that can't differentitate the 'real' ACs which may be revealed by reducing the uncertainty to a minimum level (Sheridan *et al.*, 2020). According to the literature, the elucidation of ACs is possible by investigating the way molecules bound with receptors with various poses. If it is possible, none QSAR model alone can interfere with the unpredictability of ACs no matter what method or descriptor is used (Figure 4) (Sheridan *et al.*, 2020).

19. Nonadditivity analysis

The determination of potential SAR outliers, nonadditivity (NA) and upper limit estimation of data set experimental uncertainty are the contributions of NA analysis. The term nonadditivity refers to the varied results obtained by the fusion of two new fragments as compared to the sum of their individual effect (Kramer, 2019). In case of linear SAR, NA can be an issue; but, if it is used intentionally, NA can occupy a crucial place in drug discovery. Additionally, NA analysis can help apprehend possible experimental noise and provides deep structural insights (Gogishvili *et al.*, 2021). Nonadditivity in compounds can be caused by the variation in hydrophobicity and hydrophilicity, residual mobility, internal hydrogen bonds and clear conformational changes which causes NA above 2 log units (Baum *et al.*, 2010; Kramer *et al.*, 2015). The calculation of NA is done by double transformation cycles that consist of four compounds linked by two indistinguishable transformations as shown in Figure 5. However, some experimental uncertainty is present in the values measured for each compound and the addition of these uncertainties provide apparent and false NA. Thus, this save the time of the researcher from finding explanations for the false non existing effects (Kramer, 2019; Cockroft & Hunter, 2007; Fischer *et al.*, 2007).

20. Conclusions

ACs have been found to be the heart of medicinal and computational chemistry as they play a conspicuous role in drug discovery. A large number of ACs have been recognized and studied by SARs

analysis utilizing AL modeling. The occurrence of ACs limits the use of QSAR and likeness-based strategies, which support the concept that 'structurally similar molecules are also potentially identical'. Today, ML has been used for the accurate analysis of ACs and plays a role in optimization for the lead-compound in drug designing by forming various algorithms. Despite all the efforts till now, many algorithms are still required to further ameliorate the analysis.

21. Statements and Declarations

Conflict of Interest: No conflict of interest.

Funding: No funding has been received from any source; the work is self-supported.

Author's Contribution: Single author contribution.

References

Agrafiotis, D. K., Wiener, J. J., Skalkin, A., & Kolpak, J. (2011). Single R-group polymorphisms (SRPs) and R-cliffs: An intuitive framework for analyzing and visualizing activity cliffs in a single analog series. *Journal of Chemical Information and Modeling*, *51*(5), 1122-1131.

Aguayo-Ortiz, R., Pérez-Villanueva, J., Hernández-Campos, A., Castillo, R., Meurice, N., Medina-Franco, J. L. (2014). Chemoinformatic characterization of activity and selectivity switches of antiprotozoal compounds. *Future Medicinal Chemistry*, *6*(3), 281-294.

Aguiar-Pulido, V., Gestal, M., Cruz-Monteagudo, M., Rabuñal, J. R., Dorado, J., & Munteanu, C. R. (2013). Evolutionary computation and QSAR research. *Current Computer-Aided Drug Design*, *9*(2), 206-225.

Bajorath, J. (2017). Representation and identification of activity cliffs. *Expert Opinion on Drug Discovery*, *12*(9), 879-883.

Bajorath, J., Peltason, L., Wawer, M., Guha, R., Lajiness, M. S., & Van Drie, J. H. (2009). Navigating structure–activity landscapes. *Drug Discovery Today*, *14*(13-14), 698-705.

Blaschke, T., Feldmann, C., & Bajorath, J. (2021). Prediction of promiscuity cliffs using machine learning. *Molecular Informatics*, *40*(1), 2000196.

Baum, B., Muley, L., Smolinski, M., Heine, A., Hangauer, D., & Klebe, G. (2010). Non-additivity of functional group contributions in protein–ligand binding: a comprehensive study by crystallography and isothermal titration calorimetry. *Journal of Molecular Biology*, *397*(4), 1042-1054.

Cockroft, S. L., & Hunter, C. A. (2007). Chemical double-mutant cycles: dissecting non-covalent interactions. *Chemical Society Reviews*, *36*(2), 172-188.

De la Vega de León, A., & Bajorath, J. (2014). Prediction of compound potency changes in matched molecular pairs using support vector regression. *Journal of Chemical Information and Modeling*, *54*(10), 2654-2663.

