Drug discovery with explainable artificial intelligence

José Jiménez-Luna[+], Francesca Grisoni[+], and Gisbert Schneider[*]

ETH Zurich, Department of Chemistry and Applied Biosciences, RETHINK, Vladimir-Prelog-Weg 4, 8093 Zurich, Switzerland.

[*]Correspondence: gisbert@ethz.ch
[+]These authors contributed equally to this work

Abstract

Deep learning bears promise for drug discovery, including advanced image analysis, prediction of molecular structure and function, and automated generation of innovative chemical entities with bespoke properties. Despite the growing number of successful prospective applications, the underlying mathematical models often remain elusive to interpretation by the human mind. There is a demand for 'explainable' deep learning methods to address the need for a new narrative of the machine language of the molecular sciences. This review summarizes the most prominent algorithmic concepts of explainable artificial intelligence, and dares a forecast of the future opportunities, potential applications, and remaining challenges.

1 Introduction

Several concepts of 'artificial intelligence' (AI) have been adopted for computer-assisted drug discovery.[1–7] Deep learning algorithms, *i.e.*, artificial neural networks with multiple processing layers,[8] are currently receiving particular attention, owing to their capacity to model complex nonlinear input-output relationships, and perform pattern recognition and feature extraction from low-level data representations.[8,9]

Certain deep learning models have been shown to match or even exceed the performance of the existing machine-learning and quantitative structure-activity relationship (QSAR) methods for drug discovery.[6,10–13] Moreover, deep learning has boosted the potential and broadened the applicability of computer-assisted discovery, *e.g.,* for molecular de novo design,[14–16] chemical synthesis planning,[17–19] protein structure prediction,[20,21] and macromolecular target identification.[22–24] The ability to capture intricate nonlinear relationships between input data (*e.g.*, chemical structure representations) and the associated output (*e.g.*, assay readout) often comes at the price of limited comprehensibility of the resulting model. While there have been efforts to explain QSARs in terms of algorithmic insights and molecular descriptor analysis,[25–34] deep neural network models notoriously elude immediate accessibility by the human mind.[35,36]

In an effort to mitigate the lack of interpretability of deep learning models, attention has been drawn to explainable artificial intelligence (XAI).[37,38] Providing informative explanations of AI models aims to (i) render the underlying decision-making process transparent ('understandable'),[39] (ii) avoid correct predictions for the wrong reasons (the so-called 'clever Hans effect'[40]), (iii) avert unfair biases or unethical discrimination,[41] and (iv) bridge the gap between the machine learning community and other scientific disciplines. In medicinal chemistry, XAI already enables the mechanistic interpretation of drug action,[42,43] and contributes to drug safety enhancement[44] and organic synthesis planning.[17,45] Intensified efforts towards interpreting deep learning models will help increase their reliability and foster their acceptance and usage in drug discovery and medicinal chemistry projects.[46–49] The availability of 'rule of thumb' scores correlating biological effects with physicochemical properties[50–54] underscores the willingness, in certain situations, to sacrifice accuracy in favour of models that better fit human intuition. Thus, blurring the lines between the 'two QSARs'[55] (*i.e.*, mechanistically interpretable *vs.* highly accurate models) can be the key for accelerated drug discovery.[56]

While the exact definition of XAI is still under debate,[41,57] several aspects of XAI are certainly desirable in drug design applications:[37] (i) *transparency*, *i.e.,* knowing how the system reached a particular answer, (ii) *justification*, *i.e.,* elucidating why the answer provided by the model is acceptable, (iii) *informativeness*, *i.e.,* providing new information to human decision-makers, and (iv) *uncertainty estimation*, *i.e.,* the quantification of how reliable a prediction is. Moreover, AI model explanations can be *global* (*i.e.*, summarizing the relevance of input features in the model) or *local* (*i.e.* on individual predictions).[58] Furthermore, XAI can be model-dependent or agnostic, which in turn affects the potential applicability of each method.

The field of XAI is still in its infancy but moving forward at a fast pace, and we expect an increase of its relevance in the years to come. In this review, we aim to provide a comprehensive overview of recent XAI research, including its benefits, limitations and future opportunities for drug discovery. In what follows, after providing an introduction to the most relevant XAI methods structured into conceptual categories, the existing and some of the potential applications to drug discovery are highlighted. Finally, we discuss the limitations of contemporary XAI and point to the potential methodological improvements needed to foster practical applicability of these techniques to pharmaceutical research.

2 State of the art

This section aims to provide a concise overview of modern XAI approaches, and exemplify their use in computer

1

Table 1: Computational approaches toward explainable AI, categorized according to the respective methodological concept. For each family of approaches, a brief description of its aim is provided, its specific methods, and reported applications in drug discovery. '*Not reported*' refers to families of methods that, to the best of our knowledge, have not been yet applied in drug discovery. Potential applications of these are discussed in the main text.

Family	Aim	Methods	Reported applications in drug discovery
Feature attribution	Determine local feature importance towards a prediction.	- Gradient-based - Surrogate models - Perturbation-based	Ligand pharmacophore identification,[59–61] structural alerts for adverse effect,[62] protein-ligand interaction profiling.[63]
Instance-based	Compute a subset of features that need to be present or absent to guarantee or change a prediction.	- Anchors - Counterfactual instances - Contrastive explanations	*Not reported*
Graph-convolution-based	Interpret models within the message-passing framework.	- Sub-graph approaches - Attention-based	Retrosynthesis elucidation,[64] toxicophore and pharmacophore identification,[49] ADMET,[65,66] and reactivity prediction.[19]
Self-explaining	Develop models that are explainable by design.	- Prototype-based - Self-explaining neural networks - Concept learning - Natural language explanations	*Not reported*
Uncertainty estimation	Quantify the reliability of a prediction.	- Ensemble-based - Probabilistic - Other approaches	Reaction prediction,[67] active learning,[68] molecular activity prediction.[69]

vision, natural-language processing, and discrete mathematics. We then highlight selected case studies in drug discovery and propose potential future areas of XAI research. A summary of the methodologies and their goals, along with reported applications is provided in Table 1. In what follows, without loss of generality, f will denote a model (in most cases a neural network); $x \in \mathcal{X}$ will be used to denote the set of features describing a given instance, which are used by f to make a prediction $y \in \mathcal{Y}$.

2.1 Feature attribution methods

Given a regression or classification model $f : \boldsymbol{x} \in \mathbb{R}^K \to \mathbb{R}$, a feature attribution method is a function $\mathcal{E} : \boldsymbol{x} \in \mathbb{R}^K \to \mathbb{R}^K$ that takes the model input and produces an output whose values denote the relevance of every input feature for the final prediction computed with f. Feature attribution methods can be grouped into the following three categories (Figure 1):

- *Gradient-based feature attribution.* These approaches measure how much a change around a local neighborhood of the input $\boldsymbol{x}$ corresponds to a change in the output $f(\boldsymbol{x})$. A common approach among deep-learning practitioners relies on the use of the derivative of the output of the neural network with respect to the input (*i.e.*, $\frac{\delta f}{\delta \boldsymbol{x}}$) to determine feature importance.[70–72] Its popularity arises partially from

the fact that this computation can be performed via backpropagation,[73] the main way of computing partial first-order derivatives in neural network models. While the use of gradient-based feature attribution may seem straightforward, several methods relying on this principle have been shown to lead to only partial reconstruction of the original features,[74] which is prone to misinterpretation.

- *Surrogate-model feature attribution.* Given a model f, these methods aim to develop a surrogate explanatory model g, which is constructed in such a way that: (i) g is interpretable, and (ii) g approximates the original function f. A prominent example of this concept is the family of additive feature attribution methods, where the approximation is achieved through a linear combination of binary variables z_i:

$$g\left(z_i^{'}\right) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i, \qquad (1)$$

where $z_i \in \{0,1\}^M$, M is the number of original input features, $\phi_i \in \mathbb{R}$ are coefficients representing the importance assigned to each i-th binary variable and ϕ_0 is an intercept. Several notable feature attribution methods belong to this family,[75,76] such as Local Interpretable Model-Agnostic Explanations (LIME),[77] Deep Learning Important FeaTures (DeepLIFT),[78]
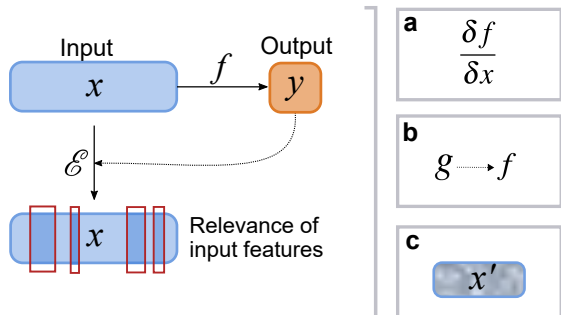
Figure 1: Feature attribution methods. Given a neural network model f, which computes the prediction $y = f(x)$ for input sample x, a feature attribution method $\mathcal{E}$ outputs the relevance of every input feature of x for the prediction. There are three basic approaches to determine feature relevance: (a) gradient-based methods, computing the gradient of the network f *w.r.t.* the input x, (b) surrogate methods, which approximate f with a human-interpretable model g, and (c) perturbation-based methods, which modify the original input to measure the respective changes in the output.

