Explaining compound activity predictions with a substructure-aware loss for graph neural networks

Kenza Amara,[†,‡] Raquel Rodríguez-Pérez,[*,¶] and José Jiménez-Luna[*,§]

†*Microsoft Research AI4Science, CB1 2FB Cambridge, United Kingdom*

‡*Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland*

¶*Novartis Institutes for Biomedical Research, 4002 Basel, Switzerland*

§*Microsoft Research Cambridge, CB1 2FB Cambridge, United Kingdom*

E-mail: raquel.rodriguez_perez@novartis.com; jjimenezluna@microsoft.com

Abstract

Explainable machine learning is increasingly used in drug discovery to help rationalize compound property predictions. Feature attribution techniques are popular choices to identify which molecular substructures are responsible for a predicted property change. However, established molecular feature attribution methods have so far displayed low performance for popular deep learning algorithms such as graph neural networks (GNNs), especially when compared with simpler modeling alternatives such as random forests coupled with atom masking. To mitigate this problem, a modification of the regression objective for GNNs is proposed to specifically account for common core structures between pairs of molecules. The presented approach shows higher accuracy on a recently-proposed explainability benchmark. This methodology has the potential to assist with model explainability in drug discovery pipelines, particularly in lead optimization efforts where specific chemical series are investigated.

Introduction

Drug discovery is one of the many fields where deep learning techniques have found extensive applicability in the last few years.[1] While the history behind traditional machine learning (ML) in cheminformatics can be traced as far back to the 1960s,[2,3] some recently-adopted deep learning paradigms have become increasingly popular across many tasks (*e.g.*, *de novo* molecular design, synthesis prediction). Specifically, *in silico* molecular property prediction (also commonly referred to as quantitative structure-property relationship modeling) is a central challenge in drug discovery where graph neural networks (GNNs)[4] have shown promising performance. Among the many factors that contributed to the popularity of GNNs in chemistry and other areas, we can highlight their suitability to naturally perform automatic feature extraction on arbitrarily-sized graphs and their scalability to existing commodity hardware. In chemistry, GNNs can take advantage of the natural description of molecules as graphs, where atoms and bonds can be represented as nodes and edges, respectively. Recent applications of GNN for molecular property prediction include *in vivo* brain penetration,[5] *in vitro* intrinsic clearance,[6] among others .[7–9]

However, the popularity of GNNs has also been accompanied by an increasing need for explainability,[10–18] as these models have been notoriously known for their black-box character. Towards this goal, explainable artificial intelligence techniques, such as feature attribution analyses, have become relevant tools. These analyses provide an importance value for every input feature, atom or bond in a molecular graph. Such importance values are often visualized through atom or bond coloring, where the structural patterns that drive a prediction are highlighted on top of the two-dimensional molecular representation of the compound of interest.[19]

Towards disentangling what structural patterns are exploited by GNNs in compound property predictions, a variety of feature attribution techniques have been previously reported

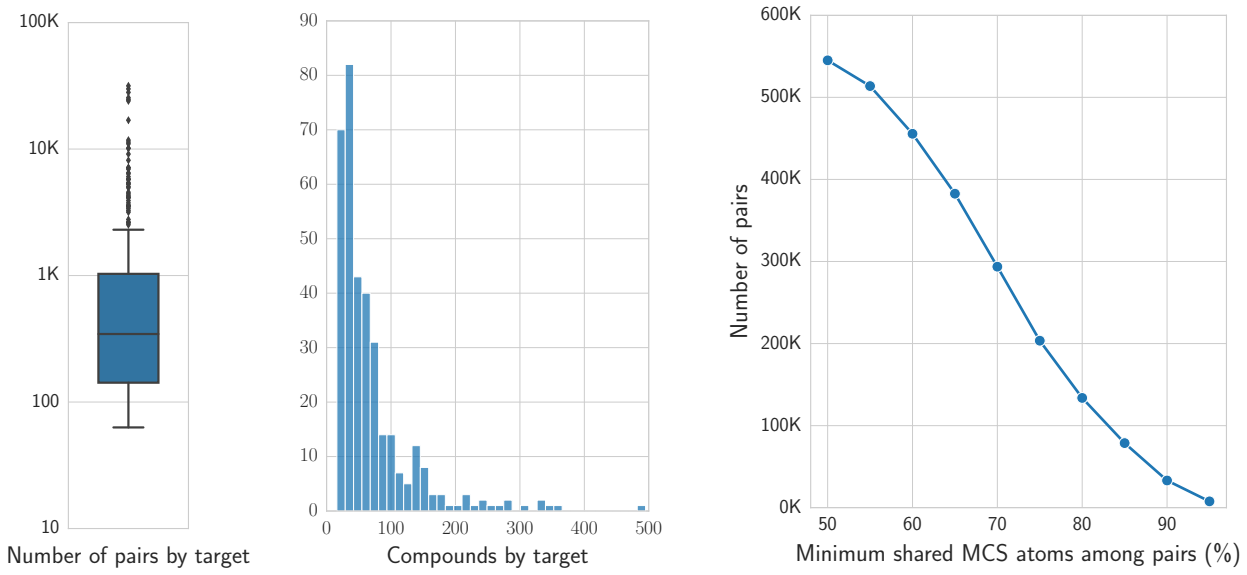
in the literature.[20] Importantly, many research efforts have focused on benchmarking feature attribution techniques, exploring their consistency and quality in atom coloring, and providing recommendations.[21–24] In particular, one such study proposed a quantitative benchmark based on publicly-available activity data for congeneric series and evaluated the performance of several GNN architectures and feature attribution techniques.[25] Therein, it was shown that GNNs did exhibit some degree of accordance with the predefined colors of the benchmark, but their explainability performance fell markedly behind simpler techniques such as atom masking[26] in combination with more traditional machine learning methods such as random forests (RF).

In order to mitigate this issue, in this paper we propose a training loss modification for GNNs that improves explainability performance on the aforementioned benchmark. Our method takes advantage of the fact that lead optimization efforts focus on specific compound series, where molecules share structural cores (*i.e.*, scaffolds). The explicit consideration of the molecular scaffold formalism can be leveraged to appropriately assign importance of the uncommon substructures responsible for a property change during model training. We show that the proposed approach is beneficial towards closing the explainability performance gap previously reported between GNNs and other classical methods. The architecture is inspired by recent work on molecular representation learning based on reaction data that explicitly encourage the similarity of reactants and reagents in embedding space.[27] To foster reproducibility, all code and data are made available through a permissive open-source license.