Dimova, D., Heikamp, K., Stumpfe, D., & Bajorath, J. (2013). Do medicinal chemists learn from activity cliffs? A systematic evaluation of cliff progression in evolving compound data sets. *Journal of Medicinal Chemistry*, *56*(8), 3339-3345.

Dimova, D., Stumpfe, D., Hu, Y., & Bajorath, J. (2015). Activity cliff clusters as a source of structure–activity relationship information. *Expert Opinion on Drug Discovery*, *10*(5), 441-447.

Evans, B. E., Rittle, K. E., Bock, M. G., DiPardo, R. M., Freidinger, R. M., Whitter, W. L., Hirshfield, J. (1988). Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *Journal of Medicinal Chemistry*, *31*(12), 2235-2246.

Faller, B., Ottaviani, G., Ertl, P., Berellini, G., & Collis, A. (2011). Evolution of the physicochemical properties of marketed drugs: can history foretell the

future?. *Drug Discovery Today*, *16*(21-22), 976-984.

Fischer, F. R., Schweizer, W. B., & Diederich, F. (2007). Molecular torsion balances: Evidence for favorable orthogonal dipolar interactions between organic fluorine and amide groups. *Angewandte Chemie International Edition*, *46*(43), 8270-8273.

Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics*, *20*(15), 2479-2481.

Ganesan, A. (2008). The impact of natural products upon modern drug discovery. *Current Opinion in Chemical Biology*, *12*(3), 306-317.

Golbraikh, A., Muratov, E., Fourches, D., & Tropsha, A. (2014). Data set modelability by QSAR. *Journal of Chemical Information and Modeling*, *54*(1), 1-4.

Guha, R., & Van Drie, J. H. (2008). Assessing How Well a Modeling Protocol Captures a Structure− Activity Landscape. *Journal of Chemical Information and Modeling*, *48*(8), 1716-1728.

Guha, R. (2011). The ups and downs of structure–activity landscapes. *Chemoinformatics and Computational Chemical Biology*, 101-117.

Guha, R., & Van Drie, J. H. (2008). Structure− activity landscape index: identifying and quantifying activity cliffs. *Journal of Chemical Information and Modeling*, *48*(3), 646-658.

Heikamp, K., Hu, X., Yan, A., & Bajorath, J. (2012). Prediction of activity cliffs using support vector machines. *Journal of Chemical Information and Modeling*, *52*(9), 2354-2365.

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. *Journal of Chemical Information and Modeling*, *52*(5), 1138-1145.

Hu, Y., & Bajorath, J. (2012). Exploration of 3D activity cliffs on the basis of compound binding modes and comparison of 2D and 3D cliffs. *Journal of Chemical Information and Modeling*, *52*(3), 670-677.

Hu, H., & Bajorath, J. (2020). Increasing the public activity cliff knowledge base with new categories of activity cliffs. *Future Science OA*, *6*(5), FSO472.

Hu, H., Stumpfe, D., & Bajorath, J. (2019). Systematic identification of target set-dependent activity cliffs. *Future Science OA*, *5*(2), FSO363.

Hu, Y., de León, A. D. L. V., Zhang, B., & Bajorath, J. (2014). Matched molecular pair-based data sets for computer-aided medicinal chemistry. *F1000Research*, *3*.

Hu, H., Stumpfe, D., & Bajorath, J. (2018). Rationalizing the formation of activity cliffs in different compound data sets. *ACS Omega*, *3*(7), 7736-7744.

Hu, H., Stumpfe, D., & Bajorath, J. (2019). Second-generation activity cliffs identified on the basis of target set-dependent potency difference criteria. *Future Medicinal Chemistry*, *11*(5), 379-394.

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. *Journal of Chemical Information and Modeling*, *52*(7), 1806-1811.

Hu, Y., Furtmann, N., Gütschow, M., & Bajorath, J. (2012). Systematic identification and classification of three-dimensional activity cliffs. *Journal of Chemical Information and Modeling*, *52*(6), 1490-1498.

Hussain, J., & Rea, C. (2010). Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *Journal of chemical information and modeling*, *50*(3), 339-348.

Horvath, D., Marcou, G., Varnek, A., Kayastha, S., de la Vega de León, A., & Bajorath, J. (2016). Prediction of activity

cliffs using condensed graphs of reaction representations, descriptor recombination, support vector machine classification, and support vector regression. *Journal of Chemical Information and Modeling*, *56*(9), 1631-1640.

