JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

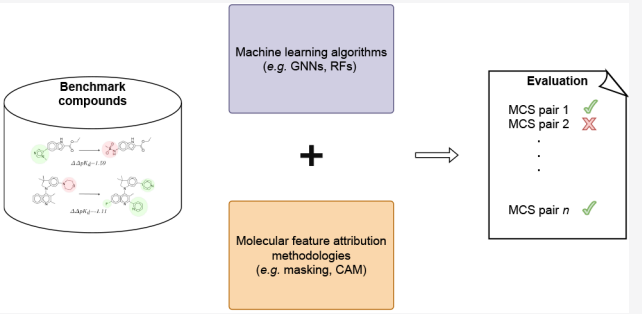
Benchmarking Molecular Feature Attribution Methods with Activity Cliffs

José Jiménez-Luna,* Miha Skalic, and Nils Weskamp

Cite This: J. Chem. Inf. Model. 2022, 62, 274−283 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Feature attribution techniques are popular choices within the explainable artificial intelligence toolbox, as they can help elucidate which parts of the provided inputs used by an underlying supervised-learning method are considered relevant for a specific prediction. In the context of molecular design, these approaches typically involve the coloring of molecular graphs, whose presentation to medicinal chemists can be useful for making a decision of which compounds to synthesize or prioritize. The consistency of the highlighted moieties alongside expert background knowledge is expected to contribute to the understanding of machine-learning models in drug design. Quantitative evaluation of such coloring approaches, however, has so far been limited to substructure identification tasks. We here present an approach that is based on maximum common substructure algorithms applied to experimentally-determined activity cliffs. Using the proposed benchmark, we found that molecule coloring approaches in conjunction with classical machine-learning models tend to outperform more modern, graph-neural-network alternatives. The provided benchmark data are fully open sourced, which we hope will facilitate the testing of newly developed molecular feature attribution techniques.



INTRODUCTION

Deep learning has quickly become a defacto first-class citizen modeling approach in drug discovery applications, with the main advantage compared to other classical machine-learning (ML) methods being their automatic feature extraction capabilities.[1,2] Among those approaches, message-passing methods, also known as graph-neural-network models[3,4] (GNNs) have recently become increasingly popular in chemoinformatics for relevant tasks such as molecular property prediction,[5] de novo generative design,[6,7] or synthesis prediction.[8]

The rise of complex deep-learning methodologies in the discussed and related fields has also been accompanied by an increasing demand of explainability, as their inner workings continue to remain elusive to interpretation among field experts.[9] Additionally, while these models have been shown to provide impressive predictive capabilities in many use cases, their performance, especially in fields that feature heavy experimental uncertainty, has been far from perfect, making it necessary to critically assess and rationalize their predictions before decision making. As a consequence, explainable artificial intelligence (XAI) has become a very active topic of research in theoretical ML,[10] as well as within other more applied fields such as computer vision and natural language understanding.[11,12] In the specific context of chemoinformatics, several attempts have been made in recent years with the aim of uncovering black-box ML algorithms in property prediction tasks.[13−15] In particular, while some studies go to great lengths to show how several modern feature attribution methods can be used to some extent to identify structural motifs[16,17] or property cliffs,[18] it is hard to evaluate which feature attribution methods work best and under which specific conditions. Along these lines, a study by Sánchez-Lengeling et al.[19] proposed a quantitative benchmark for several well-known feature attribution techniques in conjunction with GNNs. While it was shown that some modern feature attribution techniques can correctly highlight certain motifs, the said benchmark was mostly limited to synthetic tasks where the evaluation procedure consisted of identifying whether certain molecules contained a set of predefined molecular substructures. Furthermore, nondeep-learning approaches alongside other classical coloring techniques were not considered in the study, while these have the advantage of working under a wider umbrella of ML models and descriptors.[13,20−23]

In real drug discovery settings, however, one is usually interested in the explicit prediction of pharmacologically

Received: September 23, 2021
Published: January 12, 2022

https://doi.org/10.1021/acs.jcim.1c01163
J. Chem. Inf. Model. 2022, 62, 274−283

relevant end points, such as potency, or complementary ones, such as absorption, distribution, metabolism, excretion, and toxicity (ADMET),[24] which in practice imply a certain degree of inherent experimental uncertainty.[25] With the goal of overcoming the limitations of previous studies, in this work, we propose what we believe to be a more realistic approach to evaluate feature attribution methods for in silico drug discovery. We rely on maximum common substructure (MCS) algorithms to build a large data collection of biologically active pairs of closely related compounds and use their associated activity information as a proxy for producing "ground-truth" colorings. We believe that this systematic, large-scale identification of examples yields a more relevant, comprehensive, and less biased analysis than a purely qualitative validation that is based on manually selected test cases or well-known pharmacophores extracted from the literature.[18] Given their rising popularity, we evaluate several popular graph neural network architectures, as well as different associated coloring procedures, and benchmark them against other classical techniques. To our surprise, we find that a comparatively simple approach reported by Sheridan,[20] which uses a random forest as the underlying machine-learning model, significantly outperforms all of the modern GNN-based feature attribution techniques in the proposed benchmark when the ligands present in the latter are also contained in the training sets. However, we also find out that none of the considered approaches manages to achieve satisfactory coloring performance on previously unseen examples. Finally, we investigate the obtained results to rationalize the observed performance differences, describe possible directions for future research, and provide usage recommendations.

MATERIALS AND METHODS

We used two databases for different and complementary purposes, namely the BindingDB protein−ligand validation sets[26,27] (accessed January 2021) and the ChEMBL[28] database of drug-like molecules (version 27). The BindingDB protein−ligand validation sets were used as a starting point to build the proposed benchmark, which features 1222 molecular congeneric series of sizes ranging between 10 and 50 compounds. Having obtained an evaluation set based on closely related compounds, additional activity data per target were necessary in order to train all underlying supervised ML models. Toward that end, we used the UniProt[29] identifier associated with each of the targets considered in the benchmark data and correspondingly ran a compound search in a locally installed PostgreSQL instance of the ChEMBL database. Several selection criteria were applied: Only noncensored activity information in either IC_{50}, K_d, or K_i units was considered, and only training sets with at least 100 activity data points were kept. After applying these filters, 997 training sets could be successfully extracted.

