Sampling Chemical Space: Activity Cliffs, Extended Similarity, and ML Performance

Kenneth López-Pérez[1], Ramón Alain Miranda-Quintana[1*]

[1]*Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611, USA.*

Email: quintana@chem.ufl.edu

Abstract

The presence of Activity Cliffs (ACs) has been known to represent a challenge for QSAR modeling. With its data high dependency, Machine Learning QSAR models will be highly influenced by the activity landscape of the data. We propose several extended similarity and extended SALI methods to study the implications of ACs distribution on the training and test sets on the model's errors. Non-uniform ACs and chemical space distribution tends to lead to worse models than the proposed uniform methods. ML modeling on AC-rich sets needs to be analyzed case-by-case. Proposed methods can be used as a tool to study the dataset, with random and uniform splitting being the better overall data splitting alternatives.

Keywords: chemical space; drug design; similarity; extended similarity; activity cliffs; structure activity relationships

Introduction

The existence of highly similar compounds with a large difference in bioactivity has been pointed out in medicinal chemistry for several decades.[1] The term *Activity Cliffs* (ACs) has been adopted to describe pairs of molecules that exhibit this behavior, as shown in Figure 1's example.[1,2] ACs break the "*similar molecule have similar properties*" principle[3], making them of special attention in Quantitative Structure Activity Relationship (QSAR) modeling since it represents a challenge to overcome.[4] The analysis of large datasets with presence of ACs has

brought up the concept of activity landscape. Activity landscapes are dependent on the chemical space representation used for the compounds, similarity definition, and the activity essay to measure the bioactivity for a certain target. [4,5]

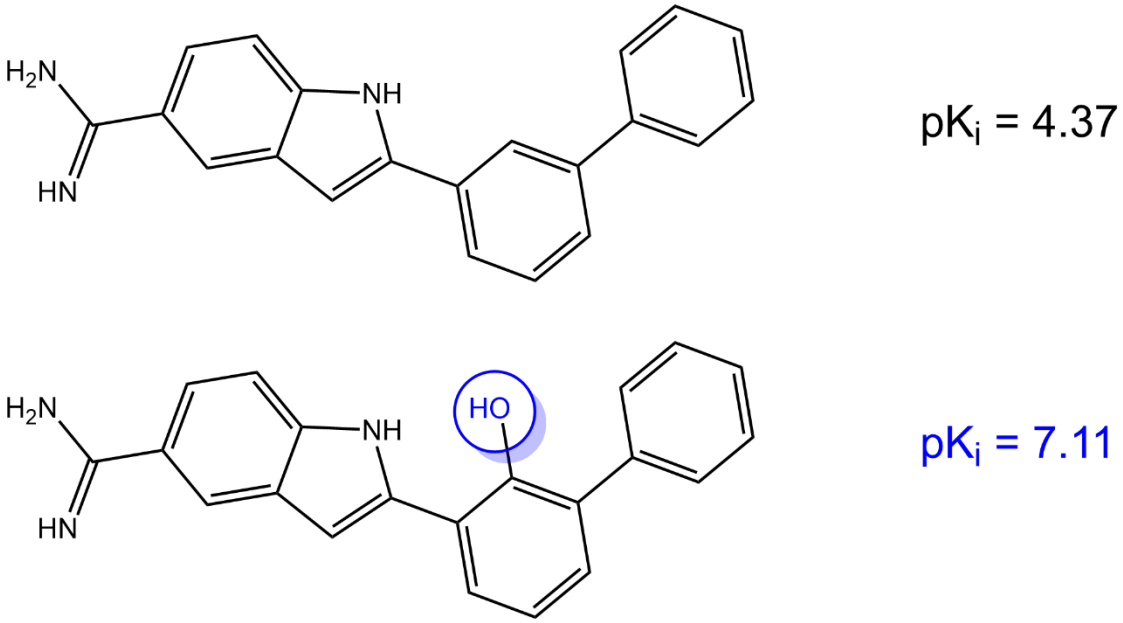

Figure 1. Structures of an activity cliff example for blood coagulation facto Xa inhibitors. [6]

Several quantitative approaches have been proposed to study ACs. Structure-Activity Similarity (SAR) maps were one of the earliest methods to identify activity cliffs in data sets by plotting molecular similarities in the X-axis and difference in potency the Y-axis of all possible pairs in the set.[7] In SAR maps the points located in the upper right quadrant (large similarity, large activity difference) will correspond to activity cliffs and points in the lower right quadrant to a smooth landscape (large similarity, low activity difference).[7,8] The Structure-Activity Landscape Index (SALI), shown on equation 1, is a straightforward method that generates a SALI matrix were the activity cliff can be identified by finding the largest values.[9] Another index is the Structure-Activity Relationship Index (SARI) that uses a potency weighted and average potency differences for pairs to identify continuous, discontinuous, and heterogeneous SARs. [10]

$$SALI(i,j) = \frac{|P_i - P_j|}{1 - sim(i,j)} \quad [1]$$

All these well stablished methods require pairwise comparisons between all the molecules in the set, but this can quickly become very computationally demanding for large