Shapley Additive Explanations (SHAP)[76] and Layer-Wise Relevance Propagation.[79] Both gradient-based methods and the additive subfamily of surrogate attribution methods provide local explanations (*i.e.*, each prediction needs to be examined individually), but they do not offer a general understanding of the underlying model f. Global surrogate explanation models aim to fill this gap by generically describing f via a decision tree or decision set[80] model. If such an approximation is precise enough, these aim to to mirror the computation logic of the original model. While early attempts limited f to the family of tree-based ensemble methods (*e.g.*, random forests[81]), more recent approaches are readily applicable to arbitrary deep learning models.[82]

- *Perturbation-based methods* modify or remove parts of the input aiming to measure its corresponding change in the model output; this information is then used to assess the feature importance. Alongside the well-established step-wise approaches,[83–85] methods such as feature masking,[86] perturbation analysis,[87] response randomization,[88] and conditional multivariate models[89] belong to this category. While perturbation-based methods have the advantage of directly estimating feature importance, they are computationally slow when the number of input features increases,[89] and the final result tends to be strongly influenced by the number of features that are perturbed altogether.[90]

Feature attribution methods have been used for ligand- and structure-based drug-discovery. For instance, McCloskey *et al.*[59] employed gradient-based attribution[70] to detect ligand pharmacophores relevant for binding. The study showed that, despite good performance of the models on held-out data, these still can learn spurious correlations.[59] Pope *et al.*[62] adapted gradient-based feature attribution[91,92] for the identification of relevant functional groups in adverse effect prediction.[93] Recently, SHAP[76]

was used to interpret relevant features for compound potency and multiâĂŚtarget activity prediction.[60,61] Hochuli *et al.*[63] compared several feature attribution methodologies, showing how the visualization of attributions assists in parsing and interpreting of protein-ligand scoring with 3D convolutional neural networks.[94,95]

It should be noted that the interpretability of feature attribution methods is limited by the original set of features (model input). In drug discovery the interpretability of machine learning methods is often hampered by the use of complex input molecular descriptors. When making use of feature attribution approaches, it is advisable to choose molecular descriptors or representations for model construction that one considers to bear 'interpretable meaning' (Box 1).

2.2 Instance-based approaches

Instance-based approaches compute a subset of relevant features (instances) that must be present to retain (or absent to change) the prediction of a given model (Figure 2). An instance can be real (*i.e.*, drawn from the set of data) or generated for the purposes of the method. Instance-based approaches have been argued to provide 'natural' model interpretations for humans, because they resemble counterfactual reasoning (*i.e.* producing alternative sets of action to achieve a similar or different result).[96]

- *Anchor* algorithms[97] offer model-agnostic interpretable explanations of classifier models. They compute a subset of *if-then* rules based on one or more features that represent conditions to sufficiently guarantee a certain class prediction. In contrast to many other local explanation methods,[77] anchors therefore explicitly model the 'coverage' of an explanation. Formally, an anchor A is defined as a set of rules such that, given a set of features x from a sample, they return $A(x) = 1$ if said rules are met, while guaranteeing the desired predicted class from f with a certain probability τ:

$$\mathbb{E}_{\mathcal{D}(z|A)} \left[\mathbb{1}_{f(x)=f(z)} \right] \geq \tau, \qquad (2)$$

where $\mathcal{D}(z|A)$ is defined as the conditional distribution on samples where anchor A applies. This methodology has successfully been applied in several tasks such as image recognition, text classification, and visual question answering.[97]

- *Counterfactual instance search.* Given a classifier model f and an original data point x, counterfactual instance search[98] aims to find examples x' that (i) are as close to x as possible, and (ii) for which the classifier produces a different class label from the label assigned to x. In other words, a counterfactual describes small feature changes in sample x such that it is classified differently by f. The search for the set of instances x' may be cast into an optimization problem:

$$\min_{x'} \max_{\lambda} \left(f_t - p_t \right)^2 + \lambda L_1 \left(x', x \right), \qquad (3)$$

where f_t is the prediction of the model for the t-th class, p_t a user-defined target probability for the same
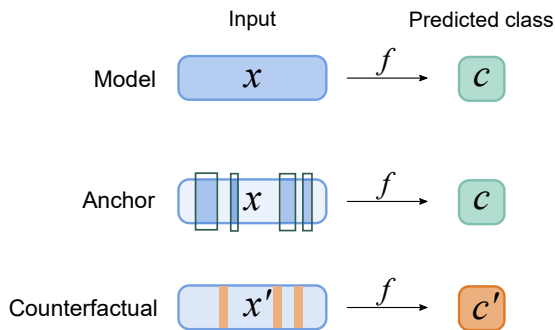
Figure 2: Instance-based model interpretation. Given a model f, input instance x and the respective predicted class c, so-called anchor algorithms identify a minimal subset of features of x that are sufficient to preserve the predicted class assignment c. Counterfactual search generates a new instance x' that lies close in feature space to x but is classified differently by the model, as belonging to class c'.

class, L_1 the Manhattan distance between the proposed x' and the original sample x, and λ an optimizable parameter that controls the contribution of each term in the loss. The first term in this loss encourages the search towards points that change the prediction of the model, while the second ensures that both x and x' lie close to each other in their input manifold. While in the original paper this approach was shown to successfully obtain counterfactuals in several data sets, the results revealed a tendency to look artificial. A recent methodology[99] mitigates this problem by adding extra terms to the loss function with an autoencoder architecture,[100] to better capture the original data distribution. Importantly, counterfactual instances can be evaluated using trust scores (*cf.* the section on uncertainty estimation). One can interpret a high trust score as the counterfactual being far from the initially predicted class of x compared to the class assigned to the counterfactual x'.

- *Contrastive explanation* methods[101] provide instance-based interpretability of classifiers by generating 'pertinent positive' (PP) and 'pertinent negative' (PN) sets. This methodology is, therefore, related to both anchors and counterfactual search approaches. PPs are defined as the smallest set of features that should be present in an instance in order for the model to predict a 'positive' result (similar to anchors). Conversely, PNs identify the smallest set of features that should be absent in order for the model to be able to sufficiently differentiate from the other classes (similar to a counterfactual instance). This method generates explanations of the form *"An input x is classified as class y because a subset of features $x_1, \ldots, x_k$ is present, and because a subset of features $x_m, \ldots, x_p$ is absent."*[96,102]. Contrastive explanation methods find such sets by solving two separate optimization problems, namely by (i) perturbing the original instance until it is predicted differently than its original class, and (ii) searching for critical features in the original

input (*i.e.,* those features that guarantee a prediction with a high degree of certainty). The proposed approach uses an elastic net regularizer,[103] and optionally a conditional autoencoder model[104] so that the found explanations are more likely to lie closer to the original data manifold.

To the best of our knowledge, instance-based approaches have yet to be applied to drug-discovery. In our opinion, they bear promise in several areas of de novo molecular design, such as (i) activity cliff prediction,[105,106] as they can help identify small structural variations in molecules that cause large bioactivity changes, (ii) fragment-based virtual screening,[107] by highlighting a minimal subset of atoms responsible for a given observed activity), and (iii) hit-to-lead optimization, by helping identify the minimal set of structural changes required to improve one or more biological or physicochemical properties.

2.3 Graph-convolution-based methods

Molecular graphs are a natural mathematical representation of molecular topology, with nodes and edges representing atoms and chemical bonds, respectively[108] (Figure 3a). Furthermore, their usage has been commonplace in chemoinformatics and mathematical chemistry since the late 1970s.[109–112] Thus, it does not come as a surprise in these fields to witness the increasing application of novel graph-convolution neural networks,[113] which formally fall under the umbrella of neural-message passing algorithms.[114–116] Generally speaking, convolution refers to a mathematical operation on two functions that produces a third function expressing how the shape of one is modified by the other. This concept is widely used in convolutional neural networks for image analysis. Graph convolutions naturally extend the convolution operation typically used in computer vision[117] or in natural language processing[118] applications to arbitrarily-sized graphs. In the context of drug discovery, graph convolutions have been applied to molecular property prediction[119,120] and in generative models for de novo drug design.[121]

Exploring the interpretability of models trained with graph-convolution architectures is currently a particularly active research topic. For the purpose of this review, XAI methods based on graph-convolution are grouped into the following two categories:

- *Sub-graph identification approaches* aim to identify one or more parts of a graph that are responsible for a given prediction (Figure 3b). GNNExplainer[122] is a model-agnostic example of such category, and provides explanations for any graph-based machine learning task. Given an individual input graph, GNNExplainer identifies a connected sub-graph structure, as well a set of node-level features that are relevant for a particular prediction. The method can also provide such explanations for a group of data points belonging to the same class. GNNExplainer is formulated as an optimization problem, where a mutual information objective between the prediction of a graph neural network and the distribution of feasible sub-graphs is maximized. Mathematically, given a node v, the goal is to identify a sub-graph $G_S \subseteq G$ with

associated features $X_S = \{x_j | v_j \in G_S\}$ that are relevant in explaining a target prediction $\hat{y} \in Y$ via a mutual information measure MI:

$$\max_{G_S} \text{MI}\left(Y, (G_S, X_S)\right) = H(Y) -$$
$$H(Y | G = G_S, X = X_S). \quad (4)$$

In practice, however, this objective is not mathematically tractable, and several continuity and convexity assumptions have to be made. GNNExplainer was tested on a set of molecules labeled for their mutagenic effect on *Salmonella typhimurium*,[123] and identified several known mutagenic functional groups (*i.e.*, certain aromatic and heteroaromatic rings and amino/nitro groups[123]) as relevant.