Determining Ground-Truth Colors. In order to determine ground-truth atom-level color labels for the considered benchmark sets, an implementation of the FMCS[30] maximum common substructure algorithm was used, as available in the rdFMCS module of the RDKit software package.[31] MCS calculations were run for all compound pairs in each benchmark series whose activity difference exceeded 1 log units. We excluded those cases where at least one compound had a molecular weight higher than 800 Da, those pairs whose fraction of common substructure atoms was below 50%, and those whose MCS calculation time exceeded 5 min, for computational expense reasons. This procedure resulted in 729 series featuring at least one colored pair of compounds. Histograms describing the distribution of both the number of benchmark pairs and training compounds considered per set, as well as the percentage of benchmark compounds that were initially present in the training sets, are provided in Figure 1. Additionally, we considered benchmark pairs of compounds at different thresholds depending on their percentage of shared atoms, as determined by the MCS procedure. The number of pairs considered at each of these thresholds can be checked on Figure 2.

Each identified pair represents an activity cliff, and we assume that the observed potency difference has to be the result of the structural variation between the two compounds.[32] Moieties present only in the more active compound are expected to receive a positive feature attribution and vice versa. Atomic labels were therefore assigned depending on the sign of the activity difference in a pair with the common substructure considered neutral (see Figure 3 for examples). We then assess the performance of the examined methods based on how well the sign of the atom-level coloring agrees with the assigned ground truth. Finally, these scores are averaged out on a molecule level so that side chains of different sizes have equal contributions toward the final accuracy metric. However, since the described procedure can result in the same compound being present in different constructed pairs, color inconsistencies can be produced. Specifically, this issue can happen if the sign of the activity difference differs for a compound present in different benchmark pairs, which is the case here with a per-target ratio average of 32% offending compounds. Since removing such problematic pairs is not an adequate workaround as there is no clear way of determining which ones to select, we provide additional ways of evaluation. In the second part of the benchmark, instead of focusing on the atom-level absolute sign of the predicted coloring, for each pair we instead check whether the average noncommon atomic contributions defined by the MCS of each molecule pair agrees with the direction of their activity ranking, which circumvent the previous issue. Specifically, if $\mathcal{M}_i$ and $\mathcal{M}_j$ are the noncommon atomic sets defined by the MCS of two molecules m_i, m_j extracted from the same congeneric series such that potency (m_1) > potency (m_2), and ψ: $\mathcal{M} \rightarrow \mathbb{R}^p$ is an atom-level feature attribution method, we simply check whether

$$\text{avg}(\psi(\mathcal{M}_i)) > \text{avg}(\psi(\mathcal{M}_j)) \quad (1)$$

where avg is a simple average function applied over the attributions assigned to the noncommon atomic sets defined by the MCS of each benchmark pair.

Models. We make use of the GNN implementations provided by Sánchez-Lengeling et al.,[19] which includes four popular variants: GraphNets,[33] Graph-Convolutional Neural Networks (GCNs),[4] Message-Passing Neural Networks (MPNNs),[3] and Graph-Attention Neural Networks (GATs).[34] Arguably, all of these architectures fall under the umbrella of message-passing algorithms, which for completeness we briefly summarize here.

Given a graph $\mathcal{G} = (V, E, u)$ with vertices $v \in V$, edges $e \in E$, and optional global graph information $u \in \mathbb{R}^k$, a graph neural network is a function f that takes a graph as an input and whose output is another graph with equal topology but with updated (i.e., latent) node, edge and global information. In practice, updated representations are aggregated via a readout

https://doi.org/10.1021/acs.jcim.1c01163
J. Chem. Inf. Model. 2022, 62, 274−283

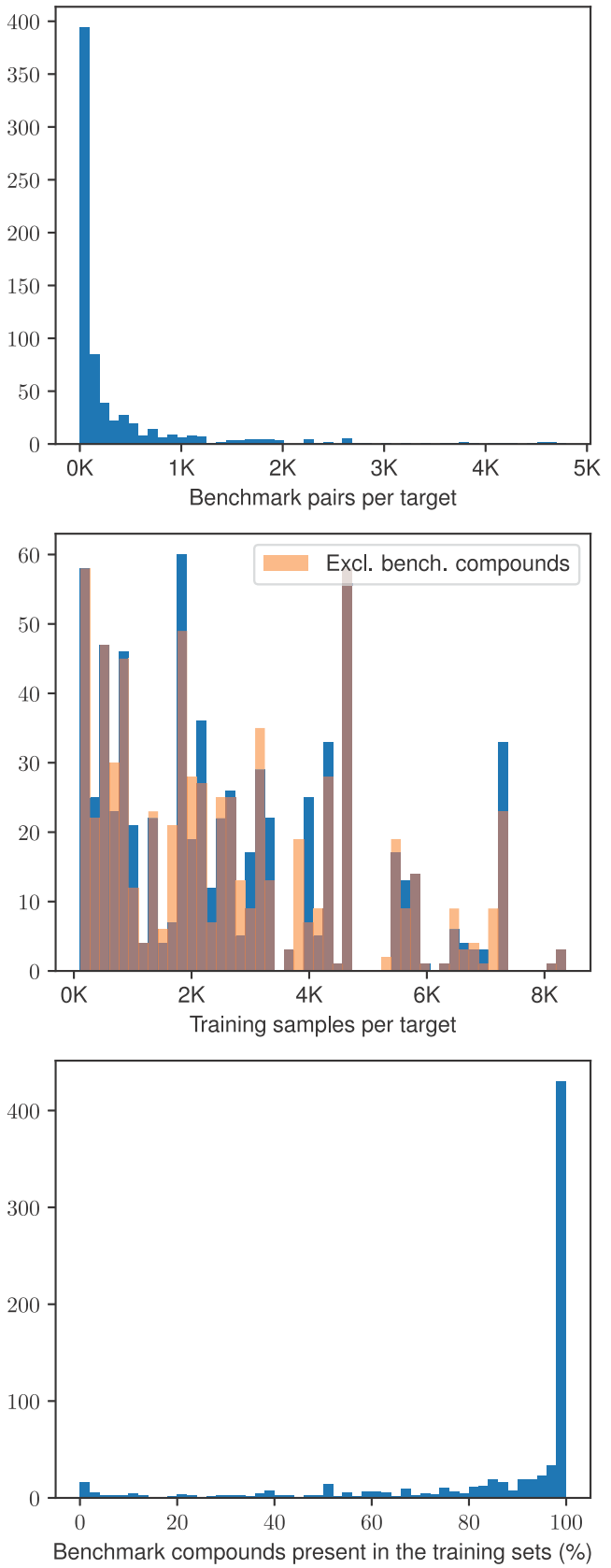


Figure 1. Histograms portraying the distribution of the number of benchmark pairs, training compounds available per target considered in the BindingDB protein−ligand validation sets (before and after removing identical ligands present in their respective benchmark sets), and percentage of benchmark ligands initially present in the training sets.

