# A Perspective on Explanations of Molecular Prediction Models

A Preprint

**Geemi P. Wellawatte**
Department of Chemistry
University of Rochester
Rochester, NY, 14627
gwellawa@ur.rochester.edu

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

**Aditi Seshadri**
Department of Chemical Engineering
University of Rochester
Rochester, NY, 14627
aseshad4@u.rochester.edu

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

December 8, 2022

**Abstract**

Chemists can be skeptical in using deep learning (DL) in decision making, due to the lack of interpretability in "black-box" models. Explainable artificial intelligence (XAI) is a branch of AI which addresses this drawback by providing tools to interpret DL models and their predictions. We review the principles of XAI in the domain of chemistry and emerging methods for creating and evaluating explanations. Then we focus methods developed by our group and their application to predicting solubility, blood-brain barrier permeability, and the scent of molecules. We show that XAI methods like chemical counterfactuals and descriptor explanations can both explain DL predictions and give insight into structure-property relationships. Finally, we discuss how a two step process of highly accurate black-box modeling and then creating explanations gives both highly accurate predictions and clear structure-property relationships.

## 1 Introduction

Deep learning (DL) is advancing the boundaries of computational chemistry because it can accurately model non-linear structure-function relationships.[1–3] Applications of DL can be found in a broad spectrum spanning from quantum computing[4,5] to drug discovery[6–10] to materials design.[11,12] The rationale of DL predictions is not always apparent due to the architecture and large parameter count of DL models.[13,14] DL models are thus often termed"black box" models. We can only reason about the input and output to DL model, not the underlying process that leads to a specific prediction.

The black box nature has practical consequences and is a limitation of DL. Users are more likely to trust and use predictions from a model if they can understand why the prediction was made.[15] Explaining predictions can help developers of DL models ensure the model is not learning spurious correlations.[16,17] Infamous examples are neural networks that learned to recognize horses by looking for a photographer's watermark[18] and neural networks that predicted a COVID-19 diagnoses by looking at font choice on medical images.[19] It is routine in chemistry now for DL to exceed human level performance — humans are not good at predicting solubility from structure for example[20] [21] — and so understanding how a model makes predictions can guide hypotheses. This is in contrast to a topic like finding a stop sign in an image, where there is little new to be learned about visual perception by explaining a DL model.

*Corresponding Author

Finally, there is an emerging regulatory framework for when any computer algorithms impact humans.[22–24] Although we know of no examples yet in chemistry, the use of AI in predicting toxicity, carcinogenicity, and environmental persistence may require rationale for the predictions to have regulatory consequences.

EXplainable Artificial Intelligence (XAI) is a field of growing importance that aims to address these limitations of DL. The main goals are to provide model interpretations of DL predictions. Miller[25] defines that interpretability of a model refers to the degree of human understandability intrinsic within the model. Murdoch et al.[26] clarify that interpretability can be perceived as "knowledge" which provide insight to a particular problem. Justifications are quantitative metrics tell the users "why the model should be trusted," like test error.[27] Justifications are evidence which defend why a prediction is trustworthy.[25] An "explanation" is a description on why a certain prediction was made.[9,28] Interpretability and explanation are often used interchangeably. Arrieta et al.[13] distinguish that interpretability is a passive characteristic of a model, whereas explainability is an active characteristic which is used to clarify the internal decision-making process. Namely, an explanation is extra information that is attached to each prediction and gives the context and a cause for the prediction.[29] We adopt this nomenclature: interpretability is intrinsic attribute of a model and an explanation is additional information for a prediction (or occasionally a group of predictions).

There is often a trade-off between model accuracy and interpretability. For example, linear regression is interpretable but inaccurate. DL models are accurate but not interpretable.[28,30] XAI provides a way to avoid that trade-off in chemical property prediction. We develop an accurate but uninterpretable DL model first. Then add explanations to predictions. If the DL model is correctly capturing the data generating process, then the explanations should give insight into the underlying mechanism. In the remainder of this article, we review recent approaches for XAI of chemical property prediction and then focus on specific examples with our own recent XAI work based around creating local chemical spaces with SELF-referencing embedded string (SELFIES).[31–33] We show how in various systems, this two step approach of developing an accurate DL model first and then explaining it yields explanations that are consistent with known and mechanism supported structure-property relationships.

## 2 Theory

There is a lack of consensus on how to classify and evaluate XAI.[34,35] Das and Rad[36] propose a taxonomy based on three XAI classifications. The first is what is being explained: the entire model (global interpretations) or an individual outcome (instance interpretations). The second is the relation between the model and the interpretation: post-hoc (extrinsic) or intrinsic to the model.[36,37] The last is the methodology used. For example, if backpropagation or perturbations are used.

An intrinsic XAI method is part of the model and thus may not generalize.[36] An extrinsic method is one that can be applied post-training to any model.[37] Sometimes these are called model agnostic. Intrinsic models are self-explanatory. Two examples are linear models and decision trees. These are also referred to as white-box models due to contrast them with uninterpretable black box models.[28] Post-hoc methods found in literature focus on interpreting models through (1) training data[38] and feature attribution[39] (2) surrogate models[10] and (3) counterfactual[9] or contrastive explanations.[40]

There is a lack of consensus how do assess correctness in XAI. What is a "good" explanation and what are the required components of an explanation are debated.[36,41]Palacio et al.[29] state the lack of a standard framework for XAI has caused the inability to agree if a model is interpretable or not. For example, Shapley values are Jin et al.[41] argue that explanations fundamentally for humans and their evaluation depends on "complex human factors and application scenarios." In physical sciences, we may instead consider if the explanations somehow reflect the underlying physics. For example, Oviedo et al.[42] propose that a model explanation can be evaluated by considering its agreement with the underlying physical system, which they term "correctness."

Therefore, we can expect a trade-off between the degree of understandability and completeness of an explanation. For example, an explanation can be described through the trainable parameters, but the understandability reduces with increasing non-linearity. Additionally, authors propose that "correctness" should be another property to evaluate an explanation.[42] However, this is an intrinsic property of the model which measures its scientific accuracy.

Another challenge in XAI is the lack of an agreed-upon framework to evaluate an interpretation or explanation.[36,41,43] Some attributes of an explanation are:

- *Actionable.* Is it clear how we could change the input features to modify the output?
- *Complete.* Are all contributing features explained?
- *Correct.* Does the explanation agree with hypothesized or known underlying physical mechanism?[42]

- *Domain Applicable.* Does the explanation use language and concepts of domain experts?
- *Fidelity/Faithful.* Does the explanation agree with the black box model?
- *Robust.* Does the explanation change significantly with small changes to the model or instance being explained?
- *Sparse/Succinct.* Is the explanation succinct?

For example, Shapley values are proposed as an explanation method because they offer a complete explanation.[44] Completeness is a clearly measurable and well-defined metric and yields explanations with many components. Yet Shapley values are not actionable nor sparse. Ribeiro et al.[39] proposed a surrogate model method that aims to provide sparse/succinct explanations that have high fidelity to the original model. We argued in Wellawatte et al.[9] that counterfactuals are more useful explanations because they are actionable. Jin et al.[41] posit that explanations are fundamentally for humans and their evaluation depends on "complex human factors and application scenarios." In physical sciences, we consider if the explanations somehow reflect the underlying physics. Oviedo et al.[42] proposed that a model explanation can be evaluated by considering its agreement with the underlying physical system, which they term "correctness."

### 2.1 Self-explaining models

A self-explanatory model is one that is intrinsically interpretable to an expert.[45] Two common examples found in literature are linear regression models and decision trees (DT). A linear model is described by the equation 1 where, $W's$ are the weight parameters and $x's$ are the input features associated with the prediction $\hat{y}$. The trained parameters provide a complete explanation of the model. Accodring to Molnar et al.[45], trained weights quantify the importance of each feature, thereby uncovering the rational for a prediction.

$$\hat{y} = \Sigma_i W_i x_i \tag{1}$$

DT models are another type of self-explaining models which have been used in classification and high-throughput screening tasks. Gajewicz et al.[46] used DT models to classify nanomaterials that identify structural features responsible for surface activity. In another study by Han et al.[47], a DT model was developed to filter compounds by their bioactivity based on the chemical fingerprints.

