EXPLAINING STRUCTURE–ACTIVITY RELATIONSHIPS USING LOCALLY FAITHFUL SURROGATE MODELS

A PREPRINT

Heta A. Gandhi
Department of Chemical Engineering
University of Rochester
Rochester, NY, 14627
hgandhi@ur.rochester.edu

Andrew D. White*
Department of Chemical Engineering
University of Rochester
Rochester, NY, 14627
andrew.white@rochester.edu

ABSTRACT

We present a model-agnostic method that gives structure-activity explanations of black-box models. Machine learning models are now common for molecular property prediction and chemical design. They typically are black boxes – having no explanation for predictions. Our method uses surrogate models to attribute predictions to chemical descriptors and molecular substructures, independent of the black box model inputs. Our approach provides explanations consistent with chemical reasoning, like connecting existence of a functional group or molecular polarity.

1 Introduction

Understanding the link between chemical or biological activity and molecular structure is central to aspects of drug discovery and medicinal chemistry.[1] Quantitative structure–activity relationship (QSAR) modeling aims to model the variations in biological or pharmacokinetic properties caused by a variation in structural properties. As a result, QSAR modeling has been applied across disciplines to comprehend, rationalize and predict biological activity and physicochemical properties of molecules.[2–4] Some specific use cases include chemical property prediction,[5] computer-aided drug design,[6] and lead optimization.[7]

In recent years, deep learning (DL) methods have gained popularity for QSAR modeling.[8–11] While these methods may be highly accurate in their predictions, most DL models are black box functions that provide little explanation or scientific insight for their predictions.[12–17] For many of the chemistry and biochemistry applications, especially in healthcare and drug discovery, predictions from DL models may be used to make high-stakes decisions.[18–20] Thus, it is crucial that prediction accuracy comes from learning relevant relationships between data features rather than from picking up potential biases in the data, also known as the so-called Clever Hans effect.[21] For example, Chuang and Keiser[22] found spurious correlations in a black box model used to predict C-N cross-coupling reaction yields. Understanding what the model is learning and what factors impact model predictions assists in avoiding the Clever Hans effect by detecting model bias and in determining whether to trust the predictions while making decisions.

Explainable artificial intelligence (XAI) has emerged as a field to better comprehend what DL models learn and to gain scientific insight into model predictions.[23,24] Two broad approaches are typically used for model interpretability — intrinsic interpretability and post-hoc approaches.[25] Intrinsic interpretability comes from using models that are considered inherently interpretable or self-explaining. These models are generally simple, and model weights can be used to draw relationships between model outputs and input features. Examples of interpretable models are linear models, decision trees, k-nearest neighbors. However, as the complexity of models increases, they typically become less interpretable. Post-hoc methods are instead applied as an extra step after model training to explain predictions.

Interpretability methods and model interpretations may be categorized in multiple ways.[14,25] One categorization is based on whether the method is model-specific or model-agnostic. Model-specific methods are applicable only

*Corresponding Author

to certain models for which they are devised. Self-explaining methods are always model-specific. Model-agnostic methods can be applied to any deep learning model, generally in a post-hoc fashion.

The scope of model interpretation can either be global or local. Global interpretations focus on general model decisions and provide insight into how the model learns. Miller[26] defines such interpretability as "the degree to which an observer can understand the cause of a decision." Global interpretations provide a generic understanding of the model, while local explanations capture input-output relationships that may not be apparent from global trends. By Miller's definition, an explanation is "a presentation of information intended for humans that gives the context and cause for an outcome."

Here, we focus on developing a post-hoc model-agnostic local explanation method. The desire for post-hoc is because self-explaining models cannot compete with deep learning and other black box methods in accuracy. The motivation for local explanations is both because of their better agreement with the model being explained and because of the well-known activity cliffs in structure–activity relationships (SAR).[27–29] An activity cliff is the often observed effect of SARs breaking down as a model leaves one region of chemical space. Since our desire in this work is explanations rooted in SARs, we focus on local explanations to avoid activity cliffs.

Commonly used local explanation approaches include counterfactual analysis, feature importance, training data importance, and surrogate models.[30] Polishchuk[31] and Rodríguez-Pérez and Bajorath[32] provide a good review on application of these approaches in chemical property prediction and QSAR modeling. Humer et al.[33] compare visualization based XAI methods to get per atom attributions in their interpretability visualization tool called cheminformatics model explorer (CIME). Jiménez-Luna et al.[24] discuss various XAI methods, as applicable to drug discovery.

Counterfactual analysis are instance-based approaches that rely on creating counter examples for molecules of interest. Counterfactuals of a prediction are similar molecules with a different outcome.[34,35] Figure 1a shows an example of counterfactuals. Although counterfactuals are intuitive and provide insight into chemical predictions, these explanations are not complete since they cannot quantify the effect of a structural change on the prediction. An expert needs to examine multiple counterfactual molecules to deduce a SAR. Contrastive explanations are a similar approach, except the counterexamples give information of pertinent or missing features that influence predictions positively and negatively.[36–38] An example is seen in Figure 1b where pertinent negative features for triphenylphosphine oxide are shown. Contrastive explanations, like counterfactuals, provide chemical intuition but do not provide a quantitative description.

Feature importance or feature attribution methods assign a numerical score to each input feature to indicate how important it is for the prediction. Figure 1c(i) shows a visualization technique called similarity maps that use fingerprint similarity to compare chemical structures and highlight atomic contributions to DL predictions. Rasmussen et al.[42] recently developed benchmarks for visualization-based feature attribution methods. Gradient-based feature attribution techniques and layerwise relevance propagation (LRP) are most frequently used to explain predictions by assigning feature importance.[39,41,43,44] McCloskey et al.[39] use integrated gradients for substructure attribution to understand protein-ligand binding, see Figure 1b(ii). Jiménez-Luna et al.[41] proposed a graph architecture to get attribution scores for molecular features using integrated gradients, shown in Figure 1e. Payne et al.[45] use an attention-based transformer model to get per atom contributions. Feature attributions can provide valuable model insight, but they are only partial explanations because it is difficult to act on them (know how to modify structure to get a different outcome) and connect to an underlying structure–activity relationship[46].

Surrogate models have been widely used to explain model predictions. Locally interpretable model-agnostic explanations (LIME) uses a surrogate interpretable model to approximate the black box function and provides per-instance explanations by perturbing the input features of that instance.[47] They have been seen in chemistry too. For example, Whitmore et al.[48] show a model-specific application of LIME to interpret research octane number predictions coming from a random forest classifier. A related, popular approach is SHapley Additive exPlanations (SHAP)[49]. SHAP is a kernel-based approach that gives features contributions using Shapley value explanations. The concept of Shapley values originated in game theory to fairly distribute gains and costs among players depending on their contributions.[50] Rodríguez-Pérez and Bajorath[51] showed how SHAP can be used to generate local explanations for compound activity predictions. Wojtuch et al.[40] used SHAP to understand metabolic activity of compounds using Molecules are described using Molecular ACCess System (MACCS) fingerprints. Figure 1d is an example of feature importances extracted using SHAP. Although accurate and consistent, SHAP ignores feature dependence, can be computationally expensive because of the combinatorial scaling of coalitions, and result in feature attributions, thereby having the same drawbacks of unactionability and difficulty in connecting to chemical concepts.

Explanation methods do not necessarily provide contextual and scalable outcomes. Domain knowledge needs to be incorporated to make explanations contextual and usable. Counterfactuals, for example, are actionable and contextual since they provide exact changes that need to be made to a molecule to change its activity. They are agnostic to input

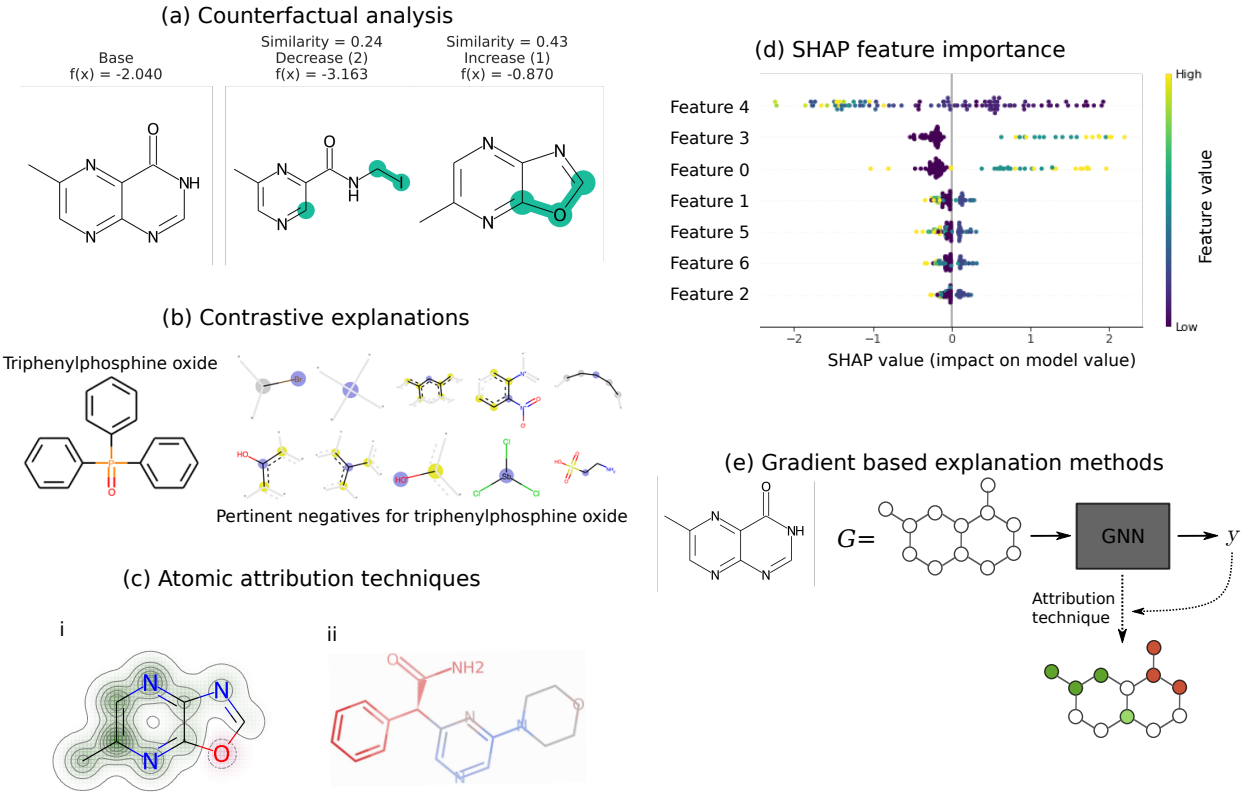

Figure 1: Different explanation methods from literature. (a) Counterfactuals give the smallest possible change that changes the activity[35] (b) Contrastive explanations identify the missing features that may influence the prediction, image taken from Lim et al.[37] (c)Atomic attribution techniques give scores for each contribution of atoms and subgroups[33,39] (d) SHAP uses Shapley values to give feature attributions[40] (e) Gradient based methods for graph attribution[41]

features. Feature attribution and weighting methods are limited by the original set of features or model inputs. This often hinders interpretability when input features are complex and do not incorporate chemistry knowledge.[24] We aim to develop an intuitive understanding of local SAR for chemical data by attributing descriptors that are independent of model features and use concepts that are of interest to users of molecular data.

