A Perspective on Explanations of Molecular Prediction Models

Geemi P. Wellawatte,[†] Heta A. Gandhi,[‡] Aditi Seshadri,[‡] and Andrew D. White[*,‡]

[†]*Department of Chemistry, University of Rochester, Rochester, NY, 14627*

[‡]*Department of Chemical Engineering, University of Rochester, Rochester, NY, 14627*

[¶]*Vial Health Technology, Inc., San Francisco, CA 94111*

E-mail: andrew.white@rochester.edu

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 on methods developed by our group and their applications in predicting solubility, blood-brain barrier permeability, and the scent of molecules. We show that XAI methods like chemical counterfactuals and descriptor explanations can explain DL predictions while giving insight into structure-property relationships. Finally, we discuss how a two-step process of developing a black-box model and explaining predictions can uncover structure-property relationships.

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] According to Kre[13], DL models can contribute to scientific discovery in three "dimensions" - 1) as a 'computational microscope' to gain insight which are not attainable through experiments 2) as a 'resource of inspiration' to motivate scientific thinking 3) as an 'agent of understanding' to uncover new observations. However, the rationale of a DL prediction is not always apparent due to the model architecture consisting a large parameter count.[14,15] DL models are thus often termed"black box" models. We can only reason about the input and output of an DL model, not the underlying cause that leads to a specific prediction.

It is routine in chemistry now for DL to exceed human level performance — humans are not good at predicting solubility from structure for example[161] — 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. However, the black box nature of DL has its own limitations. Users are more likely to trust and use predictions from a model if they can understand why the prediction was made.[17] Explaining predictions can help developers of DL models ensure the model is not learning spurious correlations.[18,19] Two infamous examples are, 1)neural networks that learned to recognize horses by looking for a photographer's watermark[20] and, 2) neural networks that predicted a COVID-19 diagnoses by looking at the font choice on medical images.[21] As a result, there is an emerging regulatory framework for when any computer algorithms impact humans.[22–24] Although we know of no examples yet in chemistry, one can assume the use of AI in predicting toxicity, carcinogenicity, and environmental persistence will require rationale for the predictions due to regulatory consequences.

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

EXplainable Artificial Intelligence (XAI) is a field of growing importance that aims to provide model interpretations of DL predictions Three terms highly associated with XAI are, interpretability, justifications and explainability. 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.[14] 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 gives the context and a cause for one or more predictions.[29] We adopt the same nomenclature in this perspective.

Accuracy and interpretability are two attractive characteristics of DL models. However, DL models are often highly accurate and less interpretable.[28,30] XAI provides a way to avoid that trade-off in chemical property prediction. XAI can be viewed as a two-step process. First, we develop an accurate but uninterpretable DL model. Next, we add explanations to predictions. Ideally, if the DL model has correctly learned the input-output relations, 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 while drawing specific examples from our recent XAI work.[9,10,31] We show how in various systems these methods yield explanations that are consistent with known and mechanisms in structure-property relationships.

Theory

In this work, we aim to assemble a common taxonomy for the landscape of XAI while providing our perspectives. We utilized the vocabulary proposed by Das and Rad [32] to classify XAI. According to their classification, interpretations can be categorized as global or local interpretations on the basis of "what is being explained?". For example, counterfactuals are local interpretations, as these can explain only a given instance. The second classification is based on the relation between the model and the interpretation – is interpretability post-hoc (extrinsic) or intrinsic to the model?.[32,33] An intrinsic XAI method is part of the model and is self-explanatory[32] These are also referred to as white-box models to contrast them with non-interpretable black box models.[28] An extrinsic method is one that can be applied post-training to any model.[33] Post-hoc methods found in the literature focus on interpreting models through 1) training data[34] and feature attribution,[35] 2) surrogate models[10] and, 3) counterfactual[9] or contrastive explanations.[36]

Often, what is a "good" explanation and what are the required components of an explanation are debated.[32,37,38] Palacio et al.[29] state that the lack of a standard framework has caused the inability to evaluate the interpretability of a model. In physical sciences, we may instead consider if the explanations somehow reflect and expand our understanding of physical phenomena. For example, Oviedo et al.[39] propose that a model explanation can be evaluated by considering its agreement with physical observations, which they term "correctness." For example, if an explanation suggests that polarity affects solubility of a molecule, and the experimental evidence strengthen the hypothesis, then the explanation is assumed "correct". In instances where such mechanistic knowledge is sparse, expert biases and subjectivity can be used to measure the correctness.[40] Other similar metrics of correctness such as "explanation satisfaction scale" can be found in the literature.[41,42] In a recent study, Humer et al.[43] introduced CIME an interactive web-based tool that allows the users to inspect model explanations. The aim of this study is to bridge the gap between analysis of XAI methods. Based on the above discussion, we identify that an agreed upon

evaluation metric is necessary in XAI. We suggest the following attributes can be used to evaluate explanations. However, the relative importance of each attribute may depend on the application - actionability may not be as important as faithfulness when evaluating the interpretability of a static physics based model. Therefore, one can select relative importance of each attribute based on the application.

- *Actionable.* Is it clear how we could change the input features to modify the output?
- *Complete.* Does the explanation completely account for the prediction? Did features not included in the explanation really contribute zero effect to the prediction? [44]
- *Correct.* Does the explanation agree with hypothesized or known underlying physical mechanism? [39]
- *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?

We present an example evaluation of the SHAP explanation method based on the above attributes. [44] Shapley values were proposed as a local explanation method based on feature attribution, as they offer a complete explanation - each feature is assigned a fraction of the prediction value. [44,45] Completeness is a clearly measurable and well-defined metric, but yields explanations with many components. Yet Shapley values are not actionable nor sparse. They are non-sparse as every feature has a non-zero attribution and not-actionable because they do not provide a set of features which changes the outcome. [46] Ribeiro et al. [35] proposed a surrogate model method that aims to provide sparse/succinct explanations that have high

fidelity to the original model. In Wellawatte et al.[9] we argue that counterfactuals are "better" explanations because they are actionable and sparse. We highlight that, evaluation of explanations is a difficult task because explanations are fundamentally for and by humans. Therefore, these evaluations are subjective, as they depend on "complex human factors and application scenarios."[37]

Self-explaining models

A self-explanatory model is one that is intrinsically interpretable to an expert.[47] Two common examples found in the literature are linear regression models and decision trees (DT). Intrinsic models can be found in other XAI applications acting as surrogate models (proxy models) due to their transparent nature.[48,49] 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 $\hat{y}$. Therefore, we observe that the weights can be used to derive a complete explanation of the model - trained weights quantify the importance of each feature.[47] DT models are another type of self-explaining models which have been used in classification and high-throughput screening tasks. Gajewicz et al.[50] used DT models to classify nanomaterials that identify structural features responsible for surface activity. In another study by Han et al.[51], a DT model was developed to filter compounds by their bioactivity based on the chemical fingerprints.

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

Regularization techniques such as EXPO[52] and RRR[53] are designed to enhance the black-box model interpretability.[54] 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[55] 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 DT's 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. On the other hand, a linear model is not innately accurate as they fail to capture non-linear relationships in data, limiting is applicability. Similarly, a DT is a rule-based model and lacks physics informed knowledge. Therefore, an existing drawback is the trade-off between the degree of understandability and the accuracy of a model. For example, an intrinsic model (linear regression or decision trees) can be described through the trainable parameters, but it may fail to "correctly" capture non-linear relations in the data.

