Coloring Molecules with Explainable Artificial Intelligence for Preclinical Relevance Assessment

Coloring molecules with explainable artificial intelligence for preclinical relevance assessment

José Jiménez-Luna,[*,†] Miha Skalic,[‡] Nils Weskamp,[‡] and Gisbert Schneider[*,†]

[†]*Department of Chemistry and Applied Biosciences, RETHINK, ETH Zurich, 8049 Zurich, Switzerland*

[‡]*Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Straße 65, 88397 Biberach an der Riss, Germany*

E-mail: jose.jimenez@rethink.ethz.ch; gisbert@ethz.ch

Abstract

Graph neural networks are able to solve certain drug discovery tasks such as molecular property prediction and *de novo* molecule generation. However, these models are considered 'black-box' and 'hard-to-debug'. This study aimed to improve modeling transparency for rational molecular design by applying the integrated gradients explainable artificial intelligence (XAI) approach for graph neural network models. Models were trained for predicting plasma protein binding, hERG channel inhibition, passive permeability, and cytochrome P450 inhibition. The proposed methodology highlighted molecular features and structural elements that are in agreement with known pharmacophore motifs, correctly identified property cliffs, and provided insights into unspecific ligand-target interactions. The developed XAI approach is fully open-sourced and can be used by practitioners to train new models on other clinically-relevant endpoints.

Introduction

Medicinal chemists have to solve multidimensional optimization problems, that is, the simultaneous optimization of several different compound parameters.[1] Successful drug candidates should not only possess sufficient activity towards a certain target protein or pathway but also suitable overall absorption, distribution, metabolism, and excretion (ADME) properties while holding an acceptable safety profile. Quantitative structure-property relationship (QSPR) approaches[2] have been extensively used to close the gap between *in silico* experiments and more cost- and time-intensive *in vitro* data.[3,4] Currently, deep-learning approaches are among the most popular machine-learning QSPR methodologies, as these have proven useful for improved ligand-[5,6] and structure-based property prediction,[7] target identification,[8,9] de novo molecule generation,[10,11] and chemical synthesis planning,[12] to name some of its most prominent applications.

Among these learning algorithms, message-passing neural networks, commonly referred to as graph neural networks,[13] have shown good capabilities in ligand-based molecular property prediction[14] despite their increased computational cost.[15] Since one of the advantages of deep-learning approaches against more classical machine-learning methods, is their ability to approximate highly non-linear functions from representations that are closer to the data source,[16] graph neural networks have the potential of replacing decades-old hand-crafted molecular fingerprint representations.[17] Despite their promise, the practical utility and acceptance of graph neural network models in drug discovery is limited owing to their lack of interpretability regarding the established chemical language.[18] While previous efforts have been made to mitigate these issues,[19,20] this limitation is further exacerbated by the fact that deep neural networks are notorious for producing correct answers for the wrong reasons (*i.e.*, the Clever Hans effect),[21] and for making overly confident erroneous predictions.[22] 'Explainable' artificial intelligence (XAI) aims to overcome some of these limitations by rendering the decision-making process of machine learning methods more transparent for the human mind.[23,24]

In the context of drug discovery-related applications, in particular for property prediction tasks, XAI methods can potentially help rationalize deep, as well as classical machine-learning models by highlighting molecular substructures that are critical for a given prediction.[25–31] An alternative paradigm is to develop models which are inherently interpretable, although it is debatable whether there is a trade-off with predictive performance.[32,33] Analysis of the physicochemical properties of compounds can provide a complementary perspective. Several studies have examined the influence of such 'global' properties on drug-likeness estimations and other aspects of chemical compounds.[34–36] Herein, an established structure- and property-based XAI approach, the integrated gradients feature attribution technique,[19] was used to examine its practical utility for a number of ADME and safety-related endpoints. Towards that goal, and building upon previous related work,[20,25] we propose several complementary assessments of model interpretations that leverage known structure-property relationships, property cliffs, as well as unspecific molecular interactions. Additionally, to the best of our knowledge, we provide the first open-source implementation of this XAI approach in combination with message-passing neural networks in the context of chemical property prediction. We furthermore make available all trained models and evaluation code, so that other researchers reproduce the results shown, test on novel examples, and adapt the proposed XAI approach to their own message-passing models.

Table 1: Data sets used for each pharmacological endpoint considered.

Endpoint	No. compounds	Task	References
Plasma protein binding	4,119	Regression	41–46
Caco-2 passive permeability	239	Regression	47,48
hERG inhibition	6,993	Regression	49
P450 inhibition	9,120	Binary classification	50,51

Data sets

Four pharmacologically relevant parameters – plasma protein binding (PPB),[37] human ether-a-go-go-related gene (hERG) potassium channel inhibition,[38] passive drug permeability (Caco-2 assay),[39] and cytochrome P450 inhibition (CYP3A4 isoform) – were evaluated.[40] To ensure that prospective users could explore the applicability of the proposed XAI approach and make use of the trained models, a literature survey was conducted to collect publicly available data on these four endpoints (Table 1, Figure 1).

Plasma protein binding

The capacity of a compound to bind to serum proteins, such as albumin and alpha-1-acid glycoprotein, critically affects its pharmacokinetic and pharmacodynamic profile and the disposition of the drug (*e.g.*, bioavailability, distribution, and clearance).[52] High-affinity compounds for these targets may, in practice, require higher dosing to achieve effective concentrations in patients.[53] In the present study, data from six different studies,[41–46] comprising 4,119 drugs, were combined in order to construct a training set for predicting the fraction bound (f_b) in plasma.

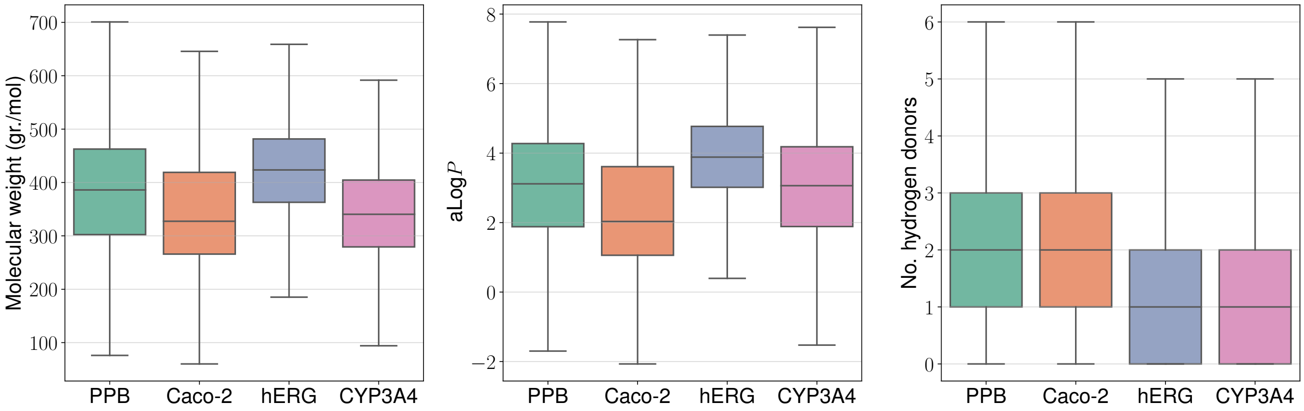


Figure 1: Box-whiskers plots concerning the distributions of molecular weight, calculated log$P$ (aLog$P$) values, and the number of hydrogen donors. Caco-2, passive permeability; CYP3A4, cytochrome P450 inhibition; hERG, human ether-a-go-go cardiac potassium channel inhibition; PPB, plasma protein binding.

Caco-2 cell passive permeability

Drugs administered orally must cross cell membranes to perform their function.[54] Such performance can be determined in vivo with radiolabeled compounds,[55] whereas the Caco-2 cell line is considered the in vitro gold standard proxy for studying pharmaceutical drug transport across cellular barriers.[56] For this endpoint, passive permeability data from 239 compounds was collected from two independent studies.[47,48] Passive permeability values (P_{app}) were collected (in cm s^{-1}) and converted to the $\log_{10}$ scale for numerical stability during model training.

hERG potassium channel inhibition

hERG inhibition is associated with the prolongation of the cardiac QT interval, which may lead to cardiac conditions such as arrhythmia.[57,58] For this endpoint, data compiled by Sato *et al.* was used,[49] among which 6,993 compounds with reported activity (IC_{50} values) in the nanomolar range were selected. IC_{50} values were transformed into the $p\text{IC}_{50}$ scale for numerical stability during model training.

Cytochrome P450 inhibition

The family of metabolic cytochrome P450 enzymes are relevant for drug clearance and the oxidation of xenobiotics, steroids, fatty acids, as well as for hormone synthesis.[59] For this endpoint, data compiled by Nembri *et al.* was used,[50] encompassing 9,120 CYP3A4 inhibitors and substrates with binary activity information (active/inactive), as determined by Veith *et al.*[51] via high-throughput screening with a bioluminescent assay. Substrate data – while known to potentially pose several modeling problems from a pharmacological point of view[60] – was not explicitly labeled and hence could not be removed.