In this paper, we present a post-hoc model-agnostic explanation method for providing locally faithful, meaningful quantitative explanations for predictions from DL models of molecules using domain ontology. We aim to develop an intuitive understanding of local SAR for chemical data by attributing descriptors, independent of model features. Molecules are represented using interpretable chemical fingerprints and Rdkit descriptors. Chemical fingerprints encode structural characteristics of molecules into a vector. In graph terminology, fingerprints are k-neighborhood subgraph counts[52–54]. A simple linear surrogate model based on LIME is used to get attributions for these descriptors. For perturbation of the input features, we use the Superfast Traversal, Optimization, Novelty, Exploration and Discovery (STONED) algorithm which allows for generation of chemically similar, valid molecules without the need for a pretrained generative model.[55] Molecules are described using MACCS fingerprints[56] (referred to as MACCSfps hereafter), Extended Connectivity FingerPrints (ECFP),[52] and Rdkit descriptors[57] since these have been widely used in chemistry and are interpretable to domain experts. They each serve a slightly different purpose: ECFP works well when a molecule can be broken into subgraphs. MACCS works well on small molecules or very large molecules that cannot be broken-up. Rdkit descriptors provide complementary "whole-molecule" information. We test the explanations for soundness, completeness and coherence.[58] The explanations obtained are quantitative and give insights into the influence of molecular substructures and descriptors on a model prediction, thereby giving structure–activity relationships.

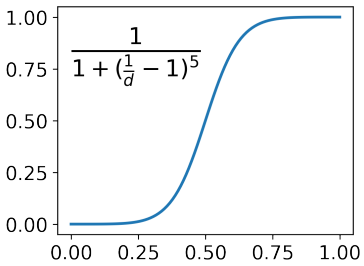

Figure 2: Conversion from Tanimoto similarity score to weights for regression. Shifted sigmoid curve generated using Equation 3

2 Methods

Our method compute QSARs for molecular structure properties, independent of features used for model predictions. We use LIME to compute these because it is locally faithful, can compute QSARs, and is model agnostic.

LIME is a model-agnostic, perturbation based method that aims to explain a specific model prediction using an interpretable surrogate model.[47] Let f be the original black box model to be explained and let g be the surrogate explanation model. Let $\vec{x}$ be the feature vector for a given instance. The objective of the local surrogate model is to fit the perturbed inputs around an instance $\vec{x}$ and corresponding model predictions from f, such that predictions from g match those from f closely. The explanation ξ for a given instance $\vec{x}$ is given by Equation 1.

$$\xi(\vec{x}) = \arg\min_{g \in G} \mathcal{L}(f, g, w) + \Omega(g) \tag{1}$$

Where the explanation model g minimizes the loss $\mathcal{L}$ which is a measure of how closely g approximates f. G is a class of interpretable models. w represents the similarity between $\vec{x}$ and it's perturbed input $\vec{x}'$, and $\Omega(g)$ is an optional parameter that controls complexity of g. Ω could be a regularization term, like L1 used in lasso or L2 used in ridge regression. We use a linear model fit with weighted least squares (WLS) regression[59] with a Tikhonov regularization[60] term for our surrogate model g, because linear models are self-explaining and have been shown to be comparable to sophisticated explanation strategies[61]. The regularization term is added to alleviate the problem of multicollinearity due to correlated features.

The weights represent distance from the instance we are trying to explain and are computed by Tanimoto similarity.[62] ECFP4 fingerprints[52] are used to calculate tanimoto similarity between the instance to be explained $\vec{x}$ and the points $\vec{x}'$. ECFP4 fingerprints capture the entire molecular structure and hence, ensure accurate comparison of molecular structures.

$$g(\vec{x}) = \boldsymbol{\beta}\mathbf{X}, \quad \boldsymbol{\beta} = (\mathbf{X^T W X} + \lambda \mathbf{I})^{-1} \mathbf{X^T W Y} \tag{2}$$

$$w = \frac{1}{1 + (\frac{1}{d(\vec{x}, \vec{x}')} - 1)^k} \tag{3}$$

Mathematically, WLS is given by Equation 2 where the regression coefficients β_i indicate how much $\hat{y}$ changes if feature $\vec{x}_i$ is changed while other features are kept constant. $\lambda\mathbf{I}$ is the Tikhonov regularization term. The features can be ranked using the regression coefficients by finding the t-statistic for each β_i. Equation 4 is used to calculate feature t-statistics. It is a ratio of β_i and standard error in β_i. In Equation 2, weighted tanimoto similarities are used for W. Tanimoto similarities are weighted using a shifted sigmoid function (Equation 3) so that molecules that are dissimilar to the base molecule are disregarded in determining the explanation. In Equation 3, $d(\vec{x}, \vec{x}')$ denotes tanimoto similarity between molecules represented by $\vec{x}$ and $\vec{x}'$ and k is a parameter that is used to adjust the slope of the curve. Figure 2 shows a plot of the shifted sigmoid curve. Figures S4, S5, S6 show how ignoring the regularization term and using unweighted tanimoto similarity as WLS weights affects the regression fit and descriptor explanations. For some molecules, considering dissimilar molecules doesn't affect the regression fit, but leads to misleading explanations.

$$t_i = \frac{\beta_i}{S_{\beta_i}}, \; S^2_{\beta_i} = \frac{1}{N - D} \sum_j \frac{(\hat{y}_j - y_j)^2}{(x_{ij} - \bar{x}_i)^2} \tag{4}$$

In Equation 4, N is the number of examples and D is the number of features. Standard error, S_{β_i}, is a ratio of the prediction accuracy to feature variance. Here, prediction accuracy refers to how closely $g(\vec{x})$ approximates $f(\vec{x})$. Finding the t-statistic removes sensitivity coming from units and magnitudes of the features.

Using LIME explanations allows the use of any interpretable representation of the inputs as features for the surrogate model.[47] The surrogate model's inputs do not have to be the same as features used to train the underlying model. This makes the explanations focused on the features of interest and accessible to domain-specific experts. To make our molecular explanation method model agnostic and widely applicable, we use MACCSfps, ECFP fingerprints, and RdKit chemical descriptors to represent molecules when generating explanations. MACCSfps are binary vectors that encode the presence of predefined substructures in a molecule.[56] They are fixed size vectors that contain a total of 166 keys. ECFP are binary vectors that encode instance-based substructures. It is also possible to obtain atomic contributions using ECFP descriptors, however this results in the loss of interpretability that substructures provide. While MACCSfps and ECFP account for structural characteristics, Rdkit descriptors constitute physical and chemical properties.

To get perturbed input features around a molecule of interest, we create a chemical space around the instance using the STONED algorithm.[55] STONED creates a chemical space by mutating the SELFIES representation of the instance being explained. SMILES strings are not used for this because mutations of a SMILES string do not always correspond to valid molecules. SELFIES (SELF-referencIng Embedded Strings), introduced by Krenn et al.[63], are surjective in nature and any mutation made to a SELFIES string gives a valid molecule. Hence, the resulting chemical space from STONED contains all valid molecules. However, the chemical stability and synthesizability is not guaranteed. Wellawatte et al.[35] utilized STONED to generate a chemical space that was used to identify counterfactuals in their method, called MMACE.

To test our method, we applied it to small molecule solubility prediction. Aqueous solubility is a key physicochemical property in drug design and development, since it has an impact on drug uptake and bioavailability.[64] Hence, many predictive models have been developed to predict solubility of molecules.[65–67] We use the AqSolDB database curated by Sorkun et al.[68] to build a DL model and then draw explanations for its predictions. AqSolDB contains 9982 small molecules along with their experimental aqueous solubility and has been of interest in developing several DL solubility prediction models.[69–72]

3 Results and Discussion

We evaluate the explanations obtained from our method. We investigate whether our model can recover known SARs. Next, we evaluate how well the linear regression fits the original model and if it can provide local explanations that match chemical intuition about SARs for real data. Finally, we check if the method is robust to the sampling method.

3.1 Can the method recover a known SAR?

To evaluate if our method can explain known SARs in the vicinity of a given instance, we used the same features for the model and explanation. A random forest (RF) regression model was trained using three calculated features for the AqSolDB dataset. These features were picked randomly from the list of ten Rdkit descriptors. The RF model was implemented in Scikit-learn[73] using 100 decision trees with a maximum tree depth of 10 and mean squared error as the loss function. The data was split using a 10% train/test split. A correlation coefficient of 0.82 was obtained (see Figure S2). Correlation is not expected to be high, since we are using few features. The described method was used to generate descriptor explanations for one of the molecules, and we check if the features used for training were recovered as important. Figure 3 shows the outcome of this analysis. Features in purple font were used to train the RF model and, as can be seen, they are reflected in the set of important descriptors calculated. Thus, our method can recover the known model features.

