A substructure-aware loss for feature attribution in drug discovery

Kenza Amara[1,2]
kenza.amara@ai.ethz.ch

Raquel Rodríguez-Pérez[3]
raquel.rodriguez_perez@novartis.com

José Jiménez-Luna[1]
jjimenezluna@microsoft.com

[1] Microsoft Research Cambridge, CB1 2FB Cambridge, United Kingdom
[2] Department of Computer Science, ETH Zurich, 8092 Zurich, Switzerland
[3] Novartis Institutes for Biomedical Research, 4002 Basel, Switzerland

Abstract

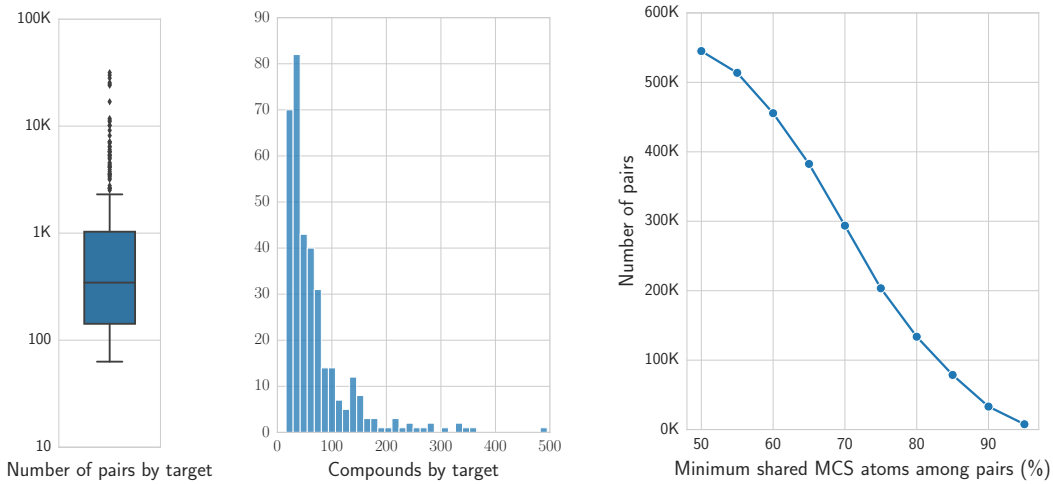
Explainable machine learning is increasingly used in drug discovery to help rationalize compound property predictions. Feature attribution techniques are popular choices within this toolbox that aim to identify which substructures of a molecule are responsible for a predicted property change. However,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

1 Introduction

Drug discovery is one of the many fields where deep learning techniques have found extensive applicability in the last few years [Chen et al., 2018]. While the history behind traditional applied machine-learning (ML) research in cheminformatics can be traced as far back to the 1960s [Muratov et al., 2020, Cherkasov et al., 2014], some recently-adopted deep-learning paradigms have become increasingly popular across many tasks (*e.g.*, de novo molecular design, synthesis prediction). Specifically, in silico property prediction (also commonly referred to as quantitative structure-property relationship modeling) is one of those central challenges in drug discovery where graph neural networks (GNNs) [Gilmer et al., 2017] have shown promising performance. Among the many factors that contributed to the popularity of GNNs in chemistry as well as in other areas, we can highlight their suitability to naturally perform automatic feature extraction on arbitrarily-sized graphs and their scalability to existing commodity hardware. Specifically in chemistry, GNNs can take advantage of the fact that molecules can be naturally described as graphs where atoms and bonds correspond to nodes and edges, respectively.

This rise in popularity has also been accompanied by an increasing need for explainability [Jiménez-Luna et al., 2020, Rodríguez-Pérez and Bajorath, 2021a,b], as these models have been notoriously known for their black-box character. Towards this goal, explainable artificial intelligence techniques,

Preprint. Under review.



(a) (b) (c)

Figure 1: Descriptive statistics of the benchmark used in this work. (a-b) Distribution of number of pairs and compounds per protein target. (c) Total number of available compound pairs considered at each MCS threshold.

such as feature attribution analyses, have become popular tools. These analyses are attractive since they provide an importance value for every input feature in an atom or bond in a molecular graph. Such values are then usually visualized through atom coloring, where the structural patterns driving a prediction are highlighted on top of the two-dimensional molecular representation of the compound of interest [Riniker and Landrum, 2016].

Towards disentangling what structural patterns are exploited by GNNs in activity or other property prediction tasks, the use of a variety of feature attribution techniques has been previously reported in the literature [McCloskey et al., 2019] and continues to remain an active topic of research. Importantly, many research efforts have focused on benchmarking feature attribution techniques, exploring their consistency and quality in atom coloring, and providing recommendations. [Matveieva and Polishchuk, 2021, Sanchez-Lengeling et al., 2020, Rasmussen et al., 2022, Rao et al., 2021] In particular, one of such studies 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 [Jiménez-Luna et al., 2022]. 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 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 considerably improves performance on the aforementioned benchmark. Specifically, 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 significant improvements for many gradient-based as well as simpler perturbation-based attribution methods. We furthermore manage to close the attribution performance gap for GNNs against the best performing combination of classical ML method and feature attribution technique.