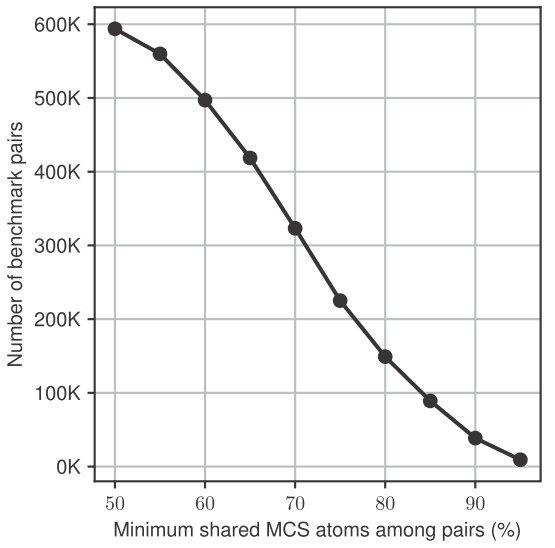


Figure 2. Number of benchmark pairs per each of the considered minimum percentage of common atoms, as determined by the FMCS algorithm.

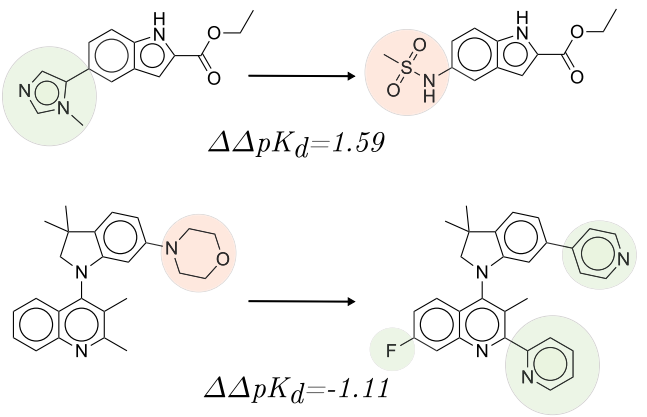


Figure 3. Two example ground-truth colorings for ligand pairs extracted from the 5BOT-4UM and 4F5Z-4F5 BindingDB congeneric series, respectively. MCS calculations between the aforementioned pairs were carried out, and noncommon atoms were assigned ground-truth labels according to their activity difference sign.

function into a single latent vector that can be then forward propagated to a single scalar, so that $f: \mathcal{G} \rightarrow \mathbb{R}$. The previous architecture variants mainly differ in the choice of message and update functions (i.e., the strategies which determine how node and edge latent spaces are computed). For all GNN block types considered, nodes and edges in the molecular graphs were featurized with the descriptors detailed in Table 1. As baselines, we furthermore consider a three-hidden-layer fully connected neural network with ReLU nonlinearities, and

Table 1. Node and Edge Molecular Graph Features Used in Training of GNN Models, as Computed with the RDKit[31] Software Package

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

https://doi.org/10.1021/acs.jcim.1c01163
J. Chem. Inf. Model. 2022, 62, 274−283

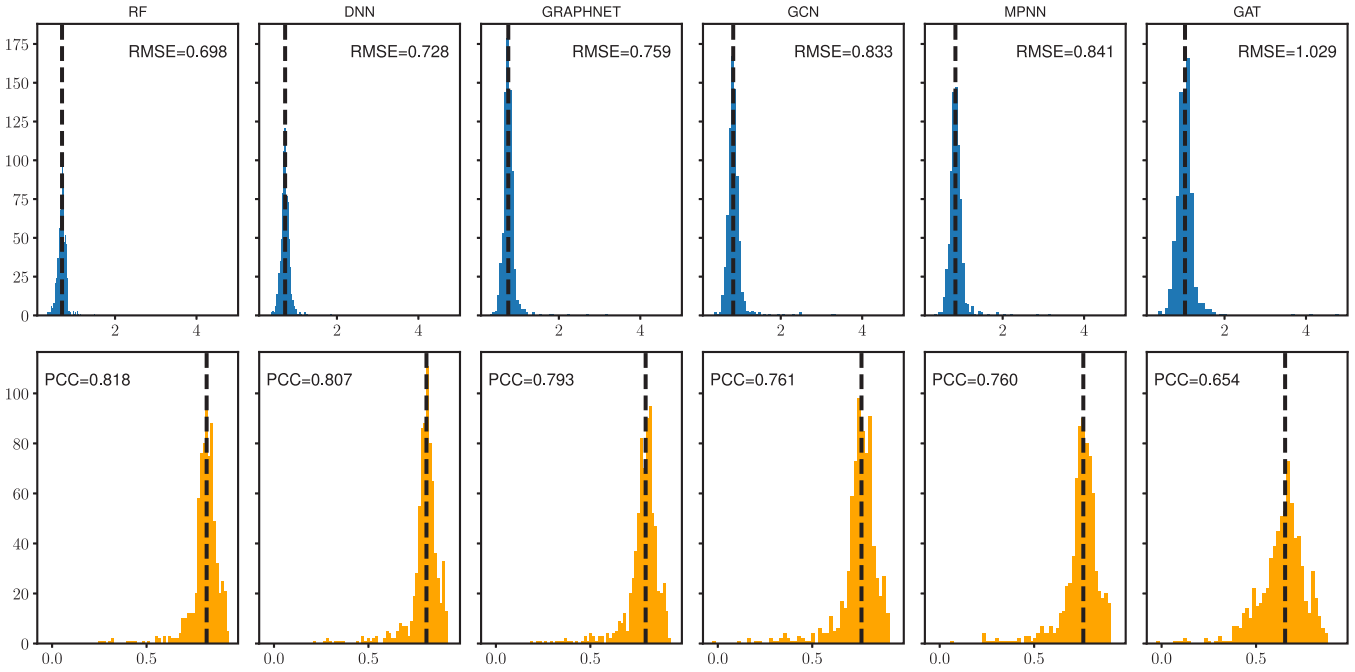


Figure 4. Predictive performance, as measured via root-mean-squared error (RMSE, above) and Pearson's correlation coefficient (PCC, below), for all the six considered model types and 729 training sets considered in this study, using a 20% random test split in each. Dotted vertical black lines mark the median value for each model and metric combination.

a random forest model, both using Extended Connectivity Fingerprints (ECFP4)[35] as input descriptors.