Materials and methods

Benchmark data

Molecular scaffolds. A scaffold is defined as the core of the molecule where one or several



(a) (b) (c)

Figure 1: Benchmark descriptive analyses. Reported are (a) the distribution of number of pairs per protein target, (b) the number of compounds per protein target, and (c) the number of compound pairs considered at varying scaffold size (different thresholds of minimum shared MCS among pairs).

functional groups can be attached. Herein, the maximum common substructure (MCS) formalism was used to define a molecular scaffold[28] between pairs of compounds binding to a specific target. To consider that two compounds share a molecular scaffold, such common part should encompass a minimum fraction of their structure. Taking this into consideration and in line with previous work, different thresholds of minimum shared substructures were examined.[25] For the development and evaluation of our methodology, MCS pairs were computed using the FMCS[29] algorithm, as available in the RDKit *rdFMCS* module.[30]

Data preparation. The benchmark data from a recently proposed study on feature attribution[25] was used, which consisted of 723 protein targets with associated small molecule activity data (half maximal inhibitory concentration, IC_{50}). The dataset was initially constructed using the BindingDB protein-ligand validation sets,[31] which contains binding affinities for a large number of targets and across different molecular scaffolds. In said data set, ground-truth atom-level feature attribution labels were determined via the concept of

activity cliffs.[32,33] Specifically, these were defined as pairs of compounds in one or multiple congeneric series sharing a molecular scaffold and with at least 1 log unit activity difference. Compounds for each protein target were randomly divided into training (80%) and test (20%) sets. Only protein targets with at least 50 compound pairs in the training set were kept. To avoid data leakage, the same compound was not allowed to be present in different pairs in training and test sets, resulting in a final selection of 350 protein targets. Figure 1 shows the distribution of the number of pairs and compounds per target at the minimum considered MCS threshold of 50%, as well as the number of pairs sharing molecular scaffolds at different minimum thresholds.

Models and feature attribution techniques

Models. Message-passing GNN[34] models were trained to predict compound activity against all available protein targets. In most molecular property prediction scenarios, these are models $f \in \mathcal{F}$ that map molecular graphs to real values $f : \mathcal{G}(\mathcal{V}, \mathcal{E}) \rightarrow \mathbb{R}$, with $v \in \mathcal{V}, e \in \mathcal{E}$ representing atoms and bonds, respectively. They do so by iteratively learning and updating internal node latent representations using the information from neighboring atom and bond latent spaces (for a more comprehensive description a canonical reference is provided in Gilmer *et al.*[4]). In this work GNNs were optimized to minimize at least one of the following loss functions: (i) mean squared error (MSE) between observed and predicted binding affinities (in logarithmic scale), (ii) a relative affinity loss computed on pairs of related compounds, hereby referred to as activity cliff (AC) loss, and (iii) the proposed uncommon node loss (UCN). Both AC and UCN losses were considered on top of the standard MSE loss with a fixed weighting term (see *Substructure-aware loss* Section). As a control, random forest (RF) models trained with extended-connectivity fingerprints (ECFP4) were also considered. Additional details regarding neural network hyperparameters, featurization, and optimization details are provided in Section S4.

Feature attribution techniques.. In the context of this work, feature attribution metrics consist of a function that takes a molecular graph as well as a trained property model and produces a real number (*i.e.*, a coloring) for each atom in the graph representing their importance in the prediction. $e : (\mathcal{G}, \mathcal{F}) \to \mathbb{R}^{\mathcal{V}}$. A variety of feature attribution methods that enable the estimation of positive and negative atom contributions were investigated. Class Activation Maps (CAM)[35] and gradient-based methods, namely GradInput,[36] Integrated Gradients,[37] and Grad-CAM[38] were utilized. Additionally, other perturbation-based approaches such as node masking, where the contribution of each atom is determined as the difference in prediction upon its artificial modification, were considered. For the presented GNN models, node masking iteratively set node features to zero. For RF models, each atom was assigned an atom type that was not present in the benchmark sets, and molecular features re-calculated.[26] Section S5 reports additional details on the hyperparameters used in each feature attribution technique.

Substructure-aware loss

A supervised learning problem was considered where a GNN model was trained to predict compound activity against a specific protein target. Motivated by the fact that several drug discovery efforts tend to focus on congeneric series (*e.g.,* lead optimization), we propose a loss that focuses on the uncommon structural motifs between ligand pairs. A schematic representation of this procedure is provided in Figure 2. During training, compound pairs with a common scaffold are sampled and the difference in predicted activity is attributed to the uncommon node latent spaces. For each pair k of compounds i, j, with corresponding molecular graphs $c_i, c_j \in C$ and experimental activities $y_i, y_j \in \mathbb{R}$, the proposed uncommon node loss is computed as:

$$\mathcal{L}_{\text{UCN}}\left(c_i, c_j, k\right) := \left\| \left(\xi\left(\phi\left(M_i^k\left(\boldsymbol{h}_i\right)\right)\right) - \xi\left(\phi\left(M_j^k\left(\boldsymbol{h}_j\right)\right)\right)\right) - \left(y_i - y_j\right) \right\|^2, \tag{1}$$

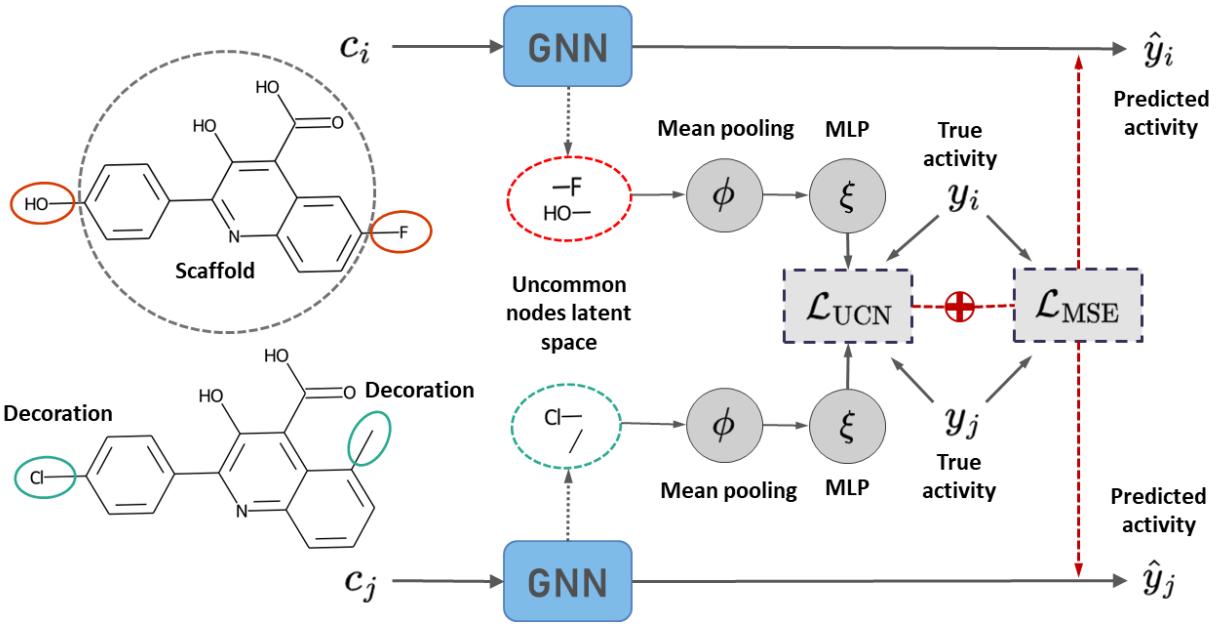


Figure 2: Schema of the proposed UCN loss. Two compounds sharing a scaffold are sampled from the training set, and their atom latent spaces computed via a forward pass of a GNN model. The uncommon latent nodes are used for the loss computation, targeting the activity difference between the compound pairs. In the illustrated example, the compound pair is composed by c_i and c_j, with a large MCS and two substitution sites, highlighted in red for c_i and green for c_j. Substituents (or decorations) differ for both compounds, and correspond to the uncommon nodes in the latent space.