2 Materials and methods

2.1 Benchmark

Molecular scaffolds. A scaffold is defined as the core of the molecule where one or several functional groups can be attached. Herein, the maximum common substructure (MCS) formalism was used to define a molecular scaffold [Hu et al., 2016] between pairs of compounds binding to a specific target. To consider that two compounds share a molecular scaffold, such common structural part should encompass a minimum fraction of their structure. As further described below, different thresholds of minimum shared MCS were examined [Jiménez-Luna et al., 2022]. For the development and evaluation of our methodology, MCS pairs were computed using the FMCS [Dalke and Hastings, 2013] algorithm, as available in the RDKit rdFMCS module [Landrum, 2013].

Data preparation. We used the benchmark data proposed by Jiménez-Luna et al. [2022], which consisted of 723 protein targets with associated small molecule activity data (half maximal inhibitory concentration, IC_{50}). In particular, we focus on congeneric (*i.e.*, related) compounds pairs that share a molecular scaffold (MCS) and have at least 1 log activity difference. Compound sets for each protein target were randomly divided into training (80%) and test (20%) sets. Only those protein targets with at least 50 compound pairs in the training set were kept. To avoid data leakage, we ensured that compounds did not have analogs (pairs) both in the training and test sets. Such filtering procedure resulted in a final data set of 350 protein targets. Figure 1 shows the distribution of the number of pairs and compounds per target at a minimum MCS threshold of 50%. Figure 1 also reports the number of pairs sharing a molecular scaffold at different minimum MCS thresholds.

2.2 Models and feature attribution techniques

Models. Message-passing GNN [Simonovsky and Komodakis, 2017] models were generated to predict compound activity against all available protein targets. GNN models were trained to minimize at least one of the following loss functions: (i) mean squared error (MSE) between observed and predicted binding affinities (in log scale), (ii) a relative affinity loss computed on pairs of related compounds, hereby referred to as activity cliff (AC), 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 Section 2.3). As a control, random forest (RF) models trained with extended-connectivity fingerprints (ECFP4) were also considered.

Feature attribution techniques. A variety of feature attribution methods that enable the estimation of positive and negative atom contributions were investigated. Methods such as Class Activation Maps (CAM) [Zhou et al., 2016], gradient-based ones such as GradInput [Shrikumar et al., 2017], Integrated Gradients [Sundararajan et al., 2017] and Grad-CAM [Selvaraju et al., 2017] were utilized. Additionally, we considered other perturbation-based approaches such as node masking, where the contribution of each atom is determined as a function of the difference in prediction after it has been perturbed. In the case of the GNN models used here, this approach iteratively zeroed-out node features. For the RF models, instead, each atom was assigned a chemical species (*i.e.*, an atom type) that was not present in the benchmark sets, and followed by molecular re-featurization [Sheridan, 2019].

2.3 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. Specifically, during training we sample pairs of compounds with a common substructure and attribute the difference in predicted activity 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:

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

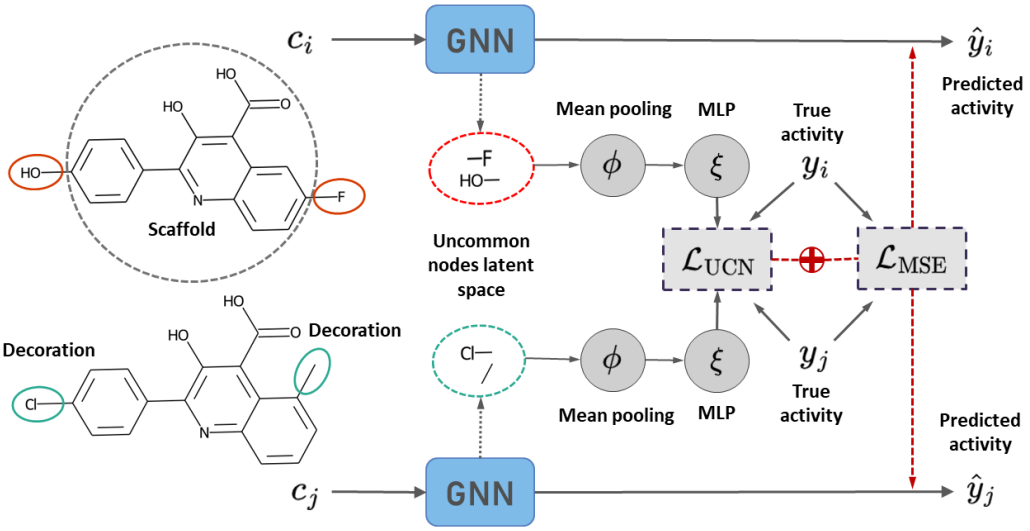


Figure 2: The proposed UCN loss is schematized. Two compounds sharing a common substructure are sampled from the training set and node latent spaces are computed via a forward pass of a GNN model. The uncommon latent nodes are then used to compute a loss targeting the activity difference between the compound pairs.