Molecules from different sources were converted to InChI strings and sanitized using RDKit.[61] For all previous endpoints, if several measurements were available for the same compound, we considered their arithmetic average as the target value to predict for simplicity. Since the use of experimental data from various sources can lead to an increase in noise when modeling,[62–64] we provide the overlap percentage of compounds between data sources, as well as the mean and median standard deviation between different measurements in Table S1. While there is a high degree of compounds present in different sources, the overall agreement between reported measurements also appears to be high (*e.g.* 0.27 or 0.1 median standard deviations for the Caco-2 and the hERG endpoints, respectively, in log units).

Methods

Message-passing neural networks

Message-passing neural networks (MPNNs) belong to the family of graph convolutional neural networks (GCNs). In this context, a molecule is considered a graph $\mathcal{G}$ with a set of vertices and edges $\mathcal{G} = (V, E)$, representing the atoms $v \in V$ and bonds $e \in E$ of a two-dimensional molecular graph. The general MPNN framework assumes that both the vertices and edges are characterized by feature vectors $x_v \in \mathbb{R}^{d_1}$ and $w_e \in \mathbb{R}^{d_2}$, respectively. Message passing is performed iteratively across each pair of vertices u, v according to the following equations:

$$m_e^{(t+1)} = \phi\left(x_v^{(t)}, x_u^{(t)}, w_e^{(t)}\right), \tag{1}$$

$$x_v^{(t+1)} = \psi\left(x_v^{(t)}, \rho\right), \tag{2}$$

for $(u, v, e) \in \mathcal{G}$. Here, ϕ is a message function that is defined on each edge and combines its features with those of its neighboring nodes. ψ is an update function, which updates the node features by aggregating the information of the neighboring messages m_e using a reduction function ρ. The outcome of performing these iterative message passing steps, until a prespecified maximum number of iterations, is to generate a vector representing the entire molecular graph, that may subsequently be passed through additional fully-connected layers of the network to generate predictions. The different combinations of message, update, and reduction functions result in different MPNN architectures. The message and update functions contain weights that are learnable by backpropagation. In the present study, the MPNN architecture proposed by Gilmer *et al.*.[13] was applied, which combines a graph convolutional network and a Set2Set submodel[65] to embed molecules and compute a prediction. This model and other MPNN variations were shown to perform well on several ligand-based tasks.[14]

Table 2: Vertex, bond, and 'global' molecular graph features computed with RDKit[61]

Description level	Features
Atom	atom type, chirality, valence, formal charge, hybridization, bond degree, presence in ring, aromaticity, number of hydrogens, number of radical electrons, atomic mass, van der Waals radius
Bond	bond type, bond stereo, conjugation, presence in ring
Global	molecular weight, calculated octanol-water partition coefficient (aLog$P$), topological polar surface area (TPSA), number of hydrogen-bond donors

Furthermore, to account for unspecific molecular interactions, a fully-connected neural network sub-architecture was also included for the consideration of computed physicochemical features $x \in \mathbb{R}^{d_3}$. A visual representation of how the different kinds of information were used to generate the predictions, including latent graph vectors generated via message passing and the vector of calculated global molecular properties, can be found in Figure 2. Additional details presented in this figure are explained in further sections. Selected vertex, bond, and global features were computed with the RDKit software (Table 2).[61] Details on the network architecture and hyperparameter selection are fully disclosed in the associated code repository.

Model training

A $k = 10$ cross-validation scheme was used to estimate the model performance. The compounds were randomly shuffled and each model was trained on $k - 1$ non-overlapping subsets, and evaluated on the remaining one, for a total of k repetitions. While a random cross-validation scheme is known to produce overly-optimistic results, when compared to scaffold or time-based splits, here we only sanity-check the satisfactory training of the underlying models, as our goal here is to explore whether modern message-passing neural networks can provide meaningful explanations, rather than evaluating their predictive performance under more exhaustive scenarios. We trained models on each

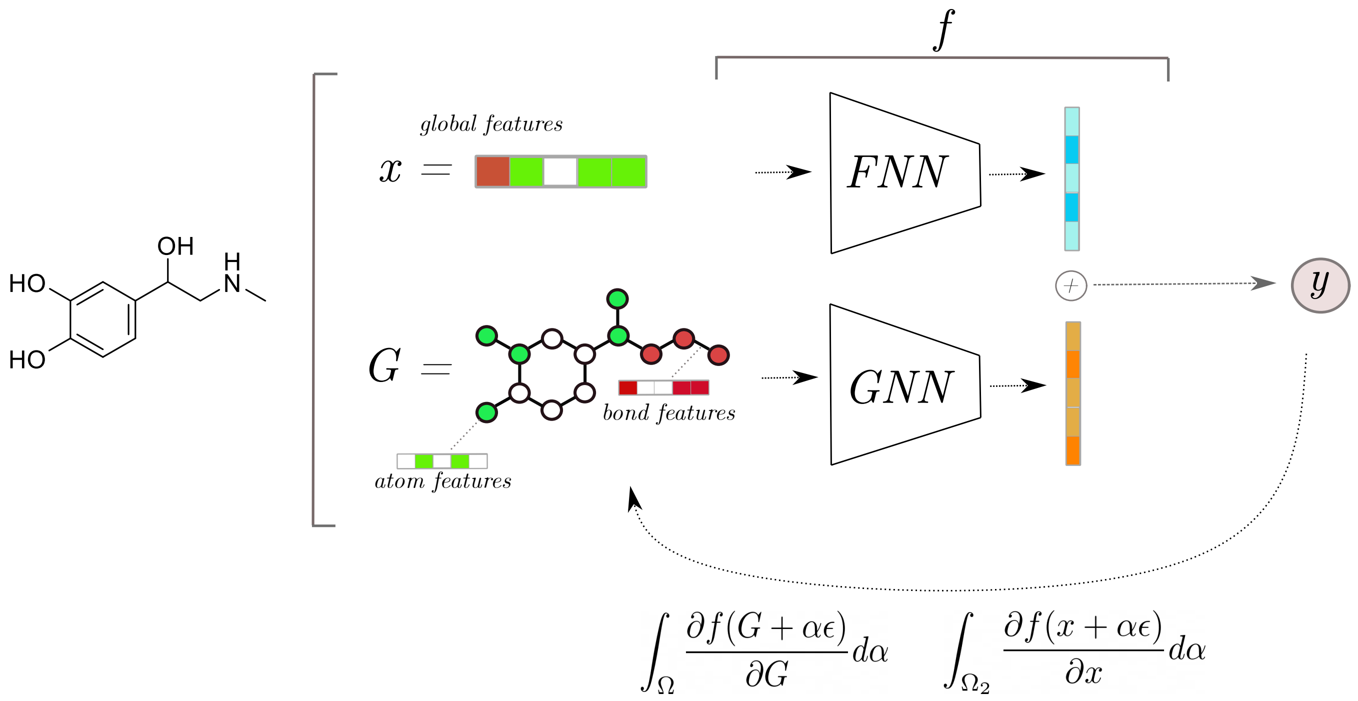

f
global features
$x =$
FNN
OH
HO
H
N
HO
$G =$
GNN
bond features
atom features
y
$\int_{\Omega} \frac{\partial f(G+\alpha\epsilon)}{\partial G} d\alpha$
$\int_{\Omega_2} \frac{\partial f(x+\alpha\epsilon)}{\partial x} d\alpha$

Figure 2: Schematic of the XAI methodology and neural network architecture. A message-passing graph neural network (GNN) and a forward fully-connected neural network (FNN) were combined to process an input presented as a molecular graph with atom, bond, and computed global properties (*e.g.*, octanol-water partition coefficient and topological polar surface area). The integrated gradients method[19] was then applied to compute atom, bond, and global importance scores.

data split for 250 epochs, with a batch size of 32 samples, and employed the Adam stochastic optimizer[66] with default momentum parameters ($\beta_1 = 0.9, \beta_2 = 0.999$) and a starting learning rate of 10^{-4}. Regression and binary classification tasks were optimized using a mean-squared and binary cross-entropy criterion, respectively.

Feature attribution

The MPNN model can be denoted as a function that maps tuples of graphs and global features to arbitrary target values $f : (\mathcal{G}, \mathcal{X}) \rightarrow \mathcal{Y}$. Given this notation, a feature attribution approach for graphs can be defined as a function that, using a trained MPNN model, takes a graph with featured vertices and edges, as well as a set of global features, and produces an importance score $\mathcal{E} : (\mathcal{G}, \mathcal{X}) \rightarrow c_v, b_{u,v}, z$, for each $u, v \in \mathcal{G}$, and $z \in \mathcal{X}$ (*i.e.* assigns importance scores to atoms and calculated global molecular properties in the current context.). This process can be performed by gradient backpropagation to the input features of the nodes, edges, and global features:[67,68] $\left(i.e. \ \frac{\partial f}{\partial x_v}, \frac{\partial f}{\partial w_e}, \frac{\partial f}{\partial x} \right)$.