where $\boldsymbol{h}_i \in \mathbb{R}^{N_i \times d}$ is the latent node representation of compound c_i, $M_i^k : \mathbb{R}^{N_i \times d} \to \mathbb{R}^{n_i \times d}$ is a masking function over nodes that retrieves those uncommon for compound i in the context of pair k, $\phi : \mathbb{R}^{n \times d} \to \mathbb{R}^d$ is a mean readout function over nodes, $\xi : \mathbb{R}^d \to \mathbb{R}$ is a multilayer perceptron with linear activation, and $\|\cdot\|$ is the vector Frobenius norm. During model training, the UCN term was used alongside of a standard mean squared error (MSE) loss on the absolute predicted versus experimental binding affinities of pair k:

$$\mathcal{L}_{\text{MSE}}\left(c_i, c_j\right) := \left\|y_i - \hat{y}_i\right\|^2 + \left\|y_j - \hat{y}_j\right\|^2, \tag{2}$$

where $\hat{y}_i$ is an absolute activity prediction output that aggregates over all available nodes in each pair (*i.e.*, both common and uncommon). Since sampling compound pairs results in an

augmented data set that could artificially boost performance, additional models were trained to minimize a relative binding affinity loss:

$$\mathcal{L}_{\mathrm{AC}}\left(c_i, c_j\right) := \left\|(y_i - y_j) - (\hat{y}_i - \hat{y}_j)\right\|^2. \tag{3}$$

Specifically, the models considered in this study were trained to minimize either $\mathcal{L}_{\mathrm{MSE}}$ or one of the two combinations $\mathcal{L}_{\mathrm{MSE+AC}} := \mathcal{L}_{\mathrm{MSE}} + \lambda\mathcal{L}_{\mathrm{AC}}$, $\mathcal{L}_{\mathrm{MSE+UCN}} := \mathcal{L}_{\mathrm{MSE}} + \lambda\mathcal{L}_{\mathrm{UCN}}$. For all training and testing purposes in this study we fix $\lambda = 1$.

Evaluation metrics

Predictive performance. Regression model performance was evaluated with the root mean squared error (RMSE) and Pearson's correlation coefficient (PCC). RMSE and PCC metrics were calculated to evaluate activity prediction against individual targets. To aggregate results across all targets in the data set, both the unweighted (simple) and weighted average values were calculated. For the weighted average calculation, RMSE or PCC values were weighted by the number of compounds pairs in the test set of each target.

Explainability. The performance of the feature attribution methods was evaluated using *global direction* and *atom-level accuracy* metrics.[25] Global direction is a binary metric assessing whether average feature attribution across the uncommon nodes in a pair k of compounds preserves the direction of the activity difference. Assuming $\psi : C \to \mathbb{R}^{N\times d}$ is a feature attribution function that assigns a score to each node feature in an input graph, the metric for a single pair is computed as:

$$g_{\mathrm{dir}}\left(c_i, c_j\right) = \mathbb{1}\left[\mathrm{sign}\left(\Phi\left(M_i^k\left(\psi\left(c_i\right)\right)\right) - \Phi\left(M_j^k\left(\psi\left(c_j\right)\right)\right)\right) = \mathrm{sign}\left(y_i - y_j\right)\right], \tag{4}$$

where $\Phi : \mathbb{R}^{N\times d} \to \mathbb{R}$ is a mean aggregator over nodes and features. The score is averaged over all pairs in the benchmark test sets.

Atom-level accuracy, also hereby referred to as *color agreement*, measures whether the

feature attribution assigned to a node has the same sign as the experimental activity difference of the compound pair (ground truth). In previous work, ground-truth atom attribution labels were obtained by assuming that the structural changes between a pair of compounds were responsible for the observed potency changes.[25] Therefore, structural parts in the most potent compound of the pair were assigned a positive feature attribution, and vice versa. For every atom in a compound with corresponding molecular graph c_i with m_i common atoms in pair k, and with ground truth atom color $\boldsymbol{t}_i^k \in \{-1, 1\}^{m_i}$, the (vector-valued) metric is defined as:

$$g_{\text{atom}}(c_i) := \mathbb{1}_{m_i} \left[\text{sign} \left(\eta \left(M_i^k \left(\psi \left(c_i \right) \right) \right) \right) = \boldsymbol{t}_i^k \right], \tag{5}$$

where $\eta : C \to \mathbb{R}^N$ is a mean aggregation function over features and $\mathbb{1}_{m_i}$ is an indicator vector with m_i binary entries. The mean value $\bar{g}_{\text{atom}}$ is then used as a summary of the color accuracy for compound c_i.

Jiménez-Luna *et al.*[25] noted that the ground-truth colors assigned by g_{atom} can be ill-defined for a compound, since they are dependent on the other compound in the pair (*i.e.,* the assigned colors to one compound could either be positive or negative depending on the specific comparison). In contrast, g_{dir} does not suffer from this problem. For this reason, the analyses reported here focus on the g_{dir} evaluation metric and, for completeness, g_{atom} results are reported in the Supporting Information.