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. Feature attributions provide local explanations, but can be averaged or combined to explain multiple instances. Atom-based numerical assignments are commonly referred to as heatmaps.[56] Sheridan[57] describes an atom-wise attribution method for interpreting QSAR models. Recently, Rasmussen et al.[58] showed that Crippen logP models serve as a benchmark for heatmap approaches. Other most widely used feature attribution approaches in the literature are gradient based methods,[59,60] Shapley Additive exPlanations (SHAP),[44] and layer-wise relevance prorogation.[61]

Gradient based approaches are based on the hypothesis that gradients for neural networks are analogous to coefficients for regression models.[62] Class activation maps (CAM),[63] gradCAM,[64] smoothGrad,,[65] and integrated gradients[62] are examples of this method. The main idea behind feature attributions with gradients can be represented with equation 2.

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

where $\hat{f}(x)$ is the black-box model and $\frac{\Delta \hat{f}(\vec{x})}{\Delta x_i}$ are used as our attributions. The left-hand side of equation 2 says that we attribute each input feature x_i by how much one unit change in it would affect the output of $\hat{f}(x)$. If $\hat{f}(x)$ is a linear surrogate model, then this method reconciles with LIME.[35] In DL models, $\nabla_x f(x)$, suffers from the shattered gradients problem.[62] This means directly computing the quantity leads to numeric problems. The different gradient based approaches are mostly distinguishable based on how the gradient is approximated.

Gradient based explanations have been widely used to interpret chemistry predictions.[60,66–70] McCloskey et al.[60] used graph convolutional networks (GCNs) to predict protein-ligand binding and explained the binding logic for these predictions using integrated gradients. Pope et al.[66] and Jiménez-Luna et al.[67] show application of gradCAM and integrated gradients to explain molecular property predictions from trained graph neural networks (GNNs). Sanchez-Lengeling et al.[68] present comprehensive, open-source XAI benchmarks to explain GNNs and other graph based models. They compare the performance of class activation maps (CAM),[63] gradCAM,[64] smoothGrad,,[65] integrated gradients[62] and attention mechanisms for explaining outcomes of classification as well as regression tasks. They concluded that CAM and integrated gradients perform well for graph based models. Another attempt at creating XAI benchmarks for graph models was made by Rao et al.[70]. They compared these gradient based methods to find subgraph importance when predicting activity cliffs and concluded that gradCAM and integrated gradients provided the most interpretability for GNNs. The GNNExplainer[69] is an approach for generating explanations (local and global) for graph based models. This method focuses on identifying which sub-graphs contribute most to the prediction by maximizing mutual information between the prediction and distribution of all possible sub-graphs. Ying et al.[69] show that GNNExplainer can be used to obtain model-agnostic explanations. SubgraphX is a similar method that explains GNN predictions by identifying important subgraphs.[71]

Another set of approaches like DeepLIFT[72] and Layerwise Relevance backPropagation[73]

(LRP) are based on backpropagation of the 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 [74] showed 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. [44,46] SHAP has been popularly used in explaining molecular prediction models. [75–78] 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 ϕ_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. [79,80]

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

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 $i_t h$ 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 $\phi_0 + \sum_{i=1} \phi_i(\vec{x}) = \hat{f}(\vec{x})$.

Visualization based feature attribution has also been used for molecular data. In computer science, saliency maps are a way to measure spatial feature contribution. [81] Simply put, saliency maps draw a connection between the model's neural fingerprint components (trained weights) and input features. Weber et al. [82] used saliency maps to build an explainable GCN architecture that gives subgraph importance for small molecule activity prediction. On the other hand, similarity maps compare model predictions for two or more molecules based on their chemical fingerprints. [83] Similarity maps provide atomic weights or predicted probabil-

ity difference between the molecules by removing one atom 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.[84]

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. Generally, one model per explanation is trained. However, if we could find one surrogate model that explained the whole DL model, then we would simply have a globally accurate interpretable model. This means that the black-box model is no longer needed.[79] In the work by White[79], a weighted least squares linear model is used as the surrogate model. This model provides natural language based descriptor explanations by replacing input features with chemically interpretable descriptors. This approach is similar to the concept-based explanations approach used by McGrath et al.[85], where human understandable concepts were 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).[35] LIME creates perturbations around the example of interest and fits an interpretable model to these local perturbations. Ribeiro et al.[35] mathematically define an explanation ξ for an example $\vec{x}$ using Equation 4.

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

Here f is the black box model and $g \in G$ is the interpretable explanation model. G is a class of potential interpretable models (e.g.: linear models). π_x is a similarity measure

between original input $\vec{x}$ and it's perturbed input $\vec{x}'$. In context of molecular data, this can be a chemical similarity metric like Tanimoto[86] similarity between fingerprints. The goal for LIME is to minimize the loss, $\mathcal{L}$, such that f is closely approximated by g. Ω is a parameter that controls the complexity (sparsity) of g. Ribeiro et al.[35] termed the agreement (how low the loss is) between f and g as the "fidelity".

GraphLIME[87] and LIMEtree[88] 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.[89] who used LIME to explain octane number predictions of molecules from a random forest classifier. Mehdi and Tiwary[90] 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[90], 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.

Counterfactual explanations

Counterfactual explanations can be found in many fields such as statistics, mathematics and philosophy.[91–94] According to Woodward and Hitchcock[92], 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 the difference between the two instances is based on the existence of similar worlds in philosophy,[95] 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[96] between ECFP4 fingerprints.[97] Additionally, Mohapatra et al.[98] introduced a chemistry-informed graph representation for computing macromolecular similarity. Contrastive explanations are peripheral to counterfac-

tual 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.[36,99]

A counterfactual x' of an instance x is one with a dissimilar prediction $\hat{f}(x)$ in classification tasks. As shown in equation 5, counterfactual generation can be thought of as a constrained optimization problem which minimizes the vector distance $d(x, x')$ between the features.[9,100]

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

For regression tasks, equation 6 adapted from equation 5 can be used. Here, 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 predictions and are able to uncover spurious relationships in training data.[101] Counterfactuals create local (instance-level), actionable explanations. Actionability of an explanation suggest which features can be altered to change the outcome. For example, changing a hydrophobic functional group in a molecule to a hydrophilic group to increase solubility.

Counterfactual generation is a demanding task as it requires gradient optimization over discrete features that represents a molecule. Recent work by Fu et al.[102] and Shen et al.[103] 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.