In practice, however, this approach has several limitations, such as gradient saturation.[69] It also ignores two desirable aspects, namely model sensitivity and implementation invariance. Sensitivity refers to the fact that if two models had different predictions but differed on a single feature, then this feature should be assigned a non-zero attribution, while invariance ensures that two functionally identical models produce the same attributions. As previously discussed by McCloskey *et al.*,[20] the integrated gradients method[19] was herein employed to address these issues. This approach aggregates the gradient of the output with respect to the node features that fall on the straight line between user-defined baselines x'_v and the input x_v as follows:

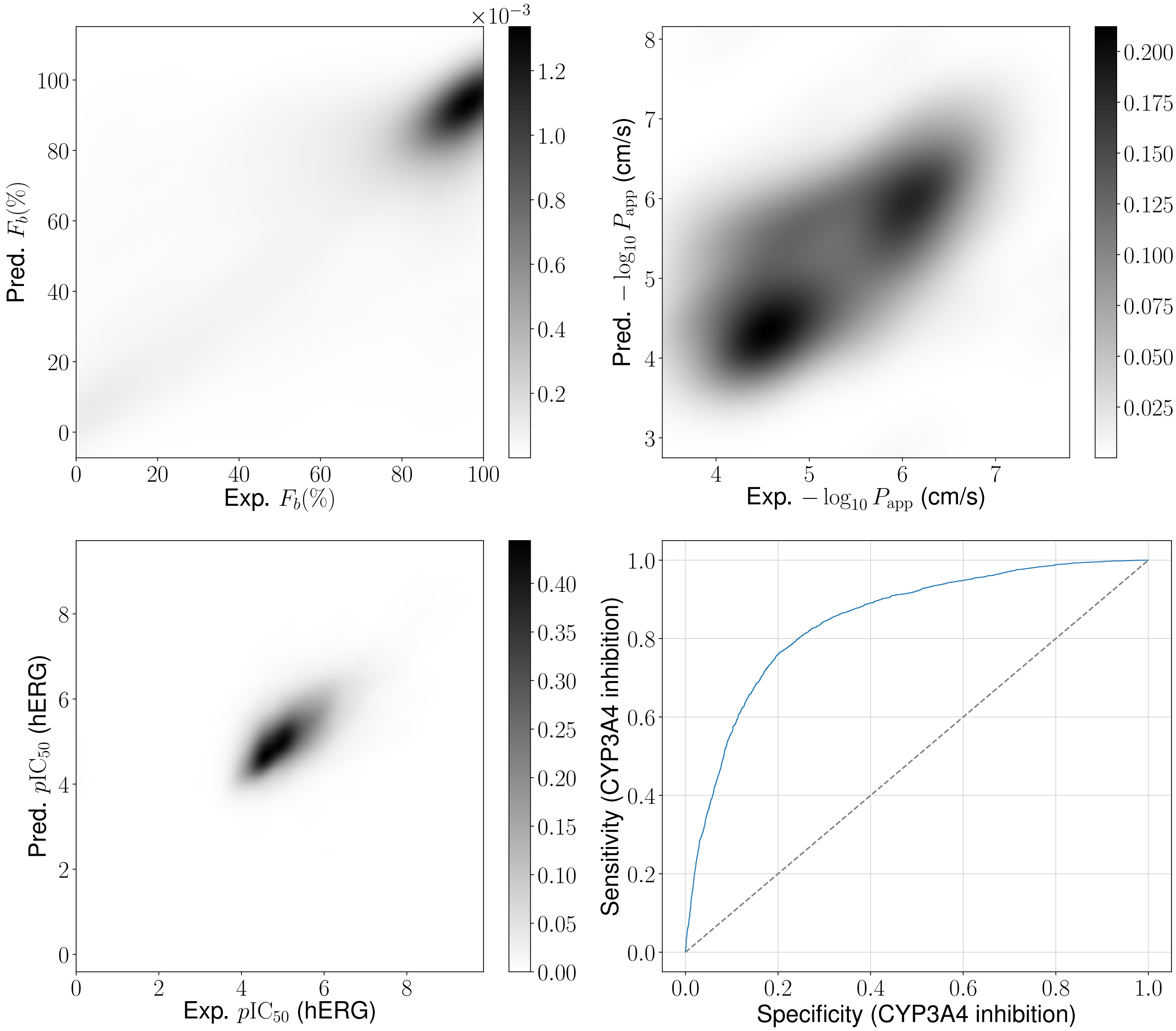


Figure 3: Model performance. A $k = 10$ random cross-validation scheme was used. Two-dimensional density plots (in lieu of scatterplots, due to large dataset sizes and the concentration of a large number of experimental values in small numeric intervals) portraying experimental vs. predicted values for the plasma protein binding, passive permeability, and hERG inhibition data sets. For the CYP 3A4 P450 data set, a receiver operating characteristic (ROC) curve is reported given its binary activity label (active/inactive).

$$\mathrm{IG}(x_v) = (x_v - x'_v) \int_{\Omega} \frac{\partial f\left(x'_v + \alpha\left(x_v - x'_v\right)\right)}{\partial x_v} d\alpha, \tag{3}$$

where α is used to integrate over the continuous points (in $\mathbb{R}^{d_1}$) in the path between the baseline and the input, and Ω is the input domain of integration. Because the integral in Equation 3 is non-tractable; it was computed with a Riemann approximation in accordance with:

$$\mathrm{IG}(x_v) \approx \frac{(x_v - x'_v)}{m} \sum_{r=1}^{m} \frac{\partial f\left(x'_v + \frac{r}{m}\left(x_v - x'_v\right)\right)}{\partial x_v}, \tag{4}$$

where r is the approximation variable. Equations 3 and 4 can be subsequently applied in the same manner to edge features w_e, and global input features x. Equation 4 was iterated over

$m = 50$ steps, and utilized baselines corresponding to zeroed-out vertex, edge, and global feature tensors. In order to obtain a single score for atoms and bonds, the gradients associated with each atomic and bond feature are mean-pooled. For visualization, computed edge importance values $b_{u,v}$ were evenly distributed among the importances c_v of their connecting vertices:

$$c'_v = c_v + \sum_{i \in \mathcal{N}(v)} b_{i,v}/2, \qquad (5)$$

where $\mathcal{N}(v)$ is the set of neighboring vertices at one bond distance from vertex v. Since the score assigned to each vertex is a monotonic function of the output gradients, in a regression setting a positive value marks a vertex that is contributing towards an increasing output value and vice versa. As depicted in Figure 2, each atom position (vertex) was represented with its assigned color depending on the sign of the respective importance value (green and red colors indicate a positive and negative contribution, respectively), and with a radius proportional to the magnitude of the importance value. This visualization style contrasts with that of previous approaches,[25] which took score magnitude into account in the form of color shading. Atomic importances below a user-defined value of 10^{-4} were not considered. Bonds (edges) were colored according to whether the color of their connecting nodes matched.

We note that in the present work we mainly deal with either regression or binary classification tasks, yielding the interpretation of the resulting atomic contribution score straightforward (*i.e.* since the value is a monotonic function of the gradient, a positive value indicates regions of increasing target value, either on the real line or towards the positive class, and vice versa). However, special care needs to be taken when modeling multi-class tasks, as the sign of the score only bears meaning towards a specific class. Interpretation can be particularly problematic in datasets where a real-valued target variable is binned into several non-overlapping categories.

Table 3: Predictive performance of the $k = 10$ cross-validation scheme for the endpoints considered. Pearson's correlation coefficient R, coefficient of determination R^2, and RMSE ±1 standard deviation) between experimental and predicted values are reported for the regression models; AUC (±1 standard deviation) for the classifier model.

Endpoint	**Pearson's R**	R^2	**RMSE**	**AUC**
Plasma protein binding	0.74 ± 0.03	0.523 ± 0.05	20.79 ± 0.9	-
Passive permeability	0.53 ± 0.1	0.26 ± 0.23	0.87 ± 0.09	-
hERG inhibition	0.63 ± 0.03	0.29 ± 0.03	0.76 ± 0.03	
P450 inhibition	-	-	-	0.85 ± 0.01

RMSE, root mean square error; AUC, area under receiver-operator characteristic curve. RMSE values reported in percentage units for the fraction bound F_b in the plasma protein binding dataset, in $\log_{10}$ P_{app} units for the passive permeability dataset and in $p\text{IC}_{50}$ units for the hERG inhibition dataset.