- *Attention-based approaches.* The interpretation of graph-convolutional neural networks can benefit from attention mechanisms,[124] which borrow from the natural language processing field, where their usage has become standard.[125] The idea is to stack several message-passing layers to obtain hidden node-level representations, by first computing attention coefficients associated with each of the edges connected to the neighbors of a particular node in the graph (Figure 3c). Mathematically, for a given node, an attention-based graph convolution operation obtains its updated hidden representation via a normalized sum of the node-level hidden features of the topological neighbours:

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^l W^{(l)} h_j^{(l)}\right), \quad (5)$$

where $\mathcal{N}(i)$ is the set of topological neighbours of node i with a one-edge distance, α_{ij}^l are learned attention coefficients over those neighbours, σ is a nonlinear activation function, and $W^{(l)}$ is a learnable feature matrix for layer l. The main difference between this approach and a standard graph convolution update is that, in the latter, attention coefficients are replaced by a fixed normalization constant $c_{ij} = \sqrt{|\mathcal{N}(i)|}\sqrt{|\mathcal{N}(j)|}$.

A recent study[49] describes how the interpretation of filters in message-passing networks can lead to the identification of relevant pharmacophore- and toxicophore-like substructures, showing consistent findings with literature reports. Gradient-based feature attribution techniques, such as Integrated Gradients,[70] were used in conjunction with graph-convolutional networks to analyze retrosynthetic reaction predictions and highlight the atoms involved in each reaction step.[64] Attention-based graph convolutional neural networks have also been used for the prediction of solubility, polarity, synthetic accessibility, and photovoltaic efficiency, among other properties,[65,66] leading to the identification of relevant molecular substructures for the target properties. Finally, attention-based graph architectures have also been used in chemical reactivity prediction,[19] pointing to structural motifs that are consistent
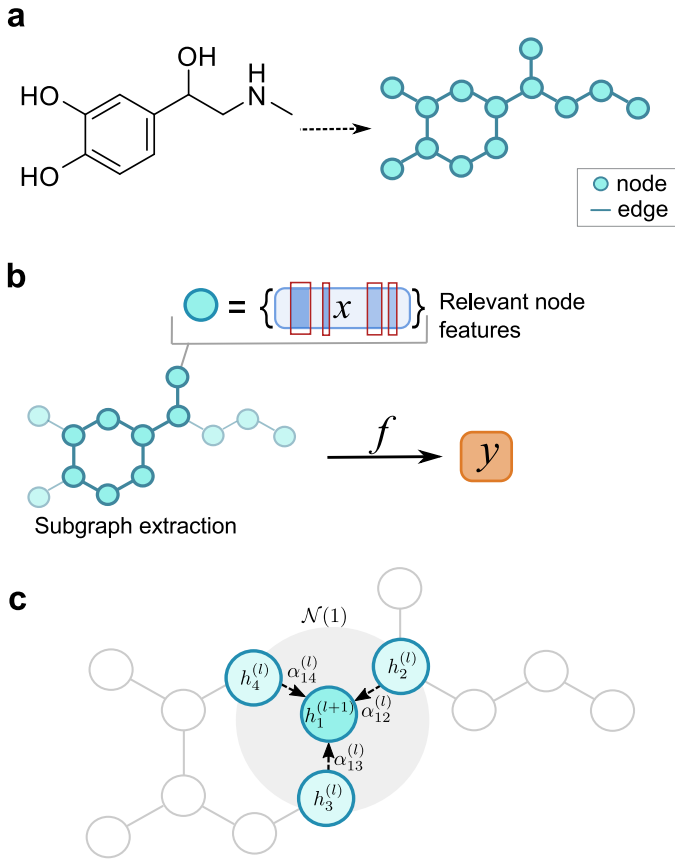

Figure 3: Graph-based model interpretation. (a) Kekulé structure of adrenaline and its respective molecular graph; atoms and bonds constitute nodes and edges, respectively. (b) Given an input graph, approaches like GNNExplainer[122] aim to identify a connected, compact subgraph, as well as node-level features that are relevant for a particular prediction y of a graph-neural network model f. (c) Attention mechanisms can be used in conjunction with message-passing algorithms in order to learn coefficients $\alpha_{ij}^{(l)}$ for the l-th layer, which assign 'importance' to the set of neighbours $\mathcal{N}(i)$ (*e.g.*, adjacent atoms) of a node i. These coefficients are an explicit component in the computation of new hidden-node representations $h_i^{(l+1)}$ (Eq. (5)) in attention-based graph-convolutional architectures. Such learned attention coefficients can be then used to highlight the predictive relevance of certain edges and nodes.

with a chemist's intuition in the identification of suitable reaction partners and activating reagents.

Due to their intuitive connection with the two-dimensional representation of molecules, graph-convolution-based XAI bears the potential of being applicable in many common modeling tasks in drug discovery. Some of the applications that might benefit most from the usage of these approaches are those aiming to identify relevant structural molecular motifs, *e.g.*, for structural alert identification,[126] site of reactivity[127] or metabolism[128] prediction, and the selective optimization of side activities.[129]

2.4 Self-explaining approaches

The XAI methods introduced so far produce *a posteriori* explanations of deep learning models. Although such *post-*

hoc interpretations have been shown to be useful, some argue that, ideally, XAI methods, should automatically offer human-interpretable explanation alongside their predictions.[130] Such approaches (herein referred to as '*self-explaining*') would promote verification and error analysis, and be directly linkable with domain knowledge.[131] While the term self-explaining has been coined to refer to a specific neural network architecture – self-explaining neural networks,[131] described below – in this review, the term is used in a broader sense, so as to identify methods that feature interpretability as a central part of their design. Self-explaining XAI approaches can be grouped into the following categories:

- *Prototype-based reasoning* refers to the task of forecasting future events (*i.e.*, novel samples) based on particularly informative known data points. Usually, this is done by identifying *prototypes*, *i.e.*, representative samples, which are adapted (or used directly) to make a prediction. These methods are motivated by the fact that predictions based on individual, previously seen examples mimic human decision making.[132] The Bayesian case model,[133] is a *pre*-deep-learning approach that constitutes a general framework for such prototype-based reasoning. A Bayesian case model learns to identify observations that best represent clusters in a dataset (*i.e.*, prototypes), along with a set of defining features for that cluster. Joint inference is performed on cluster labels, prototypes, and extracted relevant features, thereby providing interpretability without sacrificing classification accuracy.[133] Recently, Li *et al.*[134] developed a neural network architecture composed of an autoencoder and a therein named 'prototype layer', whose units store a learnable weight vector representing an encoded training input. Distances between the encoded latent space of new inputs and the learned prototypes are then used as part of the prediction process. This approach was later expanded by Chen *et al.*[135] to convolutional neural networks for computer vision tasks.

- *Self-explaining neural networks*[131] aim to associate input or latent features with semantic concepts. They jointly learn a class prediction and generate explanations using a feature-to-concept mapping. Such a network model consists of (i) a sub-network that maps raw inputs into a predefined set of explainable concepts, (ii) a parameterizer that obtains coefficients for each individual explainable concept, and (iii) an aggregation function that combines the output of the previous two components to produce the final class prediction.

- *Human-interpretable concept learning* refers to the task of learning a class of concepts, *i.e.* high-level combinations of knowledge elements[136]) from data, aiming to achieve human-like generalization ability. The Bayesian Program Learning approach[137] was proposed with the goal of learning visual concepts in computer vision tasks. Such concepts were represented as probabilistic programs expressed as structured procedures in an abstract description language.[138] The model then composes more complex programs using the elements of previously learned ones using

a Bayesian criterion. This approach was shown to reach human-like performance in one-shot learning tasks.[139,140]

- *Testing with concept activation vectors*[141] computes the directional derivatives of the activations of a layer *w.r.t.* its input, towards the direction of a concept. Such derivatives quantify the degree to which the latter is relevant for a particular classification (*e.g.*, how important the concept 'stripes' is for the prediction of the class 'zebra'). It does so by considering the mathematical association between the internal state of a machine learning model – seen as a vector space E_m spanned by basis vectors e_m that correspond to neural activations – and human-interpretable activations residing in a different vector space E_h spanned by basis vectors e_h. A linear function is computed that translates between these vector spaces ($g : E_m \rightarrow E_h$). The association is achieved by defining a vector in the direction of the values of a concept's set of examples, and then training a linear classifier between those and random counterexamples, to finally take the vector orthogonal to the decision boundary.