CF-GNNExplainer[104] is a counterfactual explanation generating method based on GNNExplainer[69] for graph data. This method generate counterfactuals by perturbing the input data (removing edges in the graph), and keeping account of perturbations which 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[105] 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 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 firstly populates a local chemical space through random string mutations of SELFIES[106] molecular representations using the STONED algorithm.[107] Next, the labels (predictions) of the molecules in the local space are generated using the model that needs to be explained. Finally, the counterfactuals are identified and sorted by their similarities – Tanimoto distance[96] between ECFP4 fingerprints.[97] Unlike the CF-GNNExplainer[104] and MEG[105] methods, the MMACE algorithm ensures that generated molecules are valid, owing to the surjective property of SELFIES. Additionally, the MMACE method can be applied to both regression and classification models. However, like most XAI methods for molecular prediction, MMACE does not account for the chemical stability of predicted counterfactuals. To circumvent this drawback, Wellawatte et al.[9] propose another approach, which identift counterfactuals through a similarity search on the PubChem database.[108]

Similarity to adjacent fields

Tangential examples to counterfactual explanations are adversarial training and matched molecular pairs. Adversarial perturbations are used during training to deceive the model to expose the vulnerabilities of a model[109,110] whereas counterfactuals are 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.[100] Grabocka et al.[111] have developed a method named Adversarial Training on EXplanations (ATEX) which improves model robustness via exposure to adversarial examples. While there are conceptual disparities, we note that the counterfactual and adversarial explanations are equivalent mathematical objects.

Matched molecular pairs (MMPs) are pairs of molecules that differ structurally at only one site by a known transformation.[112,113] MMPs are widely used in drug discovery and medicinal chemistry as these facilitate fast and easy understanding of structure-activity relationships.[114–116] Counterfactuals and MMP examples intersect if the structural change is associated with a significant change in the properties. In the case the associated changes in the properties are non-significant, the two molecules are known as bioisosteres.[117,118] The connection between MMPs and adversarial training examples has been explored by van Tilborg et al.[119]. MMPs which belong to the counterfactual category are commonly used in outlier and activity cliff detection.[113] This approach is analogous to counterfactual explanations, as the common objective is to uncover learned knowledge pertaining to structure-property relationships.[70]

Applications

Model interpretation is certainly not new and a common step in ML in chemistry, but XAI for DL models is becoming more important[60,66–69,73,88,104,105] 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.[31]

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.[120] Small molecule blood-brain barrier (BBB) permeation is a classification problem routinely assessed with DL models.[121,122] To explain why DL models work, we trained two models a random forest (RF) model[123] and a Gated Recurrent Unit Recurrent Neural Network (GRU-RNN). Then we explained the RF model with generated counterfactuals explanations using the MMACE[9] and the GRU-RNN with descriptor explanations.[10] Both the models were trained on the dataset developed by Martins et al.[124]. The RF model was implemented in Scikit-learn[125] using Mordred molecular descriptors[126] as the input features. The GRU-RNN model was implemented in Keras.[127] See Wellawatte et al.[9] and Gandhi and White[10] for more details.

According to the counterfactuals of the instance molecule in figure 1, we observe that the modifications to the carboxylic acid group enable the negative example molecule to permeate the BBB. Experimental findings by Fischer et al.[120] 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. This proves the advantage of using counterfactual explanations, as they suggest 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 is 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 is proposed in more mechanistic studies.[128–130] 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.

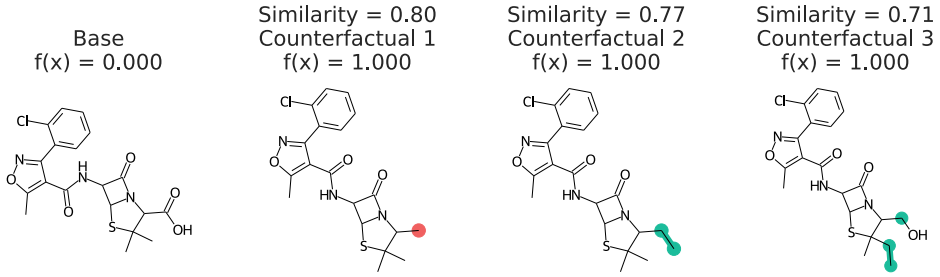

Figure 1: Counterfactuals of a molecule which cannot permeate the blood-brain barrier. Similarity is the Tanimoto similarity of ECFP4 fingerprints.[131] Red indicates deletions and green indicates substitutions and addition of atoms. Republished from Ref.[9] with permission from the Royal Society of Chemistry.

Solubility prediction

Small molecule solubility prediction is a classic cheminformatics regression challenge and is important for chemical process design, drug design and crystallization.[133–136] In our previous works,[9,10] we implemented and trained an RNN model in Keras to predict solubilities (log molarity) of small molecules.[127] The AqSolDB curated database[137] was used to train the RNN model.

In this task, counterfactuals are based on equation 6. Figure 3 illustrates the generated local chemical space and the top four counterfactuals. 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.[134] Figure 4 shows a quantitative measurement of how substructures are contributing to the pre-

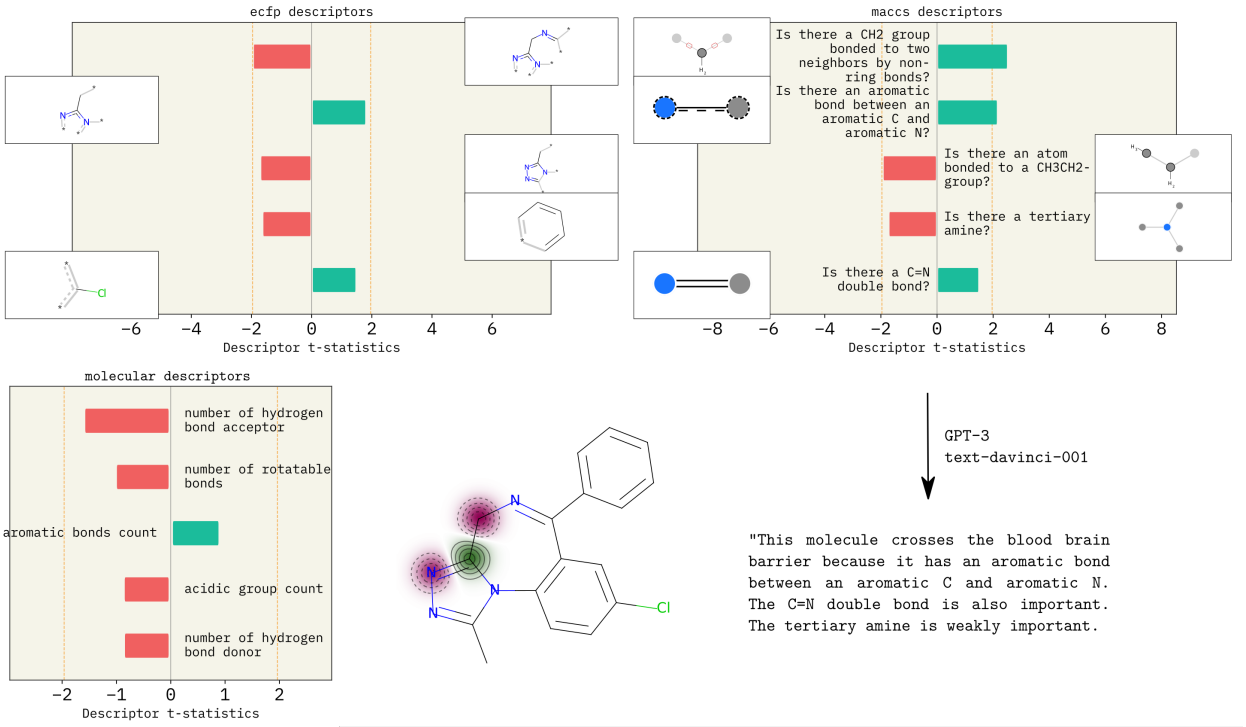

Figure 2: Descriptor explanations along with natural language explanation obtained for BBB permeability of Alprozolam molecule. The green and red bars show descriptors that influence predictions positively and negatively, respectively. Dotted yellow lines show significance threshold ($\alpha = 0.05$) for the t-statistic. Molecular descriptors show molecule-level properties that are important for the prediction. 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. 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.[132].

diction. For example, we see that adding acidic and basic groups as well as hydrogen bond acceptors, increases solubility. Substructure importance from ECFP[97] and MACCS[138] descriptors indicate that adding heteroatoms increases solubility, while adding rings structures makes the molecule less soluble. Although these are established hypotheses, it is interesting to see they can be derived purely from the data via DL and XAI.