Results and discussion

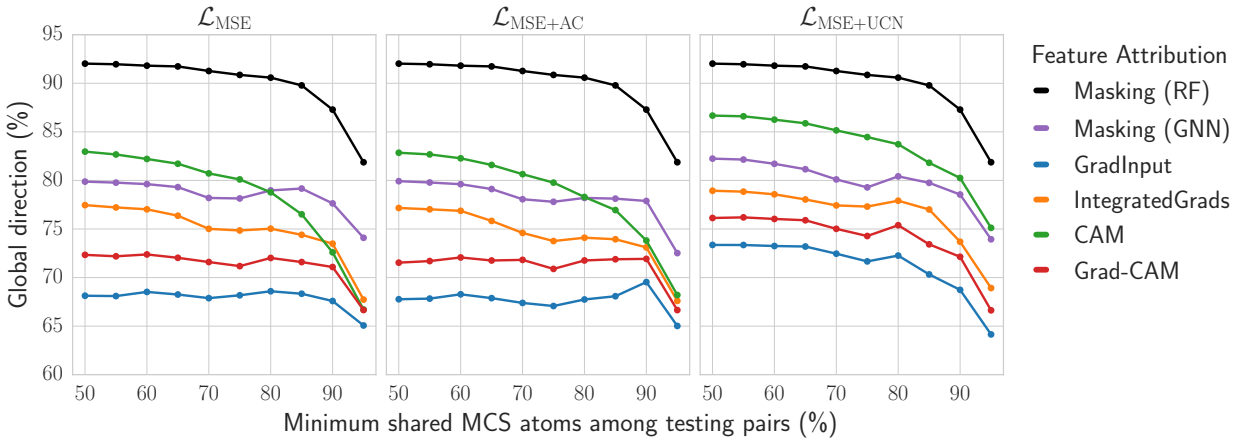
ML models were generated to predict compound potency against 350 protein targets. Message-passing GNNs were trained to minimize different loss functions, including the standard MSE loss, and its linear combination with relative (AC), or uncommon node (UCN) losses. Moreover, RF models were built for comparison. First, prediction performance was assessed for all GNN and RF models. Next, model explainability was benchmarked and the influence of the UCN loss analyzed for individual targets.

Table 1: Test set predictive performance. Reported are the average (Avg.) and weighted average (W. Avg., over number of compounds per target) of root mean squared error (RMSE) and Pearson's correlation coefficient (PCC) values (± 1 standard deviation).

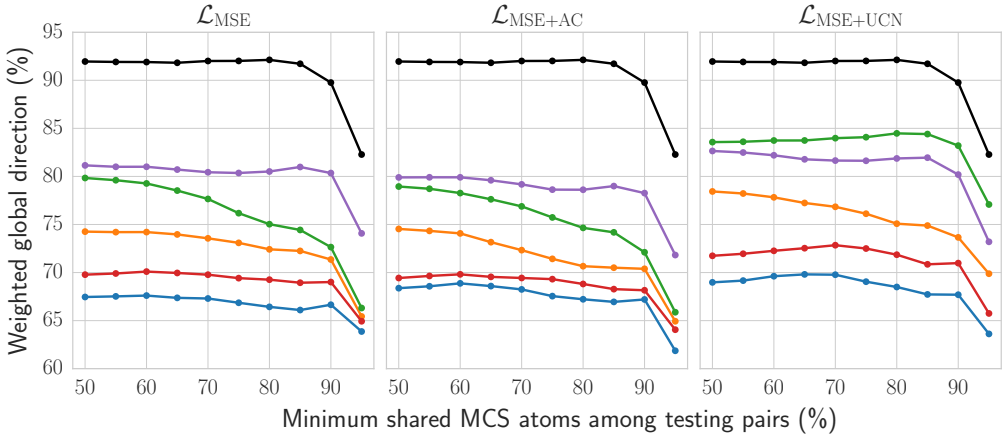
	Avg. RMSE	**W. Avg. RMSE**	**Avg. PCC**	**W. Avg. PCC**
RF	0.32 (±0.11)	0.30 (±0.08)	0.95 (±0.07)	0.96 (±0.04)
GNN $\mathcal{L}_{\text{MSE}}$	0.34 (±0.26)	0.24 (±0.14)	0.89 (±0.23)	0.96 (±0.08)
GNN $\mathcal{L}_{\text{MSE+AC}}$	0.34 (±0.27)	0.23 (±0.11)	0.89 (±0.23)	0.97 (±0.08)
GNN $\mathcal{L}_{\text{MSE+UCN}}$	0.43 (±0.29)	0.30 (±0.14)	0.86 (±0.24)	0.95 (±0.09)

Predictive performance.

There is a known trade-off between model interpretability and accuracy.[39] Moreover, only explanations from well-performing methods can be used to assist in the interpretation of predictions, and thus drug design. Therefore, prediction performance was evaluated for all GNN and RF models. Table 1 reports the simple and weighted average values for root mean squared error (RMSE) and Pearson's correlation coefficient (PCC) metrics. Results are shown for GNNs built with different loss functions, *i.e.,* solely MSE loss ($\mathcal{L}_{\text{MSE}}$), MSE in combination with AC ($\mathcal{L}_{\text{MSE+AC}}$) or UCN losses ($\mathcal{L}_{\text{MSE+UCN}}$), and RF. Average RMSE values across all targets ranged from 0.32 (RF) to 0.43 (GNN with $\mathcal{L}_{\text{MSE+UCN}}$). Average correlation between predicted and experimental potency values ranged from 0.86 (GNN with $\mathcal{L}_{\text{MSE+UCN}}$) to 0.95 (RF). Importantly, performance differences between methods were smaller when considering a weighted average across targets. In such case, per-target performance results are weighted by the number of compound pairs in each test set. The smallest and largest weighted average RMSE were 0.23 (GNN with $\mathcal{L}_{\text{MSE+AC}}$)and 0.30 (RF and GNN with $\mathcal{L}_{\text{MSE+UCN}}$), respectively. In addition, weighted average correlation values were between 0.95 (GNN with $\mathcal{L}_{\text{MSE+UCN}}$) and 0.97 (GNN with $\mathcal{L}_{\text{MSE+AC}}$). Only minor differences favouring the simpler $\mathcal{L}_{\text{MSE}}$ loss for both RMSE and PCC values were observed, with most results lying within one standard deviation of each other.



(a)



(b)

Figure 3: Global direction at varying scaffold size and across feature attribution methods. (a) Global direction and (b) weighted global direction values are reported at different thresholds of minimum shared MACS among testing pairs (%). In (b), global direction is weighted by the number of pairs per each target. Results are shown for three loss functions, *i.e.* $\mathcal{L}_{\text{MSE}}$ (left panel), $\mathcal{L}_{\text{MSE+AC}}$ (middle panel), and $\mathcal{L}_{\text{MSE+UCN}}$ (right panel). Colors report different feature attribution methods, five for GNN models and atom masking for RF models. Since the three losses functions are only applied to GNN models, RF results are equivalent in the three panels.

Explainability evaluation at varying scaffold size.

