Decoding Chemical Predictions: Group Contribution Methods for XAI

Gabriel Cathoud [1] **Vignesh Ram Somnath** [2] **Luis Macedo** [1] **Kjell Jorner** [3]

Abstract

Graph Neural Networks (GNNs) have recently shown great promise for modeling chemical systems. However, beyond the accuracy and performance of these models, understanding their underlying mechanisms is also crucial. While many general GNN explainers exist, incorporating domain-specific knowledge can enhance the development of explainers tailored to chemical applications. In this study, we developed an approach based on the well-established concept of group contributions, providing additional explanations without compromising model accuracy. Our results indicate that different GNN models may learn distinct patterns from the molecules. Furthermore, by applying a custom loss function, we successfully aligned the learning process of the models with desired group contributions while maintaining the overall model performance.

1. Introduction

The ability to predict chemical properties has long been a desire in both academic and industrial contexts. Recently, Artificial Intelligence (AI) and Machine Learning (ML) have shown considerable promise in chemistry (Janet & Kulik, 2020; Anstine & Isayev, 2023). Graph Neural Networks (GNNs) hold tremendous potential in this field, as they naturally align with the intrinsic graph structure of a molecule. Numerous GNN-based models have reached state-of-the-art accuracy (Bihani et al., 2024).

However, as models become more accurate, other important debates have emerged. Some argue that ML models merely capture correlations within the data without understanding the underlying chemical principles. Due to the nature of ML models and their reliance on numerous non-linear operations, their internal workings are often complex and challenging for humans to interpret. In light of these discussions, it is crucial to focus not only on the accuracy and performance of models but also on understanding their mechanisms.

Explainable Artificial Intelligence (XAI) plays a vital role in this context. XAI is dedicated to developing methods that clarify how models function and justify their predictions. XAI tools serve three essential functions in the field of chemistry. First, they foster trust among skeptical users. Second, by elucidating the model's mechanisms, they enable specialists to use their knowledge of chemistry to improve and refine the models. Third, they facilitate informed decision-making and help justify the validity of the choices resulting from model predictions. While various XAI techniques for explaining GNN models exist (Yuan et al., 2022; Li et al., 2022; Kakkad et al., 2023), many are too general and require customization for specific model applications. A critical discussion centers on the potential benefits of developing explainers that utilize domain-specific knowledge to craft more relevant explanations. Additionally, this can help establish a more accurate ground truth for what an explanation should entail. Recognizing this possibility, some researchers have begun exploring this avenue by creating explainers leveraging from chemical knowledge (Jiménez-Luna et al., 2020; Rodríguez-Pérez & Bajorath, 2021; Wellawatte et al., 2023)

In our study, we revisited an established concept in chemistry, group contribution (GC) methods (Gani, 2019), to explain GNN models. GC methods segment molecules into groups and calculate the molecule's properties by summing each group's contribution. Typically, these contributions are estimated using linear regression, a simple and transparent model. Since the groups represent parts of the molecule that chemists are familiar with, they are intuitively understood within the chemical community. By altering the aggregation approach of the GNN models, we successfully extracted group contributions in addition to the predictions. Our objective was to implement these modifications with minimal changes to the original models, ensuring that their accuracy was preserved while simultaneously providing interpretability. Fig. 1 presents a schematic of our approach, illustrating how these modifications were integrated into the existing model framework.

[1]Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal [2]Department of Computer Science, ETH Zurich, Zurich, Switzerland [3]Department of Chemistry and Applied Biosciences, ETH Zurich, Zurich, Switzerland. Correspondence to: Kjell Jorner <kjell.jorner@chem.ethz.ch>.

Accepted at the 1st Machine Learning for Life and Material Sciences Workshop at ICML 2024.