Assessment of model interpretations

To enable methodology evaluation, 25 external molecular series were extracted and compiled from available literature (provided in Supporting Data and colored in the accompanying code repository of this work). These series represent background knowledge and contain examples that are known to be relevant for the pharmacological endpoints considered in this study, most of which were external to the used training sets. Furthermore, a range of different approaches were considered in order to check if the models (i) were able to highlight relevant pharmacophore motifs, (ii) successfully detected property cliffs in the considered data sets (*i.e.*, small structural changes that result in a marked property or activity change[70]), and (iii) were able to identify 'unspecific' ligand-protein interactions mediated by molecular properties (*e.g.*, log$P$, TPSA). In contrast to previous research, which presented validations for explainable machine-learning models either by assessing the additivity of atomic contributions,[25] or by quantifying the quality of the provided molecular colorings by comparing them with synthetically-generated structure-activity relationships,[20] the presented work provides complementary information towards the development of increasingly-objective evaluation frameworks for interpretable artificial-intelligence in QSAR/QSPR tasks.

Results and discussion

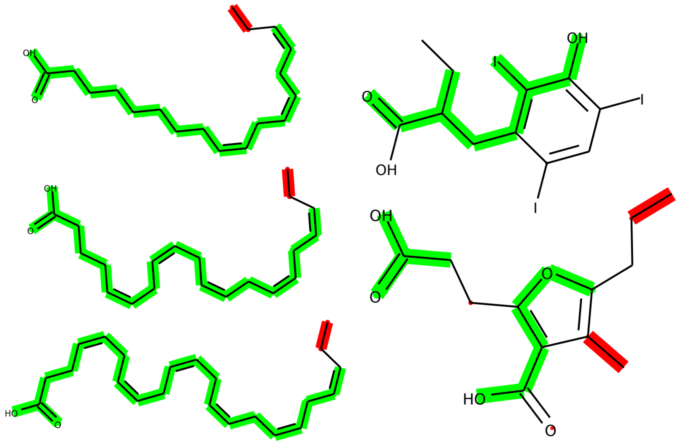
Model performance

While the main goal of the study is not to evaluate the predictive performance of graph neural networks compared to other machine-learning models, in order to assess whether the proposed feature attribution approach was able to extract meaningful relationships between structural motifs and the respective pharmacological endpoints, a rigorous evaluation was mandatory since explanations generated by a model with limited predictive capability bear little trust. Results of a quantitative benchmark are presented in Figure 3 and Table 3, where the root mean squared error (RMSE), Pearson's correlation, and determination coefficients between experimental and predicted values, and the receiver-operator characteristic area under the curve (AUC) are reported. In the hopes of avoiding unnecessary model selection bias,[71] no explicit hyperparameter optimization was performed in any of the training folds. These were chosen mainly taking into account sufficient network capacity and reasonable computational cost on commodity hardware, and can be checked on the accompanying code repository of this work.

All trained models showed predictive capabilities, with R values ranging between 0.53 and 0.74 for the three regression models, and AUC = 0.85 for the binary classifier. These values suggest that the training tasks varied in difficulty. Although none of the models exhibited perfect predictive capabilities, the results obtained were markedly better than random, suggesting that meaningful molecular graph features were identified in the learning process.

Pharmacophore motif recognition

Two relevant features were analyzed to assess plasma protein binding potential, namely fatty acid character[75] and a pharmacophore motif[44] consisting of two acidic groups separated by a hydrophobic part of five bond units (Figure 4).



(a) (b)

Figure 4: Recognized motifs from the plasma protein binding data set. (a) Fatty acids; (b) Iophexonate and 3-carboxy-4-methyl-5-propyl2-furanpropionicacid (CMPF). The latter compounds feature two acidic groups separated by a hydrophobic part of five bond units[44] (considering phenol as a weak acid,[72] 2,4,6-triiodophenol moiety pK_a = 5.97 as computed with MoKa 3.2.2[73]), which are partially highlighted. Green and red areas represent structural positive and negative contributions, respectively, w.r.t. the ligand fraction bound f_b

For the hERG endpoint, two cases are shown in which the XAI was able to reproduce activity changes that were previously reported in the literature. Figure 5a highlights the effect of a negatively ionizable substructure, such as a carboxylate group, which abolished the activity of the compound.[76] This effect could be explained by the fact that the ligand-accommodating cavity of the hERG potassium channel stabilizes positive charges. In this case, while the experimental activity difference spanned over 2 orders of magnitude, the underlying model was unable to fully capture this range, illustrating the potential of the model to debug less-than-ideal cases. The second example illustrates the introduction of an activity cliff by another replacement[77] (Figure 5b). However, these examples also highlight potential limita-

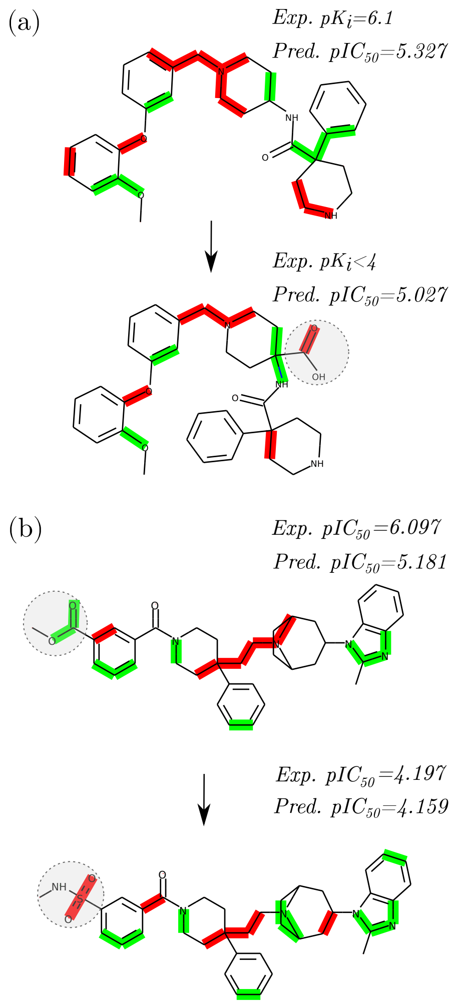


Figure 5: Examples of motifs indicating hERG inhibition. (a) Addition of a negative charge and (b) other replacements resulting in activity cliffs. Green and red colors represent structural positive and negative contributions towards hERG inhibition, respectively

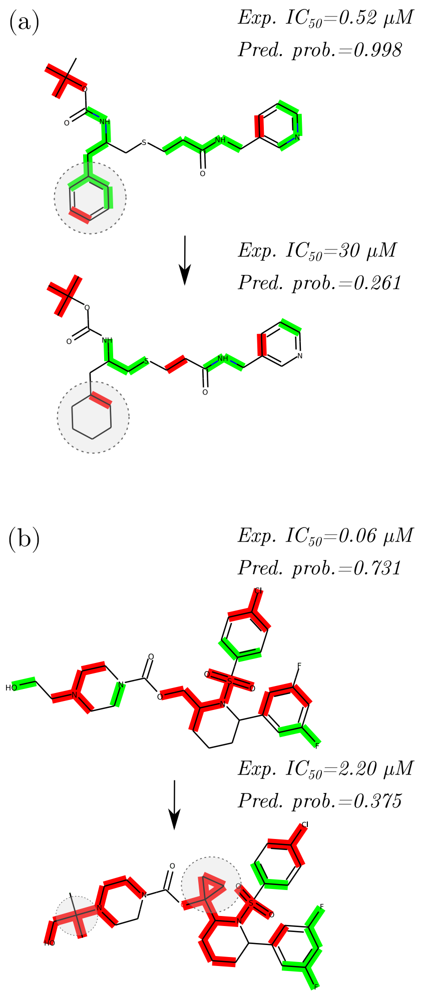


Figure 6: Cytochrome (CYP) inhibition motif replication examples. (a) Structure-based pharmacophore developed by Kaur *et al.*[74] (b) Decrease in activity caused by the addition of two extra methyl groups. Green and red areas represent structural positive and negative contributions, respectively, towards CYP3A4 inhibition.

tions of the proposed methodology. In particular other highlighted patterns are at odds with established hERG structure-activity relationships.[78,79] For instance, in the first example, the tertiary amine is suggested to contribute negatively to activity, as well as a carbonyl group and an arbitrary aliphatic ring. Additionally the approach does not fully highlight the acidic moeity negatively. On a similar note, in the second example, the tertiary basic amine and the tertiary amide were assigned colors incorrectly upon replacement. Further coloring examples for hERG, such as the effect of other bioisosteric replacements, changes in amine-nitrogen environments, and topological polar surface area differences are available in the accompanying code repository.