Intrinsic interpretability can also be improved by regularizing the input gradients as they can identify which feature descriptors contributed towards a prediction.[48] Regularization techniques such as EXPO[49] and RRR[50] are designed to enhance the black-box model interpretability. Although one can argue that "simplicity" of models are positively correlated with interpretability, this is based on how the interpretability is evaluated. For example, Lipton[35] argue that, from the notion of "simulatability" (the degree to which a human can predict the outcome based on inputs), self-explanatory linear models, rule-based systems, and DTs can be claimed uninterpretable. A human can predict the outcome given the inputs only if the input features are interpretable. Therefore, a linear model which takes in non-descriptive inputs may not be as transparent. Based on the correctness of a model, a linear model is not inherently accurate as they fail to capture non-linear relationships in data limiting is applicability. Similarly, a DT is a rule-based model which may lack physics informed knowledge. However, intrinsic models are commonly accepted as transparent models They can be found in other XAI applications in interpreting black-box models acting as surrogate models (proxy models).[51,52]

### 2.2 Attribution methods

Feature attribution methods explain black box predictions by assigning each input feature a numerical value which indicates its importance or contribution to the prediction. These atom-based numerical assignments are commonly referred to as heatmaps.[53] Recently, Rasmussen et al.[54] showed that Crippen logP models serve as a benchmark for heatmap approaches. However, some argue if these heatmaps (attribution methods) provide actual explanations.[35] Feature attributions provide local explanations, but can be averaged or combined explain multiple instances. The Most widely used feature attribution approaches in literature are gradient based methods,[55,56] Shapley Additive exPlanations (SHAP),[57] and layerwise relevance propogation.[58]

Gradient based approaches are based on the idea that gradients for neural networks are analogous to coefficients for regression models.[59] Class activation maps (CAM),[60] gradCAM,[61] smoothGrad,,[62] and integrated gradients[59] are examples of gradient methods for feature attribution. The principle of feature attributions with gradients is that:

$$\frac{\Delta \hat{f}(\vec{x})}{\Delta x_i} \approx \frac{\partial \hat{f}(\vec{x})}{\partial x_i} \tag{2}$$

where $\hat{f}(x)$ our black box models and $\frac{\Delta \hat{f}(\vec{x})}{\Delta x_i}$ are used as our attributions. The left hand says that we attribute each input feature $x_i$ by how much one unit change in it would affect the output of $\hat{f}(x)$. You can also view this as assuming a linear surrogate model and this can be reconciled with LIME.[39] For DL models, $\nabla_x f(x)$, suffers from something called the shattered gradients problem[59] – which means directly computing it leads to numeric problems. The different gradient based approaches are mostly distinguishable about how they approximate this gradient.

Gradient based explanations have been used widely to interpret chemistry predictions.[56,63–67] McCloskey et al.[56] used graph convolutional networks (GCNs) to predict protein-ligand binding and explained the binding logic for these predictions using integrated gradients. Pope et al.[63] and Jiménez-Luna et al.[64] show application of gradCAM and integrated gradients to explain molecular property predictions from trained graph neural networks (GNNs). Sanchez-Lengeling et al.[65] present comprehensive, open-source XAI benchmarks to explain GNNs and other graph models. They compare the performance of class activation maps (CAM),[60] gradCAM,[61] smoothGrad,,[62] integrated gradients[59] and attention mechanisms for explaining outcomes of classification as well as regression tasks. They concluded that CAM and integrated gradients perform well for graph models. Another attempt at creating XAI benchmarks for graph models was made by Rao et al.[67]. They compare these gradient based methods to find subgraph importance when predicting activity cliffs and conclude that gradCAM and integrated gradients provided the most interpretability for GNNs. GNNExplainer[66] focuses on identifying the most influential subgraph and node feature contributions that maximize mutual information between the GNN prediction and the distribution of possible subgraphs. Ying et al.[66] show that GNNExplainer can be used to obtain model-agnostic local as well as global explanations. SubgraphX is another method that explains GNN predictions by identifying important subgraphs.[68]

Another set of approaches like DeepLIFT[69] and Layerwise Relevance backPropagation[70] (LRP) depends on backpropagation of prediction scores through each layer of the neural network. The specific backpropagation logic across various activation functions differs in these approaches, which means each layer must have its own implementation. Baldassarre and Azizpour[71] show application of LRP to explain aqueous solubility prediction for molecules.

SHAP is a model-agnostic feature attribution method that is inspired from the game theory concept of Shapley values.[57,72] It's an additive feature contribution approach which assumes that an explanation model is a linear combination of binary variables $z$. If the Shapley value for the $i^{th}$ feature is $\phi_i$, then the explanation is $\hat{f}(\vec{x}) = \sum_i \phi_i(\vec{x}) z_i(\vec{x})$. Shapley values for features are computed using Equation 3.[73,74]

$$\phi_i(\vec{x}) = \frac{1}{M} \sum^{M} \hat{f}\left(\vec{z}_{+i}\right) - \hat{f}\left(\vec{z}_{-i}\right) \tag{3}$$

Here $\vec{z}$ is a fabricated example created from the original $\vec{x}$ and a random perturbation $\vec{x}'$. $\vec{z}_{+i}$ has the feature $i$ from $\vec{x}$ and $\vec{z}_{-i}$ has the $i_th$ feature from $\vec{x}'$. Some care should be taken in constructing $\vec{z}$ when working with molecular descriptors to ensure that an impossible $\vec{z}$ is not sampled (e.g., high count of acid groups but no hydrogen bond donors). $M$ is the sample size of perturbations around $\vec{x}$. Shapley value computation is expensive, hence $M$ is chosen accordingly. Equation 3 is an approximation and gives contributions with an expectation term as $\phi_0 + \sum_{i=1} \phi_i(\vec{x}) = \hat{f}(\vec{x})$. SHAP has been popularly used in explaining molecular prediction models.[75–78]

Visualization based feature attribution has also been used for molecular data. In computer science, saliency maps is a way to measure spatial feature contribution.[79] Simply put, saliency maps draw a connection between the model's neural fingerprint components (trained weights) and input features. Weber et al.[80] use saliency maps to build an explainable GCN architecture that gives subgraph importance for small molecule activity prediction. Similarity maps, on the other hand, compare model predictions for two or more molecules based on their chemical fingerprints.[81] Similarity maps provide atomic weights or predicted probability difference between the molecules by removing atoms one at a time. These weights can then be used to color the molecular graph and give a visual presentation. ChemInformatics Model Explorer (CIME) is an interactive web based toolkit which allows visualization and comparison of different explanation methods for molecular property prediction models.[82]

### 2.3 Surrogate models

One approach to explain black box predictions is to fit a self-explaining or interpretable model to the black box model, in the vicinity of one or a few specific examples. These are known as surrogate models. We will make one model

per explanation. If we could find one that explained the whole model, then we would simply have a globally accurate interpretable model and no longer need the black box model.[73] In the work by White[73] a weighted least squares linear model is used as the surrogate model. This model provides natural language based descriptor explanations by replaced input features with chemically interpretable descriptors. This approach is similar to the concept-based explanations approach used by McGrath et al.[83] where human understandable concepts are used in place of input features in acquisition of chess knowledge in AlphaZero. Any of the self-explaining models detailed in the Self-explaining models section can be used as a surrogate model.

The most commonly used surrogate model based method is Locally Interpretable Model Explanations (LIME).[39] LIME creates perturbations around the example of interest and fits an interpretable model to these local perturbations. Ribeiro et al.[39] mathematically define an explanation $\xi$ for an example $\vec{x}$ using Equation 4.

$$\xi(\vec{x}) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g) \tag{4}$$

Where $f$ is the black box model and $g \in G$ is the interpretable explanation model. $G$ is a class of potential interpretable models, such as linear models, decision trees, and so on. $\pi_x$ is a similarity measure between original input $\vec{x}$ and it's perturbed input $\vec{x}'$. In context of molecular data, this could be a chemical similarity coefficient like Tanimoto[84] similarity, cosine similarity and so on. The goal for LIME is to minimize the loss, $\mathcal{L}$, so that $g$ closely approximates $f$. $\Omega$ is a parameter that controls the complexity (sparsity) of $g$. Ribeiro et al.[39] terms the agreement (how low the loss is) between $f$ and $g$ as the fidelity.

GraphLIME[85] and LIMEtree[86] are modifications to LIME as applicable to graph neural networks and regression trees, respectively. LIME has been used in chemistry previously, such as Whitmore et al.[87] who used LIME to explain octane number predictions of molecules from a random forest classifier. Mehdi and Tiwary[88] used LIME to explain thermodynamic contributions of features. Gandhi and White[10] use an approach similar to GraphLIME, but use chemistry specific fragmentation and descriptors to explain molecular property prediction. Some examples are highlighted in the Applications section. In recent work by Mehdi and Tiwary[88], a thermodynamic-based surrogate model approach was used to interpret black-box models. The authors define an "interpretation free energy" which can be achieved by minimizing the surrogate model's uncertainty and maximizing simplicity.