- *Natural language explanation generation.* Deep networks can be designed to generate human-understandable explanations in a supervised manner.[142] In addition to minimizing the loss of the main modeling task, several approaches synthesize a sentence using natural language that explains the decision performed by the model, by simultaneously training generators on large data sets of human-written explanations. This approach has been applied to generate explanations that are both image- and class-relevant.[143] Another prominent application is visual question answering.[144] To obtain meaningful explanations, however, this approach requires a significant amount of human-curated explanations for training, and might, thus, find limited applicability in drug discovery tasks.

To the best of our knowledge, self-explaining deep learning has not been applied to chemistry or drug design yet. Including interpretability by design could help bridge the gap between machine representation and the human understanding of many types of problems in drug discovery. For instance, prototype-reasoning bears promise in the modeling of heterogeneous sets of chemicals with different modes of action, allowing to preserve both mechanistic interpretability and predictive accuracy. Explanation generation (either concept- or text-based) is another potential solution to include human-like reasoning and domain knowledge in the model building task. In particular, explanation generation approaches might be applicable to certain decision-making processes, such as the replacement of animal testing and *in-vitro* to *in-vivo* extrapolation, where human-understandable generated explanations constitute a crucial element.

2.5 Uncertainty estimation

Uncertainty estimation, *i.e.* the quantification of epistemic error in a prediction, constitutes another approach to

model interpretation. While some machine learning algorithms, such as Gaussian processes,[145] provide built-in uncertainty estimation, deep neural networks are known for being poor at quantifying uncertainty.[146] This is one of the reasons why several efforts have been devoted to specifically quantify uncertainty in neural network-based predictions. Uncertainty estimation methods can be grouped into the following categories:

- *Ensemble approaches.* Model ensembles improve the overall prediction quality and have become a standard for uncertainty estimates.[147] Deep ensemble averaging[148] is based on m identical neural network models that are trained on the same data and with a different initialization. The final prediction is obtained by aggregating the predictions of all models (*e.g.,* by averaging), while an uncertainty estimate can be obtained from the respective variance (Figure 4a). Similarly, the sets of data on which these models are trained can be generated via bootstrap re-sampling.[149] A disadvantage of this approach is its computational demand, as the underlying methods build on m independently trained models. Snapshot ensembling[150] aims to overcome this limitation by periodically storing model states (*i.e.* model parameters) along the the training optimization path. These model 'snapshots' can be then used for constructing the ensemble.

- *Probabilistic approaches* aim to estimate the posterior probability of a certain model output or to perform *post-hoc* calibration. Many of these methods treat neural networks as Bayesian models, by considering a prior distribution over its learnable weights, and then performing inference over their posterior distribution with various methods (Figure 4b), *e.g.,* Markov Chain Monte Carlo[151] or variational inference.[152, 153] Gal *et al.*,[154] suggested the usage of dropout regularization to perform approximate Bayesian inference, which was later extended[155] to compute epistemic (*i.e.*, caused by model mis-specification) and aleatory uncertainty (inherent to the noise in the data). Similar approximations can also be made via batch normalization.[156] Mean variance estimation[157] considers a neural network designed to output both a mean and variance value, to then train the model using a negative Gaussian log-likelihood loss function. Another subcategory of approaches consider asymptotic approximations of a prediction by making Gaussian distributional assumptions of its error, such as the delta technique.[158, 159]

- *Other approaches.* The LUBE (lower upper bound estimation)[160] approach trains a neural network with two outputs, corresponding to the upper and lower bounds of the prediction. Instead of quantifying the error of single predictions, LUBE uses simulated annealing and optimizes the model coefficients to achieve (a) maximum coverage (probability that the real value of the i-th sample will fall between the upper and the lower bound) of training measurements and (b) minimum prediction interval width. Ak *et al.* suggested to quantify the uncertainty in neural network models by directly modelling interval-valued data.[161] Trust scores[162] measure the agreement between a neural network and a k-nearest neighbour
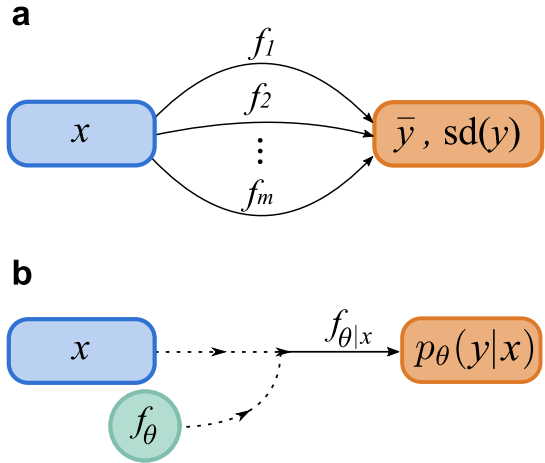
Figure 4: Uncertainty estimation. (a) Ensemble-based methods aggregate the output of m identical, but differently-initialized, models f_i. The final prediction is obtained by aggregating the predictions of all models (*e.g.,* as the average, $\bar{y}$), while an uncertainty estimate can be obtained from the respective predictive variance, *e.g.,* in the form of a standard deviation, $\mathrm{sd}(y)$. (b) Bayesian probabilistic approaches consider a prior $p(\theta)$ over the learnable weights of a neural network model f_θ, and make use of approximate sampling approaches to learn a posterior distribution over both the weights $p(\theta|x)$ and the prediction $p_\theta(y|x)$. These distributions can be then sampled from to obtain uncertainty estimates over both the weights and the predictions.

classifier that is trained on a filtered subset of the original data. The trust score considers both the distance between the instance of interest to the nearest class that is different from the original predicted one and its distance towards the predicted class. Union-based methods[163] first train a neural network model and then feed its embeddings to a second model that handles uncertainty, such as a Gaussian process or a random forest. Distance-based approaches[164] aim to estimate the prediction uncertainty of a new sample x' by measuring the distance to the closest sample in the training set, either using input features[165] or an embedding produced by the model.[166]

Many uncertainty estimation approaches have been successfully implemented in drug discovery applications,[167] mostly in traditional QSAR modeling, either by the use of models that naturally handle uncertainty[168] or post-hoc methods.[169–171] Attention has recently been drawn toward the development of uncertainty-aware deep-learning applications in the field. 'Snapshot ensembling' was applied to model 24 bioactivity datasets,[172] showing that it performs *on par* with random forest and neural network ensembles, and also leads to narrower confidence intervals. Schwaller *et al.*[67] proposed a transformer model[173] for the task of forward chemical reaction prediction. This approach implements uncertainty estimation by computing the product of the probabilities of all predicted tokens in a SMILES sequence representing a molecule. Zhang *et al.*[68] have recently proposed a Bayesian treatment of a semi-supervised graph-neural network for uncertainty-calibrated predictions of molecular properties, such as the

melting point and aqueous solubility. Their results suggest that this approach can efficiently drive an active learning cycle, particularly in the low-data regime – by choosing those molecules with the largest estimated epistemic uncertainty.

Importantly, a recent comparison of several uncertainty estimation methods for physicochemical property prediction showed that none of the methods systematically outperformed all others.[174]

3 Available software

Given the attention deep learning applications are currently receiving, several software tools have been developed to facilitate model interpretation. A prominent example is Captum,[175] an extension of the PyTorch[176] deep-learning and automatic differentiation package that provides support for most of the feature attribution techniques described in this work. Another popular package is Alibi,[177] which provides instance-specific explanations for certain models trained with the scikit-learn[178] or TensorFlow[179] packages. Some of the explanation methods implemented include anchors, contrastive explanations, and counterfactual instances.

4 Conclusions and outlook

Automated analysis of medical and chemical knowledge to extract and present features in a human-readable format dates back to the 1990s,[180,181] but it's receiving increasing attention in the last years due to the *renaissance* of neural networks in chemistry and healthcare. Given the current pace of AI in drug discovery and related fields, there will be an increased demand for methods that help us understand and interpret deep learning models.

For drug discovery and design in particular, XAI will foster collaborations between medicinal chemists, chemoinformaticians and data scientists. If successful, XAI bears the potential to provide fundamental support in the analysis and interpretation of increasingly more complex chemical data, as well as in the formulation of new pharmacological hypotheses, while avoiding human biases.[182,183] At the same time, novel drug discovery challenges might boost the development of application-tailored XAI approaches, to promptly respond to specific scientific questions related to human biology and pathophysiology.

It will be important to explore the opportunities and limitations of the established chemical language for representing the decision space of these models. Additionally, we have to concede our incomplete understanding of the human pathology at the molecular level, with all its individual idiosyncrasies. In this context, full comprehensibility may be hard to achieve, given, for instance, the often nonlinear relationships between chemical structure and pharmacological activity, the presence of error, and limited predictability.[184]

XAI also poses technical challenges, given the multiplicity of possible explanations and methods applicable to a given task.[185] Most approaches do not come as readily usable out-of-the-box solutions, but need to be tailored to each individual application. Additionally, profound knowledge of the problem domain is crucial to identify which model decisions demand further explanations, which type of answers are meaningful to the user and which are instead trivial or expected.[186] For human decision making, the explanations generated with XAI have to be *non-trivial*, *non-artificial*, and *sufficiently informative* for the respective scientific community. At least for the time being, finding such solutions will require the collaborative efforts of deep-learning experts, chemoinformaticians and data scientists, chemists, biologists and other domain experts, to ensure that XAI methods serve their intended purpose and deliver reliable answers.