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:

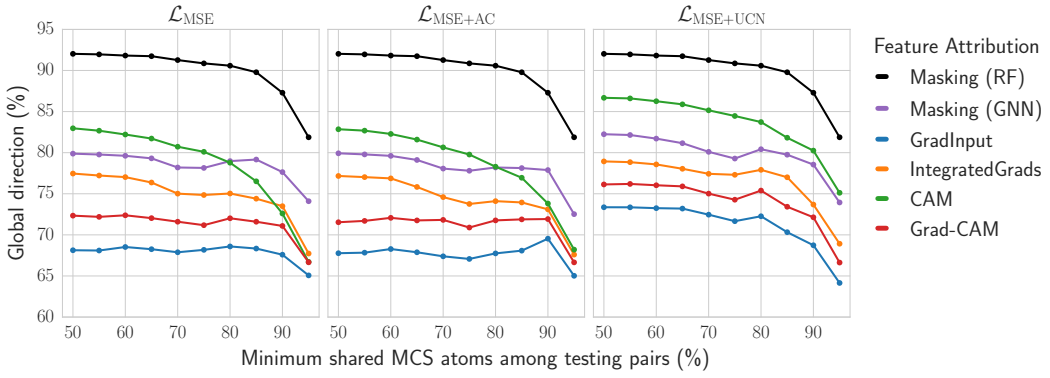
$$\mathcal{L}_{\text{AC}}\left(c_i, c_j\right) = \left\|\left(y_i - y_j\right) - \left(\hat{y}_i - \hat{y}_j\right)\right\|^2. \tag{3}$$

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

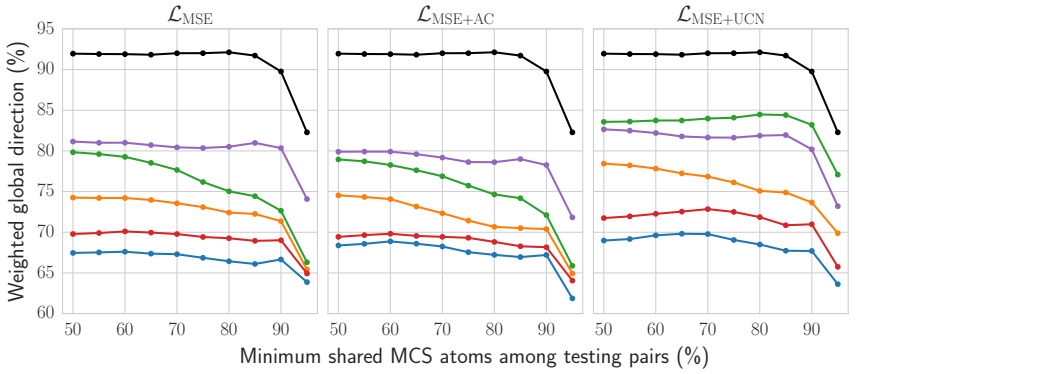
2.4 Evaluation

The performance of the feature attribution methods considered in this study was evaluated using the *global direction* and *atom-level accuracy* metrics proposed in Jiménez-Luna et al. [2022]. Global direction is a binary metric assessing whether the 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:

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



(a)



(b)

Figure 3: (a) Global direction and, (b) weighted global direction by the number of pairs present in each target at different MCS thresholds. Results are shown for the test sets of 350 protein targets.

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 a ground truth determined by the experimental activity difference of the compound pair. In the original benchmark study, the ground truth of atom contributions was obtained by assuming that for each compound pair, structural changes were responsible for the observed potency changes. 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:

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

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.

Regarding g_{atom}, it was previously noted in the original benchmark study that the ground-truth colors assigned by this metric to a given compound can be ill-defined, since they are dependent on the other compound in the pair. In contrast, the other proposed metric, g_{dir}, does not suffer from this problem. For this reason, results are shown for g_{dir} in the experiments section of the main manuscript while, for completeness purposes, g_{atom} results are reported in the Section A.4.