### 2.4 Counterfactual explanations

Counterfactual explanations can be found in many fields such as statistics, mathematics and philosophy.[89–92] According to Woodward and Hitchcock[90], a counterfactual is an example with minimum deviation from the initial instance but with a contrasting outcome. They can be used to answer the question, "which smallest change could alter the outcome of an instance of interest?" While this deviation between the two instances is based on the existence of similar worlds in philosophy,[93] a distance metric based on molecular similarity is employed in XAI for chemistry. For example, in the work by Wellawatte et al.[9] distance between two molecules is defined as the Tanimoto distance[94] between ECFP4 fingerprints.[95] Contrastive explanations are tangential to counterfactual explanations. Unlike the counterfactual approach, contrastive approach employ a dual optimization method, which works by generating a similar and a dissimilar (counterfactuals) example. Contrastive explanations can interpret the model by identifying contribution of presence and absence of subsets of features towards a certain prediction.[40,96]

A counterfactual $x'$ of an input $x$ is one with an dissimilar prediction $\hat{f}(x)$. Therefore, counterfactual generation can be thought of as a constrained optimization problem which minimizes the vector distance $d(x, x')$ between the features.[9,97]

$$\begin{aligned} &\text{minimize} \quad d(x, x') \\ &\text{such that} \quad \hat{f}(x) \neq \hat{f}(x') \end{aligned} \tag{5}$$

Equation 5 can be adapted for regression tasks as shown below in equation 6, where a counterfactual is one with a defined increase or decrease in the prediction.

$$\begin{aligned} &\text{minimize} \quad d(x, x') \\ &\text{such that} \quad \left| \hat{f}(x) - \hat{f}(x') \right| \geq \Delta \end{aligned} \tag{6}$$

Counterfactuals explanations have become a useful tool for XAI in chemistry, as they provide intuitive understanding of a decision and are able to uncover spurious relationships in training data[98] – helps to uncover physicochemical

mechanisms based on data. Counterfactuals create local (instance-level), actionable explanations. Actionability of an explanation is the degree to which the explanation can be acted on. For example, if an explanation claim that the the electronegativity of an atom contributes towards a molecule's solubility, this is non-actionable.

Counterfactual generation is an demainding task as it requires gradient optimization over discrete features that represents a molecule. Recent work by Fu et al.[99] and Shen et al.[100] present two techniques which allow continuous gradient-based optimization. Although, these methodologies are shown to circumvent the issue of discrete molecular optimization, counterfactual explanation based model interpretation still remains unexplored compared to other post-hoc methods. The GNNExplainer[66] is another approach for generating local explanations for graph based models. This method focuses on distinguishing which sub-graphs contribute most to the prediction by maximizing mutual information between the prediction and distribution of possible sub-graphs. However, this method fits under the feature attribution category than the counterfactual explanation category.

CF-GNNExplainer is a method based on the GNNExplainer which generates counterfactual explanations for graph based data.[101] This method generate informative and non-adversarial counterfactuals by perturbing the input data (removing edges in the graph), and keeping account of which perturbations lead to changes in the output. However, this method is only applicable to graph-based models and can generate infeasible molecular structures. Another related work by Numeroso and Bacciu[102] focus on generating counterfactual explanations for deep graph networks. Their method MEG (Molecular counterfactual Explanation Generator) uses a reinforcement learning based generator to create molecular counterfactuals (molecular graphs). While this method is able to generate counterfactuals through a multi-objective reinforcement learner, this is not a universal approach (only for graph based models) and requires training the generator for each task.

Work by Wellawatte et al.[9] present a model agnostic counterfactual generator MMACE (Molecular Model Agnostic Counterfactual Explanations) which does not require training or computing gradients. This method employs a molecular generator based on the STONED algorithm[33] to populate a local chemical space through random string mutations of SELFIES[103] representations. Unlike the CF-GNNExplainer[101] and MEG[102] methods, the MMACE algorithm ensures that generated molecules are valid, owing to the surjective property of SELFIES. The MMACE method can be applied to both regression and classification models. However, one limitation is that, it does not account for the chemical stability of predicted counterfactuals. To circumvent this drawback, the Wellawatte et al.[9] propose another approach, which lists counterfactuals through a similarity search on the PubChem database[104] instead of a generated chemical space.

### 2.4.1 Similarity to adjacent fields

Tangential examples to counterfactual explanations are adversarial training and matched molecular pairs. Adversarial perturbations are used in training to deceive the model such that the vulnerabilities of the model are exposed.[105,106] Instead counterfactual examples are used to explain the model – an XAI tool applied post-hoc. Therefore, the main difference between adversarial and counterfactual examples are in the application, although both are derived from the same optimization problem.[97] Grabocka et al.[107] have developed a method named Adversarial Training on EXplanations (ATEX) which improves model robustness which improves model's stability to adversarial explanations. While there are conceptual disparities between the similarity between the two, we agree that the counterfactual and adversarial explanations are equivalent mathematical objects.

Matched molecular pairs (MMPs) are a pair of molecules that differ structurally at only one site by a known transformation.[108,109] MMPs are widely used in drug discovery and medicinal chemistry as they facilitate fast and easy understanding of structure-activity relationships.[110–112] Counterfactual and MMP analysis intersect if the structural change is associated with a drastic change in the properties. These MMPs are then counterfactual pairs. In the case the associated changes in the properties are non-significant, they are known as bioisosteres.[113,114] The connection between MMPs and adversarial training examples has also been explored by van Tilborg et al.[115]. MMPs which belong to the counterfactual category are commonly used in outlier and activity cliff detection.[109] This approach is analogous to counterfactual explanations in XAI as the common objective is to uncover if the models have learned non-linear relationships by identifying subgraphs associated with certain properties of the molecules.[67]

## 3 Applications

Model interpretation is certainly not new and a common step in ML in chemistry, but XAI for DL models is becoming more important[56,63–66,70,86,101,102] Here we illustrate some practical examples drawn from our published work on how model-agnostic XAI can be utilized to interpret black-box models and connect the explanations to structure-property relationships. The methods are "Molecular Model Agnostic Counterfactual Explanations" (MMACE)[9] and

"Explaining molecular properties with natural language".[10] Then we demonstrate how counterfactuals and descriptor explanations can propose structure-property relationships in the domain of molecular scent.[116]

### 3.1 Blood-brain barrier permeation prediction

The passive diffusion of drugs from the blood stream to the brain is a critical aspect in drug development and discovery.[117] Predicting if a small molecule can permeate the blood-brain barrier (BBB) is routinely conducted with DL.[118,119] Here we see if we can explain why the DL models work. We trained a random forest (RF) model[120] for this classification task of BBB permeation prediction and generated counterfactuals explanations using the MMACE.[9] Then we generate descriptor explanations to interpret a Gated Recurrent Unit Recurrent Neural Network (GRU-RNN) trained for the same task. Both the models were trained on the dataset developed by Martins et al.[121]. The RF model was implemented in Scikit-learn[122] using Mordred molecular descriptors[123] as the input features. The GRU-RNN model was implemented in Keras.[124] See Wellawatte et al.[9] for complete details

Figure 1 shows generated counterfactuals of a negative example; this molecule is predicted to not cross the BBB. According to the counterfactuals of the instance molecule, we observe that the modifications to the carboxylic acid group enable the example molecule to permeate the BBB. Experimental findings by Fischer et al.[117] show that the BBB permeation of molecules are governed by hydrophobic interactions and surface area. The carboxylic group is a hydrophilic functional group which hinders hydrophobic interactions and addition of atoms enhances the surface area. The counterfactual provides an actionable modification to the molecule to make it cross the BBB.

In Figure 2 we show descriptor explanations generated for Alprozolam, a molecule that permeates the BBB, using the method described by Gandhi and White[10]. We see that predicted permeability positively correlated with the aromaticity of the molecule while negatively correlated with the number of hydrogen bonds donors and acceptors. A similar structure-property relationship for BBB permeability in more mechanistic studies.[125–127] The substructure attributions indicates a reduction in hydrogen bond donors and acceptors. These descriptor explanations are quantitative and interpretable by chemists. Finally, we can use a natural language model to summarize the findings into a written explanation, as shown in the printed text in Figure 2. This matches how a chemist might describe the findings after inspecting the figures and molecular structure.

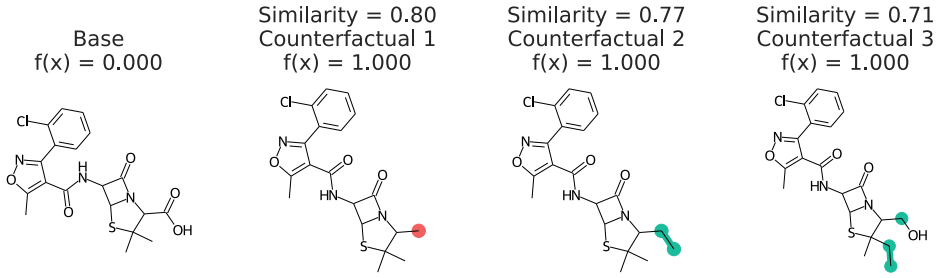


Figure 1: Counterfactuals of a molecule which cannot permeate the blood-brain barrier.Similarity is computed from Tanimoto similarity of ECFP4 fingerprints.[128] Red indicates deletions and green indicates substitutions and addition of atoms. Republished from Ref.[9] with permission from the Royal Society of Chemistry.