Feature Attribution Techniques. Several popular deep-learning feature attribution methods were used, as available in the accompanying code repository of Sánchez-Lengeling et al.:[19] GradInput,[36] Class Activation Maps (CAM),[37] Gradient Class Activation Maps (GradCAM),[38] Integrated Gradients,[39] and Attention Weights.[34] For completeness, a masking-like approach where the atom features in each node in the graph are sequentially zeroed out, and its corresponding graph then forward passed through the model, was also implemented (this method is referred to as "diff" in what follows). Furthermore, for baseline purposes, the fingerprint-based masking method proposed by Sheridan[20] was also implemented, where the types in each atom of a molecule are sequentially changed to one that is not present in the training set, and the difference between the predictions using the unmodified and modified fingerprints is taken as a proxy for atom importance. ECFP4 fingerprints are computed for these modified molecules and then used for prediction using either a random forest or a fully connected neural-network model trained with the hyper-parameters specified in the previous section. While there is a plethora of ML models that can also be used also in combination with other molecular fingerprinting strategies (e.g., Daylight fingerprints[40]), we consider only the previous two combinations mainly due to their simplicity and practical popularity.

Training and Other Details. All GNN and fully connected-layer models models were trained for a fixed number of 300 epochs, using a learning rate of 3×10^{-4} and a batch size of 32 samples. The rest of the hyperparameters were set as the default ones detailed in Sánchez-Lengeling et al.[19] Three hidden layers with a size of 64 units were used for the node-update multilayer perceptrons in the GNN architectures, while for the fully connected models we used three hidden layers with a size of 256 units. Random forest models were trained with 1000 base trees, and ECFP4 fingerprints with a bond radius of 2 units, computed via the RDKit software. For the Integrated Gradients feature attribution method, 500 Riemann integral approximation steps were used. Since most of the deep-learning coloring approaches produce both a score per node c_v as well as per edge $b_{u,v}$, and the proposed benchmark only considers the first, edge contributions were evenly distributed among their connecting nodes according to

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

where $\mathcal{N}(v)$ is the set of neighboring vertices at one bond distance from vertex v.

RESULTS

Model Predictive Performance. A summary of the predictive performance of all the considered machine-learning models (using a 20% random test-set split) and for each target considered is provided in Figure 4. Most model types show satisfactory predictive capabilities, with median Pearson correlation coefficient (PCC) values above 0.7, although the GAT model type falls slightly behind with a median PCC of 0.65. Furthermore, with a median root-mean-squared error (RMSE) and PCC values between experimental and predicted values of 0.7 and 0.82, respectively, the random forest model significantly outperforms the second best-performing alternative, namely, the fully connected neural network that uses the same descriptor type (Wilcoxon paired signed-rank test, p-values < 0.01), and consequently the rest of the other competing GNN-based models. These results are in line with some conclusions drawn from previous related research[41,42] where it was shown that bagging and boosting-based[43−45] ML models performed at least on par with more modern deep-learning alternatives at a fraction of their computational cost.

https://doi.org/10.1021/acs.jcim.1c01163
J. Chem. Inf. Model. 2022, 62, 274−283

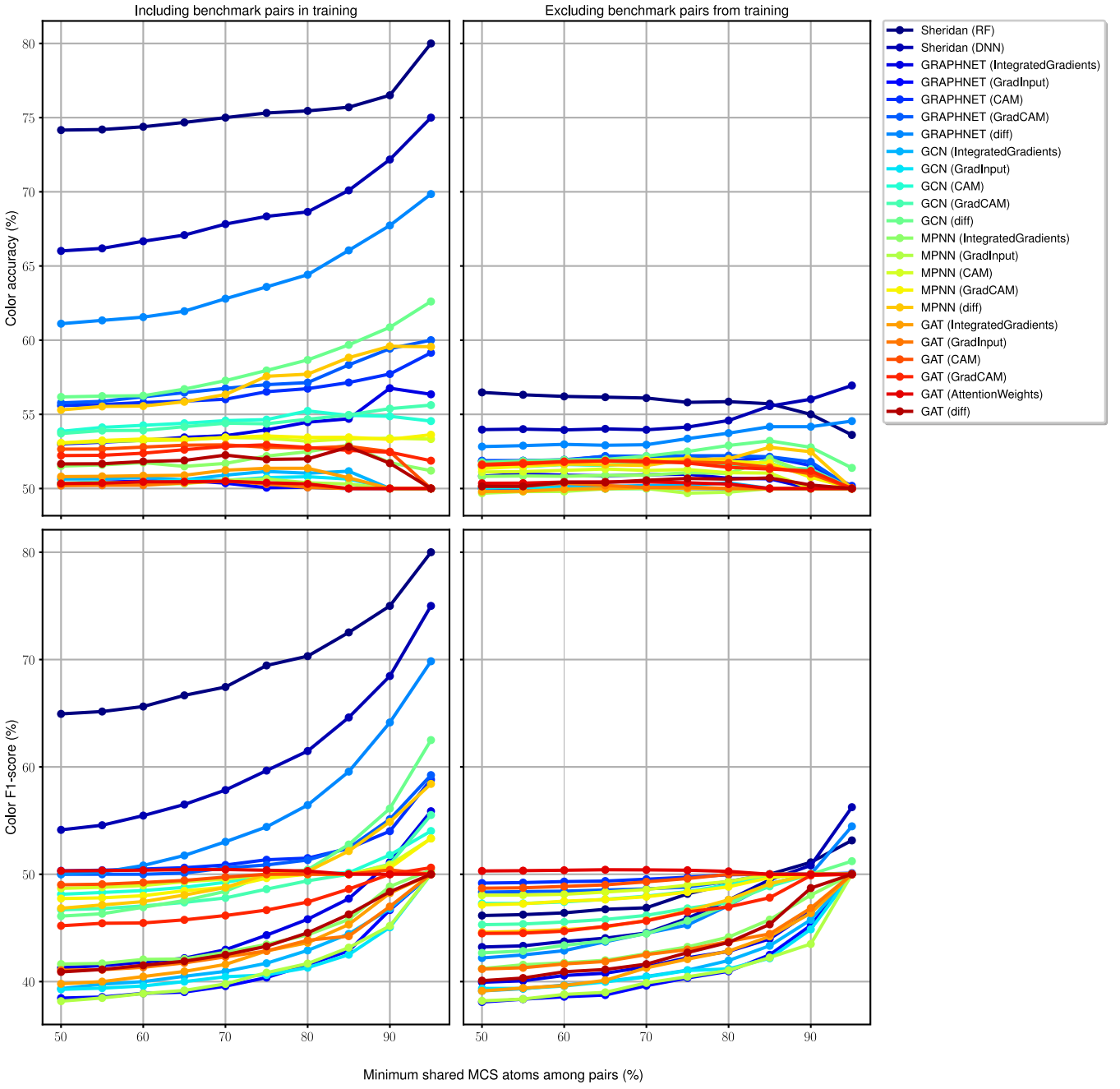


Figure 5. Median color accuracy and median F_1 scores (geometric average of precision and recall) at different MCS thresholds of common atom percentages for all the molecular pairs considered in the benchmark.