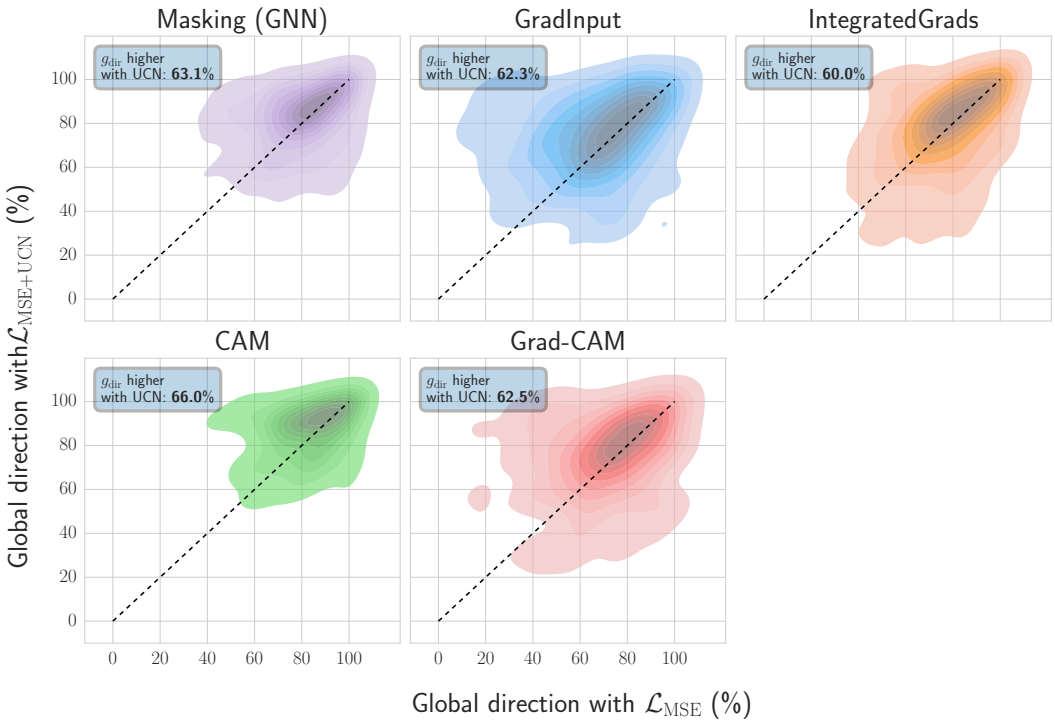


Figure 4: 2D kernel density plot displaying target specific global direction at the 50% MCS threshold for all molecular test pairs, for both the simple absolute MSE loss and the hereby proposed UCN loss. Text-box quantifying the proportion of targets for which g_{dir} was higher with $\mathcal{L}_{\texttt{MSE+UCN}}$

3 Experiments

Figure 3 shows the global direction values for all test pairs and targets considered in the study. Many feature attribution methods trained with GNNs using the proposed UCN objective exhibited a global direction improvement over both the absolute MSE and relative MSE loss (*i.e.* AC). Improvements are shown for most methods, and are most marked in the case of CAM, Grad-CAM and GradInput. Additionally, the GNN-based masking method also exhibits a slight increase. Most importantly, this improvement holds across different thresholds of minimum MCS between pairs. Results for the weighted color direction metric are also reported on Figure 3b, where similar conclusions can be drawn, and with Integrated Gradients also showing a much more pronounced improvement than in the non-weighted analyses.

In Figure 4 the influence of the UCN loss on each specific feature attribution method was investigated. Global direction values are reported for models trained with and without using the UCN loss term per each target. The percentage of targets for which the global direction values were larger when using the MSE+UCN loss are also reported. Most methods showed improvements with the inclusion of the UCN loss term, as well as the GNN-based masking. At least 60% of the targets considered in the testing set benefited from the UCN loss with any of the feature attribution methods considered, with CAM showing the highest improvements (66%). Additional plots and analyses can be found in Sections A.2 and A.3, where we can highlight that the performance of CAM is observed to be very similar to that of the RF masking method when evaluated on the training sets.

4 Discussion

In this study, we have proposed a substructure-aware loss for GNN models in the context of congeneric series data in drug discovery. We have evaluated the influence of this loss on a previously-reported benchmark for molecular ML explainability and show that many GNN-based feature attribution techniques markedly benefit from its usage. Our results showed that the average global direction as well as the percentage of targets with global direction improvement were superior

with the consideration of the UCN loss during GNN training. However, despite the observed improvements in all GNN-based feature attribution methods, the RF model with an atom masking approach still remains the best combination for explainability in chemical property prediction settings [Sheridan, 2019].

Since lead optimization efforts in drug discovery are centered on specific chemical series, we expect that the presented explainability approach will help rationalize GNN-based model decisions. Finally, a main limitation of the proposed methodology is the requirement of precomputed common substructures between pairs of compounds in the series of interest. This calculation can be computationally expensive with exact common substructure algorithms such as FMCS, but may be bypassed by the use of approximate MCS techniques or matched molecular pair analyses [Griffen et al., 2011]. Code to replicate the results in this paper is provided in `https://github.com/microsoft/molucn`

Acknowledgments and Disclosure of Funding

K. Amara acknowledges financial support from Microsoft Research. We also thank K. Maziarz and M. Segler for helpful discussions during K. Amara's internship at MSR Cambridge. The authors declare no conflict of interest.

References

Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug Discov. Today*, 23(6):1241–1250, 2018.

Eugene N Muratov, Jürgen Bajorath, Robert P Sheridan, Igor V Tetko, Dmitry Filimonov, Vladimir Poroikov, Tudor I Oprea, Igor I Baskin, Alexandre Varnek, Adrian Roitberg, et al. QSAR without borders. *Chem. Soc. Rev.*, 49(11):3525–3564, 2020.