Explainability was primarily evaluated using the global direction score, which focuses on the uncommon decorations in a pair of compounds and assesses whether the direction of the activity difference is preserved. Global direction values were calculated at varying MCS thresholds among compound pairs. Figure 3 shows the global direction values for all test

pairs and targets considered in the study. Many feature attribution methods applied to GNNs with the proposed UCN objective ($\mathcal{L}_{\text{MSE+UCN}}$) exhibited larger global direction values over the absolute MSE ($\mathcal{L}_{\text{MSE}}$) and relative MSE ($\mathcal{L}_{\text{MSE+AC}}$) losses. Improvements were observed for most methods, but were more pronounced for CAM, Grad-CAM, and GradInput. Additionally, the GNN-based masking method also exhibited a slight increase. Most importantly, this improvement held across different thresholds of minimum MCS between pairs. Figure 3b reports the results with the weighted color direction metric, where similar conclusions can be drawn. In this case, Integrated Gradients showed larger improvements compared to the non-weighted analyses. Despite the improvement of the global direction metric for GNNs using $\mathcal{L}_{\text{MSE+UCN}}$ loss, RF models with an atom masking approach achieved larger values. Among the GNN methods, the CAM and masking approaches provided top-performing global direction results. Global direction values were overall stable across different scaffold size. Only when the uncommon structural parts in compound pairs were small (>85-90% thresholds), global direction values significantly decreased for all methods.

Explainablity for individual protein targets.

In the previous section, explainability methods were benchmarked using the average global direction across all targets. Nevertheless, for specific protein targets, the best explainability method might differ. To evaluate how often this is the case, global direction with $\mathcal{L}_{\text{MSE}}$ and $\mathcal{L}_{\text{MSE+UCN}}$ loss functions were compared on a target-by-target basis (Figure 4). From this analysis, we observed that global direction values were higher for 60-66% of the targets when including the UCN loss. Additionally, most feature attribution methods showed improvements with the inclusion of the UCN loss, with CAM exhibiting the largest improvements (66%). Additional plots and analyses can be found in Sections S1 and S2, where we highlight that CAM approached performance close to the RF masking method when evaluated on the training sets. Section S3 reports results with color agreement as an alternative metric. In this case, the proposed UCN loss produced an improvement for several of the feature attribution

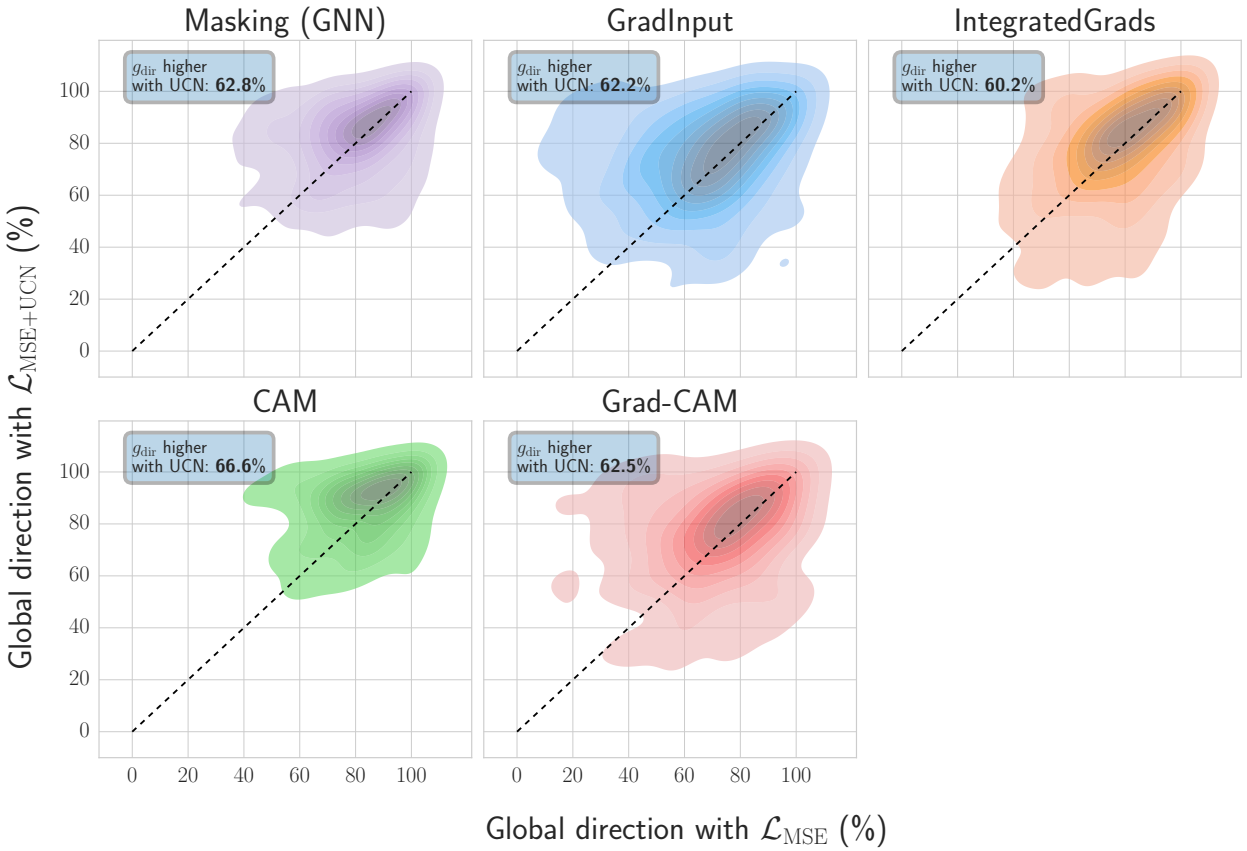


Figure 4: Per-target comparison of global direction values. The two-dimensional kernel density plot shows the target-specific global direction values with $\mathcal{L}_{\texttt{MSE}}$ (x-axis) and $\mathcal{L}_{\texttt{MSE+UCN}}$ (y-axis) loss functions. The text-box reports the percentage of protein targets for which global direction (g_{dir}) was larger with $\mathcal{L}_{\texttt{MSE+UCN}}$ loss. Compound pairs considered at the minimum 50% MCS threshold.

methods evaluated in both training and test sets, albeit the advantage was less observed to be less pronounced than for the global direction metric.

Figure 5 reports the number of targets for which the addition of the UCN loss term led to a negligible ($\leq$5%), small (between 5% and 10%), medium (between 10% and 20%), or large ($\geq$20%) global direction improvement. Results indicate that GNNs with $\mathcal{L}_{\texttt{MSE+UCN}}$ loss led to larger global direction values for the same or higher number of targets than GNNs with the standard $\mathcal{L}_{\texttt{MSE}}$ loss. Interestingly, differences across loss functions became larger when considering targets with medium to large global direction improvements in their explanations. CAM, GradInput, and Grad-CAM were the methods showing the largest benefit of UCN loss