This is usually the case in small-to-medium sample-sized scenarios, which are arguably commonplace in hit-to-lead or lead optimization campaigns.

Molecular Coloring Benchmark. The first results of the proposed benchmark can be seen in Figure 5, where color accuracy and F_1 scores are computed only taking into account the sign of the computed atomic attribution and ignoring its magnitude. To evaluate the generalization capabilities of the different XAI approaches, we report results for both the case where benchmark ligands are present during the training stage and after removing them. To investigate whether the size of substructural change between pairs has an effect on the performance of the different feature attribution methods, the proposed benchmark was further studied at different percentage thresholds of shared atoms between the considered pairs.

Looking at the case where benchmark ligands had not been removed from the training sets, it is a surprising result that the best performing molecular feature attribution approach, by a considerable margin, is the one proposed by Sheridan,[20] particularly in combination with an underlying random forest model and ECFP4 fingerprints. The next best-performing approach is a feed-forward neural network model using an identical featurization schema. The rest of the deep-learning-based approaches (Integrated Gradients, GradInput, CAM, GradCAM, diff) fall significantly behind these two, with performances only marginally better than random color assignment (i.e., 50% accuracy). With a margin of five absolute

https://doi.org/10.1021/acs.jcim.1c01163
J. Chem. Inf. Model. 2022, 62, 274−283

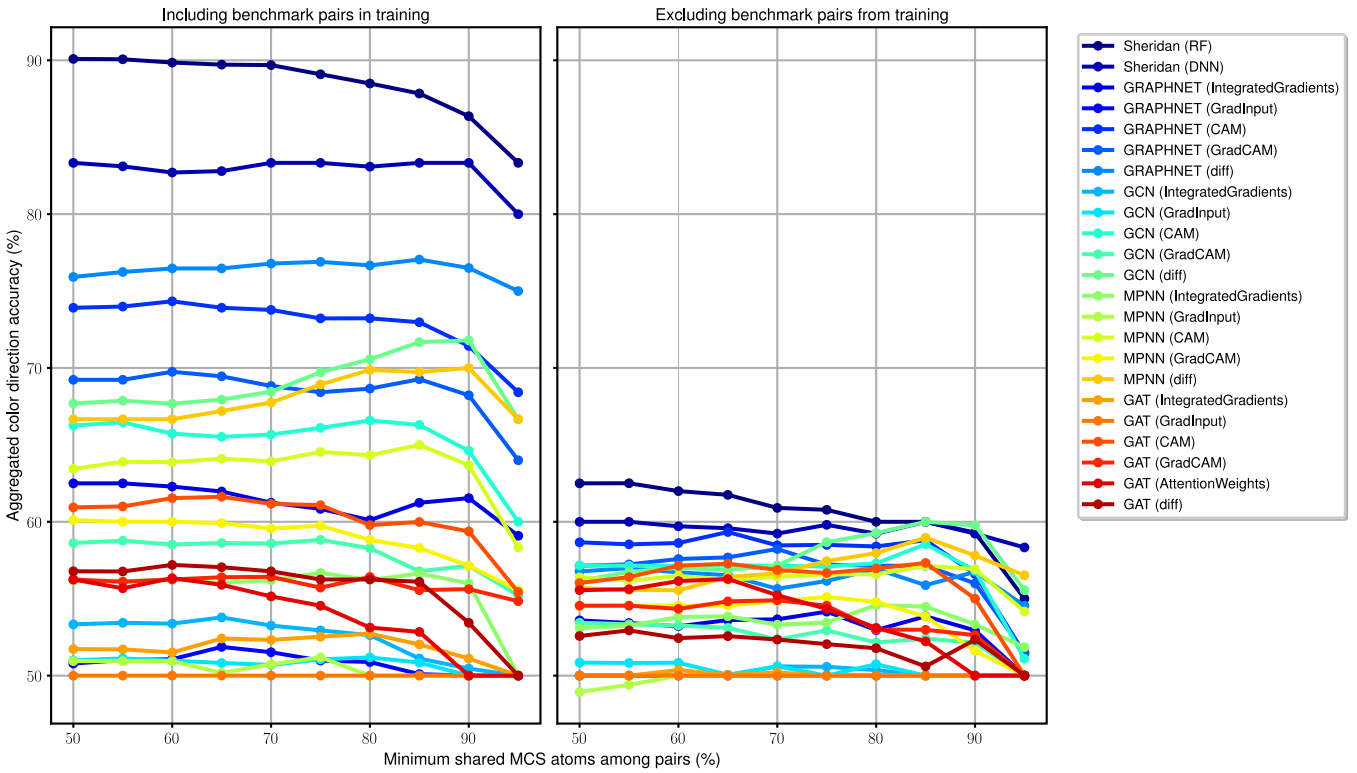


Figure 6. Median aggregated color direction accuracy scores at different MCS thresholds of common atom percentages for all the molecular pairs considered in the benchmark.