A crucial challenge for future AI-assisted drug discovery is the data representation used for machine and deep learning. In contrast to many other areas in which deep learning has been shown to excel, such as natural language processing and image recognition, there is no naturally applicable, complete, 'raw' molecular representation. After all, molecules – as scientists conceive them – are models themselves. Such 'induction-based' approach, which builds higher-order (*e.g.*, deep learning) models from lower-order ones (*e.g.*, molecular representations or descriptors based on observational statements) is therefore philosophically debatable.[187] The question as to how to represent molecules for deep learning thus constitutes still one of the fundamental challenges of XAI in drug discovery and design.

One step forward towards explainable AI is to build on interpretable 'low-level' molecular representations that have direct meaning for chemists (*e.g.* SMILES strings,[188,189] amino acid sequences,[190,191] and spatial 3D-voxelized representations[94,192]). Many recent studies also rely on well-established molecular descriptors, such as hashed binary fingerprints[193,194] and topochemical and geometrical descriptors,[195,196] which capture structural features defined *a priori*. Often, molecular descriptors, while being relevant for subsequent modeling, capture complex chemical information in a non readily-interpretable way. Consequently, when striving for XAI, there is an understandable tendency to employ molecular representations that are more easily interpretable in terms of the known language of chemistry.[197] It therefore goes without saying that the development of novel interpretable molecular representations for deep learning will constitute a critical area of research for the years to come, including the development of self-explaining approaches to overcome the hurdles of non-interpretable but information-rich descriptors, by providing human-like explanations alongside sufficiently accurate predictions.

Most of the deep learning models in drug discovery do not consider applicability domain issues,[198] *i.e.*, the region of chemical space where statistical learning assumptions are met.[199] Inclusion of applicability domain restrictions – in an interpretable way, whenever possible – should be considered as an integral element of explainable AI. Knowing when to apply which particular model will, in fact, help address the problem of high confidence of deep learning models on wrong predictions,[146,200,201] avoiding unnecessary extrapolations at the same time. Assessing a model's applicability domain and rigorous determination of model accuracy might arguably be more important for decision making than the particular modeling approach chosen.[69]

At present, XAI in drug discovery is lacking open-community platforms, where software code, model inter-

pretations, and the respective training data can be shared and improved by synergistic efforts of researchers with different scientific backgrounds. Initiatives like MELLODDY (*Machine Learning Ledger Orchestration for Drug Discovery*) for decentralized, federated model development and secure data handling across pharmaceutical companies constitute a first step in the right direction.[202] Such kinds of collaboration will hopefully foster the validation and acceptance of XAI approaches and the associated explanations they provide.

Especially in time- and cost-sensitive scenarios like drug discovery, the users of deep learning methods have the responsibility to cautiously inspect and interpret the predictions made by such a computational model. Keeping in mind the possibilities and limitations of drug discovery with XAI, it is reasonable to assume that the continued development of alternative models that are more easily comprehensible and computationally affordable will not lose its importance.

Conflict of interest

Author contributions

All authors contributed equally to this manuscript.

Acknowledgements

Related links

- PyTorch Captum: `captum.ai`
- Alibi: `github.com/SeldonIO/alibi`
- MELLODDY Consortium: `melloddy.eu`

References

[1] Gawehn, E., Hiss, J. A. & Schneider, G. Deep learning in drug discovery. *Molecular Informatics* **35**, 3–14 (2016).

[2] Zhang, L., Tan, J., Han, D. & Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today* **22**, 1680–1685 (2017).

[3] Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **23**, 1241 – 1250 (2018).

[4] Tang, W., Chen, J., Wang, Z., Xie, H. & Hong, H. Deep learning for predicting toxicity of chemicals: A mini review. *Journal of Environmental Science and Health, Part C* **36**, 252–271 (2018).

[5] Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chemical Reviews* **119**, 10520–10594 (2019).

[6] Muratov, E. N. *et al.* QSAR without borders. *Chemical Society Reviews* **49**, 3525–3564 (2020).

[7] Hemmerich, J. & Ecker, G. F. In silico toxicology: From structure-activity relationships towards deep learning and adverse outcome pathways. *WIREs Computational Molecular Science* **10**, e1475 (2020).

[8] LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

[9] Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Networks* **61**, 85–117 (2015).

[10] Lenselink, E. B. *et al.* Beyond the hype: Deep neural networks outperform established methods using a chembl bioactivity benchmark set. *Journal of Cheminformatics* **9**, 1–14 (2017).

[11] Goh, G. B., Siegel, C., Vishnu, A., Hodas, N. O. & Baker, N. Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. *arXiv preprint arXiv:1706.06689* (2017).

[12] Unterthiner, T. *et al.* Deep learning as an opportunity in virtual screening. In *Proceedings of the Deep Learning Workshop at NIPS*, vol. 27, 1–9 (2014).

[13] Wallach, I., Dzamba, M. & Heifets, A. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855* (2015).

[14] Méndez-Lucio, O., Baillif, B., Clevert, D.-A., Rouquié, D. & Wichard, J. De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. *Nature Communications* **11**, 1–10 (2020).

[15] Merk, D., Friedrich, L., Grisoni, F. & Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Molecular Informatics* **37**, 1700153 (2018).

[16] Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology* **37**, 1038–1040 (2019).

[17] Schwaller, P., Gaudin, T., Lanyi, D., Bekas, C. & Laino, T. 'Found in translation': Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chemical Science* **9**, 6091–6098 (2018).

[18] Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Accounts of Chemical Research* **51**, 1281–1289 (2018).

[19] Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical Science* **10**, 370–377 (2019).

[20] Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

[21] Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences U.S.A.* **117**, 1496–1503 (2020).

[22] Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: Deep drug–target binding affinity prediction. *Bioinformatics* **34**, i821–i829 (2018).

[23] Zeng, X. *et al.* Target identification among known drugs by deep learning from heterogeneous networks. *Chemical Science* **11**, 1775–1797 (2020).

[24] Jimenez, J. *et al.* Pathwaymap: Molecular pathway association with self-normalizing neural networks. *Journal of Chemical Information and Modeling* **59**, 1172–1181 (2018).

[25] Marchese Robinson, R. L., Palczewska, A., Palczewski, J. & Kidley, N. Comparison of the predictive performance and interpretability of random forest and linear models on benchmark data sets. *Journal of Chemical Information and Modeling* **57**, 1773–1792 (2017).

[26] Webb, S. J., Hanser, T., Howlin, B., Krause, P. & Vessey, J. D. Feature combination networks for the interpretation of statistical machine learning models: Application to Ames mutagenicity. *Journal of Cheminformatics* **6**, 1–21 (2014).

[27] Grisoni, F., Consonni, V. & Ballabio, D. Machine learning consensus to predict the binding to the Androgen receptor within the CoMPARA project. *Journal of Chemical Information and Modeling* **59**, 1839–1848 (2019).

[28] Polishchuk, P. *et al.* Structural and physico-chemical interpretation (SPCI) of QSAR models and its comparison with matched molecular pair analysis. *Journal of Chemical Information and Modeling* **56**, 1455–1469 (2016).

[29] Kuz'min, V. E., Polishchuk, P. G., Artemenko, A. G. & Andronati, S. A. Interpretation of QSAR models based on random forest methods. *Molecular Informatics* **30**, 593–603 (2011).

[30] Chen, Y., Stork, C., Hirte, S. & Kirchmair, J. NP-scout: Machine learning approach for the quantification and visualization of the natural product-likeness of small molecules. *Biomolecules* **9**, 43 (2019).

[31] Rosenbaum, L., Hinselmann, G., Jahn, A. & Zell, A. Interpreting linear support vector machine models with heat map molecule coloring. *Journal of Cheminformatics* **3**, 11 (2011).

[32] Riniker, S. & Landrum, G. A. Similarity maps-a visualization strategy for molecular fingerprints and machine-learning methods. *Journal of Cheminformatics* **5**, 43 (2013).

[33] Grisoni, F., Consonni, V., Vighi, M., Villa, S. & Todeschini, R. Investigating the mechanisms of bioconcentration through QSAR classification trees. *Environment International* **88**, 198–205 (2016).

[34] Todeschini, R., Lasagni, M. & Marengo, E. New molecular descriptors for 2D and 3D structures. Theory. *Journal of Chemometrics* **8**, 263–272 (1994).

[35] Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215 (2019).

[36] Schneider, G. Mind and machine in drug design. *Nature Machine Intelligence* **1**, 128–130 (2019).

[37] Lipton, Z. C. The mythos of model interpretability. *Queue* **16**, 31–57 (2018).

[38] Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R. & Yu, B. Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592* (2019).

[39] Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[40] Lapuschkin, S. *et al.* Unmasking clever Hans predictors and assessing what machines really learn. *Nature Communications* **10**, 1–8 (2019).

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

[42] Xu, Y., Pei, J. & Lai, L. Deep learning based regression and multiclass models for acute oral toxicity prediction with automatic chemical feature extraction. *Journal of Chemical Information and Modeling* **57**, 2672–2685 (2017).

[43] Ciallella, H. L. & Zhu, H. Advancing computational toxicology in the big data era by artificial intelligence: Data-driven and mechanism-driven modeling for chemical toxicity. *Chemical Research in Toxicology* **32**, 536–547 (2019).