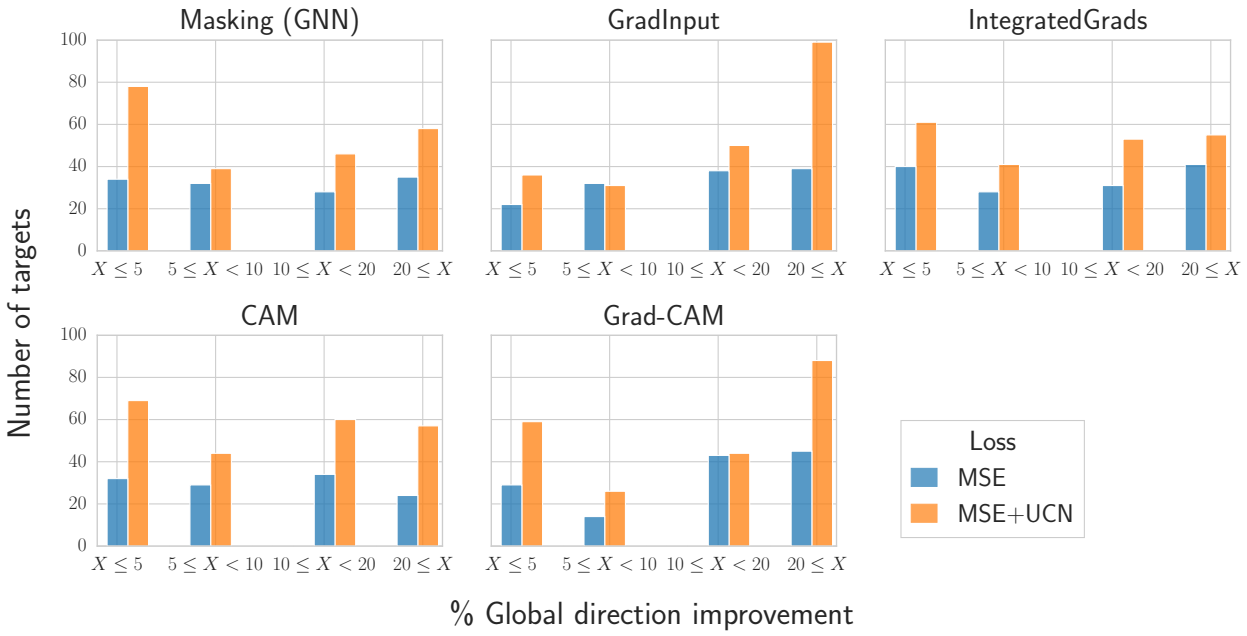


Figure 5: Protein targets with global direction improvements. Reported are the number of targets (y-axis) displaying a given improvement of the global direction metric g_{dir} using the proposed $\mathcal{L}_{\text{MSE+UCN}}$ loss compared to $\mathcal{L}_{MSE}$ (x-axis). Global direction improvements were binned into $\leq$5%, between 5 and 10%, between 10 and 20%, and $\geq$20%. Colors indicate the loss function utilized during GNN training ($\mathcal{L}_{MSE}$, blue; $\mathcal{L}_{\text{MSE+UCN}}$, orange). A minimum threshold of 50% MCS was considered for this analysis.

inclusion, with many targets showing global direction improvements higher than 20% (133 for Grad-CAM, 138 for GradInput, and 81 for CAM).

Finally, as a way of exemplifying how the proposed methodology can be used in practice, in Figure 6 we showcase two cases where the proposed loss function yields better explanations than the simpler MSE loss using the Integrated Gradients method. Specifically, we consider protein targets with associated PDB identifiers 1F0R and 1D3G, and average attributions for the uncommon scaffolds as in the color direction metric.

Potential factors influencing explainability.

As a way of elucidating which factors contribute to a successful feature attribution assignment, the benchmark was extended by evaluating whether g_{dir} is affected by (i) the number of substituent sites in the compound pair, or (ii) the chemical diversity within each

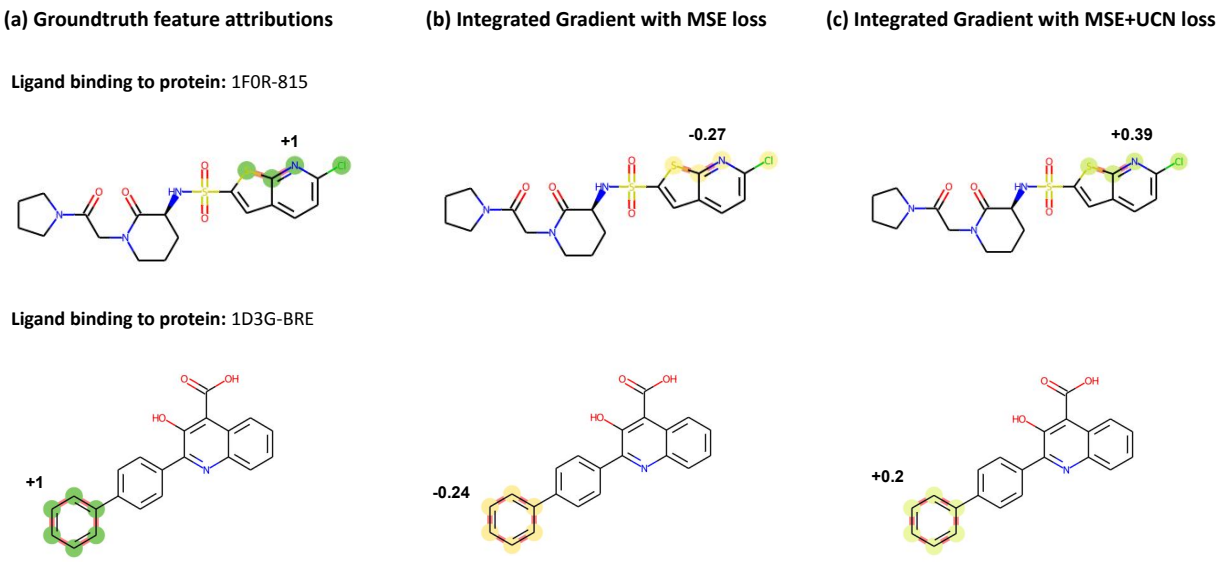


Figure 6: Two example molecules where the explanations provided by the Integrated Gradients feature attribution method benefit from the addition of the proposed UCN loss. Attributions for the uncommon motifs averaged to provide a single attribution scalar. In these two cases, the model trained with the simpler MSE loss fails to correctly capture the direction of the activity change in the respective pairs.

target. Figure 7 reports the global direction values for compound pairs that differ by one or at least two substitution sites. Results suggested that feature attribution methods did not showcase an overall higher performance for compounds pairs that differ in a single substitution site. Additionally, chemical diversity was estimated via the Bemis-Murcko scaffold [40] formalism (Figure 8). In more detail, chemical diversity was defined as the total number of scaffolds divided by the number of compounds available for each target. Apart from a slightly higher concentration of targets around areas where both the number of scaffolds is low and g_{dir} is high, no significant correlation between these values was observed.