### 3.2 Solubility prediction

Small molecule solubility prediction is a classic cheminformatics challenge and is important for chemical process design, drug design and crystallization.[130–133] In our previous works,[9,10] we implemented and trained an RNN model in Keras to predict solubilities (log molarity) of small molecules.[124] The AqSolDB curated database[134] was used to train the model.

We used 6 to define counterfactuals. Figure 3 is the generated local chemical space and the top four counterfactuals of the example molecule. Based on the counterfactuals we observe that the modifications to the ester group and other heteroatoms play an important role in solubility. These findings align with known experimental and basic chemical intuition.[131] Figure 4 shows a more direct measurement of what part of the structure is important. Increasing the polarity by adding acidic and basic groups as well as hydrogen bond acceptors, increases solubility. Substructure importance from ECFP[95] and MACCS[135] descriptors indicate that adding heteroatoms increases solubility while adding more rings makes the molecule less soluble. These are all well-known effects, but it is interesting to see they can be derived purely from the data via DL and XAI.

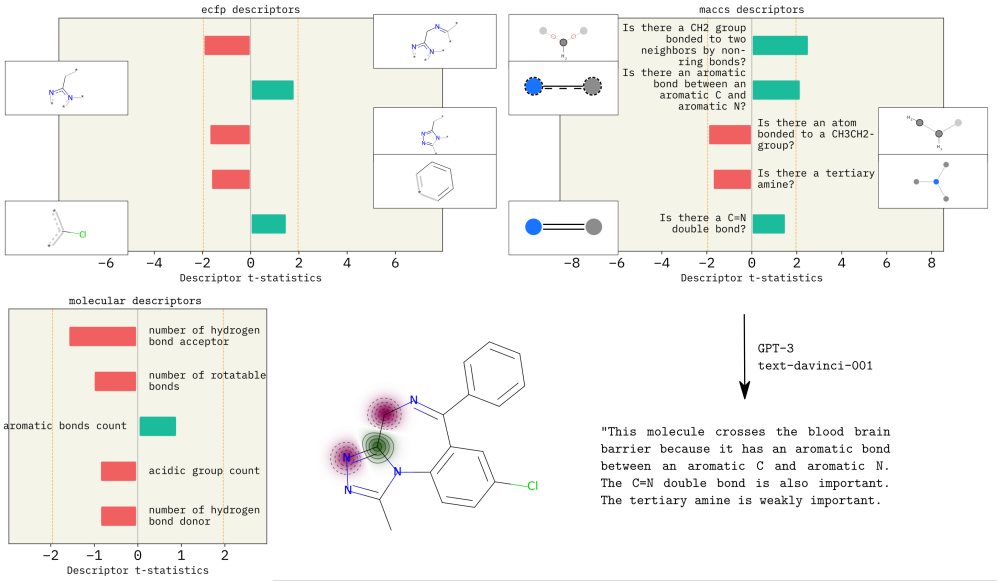


Figure 2: Descriptor explanations along with natural language explanation obtained for BBB permeability of Alprozolam molecule. he green and red bars show descriptors that influence predictions positively and negatively, respectively. Dotted yellow lines show significance threshold ($\alpha = 0.05$) for the t-statistic. Molecular descriptors show molecule-level properties that are important for prediction, and ECFP and MACCS descriptors indicate which substructures influence model predictions. MACCS explanations lead to text explanations as shown. Republished from Ref.[10] with permission from authors.[129]

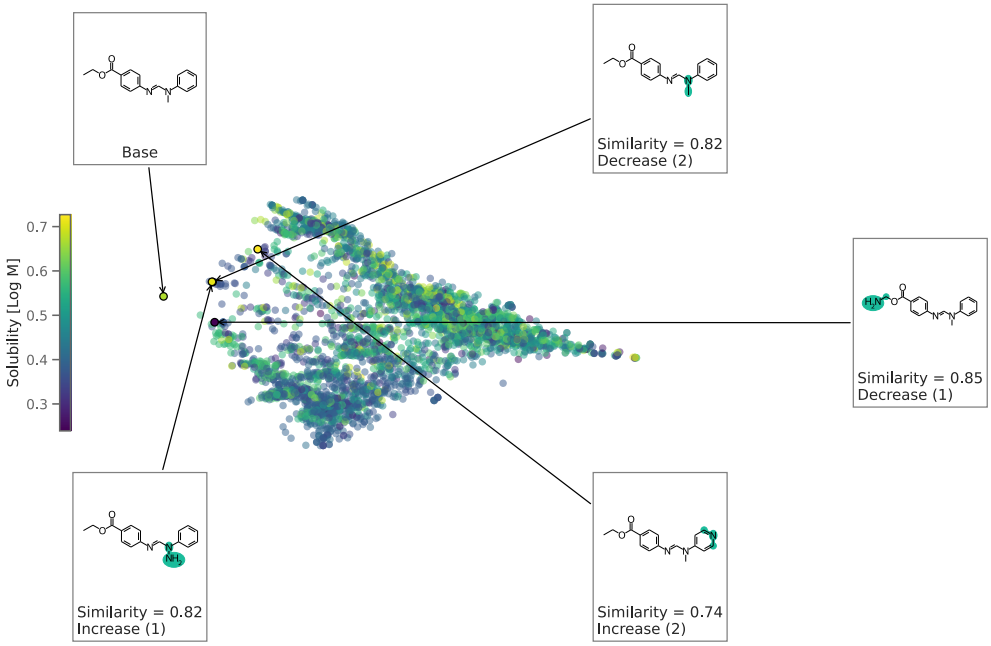


Figure 3: Generated chemical space for solubility prediction using the RNN model. The chemical space is a 2D projection of the pairwise Tanimoto similarities of the local counterfactuals. Each data point is colored by solubility. Top 4 counterfactuals are shown here. Republished from Ref.[9] with permission from the Royal Society of Chemistry.

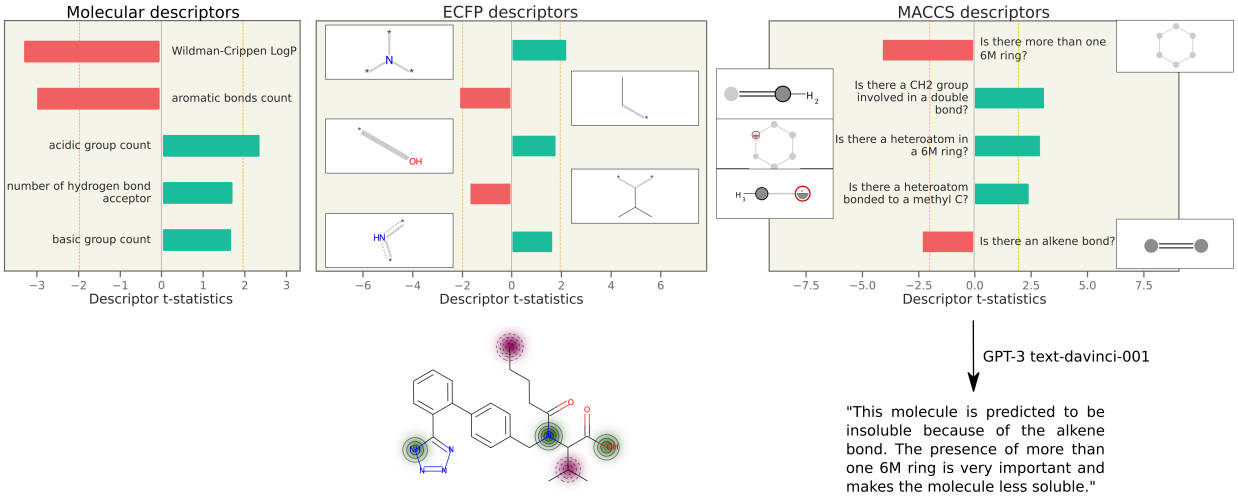


Figure 4: Descriptor explanations for solubility prediction model. The green and red bars show descriptors that influence predictions positively and negatively, respectively. Dotted yellow lines show significance threshold ($\alpha = 0.05$) for the t-statistic. The MACCS and ECFP descriptors indicate 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. MACCS descriptor are used to obtain text explanations as shown. Republished from Ref.[10] with permission from authors.[136]

### 3.3 Generalizing XAI – interpreting scent-structure relationships

Here we show how to learn non-local structure-property relationships with XAI and DL across multiple molecules. Molecular scent prediction is multi-label classification task because a molecule can be described by more than one scent. For example, the molecule jasmone can be described as having 'jasmine,' 'woody,' 'floral,' and 'herbal' scents.[137] The scent-structure relationship is not very well understood,[138] but some relationships are known. For example, molecules with an ester functional group often have a 'fruity' scent. There are some exceptions though, like tert-amyl acetate which has a 'camphoraceous' rather than 'fruity' scent.[138,139]

In Seshadri et al.[116], we trained a GNN model to predict the scent of molecules and utilized counterfactuals and descriptor explanations to quantify scent-structure relationships. The MMACE method was modified to account for the multi-label aspect of scent prediction. This modification defines molecules that differed from the instance molecule by only the selected scent as counterfactuals. For instance, counterfactuals of the jasmone molecule would be false for the 'jasmine' scent but would still be positive for 'woody,' 'floral' and 'herbal' scents.