Iqbal, J., Vogt, M., & Bajorath, J. (2021). Prediction of activity cliffs on the basis of images using convolutional neural networks. *Journal of Computer-Aided Molecular Design*, 1-8.

Johnson, M. A., & Maggiora, G. M. (1990). Concepts and applications of molecular similarity. *(No Title)*.

Kramer, C., Fuchs, J. E., & Liedl, K. R. (2015). Strong nonadditivity as a key structure–activity relationship feature: distinguishing structural changes from assay artifacts. *Journal of Chemical Information and Modeling*, *55*(3), 483-494.

Kramer, C. (2019). Nonadditivity analysis. *Journal of Chemical Information and Modeling*, *59*(9), 4034-4042.

LeDonne, N. C., Rissolo, K., Bulgarelli, J., & Tini, L. (2011). Use of structure-activity landscape index curves and curve integrals to evaluate the performance of multiple machine learning prediction models. *Journal of Cheminformatics*, *3*(1), 1-12.

López-Vallejo, F., Giulianotti, M. A., Houghten, R. A., & Medina-Franco, J. L. (2012). Expanding the medicinally relevant chemical space with compound libraries. *Drug Discovery Today*, *17*(13-14), 718-726.

Maggiora, G. M. (2006). On outliers and activity cliffs why QSAR often disappoints. *Journal of Chemical Information and Modeling*, *46*(4), 1535-1535.

Medina-Franco, J. L. (2013). Activity cliffs: facts or artifacts?. *Chemical Biology & Drug Design*, *81*(5), 553-556.

Medina-Franco, J. L., Martínez-Mayorga, K., Bender, A., Marín, R. M., Giulianotti, M. A., Pinilla, C., & Houghten, R. A. (2009). Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. *Journal of Chemical Information and Modeling*, *49*(2), 477-491.

Méndez-Lucio, O., Pérez-Villanueva, J., Castillo, R., & Medina-Franco, J. L. (2012). Identifying activity cliff generators of PPAR ligands using SAS maps. *Molecular Informatics*, *31*(11-12), 837-846.

Medina-Franco, J. L., Edwards, B. S., Pinilla, C., Appel, J. R., Giulianotti, M. A., Santos, R. G., Houghten, R. A. (2013). Rapid scanning structure–activity relationships in combinatorial data sets: identification of activity switches. *Journal of Chemical Information and Modeling*, *53*(6), 1475-1485.

Muley, L., Baum, B., Smolinski, M., Freindorf, M., Heine, A., Klebe, G., & Hangauer, D. G. (2010). Enhancement of hydrophobic interactions and hydrogen bond strength by cooperativity: synthesis, modeling, and molecular dynamics simulations of a congeneric series of thrombin inhibitors. *Journal of Medicinal Chemistry*, *53*(5), 2126-2135.

Nguyen, K. T., Blum, L. C., Van Deursen, R., & Reymond, J. L. (2009). Classification of organic molecules by molecular quantum numbers. *ChemMedChem: Chemistry Enabling Drug Discovery*, *4*(11), 1803-1805.

Perez-Villanueva, J., Méndez-Lucio, O., Soria-Arteche, O., & Medina-Franco, J. L. (2015). Activity cliffs and activity cliff generators based on chemotype-related activity landscapes. *Molecular Diversity*, *19*, 1021-1035.

Peltason, L., Iyer, P., & Bajorath, J. (2010). Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. *Journal of Chemical Information and Modeling*, *50*(6), 1021-1033.

Peltason, L., & Bajorath, J. (2007). SAR index: Quantifying the nature of structure− activity relationships. *Journal of Medicinal Chemistry*, *50*(23), 5571-5578.

Rabal, O., & Oyarzabal, J. (2012). Using novel descriptor accounting for ligand–receptor interactions to define and visually explore biologically relevant chemical space. *Journal of Chemical Information and Modeling*, *52*(5), 1086-1102.

Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742-754.

Rose, J. (2003). Methods for data analysis. *Handbook of Chemoinformatics. Weinheim: Wiley-VCH*. 1081–1097.

Silipo, C., & Vittoria, A. (1991). QSAR, rational approaches to the design of bioactive compounds. In *European Symposium on Quantitative Structure-Activity Relationships 1990: Sorrento, Italy)*. Distributors for the US and Canada, Elsevier Science.