Artem Cherkasov, Eugene N Muratov, Denis Fourches, Alexandre Varnek, Igor I Baskin, Mark Cronin, John Dearden, Paola Gramatica, Yvonne C Martin, Roberto Todeschini, et al. QSAR modeling: Where have you been? Where are you going to? *J. Med. Chem.*, 57(12):4977–5010, 2014.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pages 1263–1272. PMLR, 2017.

José Jiménez-Luna, Francesca Grisoni, and Gisbert Schneider. Drug discovery with explainable artificial intelligence. *Nat. Mach. Intell.*, 2(10):573–584, 2020.

Raquel Rodríguez-Pérez and Jürgen Bajorath. Explainable machine learning for property predictions in compound optimization. *J. Med. Chem.*, 64(24):17744–17752, 2021a.

Raquel Rodríguez-Pérez and Jürgen Bajorath. Chemistry-centric explanation of machine learning models. *Artificial Intelligence in the Life Sciences*, 1:100009, 2021b. doi: 10.1016/j.ailsci.2021.100009.

Sereina Riniker and Gregory Landrum. Similarity maps - A visualization strategy for molecular fingerprints and machine-learning methods. *J. Cheminformatics*, 5(43), 2016. doi: 10.1186/1758-2946-5-43.

Kevin McCloskey, Ankur Taly, Federico Monti, Michael P Brenner, and Lucy J Colwell. Using attribution to decode binding mechanism in neural network models for chemistry. *Proc. Natl. Acad. Sci. U.S.A.*, 116(24):11624–11629, 2019.

Mariia Matveieva and Pavel Polishchuk. Benchmarks for interpretation of QSAR models. *J. Cheminformatics*, 13(1):1–20, 2021.

Benjamin Sanchez-Lengeling, Jennifer Wei, Brian Lee, Emily Reif, Peter Wang, Wesley Qian, Kevin McCloskey, Lucy Colwell, and Alexander Wiltschko. Evaluating attribution for graph

neural networks. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5898–5910. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/417fbbf2e9d5a28a855a11894b2e795a-Paper.pdf.

Maria H Rasmussen, Diana S Christensen, and Jan H Jensen. Do machines dream of atoms? A quantitative molecular benchmark for explainable AI heatmaps. 2022.

Jiahua Rao, Shuangjia Zheng, and Yuedong Yang. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *arXiv preprint arXiv:2107.04119*, 2021.

José Jiménez-Luna, Miha Skalic, and Nils Weskamp. Benchmarking molecular feature attribution methods with activity cliffs. *J. Chem. Inf. Model.*, 62(2):274–283, 2022.

Ye Hu, Dagmar Stumpfe, and Juergen Bajorath. Computational exploration of molecular scaffolds in medicinal chemistry. *J. Med. Chem.*, 59(9):4062–4076, 2016. doi: 10.1021/acs.jmedchem.5b01746.

Andrew Dalke and Janna Hastings. FMCS: a novel algorithm for the multiple MCS problem. *J. Cheminformatics*, 5(1):1–1, 2013.

Greg Landrum. RDKit documentation. *Release*, 1(1-79):4, 2013.

Martin Simonovsky and Nikos Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3693–3702, 2017.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Robert P Sheridan. Interpretation of QSAR models by coloring atoms according to changes in predicted activity: How robust is it? *J. Chem. Inf. Model.*, 59(4):1324–1337, 2019.

Ed Griffen, Andrew G Leach, Graeme R Robb, and Daniel J Warner. Matched molecular pairs as a medicinal chemistry tool: Miniperspective. *J. Med. Chem.*, 54(22):7739–7750, 2011.

Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch Geometric. *arXiv preprint arXiv:1903.02428*, 2019.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 2010.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.

A Appendix

A.1 Predictive performance

Predictive performance results are reported in Table A.1 for GNN models trained with the proposed $\mathcal{L}_{\text{MSE}}$, $\mathcal{L}_{\text{MSE+AC}}$, and $\mathcal{L}_{\text{MSE+UCN}}$ losses. As a control, RF models were also evaluated. Average performance values were lower for the GNN model trained with the UCN loss, albeit results lied within one standard deviation of the other methods.

Table A.1: Test set prediction performance. Reported are the average root mean squared error (RMSE) and Pearson correlation coefficient (PCC) values ($\pm$ standard deviation). Results are shown for a simple average over targets and a weighted average, weighing by the number of test pairs per target.

	RMSE	**Weighted RMSE**	**PCC**	**Weighted PCC**
Random Forest	0.32 ($\pm$0.11)	0.3 ($\pm$0.08)	0.95 ($\pm$0.07)	0.96 ($\pm$0.04)
$\mathcal{L}_{\text{MSE}}$	0.34 ($\pm$0.26)	0.24 ($\pm$0.14)	0.89 ($\pm$0.23)	0.96 ($\pm$0.08)
$\mathcal{L}_{\text{MSE+AC}}$	0.34 ($\pm$0.27)	0.23 ($\pm$0.11)	0.89 ($\pm$0.23)	0.97 ($\pm$0.08)
$\mathcal{L}_{\text{MSE+UCN}}$	0.43 ($\pm$0.29)	0.3 ($\pm$0.14)	0.86 ($\pm$0.24)	0.95 ($\pm$0.09)

A.2 Additional global direction results in the test set

To complement Figure 4, heatmaps on Figure A.1 indicate the absolute number of protein targets concerned by an improvement in global direction with the UCN loss. Again we observe most protein targets (represented by darker blue squares) above the diagonal.

