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, cardiac potassium 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] 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, graph neural networks have the potential of replacing decades-old hand-crafted molecular fingerprint representations.[15] 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.[16] This 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),[17] and for making overly confident erroneous predictions.[18] '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.[19,20]

In the context of drug discovery-related applications, in particular for property prediction tasks, XAI methods can potentially help rationalize deep learning models by highlighting molecular substructures that are critical for a given prediction.[21–23] Analysis of the physicochemical properties of compounds can provide an alternative perspective. Several studies have examined the influence of such 'global' properties on drug-likeness estimations and other aspects of chemical compounds.[24–26] Herein, an established structure- and property-based XAI approach, the integrated gradients feature attribution technique,[27] was used and extended to examine its practical utility for a number of ADME and safety-related endpoints. 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.

Data sets

Four pharmacologically relevant parameters – plasma protein binding (PPB),[28] human ether-a-go-go-related gene (hERG) potassium channel inhibition,[29] passive drug permeability (Caco-2 assay),[30] and cytochrome P450 inhibition (CYP3A4 isoform) – were evaluated.[31] 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).

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

Endpoint	No. compounds	Task	References
Plasma protein binding	4,634	Regression	32–37
Caco-2 passive permeability	276	Regression	38
hERG inhibition	6,993	Regression	39,40
P450 inhibition	9,120	Binary classification	41,42

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).[43] High-affinity compounds for these targets may, in practice, require higher dosing to achieve effective concentrations in patients.[44] In the present study, data from six different studies,[32–37] comprising 4,634 drugs, were combined in order to construct a training set for predicting the fraction bound (f_b) in plasma.

Caco-2 cell passive permeability

Drugs administered orally must cross cell membranes to perform their function.[45] Such performance can be determined in vivo with radiolabeled compounds,[46] whereas the Caco-2 cell line is considered the in vitro gold standard proxy for studying pharmaceutical drug transport across cellular barriers.[47] For this endpoint, passive permeability data from 276 compounds was collected from two independent studies.[39,40] Passive permeability values (P_{app}) were collected (in cm s^{-1}) and converted to the

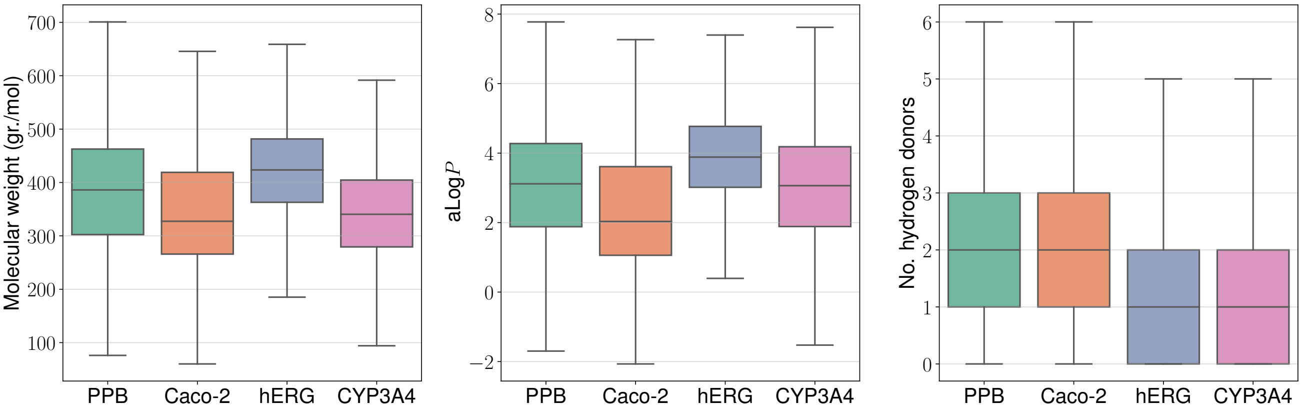


Figure 1: Box-whiskers plots of the distributions of molecular weight, calculated $\log P$ (aLog$P$) values, and the number of hydrogen donors. Caco-2, passive permeability; CYP3A4, cytochrome P450 3A4 inhibition hERG, human either-a-go-go cardiac potassium channel inhibition; PPB, plasma protein binding.

$\log_{10}$ scale for numerical stability during the model training. If several measurements were available for the same compound, we considered their arithmetic average as the P_{app} target value.