For the CYP3A4 endpoint, the respective model clearly identified motifs of a previously reported specific pharmacophore,[74] highlighting the importance of a flexible backbone, hydrogen-bond donor/acceptor moieties, and hydrophobic interactions (Figure 6a). The addition of two methyl groups was previously reported as a strategy for mitigating the CYP3A4 activity of morpholine-based N-arylsulfonamide γ-secretase inhibitors.[80] Of note, the relative importance of the corresponding structural features was correctly recognized (Figure 6b). Interestingly, lowering the overall molecular weight was also reported as a successful strategy towards decreasing CYP3A4 activity.[80] In contrast, for this particular example, the assigned global molecular weight importance is markedly positive, a fact that is in line with the empirical correlation between molecular weight and activity for this endpoint (see Global importance analysis section and Table S2). Additional examples[81–83] are provided in the accompanying code repository of this work.

Property cliff identification

To further evaluate the capabilities of the models to recognize property cliffs beyond the selected literature examples, it was evaluated whether activity cliffs exist in the training sets via a matched molecular pairs analysis.[84] The cliffs were ranked according to the structure activity landscape index (SALI).[85] This functional balances the structural similarity of a pair of compounds with their predicted property difference:

$$\mathrm{SALI}\left(\mathrm{mol}_i, \mathrm{mol}_j\right) = \frac{|p_i - p_j|}{\mathrm{sim}\left(\mathrm{mol}_i, \mathrm{mol}_j\right)}, \qquad (6)$$

where p_i, p_j are the properties of interest of molecules mol_i and mol_j, respectively, and sim is a molecular similarity function. Coloring examples with their respective SALI ranks for the endpoints considered in this study, using the entire sets as training data, as well as out-of-fold models, are presented in Figures 7 and S1, respectively. Others can be computed via the accompanying code repository of this work. It is noteworthy that the proposed approach correctly identified several structural elements that are responsible for these striking property differences, either by highlighting a positive contribution for the active molecule in the pair, when a certain structural feature is removed upon moving to the inactive molecule in the pair, or a negative contribution for the inactive molecule when the feature is added. However, similar to the discussion in the previous section, these analyses illustrate limitations of the approach: common substructural features between a similar pair of compounds could appear differently highlighted. Whether these common highlighted parts could potentially generate other property cliffs upon replacement, however, is difficult to determine in the absence of additional experimental information.

Global importance analysis

Many ADME and relevant toxicological endpoints, such as passive permeability or plasma protein binding parameters, are not solely characterized by specific structural motifs. In these cases, medicinal chemists are focused on investigating the influence of 'global' molecular properties (*e.g.* log$P$, TPSA) on the endpoint of interest to achieve optimal compounds. Plasma protein binding correlates positively with lipophilicity,[86] increasing circulation half-life, and reducing glomerular filtration. Our collected data set revealed a moderate positive correlation between aLog$P$ and the fraction bound ($R = 0.5, p < 0.01$, one-tailed Pearson's correlation test), which was confirmed by the importance assigned to the global aLog$P$ feature ($R = 0.55, p < 0.01$) by the XAI model.

P_{app}, as measured by the Caco-2 assay, is also known to correlate with global molecular properties, such as TPSA[87] (compounds with a large polar surface area are unlikely to permeate cell membranes) and lipophilicity[88] (compounds with a greater log$P$ permeate more easily). For the respective training data, we observed a moderate negative correlation between the computed TPSA and passive permeability ($R = -0.61, p < 0.01$), and a weak positive correlation with aLog$P$ ($R = 0.31, p < 0.01$). The first relationship was again correctly captured by the XAI approach, indicating a moderate negative correlation between the importance assigned to the TPSA global feature and the P_{app} endpoint ($R = -0.59, p < 0.01$). All correlations between assigned global importances and each respective endpoint can be checked on Table S2.

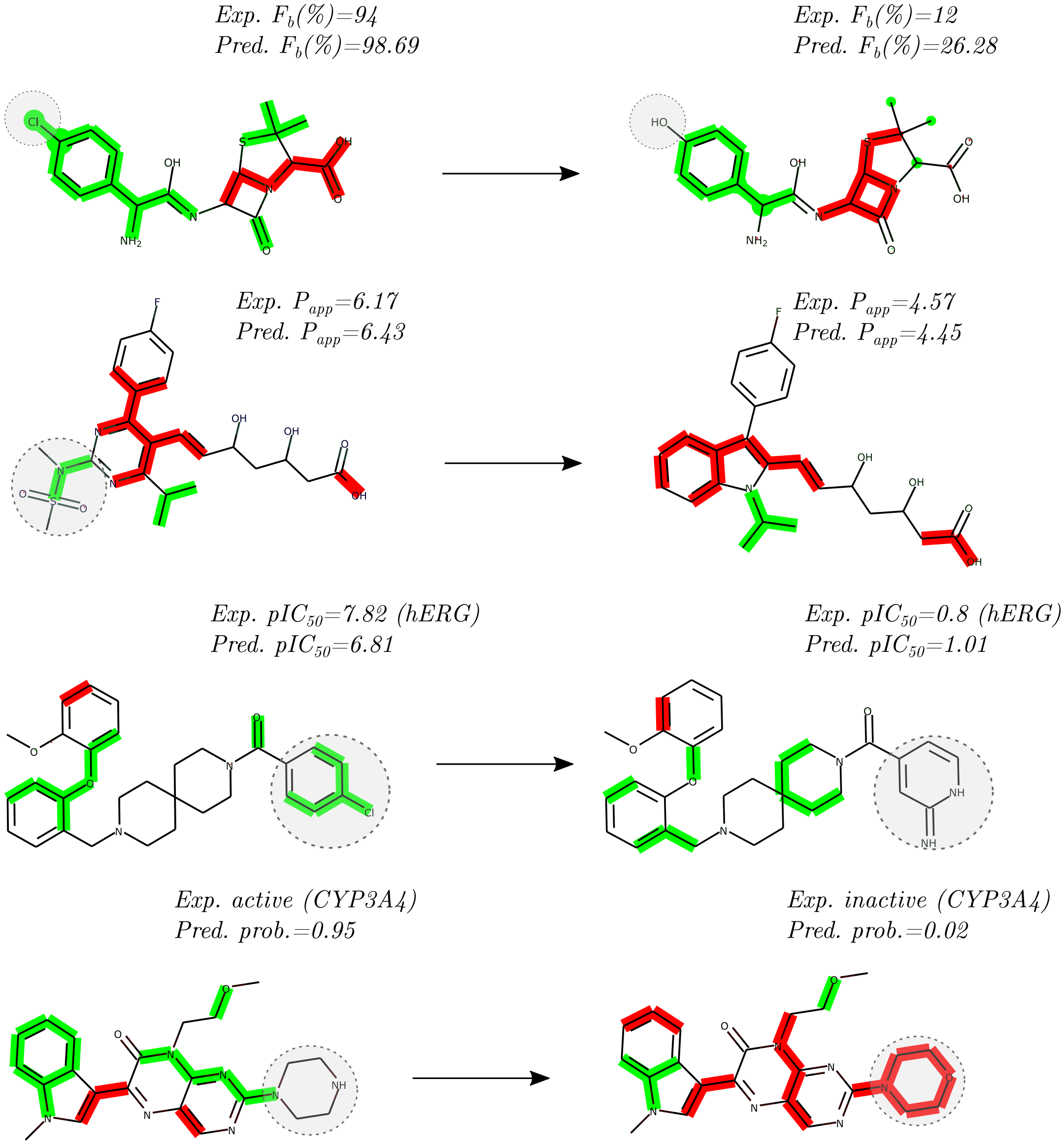


Figure 7: Some examples of the property cliffs identified by the proposed approach, selected via the SALI index (from top to bottom, ranks 78, 45, 4 and 105 out of all possible $n(n-1)/2$ pairs) for all the endpoints and data sets considered in this study. Green and red values represent positive and negative contributions, respectively, w.r.t. the considered endpoint.