datasets. Recently, eSALI, equation 2, was introduced to alleviate this problem.[11] eSALI uses the extended similarity (eSIM) framework to obtain an index that quantifies the similarity of the whole set in a linear scaling way[12,13] and the differences between each property and the average. In this way the calculation of eSALI will scale as O(*N*), providing a fast and computationally cheaper way of quantifying the activity landscape roughness. [11]

$$eSALI(i) = \frac{1}{N}\sum\frac{|P_i-\bar{P}|}{1-s_e} \quad [1]$$

The upswing of Machine Learning has also impacted the QSAR modeling. Since Machine Learning QSAR models are dependent on the training data, those models will be highly affected by the activity landscape of the data, and can often deliver underwhelming results.[4,14] Despite the detriment of ML due activity cliffs presence in the data, the issue has been understudied. Up until recently, the Roughness Index (ROGI) was proposed to quantify the roughness of the activity landscapes in data sets by monitoring the loss in dispersion when clustering with increasing thresholds. ROGI was found to correlate with the error of ML models, supporting that high presence of activity cliffs will result in less accurate ML predictions. [14] A recent study benchmarked the performance of several ML and Deep Learning (DL) algorithms on activity cliffs and tried to find new direction on how to address the issue. AC-heavy sets were clustered and then stratified split (based on if the molecule was part of an AC) into test and train sets. Results concluded that both ML and DL models struggle with ACs, even after doing the above-mentioned splitting, and to be highly dataset dependent; making it not possible to give a consensus on the AC effect on the model types. [15]

In this work we use eSIM indices and eSALI frameworks to explore the effect of the data train/test split affects in the performance in traditional Machine Learning bioactivity regression models. The splitting approaches evaluate splits based on chemical space (only structural) and activity landscape regions and are compared to the traditional randomized splitting.

Methods

Datasets

We used CHEMBL datasets for 30 different molecular targets. Curated data was sourced from van Tilborg et al. [15] Each set consisted of SMILES string had an associated K_i or EC_{50} value. For

each molecule MACCS and ECFP4 (bond radius = 2) binary fingerprints were computed using RDKIT. [16]

Data splitting methods

Several train/test data splitting methods were tested. Most of them are based on the eSIM[12,13] and eSALI frameworks.[11] A few concepts of eSIM and eSALI need to be explained for better understanding of the data splitting methods.

Extended similarity performs the column-wise summation of the fingerprints, $\sum = [\sigma_1, \sigma_2, \dots, \sigma_M]$. To then classify each column as a 1-similarity, 0-similarity or dissimilarity counter based on a threshold γ. Posterior to the classification any similarity index (i.e. Jaccard-Tanimoto, Russell-Rao, etc.) can be transformed using a weighting function, with $|2\sigma_k - N|$ as independent variable, this will take into count *partial* similarity or dissimilarity of the columns. All mentioned steps scale O(*N*), making eSIM a fast quantitative measure of the similarity of the whole set.[12,13] If one molecule is taken out from the set and eSIM is applied, the yielded value will correspond to the complementary similarity of the taken molecule. This step can be done rapidly; we only need to subtract the fingerprint of the molecule from the column-wise vector that was calculated in the first step. The complementary similarity will give information on how similar that molecule to the rest of the set was, low-complementary eSIM molecules will be medoids of the set and high- complementary eSIM outliers.[17] Molecules can be sorted based on the complementary similarity value, making it possible to explore the chemical space in many different ways. [18] eSIM can be used as a loss function (maximize or minimize, depending on the task) to pick molecules to increase or decrease diversity starting from an initial molecule. When minimizing, we perform what we call diversity selection. [13]

The eSALI corresponds to a quantitative measurement of the activity landscape roughness of a whole set.[11] However, it can also be used as a loss function to pick molecule from a set. When maximizing eSALI, presence of activity cliffs will be favored. The algorithm will start with one desired molecule and then from the remaining molecules will add the one that yields the highest eSALI value. In cases where ACs are not wanted, eSALI should be minimized, what we call anti-eSALI.

The following eight selection methods were used to pick 20% of the molecules in each set as test set. In all methods that required similarity calculations the eSIM version of the Jaccard-Tanimoto[19,20] (JT) index was used.

- Random: randomized selection using the random[21] python module.
- Medoid: selecting the molecules with the lowest complementary eSIM.
- Uniform: sorting molecules by increasing eSIM, then separating them into five batches. In their original batch order, molecules are picked alternating batches, starting from the lowest complementary eSIM batch to the highest.
- Diverse: medoid (highest complementary eSIM molecule) is identified, then a screen is done to add the molecule that will give the smallest eSIM value. Repeated until desired number of molecules is selected.
- Kennard-Stone[22]: same process as diverse selection but starting from the two furthest points away from each other.
- eSALI: starting from the medoid, then a screen is done to add the molecule that will give the largest eSALI value. Repeated until desired number of molecules is selected.
- Anti-eSALI: starting from the medoid, then a screen is done to add the molecule that will give the smallest eSALI value. Repeated until desired number of molecules is selected.
- bSALI: separated into five batches based on eSALI method, then the molecules from each batch are resorted taking one molecule from each batch in the order they are in the batch.