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.[48,49] For this endpoint, data compiled by Sato *et al.* was used,[38] 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 the 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.[50] For this endpoint, data compiled by Nembri *et al.* was used,[41] encompassing 9,120 CYP3A4 inhibitors and substrates with binary activity information (active/inactive), as determined by Veith *et al.*[42]

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) and bonds (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 edges 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, ψ is a message function that is defined on each edge and combines its features with those of its neighboring nodes. ϕ 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 different combinations of message, update, and reduction functions result in different MPNN architectures. The message and update functions contain weights that are learnable by backpropaga-

Table 2: Vertex, bond, and 'global' molecular graph features computed with RDkit[52]

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

tion. 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[51] to embed molecules and compute a prediction. This model and other MPNN variations were shown to perform well on several ligand-based tasks.[14] Furthermore, to account for unspecific molecular interactions, a fully connected neural network sub-architecture was also included for consideration of computed physicochemical features $x \in \mathbb{R}^{d_3}$. Selected vertex, bond, and global features were computed with the RDkit software (Table 2).[52] Full details on the network architecture and hyperparameter selection are included 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. We trained models on each data split for 250 epochs, with a batch size of 32 samples, and employed the Adam stochastic optimizer[53] with default momentum parameters ($\beta_1 = 0.9, \beta_2 = 0.999$) and a starting learning rate of 10^{-4}.

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 $z \in \mathcal{X}$. This process can be performed by gradient backpropagation to the input features of the nodes, edges, and global features:[54,55] $\left(i.e. \, \frac{\partial f}{\partial x_v}, \frac{\partial f}{\partial w_e}, \frac{\partial f}{\partial x} \right)$.

In practice, however, this approach has several limitations, such as gradient saturation.[56] 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 suggested by McCloskey *et al.*,[57] the integrated gradients method[27] 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:

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

Because the integral in Equation 3 is non-tractable; it was computed with a Riemann approximation in accordance with:

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

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

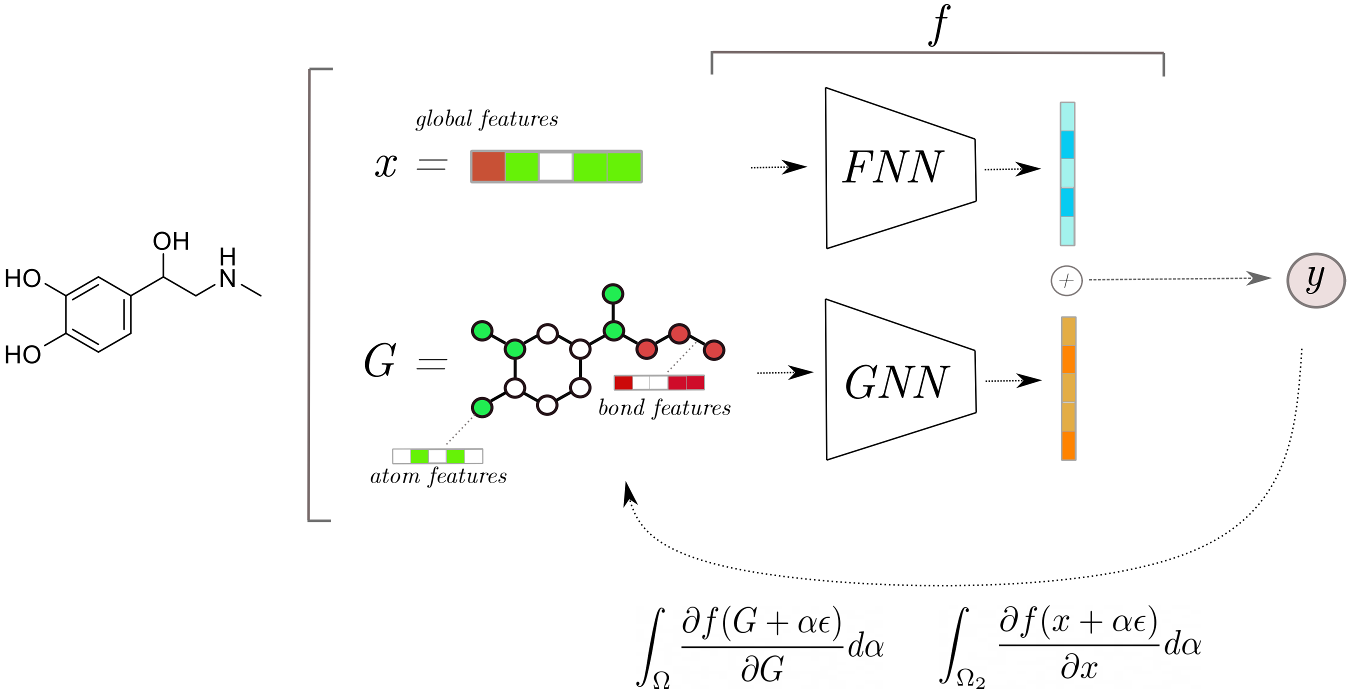

f
global features
$x =$
FNN
OH
HO
H
N
HO
$G =$
GNN
bond features
atom features
$+$
y
$\int_{\Omega} \frac{\partial f(G+\alpha\epsilon)}{\partial G} d\alpha$
$\int_{\Omega_2} \frac{\partial f(x+\alpha\epsilon)}{\partial x} d\alpha$

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[27] was then applied to compute atom, bond, and global importance scores.

iterated over $m = 50$ steps, and utilized baselines corresponding to zeroed-out vertex, edge, and global feature tensors. For visualization, computed edge importance values were evenly distributed among their connecting vertices:

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

where $\mathcal{N}(v)$ is the set of neighboring vertices at one bond distance from vertex v. 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. Bonds (edges) were colored according to whether the color of their connecting nodes matched.

Table 3: Predictive performance of the $k = 10$ cross-validation scheme for the endpoints considered. Pearson's correlation coefficient R 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**	**RMSE**	**AUC**
Plasma protein binding	0.77 ± 0.03	19.85 ± 0.9	-
Passive permeability	0.71 ± 0.09	0.68 ± 0.09	-
hERG inhibition	0.63 ± 0.03	0.76 ± 0.03	
P450 inhibition	-	-	0.85 ± 0.01

RMSE, root mean square error; AUC, area under receiver-operator characteristic curve.

Model validation

To enable preliminary model benchmarking, 25 molecular series were extracted and compiled from available literature (provided in Supporting Data). These series represent back-

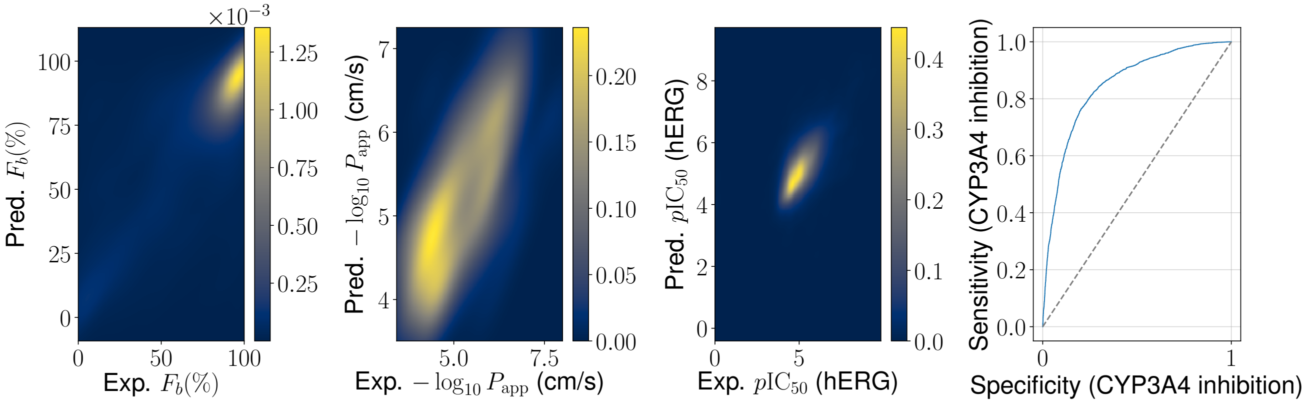


Figure 3: Model performance. A $k = 10$ cross-validation scheme was used. From left to right, two-dimensional density plots portraying experimental vs predicted values for the plasma protein binding, passive permeability, and hERG inhibition data sets, and the CYP P450 3A4 inhibition data. For the CYP data set, a receiver operating characteristic (ROC) curve is reported given its binary activity label (active/inactive).

ground knowledge and contain examples that are known to be relevant for the pharmacological endpoints considered in this study. 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), and (iii) were able to identify ‘unspecific’ ligand-protein interactions mediated by molecular properties (*e.g.*, log$P$, TPSA).