3.2 Does the method recover SAR for real data?

We use AqSolDB with another DL model to evaluate the SAR obtained. The DL model we use for this regression task is a gated recurrent unit (GRU) recurrent neural network (RNN) implemented in Keras.[75,76] Molecules are specified as SMILES[77] in the data. They are canonicalized and converted to SELFIES for model training. The model is trained for 100 epochs using the Adam optimizer[78] with a learning rate of 10^{-4}, and validated using early stopping. An 80%-10%-10% train-validation-test split is used. A correlation coefficient of 0.87 is obtained (see Figure S3), and the state-of-the-art is between 0.8 and 0.93.[66]

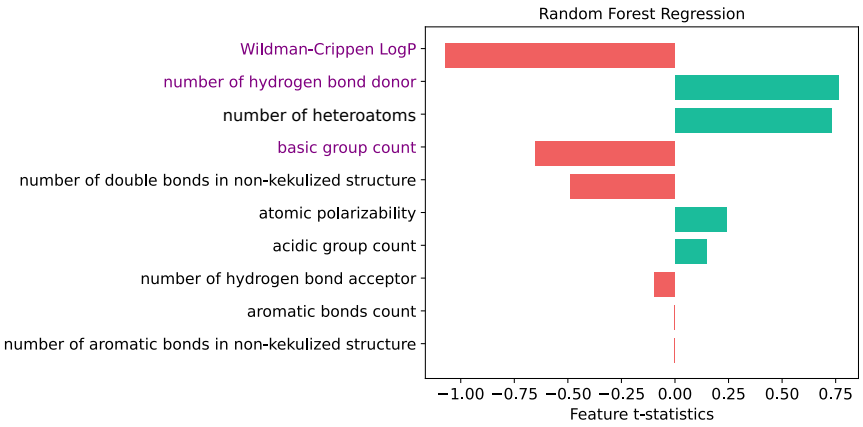

Figure 3: Descriptor explanations retrieve features used to train the model and weigh these higher than others, indicating that the XAI model is robust to training features. The descriptors highlighted in purple were used to train a random forest model, and green and red bars show descriptors that influence predictions positively and negatively, respectively. Wildman-Crippen LogP is a measure of hydrophobicity [74] and has an inverse relationship with aqueous solubility. Number of hetero atoms shows up among the important descriptors since it is correlated with the number of hydrogen bond donors.

To get the SAR for solubility data, we pick an instance to explain (referred to as 'base molecule') from AqSolDB, create a chemical space around that molecule and fit the WLS regression model to predictions for this space. Figure 4b shows the chemical space for a given base molecule. Distance in the chemical space denotes similarity to the base molecule, and the color indicates agreement between RNN predictions, $\hat{y}$ and regression model approximations, g. Notice that the regression is weighted to fit in the vicinity of the base molecule. We see that regression fit becomes poorer as we move away from the base molecule in chemical space, as desired by Equation 1. The parity plot (Figure 4a) shows the regression fit between $\hat{y}$ and g obtained for points in the chemical space, and color and transparency of the points denotes similarity to the base molecule. A correlation coefficient of 0.78 indicates a strong positive correlation between $\hat{y}$ and g, meaning we see a good agreement between the local model and the RNN prediction – the locally interpretable model is faithful.

Figure 5 shows the descriptor explanations for the base molecule shown. These attributions are calculated using Equation 4. The five highest t-statistic descriptors are shown. The yellow dotted lines indicate the significance threshold for the t-statistics. The significance threshold is set at 0.05, although this is somewhat arbitrary. Significance t-statistics help provide sparse explanations and quantify whether a descriptor is important or shows up as likely due to chance. Among the classic descriptors, acidic group count, basic group count, and number of hydrogen bond acceptors positively influence solubility predictions. By chemical intuition, acidic and basic groups as well as hydrogen bond donors and acceptors make molecules more polar, and polar compounds are more soluble in water. Wildman-Crippen LogP [79], and aromatic bonds count negatively influence the solubility. LogP is a measure of hydrophobicity and has an inverse relationship with aqueous solubility. [74] Increase in aromaticity has been shown to decrease aqueous solubility. [80,81] The classic descriptor explanations show that this approach is chemically intuitive.

The MACCSfps and ECFP explanations show which functional groups or substructures affect solubility of the molecule in water. Because these are local explanations, substructures or functional groups that show up as important are related to the base molecule and perturbations created around it. For example, in Figure 5, the pictured molecule is highly insoluble in water. MACCS descriptors suggest that multiple six member rings negatively influence the aqueous solubility. Adding heteroatoms to the ring increases solubility. This is intuitive, since addition of heteroatoms increases polarity. This is also alluded by the ECFP descriptors where substructures containing heteroatoms are shown to increase solubility. Substructure attributions, provided with statistical significance, give sparse structure–activity relationships that are locally valid.

3.3 Is the method sensitive to STONED parameters?

The STONED algorithm has a few parameters that affect the way chemical space is sampled. These parameters are number of mutations, choice of alphabet and size of chemical space. Depending on choice of parameters, chemical space creation varies. In Figure 5, the parameters are chemical space size of 2500 molecules created using the basic alphabet with up to two mutations to the base molecule. "Alphabet" implies the available tokens that may be utilized

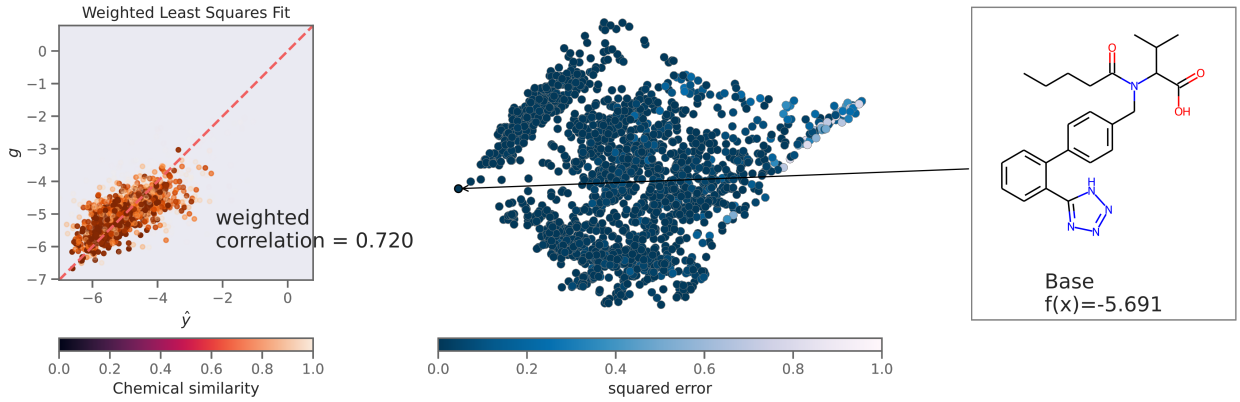

Figure 4: (a) Parity plot showing weighted least squares predictions against true values (black box predictions) and colored by chemical similarity from the base molecule. (b) Chemical space created by STONED around the base molecule, colored by the weighted least squares fit.

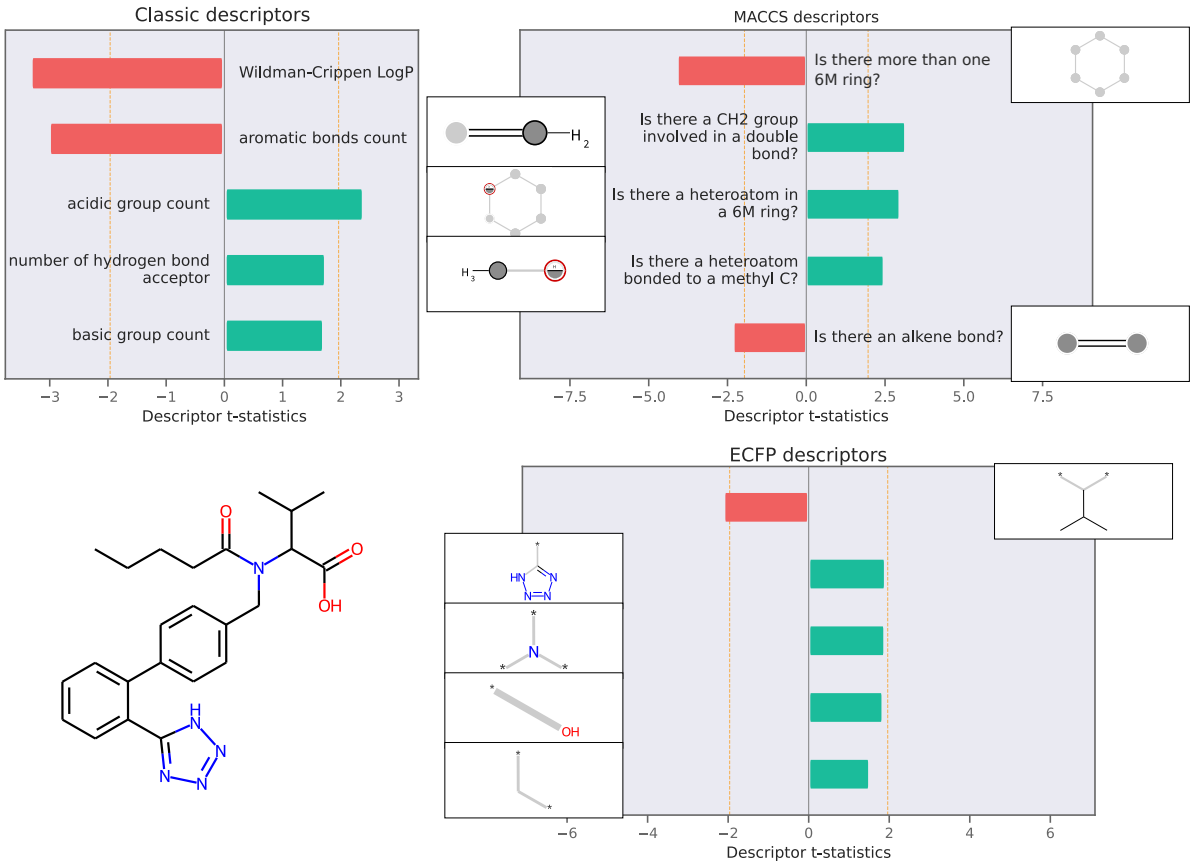

Figure 5: Descriptor t-statistics for the molecule pictured. The green and red bars show descriptors that influence predictions positively and negatively, respectively. Dotted yellow lines show significance threshold ($\alpha = 0.05$) for the t-statistic. SMARTS annotations for MACCS descriptors were created using SMARTSviewer (smartsview.zbh.uni-hamburg.de, Copyright: ZBH, Center for Bioinformatics Hamburg) developed by Schomburg et al.[82] The MACCS and ECFP descriptors indicate which substructures influence model predictions. MACCS substructures may either be present in the molecule as is or may represent a modification, and ECFP fingerprints are substructures in the molecule that affect the prediction.

for SELFIES modification in STONED. Basic alphabet restricts the available elements to B, C, N, O, S, F, Cl, Br, I. Other alphabet choices include "training data" and the SELFIES alphabet. Training data alphabet includes all unique SELFIES tokens present in the training data. 'SELFIES' alphabet includes all elements or tokens that are allowed in SELFIES representation.[63] The descriptor explanation method itself is insensitive to the choice of parameters. However, descriptor explanations depend on the chemical space and the choice of parameters affects the mutations created around the base molecule. For example, Wellawatte et al.[35] showed that increasing the number of SELFIES mutations leads to perturbed molecules being dissimilar from the base molecule. Hence, the choice of parameters should be governed by the kind of molecule mutations expected by the user. Figures 6,S7,S8 in the supplementary information shows the effect of these parameters on explanations. Notice how substructures that matter for prediction differ as parameters change.