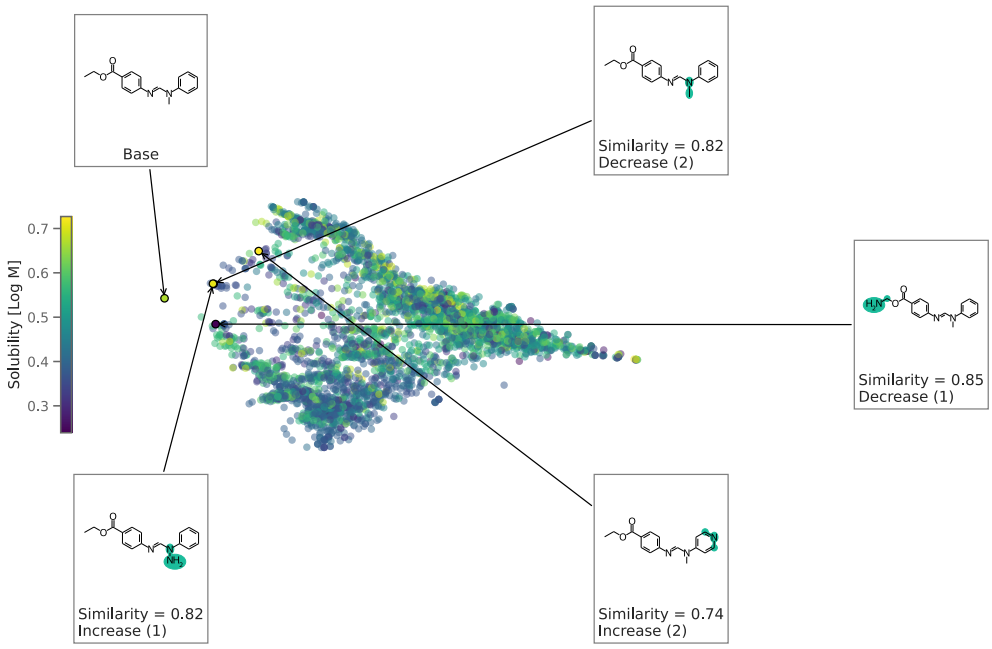

Figure 3: Generated chemical space for solubility prediction using the RNN model. The chemical space is a 2D projection of the pairwise Tanimoto similarities of the local counterfactuals. Each data point is colored by solubility. Top 4 counterfactuals are shown here. Republished from Ref.[9] with permission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

Generalizing XAI – interpreting scent-structure relationships

In this example, we show how non-local structure-property relationships can be learned with XAI across multiple molecules. Molecular scent prediction is a 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.[139] The scent-structure relationship is not very well understood,[140] although some relationships are known. For example, molecules with an ester functional group are often associated with

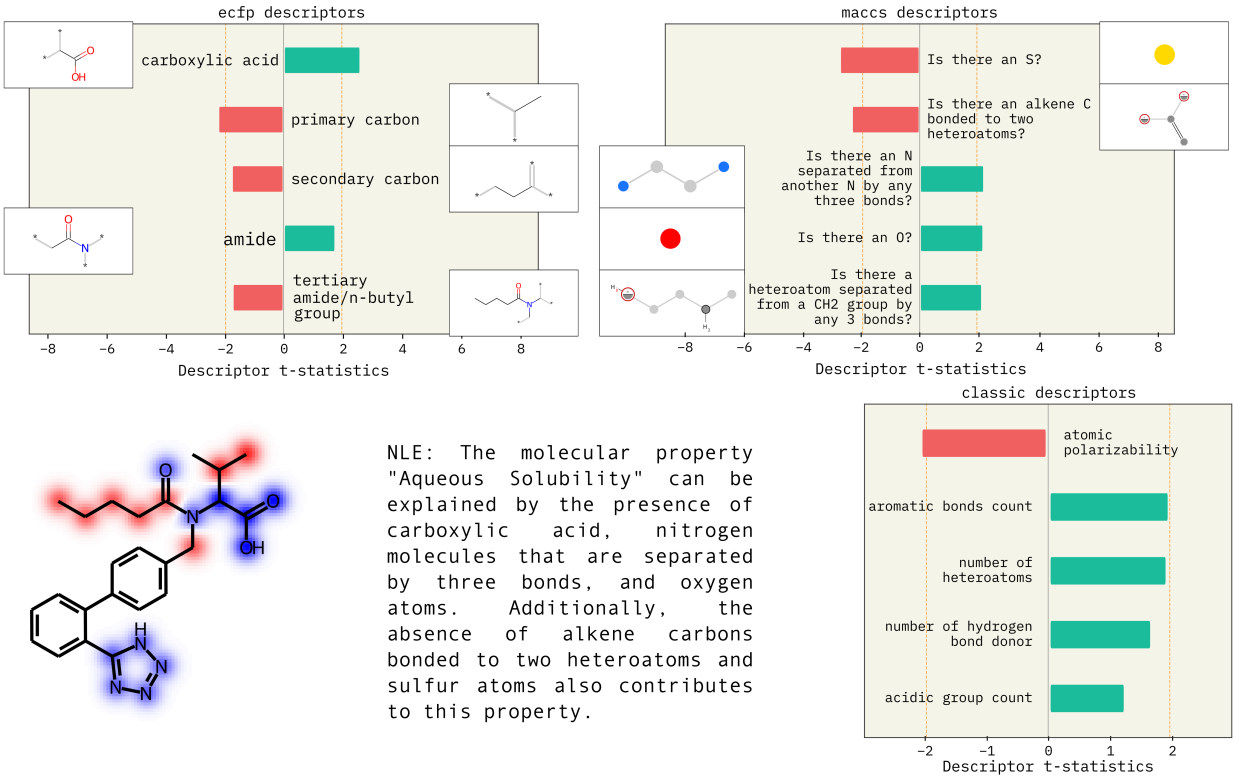

Figure 4: Descriptor explanations for solubility prediction model. The green and red bars show descriptors that influence predictions positively and negatively, respectively. Dotted yellow lines show significance threshold ($\alpha = 0.05$) for the t-statistic. The MACCS and ECFP descriptors indicate which substructures influence model predictions. MACCS substructures may either be present in the molecule as is or may represent a modification. 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. 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.[132].

the ‘fruity’ scent. There are some exceptions though, like tert-amyl acetate which has a ‘camphoraceous’ rather than ‘fruity’ scent.[140,141]

In Seshadri et al.[31], we trained a GNN model to predict the scent of molecules and utilized counterfactuals[9] and descriptor explanations[10] 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.