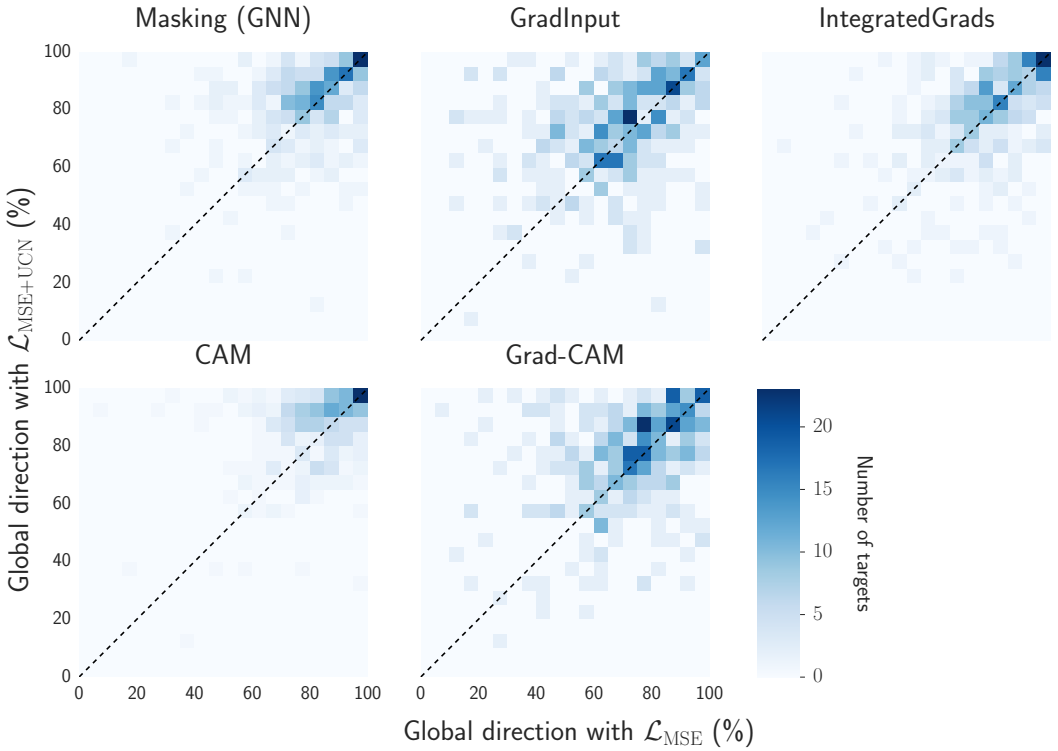


Figure A.1: Heatmap displaying target specific global direction at the 50% MCS threshold for all molecular test pairs, for both the simple absolute MSE loss and the hereby proposed UCN loss.

In Figure A.2, we further explore for how many targets there was a negligible (<5%), small (between 5 and 10%), medium (10 and 20%), or large (>20%) global direction improvement when including the UCN loss term. Results indicate that MSE+UCN consistently achieve superior global direction results

for the same or higher number of targets than MSE. Interestingly, differences across loss functions become larger when considering targets with medium to large global direction improvements. CAM, GradInput, and Grad-CAM are the methods showing the largest benefit of MSE+UCN, having a large number of targets (133 targets for Grad-CAM, 138 for GradInput and 81 for CAM) with global direction improvements higher than 20%.