3.4 Is the method robust to incomplete sampling of chemical space?

The chemical space sampled by STONED may not be complete and is sensitive to hyperparameters chosen of the method. We investigate the robustness of our method by varying chemical space size. To do this, we subsample chemical space of different sizes from a large reference chemical space sampled using STONED. The reference chemical space is sampled using two mutations, basic alphabet and a chemical space size of 7500. Descriptor explanations are calculated for each of the sampled subspaces and compared to the reference set of important descriptors using Spearman's rank-order correlation coefficient.[83] Spearman's rank-order correlation is a non-parametric measure of the monotonicity between two sets. Figure 7 shows the rank correlation for chemical subspaces as a function of increasing space size. For each size, ten chemical spaces are subsampled and the average of rank correlations found for those ten spaces is reported. The correlation between descriptor explanations of a subsampled chemical space and reference set (red dotted line in Figure 7) increases monotonically and then plateaus. Rank correlation shows how close the ranks of important descriptors in the subsampled set are to the reference set. For as low as 1000 perturbed examples, we see a rank correlation of 0.9. A high rank correlation coefficient indicates that descriptor ranks for the subsampled set and the reference set are positively correlated. We observe high correlation of ranks at chemical space size of 4000, and increasing the chemical space size beyond that doesn't change the ranks assigned to descriptors.

4 Conclusions

Machine learning models are becoming widespread in chemical and life science. It is important to understand whether these models behave as expected and provide valid predictions. The presented method is a descriptor explanation method that provides localized explanations of model predictions and quantifies the importance of certain functional groups or fragments present in the molecule. We demonstrated our method by applying it to AqSolDB. We recovered known structure-activity relationships and showed our method is robust. MACCSfps and ECFP are a set of substructures that provide insight into which parts of the molecule explain predictions, and Rdkit descriptors explain which chemical properties might be influencing predictions. Our method also provides a confidence threshold for explanations. These outcomes are intuitive, connect well to SAR and are easily interpretable by chemists. Counterfactuals are actionable explanations; however, they do not provide a quantitative view of the SAR. Descriptor explanations complement counterfactual explanations, as they provide quantitative SAR with significance statistics for important molecular substructures.

5 Data Availability

The code and data is available at https://github.com/ur-whitelab/exmol.

6 Acknowledgements

Research reported in this work is supported by the National Science Foundation under Grant number 1751471, and National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM137966. The authors thank the Center for Integrated Research Computing (CIRC) at the University of Rochester for providing computational resources and technical support.

References

[1] Rajarshi Guha. On Exploring Structure–Activity Relationships. In Sandhya Kortagere, editor, *In Silico Models for Drug Discovery. Methods in Molecular Biology*, volume 993, pages 81–94. Humana Press, Totowa, NJ, 2013.

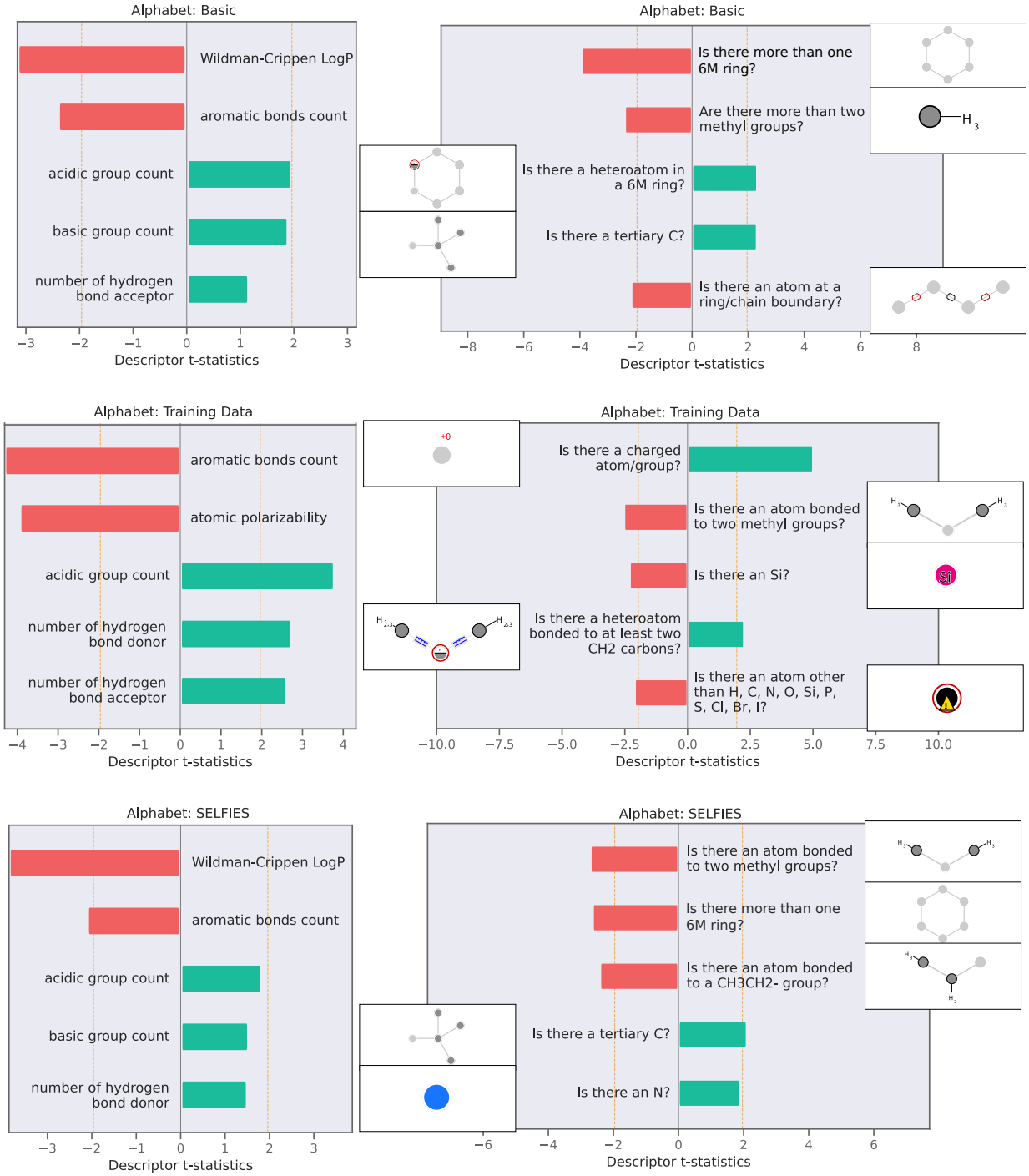

Figure 6: Descriptor explanations are insensitive to alphabet choice. However, the chemical space created around the base molecule varies with the choice of alphabet, and descriptor explanations generated depend on the chemical space. Choice of alphabet should be dictated by the kind of mutations that are of interest to the user. 'Basic' alphabet which restricts available elements to [B, C, N, O, S, F, Cl, Br, I]. 'Training data' alphabet includes all unique SELFIES tokens available in the training data examples. 'SELFIES' alphabet includes all elements that are allowed in SELFIES representation. SMARTS annotations for MACCS descriptors were created using SMARTSviewer (smartsview.zbh.uni-hamburg.de, Copyright: ZBH, Center for Bioinformatics Hamburg) developed by Schomburg et al.[82]

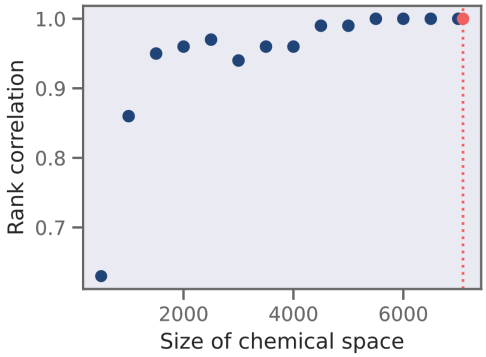

Figure 7: Rank correlation as a function of chemical space size. Spearman's rank-order correlation coefficient is calculated between the top five important descriptors in the subsampled chemical space and the reference chemical space. The red dotted line shows the size of the reference chemical space. We observe high correlation of ranks beyond a chemical space size of 4000.

ISBN 9781627033411. doi: 10.1007/978-1-62703-342-8_6. URL https://link.springer.com/protocol/10.1007/978-1-62703-342-8_6.

[2] Mark T.D. Cronin. Quantitative Structure–Activity Relationships (QSARs) – Applications and Methodology. In Tomasz Puzyn, Jerzy Leszczynski, and Mark T. D. Cronin, editors, *Recent Advances in QSAR Studies. Challenges and Advances in Computational Chemistry and Physics*, volume 8, pages 3–11. Springer, Dordrecht, 2010. doi: 10.1007/978-1-4020-9783-6_1. URL https://link.springer.com/chapter/10.1007/978-1-4020-9783-6_1.