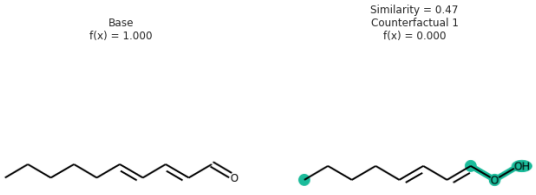

Figure 5: Counterfactual for the 2,4 decadienal molecule. The counterfactual indicates structural changes to ethyl benzoate that would result in the model predicting the molecule to not contain the ‘fruity’ scent. The Tanimoto[96] similarity between the counterfactual and 2,4 decadienal is also provided. Republished with permission from authors.[31]

The molecule 2,4-decadienal, which is known to have a ‘fatty’ scent, is analyzed in Figure 5.[142,143] The resulting counterfactual, which has a shorter carbon chain and no carbonyl groups, highlights the influence of these structural features on the ‘fatty’ scent of 2,4 decadienal. To generalize to other molecules, Seshadri et al.[31] 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." [31] The importance of a heptanyl fragment aligns with that reported in the literature, as 'fatty' molecules often have a long carbon chain.[144] Furthermore, the importance of a C=O double bond is supported by the findings reported by Licon et al.[145], where in addition to a "larger carbon-chain skeleton", they found that 'fatty' molecules also had "aldehyde or acid functions".[145] 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." [31] Esters, such as ethyl 2-methylbutyrate, are present in many pineapple volatile compounds.[146,147] 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.[146,148]

Discussion

We have shown two post-hoc XAI applications based on molecular counterfactual explanations[9] and descriptor explanations.[10] These methods can be used to explain black-box models whose input is a molecule. These two methods can be applied for both classification and regression tasks. Note that the "correctness" of the explanations strongly depends on the accuracy of the black-box model.

A molecular counterfactual is one with a minimal distance from a base molecular, but with contrasting chemical properties. In the above examples, we used Tanimoto similarity[96] of ECFP4 fingreprints[97] 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.[69,104,105]

The descriptor explanation method developed by Gandhi and White[10] fits a self-explaining

surrogate model to explain the black-box model. This is similar to the GraphLIME[87] method, although we have the flexibility to use explanation features other than subgraphs. Futhermore, we show that natural language combined with chemical descriptor attributions can create explanations useful for chemists, thus enhancing the accessibility of DL in chemistry. Lastly, we examined if XAI can be used beyond interpretation. Work by Seshadri et al.[31] 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.

Choosing among the plethora of XAI methods described here is still an open question. It remains to be seen if there will ever be a consensus benchmark, since this field sits on the intersection of human-machine interaction, machine learning, and philosophy (i.e., what constitutes an explanation?). Our current advice is to consider first the audience – domain experts or ML experts or non-experts – and what the explanations should accomplish. Are they meant to inform data selection or model building, how a prediction is used, or how the features can be changed to affect the outcome. The second consideration is what access you have to the underlying model. The ability to have model derivatives or propagate gradients to the input to models informs the XAI method.

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 principles. 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 interpretability. However, XAI in chemistry has some major open questions, 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 concept, etc?

- *Molecular distance*: in XAI approaches such as counterfactual generation, the "distance" between two molecules is minimized. Molecular distance is subjective. Possibilities are distance based on molecular properties, synthesis routes, and direct structure comparisons.

- *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. Explanations may need to be tuned for be for doctors instead of chemists or to satisfy a legal requirement.

- *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? Can an explanation consider nearby impossible molecules? How can we generate local chemical spaces centered around a specific molecule for finding counterfactuals or other instance explanations? Similarly, can "activity cliffs" be connected to explanations and the local chemical space.[149]

- *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? For example, work by Rasmussen et al.[58] attempts to focus on comparing feature attribution XAI methods via Crippen's logP scores.

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) Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J.; Holm, E.; Ong, S. P.; Wolverton, C. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials* **2022**, *8*.

(2) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chemical Reviews* **2021**, *121*, 9816–9872, PMID: 34232033.

(3) Goh, G. B.; Hodas, N. O.; Vishnu, A. Deep learning for computational chemistry. *Journal of Computational Chemistry* **2017**, *38*, 1291–1307.

(4) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Advanced Materials* **2019**, *31*, 1902765.

(5) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *Journal of Chemical Theory and Computation* **2017**, *13*, 5255–5264, PMID: 28926232.

(6) Duch, W.; Swaminathan, K.; Meller, J. Artificial Intelligence Approaches for Rational Drug Design and Discovery. *Current Pharmaceutical Design* **2007**, *13*, 1497–1508.

(7) Dara, S.; Dhamercherla, S.; Jadav, S. S.; Babu, C. M.; Ahsan, M. J.; darasuresh, S. D.; Dara, S. Machine Learning in Drug Discovery: A Review. *Artificial Intelligence Review* **123**, *55*, 1947–1999.

(8) Gupta, R.; Srivastava, D.; Sahu, M.; Tiwari, S.; Ambasta, R. K.; Kumar, P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular diversity* **2021**, *25*, 1315–1360.

(9) Wellawatte, G. P.; Seshadri, A.; White, A. D. Model agnostic generation of counterfactual explanations for molecules. *Chemical Science* **2022**, *13*, 3697–3705.

(10) Gandhi, H. A.; White, A. D. Explaining structure-activity relationships using locally faithful surrogate models. *chemrxiv* **2022**,

(11) Gormley, A. J.; Webb, M. A. Machine learning in combinatorial polymer chemistry. *Nature Reviews Materials* **2021**,

(12) Gomes, C. P.; Fink, D.; Dover, R. B. V.; Gregoire, J. M. Computational sustainability meets materials science. *Nature Reviews Materials* **2021**,

(13) On scientific understanding with artificial intelligence. *Nature Reviews Physics 2022 4:12* **2022**, *4*, 761–769.

(14) Arrieta, A. B.; Díaz-Rodríguez, N.; Ser, J. D.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; Chatila, R.; Herrera, F. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* **2019**, *58*, 82–115.

(15) Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Interpretable machine learning: definitions, methods, and applications. *ArXiv* **2019**, *abs/1901.04592*.

(16) Boobier, S.; Osbourn, A.; Mitchell, J. B. Can human experts predict solubility better than computers? *Journal of cheminformatics* **2017**, *9*, 1–14.

(17) Lee, J. D.; See, K. A. Trust in automation: Designing for appropriate reliance. *Human Factors* **2004**, *46*, 50–80.

(18) Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* **2016**, *29*.

(19) Buolamwini, J.; Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Proceedings of the 1st Conference on Fairness, Accountability and Transparency. 2018; pp 77–91.

(20) Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.-R. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications* **2019**, *10*, 1–8.

(21) DeGrave, A. J.; Janizek, J. D.; Lee, S.-I. AI for radiographic COVID-19 detection selects shortcuts over signal. *Nature Machine Intelligence* **2021**, *3*, 610–619.

(22) Goodman, B.; Flaxman, S. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* **2017**, *38*, 50–57.

(23) ACT, A. I. European Commission. *On Artificial Intelligence: A European Approach to Excellence and Trust.* **2021**, COM/2021/206.

(24) Blueprint for an AI Bill of Rights, The White House. 2022; `https://www.whitehouse.gov/ostp/ai-bill-of-rights/`.

(25) Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* **2019**, *267*, 1–38.