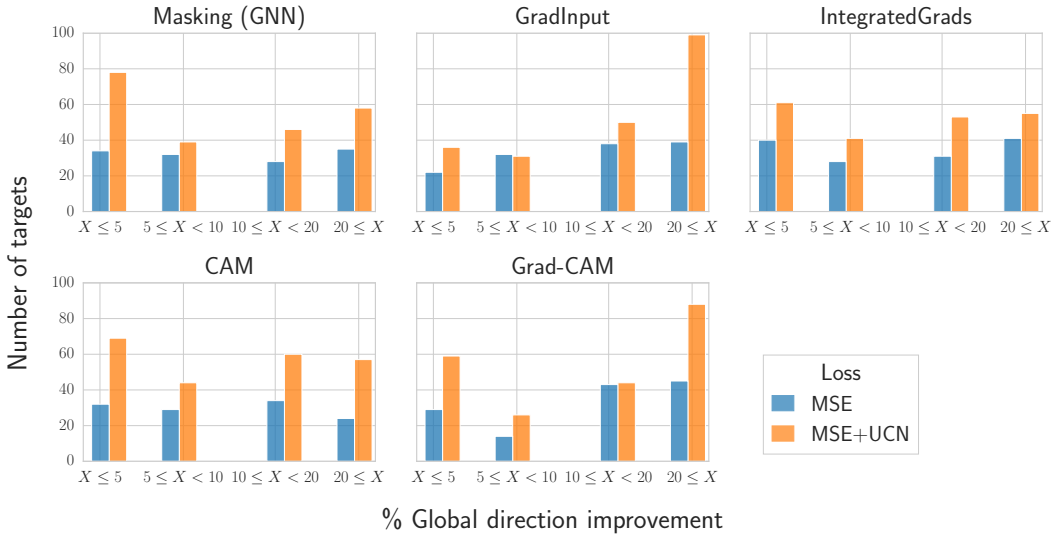


Figure A.2: Number of targets with an improvement of the global direction with the simple absolute MSE (blue) or the proposed UCN loss (orange) at different percentage levels. Here, the global direction is computed at the 50% MCS threshold for all molecular test pairs.

A.3 Global direction in the training set

For completeness, global direction results are also reported for the training set in Figure A.3. Similar trends were observed, with the UCN loss improving on this metric for all gradient-based feature attribution methods evaluated.

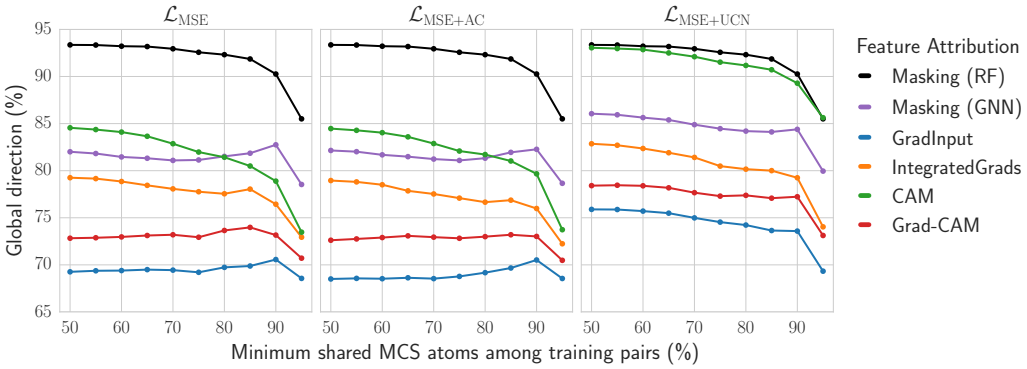
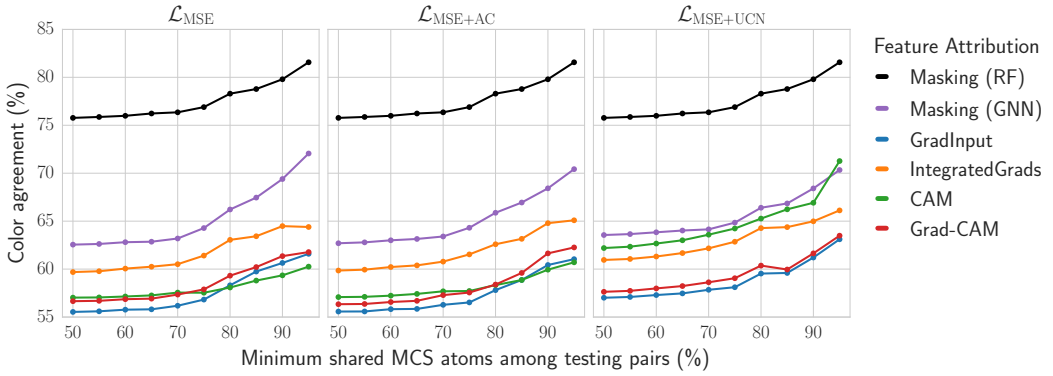


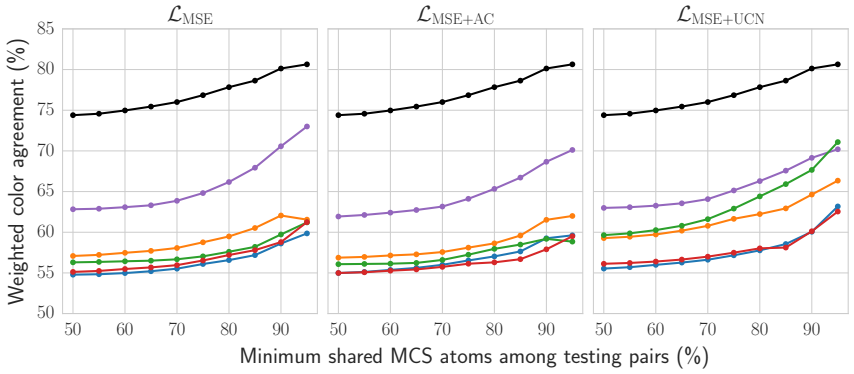
Figure A.3: Global direction metric at different MCS thresholds for the molecules present in the training pairs and for different loss functions considered.