[3] Artem Cherkasov, Eugene N. Muratov, Denis Fourches, Alexandre Varnek, Igor I. Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C. Martin, Roberto Todeschini, Viviana Consonni, Victor E. Kuz'Min, Richard Cramer, Romualdo Benigni, Chihae Yang, James Rathman, Lothar Terfloth, Johann Gasteiger, Ann Richard, and Alexander Tropsha. QSAR modeling: Where have you been? Where are you going to? *Journal of Medicinal Chemistry*, 57(12):4977–5010, jun 2014. ISSN 15204804. doi: 10.1021/JM4004285. URL https://pubs.acs.org/doi/full/10.1021/jm4004285.

[4] Giuseppina Gini. QSAR: What Else? In Orazio Nicolotti, editor, *Computational Toxicology. Methods in Molecular Biology.*, volume 1800, pages 79–105. Humana Press, New York, NY, 2018. doi: 10.1007/978-1-4939-7899-1_3. URL https://link.springer.com/protocol/10.1007/978-1-4939-7899-1_3.

[5] Mariia Matveieva, Mark T.D. Cronin, and Pavel Polishchuk. Interpretation of QSAR Models: Mining Structural Patterns Taking into Account Molecular Context. *Molecular Informatics*, 38(3):1800084, mar 2019. ISSN 1868-1751. doi: 10.1002/MINF.201800084. URL https://onlinelibrary.wiley.com/doi/full/10.1002/minf.201800084.

[6] Sunyoung Kwon, Ho Bae, Jeonghee Jo, and Sungroh Yoon. Comprehensive ensemble in QSAR prediction for drug discovery. *BMC Bioinformatics*, 20(521):1–12, oct 2019. ISSN 14712105. doi: 10.1186/S12859-019-3135-4. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-3135-4.

[7] Vivek Srivastava, Chandrabose Selvaraj, and Sanjeev Kumar Singh. Chemoinformatics and QSAR. In *Advances in Bioinformatics*, pages 183–212. Springer, Singapore, aug 2021. doi: 10.1007/978-981-33-6191-1_10. URL https://link.springer.com/chapter/10.1007/978-981-33-6191-1_10.

[8] Adam C. Mater and Michelle L. Coote. Deep learning in chemistry. *Journal of Chemical Information and Modeling*, 2019. ISSN 15205142. doi: 10.1021/ACS.JCIM.9B00266. URL https://pubs.acs.org/doi/full/10.1021/acs.jcim.9b00266.

[9] Elena L. Cáceres, Matthew Tudor, and Alan C. Cheng. Deep learning approaches in predicting ADMET properties. *Future Medicinal Chemistry*, 12(22):1995–1999, nov 2020. ISSN 17568927. doi: 10.4155/FMC-2020-0259.

[10] Qifeng Bai, Shuo Liu, Yanan Tian, Tingyang Xu, Antonio Jesús Banegas-Luna, Horacio Pérez-Sánchez, Junzhou Huang, Huanxiang Liu, and Xiaojun Yao. Application advances of deep learning methods for de novo drug design and molecular dynamics simulation. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, page

e1581, 2021. ISSN 1759-0884. doi: 10.1002/WCMS.1581. URL https://onlinelibrary.wiley.com/doi/full/10.1002/wcms.1581.

[11] W. Patrick Walters and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of Chemical Research*, 54:263–270, 1 2021. ISSN 15204898. doi: 10.1021/ACS.ACCOUNTS.0C00699. URL https://pubs.acs.org/doi/full/10.1021/acs.accounts.0c00699.

[12] Felipe Oviedo, Juan Lavista Ferres, Tonio Buonassisi, and Keith Butler. Interpretable and explainable machine learning for materials science and chemistry, 2021.

[13] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), aug 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL https://doi.org/10.1145/3236009.

[14] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2019.12.012. URL https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[15] Jenna A. Bilbrey, Joseph P. Heindel, Malachi Schram, Pradipta Bandyopadhyay, Sotiris S. Xantheas, and Sutanay Choudhury. A look inside the black box: Using graph-theoretical descriptors to interpret a continuous-filter convolutional neural network (cf-cnn) trained on the global and local minimum energy structures of neutral water clusters. *The Journal of Chemical Physics*, 153(2):024302, 2020. doi: 10.1063/5.0009933. URL https://doi.org/10.1063/5.0009933.

[16] Masanari Kimura and Masayuki Tanaka. New perspective of interpretability of deep neural networks. In *2020 3rd International Conference on Information and Computer Technologies (ICICT)*, pages 78–85, 2020. doi: 10.1109/ICICT50521.2020.00020.

[17] Kei Terayama, Masato Sumita, Ryo Tamura, and Koji Tsuda. Black-box optimization for automated discovery. *Accounts of Chemical Research*, 54(6):1334–1346, 2021. doi: 10.1021/acs.accounts.0c00713. URL https://doi.org/10.1021/acs.accounts.0c00713. PMID: 33635621.

[18] Cynthia Rudin and Berk Ustun. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *INFORMS Journal on Applied Analytics*, 48(5):449–466, 2018. doi: 10.1287/inte.2018.0957. URL https://doi.org/10.1287/inte.2018.0957.

[19] Sean Ekins, Ana C. Puhl, Kimberley M. Zorn, Thomas R. Lane, Daniel P. Russo, Jennifer J. Klein, Anthony J. Hickey, and Alex M. Clark. Exploiting machine learning for end-to-end drug discovery and development. *Nature Materials 2019 18:5*, 18:435–441, 4 2019. ISSN 1476-4660. doi: 10.1038/s41563-019-0338-z. URL https://www.nature.com/articles/s41563-019-0338-z.

[20] Maryam Ashoori and Justin D. Weisz. In ai we trust? factors that influence trustworthiness of ai-infused decision-making processes, 2019. URL https://arxiv.org/abs/1912.02675.

[21] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus Robert Müller. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications 2019 10:1*, 10(1096):1–8, mar 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4. URL https://www.nature.com/articles/s41467-019-08987-4.

[22] Kangway V. Chuang and Michael J. Keiser. Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science*, 362(6416):589–604, nov 2018. ISSN 10959203. doi: 10.1126/science.aat8603. URL https://www.science.org/doi/full/10.1126/science.aat8603.

[23] Richard Dybowski. Interpretable machine learning as a tool for scientific discovery in chemistry. *New Journal of Chemistry*, 44(48):20914–20920, dec 2020. ISSN 1369-9261. doi: 10.1039/D0NJ02592E. URL https://pubs.rsc.org/en/content/articlehtml/2020/nj/d0nj02592e.

[24] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2:573–584, 2020. doi: 10.1038/s42256-020-00236-4. URL https://doi.org/10.1038/s42256-020-00236-4.

[25] Zachary C. Lipton. The mythos of model interpretability. *CoRR*, abs/1606.03490, 2016. URL http://arxiv.org/abs/1606.03490.

[26] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[27] Yuan Hua, Wang Yongyan, and Cheng Yiyu. Local and global quantitative structure-activity relationship modeling and prediction for the baseline toxicity. *Journal of Chemical Information and Modeling*, 47(1):159–169, jan 2007. ISSN 15499596. doi: 10.1021/ci600299j. URL https://pubs.acs.org/doi/full/10.1021/ci600299j.

[28] Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. *ChemRxiv*, 2022. doi: 10.26434/chemrxiv-2022-mfq52.

[29] José Jiménez-Luna, Miha Skalic, and Nils Weskamp. Benchmarking Molecular Feature Attribution Methods with Activity Cliffs. *Journal of Chemical Information and Modeling*, 62(2):274–283, 2022. doi: 10.1021/acs.jcim.1c01163. URL https://doi.org/10.1021/acs.jcim.1c01163.

[30] Andrew D. White. *Deep Learning for Molecules and Materials*. 2021. URL https://dmol.pub.

[31] Pavel Polishchuk. Interpretation of quantitative structure–activity relationship models: Past, present, and future. *Journal of Chemical Information and Modeling*, 57:2618–2639, 11 2017. ISSN 15205142. doi: 10.1021/ACS.JCIM.7B00274. URL https://pubs.acs.org/doi/abs/10.1021/acs.jcim.7b00274.

[32] Raquel Rodríguez-Pérez and Jürgen Bajorath. Explainable machine learning for property predictions in compound optimization. *Journal of Medicinal Chemistry*, 64:17744–17752, 12 2021. ISSN 0022-2623. doi: 10.1021/ACS.JMEDCHEM.1C01789. URL https://pubs.acs.org/doi/abs/10.1021/acs.jmedchem.1c01789.

[33] Christina Humer, Henry Heberle, Floriane Montanari, Thomas Wolf, Florian Huber, Ryan Henderson, Julian Heinrich, and Marc Streit. Cheminformatics model explorer (CIME): Exploratory analysis of chemical model explanations. *ChemRxiv*, 2021. doi: 10.26434/chemrxiv-2021-crpd0.

[34] Danilo Numeroso and Davide Bacciu. Explaining deep graph networks with molecular counterfactuals, 2020.

[35] Geemi P Wellawatte, Aditi Seshadri, and Andrew D White. Model agnostic generation of counterfactual explanations for molecules. *Chem. Sci.*, 13(13):3697–3705, 2022. doi: 10.1039/D1SC05259D. URL http://dx.doi.org/10.1039/D1SC05259D.

[36] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 590–601, Red Hook, NY, USA, 2018. Curran Associates Inc.

[37] Kar Wai Lim, Bhanushee Sharma, Payel Das, Vijil Chenthamarakshan, and Jonathan S. Dordick. Explaining chemical toxicity using missing features, 2020.

[38] Bhanushee Sharma, Vijil Chenthamarakshan, Amit Dhurandhar, Shiranee Pereira, James A. Hendler, Jonathan S. Dordick, and Payel Das. Accurate clinical toxicity prediction using multi-task deep neural nets and contrastive molecular explanations, 2022. URL https://arxiv.org/abs/2204.06614.

[39] Kevin McCloskey, Ankur Taly, Federico Monti, Michael P. Brenner, and Lucy J. Colwell. Using attribution to decode binding mechanism in neural network models for chemistry. *Proceedings of the National Academy of Sciences*, 116(24):11624–11629, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1820657116. URL https://www.pnas.org/content/116/24/11624.