Comparison to other coloring approaches

Lastly, the XAI approach herein proposed was compared to the molecular coloring method published by Sheridan,[26] which is model-agnostic and can be used for either regression or classification tasks. In order to highlight the importance of a particular atom, this approach iteratively 'masks' individual atoms and computes a molecular fingerprint. These fingerprints are then combined with a machine-learning model, and the difference between the model prediction with and without masked atoms serves as a proxy for atom importance. Figure 8 shows molecular structures for which the fingerprint-based model identified motifs corresponding to known pharmacophores of the hERG and CYP3A4 endpoints, using a well-known and robust industry standard, a random forest model with 1000 trees and ECFP4 fingerprinting featurization. Surprisingly, Figure 8a displays a bioisosteric ring substitution, involving the replacement of a pyrimidine with a thiazole, which in turn increases overall molecular weight and solubility, factors linked towards

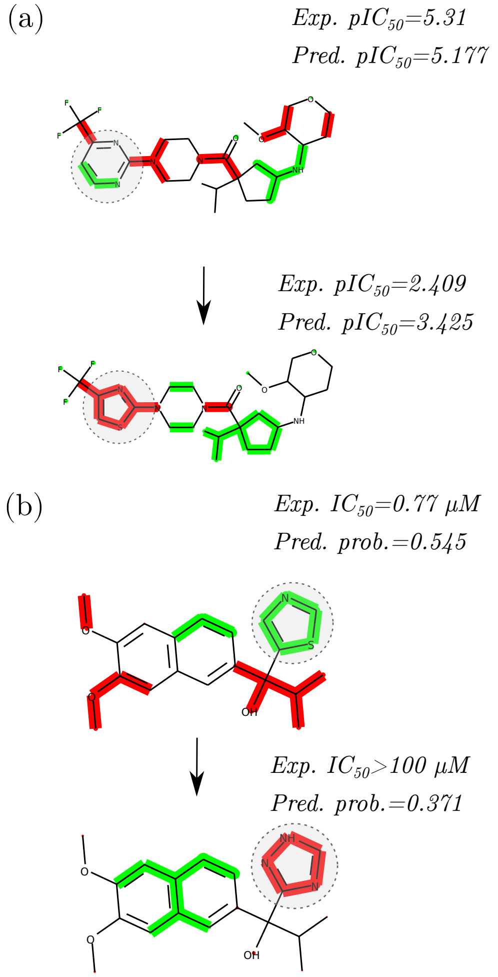


Figure 8: Examples using the approach of Sheridan[26] for the (a) hERG endpoint, involving a bioisosteric ring transformation, and for the (b) CYP3A4 endpoint, involving a heme-binding group substitution. Green and red colors represent positive and negative contributions, respectively, w.r.t. the considered endpoint.

increased hERG inhibition. Figure 8b features a heme-binding group replacement example. Interestingly, the gradient-based approach proposed in this work failed for these presented examples, whereas the fingerprint-based approach was unable to reproduce any of the other coloring examples presented in this study (Figures 4-7), suggesting that their appropriateness may be case-dependent. Further comparative examples are provided in the supporting code accompanying this article.

Given the lack of an established quantitative benchmark for atom coloring approaches in chemoinformatics, the superiority of either method remains to be determined. In particular, while the integrated gradients approach proposed here is well-grounded in theory, and fulfils several desirable feature attribution axioms, it requires a fully-differentiable model, such as the used message passing networks. On the other hand, the approach proposed by Sheridan, albeit simpler in nature, is model-agnostic. Additionally, we have observed limited agreement between the substructures highlighted by the two different methods, advocating the use of multiple models in parallel. With the aim of facilitating further evaluation, an implementation of the approach proposed by Sheridan, using a random forest model featured with ECFP4 fingerprints, using a bond radius of 2 units, is provided in the accompanying code repository of this work, together with trained models for all of the endpoints considered here. A similar approach, developed by Riniker and Landrum[31] is also available in the RDKit[61] software package.

Conclusion

Herein, we described the application of a popular XAI framework, the integrated gradients feature attribution technique, to four pharmacologically relevant ADME endpoints. The results show that the proposed approach correctly replicated motifs corresponding to known pharmacophore patterns, identified property cliffs, and detected non-specific ligand-receptor interactions mediated by global molecular properties. However, there are certain limitations to its applicability. First, the proposed methodology suffers from multi-collinearity, meaning that it is unable to correctly assign importance values to a pair of strongly-correlated molecular features. This issue is not exclusive to this particular methodology but is a limitation of many machine learning approaches.[89] Second, while we have shown that the described approach can successfully identify some structure-property relationships, either in the form of known motifs or property cliffs, several examples also exhibited some degree of attribution instability between closely-related pair of compounds, which debatably could be due to insufficient training data, or a direct consequence of the Clever Hans effect,[21] among other reasons. Third, this study would have benefited from a suitable XAI benchmark. Although several chemical series were provided to qualitatively evaluate the developed approach, the lack of suitable quantitative evaluation sets for XAI in chemistry and cheminformatics renders the evaluation of newly-developed approaches arduous. The first steps have been made in this direction in other research fields,[90,91] as well as chemoinformatics.[20,25,26] Nonetheless, further development of XAI applications in chemistry would greatly benefit from meaningful benchmarking, which will require close collaboration between medicinal chemists and computer scientists in order to prove their usefulness in prospective settings. Furthermore, when using the proposed approach, standard machine-learning and QSAR advice applies, as underlying models trained with larger and more chemically-diverse datasets will be more likely to produce better explanations. This is supported by the fact that, even with the larger sets considered in this study, such as those related to P450 and hERG inhibition, many literature-extracted motifs were not correctly captured, suggesting that either the underlying model did not learn the expected pattern or the feature attribution technique did not correctly capture what the model had learned. This also hints that current XAI approaches cannot yet be used as a recipe for compound optimization, requiring significant human expertise for correct interpretation. The reasons for these phenomena remain a topic of further study.

Implementation and code availability

The graph neural-network models were trained with the Deep Graph Library Python (DGL) package (version 0.4.3)[92] and the dgllife extension (github.com/awslabs/dgl-lifesci) that run on top of the PyTorch tensor manipulation library (version 1.4.0).[93] Molecular structures were handled using RDKit.[61] Users can retrieve the complete program code for replication of the experiments, training of new models, and molecular importance map generation from an AGPL-3 licensed repository on GitHub (github.com/josejimenezluna/molgrad). All models trained with publicly available data are also available.

All models reported in this work were trained using the Leonhard computer cluster at ETH Zurich, Switzerland.

Acknowledgement This work was financially supported by the ETH RETHINK initiative, the Swiss National Science Foundation (grant no. 205321_182176), and Boehringer Ingelheim Pharma GmbH & Co. KG.

Conflict of interest statement. G.S. is a cofounder of inSili.com LLC, Zurich, and a consultant to the pharmaceutical industry.

References

(1) Nicolaou, C. A.; Brown, N. Multi-objective Optimization Methods in Drug Design. *Drug Discov. Today: Technologies* **2013**, *10*, e427 – e435.

(2) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR Without Borders. *Chem. Soc. Rev* **2020**, *49*, 3525–3564.

(3) Gedeck, P.; Lewis, R. A. Exploiting QSAR Models in Lead Optimization. *Curr. Opin. Drug Discov. Devel.* **2008**, *11*, 569.

(4) Lewis, R. A. A General Method for Exploiting QSAR Models in Lead Optimization. *J. Med. Chem.* **2005**, *48*, 1638–1648.

(5) Lenselink, E. B.; Ten Dijke, N.; Bongers, B.; Papadatos, G.; Van Vlijmen, H. W.; Kowalczyk, W.; IJzerman, A. P.; Van Westen, G. J. Beyond

the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminformatics* **2017**, *9*, 1–14.

(6) Unterthiner, T.; Mayr, A.; Klambauer, G.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Hochreiter, S. Deep Learning as an Opportunity in Virtual Screening. Proceedings of the Deep Learning Workshop at Neural Information Processing Systems. 2014; pp 1–9.

(7) Jiménez-Luna, J.; Pérez-Benito, L.; Martínez-Rosell, G.; Sciabola, S.; Torella, R.; Tresadern, G.; De Fabritiis, G. DeltaDelta Neural Networks for Lead Optimization of Small Molecule Potency. *Chem. Sci.* **2019**, *10*, 10911–10918.

(8) Öztürk, H.; Özgür, A.; Ozkirimli, E. DeepDTA: Deep Drug–target Binding Affinity Prediction. *Bioinformatics* **2018**, *34*, i821–i829.

(9) Jimenez, J.; Sabbadin, D.; Cuzzolin, A.; Martinez-Rosell, G.; Gora, J.; Manchester, J.; Duca, J.; De Fabritiis, G. PathwayMap: Molecular Pathway Association with Self-normalizing Neural Networks. *J. Chem. Inf. Model.* **2018**, *59*, 1172–1181.

(10) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inform.* **2018**, *37*, 1700153.

(11) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. *arXiv preprint arXiv:1802.04364* **2018**,

(12) 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. *Chem. Sci.* **2018**, *9*, 6091–6098.

(13) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *arXiv preprint arXiv:1704.01212* **2017**,

(14) 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. *Chem. Sci.* **2018**, *9*, 513–530.

(15) Ching, T.; Himmelstein, D. S.; Beaulieu-Jones, B. K.; Kalinin, A. A.; Do, B. T.; Way, G. P.; Ferrero, E.; Agapow, P.-M.; Zietz, M.; Hoffman, M. M.; Xie, W.; Rosen, G. L.; Lengerich, B. J.; Israeli, J.; Lanchantin, J.; Woloszynek, S.; Carpenter, A. E.; Shrikumar, A.; Xu, J.; Cofer, E. M.; Lavender, C. A.; Turaga, S. C.; Alexandari, A. M.; Zhiyong, L.; Harris, D. J.; DeCaprio, D.; Qi, Y.; Kundaje, A.; Peng, Y.; Wiley, L. K.; Segler, M. H.; Boca, S. M.; Swamidass, S. J.; Huang, A.; Gitter, A.; Greene, C. S. Opportunities and Obstacles for Deep Learning in Biology and Medicine. *J. R. Soc. Interface* **2018**, *15*, 20170387.