Results and discussion

Model performance

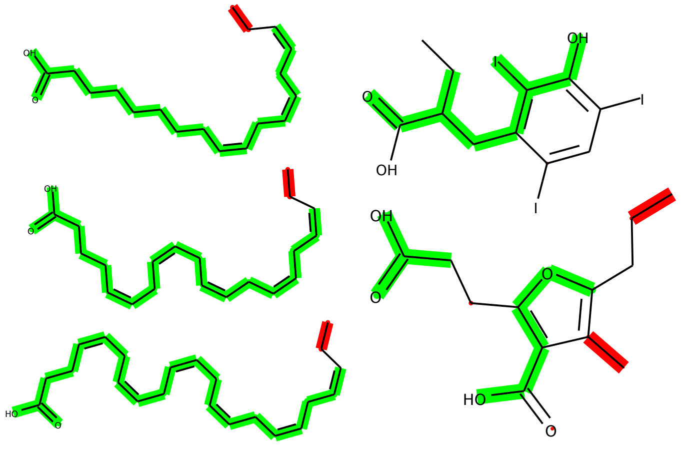
To assess whether the proposed feature attribution approach was able to extract meaningful relationships between structural motifs and the respective pharmacological endpoints, a rigorous performance evaluation was mandatory since explanations provided using a model with limited predictive capability should not be trusted. Results of a quantitative benchmark are presented in Figure 3 and Table 3, where the root mean squared error (RMSE), Pearson’s correlation coefficient R between experimental and predicted values, and the receiver-operator characteristic area under the curve (AUC) are reported.

All trained models showed predictive capabilities, with R values ranging between 0.63 and 0.77 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 predictions, 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[59] and a pharmacophore motif[35] consisting of two acidic groups separated by a hydrophobic part of five bond units (Figure 4).

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.[60] This effect could be explained by the fact that the ligand-accommodating cavity of the hERG potassium channel stabilizes



(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). These compounds feature two acidic groups separated by a hydrophobic part of five bond units. Green and red areas represent structural positive and negative contributions, respectively, w.r.t the ligand fraction bound f_b

positive charges. The second example illustrates the introduction of an activity cliff by bioisosteric replacement[61] (Figure 5b). Further examples of hERG, such as the effect of 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,[58] highlighting the importance of a flexible backbone, hydrogen-bond donor/acceptor moieties, and hydrophobic interactions (Figure 6a). Lowering the global molecular weight and crowding a basic amine was previously reported as a strategy for mitigating the CYP3A4 activity of morpholine-based N-arylsulfonamide γ-secretase inhibitors.[62] Of note, the relative importance of the corresponding structural features was correctly recognized (Figure 6b). Additional examples[63–65] are provided in the Supporting Data and the accompanying code repository of this work.

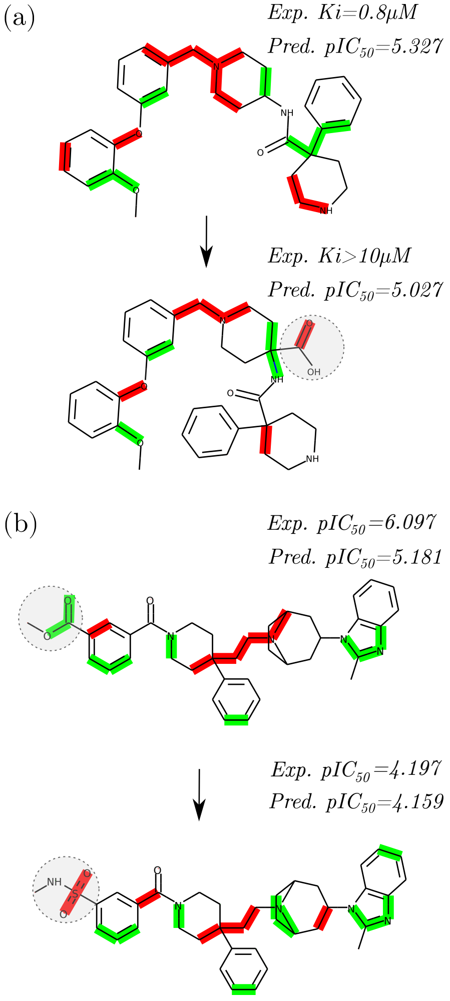


Figure 5: Examples of motifs indicating hERG inhibition. (a) Addition of a negative charge and (b) bioisosteric replacements causing activity cliffs. Green and red colors represent structural positive and negative contributions towards hERG inhibition, respectively