[40] Agnieszka Wojtuch, Rafał Jankowski, and Sabina Podlewska. How can shap values help to shape metabolic stability of chemical compounds? *Journal of Cheminformatics*, 13:1–20, 12 2021. ISSN 17582946. doi: 10.1186/S13321-021-00542-Y. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-021-00542-y.

[41] José Jiménez-Luna, Miha Skalic, Nils Weskamp, and Gisbert Schneider. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. *Journal of Chemical Information and Modeling*, 61:1083–1094, 3 2021. ISSN 15205142. doi: 10.1021/ACS.JCIM.0C01344. URL https://pubs.acs.org/doi/full/10.1021/acs.jcim.0c01344.

[42] Maria H. Rasmussen, Diana S. Christensen, and Jan H. Jensen. Do machines dream of atoms? a quantitative molecular benchmark for explainable ai heatmaps. *ChemRxiv*, 2022. doi: 10.26434/chemrxiv-2022-gnq3w.

[43] Lars Carlsson, Ernst Ahlberg Helgee, and Scott Boyer. Interpretation of nonlinear qsar models applied to ames mutagenicity data. *Journal of Chemical Information and Modeling*, 49:2551–2558, 11 2009. ISSN 15499596. doi: 10.1021/CI9002206. URL https://pubs.acs.org/doi/full/10.1021/ci9002206.

[44] Hyeoncheol Cho, Eok Kyun Lee, and Insung S. Choi. Layer-wise relevance propagation of interaction-net explains protein–ligand interactions at the atom level. *Scientific Reports 2020 10:1*, 10:1–11, 12 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-78169-6. URL https://www.nature.com/articles/s41598-020-78169-6.

[45] Josh Payne, Mario Srouji, Dian Ang Yap, and Vineet Kosaraju. Bert learns (and teaches) chemistry, 2020.

[46] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, San Francisco, California, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

[48] Leanne S. Whitmore, Anthe George, and Corey M. Hudson. Mapping chemical performance on molecular structures using locally interpretable explanations, 2016. URL https://arxiv.org/abs/1611.07443.

[49] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[50] L. S. Shapley. *Contributions to the Theory of Games (AM-28)*, volume II, chapter 17. A Value for n-Person Games, pages 307–318. Princeton University Press, 5 2016. doi: 10.1515/9781400881970-018/HTML.

[51] Raquel Rodríguez-Pérez and Jürgen Bajorath. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *Journal of Medicinal Chemistry*, 63: 8761–8777, 8 2019. ISSN 15204804. doi: 10.1021/ACS.JMEDCHEM.9B01101. URL https://pubs.acs.org/doi/full/10.1021/acs.jmedchem.9b01101.

[52] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. ISSN 15499596. doi: 10.1021/CI100050T. URL https://pubs.acs.org/doi/full/10.1021/ci100050t.

[53] Daniel Probst and Jean Louis Reymond. A probabilistic molecular fingerprint for big data settings. *Journal of Cheminformatics*, 10:1–12, 12 2018. ISSN 17582946. doi: 10.1186/S13321-018-0321-8/FIGURES/8. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-018-0321-8.

[54] Louis Bellmann, Patrick Penner, and Matthias Rarey. Connected Subgraph Fingerprints: Representing Molecules Using Exhaustive Subgraph Enumeration. *Journal of Chemical Information and Modeling*, 59(11):4625–4635, 2019. doi: 10.1021/acs.jcim.9b00571. URL https://doi.org/10.1021/acs.jcim.9b00571.

[55] AkshatKumar Nigam, Robert Pollice, Mario Krenn, Gabriel dos Passos Gomes, and Alan Aspuru-Guzik. Beyond generative models: Superfast traversal, optimization, novelty, exploration and discovery (stoned) algorithm for molecules using selfies. *ChemRxiv*, 2021. doi: 10.26434/chemrxiv.13383266.v2.

[56] Liangxu Xie, Lei Xu, Ren Kong, Shan Chang, and Xiaojun Xu. Improvement of Prediction Performance With Conjoint Molecular Fingerprint in Deep Learning. *Frontiers in Pharmacology*, 11:2148, 2020. ISSN 1663-9812. doi: 10.3389/fphar.2020.606668. URL https://www.frontiersin.org/article/10.3389/fphar.2020.606668.

[57] Rdkit: Open-source cheminformatics. http://www.rdkit.org.

[58] Kacper Sokol and Peter Flach. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 56–67, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369367. doi: 10.1145/3351095.3372870. URL https://doi.org/10.1145/3351095.3372870.

[59] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[60] A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, and A. G. Yagola. *Numerical Methods for the Solution of Ill-Posed Problems*. Springer, Dordrecht, Netherlands, 1 edition, 1995. ISBN 978-94-015-8480-7. doi: 10.1007/978-94-015-8480-7.

[61] Siddhant Arora, Danish Pruthi, Norman Sadeh, William W Cohen, Zachary C Lipton, and Graham Neubig. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. *arXiv preprint arXiv:2112.09669*, 2021.

[62] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7:1–13, 12 2015. ISSN 17582946. doi: 10.1186/S13321-015-0069-3. URL https://jcheminf.biomedcentral.com/articles/10.1186/s13321-015-0069-3.

[63] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1:045024, 10 2020. ISSN 2632-2153. doi: 10.1088/2632-2153/ABA947. URL https://iopscience.iop.org/article/10.1088/2632-2153/aba947.

[64] David Elder and René Holm. Aqueous solubility: Simple predictive methods (in silico, in vitro and bio-relevant approaches). *International Journal of Pharmaceutics*, 453(1):3–11, 2013. ISSN 0378-5173. doi: 10.1016/j.ijpharm.2012.10.041. URL https://www.sciencedirect.com/science/article/pii/S0378517312009817.

[65] John S. Delaney. ESOL: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005, may 2004. ISSN 00952338. doi: 10.1021/ci034243x. URL https://pubs.acs.org/doi/full/10.1021/ci034243x.

[66] Samuel Boobier, David R.J. Hose, A. John Blacker, and Bao N. Nguyen. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature Communications*, 11(5753):1–10, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19594-z. URL https://www.nature.com/articles/s41467-020-19594-z.

[67] Qiuji Cui, Shuai Lu, Bingwei Ni, Xian Zeng, Ying Tan, Ya Dong Chen, and Hongping Zhao. Improved Prediction of Aqueous Solubility of Novel Compounds by Going Deeper With Deep Learning. *Frontiers in Oncology*, 10, 2020. ISSN 2234-943X. doi: 10.3389/fonc.2020.00121. URL https://www.frontiersin.org/article/10.3389/fonc.2020.00121.

[68] Murat C. Sorkun, Abhishek Khetan, and Süleyman Er. AqSolDB, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds. *Scientific Data*, 6:1–8, 8 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0151-1. URL https://www.nature.com/articles/s41597-019-0151-1.

[69] Antonio Llinas, Ioana Oprisiu, and Alex Avdeef. Findings of the second challenge to predict aqueous solubility. *Journal of Chemical Information and Modeling*, 60(10):4791–4803, oct 2020. ISSN 15205142. doi: 10.1021/acs.jcim.0c00701. URL https://pubs.acs.org/doi/full/10.1021/acs.jcim.0c00701.

[70] Murat Cihan Sorkun, J. M.Vianney A. Koelman, and Süleyman Er. Pushing the limits of solubility prediction via quality-oriented data selection. *iScience*, 24(1):101961, jan 2021. ISSN 2589-0042. doi: 10.1016/J.ISCI.2020.101961.

[71] Paul G. Francoeur and David R. Koes. SolTranNet-A Machine Learning Tool for Fast Aqueous Solubility Prediction. *Journal of Chemical Information and Modeling*, 61(6):2530–2536, jun 2021. ISSN 15205142. doi: 10.1021/acs.jcim.1c00331. URL https://pubs.acs.org/doi/full/10.1021/acs.jcim.1c00331.

[72] Camille Bilodeau, Wengong Jin, Hongyun Xu, Jillian A. Emerson, Sukrit Mukhopadhyay, Thomas H. Kalantar, Tommi Jaakkola, Regina Barzilay, and Klavs F. Jensen. Generating molecules with optimized aqueous solubility using iterative graph translation. *Reaction Chemistry & Engineering*, 7(2):297–309, feb 2022. ISSN 2058-9883. doi: 10.1039/D1RE00315A. URL https://pubs.rsc.org/en/content/articlelanding/2022/re/d1re00315a.

[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[74] Neera Jain and Samuel H. Yalkowsky. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *Journal of Pharmaceutical Sciences*, 90(2):234–252, 2001. doi: 10.1002/1520-6017(200102)90:2⟨234::AID-JPS14⟩3.0.CO;2-V. URL https://www.sciencedirect.com/science/article/abs/pii/S0022354916307158.

[75] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning, 2015. URL https://arxiv.org/abs/1506.00019.

[76] François Chollet et al. Keras. https://keras.io, 2015.

[77] David Weininger. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, feb 1988. ISSN 00952338. doi: 10.1021/CI00057A005. URL https://pubs.acs.org/sharingguidelines.

[78] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014. URL https://arxiv.org/abs/1412.6980.

[79] Scott A. Wildman and Gordon M. Crippen. Prediction of Physicochemical Parameters by Atomic Contributions. *Journal of Chemical Information and Computer Sciences*, 39(5):868–873, 1999. ISSN 00952338. doi: 10.1021/CI990307L. URL https://pubs.acs.org/doi/full/10.1021/ci9903071.

[80] Timothy J. Ritchie and Simon J.F. Macdonald. The impact of aromatic ring count on compound developability – are too many aromatic rings a liability in drug design? *Drug Discovery Today*, 14(21-22):1011–1020, Nov 2009. ISSN 1359-6446. doi: 10.1016/J.DRUDIS.2009.07.014.

[81] Michael A. Walker. Improvement in aqueous solubility achieved via small molecular changes. *Bioorganic & Medicinal Chemistry Letters*, 27(23):5100–5108, dec 2017. ISSN 0960-894X. doi: 10.1016/J.BMCL.2017.09.041.

[82] Karen Schomburg, Hans Christian Ehrlich, Katrin Stierand, and Matthias Rarey. From structure diagrams to visual chemical patterns. *Journal of Chemical Information and Modeling*, 50(9):1529–1535, sep 2010. ISSN 15499596. doi: 10.1021/ci100209a. URL `https://pubs.acs.org/doi/full/10.1021/ci100209a`.