(26) Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences of the United States of America* **2019**, *116*, 22071–22080.

(27) Gunning, D.; Aha, D. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine* **2019**, *40*, 44–58.

(28) Biran, O.; Cotton, C. Explanation and justification in machine learning: A survey. IJCAI-17 workshop on explainable AI (XAI). 2017; pp 8–13.

(29) Palacio, S.; Lucieri, A.; Munir, M.; Ahmed, S.; Hees, J.; Dengel, A. Xai handbook: Towards a unified framework for explainable ai. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021; pp 3766–3775.

(30) Kuhn, D. R.; Kacker, R. N.; Lei, Y.; Simos, D. E. Combinatorial Methods for Explainable AI. *2020 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* **2020**, 167–170.

(31) Seshadri, A.; Gandhi, H. A.; Wellawatte, G. P.; White, A. D. Why does that molecule smell? *ChemRxiv* **2022**,

(32) Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* **2020**,

(33) Machlev, R.; Heistrene, L.; Perl, M.; Levy, K. Y.; Belikov, J.; Mannor, S.; Levron, Y. Explainable Artificial Intelligence (XAI) techniques for energy and power systems: Review, challenges and opportunities. *Energy and AI* **2022**, *9*, 100169.

(34) Koh, P. W.; Liang, P. Understanding black-box predictions via influence functions. International Conference on Machine Learning. 2017; pp 1885–1894.

(35) Ribeiro, M. T.; Singh, S.; Guestrin, C. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international

conference on knowledge discovery and data mining. San Diego, CA, USA, 2016; pp 1135–1144.

(36) Dhurandhar, A.; Chen, P.-Y.; Luss, R.; Tu, C.-C.; Ting, P.; Shanmugam, K.; Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems* **2018**, *31*.

(37) Jin, W.; Li, X.; Hamarneh, G. Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements? *Proceedings of the AAAI Conference on Artificial Intelligence* **2022**, *36*, 11945–11953.

(38) Zhang, Y.; Xu, F.; Zou, J.; Petrosian, O. L.; Krinkin, K. V. XAI Evaluation: Evaluating Black-Box Model Explanations for Prediction. 2021 II International Conference on Neural Networks and Neurotechnologies (NeuroNT). 2021; pp 13–16.

(39) Oviedo, F.; Ferres, J. L.; Buonassisi, T.; Butler, K. T. Interpretable and Explainable Machine Learning for Materials Science and Chemistry. *Accounts of Materials Research* **2022**, *3*, 597–607.

(40) Yalcin, O.; Fan, X.; Liu, S. Evaluating the correctness of explainable AI algorithms for classification. *arXiv preprint arXiv:2105.09740* **2021**,

(41) Hoffman, R. R.; Mueller, S. T.; Klein, G.; Litman, J. Metrics for Explainable AI: Challenges and Prospects. **2018**,

(42) Mohseni, S.; Zarei, N.; Ragan, E. D. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems* **2018**, *11*, 46.

(43) Humer, C.; Heberle, H.; Montanari, F.; Wolf, T.; Huber, F.; Henderson, R.; Heinrich, J.; Streit, M. ChemInformatics Model Explorer (CIME): exploratory analysis of chemical model explanations. *Journal of Cheminformatics* **2022**, *14*, 1–14.

(44) Lundberg, S. M.; Lee, S.-I. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 4765–4774.

(45) Štrumbelj, E.; Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems* **2014**, *41*, 647–665.

(46) Shapley, L. S. *A Value for N-Person Games*; RAND Corporation: Santa Monica, CA, 1952.

(47) Molnar, C.; Casalicchio, G.; Bischl, B. Interpretable machine learning–a brief history, state-of-the-art and challenges. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. 2020; pp 417–431.

(48) Lou, Y.; Caruana, R.; Gehrke, J. Intelligible models for classification and regression. Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. 2012; pp 150–158.

(49) Bastani, O.; Kim, C.; Bastani, H. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504* **2017**,

(50) Gajewicz, A.; Puzyn, T.; Odziomek, K.; Urbaszek, P.; Haase, A.; Riebeling, C.; Luch, A.; Irfan, M. A.; Landsiedel, R.; van der Zande, M.; Bouwmeester, H. Decision tree models to classify nanomaterials according to the DF4nanoGrouping scheme. *Nanotoxicology* **2018**, *12*, 1–17.

(51) Han, L.; Wang, Y.; Bryant, S. H. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high–throughput screening data in PubChem. *BMC Bioinformatics* **2008**, *9*, 401.

(52) Plumb, G.; Al-Shedivat, M.; Cabrera, Á. A.; Perer, A.; Xing, E.; Talwalkar, A. Regu-

larizing black-box models for improved interpretability. *Advances in Neural Information Processing Systems* **2020**, *33*, 10526–10536.

(53) Shao, X.; Skryagin, A.; Stammer, W.; Schramowski, P.; Kersting, K. Right for better reasons: Training differentiable models by constraining their influence functions. Proceedings of the AAAI Conference on Artificial Intelligence. 2021; pp 9533–9540.

(54) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials* **2018**, *2*, 083802.

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

(56) Harren, T.; Matter, H.; Hessler, G.; Rarey, M.; Grebner, C. Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence. *Journal of Chemical Information and Modeling* **2022**, *62*, 447–462.

(57) Sheridan, R. P. Interpretation of QSAR Models by Coloring Atoms According to Changes in Predicted Activity: How Robust Is It? *Journal of Chemical Information and Modeling* **2019**, *59*, 1324–1337.

(58) Rasmussen, M. H.; Christensen, D. S.; Jensen, J. H. Do machines dream of atoms? Crippen's logP as a quantitative molecular benchmark for explainable AI heatmaps. **2022**,

(59) Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. SmoothGrad: removing noise by adding noise. 2017; `https://arxiv.org/abs/1706.03825`.

(60) McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. 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 **2018**, *116*, 11624–11629.

(61) Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* **2015**, *10*, e0130140.

(62) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. International Conference on Machine Learning. 2017; pp 3319–3328.

(63) Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. 2015; `https://arxiv.org/abs/1512.04150`.

(64) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* **2019**, *128*, 336–359.

(65) Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825* **2017**,

(66) Pope, P.; Kolouri, S.; Rostrami, M.; Martin, C.; Hoffmann, H. Discovering Molecular Functional Groups Using Graph Convolutional Neural Networks. 2018; `https://arxiv.org/abs/1812.00265`.

(67) Jiménez-Luna, J.; Skalic, M.; Weskamp, N.; Schneider, G. Coloring molecules with explainable artificial intelligence for preclinical relevance assessment. *Journal of Chemical Information and Modeling* **2021**, *61*, 1083–1094.

(68) Sanchez-Lengeling, B.; Wei, J.; Lee, B.; Reif, E.; Wang, P. Y.; Qian, W. W.; McCloskey, K.; Colwell, L.; Wiltschko, A. Evaluating Attribution for Graph Neural Networks. Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, NY, USA, 2020.

(69) Ying, R.; Bourgeois, D.; You, J.; Zitnik, M.; Leskovec, J. GNNExplainer: Generating Explanations for Graph Neural Networks. *Advances in neural information processing systems* **2019**, *32*, 9240–9251.