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.[66] The cliffs were ranked according to the structure activity landscape index (SALI).[67] This functional balances the structural similarity of a pair of compounds with their predicted property difference:

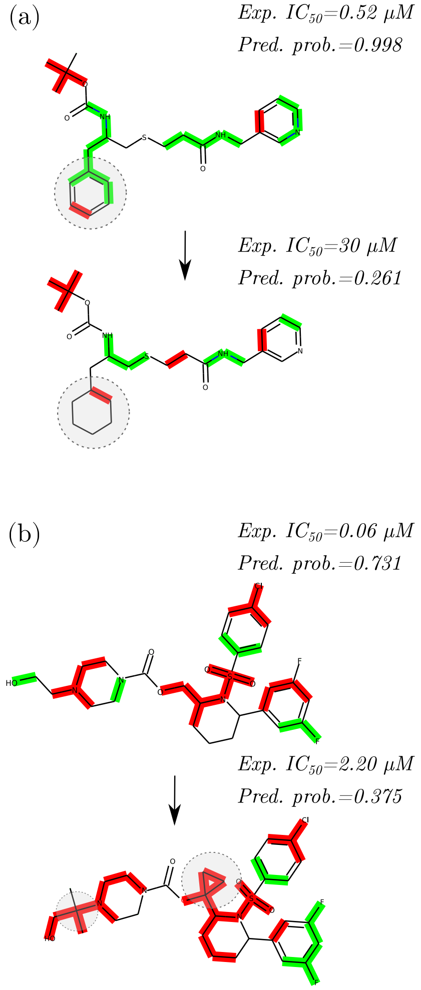


Figure 6: Cytochrome (CYP) inhibition motif replication examples. (a) Structure-based pharmacophore developed by Kaur *et al.*[58] (b) Activity cliffs caused by crowding a basic amine and lowering the overall molecular weight. Green and red areas represent structural positive and negative contributions, respectively, towards CYP3A4 inhibition.

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

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. Examples of SALI ranking for the endpoints considered in this study are presented in Figure 7 and in the accompanying code repository. 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 when a certain structural feature is present, or a 'negative' contribution in its absence.

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,[68] increasing circulation half-life, and reducing glomerular filtration. Our collected data set revealed a moderate positive correlation between $\mathrm{aLog}P$ and the fraction bound ($R = 0.5, p < 0.01$, one-tailed Pearson's correlation test), which was confirmed by the importance assigned to $\mathrm{aLog}P$ ($R = 0.55, p < 0.01$) 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[69] (meaning that compounds with a large polar surface area are unlikely to permeate cell membranes) and lipophilicity[70] (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 $\mathrm{aLog}P$ ($R = 0.31, p < 0.01$). The first relationship was again correctly captured by the XAI approach, indicating a moderate negative correlation between the importance assigned to TPSA and the P_{app} endpoint ($R = -0.59, p < 0.01$).

Comparison to other coloring approaches

Lastly, the XAI approach herein proposed was compared to the molecular coloring method published by Sheridan *et al.*,[22] which is model-agnostic and can be used for either regression or classification tasks. In order to highlight