(16) Barnard, A.; Motevalli, B.; Parker, A.; Fischer, J.; Feigl, C.; Opletal, G. Nanoinformatics, and the Big Challenges for the Science of Small Things. *Nanoscale* **2019**, *11*, 19190–19201.

(17) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput. Aided Mol. Des.* **2016**, *30*, 595–608.

(18) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug Discovery with Explainable Artificial Intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573–584.

(19) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. Proceedings of the 34th International Conference on Machine Learning, 70. 2017; pp 3319–3328.

(20) McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. J. Using Attribution to Decode Binding Mechanism in Neural Network Models for Chemistry. *Proc. Natl. Acad. Sci. U.S.A* **2019**, *116*, 11624–11629.

(21) Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.-R. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nat. Commun* **2019**, *10*, 1–8.

(22) Nguyen, A.; Yosinski, J.; Clune, J. Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015; pp 427–436.

(23) Lipton, Z. C. The Mythos of Model Interpretability. *Queue* **2018**, *16*, 31–57.

(24) 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**,

(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. *J. Chem. Inf. Model.* **2017**, *57*, 1773–1792.

(26) Sheridan, R. P. Interpretation of QSAR Models by Coloring Atoms According to Changes in Predicted Activity: How Robust Is It? *J. Chem. Inf. Model.* **2019**, *59*, 1324–1337.

(27) Rodríguez-Pérez, R.; Bajorath, J. Interpretation of Compound Activity Predictions from Complex Machine Learning Models Using Local Approximations and Shapley Values. *J. Med. Chem.* **2020**, *63*, 8761–8777.

(28) Polishchuk, P. G.; Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. Universal Approach for Structural Interpretation of QSAR/QSPR Models. *Mol. Inform.* **2013**, *32*, 843–853.

(29) Polishchuk, P. Interpretation of Quantitative Structure–activity Relationship Models: Past, Present, and Future. *J. Chem. Inf. Model.* **2017**, *57*, 2618–2639.

(30) 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. *J. Chem. Inf. Model.* **2016**, *56*, 1455–1469.

(31) Riniker, S.; Landrum, G. A. Similarity Maps-A Visualization Strategy for Molecular Fingerprints and Machine-learning Methods. *J. Cheminformatics* **2013**, *5*, 43.

(32) Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215.

(33) Barber, C.; Cayley, A.; Hanser, T.; Harding, A.; Heghes, C.; Vessey, J. D.; Werner, S.; Weiner, S. K.; Wichard, J.; Giddings, A.; Glowienke, S.; Parenty, A.; Brigo, A.; Spirkl, H.-P.; Amberg, A.; Kemper, R.; Greene, N. Evaluation of a Statistics-based Ames Mutagenicity QSAR Model and Interpretation of the Results Obtained. *Regul. Toxicol. Pharmacol.* **2016**, *76*, 7–20.

(34) Lipinski, C. A. Lead- and Drug-like Compounds: The Rule-of-five Revolution. *Drug Discov. Today: Technologies* **2004**, *1*, 337 – 341.

(35) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is There a Difference Between Leads and Drugs? A Historical Perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308–1315.

(36) Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. Moving Beyond Rules: The

Development of a Central Nervous System Multiparameter Optimization (CNS MPO) Approach to Enable Alignment of Druglike Properties. *ACS Chem. Neurosci* **2010**, *1*, 435–449.

(37) Olson, R. E.; Christ, D. D. *Annual Reports in Medicinal Chemistry*; Elsevier, 1996; Vol. 31; pp 327–336.

(38) Curran, M. E.; Splawski, I.; Timothy, K. W.; Vincen, G. M.; Green, E. D.; Keating, M. T. A Molecular Basis for Cardiac Arrhythmia: hERG Mutations Cause Long QT Syndrome. *Cell* **1995**, *80*, 795–803.

(39) Hidalgo, I. J.; Raub, T. J.; Borchardt, R. T. Characterization of the Human Colon Carcinoma Cell Line (Caco-2) as a Model System for Intestinal Epithelial Permeability. *Gastroenterology* **1989**, *96*, 736–749.

(40) Hashimoto, H.; Toide, K.; Kitamura, R.; Fujita, M.; Tagawa, S.; Itoh, S.; Kamataki, T. Gene Structure of CYP3A4, an Adult-specific Form of Cytochrome P450 in Human Livers, and Its Transcriptional Control. *Eur. J. Inorg. Chem.* **1993**, *218*, 585–595.

(41) Zhu, X.-W.; Sedykh, A.; Zhu, H.; Liu, S.-S.; Tropsha, A. The Use of Pseudo-equilibrium Constant Affords Improved QSAR Models of Human Plasma Protein Binding. *Pharm. Res.* **2013**, *30*, 1790–1798.

(42) Ingle, B. L.; Veber, B. C.; Nichols, J. W.; Tornero-Velez, R. Informing the Human Plasma Protein Binding of Environmental Chemicals by Machine Learning in the Pharmaceutical Space: Applicability Domain and Limits of Predictability. *J. Chem. Inf. Model.* **2016**, *56*, 2243–2252.

(43) Sun, L.; Yang, H.; Li, J.; Wang, T.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Compounds Binding to Human Plasma Proteins by QSAR Models. *ChemMedChem* **2018**, *13*, 572–581.

(44) Kratochwil, N. A.; Huber, W.; Müller, F.; Kansy, M.; Gerber, P. R. Predicting Plasma Protein Binding of Drugs: A New Approach. *Biochem. Pharmacol.* **2002**, *64*, 1355–1374.

(45) Votano, J. R.; Parham, M.; Hall, L. M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A. QSAR Modeling of Human Serum Protein Binding with Several Modeling Techniques Utilizing Structure- information Representation. *J. Med. Chem.* **2006**, *49*, 7169–7181.

(46) Watanabe, R.; Esaki, T.; Kawashima, H.; Natsume-Kitatani, Y.; Nagao, C.; Ohashi, R.; Mizuguchi, K. Predicting Fraction Unbound in Human Plasma from Chemical Structure: Improved Accuracy in the Low Value Ranges. *Mol. Pharm.* **2018**, *15*, 5302–5311.

(47) Bittermann, K.; Goss, K.-U. Predicting Apparent Passive Permeability of Caco-2 and MDCK Cell-monolayers: A Mechanistic Model. *PloS One* **2017**, *12*, e0190319.

(48) O'Hagan, S.; Kell, D. B. The Apparent Permeabilities of Caco-2 Cells to Marketed Drugs: Magnitude, and Independence from Both Biophysical Properties and Endogenite Similarities. *PeerJ* **2015**, *3*, e1405.

(49) Sato, T.; Yuki, H.; Ogura, K.; Honma, T. Construction of an Integrated Database for hERG Blocking Small Molecules. *PloS One* **2018**, *13*, e0199348.

(50) Nembri, S.; Grisoni, F.; Consonni, V.; Todeschini, R. In Silico Prediction of Cytochrome P450-drug Interaction: QSARs for CYP3A4 and CYP2C9. *Int. J. Mol. Sci.* **2016**, *17*, 914.

(51) Veith, H.; Southall, N.; Huang, R.; James, T.; Fayne, D.; Artemenko, N.; Shen, M.; Inglese, J.; Austin, C. P.; Lloyd, D. G.; Auld, D. S. Comprehensive Characterization of Cytochrome P450

Isozyme Selectivity across Chemical Libraries. *Nat. Biotechnol.* **2009**, *27*, 1050–1055.

(52) Schmidt, S.; Gonzalez, D.; Derendorf, H. Significance of Protein Binding in Pharmacokinetics and Pharmacodynamics. *J. Pharm. Sci.* **2010**, *99*, 1107–1122.

(53) Mehvar, R. Role of Protein Binding in Pharmacokinetics. *Am. J. Pharm. Educ.* **2005**, *69*, 103.

(54) Lin, L.; Wong, H. Predicting Oral Drug Absorption: Mini Review on Physiologically-based Pharmacokinetic Models. *Pharmaceutics* **2017**, *9*, 41.

(55) Koehn, L.; Habgood, M.; Huang, Y.; Dziegielewska, K.; Saunders, N. Determinants of Drug Entry into the Developing Brain. *F1000Research* **2019**, *8*, 1372.

(56) Van De Waterbeemd, H. Which In Vitro Screens Guide the Prediction of Oral Absorption and Volume of Distribution? *Basic Clin. Pharmacol. Toxicol.* **2005**, *96*, 162–166.

(57) Czodrowski, P. hERG Me Out. *J. Chem. Inf. Model.* **2013**, *53*, 2240–2251.

(58) De Ponti, F.; Poluzzi, E.; Montanaro, N. Organising Evidence on QT Prolongation and Occurrence of Torsades de Pointes with Non-antiarrhythmic Drugs: A Call for Consensus. *Eur. J. Clin. Pharmacol.* **2001**, *57*, 185–209.