[83] Stephen Kokoska and Daniel Zwillinger. *CRC Standard Probability and Statistics Tables and Formulae*. CRC Press, Boca Raton, 1st edition edition, mar 2000. ISBN 9780429181467. doi: 10.1201/B16923. URL `https://doi.org/10.1201/b16923`.

SUPPLEMENTAL INFORMATION: EXPLAINING STRUCTURE–ACTIVITY RELATIONSHIPS USING LOCALLY FAITHFUL SURROGATE MODELS

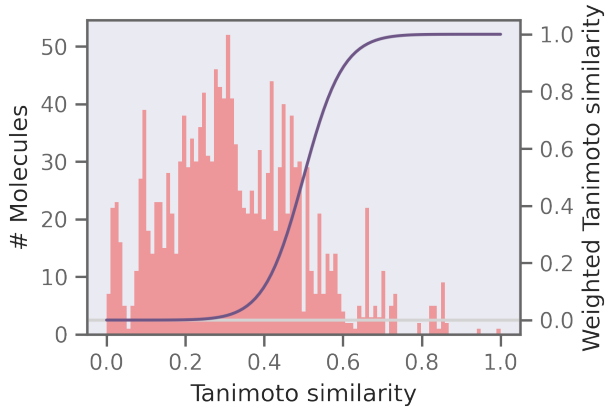

Figure S1: Tanimoto similarities of molecules are weighted using a shifted sigmoid function so that dissimilar molecules are excluded from weighted least squares regression fit. The histogram shows the distribution of tanimoto similarities and purple line shows the weighted value for tanimoto similarity.

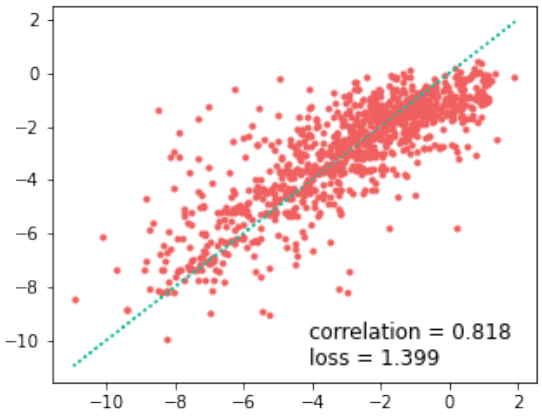

Figure S2: Random Forest Regression performance. The model was trained using 100 decision trees with a tree depth of 10. Data was split using a 90-10 train-test data split.

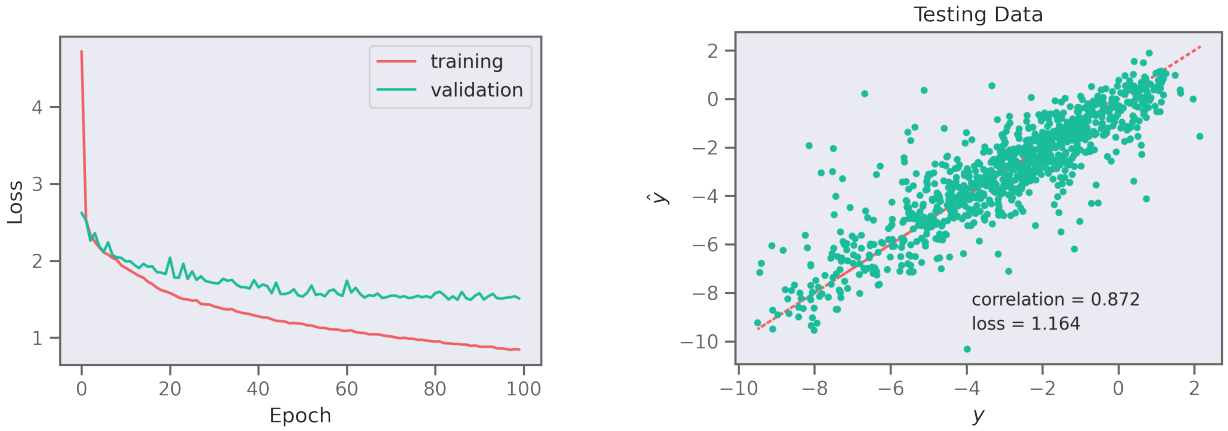

Figure S3: RNN performance. Loss curve shows training and validation loss over 100 epochs. Parity plot for testing data shows correlation between RNN predictions and experimental values.

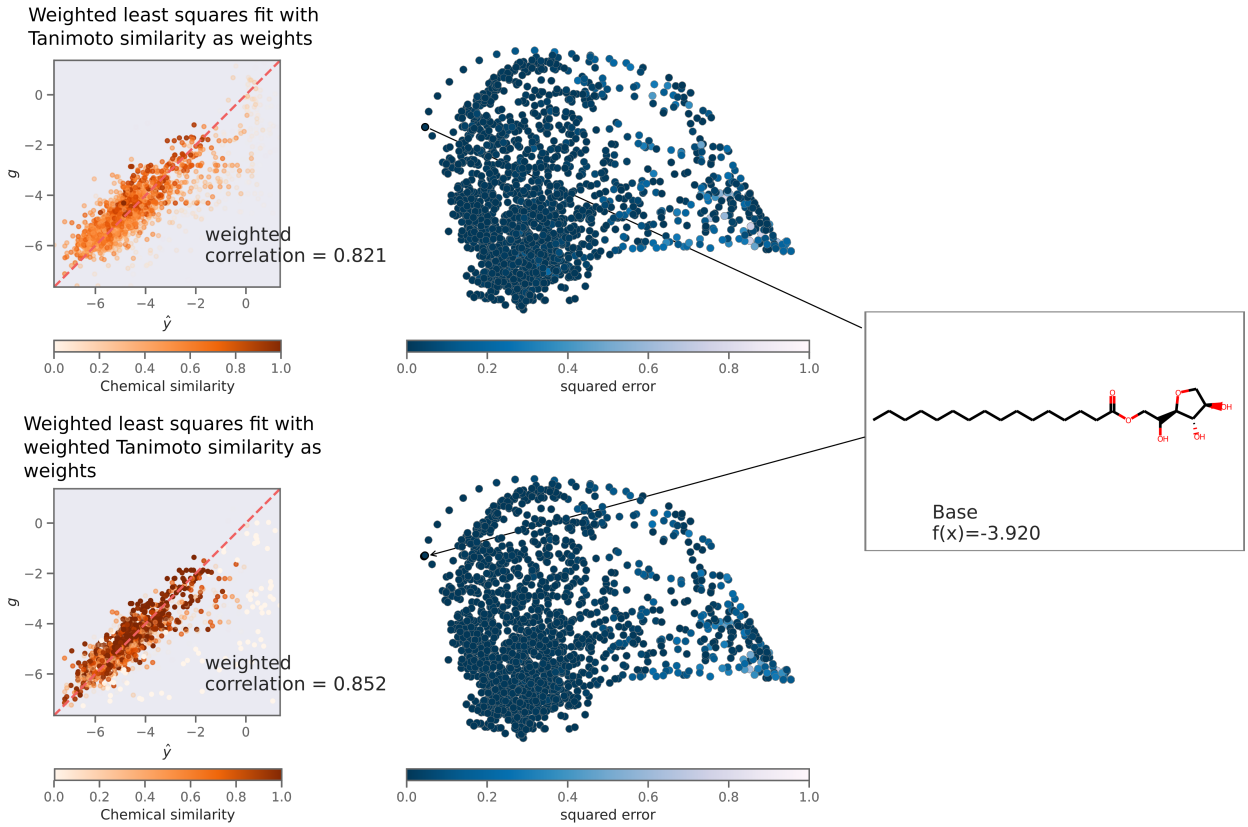

Figure S4: Comparison of using unweighted and weighted tanimoto similarities as weights for weighted least squares regression. The dissimilar molecules do not affect the regression fit, however, they end up contributing in determining descriptor explanations.