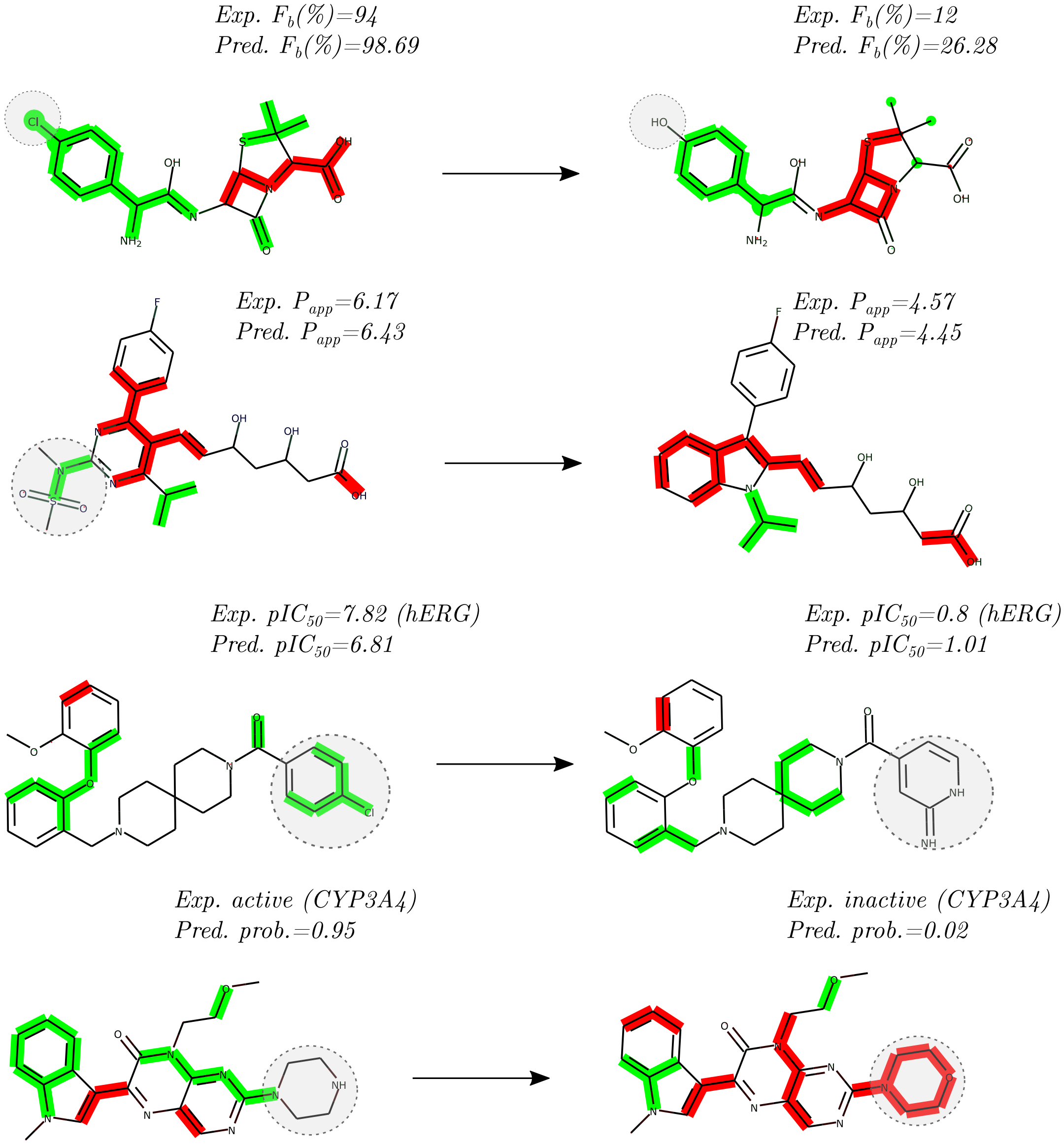


Figure 7: Some examples of the property cliffs identified by the proposed approach, selected via the SALI index for all the endpoints and data sets considered in this study. Green and red values represent positive and negative contributions, respectively, w.r.t. the considered endpoint.

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 molecules for which the fingerprint-based model identified motifs corresponding to known pharmacophores of the hERG and CYP3A4 endpoints. The approach proposed here failed for these examples, whereas the fingerprint-based approach was unable to reproduce any of the other coloring examples presented in this study (Figures 4-7). 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. Furthermore, 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 *et al.*, using a random forest model featured with ECFP4