The molecule 2,4 decadienal, which is known to have a 'fatty' scent is analyzed in Figure 5.[140,141] The resulting counterfactual, which has a shorter carbon chain and no carbonyl group, highlights the influence of these structural features on the 'fatty' scent of 2,4 decadienal. To generalize to other molecules, Seshadri et al.[116] applied the descriptor attribution method to obtain global explanations for the scents. The global explanation for the 'fatty' scent was generated by gathering chemical spaces around many 'fatty' scented molecules. The resulting natural language explanation is: "The molecular property "fatty scent" can be explained by the presence of a heptanyl fragment, two CH2 groups separated by four bonds, and a C=O double bond, as well as the lack of more than one or two O atoms."[116] The importance of a heptanyl fragment aligns with that reported in the literature, as 'fatty' molecules often have a long carbon chain.[142] Furthermore, the importance of a C=O double bond is supported by the findings reported by Licon et al.[143], where in addition to a "larger carbon-chain skeleton", they found that 'fatty' molecules also had "aldehyde or acid functions".[143] For the 'pineapple' scent, the following natural language explanation was obtained: "The molecular property "pineapple scent" can be explained by the presence of ester, ethyl/ether O group, alkene/ether O group, and C=O double bond, as well as the absence of an Aromatic atom."[116] Esters, such as ethyl 2-methylbutyrate, are present in many pineapple volatile compounds.[144,145] The combination of a C=O double bond with an ether could also correspond to an ester group. Additionally, aldehydes and ketones, which contain C=O double bonds, are also common in pineapple volatile compounds.[144,146]

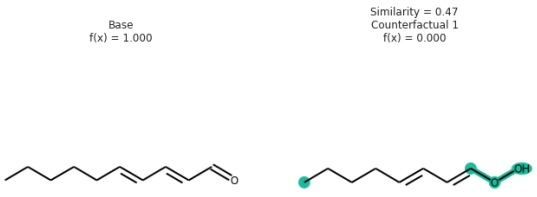


Figure 5: Counterfactual for the 2,4 decadienal molecule. The counterfactual indicates structural changes to ethyl benzoate that would result in the model predicting the molecule to not contain the 'fruity' scent. The Tanimoto[94] similarity between the counterfactual and 2,4 decadienal is also provided. Republished with permission from authors.[116]

## 4 Discussion

We have shown two post-hoc XAI methods based on molecular counterfactual explanations[9] and surrogate models.[10] These methods can be used for classification or regression black box models type whose input is a molecule. The main objective of the two XAI approaches, we present in the review, is to explain black-box models under investigation. Whether this explains the underlying physical phenomena depends strongly on the correctness of the black-box model.

A molecular counterfactual is one with a minimal distance from a base molecular, but with contrasting chemical properties. We used Tanimoto similarity[94] of ECFP4 fingreprints[95] as distance, although this should be explored in the future.. Counterfactual explanations are useful because they are represented as chemical structures (familiar to domain experts), sparse, and are actionable. A few other popular examples of counterfactual on *graph* methods are GNNExplainer, MEG and CF-GNNExplainer.[66,101,102]

The surrogate-model based method developed by Gandhi and White[10] fits a self-explaining model to explain the black-box model. This is similar to the GraphLIME[85] method, although we have the flexibility to use explanation features other than subgraphs. The surrogate-models give natural language and chemical descriptor attributions to create explanations useful for chemists. Lastly, we examined if XAI can be used beyond interpretation. Work by Seshadri et al.[116] use MMACE and surrogate model explanations to analyze the structure-property relationships of scent. They recovered known structure-property relationships for molecular scent purely from explanations, demonstrating the usefulness of a two step process: fit an accurate model and then explain it.

## 5 Conclusion and outlook

We should seek to explain molecular property prediction models because users are more likely to trust explained predictions and explanations can help assess if the model is learning the correct underlying chemical principle. We also showed that black-box modeling first, followed by XAI, is a path to structure-property relationships without needing to trade between accuracy and interepretability. However, XAI in chemistry has some major challenges, that are also related to the black-box nature of the deep learning. Some are highlighted below:

- *Explanation representation*: How is an explanation presented – text, a molecule, attributions, a plot, etc. How do we represent compounds in these explanations - for example, none of the methods above can really account for stereochemistry and thus stereochemistry cannot be proposed as an explanation.
- *Molecular distance*: in XAI approaches such as counterfactual generation, the "distance" between two molecules are minimized. Molecular distance is subjective. Possibilities are distance based on molecular properties, synthesis routes, and direct structure comparisons. Which is best?

- *Regulations*: As black-box models move from research to industry, healthcare, and environmental settings, we expect XAI to become more important to explain decisions to chemists or non-experts and possibly be legally required.
- *Chemical space*: Chemical space is the set of molecules that are realizable; "realizable" can be defined from purchasable to synthesizable to satisfied valences. What is most useful? How can we generate local chemical spaces centered around a specific molecule for finding counterfactuals and explanations? We note that this idea has a connection with the older idea of "activity cliffs" in drug design.[147]
- *Evaluating XAI*: there is a lack of a systematic framework (quantitative or qualitative) to evaluate correctness and applicability of an explanation. Can there be a universal framework, or should explanations be chosen and evaluated based on the audience and domain?

## 6 Acknowledgements

Research reported in this work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM137966. This work was supported by the NSF under awards 1751471 and 1764415. We thank the Center for Integrated Research Computing at the University of Rochester for providing computational resources.

## References

[1] Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon J.L. Billinge, Elizabeth Holm, Shyue Ping Ong, and Chris Wolverton. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1), 2022. ISSN 20573960. doi: 10.1038/s41524-022-00734-6.

[2] John A. Keith, Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Müller, and Alexandre Tkatchenko. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chemical Reviews*, 121(16):9816–9872, 2021. doi: 10.1021/acs.chemrev.1c00107. URL https://doi.org/10.1021/acs.chemrev.1c00107. PMID: 34232033.

[3] Garrett B. Goh, Nathan O. Hodas, and Abhinav Vishnu. Deep learning for computational chemistry. *Journal of Computational Chemistry*, 38(16):1291–1307, 2017. doi: https://doi.org/10.1002/jcc.24764. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.24764.

[4] Volker L. Deringer, Miguel A. Caro, and Gábor Csányi. Machine learning interatomic potentials as emerging tools for materials science. *Advanced Materials*, 31(46):1902765, 2019. doi: https://doi.org/10.1002/adma.201902765. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.201902765.

[5] Felix A. Faber, Luke Hutchison, Bing Huang, Justin Gilmer, Samuel S. Schoenholz, George E. Dahl, Oriol Vinyals, Steven Kearnes, Patrick F. Riley, and O. Anatole von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid dft error. *Journal of Chemical Theory and Computation*, 13(11):5255–5264, 2017. doi: 10.1021/acs.jctc.7b00577. URL https://doi.org/10.1021/acs.jctc.7b00577. PMID: 28926232.

[6] Wlodzislaw Duch, Karthikeyan Swaminathan, and Jaroslaw Meller. Artificial intelligence approaches for rational drug design and discovery. *Current Pharmaceutical Design*, 13:1497–1508, 5 2007. ISSN 13816128. doi: 10.2174/138161207780765954.

[7] Suresh Dara, Swetha Dhamercherla, Surender Singh Jadav, CH Madhu Babu, Mohamed Jawed Ahsan, Suresh Dara darasuresh, and S Dara. Machine learning in drug discovery: A review. *Artificial Intelligence Review*, 55:1947–1999, 123. doi: 10.1007/s10462-021-10058-4. URL https://doi.org/10.1007/s10462-021-10058-4.