The remaining 80% was used as a pool to pick the training set. Several training sets were picked to observe the effect of size of it. Training sets were picked in increases of 10% starting at 10%, picks were done with all the methods for all remaining pool after the test set, having all the possible methods combinations.

These sampling techniques can be found in: https://github.com/mqcomplab/MultipleComparisons.

Modeling

With all the possible combinations of train/test splits explained, five classic ML algorithms were used to build regression models: *Random Forests (RF)*[23], *k-Nearest Neighbors*

(kNN)[24], *Support Vector Machines (SVM)*[25], *Gradient Boosting Machine (GBM)*[26] *and Multiple Layer Perceptron Neural Networks (MLP)*[27]. Models were built with MACCS[28] and ECFP4[29] representations, except for MLP, where only ECFP4 was used. All models were constructed with the *scikit-learn*[30] Python module. The used hyperparameters for each case were the ones reported by van Tilborg et al.[15] Model accuracy was evaluated based on Root Mean Squared Error (RMSE)[31].

Results and discussion

On first instance, we analyzed the results of the models trained on data splitting using the same selection methods for the test and training sets. On Figure 2, we can see the average RMSE of all the models that used that selection methods. It can be appreciated how the eSALI method has the highest errors at any training set size, this makes sense since the test set is enriched in ACs. As naturally though, for all the methods the RMSE decreases with the size of the training set, except for the anti-eSALI method. This method is going to have a very low population of activity cliffs in the test set, then from the remaining pool at small percentages the presence of ACs in the training set will be also small. As the training set size increases the more ACs will be present because of the nature of the anti-eSALI selection method, leading to higher errors.

The diversity selection method is the second that performs the worst overall. The test set in this case will be a highly diverse set, that is hard to predict with low training data, and still a hard task when more training data is available. The same trend is observed with Kennard-Stone method, which is just a variant of diversity selection with different initialization. Medoid selection has a similar performance to random selection at small training data sizes, choosing the medoids (molecules that are the most similar to the rest of the set) benefits in this scenario. When the full rest of the data the performance is decent, the fourth lowest RMSE average. The bSALI method also shows promising results for when all the rest of the data is used for training; the selection method assures a uniform distribution of ACs in the data splitting. Similarly, uniform has this behavior, but in this case assuring a uniform distribution of the molecules from the different sectors of the chemical space. Despite all of this, the random data splitting remains very robust, being only surpassed in the low percentages cases by the anti-eSALI (with this changing after 50% of the available data is used). However, the excellent performance of anti-eSALI below 50% indicates that this sampling method could be useful in the low-data scenarios that are

relatively common in drug-design applications.[32,33] Additionally, the good performance of the uniform sampling can be seen as an alternative to random sampling, in cases where a deterministic separation between training and test sets is needed (e.g., in order to guarantee better reproducibility).