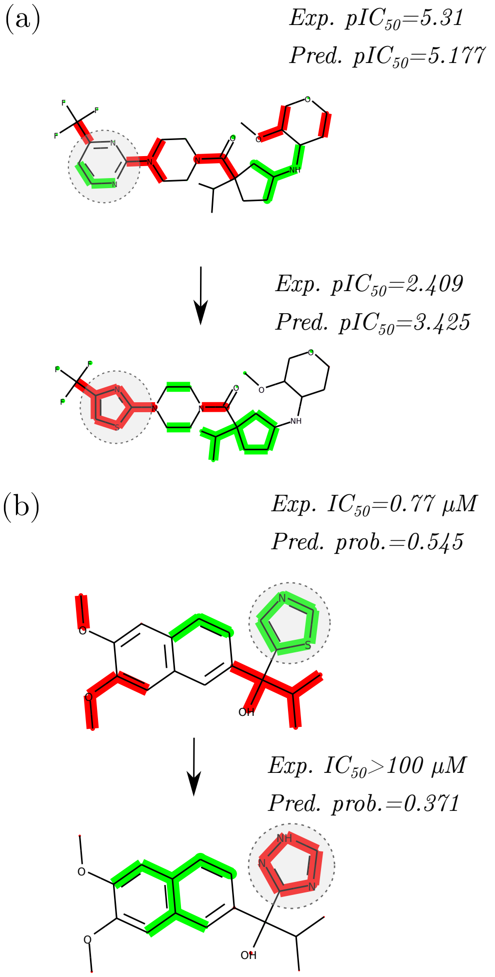


Figure 8: Examples using the approach of Sheridan *et al.*[22] for the (a) hERG endpoint, involving a bioisosteric ring transformation, and for the (b) CYP3A4 endpoint, involving a heme-binding group substitution. Green and red colors represent positive and negative contributions, respectively, w.r.t. the considered endpoint.

fingerprints is provided as supporting code, together with trained models for all of the endpoints considered here.

Conclusion

Herein, we described the extension of a popular XAI framework, the integrated gradients feature attribution technique, and its application 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 multicollinearity, 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.[71] Second, 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.[72,73] 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.

Implementation and code availability

The graph neural-network models were trained with the Deep Graph Library Python (DGL) package (version 0.4.3)[74] and the dgllife extension (github.com/awslabs/dgl-lifesci) that run on top of the PyTorch tensor manipulation library (version 1.4.0).[75] Molecular structures were handled using RDkit.[52] 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 Discovery 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. *Chemical Society Reviews* **2020**, *49*, 3525–3564.

(3) Gedeck, P.; Lewis, R. A. Exploiting QSAR models in lead optimization. *Current Opinion in Drug Discovery & Development* **2008**, *11*, 569.

(4) Lewis, R. A. A general method for exploiting QSAR models in lead optimization. *Journal of Medicinal Chemistry* **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. *Journal of 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. *Chemical Science* **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. *Journal of Chemical Information and Modeling* **2018**, *59*, 1172–1181.

(10) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De novo design of bioactive small molecules by artificial intelligence. *Molecular Informatics* **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. *Chemical Science* **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. *Chemical Science* **2018**, *9*, 513–530.

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

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

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

(18) 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.

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

(20) 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**,

(21) 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* **2017**, *57*, 1773–1792.

(22) Sheridan, R. P. Interpretation of QSAR models by coloring atoms According to changes in predicted activity: How robust is it? *Journal of Chemical Information and Modeling* **2019**, *59*, 1324–1337.

(23) Rodríguez-Pérez, R.; Bajorath, J. Interpretation of compound activity predictions from complex machine learning models using local approximations and Shapley values. *Journal of Medicinal Chemistry* **2020**, *63*, 8761–8777.

(24) Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* **2004**, *1*, 337 – 341.

(25) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *Journal of Chemical Information and Computer Sciences* **2001**, *41*, 1308–1315.

(26) 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 Chemical Neuroscience* **2010**, *1*, 435–449.

(27) 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.

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

(29) 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.

(30) 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.

(31) 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. *European Journal of Biochemistry* **1993**, *218*, 585–595.

(32) 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. *Pharmaceutical Research* **2013**, *30*, 1790–1798.

(33) 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. *Journal of Chemical Information and Modeling* **2016**, *56*, 2243–2252.

(34) 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.

(35) Kratochwil, N. A.; Huber, W.; Müller, F.; Kansy, M.; Gerber, P. R. Predicting plasma protein binding of drugs: A new approach. *Biochemical Pharmacology* **2002**, *64*, 1355–1374.

(36) 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. *Journal of Medicinal Chemistry* **2006**, *49*, 7169–7181.

(37) 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. *Molecular Pharmaceutics* **2018**, *15*, 5302–5311.