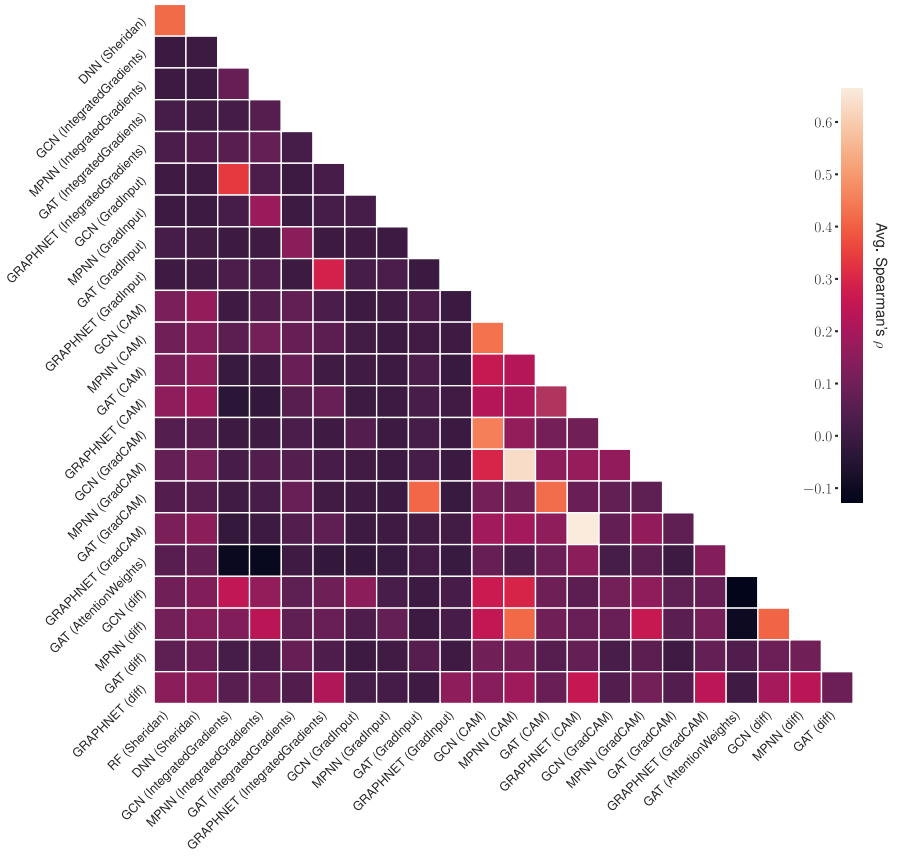


Figure 7. Average Spearman's ρ coefficient among the colors produced by all the feature attribution methods considered in this benchmark.

accuracy percent points, with respect to the second best-performing GNN-based method (considering shared pairs with at least 50% common atoms, as determined by their computed MCS), the exception is marked by the simple masking method

https://doi.org/10.1021/acs.jcim.1c01163
J. Chem. Inf. Model. 2022, 62, 274−283

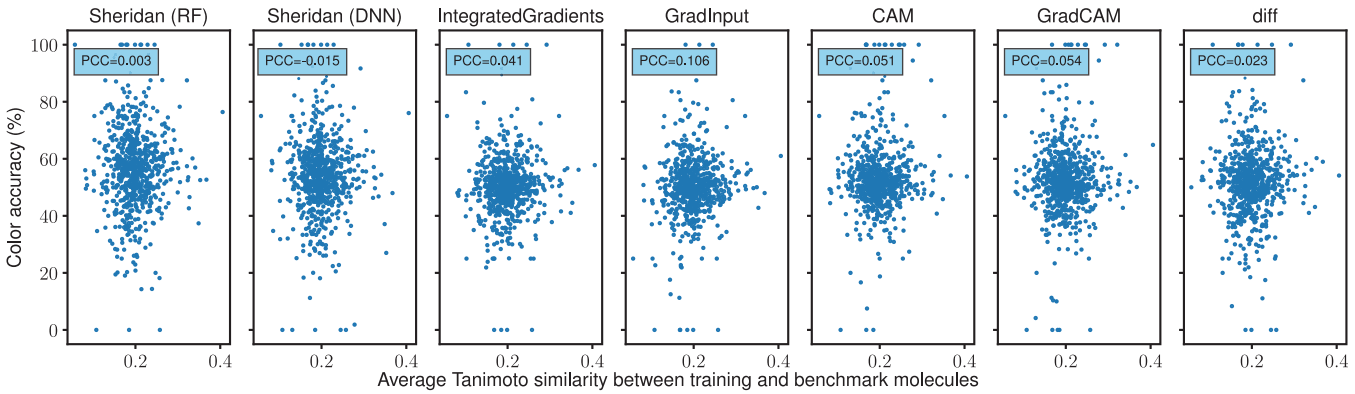


Figure 8. Scatter plots portraying the average Tanimoto similarity—calculated via ECFP4 fingerprints—and color accuracy in each of the considered benchmark sets in this study. Results presented for the masking approach proposed by Sheridan[20] and the GNN-based feature attribution methods for the GCN block.

(i.e., diff) in combination with the GraphNet block type. Similar results can be drawn from the F_1 score plot, although in this case the latter method struggles to produce scores noticeably higher than 50% for pairs of compounds whose percentage amount of common atoms is below 70%.

Conclusions drawn from the benchmark where pairs of compounds had been removed if present in their corresponding training sets are dramatically different. None of the proposed methods, including the two best-performing ones under the previous case, manage to surpass the 60% accuracy level line, and they only surpass the 50% F_1 score line in cases of minor ligand structural differences (common atom share higher than 90%, as per determined by their MCS). In particular, eight of the considered combinations even failed to produce accuracies significantly higher (one-sided t-test, α = 0.01) than random color assignment at a threshold of 50% MCS shared atoms, such as the GradInput or the Integrated Gradients approach in combination with the GCN, GraphNet, and GAT block types.

Similar results can be found using the alternative evaluation metric that avoids color inconsistencies if the same compound is present on different pairs with opposite activity difference signs (Figure 6, metric here named as aggregated color direction accuracy). When benchmark pairs are included in the training sets, masking-based methods achieve accuracies of 90% and over 80% at the 50% MCS threshold when using the random forest and the fully connected neural network as underlying models. However, as with the previous metric, these numbers dramatically decrease to barely over 60% accuracy for the two previously discussed methods when benchmark pairs are excluded from training. An interesting phenomenon is that, contrarily to the previous analysis, performance seems to decrease with a higher percentage of common atoms between benchmark pairs. We hypothesize that pairs of compounds sharing a higher quantity of atoms are also the most "cliffy" and whose overall activity difference sign may be easier to be picked up by the first analysis but whose attributions are overall harder to rank in the context of a congeneric series sorted by activity.

Overall these results seem to suggest that the studied underlying ML models struggle at true mechanistic generalization (i.e., the so-called Clever Hans effect[46]) and can only provide meaningful explanations either if (i) the compound had been previously seen during training—a fact that is fundamentally at odds with satisfactory predictive performance as evaluated on their specific test sets—or (ii) that current XAI techniques are unable to capture if the underlying ML models are learning activity cliffs. This second phenomenon could be exacerbated by the fact that activity cliffs are notoriously hard to predict, as previous studies have shown.[47] Expectedly, we observe a noticeable drop in the correlation between predicted activity differences and experimental ones when removing benchmark pairs from the training sets (Table S1).