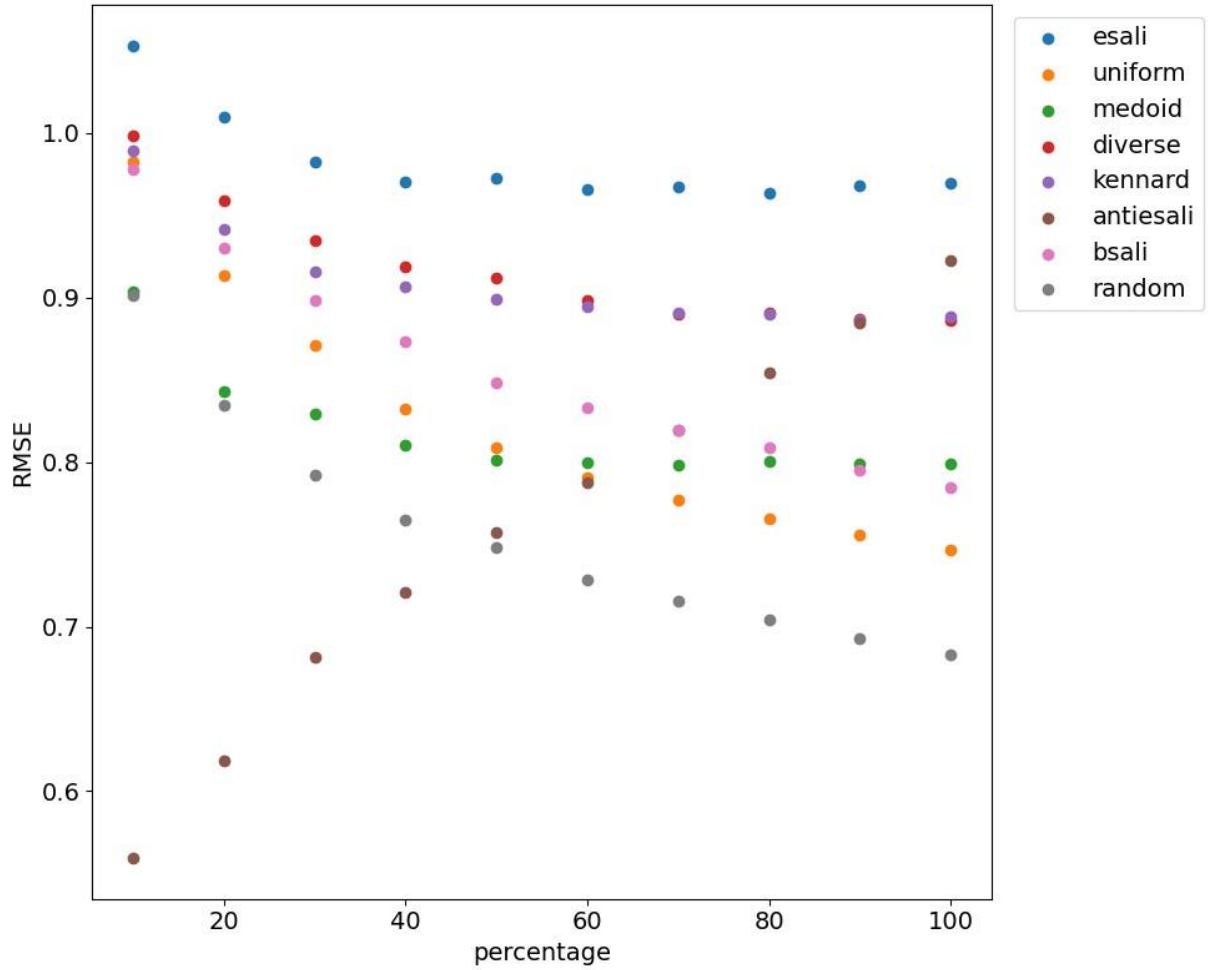

Figure 2. Average RMSE by selection method of the bioactivity ML models against percentage (of the 80% pool) used as a training set on the predictions of the test set (20%) selected by the same method.

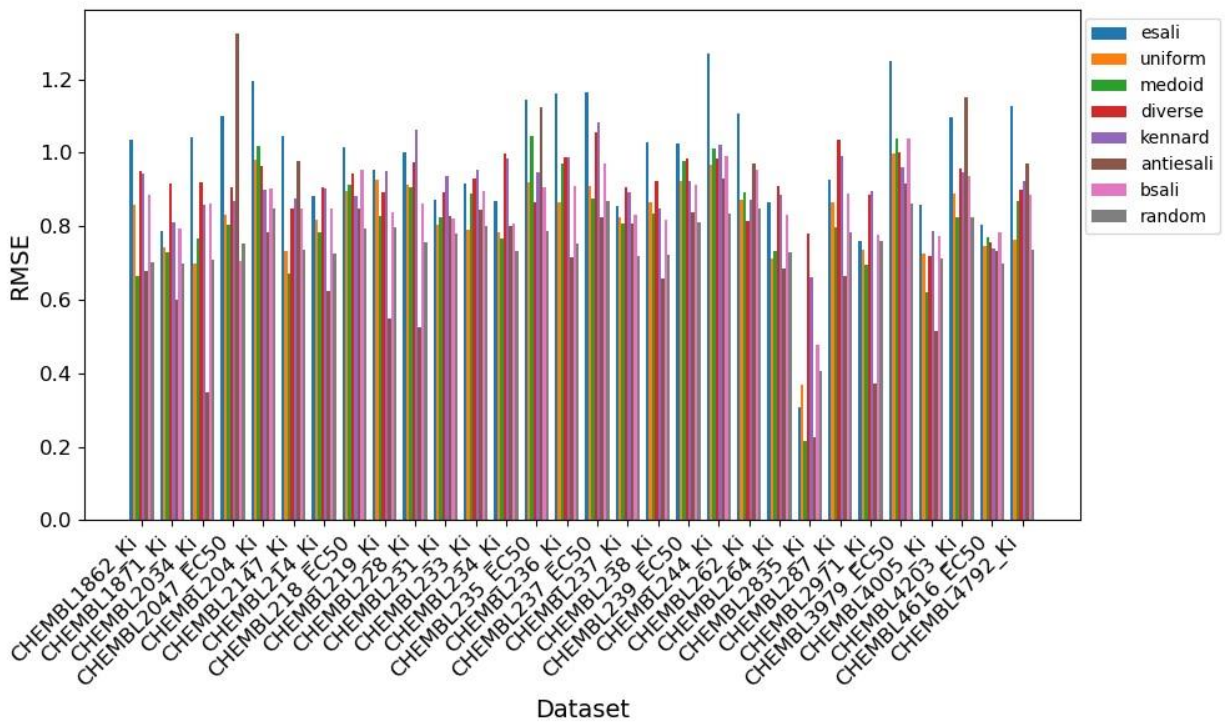

Figure 3. Average RMSE by database and using the same splitting method for the training set and test set.