[8] Rohan Gupta, Devesh Srivastava, Mehar Sahu, Swati Tiwari, Rashmi K. Ambasta, and Pravir Kumar. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular diversity*, 25:1315–1360, 8 2021. ISSN 1573-501X. doi: 10.1007/S11030-021-10217-3. URL https://pubmed.ncbi.nlm.nih.gov/33844136/.

[9] Geemi P. Wellawatte, Aditi Seshadri, and Andrew D. White. Model agnostic generation of counterfactual explanations for molecules. *Chemical Science*, 13:3697–3705, 3 2022. ISSN 2041-6539. doi: 10.1039/D1SC05259D. URL https://pubs.rsc.org/en/content/articlehtml/2022/sc/d1sc05259dhttps://pubs.rsc.org/en/content/articlelanding/2022/sc/d1sc05259d.

[10] Heta A. Gandhi and Andrew D. White. Explaining structure-activity relationships using locally faithful surrogate models. *chemrxiv*, 5 2022. doi: 10.26434/CHEMRXIV-2022-V5P6M-V2. URL https://chemrxiv.org/engage/chemrxiv/article-details/62753b1459f0d6f41a8642c4.

[11] Adam J Gormley and Michael A Webb. Machine learning in combinatorial polymer chemistry. *Nature Reviews Materials*, 2021. doi: 10.1038/s41578-021-00282-3. URL http://quantum-machine.org/datasets/.

[12] Carla P Gomes, Daniel Fink, R Bruce Van Dover, and John M Gregoire. Computational sustainability meets materials science. *Nature Reviews Materials*, 2021. doi: 10.1038/s41578-021-00348-2. URL https://ebird.org.

[13] 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, 10 2019. ISSN 15662535. doi: 10.48550/arxiv.1910.10045. URL https://arxiv.org/abs/1910.10045v2.

[14] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Interpretable machine learning: definitions, methods, and applications. *ArXiv*, abs/1901.04592, 2019.

[15] John D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors*, 46:50–80, 8 2004. ISSN 00187208. doi: 10.1518/hfes.46.1.50_30392. URL https://journals.sagepub.com/doi/10.1518/hfes.46.1.50_30392?url_ver=Z39.88-2003&rfr_id=ori%3Arid%3Acrossref.org&rfr_dat=cr_pub++0pubmed.

[16] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[17] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR, 23–24 Feb 2018. URL https://proceedings.mlr.press/v81/buolamwini18a.html.

[18] 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*, 10(1):1–8, 2019.

[19] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

[20] Samuel Boobier, Anne Osbourn, and John BO Mitchell. Can human experts predict solubility better than computers? *Journal of cheminformatics*, 9(1):1–14, 2017.

[21] there does happen to be one human solubility savant, participant 11, who matched machine performance.

[22] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[23] ARTIFICIAL INTELLIGENCE ACT. European commission. *On Artificial Intelligence: A European Approach to Excellence and Trust.*, page COM/2021/206, 2021.

[24] Blueprint for an ai bill of rights, the white house, 2022. URL https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

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

[26] W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, 116:22071–22080, 10 2019. ISSN 10916490. doi: 10.1073/PNAS.1900654116/SUPPL_FILE/PNAS.1900654116.SAPP.PDF. URL https://www.pnas.org/doi/abs/10.1073/pnas.1900654116.

[27] David Gunning and David Aha. Darpa's explainable artificial intelligence (xai) program. *AI Magazine*, 40: 44–58, 06 2019. doi: 10.1609/aimag.v40i2.2850.

[28] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8, pages 8–13, 2017.

[29] Sebastian Palacio, Adriano Lucieri, Mohsin Munir, Sheraz Ahmed, Jörn Hees, and Andreas Dengel. Xai handbook: Towards a unified framework for explainable ai. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3766–3775, 2021.

[30] D. Richard Kuhn, Raghu N Kacker, Yu Lei, and Dimitris E. Simos. Combinatorial methods for explainable ai. *2020 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pages 167–170, 2020.

[31] Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, et al. Selfies and the future of molecular string representations. *Patterns*, 3(10):100588, 2022.

[32] 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(4):045024, 2020.

[33] 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. *Chemical science*, 12:7079–7090, 2021.

[34] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. Evaluating xai: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, 291:103404, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2020.103404. URL `https://www.sciencedirect.com/science/article/pii/S0004370220301533`.

[35] Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.

[36] Arun Das and Paul Rad. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*, 2020.

[37] R. Machlev, L. Heistrene, M. Perl, K. Y. Levy, J. Belikov, S. Mannor, and Y. Levron. Explainable artificial intelligence (xai) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI*, 9:100169, 8 2022. ISSN 2666-5468. doi: 10.1016/J.EGYAI.2022.100169.

[38] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pages 1885–1894. PMLR, 2017.

[39] 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*, pages 1135–1144, San Diego, CA, USA, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-3020. URL `https://aclanthology.org/N16-3020`.

[40] 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. *Advances in neural information processing systems*, 31, 2018.

[41] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. Evaluating explainable ai on a multi-modal medical imaging task: Can existing algorithms fulfill clinical requirements? *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):11945–11953, Jun. 2022. doi: 10.1609/aaai.v36i11.21452. URL `https://ojs.aaai.org/index.php/AAAI/article/view/21452`.

[42] Felipe Oviedo, Juan Lavista Ferres, Tonio Buonassisi, and Keith T. Butler. Interpretable and explainable machine learning for materials science and chemistry. *Accounts of Materials Research*, 3(6):597–607, 2022. doi: 10.1021/accountsmr.1c00244. URL `https://doi.org/10.1021/accountsmr.1c00244`.

[43] Yuyi Zhang, Feiran Xu, Jingying Zou, Ovanes L Petrosian, and Kirill V Krinkin. Xai evaluation: Evaluating black-box model explanations for prediction. In *2021 II International Conference on Neural Networks and Neurotechnologies (NeuroNT)*, pages 13–16. IEEE, 2021.

[44] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

[45] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning–a brief history, state-of-the-art and challenges. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 417–431. Springer, 2020.

[46] Agnieszka Gajewicz, Tomasz Puzyn, Katarzyna Odziomek, Piotr Urbaszek, Andrea Haase, Christian Riebeling, Andreas Luch, Muhammad A. Irfan, Robert Landsiedel, Meike van der Zande, and Hans Bouwmeester. Decision tree models to classify nanomaterials according to the df4nanogrouping scheme. *Nanotoxicology*, 12:1–17,

1 2018. ISSN 17435404. doi: 10.1080/17435390.2017.1415388/SUPPL_FILE/INAN_A_1415388_SM4247.ZIP. URL https://www.tandfonline.com/doi/abs/10.1080/17435390.2017.1415388.

[47] Lianyi Han, Yanli Wang, and Stephen H. Bryant. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high–throughput screening data in pubchem. *BMC Bioinformatics*, 9:401, 9 2008. ISSN 14712105. doi: 10.1186/1471-2105-9-401. URL /pmc/articles/PMC2572623//pmc/articles/PMC2572623/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2572623/.

[48] Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M Ghiringhelli. Sisso: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2(8):083802, 2018.

[49] Gregory Plumb, Maruan Al-Shedivat, Ángel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. Regularizing black-box models for improved interpretability. *Advances in Neural Information Processing Systems*, 33:10526–10536, 2020.

[50] Xiaoting Shao, Arseny Skryagin, Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for better reasons: Training differentiable models by constraining their influence functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9533–9540, 2021.

[51] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158, 2012.

[52] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504*, 2017.

[53] Tobias Harren, Hans Matter, Gerhard Hessler, Matthias Rarey, and Christoph Grebner. Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence. *Journal of Chemical Information and Modeling*, 62(3):447–462, 2022.

[54] Maria H. Rasmussen, Diana S. Christensen, and Jan H. Jensen. Do machines dream of atoms? crippen's logp as a quantitative molecular benchmark for explainable ai heatmaps. 12 2022. doi: 10.26434/CHEMRXIV-2022-GNQ3W-V2. URL https://chemrxiv.org/engage/chemrxiv/article-details/6388991244ccbc1731090a96.

[55] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise, 2017. URL https://arxiv.org/abs/1706.03825.

[56] Kevin McCloskey, Ankur Taly, Federico Monti, Michael P. Brenner, and Lucy Colwell. Using attribution to decode dataset bias in neural network models for chemistry. *Proceedings of the National Academy of Sciences of the United States of America*, 116:11624–11629, 11 2018. doi: 10.1073/pnas.1820657116. URL https://arxiv.org/abs/1811.11310v3.

[57] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[58] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[59] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[60] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015. URL https://arxiv.org/abs/1512.04150.

[61] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, oct 2019. doi: 10.1007/s11263-019-01228-7. URL https://doi.org/10.1007%2Fs11263-019-01228-7.