Conclusions

In this study, we explored and quantitatively evaluated how the explainability of GNNs can be improved in the context of drug discovery. Specifically, a novel substructure-aware loss was proposed to improve GNNs' explainability for congeneric series data. This modified loss

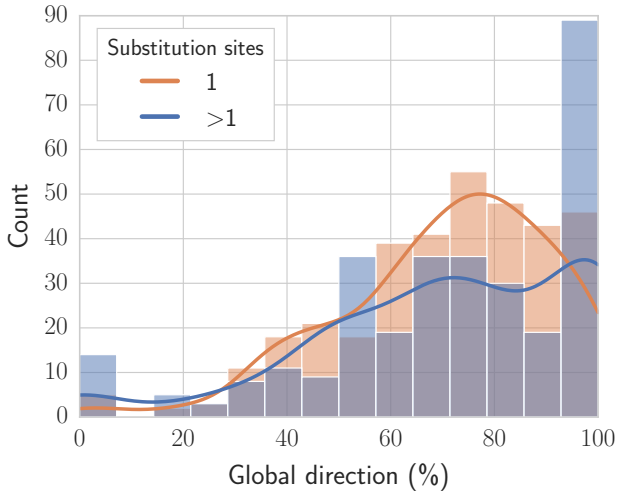


Figure 7: Effect of the number of substitution sites on global direction. Global direction (x-axis) is reported for compound pairs with a single (orange) or multiple (blue) substitution sites. For the derivation of compound pairs, a minimum 50% MCS threshold was set.

function was evaluated on a previously-reported benchmark for molecular ML explainability and it was observed that most GNN-based feature attribution techniques markedly benefited from its usage. Global direction values were used to evaluate compound explanations. Our results showed that the average global direction as well as the percentage of targets with global direction improvements were superior with the consideration of the UCN loss during GNN training. Specifically, a 66% and 63% of the targets improved global direction scores for CAM and GNN masking, respectively, which were identified as the best-performing GNN feature attribution methods. Moreover, when explaining activity predictions for a specific target protein, large global direction improvements were more likely with the newly proposed loss function. However, despite the observed superiority of the substructure-aware loss in GNN-based feature attribution methods, the RF models coupled with an atom masking approach still remained the best approach for explainability in the benchmark.[26] Nevertheless, the feature attribution performance gap between RF and GNNs was reduced with the inclusion of the proposed loss. Therefore, results on this benchmark data set support he use of the new loss function for more consistent explanations in cases where GNN is the preferred modeling approach, *e.g.* for data sets where GNNs' predictive performance is superior to RF.

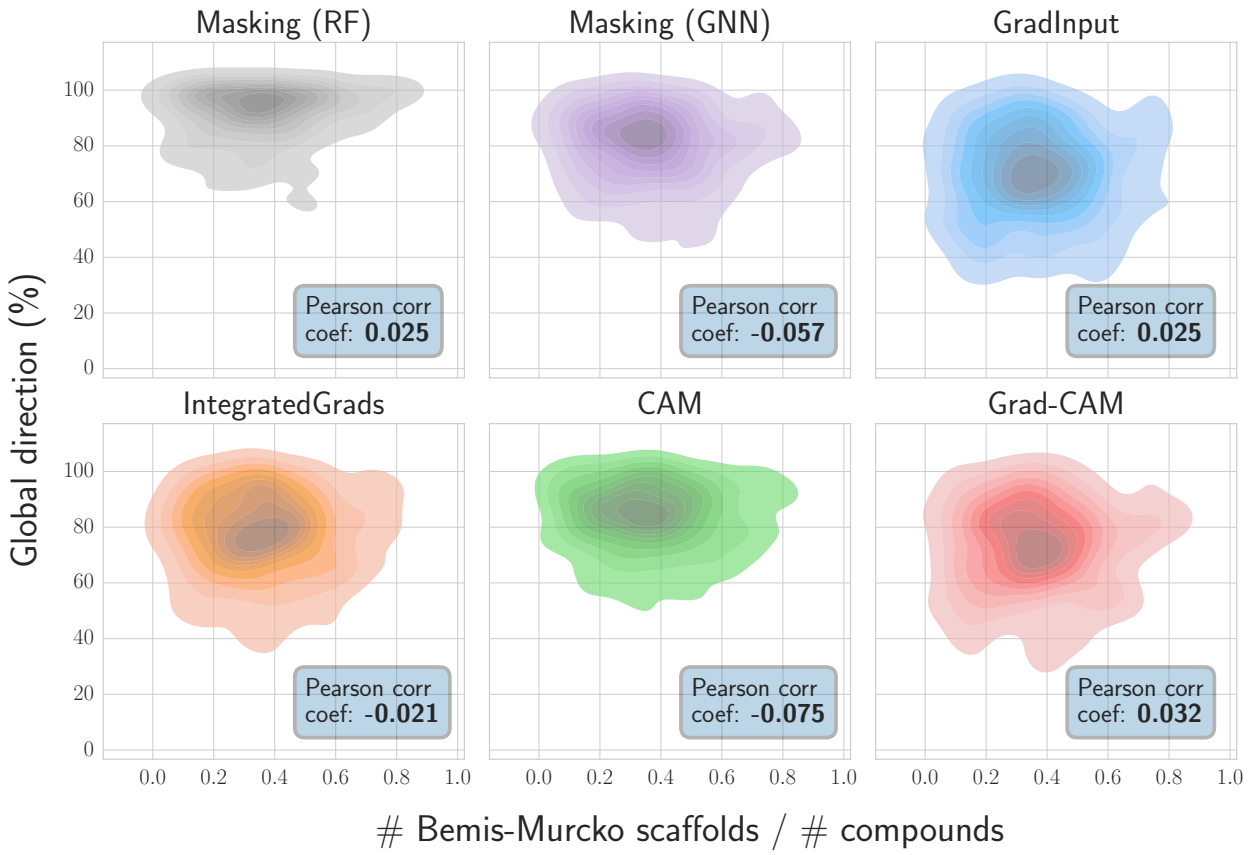


Figure 8: Effect of structural diversity on global direction. Reported are the per-target chemical diversity and global direction values per each protein target. Results reported for the minimum 50% MCS threshold.

The requirement of precomputed common substructures between pairs of compounds is admittedly a limitation of the presented method. Exact MCS algorithms are computationally expensive, but the issue may be bypassed using approximations or matched molecular pair analyses.[41,42] Moreover, we believe that feature attribution approaches may be hindered by some of the current GNN training limitations. As ventures for future research, the exploration of additional GNN architectures (*e.g.*, those that avoid the Weisfeler-Lehman graph isomorphism issue) or tackle the well-known oversmoothing effect on GNNs[43] by applying regularization,[44,45] self-supervised learning,[46,47] or pretraining techniques[48] might be promising.

All in all, a new strategy for GNN explainability was introduced, inspired by the lead optimization efforts in drug discovery, which are centered on specific chemical series. Fur-

thermore, a trade off between predictive performance and explainability was identified in the context of the proposed loss function for GNNs. Finally, we expect that the presented explainability approach will help rationalizing GNN-based model decisions in that context.