[44] Dey, S., Luo, H., Fokoue, A., Hu, J. & Zhang, P. Predicting adverse drug reactions through interpretable deep learning framework. *BMC Bioinformatics* **19**, 1–13 (2018).

[45] Yan, C. *et al.* Interpretable retrosynthesis prediction in two steps. *ChemRxiv preprint* (2020).

[46] 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* **59**, 1324–1337 (2019).

[47] Manica, M. *et al.* Toward explainable anticancer compound sensitivity prediction via multimodal attention-based convolutional encoders. *Molecular Pharmaceutics* **16** (2019).

[48] Žuvela, P., David, J. & Wong, M. W. Interpretation of ann-based QSAR models for prediction of antioxidant activity of flavonoids. *Journal of Computational Chemistry* **39**, 953–963 (2018).

[49] Preuer, K., Klambauer, G., Rippmann, F., Hochreiter, S. & Unterthiner, T. *Interpretable deep learning in drug discovery*, 331–345 (Springer, 2019). Editors: Samek W., Montavon G., Vedaldi A., and Hansen L., MÃijller KR.

[50] Gupta, M., Lee, H. J., Barden, C. J. & Weaver, D. F. The blood–brain barrier (BBB) score. *Journal of Medicinal Chemistry* **62**, 9824–9836 (2019).

[51] Rankovic, Z. CNS physicochemical property space shaped by a diverse set of molecules with experimentally determined exposure in the mouse brain: Miniperspective. *Journal of Medicinal Chemistry* **60**, 5943–5954 (2017).

[52] Wager, T. T., Hou, X., Verhoest, P. R. & Villalobos, A. Central nervous system multiparameter optimization desirability: Application in drug discovery. *ACS Chemical Neuroscience* **7**, 767–775 (2016).

[53] Ritchie, T. J., Macdonald, S. J., Peace, S., Pickett, S. D. & Luscombe, C. N. Increasing small molecule drug developability in sub-optimal chemical space. *MedChemComm* **4**, 673–680 (2013).

[54] Leeson, P. D. & Young, R. J. Molecular property design: Does everyone get it? *ACS Medicinal Chemistry Letters* **6**, 722–725 (2015).

[55] Fujita, T. & Winkler, D. A. Understanding the roles of the âĂIJtwo QSARsâĂİ. *Journal of Chemical Information and Modeling* **56**, 269–274 (2016).

[56] Schneider, P. *et al.* Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery* **19**, 353âĂŞ364 (2020).

[57] Guidotti, R. *et al.* A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* **51**, 1–42 (2018).

[58] Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 2522–5839 (2020).

[59] McCloskey, K., Taly, A., Monti, F., Brenner, M. P. & Colwell, L. J. Using attribution to decode binding mechanism in neural network models for chemistry. *Proceedings of the National Academy of Sciences U.S.A.* **116**, 11624–11629 (2019).

[60] 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* (2019).

[61] Rodríguez-Pérez, R. & Bajorath, J. Interpretation of machine learning models using Shapley values: application to compound potency and multi-target activity predictions. *Journal of Computer-Aided Molecular Design* (2020).

[62] Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E. & Hoffmann, H. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10772–10781 (2019).

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

[64] Ishida, S., Terayama, K., Kojima, R., Takasu, K. & Okuno, Y. Prediction and interpretable visualization of retrosynthetic reactions using graph convolutional networks. *Journal of Chemical Information and Modeling* **59**, 5026–5033 (2019).

[65] Shang, C. *et al.* Edge attention-based multi-relational graph convolutional networks. *arXiv preprint arXiv:1802.04944* (2018).

[66] Ryu, S., Lim, J., Hong, S. H. & Kim, W. Y. Deeply learning molecular structure-property relationships using attention-and gate-augmented graph convolutional network. *arXiv preprint arXiv:1805.10988* (2018).

[67] Schwaller, P. *et al.* Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science* **5**, 1572–1583 (2019).

[68] Zhang, Y. & Lee, A. A. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chemical Science* **10**, 8154–8163 (2019).

[69] Liu, R., Wang, H., Glover, K. P., Feasel, M. G. & Wallqvist, A. Dissecting machine-learning prediction of molecular activity: Is an applicability domain needed for quantitative structure–activity relationship models based on deep neural networks? *Journal of Chemical Information and Modeling* **59**, 117–126 (2019).

[70] Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3319–3328 (JMLR. org, 2017).

[71] Smilkov, D., Thorat, N., Kim, B., Viégas, F. & Wattenberg, M. Smoothgrad: Removing noise by adding noise. *arXiv preprint arXiv:1706.03825* (2017).

[72] Adebayo, J., Gilmer, J., Goodfellow, I. & Kim, B. Local explanation methods for deep neural networks lack sensitivity to parameter values. *arXiv preprint arXiv:1810.03307* (2018).

[73] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).

[74] Adebayo, J. *et al.* Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 9505–9515 (2018).

[75] Lipovetsky, S. & Conklin, M. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry* **17**, 319–330 (2001).

[76] Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 4765–4774 (2017).

[77] Ribeiro, M. T., Singh, S. & Guestrin, C. "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*, 1135–1144 (2016).

[78] Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3145–3153 (JMLR. org, 2017).

[79] Bach, S. *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE* **10**, 1–46 (2015).

[80] Lakkaraju, H., Kamar, E., Caruana, R. & Leskovec, J. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154* (2017).

[81] Deng, H. Interpreting tree ensembles with intrees. *International Journal of Data Science and Analytics* **7**, 277–287 (2019).

[82] Bastani, O., Kim, C. & Bastani, H. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504* (2017).

[83] Maier, H. R. & Dandy, G. C. The use of artificial neural networks for the prediction of water quality parameters. *Water Resources Research* **32**, 1013–1022 (1996).

[84] Balls, G., Palmer-Brown, D. & Sanders, G. Investigating microclimatic influences on ozone injury in clover (Trifolium subterraneum) using artificial neural networks. *New Phytologist* **132**, 271–280 (1996).

[85] Sung, A. Ranking importance of input parameters of neural networks. *Expert Systems with Applications* **15**, 405–411 (1998).

[86] Štrumbelj, E., Kononenko, I. & Šikonja, M. R. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering* **68**, 886–904 (2009).

[87] Fong, R. C. & Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, 3429–3437 (2017).

[88] Olden, J. D. & Jackson, D. A. Illuminating the âĂIJblack boxâĂİ: A randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* **154**, 135 – 150 (2002).

[89] Zintgraf, L. M., Cohen, T. S., Adel, T. & Welling, M. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595* (2017).

[90] Ancona, M., Ceolini, E., Öztireli, C. & Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104* (2017).

[91] Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626 (2017).

[92] Zhang, J. *et al.* Top-down neural attention by excitation backprop. *International Journal of Computer Vision* **126**, 1084–1102 (2018).

[93] Tice, R. R., Austin, C. P., Kavlock, R. J. & Bucher, J. R. Improving the human hazard characterization of chemicals: A Tox21 update. *Environmental Health Perspectives* **121**, 756–765 (2013).

[94] Jiménez-Luna, J., Skalic, M., Martinez-Rosell, G. & De Fabritiis, G. K_{DEEP}: Protein–ligand absolute binding affinity prediction via 3D-convolutional neural networks. *Journal of Chemical Information and Modeling* **58**, 287–296 (2018).

[95] Jiménez-Luna, J. *et al.* DeltaDelta neural networks for lead optimization of small molecule potency. *Chemical Science* **10**, 10911–10918 (2019).

[96] Doshi-Velez, F. *et al.* Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134* (2017).

[97] Ribeiro, M. T., Singh, S. & Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

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

[99] Van Looveren, A. & Klaise, J. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584* (2019).

[100] Kramer, M. A. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal* **37**, 233–243 (1991).

[101] Dhurandhar, A. *et al.* Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, 592–603 (2018).

[102] Herman, A. Are you visually intelligent? What you don't see is as important as what you do see. *Medical Daily* (2016).

[103] Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320 (2005).

[104] Mousavi, A., Dasarathy, G. & Baraniuk, R. G. Deepcodec: Adaptive sensing and recovery via deep convolutional neural networks. *arXiv preprint arXiv:1707.03386* (2017).

[105] Stumpfe, D. & Bajorath, J. Exploring activity cliffs in medicinal chemistry: Miniperspective. *Journal of Medicinal Chemistry* **55**, 2932–2942 (2012).

[106] Cruz-Monteagudo, M. *et al.* Activity cliffs in drug discovery: Dr. Jekyll or Mr. Hyde? *Drug Discovery Today* **19**, 1069–1080 (2014).

[107] Erlanson, D. A. Introduction to fragment-based drug discovery. In *Fragment-based Drug Discovery and X-ray Crystallography*, 1–32 (Springer, Berlin, 2011). Editors: Davies T., and HyvĂűnen M.

[108] Todeschini, R. & Consonni, V. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references*, vol. 41 (WileyâĂŘVCH, Weinheim, 2009). Editors: Mannhold, R., Kubinyi, H., and Folkers, G.

[109] Randić, M., Brissey, G. M., Spencer, R. B. & Wilkins, C. L. Search for all self-avoiding paths for molecular graphs. *Computers & Chemistry* **3**, 5–13 (1979).