[62] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[63] Phillip Pope, Soheil Kolouri, Mohammad Rostrami, Charles Martin, and Heiko Hoffmann. Discovering molecular functional groups using graph convolutional neural networks, 2018. URL https://arxiv.org/abs/1812.00265.

[64] 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 (3):1083–1094, 2021.

[65] Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Y. Wang, Wesley Wei Qian, Kevin McCloskey, Lucy Colwell, and Alexander Wiltschko. Evaluating attribution for graph neural networks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

[66] Rex Ying, Dylan Bourgeois, Jiaxuan You, M. Zitnik, and J. Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32:9240–9251, 2019.

[67] Jiahua Rao, Shuangjia Zheng, and Yuedong Yang. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *arXiv preprint arXiv:2107.04119*, 2021.

[68] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12241–12252. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/yuan21c.html.

[69] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. 2017. doi: 10.48550/arxiv.1704.02685. URL https://arxiv.org/abs/1704.02685.

[70] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus Robert Müller. Layer-wise relevance propagation: An overview. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11700 LNCS:193–209, 2019. ISSN 16113349. doi: 10.1007/978-3-030-28954-6_10/FIGURES/5. URL https://link.springer.com/chapter/10.1007/978-3-030-28954-6_10.

[71] Federico Baldassarre and Hossein Azizpour. Explainability techniques for graph convolutional networks, 2019. URL https://arxiv.org/abs/1905.13686.

[72] Lloyd S. Shapley. *A Value for N-Person Games*. RAND Corporation, Santa Monica, CA, 1952. doi: 10.7249/P0295.

[73] Andrew D. White. Deep learning for molecules and materials. *Living Journal of Computational Molecular Science*, 3, 2022. doi: 10.33011/LIVECOMS.3.1.1499.

[74] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41:647–665, 11 2014. ISSN 02193116. doi: 10.1007/S10115-013-0679-X/TABLES/4. URL https://link.springer.com/article/10.1007/s10115-013-0679-x.

[75] Joshua Hochuli, Alec Helbling, Tamar Skaist, Matthew Ragoza, and David Ryan Koes. Visualizing convolutional neural network protein-ligand scoring. *Journal of Molecular Graphics and Modelling*, 84:96–108, 2018. ISSN 1093-3263. doi: https://doi.org/10.1016/j.jmgm.2018.06.005. URL https://www.sciencedirect.com/science/article/pii/S1093326318301670.

[76] 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(16):8761–8777, 2020. doi: 10.1021/acs.jmedchem.9b01101. URL https://doi.org/10.1021/acs.jmedchem.9b01101. PMID: 31512867.

[77] 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.

[78] Andrea Mastropietro, Giuseppe Pasculli, Christian Feldmann, Raquel Rodríguez-Pérez, and Jürgen Bajorath. Edgeshaper: Bond-centric shapley value-based explanation method for graph neural networks. *iScience*, 25:105043, 10 2022. ISSN 25890042. doi: 10.1016/j.isci.2022.105043. URL http://www.cell.com/article/S2589004222013153/fulltexthttp://www.cell.com/article/S2589004222013153/abstracthttps://www.cell.com/iscience/abstract/S2589-0042(22)01315-3.

[79] Dumitru Erhan, Y Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *Technical Report, Univeristé de Montréal*, 9 2009.

[80] Jeffrey K. Weber, Joseph A. Morrone, Sugato Bagchi, Jan D.Estrada Pabon, Seung gu Kang, Leili Zhang, and Wendy D. Cornell. Simplified, interpretable graph convolutional neural networks for small molecule activity prediction. *Journal of Computer-Aided Molecular Design*, 36:391–404, 5 2022. ISSN 15734951. doi:

10.1007/S10822-021-00421-6/FIGURES/10. URL https://link.springer.com/article/10.1007/s10822-021-00421-6.

[81] Sereina Riniker and Gregory A. Landrum. Similarity maps - a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics*, 5:1–7, 9 2013. ISSN 17582946. doi: 10.1186/1758-2946-5-43/FIGURES/5. URL https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-5-43.

[82] 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. *Journal of Cheminformatics*, 14:1–14, 12 2022. ISSN 17582946. doi: 10.1186/S13321-022-00600-Z/FIGURES/10. URL https://link.springer.com/articles/10.1186/s13321-022-00600-zhttps://link.springer.com/article/10.1186/s13321-022-00600-z.

[83] Thomas McGrath, Andrei Kapishnikov, Nenad Tomašev, Adam Pearce, Martin Wattenberg, Demis Hassabis, Been Kim, Ulrich Paquet, and Vladimir Kramnik. Acquisition of chess knowledge in alphazero. *Proceedings of the National Academy of Sciences*, 119(47):e2206625119, 2022. doi: 10.1073/pnas.2206625119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2206625119.

[84] 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.

[85] Qiang Huang, Makoto Yamada, Yuan Tian, Dinesh Singh, Dawei Yin, and Yi Chang. Graphlime: Local interpretable model explanations for graph neural networks. *CoRR*, abs/2001.06216, 2020. URL https://arxiv.org/abs/2001.06216.

[86] Kacper Sokol and Peter A. Flach. Limetree: Interactively customisable explanations based on local surrogate multi-output regression trees. *CoRR*, abs/2005.01427, 2020. URL https://arxiv.org/abs/2005.01427.

[87] 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.

[88] Shams Mehdi and Pratyush Tiwary. Thermodynamics of interpretation. 6 2022. doi: 10.48550/arxiv.2206.13475. URL https://arxiv.org/abs/2206.13475v1.

[89] M. Höfler. Causal inference based on counterfactuals. *BMC Medical Research Methodology*, 5:1–12, 9 2005. ISSN 14712288. doi: 10.1186/1471-2288-5-28/COMMENTS. URL https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-5-28.

[90] James Woodward and Christopher Hitchcock. Explanatory generalizations, part i: A counterfactual account. *Noûs*, 37(1):1–24, 2003. doi: https://doi.org/10.1111/1468-0068.00426. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0068.00426.

[91] Mathias Florian Frisch. *Theories, models, and explanation*. University of California, Berkeley, 1998.

[92] Alexander Reutlinger. Is there a monist theory of causal and non-causal explanations? the counterfactual theory of scientific explanation. *Philosophy of Science*, 83(5):733–745, 2016. doi: 10.1086/687859.

[93] David Lewis. Causation. *The journal of philosophy*, 70(17):556–567, 1974.

[94] Taffee T Tanimoto. Elementary mathematical theory of classification and prediction. *Internal IBM Technical Report*, 1958.

[95] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t. PMID: 20426451.

[96] Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Kate Scott, Stuart Schieber, James Waldo, David Weinberger, Adrian Weller, and Alexandra Wood. Accountability of ai under the law: The role of explanation. *SSRN Electronic Journal*, 11 2017. doi: 10.48550/arxiv.1711.01134. URL https://arxiv.org/abs/1711.01134v3.

[97] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[98] José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence 2020 2:10*, 2:573–584, 10 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00236-4. URL https://www.nature.com/articles/s42256-020-00236-4.

[99] Tianfan Fu, Wenhao Gao, Cao Xiao, Jacob Yasonik, Connor W. Coley, and Jimeng Sun. Differentiable scaffolding tree for molecule optimization. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=w_drCosT76.

[100] Cynthia Shen, Mario Krenn, Sagi Eppel, and Alán Aspuru-Guzik. Deep molecular dreaming: inverse machine learning for de-novo molecular design and interpretability with surjective representations. *Machine Learning: Science and Technology*, 2:03LT02, 9 2021. ISSN 2632-2153. doi: 10.1088/2632-2153/ac09d6. URL https://iopscience.iop.org/article/10.1088/2632-2153/ac09d6.

[101] A. Lucic, Maartje ter Hoeve, Gabriele Tolomei, M. Rijke, and F. Silvestri. Cf-gnnexplainer: Counterfactual explanations for graph neural networks. *arXiv preprint arXiv:2102.03322*, 2021.

[102] Danilo Numeroso and D. Bacciu. Explaining deep graph networks with molecular counterfactuals. *arXiv preprint arXiv:2011.05134*, 2020.

[103] 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(4):045024, 2020.

[104] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research*, 49(D1):D1388–D1395, 11 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa971. URL https://doi.org/10.1093/nar/gkaa971.

[105] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines, and Mounia Lalmas. Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 465–474, 2017.

[106] Timo Freiesleben. The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines*, 32(1):77–109, 2022.

[107] Josif Grabocka, Nicolas Schilling, Martin Wistuba, and Lars Schmidt-Thieme. Learning time-series shapelets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 392–401, 2014.

[108] Peter W Kenny and Jens Sadowski. Structure modification in chemical databases. *Chemoinformatics in drug discovery*, pages 271–285, 2005.