Figure 3 shows the average RMSE by database and splitting method. In most databases the average trend, observed in Figure 1, is preserved; medoid, anti-eSALI and random are the methods with lowest RMSEs. Particular observations, like in CHEMBL2835_Ki and CHEMBL4203_Ki, support the idea that ML modeling is a case-to-case problem; these datasets show different results. In the first case we see that medoid and anti-eSALI have significant lower errors than random selection; in the latter case, the anti-eSALI method is the worst performing one. Other studies have also concluded that the best performing data splitting algorithm is data dependent and highly influenced by the data size[34]; which also comes to an agreement since CHEM2835_Ki is the smallest dataset used. The average RMSE by model is shown in Figure 4. The trends remain almost the same across the five ML algorithms, average RMSE values almost do not change with the algorithm. Once again, anti-eSALI and random provide the better results, while eSALI, diverse, and Kennard-Stone result in larger errors across the board.

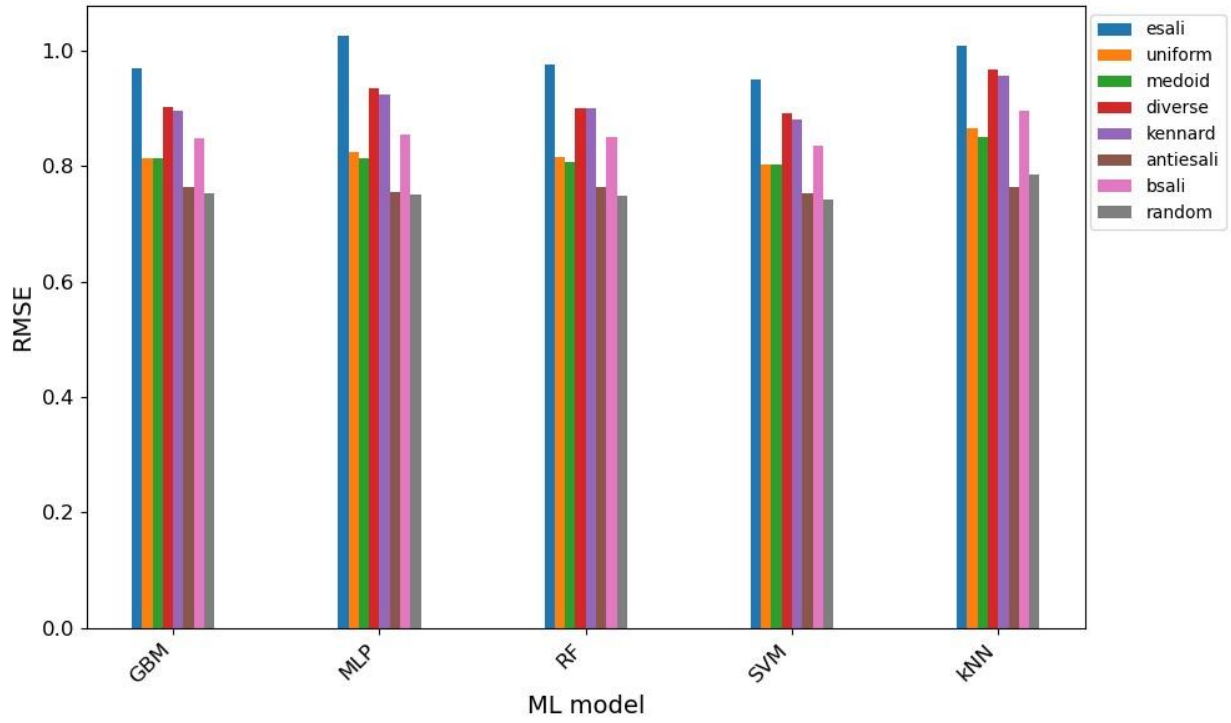

Figure 4. Average RMSE by Machine Learning model and same splitting method for the training set and test set.

It is important to mention that the proposed splitting methods will be highly influenced by representation, the similarity values can change from fingerprint to fingerprint. To analyze this, Figure 5 shows the average RMSE by fingerprint representation. In these particular cases the representation does not have a very high influence on the errors, one possible explanation might be that both correspond to binary fingerprints. In other works, differences between binary fingerprints and continuous descriptors were found in ML errors, but not between binaries (MACCS and ECFP). [15] Based on these reports, we anticipate that other representations will have more influence on the results[14,15], with the proposed selection methods being potentially dependent on the selected representation.