[110] Kier, L. B. A shape index from molecular graphs. *Quantitative Structure-Activity Relationships* **4**, 109–116 (1985).

[111] Randić, M. On unique numbering of atoms and unique codes for molecular graphs. *Journal of Chemical Information and Computer Sciences* **15**, 105–108 (1975).

[112] Bonchev, D. & Trinajstić, N. Information theory, distance matrix, and molecular branching. *The Journal of Chemical Physics* **67**, 4517–4533 (1977).

[113] Duvenaud, D. K. *et al.* Convolutional networks on graphs for learning molecular fingerprints. In *Advances in Neural Information Processing Systems*, 2224–2232 (2015).

[114] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1263–1272 (JMLR. org, 2017).

[115] Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).

[116] Nguyen, H., Maeda, S.-i. & Oono, K. Semi-supervised learning of hierarchical representations of molecules using neural message passing. *arXiv preprint arXiv:1711.10168* (2017).

[117] Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 1097–1105 (2012).

[118] Kim, Y. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[119] Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: Moving beyond fingerprints. *Journal of Computer-Aided Molecular Design* **30**, 595–608 (2016).

[120] Wu, Z. *et al.* Moleculenet: A benchmark for molecular machine learning. *Chemical Science* **9**, 513–530 (2018).

[121] Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *arXiv preprint arXiv:1802.04364* (2018).

[122] Ying, Z., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems*, 9240–9251 (2019).

[123] Debnath, A. K., Lopez de Compadre, R. L., Debnath, G., Shusterman, A. J. & Hansch, C. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry* **34**, 786–797 (1991).

[124] Veličković, P. *et al.* Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017).

[125] Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[126] Limban, C. *et al.* The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toxicology Reports* **5**, 943–953 (2018).

[127] Hughes, T. B., Miller, G. P. & Swamidass, S. J. Site of reactivity models predict molecular reactivity of diverse chemicals with glutathione. *Chemical Research in Toxicology* **28**, 797–809 (2015).

[128] Kirchmair, J. *et al.* Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics, and mechanisms. *Journal of Chemical Information and Modeling* **52**, 617–648 (2012).

[129] Wermuth, C. G. Selective optimization of side activities: The SOSA approach. *Drug Discovery Today* **11**, 160–164 (2006).

[130] Laugel, T., Lesot, M.-J., Marsala, C., Renard, X. & Detyniecki, M. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint arXiv:1907.09294* (2019).

[131] Melis, D. A. & Jaakkola, T. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, 7775–7784 (2018).

[132] Leake, D. B. *Case-based reasoning: Experiences, lessons and future directions* (MIT press, Cambridge, U.S.A., 1996). Editor: Leake, David B.

[133] Kim, B., Rudin, C. & Shah, J. A. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, 1952–1960 (2014).

[134] Li, O., Liu, H., Chen, C. & Rudin, C. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).

[135] Chen, C. *et al.* This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, 8928–8939 (2019).

[136] Goodman, N. D., Tenenbaum, J. B. & Gerstenberg, T. Concepts in a probabilistic language of thought. Tech. Rep., Center for Brains, Minds and Machines (CBMM) (2014).

[137] Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).

[138] Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature* **521**, 452–459 (2015).

[139] Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K. & Wierstra, D. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, 3630–3638 (2016).

[140] Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low-data drug discovery with one-shot learning. *ACS Central Science* **3**, 283–293 (2017).

[141] Kim, B. *et al.* Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). *arXiv preprint arXiv:1711.11279* (2017).

[142] Gilpin, L. H. *et al.* Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89 (IEEE, 2018).

[143] Hendricks, L. A. *et al.* Generating visual explanations. In *European Conference on Computer Vision*, 3–19 (Springer, 2016).

[144] Antol, S. *et al.* VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, 2425–2433 (2015).

[145] Rasmussen, C. E. Gaussian processes in machine learning. In *Summer School on Machine Learning*, 63–71 (Springer, 2003).

[146] Nguyen, A., Yosinski, J. & Clune, J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 427–436 (2015).

[147] Hansen, L. K. & Salamon, P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**, 993–1001 (1990).

[148] Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 6402–6413 (2017).

[149] Freedman, D. A. Bootstrapping regression models. *The Annals of Statistics* **9**, 1218–1228 (1981).

[150] Huang, G. *et al.* Snapshot ensembles: Train one, get m for free. *arXiv preprint arXiv:1704.00109* (2017).

[151] Zhang, R., Li, C., Zhang, J., Chen, C. & Wilson, A. G. Cyclical stochastic gradient MCMC for bayesian deep learning. *arXiv preprint arXiv:1902.03932* (2019).

[152] Graves, A. Practical variational inference for neural networks. In *Advances in Neural Information Processing Systems*, 2348–2356 (2011).

[153] Sun, S., Zhang, G., Shi, J. & Grosse, R. Functional variational bayesian neural networks. *arXiv preprint arXiv:1903.05779* (2019).

[154] Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 1050–1059 (2016).

[155] Kendall, A. & Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, 5574–5584 (2017).

[156] Teye, M., Azizpour, H. & Smith, K. Bayesian uncertainty estimation for batch normalized deep networks. *arXiv preprint arXiv:1802.06455* (2018).

[157] Nix, D. A. & Weigend, A. S. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, vol. 1, 55–60 (IEEE, 1994).

[158] Chryssolouris, G., Lee, M. & Ramsey, A. Confidence interval prediction for neural network models. *IEEE Transactions on Neural Networks* **7**, 229–232 (1996).

[159] Hwang, J. G. & Ding, A. A. Prediction intervals for artificial neural networks. *Journal of the American Statistical Association* **92**, 748–757 (1997).

[160] Khosravi, A., Nahavandi, S., Creighton, D. & Atiya, A. F. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE Transactions on Neural Networks* **22**, 337–346 (2010).

[161] Ak, R., Vitelli, V. & Zio, E. An interval-valued neural network approach for uncertainty quantification in short-term wind speed prediction. *IEEE Transactions on Neural Networks and Learning Systems* **26**, 2787–2800 (2015).

[162] Jiang, H., Kim, B., Guan, M. & Gupta, M. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, 5541–5552 (2018).

[163] Huang, W., Zhao, D., Sun, F., Liu, H. & Chang, E. Scalable gaussian process regression using deep neural networks. In *Twenty-Fourth International Joint Conference on Artificial Intelligence* (2015).

[164] Sheridan, R. P., Feuston, B. P., Maiorov, V. N. & Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *Journal of Chemical Information and Computer Sciences* **44**, 1912–1928 (2004).

[165] Liu, R. & Wallqvist, A. Molecular similarity-based domain applicability metric efficiently identifies out-of-domain compounds. *Journal of Chemical Information and Modeling* **59**, 181–189 (2018).

[166] Janet, J. P., Duan, C., Yang, T., Nandy, A. & Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chemical Science* **10**, 7913–7922 (2019).

[167] Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P. & Green, W. H. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of Chemical Information and Modeling* (2020).

[168] Obrezanova, O., Csányi, G., Gola, J. M. & Segall, M. D. Gaussian processes: a method for automatic QSAR modeling of ADME properties. *Journal of Chemical Information and Modeling* **47**, 1847–1857 (2007).

[169] Schroeter, T. S. *et al.* Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *Journal of Computer-Aided Molecular Design* **21**, 485–498 (2007).

[170] Clark, R. D. *et al.* Using beta binomials to estimate classification uncertainty for ensemble models. *Journal of Cheminformatics* **6**, 34 (2014).

[171] Bosc, N. *et al.* Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *Journal of Cheminformatics* **11**, 4 (2019).

[172] Cortés-Ciriano, I. & Bender, A. Deep confidence: A computationally efficient framework for calculating reliable prediction errors for deep neural networks. *Journal of Chemical Information and Modeling* **59**, 1269–1281 (2018).

[173] Vaswani, A. *et al.* Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008 (2017).

[174] Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R. & Coley, C. W. Uncertainty quantification using neural networks for molecular property prediction. *arXiv preprint arXiv:2005.10036* (2020).

[175] Kokhlikyan, N. *et al.* Pytorch Captum. https://github.com/pytorch/captum (2019).

[176] Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8026–8037 (2019).

[177] Klaise, J., Van Looveren, A., Vacanti, G. & Coca, A. Alibi: Algorithms for monitoring and explaining machine learning models. https://github.com/SeldonIO/alibi (2020).

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

[179] Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283 (USENIX Association, Savannah, GA, 2016).

[180] Hirst, J. D., King, R. D. & Sternberg, M. J. Quantitative structure-activity relationships by neural networks and inductive logic programming. i. the inhibition of dihydrofolate reductase by pyrimidines. *Journal of Computer-Aided Molecular Design* **8**, 405–420 (1994).

[181] Fiore, M., Sicurello, F. & Indorato, G. An integrated system to represent and manage medical knowledge. *Medinfo. MEDINFO* **8**, 931–933 (1995).

[182] Kutchukian, P. S. *et al.* Inside the mind of a medicinal chemist: The role of human bias in compound prioritization during drug discovery. *PLOS ONE* **7**, e48476 (2012).