Color Agreement and Influence of Other Variables. We first assessed to what degree the different molecular coloring approaches display any degree of agreement, as methods with little color correlation could be interpreted as orthogonal and could potentially provide different interpretations to specific property predictions. Toward that end, in Figure 7, we present the average Spearman's rank correlation coefficient ρ for all coloring methods considered in the benchmark. While most approaches show a low degree of correlation, and even negative in some cases, the two fingerprint-based approaches—which coincidentally were the ones that fared better in the presented benchmark—display a moderate agreement (ρ = 0.41). The Class Activation Map family of methods (i.e., CAM and GradCAM), despite their poor performance, also show a considerable degree of agreement ($\rho \simeq 0.5$ for some of the combinations) both within themselves and across different GNN block types (e.g., GCN, MPNN, GraphNet, GAT). Another interesting agreement with a similar level of correlation is the one found between the masking family of methods (i.e., diff) for the GCN and MPNN block types, although these are correlated to a lesser extent ($\rho \simeq 0.3$) with the best-performing GNN block type, namely, GraphNet.

We further studied whether other factors could be used to forecast the color accuracy of the different feature attribution methods. In particular, we evaluated whether the molecular similarity between the training and benchmark series, number of training examples, number of different ChEMBL assays in each of the training sets, and predictive performance on held-out data had an influence on attribution performance, as measured by the color accuracy metrics reported in the previous section. In Figure 8, we present results about the influence of chemical similarity on color accuracy, as measured by the Tanimoto coefficient between ECFP4 fingerprints corresponding to pairs of molecules extracted from the train and benchmark sets, for the fingeprint-based methods, and the ones used with the GCN block type. With PCC values

https://doi.org/10.1021/acs.jcim.1c01163
J. Chem. Inf. Model. 2022, 62, 274−283

between these two variables not noticeably higher than zero, we conclude that molecular similarity has no relevant influence on the quality of the colorings produced by any of the considered feature attribution methods. However, in this specific case, it is worth noting that all similarity values fall within the range of what can be considered chance similarity for ECFP4 fingerprints. Similar conclusions can be drawn from the rest of the block types, variables considered, and other color accuracy metrics (Figures S1−S40).

DISCUSSION

In this work we proposed a simple benchmark based on the MCS of closely related compounds featuring property cliffs to evaluate the performance of the molecular colorings produced by feature attribution models. We believe that this benchmark represents a realistic test bed that covers a wide range of compound series and is closely related to how medicinal chemists tend to think about chemical structures and how leads are typically optimized.[48,49] In molecular machine learning, activity cliffs were in the past often falsely discarded as outliers but are increasingly considered as particularly informative and relevant to assess the performance of a model.[32,50] However, one of the main limitations of this study is related to the underlying assumption that activity differences can always be attributed to structural changes. In particular, such differences can also be caused by unspecific global molecular interactions,[51−53] which we do not consider here. Additionally, while in this study we have mainly focused on different flavors of the message-passing learning framework, other methodologies, such as those making use of text-based (e.g., SMILES) molecular representations, namely, recurrent neural networks[54] and Transformers,[55] could potentially be also evaluated in combination with the hereby-tested feature attribution techniques or other ones.

We have furthermore shown that modern graph-based deep-learning methods, while able to correctly identify simple chemical motifs when trained on synthetically generated data sets, struggle to correctly highlight those in real-world lead optimization data sets, even when present in the training set. This work, at some level, further highlights the importance of testing simple baselines when evaluating newly developed approaches in molecular machine-learning research, as these were among the most performant ones. However, these conclusions need to be taken with caution, as no combination of XAI method and underlying ML model was able to successfully color previously unseen pairs of molecules in a consistent manner.

In general, based on the results obtained in this study, we discourage the overall use of current feature attribution methods in prospective lead optimization applications and particularly those that work in combination with message-passing neural networks. While some methods displayed agreement with ground-truth colorings, these were only under scenarios where the colored ligands were present in the training sets. While comparatively simpler ML models, such as random forests or fully connected networks, had shown the best results overall, if graph neural networks are to be used, the only two combinations that showed noticeably more informative results than its peers is the GraphNet block type with either the CAM technique or simple masking-based approaches. These conclusions, however, do not imply that current feature attribution methods cannot be used for other tasks within the drug discovery pipeline. In particular, these have been proven useful in the context of structural alert identification[56] (e.g., toxic moieties[57]). Alternatively, the results suggest that the machine-learning models, if providing accurate predictions, could still offer an orthogonal view to how medicinal chemists typically think of chemical structures (e.g., in terms of small changes to a common core).

Overall, and given that the results shown here indicate that there is ample room for methodological improvement (particularly regarding coloring generalization), we hope that the provided benchmark can serve as a new and more realistic starting point to evaluate explainable artificial intelligence techniques in the context of predictive molecular machine learning.

DATA AND SOFTWARE AVAILABILITY

All the results presented in this study can be reproduced with the accompanying AGPLv3-licensed code repository (https://github.com/josejimenezluna/xaibench_tf). In order to encourage the future development and testing of future molecular feature attribution methods, all pairs of compounds with their respective assigned colors are available as a compressed tarball. Instructions to download these, as well as all trained models and data sets, are also provided in the repository. Additional guidelines on how to use the proposed benchmark to test new methods are additionally included.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c01163.

Influence of variables such as molecular similarity between training and benchmark sets, training set size, number of unique ChEMBL assay identifiers per training set, and out-of-fold performance on color agreement for all model combinations (PDF)

AUTHOR INFORMATION

Corresponding Author

José Jiménez-Luna − *Department of Chemistry and Applied Biosciences, RETHINK, ETH Zurich, 8093 Zurich, Switzerland; Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, 88397 Biberach an der Riss, Germany;* orcid.org/0000-0002-5335-7834; Email: joluna@ethz.ch

Authors

Miha Skalic − *Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, 88397 Biberach an der Riss, Germany*

Nils Weskamp − *Department of Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co. KG, 88397 Biberach an der Riss, Germany*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c01163

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

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.

REFERENCES

(1) Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, 2016; Vol. *1*.

(2) Schmidhuber, J. Deep Learning in Neural Networks: An Overview. *Neural Netw.* **2015**, *61*, 85−117.

(3) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *International Conference on Machine Learning*, 2017; pp 1263−1272.

(4) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv Preprint*, arXiv:1509.09292, 2015.

(5) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discovery Today* **2018**, *23*, 1241−1250.

(6) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. *International Conference on Machine Learning*, 2018; pp 2323−2332.

(7) Jin, W.; Yang, K.; Barzilay, R.; Jaakkola, T. Learning Multimodal Graph-to-Graph Translation for Molecular Optimization. *arXiv Preprint*, arXiv:1812.01070, 2018.

(8) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10*, 370−377.

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

(10) Adadi, A.; Berrada, M. Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138−52160.

(11) Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18.

(12) Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; Müller, K.-R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Springer Nature, 2019; Vol. *11700*.

(13) 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.

(14) 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.

(15) 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.

(16) 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.