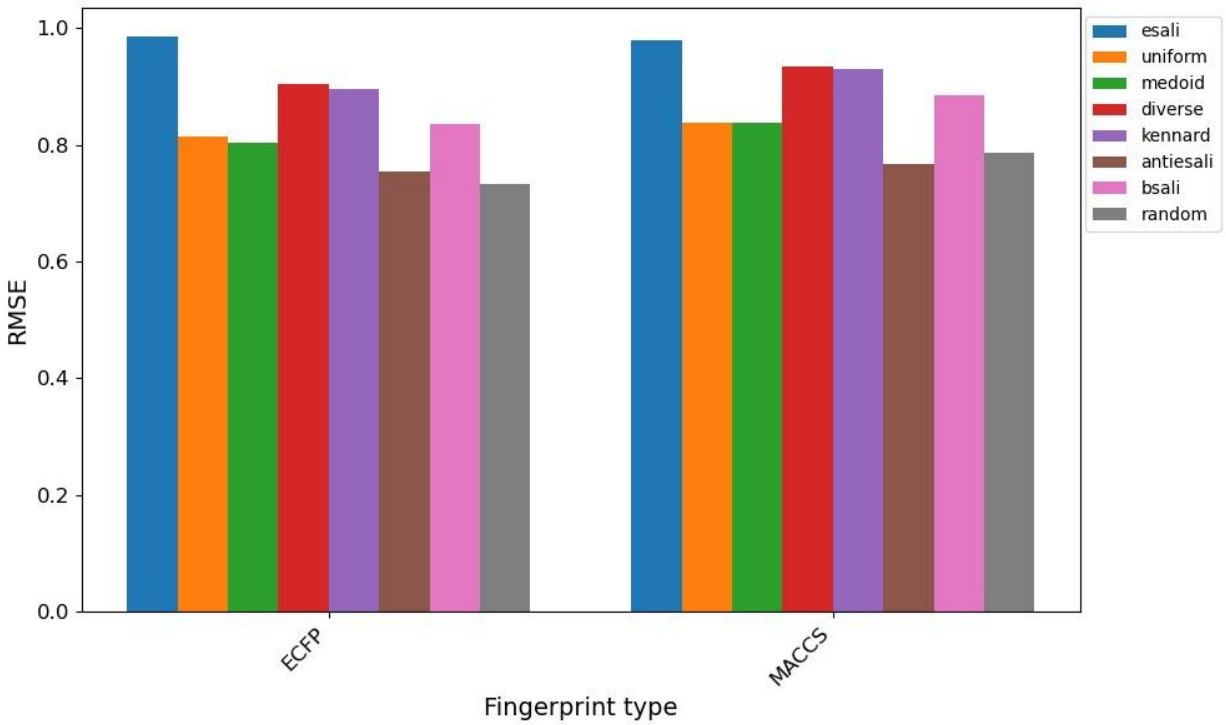

Figure 5. Average RMSE by fingerprint type and same splitting method for the training set and test set.

To further analyze the effect of ACs in the ML models, mixes between test and train selection methods were done (Fig. 6 shows the mean RMSEs). Literature has reported algorithms that show slight improvement in data splitting compared to random in RF models, the key to the reported method is the ability to match the distribution of the whole data and between the subsamples.[35] As such, the diagonal of Figure 6 has lower RMSEs, since both subsets were selected by the same method, the distributions might be similar. This underlies a key principle at the time of preparing a training set: the closer the distribution of species between the training and test sets, the better the performance of the model.

We can see that when selecting the test set with eSALI, in theory the test set with the roughest activity landscape, the rest of methods struggle overall. Even though anti-eSALI-selected test sets have the smoothest activity landscapes, this would leave a very rough landscape for the training set, provoking not so good models with the other selection methods. Models with bSALI-selected test sets have lower RMSEs than the other two eSALI based methods, this

supports the idea that is important to have similar roughness both on the test and train set.[15] By looking at the results, the models have lower errors when the test set is chosen randomly or with uniform selection. Another important note is that anti-eSALI selected training set methods have the highest errors, no matter what the test set method was. This means that the presence of ACs in the training set is crucial for the model performance. With this mixing exercise, it can be concluded that the random, uniform, and bSALI splitting still yield better models on average than any of the other methods. That is, having diversity of the different regions of the chemical space on both test and training is key to training a good model.