[183] Boobier, S., Osbourn, A. & Mitchell, J. B. Can human experts predict solubility better than computers? *Journal of Cheminformatics* **9**, 63 (2017).

[184] Schneider, P. & Schneider, G. De novo design at the edge of chaos: Miniperspective. *Journal of Medicinal Chemistry* **59**, 4077–4086 (2016).

[185] Lipton, Z. C. The doctor just won't accept that! *arXiv preprint arXiv:1711.08037* (2017).

[186] Goodman, B. & Flaxman, S. European Union regulations on algorithmic decision-making and a 'right to explanation'. *AI Magazine* **38**, 50–57 (2017).

[187] Bendassolli, P. F. Theory building in qualitative research: reconsidering the problem of induction. *Forum Qualitative Social Research* **14**, 20 (2013).

[188] Ikebata, H., Hongo, K., Isomura, T., Maezono, R. & Yoshida, R. Bayesian molecular design with a chemical language model. *Journal of Computer-Aided Molecular Design* **31**, 379–391 (2017).

[189] Segler, M. H., Kogej, T., Tyrchan, C. & Waller, M. P. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS Central Science* **4**, 120–131 (2018).

[190] Nagarajan, D. *et al.* Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *Journal of Biological Chemistry* **293**, 3492–3509 (2018).

[191] Müller, A. T., Hiss, J. A. & Schneider, G. Recurrent neural network model for constructive peptide design. *Journal of Chemical Information and Modeling* **58**, 472–479 (2018).

[192] Jiménez-Luna, J., Cuzzolin, A., Bolcato, G., Sturlese, M. & Moro, S. A deep-learning approach toward rational molecular docking protocol selection. *Molecules* **25**, 2487 (2020).

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

[194] Awale, M. & Reymond, J.-L. Atom pair 2D-fingerprints perceive 3D-molecular shape and pharmacophores for very fast virtual screening of ZINC and GDB-17. *Journal of Chemical Information and Modeling* **54**, 1892–1907 (2014).

[195] Todeschini, R. & Consonni, V. New local vertex invariants and molecular descriptors based on functions of the vertex degrees. *MATCH Commun. Math. Comput. Chem* **64**, 359–372 (2010).

[196] Katritzky, A. R. & Gordeeva, E. V. Traditional topological indexes vs electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research. *Journal of Chemical Information and Computer Sciences* **33**, 835–857 (1993).

[197] Vighi, M., Barsi, A., Focks, A. & Grisoni, F. Predictive models in ecotoxicology: Bridging the gap between scientific progress and regulatory applicabilityâĂŤremarks and research needs. *Integrated Environmental Assessment and Management* **15**, 345–351 (2019).

[198] Sahigara, F. *et al.* Comparison of different approaches to define the applicability domain of qsar models. *Molecules* **17**, 4791–4810 (2012).

[199] Mathea, M., Klingspohn, W. & Baumann, K. Chemoinformatic classification methods and their applicability domain. *Molecular Informatics* **35**, 160–180 (2016).

[200] Szegedy, C. *et al.* Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[201] Hendrycks, D. & Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016).

[202] Hale, C. The MELLODDY project. *Fierce Biotech* (2019).

[203] Nembri, S., Grisoni, F., Consonni, V. & Todeschini, R. In silico prediction of cytochrome P450-drug interaction: QSARs for CYP3A4 and CYP2C9. *International Journal of Molecular Sciences* **17**, 914 (2016).

[204] Hiratsuka, M. *et al.* Characterization of human cytochrome p450 enzymes involved in the metabolism of cilostazol. *Drug Metabolism and Disposition* **35**, 1730–1732 (2007).

[205] Foti, R. S., Rock, D. A., Wienkers, L. C. & Wahlstrom, J. L. Selection of alternative CYP3A4 probe substrates for clinical drug interaction studies using in vitro data and in vivo simulation. *Drug Metabolism and Disposition* **38**, 981–987 (2010).

[206] Waller, D., Renwick, A., Gruchy, B. & George, C. The first pass metabolism of nifedipine in man. *British Journal of Clinical Pharmacology* **18**, 951–954 (1984).

[207] Raemsch, K. D. & Sommer, J. Pharmacokinetics and metabolism of nifedipine. *Hypertension* **5**, II18 (1983).

[208] Wishart, D. S. *et al.* DrugBank 5.0: A major update to the drugbank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2018).

Box 1: Explainable AI (XAI) applied to cytochrome P450 metabolism.

This worked example showcases XAI that provides a graphical explanation in terms of molecular motifs that are considered relevant by a neural network model predicting drug interaction with cytochrome P450 (3A4 isoform, CYP3A4). The *integrated gradients feature attribution* method[70] was combined with a graph-convolutional neural network for predicting drug-CYP3A4 interaction. This network model was trained with a publicly available set of CYP3A4 substrates and inhibitors.[203] The figure shows the results obtained for two drugs that are metabolized predominantly by CYP3A4, namely the phosphodiesterase A inhibitor (antiplatelet agent) cilostazol,[204] and nifedipine,[205] an L-type calcium channel blocker.
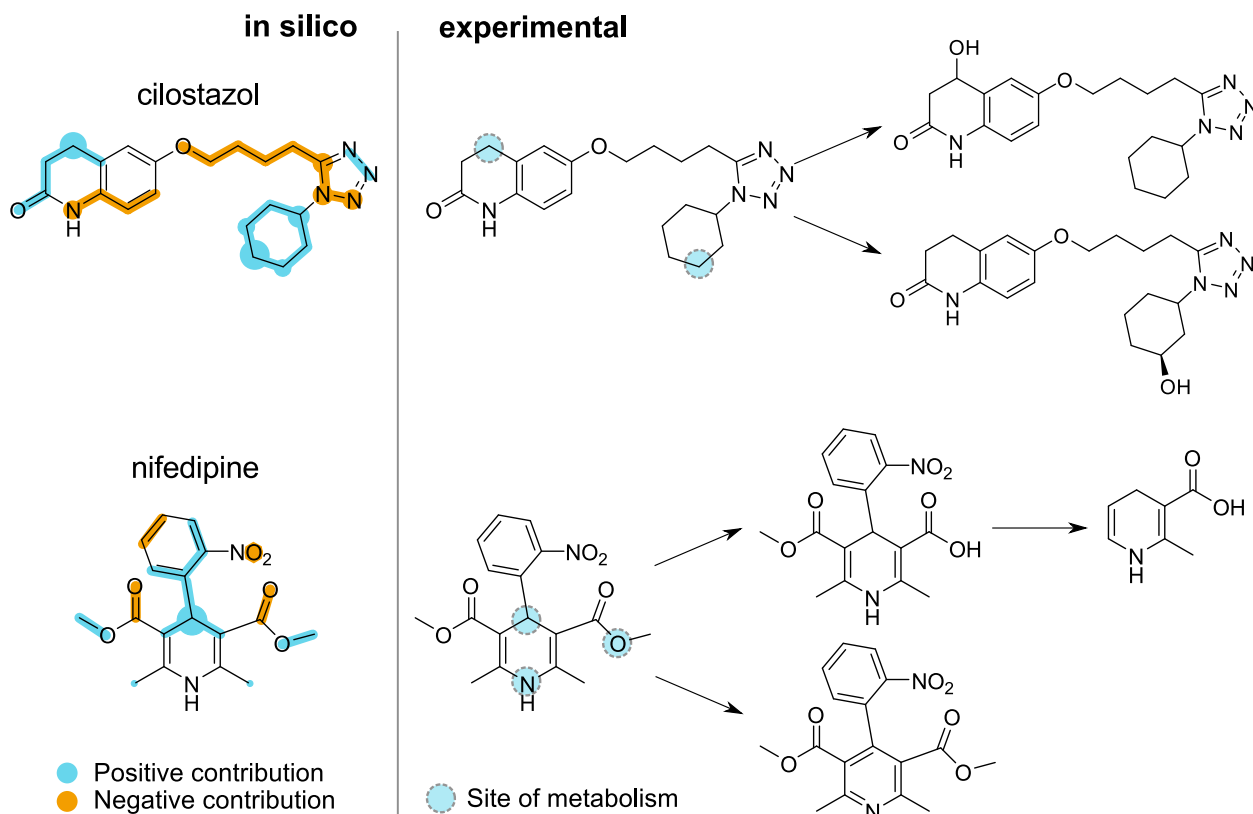
The structural features for CYP3A4-compound interaction suggested by the XAI method are highlighted in color (*left panel* 'in silico': blue, positive contribution to interaction with CYP3A4; orange, negative contribution to interaction; spot size indicates the feature relevance of the respective atom). The main sites of metabolism (dashed circles) and the known metabolites[204, 206–208] are shown in the right panel ('experimental'). Apparently, the XAI captured the chemical substructures involved in CYP3A4-mediated biotransformation and most of the known sites of metabolism. Additional generic features related to metabolism were identified, *i.e.*, (i) the tetrazole moiety and the secondary amino group in cilostazol, which are known to increase metabolic stability (*left panel*: orange, negative contribution to the CYP3A4 -cilostazol interaction), and (ii) metabolically labile groups, such as methyl and ester groups (*left panel*: blue, positive contribution to the CYP3A4-nifedipine interaction).