[109] Christian Tyrchan and Emma Evertsson. Matched molecular pair analysis in short: Algorithms, applications and limitations. *Computational and Structural Biotechnology Journal*, 15:86–90, 2017. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2016.12.003. URL https://www.sciencedirect.com/science/article/pii/S2001037016300848.

[110] Ed Griffen, Andrew G. Leach, Graeme R. Robb, and Daniel J. Warner. Matched molecular pairs as a medicinal chemistry tool. *Journal of Medicinal Chemistry*, 54(22):7739–7750, 2011. doi: 10.1021/jm200452d. URL https://doi.org/10.1021/jm200452d. PMID: 21936582.

[111] Jiazhen He, Eva Nittinger, Christian Tyrchan, Werngard Czechtizky, Atanas Patronov, Esben Jannik Bjerrum, and Ola Engkvist. Transformer-based molecular optimization beyond matched molecular pairs. *Journal of cheminformatics*, 14(1):1–14, 2022.

[112] Junhui Park, Gaeun Sung, SeungHyun Lee, SeungHo Kang, and ChunKyun Park. Acgcn: Graph convolutional networks for activity cliff prediction between matched molecular pairs. *Journal of Chemical Information and Modeling*, 2022.

[113] Sarah R. Langdon, Peter Ertl, and Nathan Brown. Bioisosteric replacement and scaffold hopping in lead generation and optimization. *Molecular Informatics*, 29(5):366–385, 2010. doi: https://doi.org/10.1002/minf.201000019. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/minf.201000019.

[114] Samo Turk, Benjamin Merget, Friedrich Rippmann, and Simone Fulle. Coupling matched molecular pairs with machine learning for virtual compound optimization. *Journal of Chemical Information and Modeling*, 57(12):3079–3085, 2017. doi: 10.1021/acs.jcim.7b00298. URL https://doi.org/10.1021/acs.jcim.7b00298. PMID: 29131617.

[115] Derek van Tilborg, Alisa Alenicheva, and Francesca Grisoni. Exposing the limitations of molecular machine learning with activity cliffs. 2022.

[116] Aditi Seshadri, Heta A. Gandhi, Geemi P. Wellawatte, and Andrew D. White. Why does that molecule smell? *ChemRxiv*, 2022. doi: 10.26434/chemrxiv-2022-vs838.

[117] H. Fischer, R. Gottschlich, and A. Seelig. Blood-brain barrier permeation: molecular parameters governing passive diffusion. *The Journal of membrane biology*, 165:201–211, 1998. ISSN 0022-2631. doi: 10.1007/S002329900434. URL https://pubmed.ncbi.nlm.nih.gov/9767674/.

[118] Lili Liu, Li Zhang, Huawei Feng, Shimeng Li, Miao Liu, Jian Zhao, and Hongsheng Liu. Prediction of the blood–brain barrier (bbb) permeability of chemicals based on machine-learning and ensemble methods. *Chemical Research in Toxicology*, 34(6):1456–1467, 2021. doi: 10.1021/acs.chemrestox.0c00343. URL https://doi.org/10.1021/acs.chemrestox.0c00343. PMID: 34047182.

[119] Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9 (2):513–530, 2018.

[120] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

[121] Ines Filipa Martins, Ana L Teixeira, Luis Pinheiro, and Andre O Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling*, 52(6):1686–1697, 2012.

[122] 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.

[123] Hirotomo Moriwaki, Yu-Shi Tian, Norihito Kawashita, and Tatsuya Takagi. Mordred: a molecular descriptor calculator. *Journal of cheminformatics*, 10(1):1–14, 2018.

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

[125] Travis T Wager, Ramalakshmi Y Chandrasekaran, Xinjun Hou, Matthew D Troutman, Patrick R Verhoest, Anabella Villalobos, and Yvonne Will. Defining Desirable Central Nervous System Drug Space through the Alignment of Molecular Properties, in Vitro ADME, and Safety Attributes. *ACS Chemical Neuroscience*, 1(6): 420–434, 2010. doi: 10.1021/cn100007x. URL https://doi.org/10.1021/cn100007x.

[126] Arup K Ghose, Torsten Herbertz, Robert L Hudkins, Bruce D Dorsey, and John P Mallamo. Knowledge-Based, Central Nervous System (CNS) Lead Selection and Lead Optimization for CNS Drug Discovery. *ACS Chemical Neuroscience*, 3(1):50–68, 2012. doi: 10.1021/cn200100h. URL https://doi.org/10.1021/cn200100h.

[127] Pavel Polishchuk, Oleg Tinkov, Tatiana Khristova, Ludmila Ognichenko, Anna Kosinskaya, Alexandre Varnek, and Victor Kuz'min. Structural and Physico-Chemical Interpretation (SPCI) of QSAR Models and Its Comparison with Matched Molecular Pair Analysis. *Journal of Chemical Information and Modeling*, 56(8):1455–1469, 2016. doi: 10.1021/acs.jcim.6b00371. URL https://doi.org/10.1021/acs.jcim.6b00371.

[128] Moises Hassan, Robert D Brown, Shikha Varma-O'Brien, and David Rogers. Cheminformatics analysis and learning in a data pipelining environment. *Molecular diversity*, 10(3):283–299, 2006.

[129] 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.[148].

[130] Ehsan Sheikholeslamzadeh and Sohrab Rohani. Solubility prediction of pharmaceutical and chemical compounds in pure and mixed solvents using predictive models. *Industrial & engineering chemistry research*, 51 (1):464–473, 2012.

[131] 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 2020 11:1*, 11: 1–10, 11 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-19594-z. URL https://www.nature.com/articles/s41467-020-19594-z.

[132] Christoph Loschen and Andreas Klamt. Solubility prediction, solvate and cocrystal screening as tools for rational crystal engineering. *Journal of Pharmacy and Pharmacology*, 67(6):803–811, 2015.

[133] Louis J Diorazio, David RJ Hose, and Neil K Adlington. Toward a more holistic framework for solvent selection. *Organic Process Research & Development*, 20(4):760–773, 2016.

[134] Murat Cihan 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):1–8, 2019.

[135] Joseph L Durant, Burton A Leland, Douglas R Henry, and James G Nourse. Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280, 2002.

[136] 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.[148].

[137] National Center for Biotechnology Information. Pubchem compound summary for cid 1549018, jasmone. URL https://pubchem.ncbi.nlm.nih.gov/compound/Jasmone. Accessed September 26, 2022.

[138] Charles S Sell. On the unpredictability of odor. *Angewandte Chemie International Edition*, 45(38):6254–6261, 2006.

[139] Manon Genva, Tierry Kenne Kemene, Magali Deleu, Laurence Lins, and Marie-Laure Fauconnier. Is it possible to predict the odor of a molecule on the basis of its structure? *International journal of molecular sciences*, 20 (12):3018, 2019.

[140] D Rowe. Aroma chemicals for savory flavors. *Perfumer and Flavorist*, 23:9–18, 1998.

[141] Silvia Mallia, Felix Escher, and Hedwig Schlichtherle-Cerny. Aroma-active compounds of butter: a review. *European Food Research and Technology*, 226(3):315–325, 2008.

[142] Henryk Jelen and Anna Gracka. Characterization of aroma compounds: Structure, physico-chemical and sensory properties. *Flavour: From food to perception*, pages 126–153, 2016.

[143] Carmen C Licon, Guillaume Bosc, Mohammed Sabri, Marylou Mantel, Arnaud Fournel, Caroline Bushdid, Jerome Golebiowski, Celine Robardet, Marc Plantevit, Mehdi Kaytoue, et al. Chemical features mining provides new descriptive structure-odor relationships. *PLoS computational biology*, 15(4):e1006945, 2019.

[144] Salma Mostafa, Yun Wang, Wen Zeng, and Biao Jin. Floral scents and fruit aromas: Functions, compositions, biosynthesis, and regulation. *Frontiers in plant science*, 13, 2022.

[145] Yukiko Tokitomo, Martin Steinhaus, Andrea Büttner, and Peter Schieberle. Odor-active constituents in fresh pineapple (ananas comosus [l.] merr.) by quantitative and sensory evaluation. *Bioscience, Biotechnology, and Biochemistry*, 69(7):1323–1330, 2005.

[146] Chang-Bin Wei, Sheng-Hui Liu, Yu-Ge Liu, Ling-Ling Lv, Wen-Xiu Yang, and Guang-Ming Sun. Characteristic aroma compounds from different pineapple parts. *Molecules*, 16(6):5104–5112, 2011.

[147] Dagmar Stumpfe and Jürgen Bajorath. Exploring activity cliffs in medicinal chemistry. *Journal of Medicinal Chemistry*, 55(7):2932–2942, 2012. doi: 10.1021/jm201706b. URL https://doi.org/10.1021/jm201706b. PMID: 22236250.

[148] 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.