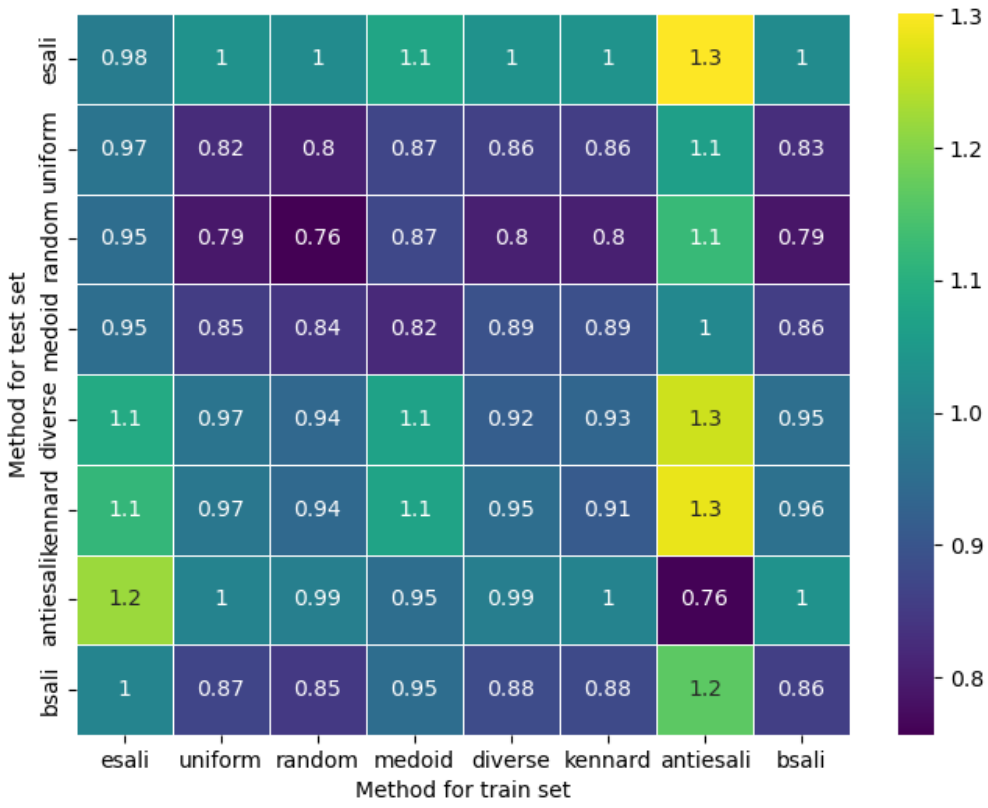

Figure 5. Average RMSE by splitting method used for training set (of the 80%) and method for test set (20%).

Conclusions

Activity cliffs are a challenging hurdle to overcome in ML bioactivity modeling. With the presented data splitting algorithms, we explored the behavior of typical ML algorithms in the presence of ACs. Presented models, on average, do not display many differences the errors with the different ML algorithms nor binary fingerprint type, the data splitting method shows a higher influence. High disparity in the number of ACs in the test/train sets lead to higher error in the bioactivity predictions overall. Random splitting performed very well overall, across different representations, ML methods, and datasets. Similarly, the uniform and bSALI methods show better performance than the others when choosing the training sets with them in cases where the test set was chosen with other methods, recalling the importance of representativity of regions of chemical space and activity roughness. The case-by-case nature of the activity landscape influence on ML models does not allow to generalize a better splitting method, at least of the ones in this study. More approaches need to be developed to study possible directions on how to manage ACs in ML.

Acknowledgements

We thank the National Institute of General Medical Sciences of the National Institutes of Health for support under award number R35GM150620.

References

(1) Silipo, C.; Vittoria, A. QSAR, Rational Approaches to the Design of Bioactive Compounds. In *European Symposium on Quantitative Structure-Activity Relationships 1990: Sorrento, Italy)*; 1991.

(2) Stumpfe, D.; Hu, Y.; Dimova, D.; Bajorath, J. Recent Progress in Understanding Activity Cliffs and Their Utility in Medicinal Chemistry. *J Med Chem* **2014**, *57* (1), 18–28. https://doi.org/10.1021/jm401120g.

(3) Johnson, M. A.; Maggiora, G. M.; others. *Concepts and Applications of Molecular Similarity*, 1st ed.; Wiley-Interscience, 1990.

(4) Maggiora, G. M. On Outliers and Activity CliffsWhy QSAR Often Disappoints. *J Chem Inf Model* **2006**, *46* (4), 1535–1535. https://doi.org/10.1021/ci060117s.

(5) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *J Med Chem* **2012**, *55* (7), 2932–2942. https://doi.org/10.1021/jm201706b.

(6) Dablander, M.; Hanser, T.; Lambiotte, R.; Morris, G. M. Exploring QSAR Models for Activity-Cliff Prediction. *J Cheminform* **2023**, *15* (1), 47. https://doi.org/10.1186/s13321-023-00708-w.

(7) Shanmugasundaram, V.; Maggiora, G. M. Characterizing Property and Activity Landscapes Using an Information-Theoretic Approach. In *ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY*; 2001; Vol. 222, pp U271–U271.

(8) Medina-Franco, J. L. Scanning Structure–Activity Relationships with Structure–Activity Similarity and Related Maps: From Consensus Activity Cliffs to Selectivity Switches. *J Chem Inf Model* **2012**, *52* (10), 2485–2493. https://doi.org/10.1021/ci300362x.

(9) Guha, R.; Van Drie, J. H. Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J Chem Inf Model* **2008**, *48* (3), 646–658. https://doi.org/10.1021/ci7004093.

(10) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure−Activity Relationships. *J Med Chem* **2007**, *50* (23), 5571–5578. https://doi.org/10.1021/jm0705713.

(11) Dunn, T. B.; López-López, E.; Kim, T. D.; Medina-Franco, J. L.; Miranda-Quintana, R. A. Exploring Activity Landscapes with Extended Similarity: Is Tanimoto Enough? *Mol Inform* **2023**, *42* (7). https://doi.org/10.1002/minf.202300056.