Acknowledgement

K. Amara acknowledges financial support during her internship at Microsoft Research. We also thank K. Maziarz and M. Segler for helpful discussions on this work. The authors declare no conflict of interest.

Supporting Information Available

Additional global direction results on training and test sets, color agreement metrics on all sets, neural network training hyperparameters, and feature attribution techniques settings, are reported in the Supporting Information to this manuscript.

Data and software availability

Code to replicate the results in this paper is provided in `https://github.com/microsoft/molucn`, and distributed under a permissive MIT license. All associated data, results and training logs are also provided.

References

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

(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) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR modeling: Where have you been? Where are you going to? 2014; `https://doi.org/10.1021/jm4004285`, PMID: 24351051.

(4) 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.

(5) Hamzic, S.; Lewis, R.; Desrayaud, S.; Soylu, C.; Fortunato, M.; Grégori, G.; Rodríguez-Pérez, R. Predicting In Vivo Compound Brain Penetration Using Multi-task Graph Neural Networks. *J. Chem. Inf. Model.* **2022**, *62*, 3180–3190.

(6) Rodríguez-Pérez, R.; Trunzer, M.; Schneider, N.; Faller, B.; Gerebtzoff, G. Multi-species Machine Learning Predictions of In Vitro Intrinsic Clearance with Uncertainty Quantification Analyses. *Mol. Pharm.* **2022**, *in press*.

(7) Montanari, F.; Kuhnke, L.; Laak, A. T.; Clevert, D.-A. Modeling Physico-Chemical ADMET Endpoints with Multitask Graph Convolutional Networks. *Molecules* **2020**, *25*, 44.

(8) 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.

(9) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.;

Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(10) Jiménez-Luna, J.; Grisoni, F.; Schneider, G. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.* **2020**, *2*, 573–584.

(11) Rodríguez-Pérez, R.; Bajorath, J. Explainable machine learning for property predictions in compound optimization. *J. Med. Chem.* **2021**, *64*, 17744–17752.

(12) Rodríguez-Pérez, R.; Bajorath, J. Chemistry-centric explanation of machine learning models. *Artif. Intell. Life Sci.* **2021**, *1*, 100009.

(13) Gandhi, H. A.; White, A. D. Explaining molecular properties with natural language. **2022**,

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

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

(16) Wellawatte, G. P.; Gandhi, H. A.; Seshadri, A.; White, A. D. A perspective on explanations of molecular prediction models. **2022**,

(17) Harren, T.; Matter, H.; Hessler, G.; Rarey, M.; Grebner, C. Interpretation of structure–activity relationships in real-world drug design data sets using explainable artificial intelligence. *J. Chem. Inf. Model.* **2022**, *62*, 447–462.

(18) Feldmann, C.; Bajorath, J. Calculation of exact Shapley values for support vector machines with Tanimoto kernel enables model interpretation. *Iscience* **2022**, *25*, 105023.

(19) Riniker, S.; Landrum, G. Similarity maps - A visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminformatics* **2016**, *5*.

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

(22) 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. Advances in Neural Information Processing Systems. 2020; pp 5898–5910.

(23) Rasmussen, M. H.; Christensen, D. S.; Jensen, J. H. Do machines dream of atoms? A quantitative molecular benchmark for explainable AI heatmaps. **2022**,

(24) Rao, J.; Zheng, S.; Yang, Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *arXiv preprint arXiv:2107.04119* **2021**,

(25) Jiménez-Luna, J.; Skalic, M.; Weskamp, N. Benchmarking molecular feature attribution methods with activity cliffs. *J. Chem. Inf. Model.* **2022**, *62*, 274–283.

(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) Wang, H.; Li, W.; Jin, X.; Cho, K.; Ji, H.; Han, J.; Burke, M. D. Chemical-reaction-aware molecule representation learning. *arXiv preprint arXiv:2109.09888* **2021**,

(28) Hu, Y.; Stumpfe, D.; Bajorath, J. Computational exploration of molecular scaffolds in medicinal chemistry. *J. Med. Chem.* **2016**, *59*, 4062–4076.

(29) Dalke, A.; Hastings, J. FMCS: A novel algorithm for the multiple MCS problem. *J. Cheminformatics* **2013**, *5*, 1–1.

(30) Landrum, G. RDKit Documentation. *Release* **2013**, *1*, 4.

(31) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research* **2007**, *35*, D198–D201.

(32) Maggiora, G. M. On outliers and activity cliffs why QSAR often disappoints. 2006.

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

(34) Simonovsky, M.; Komodakis, N. Dynamic edge-conditioned filters in convolutional neural networks on graphs. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017; pp 3693–3702.

(35) Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; Torralba, A. Learning deep features for discriminative localization. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016; pp 2921–2929.

(36) Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. International conference on machine learning. 2017; pp 3145–3153.

(37) Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic attribution for deep networks. International conference on machine learning. 2017; pp 3319–3328.

(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. Proceedings of the IEEE international conference on computer vision. 2017; pp 618–626.

(39) Johansson, U.; Sönströd, C.; Norinder, U.; Boström, H. Trade-off between accuracy and interpretability for predictive in silico modeling. *Future Med. Chem.* **2011**, *3*, 647–663.

(40) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

(41) 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.

(42) Park, J.; Sung, G.; Lee, S.; Kang, S.; Park, C. ACGCN: Graph convolutional networks for activity cliff prediction between matched molecular pairs. *J. Chem. Inf. Model.* **2022**,

(43) Chen, D.; Lin, Y.; Li, W.; Li, P.; Zhou, J.; Sun, X. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. Proceedings of the AAAI Conference on Artificial Intelligence. 2020; pp 3438–3445.

(44) Godwin, J.; Schaarschmidt, M.; Gaunt, A. L.; Sanchez-Gonzalez, A.; Rubanova, Y.; Veličković, P.; Kirkpatrick, J.; Battaglia, P. Simple GNN regularisation for 3D molecular property prediction and beyond. International conference on learning representations. 2021.

(45) You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; Shen, Y. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems* **2020**, *33*, 5812–5823.

(46) Wang, Y.; Magar, R.; Liang, C.; Barati Farimani, A. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *J. Chem. Inf. Model.* **2022**,

(47) Stärk, H.; Beaini, D.; Corso, G.; Tossou, P.; Dallago, C.; Günnemann, S.; Liò, P. 3D infomax improves GNNs for molecular property prediction. International Conference on Machine Learning. 2022; pp 20479–20502.

(48) Zaidi, S.; Schaarschmidt, M.; Martens, J.; Kim, H.; Teh, Y. W.; Sanchez-Gonzalez, A.; Battaglia, P.; Pascanu, R.; Godwin, J. Pre-training via denoising for molecular property prediction. *arXiv preprint arXiv:2206.00133* **2022**,