(38) Sato, T.; Yuki, H.; Ogura, K.; Honma, T. Construction of an integrated database for hERG blocking small molecules. *PloS One* **2018**, *13*, e0199348.

(39) Bittermann, K.; Goss, K.-U. Predicting apparent passive permeability of Caco-2 and MDCK cell-monolayers: A mechanistic model. *PloS One* **2017**, *12*, e0190319.

(40) 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.

(41) 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* **2016**, *17*, 914.

(42) 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. *Nature Biotechnology* **2009**, *27*, 1050–1055.

(43) Schmidt, S.; Gonzalez, D.; Derendorf, H. Significance of protein binding in pharmacokinetics and pharmacodynamics. *Journal of Pharmaceutical Sciences* **2010**, *99*, 1107–1122.

(44) Mehvar, R. Role of protein binding in pharmacokinetics. *American Journal of Pharmaceutical Education* **2005**, *69*, 103.

(45) Lin, L.; Wong, H. Predicting oral drug absorption: Mini review on physiologically-based pharmacokinetic models. *Pharmaceutics* **2017**, *9*, 41.

(46) Koehn, L.; Habgood, M.; Huang, Y.; Dziegielewska, K.; Saunders, N. Determinants of drug entry into the developing brain. *F1000Research* **2019**, *8*, 1372.

(47) Van De Waterbeemd, H. Which in vitro screens guide the prediction of oral absorption and volume of distribution? *Basic & Clinical Pharmacology & Toxicology* **2005**, *96*, 162–166.

(48) Czodrowski, P. hERG me out. *Journal of Chemical Information and Modeling* **2013**, *53*, 2240–2251.

(49) 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. *European Journal of Clinical Pharmacology* **2001**, *57*, 185–209.

(50) Danielson, P. á. The cytochrome P450 superfamily: Biochemistry, evolution and drug metabolism in humans. *Current Drug Metabolism* **2002**, *3*, 561–597.

(51) Vinyals, O.; Bengio, S.; Kudlur, M. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391* **2015**,

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

(53) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

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

(55) Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; Kim, B. Sanity checks for saliency maps. Advances in Neural Information Processing Systems. 2018; pp 9505–9515.

(56) 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.

(57) 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* **2019**, *116*, 11624–11629.

(58) Kaur, P.; Chamberlin, A. R.; Poulos, T. L.; Sevrioukova, I. F. Structure-based inhibitor design for evaluation of a CYP3A4 pharmacophore model. *Journal of Medicinal Chemistry* **2016**, *59*, 4210–4220.

(59) Spector, A. A. Fatty acid binding to plasma albumin. *Journal of Lipid Research* **1975**, *16*, 165–79.

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

(61) 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. *Current Topics in Medicinal Chemistry* **2014**, *14*, 1399–1415.

(62) 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 & Medicinal Chemistry Letters* **2009**, *19*, 6032 – 6037.

(63) 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. *Journal of Medicinal Chemistry* **2018**, *61*, 11158–11168.

(64) 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. *Journal of Medicinal Chemistry* **2018**, *61*, 10700–10708.

(65) 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 & Medicinal Chemistry Letters* **2019**, *29*, 2016–2024.

(66) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched molecular pairs as a medicinal chemistry tool: Miniperspective. *Journal of Medicinal Chemistry* **2011**, *54*, 7739–7750.

(67) Guha, R.; Van Drie, J. H. Structure-activity landscape index: Identifying and quantifying activity cliffs. *Journal of Chemical Information and Modeling* **2008**, *48*, 646–658.

(68) Lázníček, M.; Lázníčková, A. The effect of lipophilicity on the protein binding and blood cell uptake of some acidic drugs. *Journal of Pharmaceutical and Biomedical Analysis* **1995**, *13*, 823–828.

(69) 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. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1585–1600.

(70) Liu, X.; Testa, B.; Fahr, A. Lipophilicity and its relationship with passive drug permeation. *Pharmaceutical Research* **2011**, *28*, 962–977.

(71) Pearl, J. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016* **2018**,

(72) Holzinger, A.; Carrington, A.; Müller, H. Measuring the quality of explanations: The system causability scale (SCS). *KI-Künstliche Intelligenz* **2020**, 1–6.

(73) Hase, P.; Bansal, M. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? *arXiv preprint arXiv:2005.01831* **2020**,

(74) 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**,

(75) 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.