(12) Miranda-Quintana, R. A.; Bajusz, D.; Rácz, A.; Héberger, K. Extended Similarity Indices: The Benefits of Comparing More than Two Objects Simultaneously. Part 1: Theory and Characteristics†. *J Cheminform* **2021**, *13* (1), 32. https://doi.org/10.1186/s13321-021-00505-3.

(13) Miranda-Quintana, R. A.; Rácz, A.; Bajusz, D.; Héberger, K. Extended Similarity Indices: The Benefits of Comparing More than Two Objects Simultaneously. Part 2: Speed, Consistency, Diversity Selection. *J Cheminform* **2021**, *13* (1), 33. https://doi.org/10.1186/s13321-021-00504-4.

(14) Aldeghi, M.; Graff, D. E.; Frey, N.; Morrone, J. A.; Pyzer-Knapp, E. O.; Jordan, K. E.; Coley, C. W. Roughness of Molecular Property Landscapes and Its Impact on Modellability. *J Chem Inf Model* **2022**, *62* (19), 4660–4671. https://doi.org/10.1021/acs.jcim.2c00903.

(15) van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the Limitations of Molecular Machine Learning with Activity Cliffs. *J Chem Inf Model* **2022**, *62* (23), 5938–5951. https://doi.org/10.1021/acs.jcim.2c01073.

(16) RDKit. *RDKit: Open-source cheminformatics. https://www.rdkit.org*. https://www.rdkit.org.

(17) Chang, L.; Perez, A.; Miranda-Quintana, R. A. Improving the Analysis of Biological Ensembles through Extended Similarity Measures. *Physical Chemistry Chemical Physics* **2022**, *24* (1), 444–451. https://doi.org/10.1039/D1CP04019G.

(18) López-Pérez, K.; López-López, E.; Medina-Franco, J. L.; Miranda-Quintana, R. A. Sampling and Mapping Chemical Space with Extended Similarity Indices. *Molecules* **2023**, *28* (17), 6333. https://doi.org/10.3390/molecules28176333.

(19) Jaccard, P. THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE. *New Phytologist* **1912**, *11* (2), 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

(20) Rogers, D. J.; Tanimoto, T. T. A Computer Program for Classifying Plants. *Science (1979)* **1960**, *132* (3434), 1115–1118. https://doi.org/10.1126/science.132.3434.1115.

(21) Van Rossum, G. The Python Standard Library: Random. Python Software Foundation 2020.

(22) Kennard, R. W.; Stone, L. A. Computer Aided Design of Experiments. *Technometrics* **1969**, *11* (1), 137. https://doi.org/10.2307/1266770.

(23) Louppe, G. Understanding Random Forests: From Theory to Practice. **2014**.

(24) Kramer, O. K-Nearest Neighbors; 2013; pp 13–23. https://doi.org/10.1007/978-3-642-38652-7_2.

(25) Boyle, B. H. *Support Vector Machines : Data Analysis, Machine Learning, and Applications*; Nova Science Publishers, 2011.

(26) Natekin, A.; Knoll, A. Gradient Boosting Machines, a Tutorial. *Front Neurorobot* **2013**, *7*. https://doi.org/10.3389/fnbot.2013.00021.

(27) Fullér, R. Artificial Neural Networks. *Introduction to Neuro-Fuzzy Systems* **2000**, 133–170. https://doi.org/10.1007/978-3-7908-1852-9_2.

(28) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J Chem Inf Comput Sci* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.

(29) Glem, R. C.; Bender, A.; Arnby, C. H.; Carlsson, L.; Boyer, S.; Smith, J. Circular Fingerprints: Flexible Molecular Descriptors with Applications from Physical Chemistry to ADME. *IDrugs* **2006**, *9 3*, 199–204.

(30) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; others. Scikit-Learn: Machine Learning in Python. *the Journal of machine Learning research* **2011**, *12*, 2825–2830.

(31) Karunasingha, D. S. K. Root Mean Square Error or Mean Absolute Error? Use Their Ratio as Well. *Inf Sci (N Y)* **2022**, *585*, 609–629. https://doi.org/10.1016/j.ins.2021.11.036.

(32) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug Discovery with Explainable Artificial Intelligence. *Nat Mach Intell* **2020**, *2* (10), 573–584. https://doi.org/10.1038/s42256-020-00236-4.

(33) Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generative Molecular Design in Low Data Regimes. *Nat Mach Intell* **2020**, *2* (3), 171–180. https://doi.org/10.1038/s42256-020-0160-y.

(34) Birba, D. E. A Comparative Study of Data Splitting Algorithms for Machine Learning Model Selection. KTH, School of Electrical Engineering and Computer Science (EECS). 2020.

(35) Joseph, V. R.; Vakayil, A. SPlit: An Optimal Method for Data Splitting. *Technometrics* **2022**, *64* (2), 166–176. https://doi.org/10.1080/00401706.2021.1921037.