A.4 Color agreement

Figures A.4 and A.5 report color agreement accuracy for the training and test sets, respectively. Similar conclusions as those presented in section 3 can be drawn, with the UCN loss improving on this metric for several of the feature attribution methods evaluated, albeit the advantage is less pronounced.



(a)



(b)

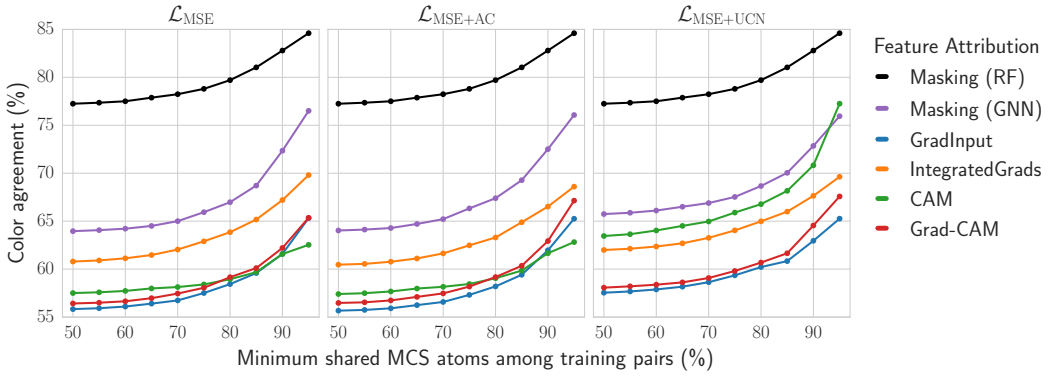
Figure A.4: (a) Color agreement and, (b) weighted color agreement accuracy at different MCS thresholds for the molecules present in the test pairs and for the different loss functions.

A.5 Neural network hyperparameters and featurization

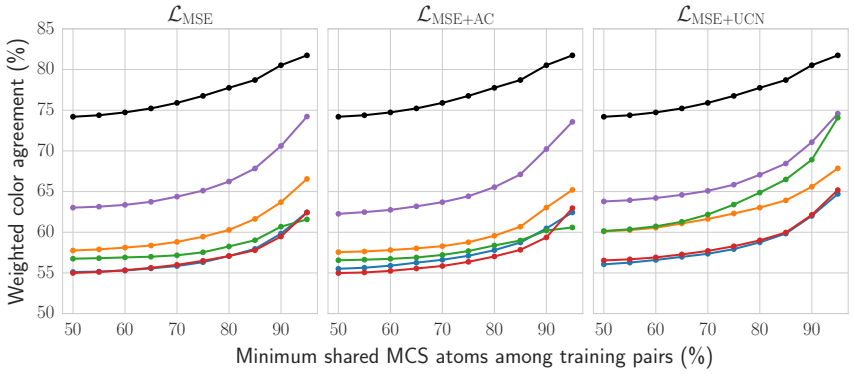
The trained GNN models featured 2 message-passing layers over the bonds of the molecular graph layers, each with 32 neurons. Graphs were featurized with the descriptors provided in Table A.2. A mean node pooling strategy was used for all losses considered in the manuscript, followed by two multi-layer perceptron modules with 32 and 16 hidden nodes, respectively and with ReLU activation functions. In the case of the CAM feature attribution method, only 1 multilayer perceptron with 32 hidden nodes was used to adhere with GNN node latent size. GNN's hyperparameters and other training parameters are described in Table A.3. All GNN models were trained using the PyTorch geometric software package [Fey and Lenssen, 2019]. RF models consisted of 1000 base tree learners and used 1024-bit binary Morgan fingerprints as molecular representation [Rogers and Hahn, 2010]. For RF, the scikit-learn implementation [Pedregosa et al., 2011] was utilized.

Table A.2: Node and edge molecular graph features used in the training of the GNN models

Description level	**Features**
Atom	atom type, number of heavy atom neighbors, formal charge, hybridization, presence in ring, aromaticity, atomic mass, van der Waals radius, covalent radius
Bond	bond type, bond stereo, conjugation, presence in ring



(a)



(b)

Figure A.5: (a) Color agreement and, (b) weighted color agreement accuracy at different MCS thresholds for the molecules present in the train pairs and for the different loss strategies evaluated.

Table A.3: Architectural details of the GNN models used throughout this study and additional training hyperparameters

Model		**Training**	
Base layer	NNConv	Batch size	16
# layers	2	Optimizer	Adam
Hidden Dim.	32	# epochs	200
Readout f.	mean	Learning rate	10^{-3}

A.6 Additional details on the feature attribution techniques

Feature attribution was computed for both edges and nodes of the input graph. Similar to prior studies [McCloskey et al., 2019], edge attributions were then halved and equally distributed to the nodes connected by the edge. For the Monte Carlo approximation of the path integral in the Integrated Gradients method, 50 steps were selected.