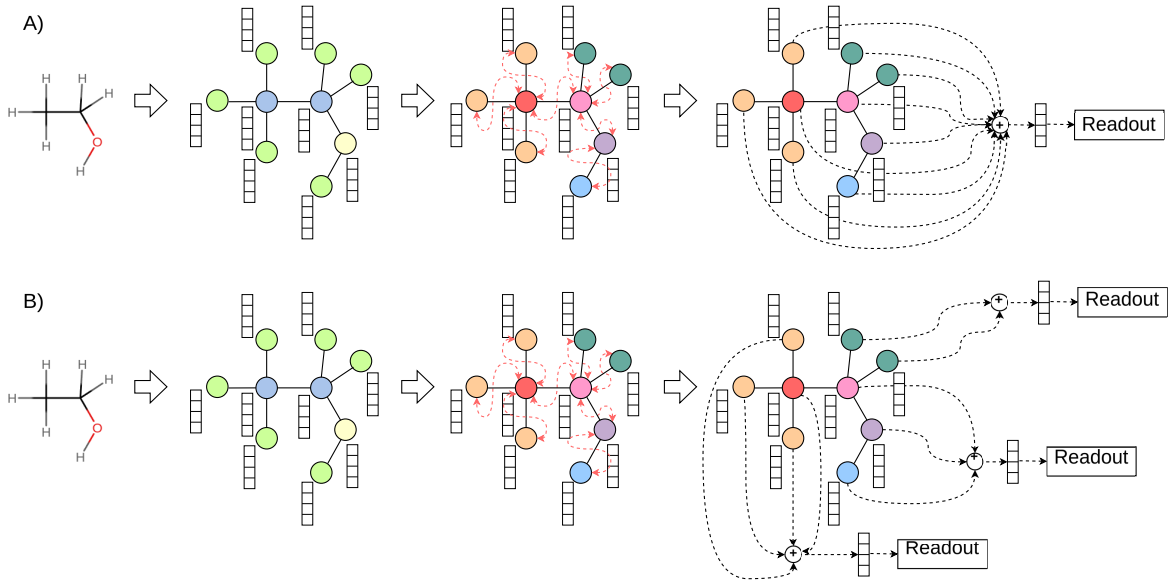

Figure 1. Schematic of a classical GNN approach (A) and our approach (B). In both approaches, a molecule is represented using a graph where each node is associated with a vector. During the message-passing process, the vectors are updated based on the neighborhood information. In the classical GNN approach (A), all vectors are aggregated after the message-passing process. In our approach (B), the vectors are aggregated based on the information about the groups.

2. Related work

Rasmussen et al. (Rasmussen et al., 2022) employed perturbation methods to determine the contributions of different molecular fragments to the predicted octanol-water partition coefficient (LogP). They have used the contributions of Crippen's LogP model, a type of GC method, to benchmark the contributions of the fragments they obtained with the ML models. Inspired by their approach, we have extended this investigation to other chemical properties, specifically the enthalpy of formation and the HOMO-LUMO gap.

Chen et al. (Chen et al., 2022) leveraged the GC concept to build their model using a GNN architecture based on 2D graphs. They incorporated GC and a GC benchmark to enhance model accuracy. While their focus was on developing a new model based on 2D graphs, we have focused on explaining the state-of-the-art 3D-based models.

Wu et al. (Wu et al., 2023) employed a masking strategy to evaluate the contribution of different molecular fragments, including BRICS substructures, Murcko substructures, and functional groups. They used these masks to identify substructures that significantly influence model predictions. While masking is effective in determining the importance of different nodes in a graph, our approach diverges by avoiding modifications to the inputs. Instead, our aim is to derive explanations directly within the prediction process itself.

Aouichaoui et al. (Aouichaoui et al., 2023) have developed a model learning embeddings at three levels: node, group, and junction tree. They concatenate these embeddings and use a multilayer perceptron for predictions. With their junction tree level, they were able to explain the predictions and obtain the influences of the different groups. Instead of a multi-level process, we directly use readouts to get scalars for each group and sum them for the final prediction.

Walter et al. (Walter et al., 2024) utilized an attention-based method to investigate which parts of a molecule contribute most to a given prediction. Their focus was on fingerprint-based models for classification tasks. In contrast, our work centers on GNN models for regression tasks, highlighting different methodological and application focuses.

Our work introduces a novel application of GC methods to explain established GNNs. The key innovation involves modifying the aggregation mechanism within the GNN, enhancing interpretability without compromising model performance. This is crucial, as traditional XAI tools often reduce model accuracy by using simpler surrogate models or decomposing inputs to gauge component influences. In contrast, we preserved the original message-passing and readout layers of the original models, altering only the aggregation logic.

3. Proposed approach

Group Aggregation: GNN models typically consist of message-passing layers followed by a readout phase. In our approach, we aggregate embeddings based on group information. Each group contains a central atom and neighbors. We sum the embeddings of all the neighbors with the embedding of the central atom within each group and then pass the resultant vector to the readout layer. This process yields a scalar value representing each group's contribution. The final prediction is obtained by summing these individual group contributions (see Fig. 1).

Group definition: Groups were defined as a central atom and its neighboring atoms. Inspired by Benson's work (Benson, 1968), we established the constraint that each group should have at least two neighbors.

Explanations: After obtaining group contributions, we used heatmaps to represent which parts of the molecule contribute more to the predicted property.

4. Experiments

Data: We used the QM9 dataset (Ramakrishnan et al., 2014), focusing on the enthalpy of formation (ΔH_f) and the HOMO-LUMO gap ($\Delta\epsilon$) as targets. The ΔH_f was corrected using reference atomic energies as already done by others (Anderson et al., 2019). The targets were scaled using the mean absolute deviation (MAD) and the median following (Satorras et al., 2021; Pinsky & Klawansky, 2023). The data was divided based on Bemis-Murcko scaffolds, resulting in a distribution of 70%, 15%, and 15% for the training, validation, and testing sets, respectively.

GNN Models: We employed the SchNet (Schütt et al., 2017) and EGNN (Satorras et al., 2021) models, as they use 3D information, contain desirable properties such as equivariance, and have been successfully applied in various material science tasks. Although newer models exist, our goal was to use well-established methods for proof-of-principle implementation. We hypothesize that the observed results would generalize to more recent methods as well. Given the fact we have used the message-passing and the readout layers of other models, we expect that the scalability and processing costs are dependent on these operations. The aggregation is a simple and fast step, given that the information about the adjacency of the groups is calculated beforehand. We used the hyperparameters from the original studies. For each model type, we created two variants: original and groups. In the groups variant, embeddings were grouped based on Benson groups, and a scalar was obtained for each group, as explained previously. In the original variant, the embeddings were aggregated using a sum operation.