(70) Rao, J.; Zheng, S.; Yang, Y. Quantitative Evaluation of Explainable Graph Neural Networks for Molecular Property Prediction. *arXiv preprint arXiv:2107.04119* **2021**,

(71) Yuan, H.; Yu, H.; Wang, J.; Li, K.; Ji, S. On Explainability of Graph Neural Networks via Subgraph Explorations. Proceedings of the 38th International Conference on Machine Learning. 2021; pp 12241–12252.

(72) Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features Through Propagating Activation Differences. **2017**,

(73) Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; Müller, K. R. Layer-Wise Relevance Propagation: An Overview. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **2019**, *11700 LNCS*, 193–209.

(74) Baldassarre, F.; Azizpour, H. Explainability Techniques for Graph Convolutional Networks. 2019; `https://arxiv.org/abs/1905.13686`.

(75) Hochuli, J.; Helbling, A.; Skaist, T.; Ragoza, M.; Koes, D. R. Visualizing convolutional neural network protein-ligand scoring. *Journal of Molecular Graphics and Modelling* **2018**, *84*, 96–108.

(76) Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *Journal of Medicinal Chemistry* **2020**, *63*, 8761–8777, PMID: 31512867.

(77) Wojtuch, A.; Jankowski, R.; Podlewska, S. How can SHAP values help to shape

metabolic stability of chemical compounds? *Journal of Cheminformatics* **2021**, *13*, 1–20.

(78) Mastropietro, A.; Pasculli, G.; Feldmann, C.; Rodríguez-Pérez, R.; Bajorath, J. EdgeSHAPer: Bond-Centric Shapley Value-Based Explanation Method for Graph Neural Networks. *iScience* **2022**, *25*, 105043.

(79) White, A. D. Deep learning for molecules and materials. *Living Journal of Computational Molecular Science* **2022**, *3*.

(80) Štrumbelj, E.; Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* **2014**, *41*, 647–665.

(81) Erhan, D.; Bengio, Y.; Courville, A.; Vincent, P. Visualizing Higher-Layer Features of a Deep Network. *Technical Report, Univeristé de Montréal* **2009**,

(82) Weber, J. K.; Morrone, J. A.; Bagchi, S.; Pabon, J. D.; gu Kang, S.; Zhang, L.; Cornell, W. D. Simplified, interpretable graph convolutional neural networks for small molecule activity prediction. *Journal of Computer-Aided Molecular Design* **2022**, *36*, 391–404.

(83) Riniker, S.; Landrum, G. A. Similarity maps - A visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics* **2013**, *5*, 1–7.

(84) Humer, C.; Heberle, H.; Montanari, F.; Wolf, T.; Huber, F.; Henderson, R.; Heinrich, J.; Streit, M. ChemInformatics Model Explorer (CIME): exploratory analysis of chemical model explanations. *Journal of Cheminformatics* **2022**, *14*, 1–14.

(85) McGrath, T.; Kapishnikov, A.; Tomašev, N.; Pearce, A.; Wattenberg, M.; Hassabis, D.; Kim, B.; Paquet, U.; Kramnik, V. Acquisition of chess knowledge in AlphaZero. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2206625119.

(86) Bajusz, D.; Rácz, A.; Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **2015**, *7*, 1–13.

(87) Huang, Q.; Yamada, M.; Tian, Y.; Singh, D.; Yin, D.; Chang, Y. GraphLIME: Local Interpretable Model Explanations for Graph Neural Networks. *CoRR* **2020**, *abs/2001.06216*.

(88) Sokol, K.; Flach, P. A. LIMEtree: Interactively Customisable Explanations Based on Local Surrogate Multi-output Regression Trees. *CoRR* **2020**, *abs/2005.01427*.

(89) Whitmore, L. S.; George, A.; Hudson, C. M. Mapping chemical performance on molecular structures using locally interpretable explanations. 2016; `https://arxiv.org/abs/1611.07443`.

(90) Mehdi, S.; Tiwary, P. Thermodynamics of Interpretation. **2022**,

(91) Höfler, M. Causal inference based on counterfactuals. *BMC Medical Research Methodology* **2005**, *5*, 1–12.

(92) Woodward, J.; Hitchcock, C. Explanatory Generalizations, Part I: A Counterfactual Account. *Noûs* **2003**, *37*, 1–24.

(93) Frisch, M. F. *Theories, models, and explanation*; University of California, Berkeley, 1998.

(94) Reutlinger, A. Is There A Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation. *Philosophy of Science* **2016**, *83*, 733–745.

(95) Lewis, D. Causation. *The journal of philosophy* **1974**, *70*, 556–567.

(96) Tanimoto, T. T. Elementary mathematical theory of classification and prediction. *Internal IBM Technical Report* **1958**,

(97) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754, PMID: 20426451.

(98) Mohapatra, S.; An, J.; Gómez-Bombarelli, R. Chemistry-informed macromolecule graph representation for similarity computation, unsupervised and supervised learning. *Machine Learning: Science and Technology* **2022**, *3*, 015028.

(99) Doshi-Velez, F.; Kortz, M.; Budish, R.; Bavitz, C.; Gershman, S.; O'Brien, D.; Scott, K.; Schieber, S.; Waldo, J.; Weinberger, D.; Weller, A.; Wood, A. Accountability of AI Under the Law: The Role of Explanation. *SSRN Electronic Journal* **2017**,

(100) Wachter, S.; Mittelstadt, B.; Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* **2017**, *31*, 841.

(101) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence 2020 2:10* **2020**, *2*, 573–584.

(102) Fu, T.; Gao, W.; Xiao, C.; Yasonik, J.; Coley, C. W.; Sun, J. Differentiable Scaffolding Tree for Molecule Optimization. International Conference on Learning Representations. 2022.

(103) Shen, C.; Krenn, M.; Eppel, S.; Aspuru-Guzik, A. Deep molecular dreaming: inverse machine learning for de-novo molecular design and interpretability with surjective representations. *Machine Learning: Science and Technology* **2021**, *2*, 03LT02.

(104) Lucic, A.; ter Hoeve, M.; Tolomei, G.; Rijke, M.; Silvestri, F. CF-GNNExplainer: Counterfactual Explanations for Graph Neural Networks. *arXiv preprint arXiv:2102.03322* **2021**,

(105) Numeroso, D.; Bacciu, D. Explaining Deep Graph Networks with Molecular Counterfactuals. *arXiv preprint arXiv:2011.05134* **2020**,

(106) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **2020**, *1*, 045024.

(107) Nigam, A.; Pollice, R.; Krenn, M.; dos Passos Gomes, G.; Aspuru-Guzik, A. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES. *Chemical science* **2021**, *12*, 7079–7090.

(108) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research* **2020**, *49*, D1388–D1395.

(109) Tolomei, G.; Silvestri, F.; Haines, A.; Lalmas, M. Interpretable predictions of tree-based ensembles via actionable feature tweaking. Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining. 2017; pp 465–474.

(110) Freiesleben, T. The intriguing relation between counterfactual explanations and adversarial examples. *Minds and Machines* **2022**, *32*, 77–109.

(111) Grabocka, J.; Schilling, N.; Wistuba, M.; Schmidt-Thieme, L. Learning time-series shapelets. Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014; pp 392–401.

(112) Kenny, P. W.; Sadowski, J. Structure modification in chemical databases. *Chemoinformatics in drug discovery* **2005**, 271–285.