Sheridan, R. P., Karnachi, P., Tudor, M., Xu, Y., Liaw, A., Shah, F., Alvarez, J. (2020). Experimental error, kurtosis, activity cliffs, and methodology: What limits the predictivity of quantitative structure–activity relationship models. *Journal of Chemical Information and Modeling*, *60*(4), 1969-1982.

Stumpfe, D., Hu, H., & Bajorath, J. (2019). Introducing a new category of activity cliffs with chemical modifications at multiple sites and rationalizing contributions of individual substitutions. *Bioorganic & Medicinal Chemistry*, *27*(16), 3605-3612.

Stumpfe, D., & Bajorath, J. (2012). Frequency of occurrence and potency range distribution of activity cliffs in bioactive compounds. *Journal of Chemical Information and Modeling*, *52*(9), 2348-2353.

Stumpfe, D., Dimova, D., & Bajorath, J. (2014). Composition and topology of activity cliff clusters formed by bioactive compounds. *Journal of Chemical Information and Modeling*, *54*(2), 451-461.

Stumpfe, D., & Bajorath, J. (2012). Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of Medicinal Chemistry*, *55*(7), 2932-2942.

Stumpfe, D., Hu, Y., Dimova, D., & Bajorath, J. (2014). Recent progress in understanding activity cliffs and their utility in medicinal chemistry: miniperspective. *Journal of Medicinal Chemistry*, *57*(1), 18-28.

Stumpfe, D., Dimova, D., Heikamp, K., & Bajorath, J. (2013). Compound pathway model to capture SAR progression: comparison of activity cliff-dependent and-independent pathways. *Journal of Chemical Information and Modeling*, *53*(5), 1067-1072.

Stumpfe, D., & Bajorath, J. (2015). Monitoring global growth of activity cliff information over time and assessing activity cliff frequencies and distributions. *Future Medicinal Chemistry*, *7*(12), 1565-1579.

Stumpfe, D., Hu, H., & Bajorath, J. (2019). Evolving concept of activity cliffs. *ACS omega*, *4*(11), 14360-14368.

Stumpfe, D., Hu, H., & Bajorath, J. (2020). Advances in exploring activity cliffs. *Journal of Computer-Aided Molecular Design*, *34*, 929-942.

Stumpfe, D., Tinivella, A., Rastelli, G., & Bajorath, J. (2017). Promiscuity of inhibitors of human protein kinases at varying data confidence levels and test frequencies. *RSC advances*, *7*(65), 41265-41271.

Thapa, B., Erickson, J., & Raghavachari, K. (2020). Quantum mechanical investigation of three-dimensional activity cliffs using the Molecules-in-Molecules fragmentation-based method. *Journal of Chemical Information and Modeling*, *60*(6), 2924-2938.

Thapa, B., Beckett, D., Jovan Jose, K. V., & Raghavachari, K. (2018). Assessment of fragmentation strategies for large

proteins using the multilayer molecules-in-molecules approach. *Journal of Chemical Theory and Ccomputation*, *14*(3), 1383-1394.

Vogt, M., Huang, Y., & Bajorath, J. (2011). From activity cliffs to activity ridges: informative data structures for SAR analysis. *Journal of Chemical Information and Modeling*, *51*(8), 1848-1856.

Wassermann, A. M., Wawer, M., & Bajorath, J. (2010). Activity landscape representations for structure− activity relationship analysis. *Journal of Medicinal Chemistry*, *53*(23), 8209-8223.

Wawer, M., Lounkine, E., Wassermann, A. M., & Bajorath, J. (2010). Data structures and computational tools for the extraction of SAR information from large compound sets. *Drug Discovery Today*, *15*(15-16), 630-639.

Witten, I. H., & Frank, E. (2002). Data mining: practical machine learning tools and techniques with Java implementations. *Acm Sigmod Record*, *31*(1), 76-77.

Wassermann, A. M., Dimova, D., & Bajorath, J. (2011). Comprehensive analysis of single-and multi-target activity cliffs formed by currently available bioactive compounds. *Chemical Biology & Drug Design*, *78*(2), 224-228.

Yongye, A. B., & Medina-Franco, J. L. (2012). Data mining of protein-binding profiling data identifies structural modifications that distinguish selective and promiscuous compounds. *Journal of Chemical Information and Modeling*, *52*(9), 2454-2461.