(59) Danielson, P. á. The Cytochrome P450 Superfamily: Biochemistry, Evolution and Drug Metabolism in Humans. *Curr. Drug Metab.* **2002**, *3*, 561–597.

(60) Deodhar, M.; Al Rihani, S. B.; Arwood, M. J.; Darakjian, L.; Dow, P.; Turgeon, J.; Michaud, V. Mechanisms of CYP450 Inhibition: Understanding Drug-drug Interactions Due to Mechanism-based Inhibition in Clinical Practice. *Pharmaceutics* **2020**, *12*, 846.

(61) Landrum, G. RDKit: Open-source Cheminformatics (Accessed Feb. 2020). `http://www.rdkit.org`.

(62) Kramer, C.; Kalliokoski, T.; Gedeck, P.; Vulpetti, A. The Experimental Uncertainty of Heterogeneous Public K_i Data. *J. Med. Chem.* **2012**, *55*, 5165–5173.

(63) Kalliokoski, T.; Kramer, C.; Vulpetti, A.; Gedeck, P. Comparability of Mixed IC$_{50}$ Data–A Statistical Analysis. *PloS One* **2013**, *8*, e61007.

(64) Wenlock, M. C.; Carlsson, L. A. How Experimental Errors Influence Drug Metabolism and Pharmacokinetic QSAR/QSPR Models. *J. Chem. Inf. Model.* **2015**, *55*, 125–134.

(65) Vinyals, O.; Bengio, S.; Kudlur, M. Order Matters: Sequence to Sequence for Sets. *arXiv preprint arXiv:1511.06391* **2015**,

(66) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(67) Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. Smoothgrad: Removing Noise by Adding Noise. *arXiv preprint arXiv:1706.03825* **2017**,

(68) Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity Checks for Saliency Maps. Adv. Neural Inf. Proc. Sys. 2018; pp 9505–9515.

(69) Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features through Propagating Activation Differences. Proceedings of the 34th International Conference on Machine Learning, 70. 2017; pp 3145–3153.

(70) Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535–1535.

(71) Baumann, D.; Baumann, K. Reliable Estimation of Prediction Errors for QSAR Models under Model Uncertainty using

Double Cross-validation. *J. Cheminformatics* **2014**, *6*, 47.

(72) Clark, J. Acidity of Phenols. 2020; `https://chem.libretexts.org/@go/page/732`, [Online; accessed 2021-01-19].

(73) Milletti, F.; Storchi, L.; Sforna, G.; Cruciani, G. New and Original $p\mathrm{K}_a$ Prediction Method Using Grid Molecular Interaction Fields. *J. Chem. Inf. Model.* **2007**, *47*, 2172–2181.

(74) Kaur, P.; Chamberlin, A. R.; Poulos, T. L.; Sevrioukova, I. F. Structure-based Inhibitor Design for Evaluation of a CYP3A4 Pharmacophore Model. *J. Med. Chem.* **2016**, *59*, 4210–4220.

(75) Spector, A. A. Fatty Acid Binding to Plasma Albumin. *J. Lipid Res* **1975**, *16*, 165–79.

(76) Leishman, D. J.; Rankovic, Z. *Tactics in Contemporary Drug Design*; Springer, 2014; pp 225–259.

(77) C Braga, R.; M Alves, V.; FB Silva, M.; Muratov, E.; Fourches, D.; Tropsha, A.; H Andrade, C. Tuning HERG Out: Antitarget QSAR Models for Drug Development. *Curr. Top. Med. Chem.* **2014**, *14*, 1399–1415.

(78) Jamieson, C.; Moir, E. M.; Rankovic, Z.; Wishart, G. Medicinal Chemistry of hERG Optimizations: Highlights and Hang-ups. *J. Med. Chem.* **2006**, *49*, 5029–5046.

(79) Marchese Robinson, R. L.; Glen, R. C.; Mitchell, J. B. Development and Comparison of hERG Blocker Classifiers: Assessment on Different Datasets Yields Markedly Different Results. *Mol. Inform.* **2011**, *30*, 443–458.

(80) Josien, H.; Bara, T.; Rajagopalan, M.; Clader, J. W.; Greenlee, W. J.; Favreau, L.; Hyde, L. A.; Nomeir, A. A.; Parker, E. M.; Song, L.; Zhang, L.; Zhang, Q. Novel Orally Active Morpholine N-arylsulfonamides γ-secretase Inhibitors with Low CYP 3A4 Liability. *Bioorganic Med. Chem. Lett.* **2009**, *19*, 6032 – 6037.

(81) Li, Y.; Pasunooti, K. K.; Li, R.-J.; Liu, W.; Head, S. A.; Shi, W. Q.; Liu, J. O. Novel Tetrazole-containing Analogues of Itraconazole as Potent Antiangiogenic Agents with Reduced Cytochrome P450 3A4 Inhibition. *J. Med. Chem.* **2018**, *61*, 11158–11168.

(82) Mandal, M.; Mitra, K.; Grotz, D.; Lin, X.; Palamanda, J.; Kumari, P.; Buevich, A.; Caldwell, J. P.; Chen, X.; Cox, K.; Favreau, L.; Hyde, L.; Kennedy, M. E.; Kuvelkar, R.; Liu, X.; Mazzola, R. D.; Parker, E.; Rindgen, D.; Sherer, E.; Wang, H.; Zhu, Z.; Stamford, A. W.; Cumming, J. N. Overcoming Time-dependent Inhibition (TDI) of Cytochrome P450 3A4 (CYP3A4) Resulting from Bioactivation of a Fluoropyrimidine Moiety. *J. Med. Chem.* **2018**, *61*, 10700–10708.

(83) Zhao, L.; Sun, N.; Tian, L.; Zhao, S.; Sun, B.; Sun, Y.; Zhao, D. Strategies for the Development of Highly Selective Cytochrome P450 Inhibitors: Several CYP Targets in Current Research. *Bioorganic Med. Chem. Lett.* **2019**, *29*, 2016–2024.

(84) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool: Miniperspective. *J. Med. Chem.* **2011**, *54*, 7739–7750.

(85) Guha, R.; Van Drie, J. H. Structure-activity Landscape Index: Identifying and Quantifying Activity Cliffs. *J. Chem. Inf. Model.* **2008**, *48*, 646–658.

(86) Lázníček, M.; Lázníčková, A. The Effect of Lipophilicity on the Protein Binding and Blood Cell Uptake of Some Acidic Drugs. *J. Pharm. Biomed. Anal.* **1995**, *13*, 823–828.

(87) Hou, T.; Zhang, W.; Xia, K.; Qiao, X.; Xu, X. ADME Evaluation in Drug Discovery. 5. Correlation of Caco-2 Permeation with Simple Molecular Properties. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1585–1600.

(88) Liu, X.; Testa, B.; Fahr, A. Lipophilicity and its Relationship with Passive Drug Permeation. *Pharm. Res* **2011**, *28*, 962–977.

(89) Pearl, J. Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution. *arXiv preprint arXiv:1801.04016* **2018**,

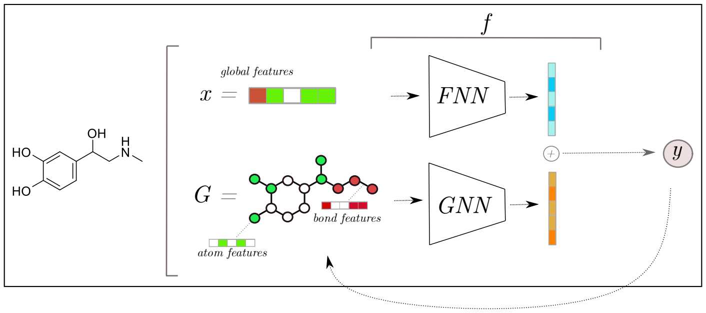
(90) Holzinger, A.; Carrington, A.; Müller, H. Measuring the Quality of Explanations: The System Causability Scale (SCS). *KI-Künstliche Intelligenz* **2020**, 1–6.

(91) Hase, P.; Bansal, M. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? *arXiv preprint arXiv:2005.01831* **2020**,

(92) Wang, M.; Zheng, D.; Ye, Z.; Gan, Q.; Li, M.; Song, X.; Zhou, J.; Ma, C.; Yu, L.; Gai, Y.; Xiao, T.; He, T.; Karypis, G.; Li, J.; Zhang, Z. Deep Graph Library: A Graph-centric, Highly-performant Package for Graph Neural Networks. *arXiv preprint arXiv:1909.01315* **2019**,

(93) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; pp 8024–8035.

Graphical TOC Entry



$$\int_{\Omega} \frac{\partial f(G + \alpha\epsilon)}{\partial G} d\alpha \qquad \int_{\Omega_2} \frac{\partial f(x + \alpha\epsilon)}{\partial x} d\alpha$$