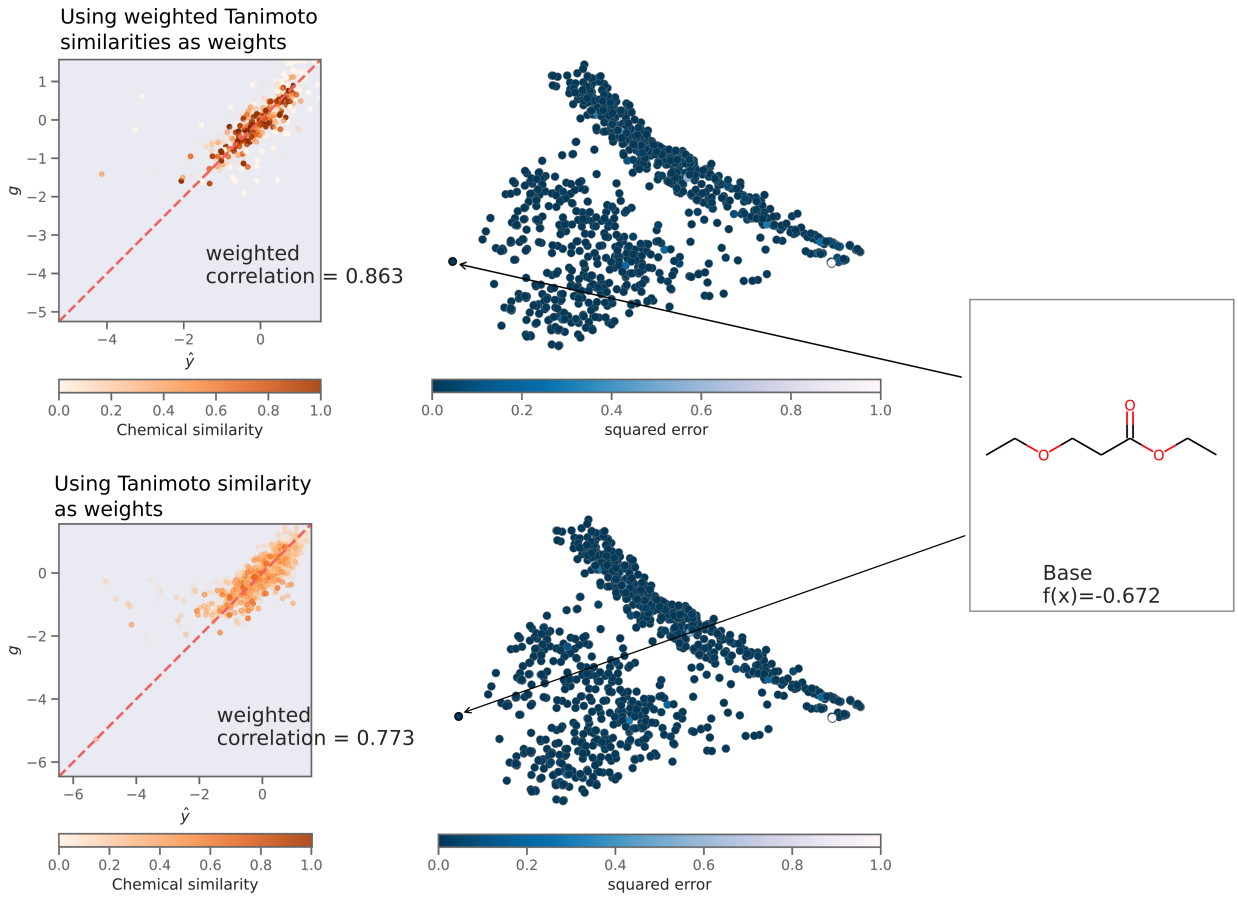

Figure S5: Comparison of using unweighted and weighted tanimoto similarities as weights for weighted least squares regression. For small molecules, the dissimilar molecules lead to poor regression fit and misleading explanations.

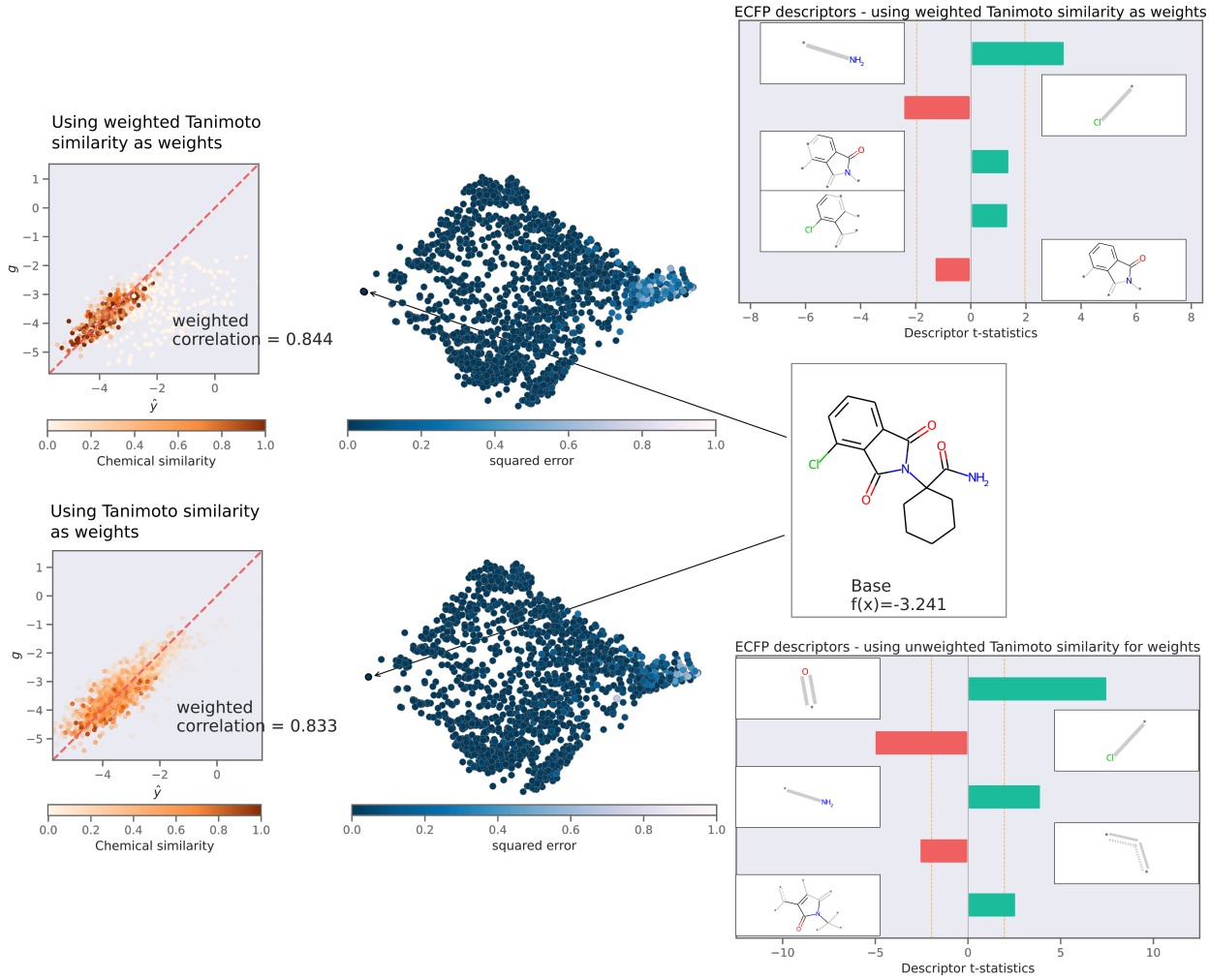

Figure S6: Comparison of using unweighted and weighted tanimoto similarities as weights for weighted least squares regression. The dissimilar molecules don't affect the regression fit for ring compounds, but affect the explanations.

Figure S7: Effect of number of mutations on descriptor explanations. Number of mutations is a STONED parameter that controls how many additions, deletions or modifications can be made to the SELFIES string. SMARTS annotations for MACCS descriptors were created using SMARTSviewer (smartsview.zbh.uni-hamburg.de, Copyright: ZBH, Center for Bioinformatics Hamburg) developed by Schomburg et al. [S82]

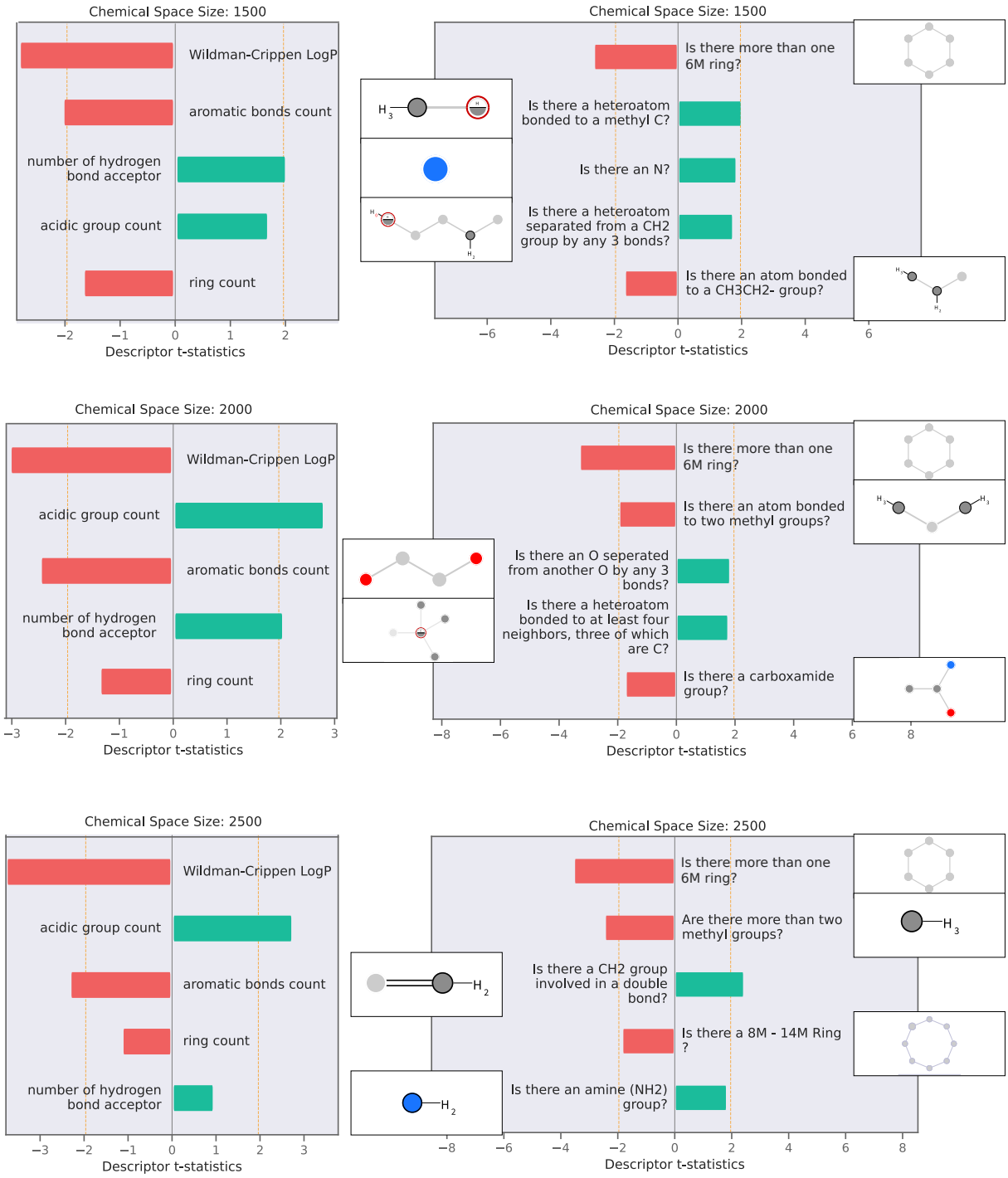

Figure S8: Effect of chemical space size on descriptor explanations. 'Chemical space size' parameter specifies how many mutated molecules of the base instance must be created. SMARTS annotations for MACCS descriptors were created using SMARTSviewer (smartsview.zbh.uni-hamburg.de, Copyright: ZBH, Center for Bioinformatics Hamburg) developed by Schomburg et al.[S82]