(17) Matveieva, M.; Polishchuk, P. Benchmarks for Interpretation of QSAR models. *J. Cheminformatics* **2021**, *13*, 1−20.

(18) Jiménez-Luna, J.; Skalic, M.; Weskamp, N.; Schneider, G. Coloring Molecules with Explainable Artificial Intelligence for Preclinical Relevance Assessment. *J. Chem. Inf. Model.* **2021**, *61*, 1083−1094.

(19) Sanchez-Lengeling, B.; Wei, J.; Lee, B.; Reif, E.; Wang, P.; Qian, W.; McCloskey, K.; Colwell, L.; Wiltschko, A. Evaluating Attribution for Graph Neural Networks. In *Advances in Neural Information Processing Systems*, 2020; pp 5898−5910.

(20) 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.

(21) Marcou, G.; Horvath, D.; Solov'Ev, V.; Arrault, A.; Vayer, P.; Varnek, A. Interpretability of SAR/QSAR Models of any Complexity by Atomic Contributions. *Mol. Inform.* **2012**, *31*, 639−642.

(22) Riniker, S.; Landrum, G. A. Similarity Maps-A Visualization Strategy for Molecular Fingerprints and Machine-Learning Methods. *J. Cheminformatics* **2013**, *5*, 1−7.

(23) Rosenbaum, L.; Hinselmann, G.; Jahn, A.; Zell, A. Interpreting Linear Support Vector Machine Models with Heat Map Molecule Coloring. *J. Cheminformatics* **2011**, *3*, 1−12.

(24) Van De Waterbeemd, H.; Gifford, E. ADMET in Silico Modelling: Towards Prediction Paradise? *Nat. Rev. Drug Discovery* **2003**, *2*, 192−204.

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

(26) Gilson, M. K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045−D1053.

(27) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: AWeb-accessible Database of Experimentally−Determined Protein−Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198−D201.

(28) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(29) UniProt Consortium.. UniProt: AHub for Protein Information. *Nucleic Acids Res.* **2015**, *43*, D204−D212.

(30) Dalke, A.; Hastings, J. FMCS: ANovel Algorithm for the Multiple MCS Problem. *J. Cheminformatics* **2013**, *5*, 1−1.

(31) Landrum, G. *RDKit Documentation*, Release 1, 1−79, 2013.

(32) Stumpfe, D.; Hu, H.; Bajorath, J. Evolving Concept of Activity Cliffs. *ACS Omega* **2019**, *4*, 14360−14368.

(33) Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V. F.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; Gülçehre, Ç.; Song, H. F.; Ballard, A. J.; Gilmer, J.; Dahl, G. E.; Vaswani, A.; Allen, K. R.; Nash, C.; Langston, V.; Dyer, C.; Heess, N.; Wierstra, D.; Kohli, P.; Botvinick, M.; Vinyals, O.; Li, Y.; Pascanu, R. Relational Inductive Biases, Deep Learning, And Graph Networks. *arXiv Preprints*, arXiv:1806.01261, 2018.

(34) Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y. Graph Attention Networks. *arXiv Preprint*, arXiv:1710.10903, 2017.

(35) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(36) Shrikumar, A.; Greenside, P.; Shcherbina, A.; Kundaje, A. Not Just a Black Box: Learning Important Features through Propagating Activation Differences. *arXiv Preprint*, arXiv:1605.01713, 2016.

(37) Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016; pp 2921−2929.

(38) Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017; pp 618−626.

(39) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In *International Conference on Machine Learning*, 2017; pp 3319−3328.

(40) James, C. A. *Daylight Theory Manual*; Daylight Chemical Information Systems, Inc., 2004.

(41) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure−Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353−2360.

(42) Wu, Z.; Zhu, M.; Kang, Y.; Leung, E. L.-H.; Lei, T.; Shen, C.; Jiang, D.; Wang, Z.; Cao, D.; Hou, T. Do We Need Different Machine Learning Algorithms for QSAR Modeling? A Comprehensive

https://doi.org/10.1021/acs.jcim.1c01163
J. Chem. Inf. Model. 2022, 62, 274−283

Assessment of 16 Machine Learning Algorithms on 14 QSAR Data Sets. *Brief. Bioinformatics* **2021**, *22*, bbaa321.
(43) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189−1232.
(44) Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016; pp 785−794.
(45) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*, 2017; Vol. *30*, pp 3146−3154.
(46) 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.
(47) Sheridan, R. P.; Karnachi, P.; Tudor, M.; Xu, Y.; Liaw, A.; Shah, F.; Cheng, A. C.; Joshi, E.; Glick, M.; Alvarez, J. Experimental Error, Kurtosis, Activity Cliffs, and Methodology: What Limits the Predictivity of Quantitative Structure−Activity Relationship Models? *J. Chem. Inf. Model.* **2020**, *60*, 1969−1982.
(48) Griffen, E.; Leach, A. G.; Robb, G. R.; Warner, D. J. Matched Molecular Pairs as a Medicinal Chemistry Tool. *J. Med. Chem.* **2011**, *54*, 7739−7750. PMID: 21936582.
(49) Tyrchan, C.; Evertsson, E. Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations. *Comput. Struct. Biotechnol.* **2017**, *15*, 86−90.
(50) Maggiora, G. M. On Outliers and Activity Cliffs: Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.
(51) 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.
(52) 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. Comp. Sci.* **2001**, *41*, 1308−1315.
(53) Lipinski, C. A. Lead-and Drug-like Compounds: The Rule-of-Five Revolution. *Drug Discovery Today Technol.* **2004**, *1*, 337−341.
(54) Chakravarti, S. K.; Alla, S. R. M. Descriptor−Free QSAR Modeling using Deep Learning with Long Short-Term Memory Neural Networks. *Front. Artif. Intell.* **2019**, *2*, 17.
(55) Karpov, P.; Godin, G.; Tetko, I. V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *J. Cheminformatics* **2020**, *12*, 1−12.
(56) Liu, R.; Yu, X.; Wallqvist, A. Data-Driven Identification of Structural Alerts for Mitigating the Risk of Drug-Induced Human Liver Injuries. *J. Cheminformatics* **2015**, *7*, 1−8.
(57) Sushko, I.; Salmina, E.; Potemkin, V. A.; Poda, G.; Tetko, I. V. ToxAlerts: A Web Server of Structural Alerts for Toxic Chemicals and Compounds with Potential Adverse Reactions. *J. Chem. Inf. Model.* **2012**, *52*, 2310−2316.

https://doi.org/10.1021/acs.jcim.1c01163
J. Chem. Inf. Model. 2022, 62, 274−283