(113) Tyrchan, C.; Evertsson, E. Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations. *Computational and Structural Biotechnology Journal* **2017**, *15*, 86–90.

(114) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *Journal of Medicinal Chemistry* **2011**, *54*, 7739–7750, PMID: 21936582.

(115) He, J.; Nittinger, E.; Tyrchan, C.; Czechtizky, W.; Patronov, A.; Bjerrum, E. J.; Engkvist, O. Transformer-based molecular optimization beyond matched molecular pairs. *Journal of cheminformatics* **2022**, *14*, 1–14.

(116) Park, J.; Sung, G.; Lee, S.; Kang, S.; Park, C. ACGCN: Graph Convolutional Networks for Activity Cliff Prediction between Matched Molecular Pairs. *Journal of Chemical Information and Modeling* **2022**,

(117) Langdon, S. R.; Ertl, P.; Brown, N. Bioisosteric Replacement and Scaffold Hopping in Lead Generation and Optimization. *Molecular Informatics* **2010**, *29*, 366–385.

(118) Turk, S.; Merget, B.; Rippmann, F.; Fulle, S. Coupling Matched Molecular Pairs with Machine Learning for Virtual Compound Optimization. *Journal of Chemical Information and Modeling* **2017**, *57*, 3079–3085, PMID: 29131617.

(119) van Tilborg, D.; Alenicheva, A.; Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. **2022**,

(120) Fischer, H.; Gottschlich, R.; Seelig, A. Blood-brain barrier permeation: molecular parameters governing passive diffusion. *The Journal of membrane biology* **1998**, *165*, 201–211.

(121) Liu, L.; Zhang, L.; Feng, H.; Li, S.; Liu, M.; Zhao, J.; Liu, H. Prediction of the Blood–Brain Barrier (BBB) Permeability of Chemicals Based on Machine-Learning and Ensemble Methods. *Chemical Research in Toxicology* **2021**, *34*, 1456–1467, PMID: 34047182.

(122) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9*, 513–530.

(123) Ho, T. K. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition. 1995; pp 278–282.

(124) Martins, I. F.; Teixeira, A. L.; Pinheiro, L.; Falcao, A. O. A Bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of chemical information and modeling* **2012**, *52*, 1686–1697.

(125) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(126) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *Journal of cheminformatics* **2018**, *10*, 1–14.

(127) Chollet, F., et al. Keras. `https://keras.io`, 2015.

(128) Wager, T. T.; Chandrasekaran, R. Y.; Hou, X.; Troutman, M. D.; Verhoest, P. R.; Villalobos, A.; Will, Y. Defining Desirable Central Nervous System Drug Space through the Alignment of Molecular Properties, in Vitro ADME, and Safety Attributes. *ACS Chemical Neuroscience* **2010**, *1*, 420–434.

(129) Ghose, A. K.; Herbertz, T.; Hudkins, R. L.; Dorsey, B. D.; Mallamo, J. P. Knowledge-Based, Central Nervous System (CNS) Lead Selection and Lead Optimization for CNS Drug Discovery. *ACS Chemical Neuroscience* **2012**, *3*, 50–68.

(130) Polishchuk, P.; Tinkov, O.; Khristova, T.; Ognichenko, L.; Kosinskaya, A.; Varnek, A.; Kuz'min, V. Structural and Physico-Chemical Interpretation (SPCI) of QSAR Models and Its Comparison with Matched Molecular Pair Analysis. *Journal of Chemical Information and Modeling* **2016**, *56*, 1455–1469.

(131) Hassan, M.; Brown, R. D.; Varma-O'Brien, S.; Rogers, D. Cheminformatics analysis and learning in a data pipelining environment. *Molecular diversity* **2006**, *10*, 283–299.

(132) Schomburg, K.; Ehrlich, H. C.; Stierand, K.; Rarey, M. From structure diagrams to visual chemical patterns. *Journal of Chemical Information and Modeling* **2010**, *50*, 1529–1535.

(133) Sheikholeslamzadeh, E.; Rohani, S. Solubility prediction of pharmaceutical and chemical compounds in pure and mixed solvents using predictive models. *Industrial & engineering chemistry research* **2012**, *51*, 464–473.

(134) Boobier, S.; Hose, D. R.; Blacker, A. J.; Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature Communications 2020 11:1* **2020**, *11*, 1–10.

(135) Loschen, C.; Klamt, A. Solubility prediction, solvate and cocrystal screening as tools for rational crystal engineering. *Journal of Pharmacy and Pharmacology* **2015**, *67*, 803–811.

(136) Diorazio, L. J.; Hose, D. R.; Adlington, N. K. Toward a more holistic framework for solvent selection. *Organic Process Research & Development* **2016**, *20*, 760–773.

(137) Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific data* **2019**, *6*, 1–8.

(138) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences* **2002**, *42*, 1273–1280.

(139) National Center for Biotechnology Information, PubChem Compound Summary for

CID 1549018, Jasmone. https://pubchem.ncbi.nlm.nih.gov/compound/Jasmone, Accessed September 26, 2022.

(140) Sell, C. S. On the unpredictability of odor. *Angewandte Chemie International Edition* **2006**, *45*, 6254–6261.

(141) Genva, M.; Kenne Kemene, T.; Deleu, M.; Lins, L.; Fauconnier, M.-L. Is It Possible to Predict the Odor of a Molecule on the Basis of its Structure? *International journal of molecular sciences* **2019**, *20*, 3018.

(142) Rowe, D. Aroma chemicals for savory flavors. *Perfumer and Flavorist* **1998**, *23*, 9–18.

(143) Mallia, S.; Escher, F.; Schlichtherle-Cerny, H. Aroma-active compounds of butter: a review. *European Food Research and Technology* **2008**, *226*, 315–325.

(144) Jelen, H.; Gracka, A. Characterization of aroma compounds: Structure, physicochemical and sensory properties. *Flavour: From food to perception* **2016**, 126–153.

(145) Licon, C. C.; Bosc, G.; Sabri, M.; Mantel, M.; Fournel, A.; Bushdid, C.; Golebiowski, J.; Robardet, C.; Plantevit, M.; Kaytoue, M., et al. Chemical features mining provides new descriptive structure-odor relationships. *PLoS computational biology* **2019**, *15*, e1006945.

(146) Mostafa, S.; Wang, Y.; Zeng, W.; Jin, B. Floral Scents and Fruit Aromas: Functions, Compositions, Biosynthesis, and Regulation. *Frontiers in plant science* **2022**, *13*.

(147) Tokitomo, Y.; Steinhaus, M.; Büttner, A.; Schieberle, P. Odor-active constituents in fresh pineapple (Ananas comosus [L.] Merr.) by quantitative and sensory evaluation. *Bioscience, Biotechnology, and Biochemistry* **2005**, *69*, 1323–1330.

(148) Wei, C.-B.; Liu, S.-H.; Liu, Y.-G.; Lv, L.-L.; Yang, W.-X.; Sun, G.-M. Characteristic aroma compounds from different pineapple parts. *Molecules* **2011**, *16*, 5104–5112.

(149) Stumpfe, D.; Bajorath, J. Exploring Activity Cliffs in Medicinal Chemistry. *Journal of Medicinal Chemistry* **2012**, *55*, 2932–2942, PMID: 22236250.