Regression Models: We performed linear regression using the ridge method (Hoerl & Kennard, 2000), using the counting of the groups within each molecule as features. The bias of the ridge models was set to zero. In this case, the group contributions were considered as the coefficients of the ridge regression. 5-fold cross-validation was employed to determine the best hyperparameter of the ridge regression.

Standard training and Testing: The models were trained using the Adam optimizer with a cosine annealing learning rate scheduler and a weight decay of 1×10^{-16}. The initial learning rate was set to 5×10^{-4}. We used the mean absolute error (MAE) as the loss function for both the training and validation phases. The models were validated every 20 epochs. We also implemented checkpointing with a patience of 10 validation cycles, meaning that if the validation loss did not improve within 10 validations, training was stopped. The maximum number of epochs was set at 1000.

Custom loss: In addition to the standard training procedure, we also trained the models using a custom loss composed of two components: the difference between the prediction and the target value and the sum of the differences between the model's group contributions and reference group contributions (see equation below). These two components were weighted by an α parameter. We used the group contribution values from the ridge regression models as reference.

$$Loss_{custom} = (1-\alpha)MAE_{pred.} + \alpha MAE_{groups}$$

The training began with an α value of 0.5, which was linearly decreased to 0. Given the dynamic nature of α, checkpointing was not employed in this phase, and the models were trained for a total of 1000 epochs without interruption. The rationale behind this approach was to initially guide the model with the group contribution bias and gradually allow the model to focus solely on minimizing the prediction loss.

XAI Plots: We used the RDKit cheminformatics toolkit (Landrum et al., 2024) to create a molecular drawing overlaid with a heatmap that was created using the group contributions attributed to the central atom of each group.

Code availability: The code developed for this work is available at https://github.com/g-cathoud/GNNXGroup.

5. Results and discussion

Accuracy of the Models: The linear regression models performed reasonably well for such a simple model, particularly for predicting ΔH_f (R^2 = 0.915). However, the performance for $\Delta\epsilon$ was poorer (R^2 = 0.815). This lower performance is likely due to the non-local distribution of the HOMO and LUMO orbitals in the molecule affecting $\Delta\epsilon$. The grouping method segments the molecules into pieces, which favors localized properties like ΔH_f.

Table 1. Accuracy metrics obtained with the test set for the different models (O - original model, G - model with group aggregation, C - model with group aggregation, trained with custom loss)

ΔH_f	RMSE / meV	MAE / meV	R^2
R.R.	303	235	0.915
EGNN (O)	43	27	0.998
EGNN (G)	50	30	0.998
EGNN (C)	101	38	0.990
SchNet (O)	41	22	0.998
SchNet (G)	46	26	0.998
SchNet (C)	**41**	**20**	**0.998**
$\Delta\epsilon$	RMSE / meV	MAE / meV	R^2
R.R.	514	401	0.815
EGNN (O)	348	222	0.917
EGNN (G)	343	214	0.919
EGNN (C)	343	222	0.919
SchNet (O)	337	211	0.922
SchNet (G)	311	198	0.934
SchNet (C)	**308**	**199**	**0.935**

GNN models significantly improved accuracy over the regression models. For ΔH_f, GNN models achieved an R^2 value of 0.998 in all cases. For $\Delta\epsilon$, GNN models also outperformed the ridge regression models, though learning $\Delta\epsilon$ remained more challenging than ΔH_f, with R^2 values ranging from 0.917 to 0.934.

Comparing our results for the original GNN models with those reported by the original authors, we observed that our mean absolute error (MAE) was twice as high for ΔH_f and four times as high for $\Delta\epsilon$. It is important to note that we used a scaffold split, whereas the original authors used a random split. Therefore, these discrepancies were anticipated. In our experiments, SchNet demonstrated slightly better accuracy than EGNN for both ΔH_f and $\Delta\epsilon$.

Interestingly, the different aggregation schemes did not affect the overall model accuracy for SchNet and EGNN. The R^2 values remained almost constant, and the MAE varied by a maximum of only 7 meV for ΔH_f and 13 meV for $\Delta\epsilon$. In fact, for $\Delta\epsilon$, the grouping method yielded better results. Typically, XAI techniques tend to decrease model accuracy, but our approach offers the significant advantage of providing enhanced interpretability without compromising model performance.

Regarding the models trained with the custom loss, the accuracy of the EGNN in predicting ΔH_f decreased, which underscores the notion that EGNN does not align with the learning process of the ridge regression model. Consequently, efforts to align the learning with the reference contributions resulted in a decline in the model's performance. In terms of the other models, the results show that there was minimal variation. In fact, the SchNet model demonstrated a slight improvement in accuracy. This confirms that employing the custom loss function does not impair the accuracy of the models. Instead, it suggests that the custom loss can be beneficial.

Explainability of the models: Using the group contributions from the different models, we were able to construct the XAI plots (see Fig. 2). Visual inspection showed good agreement in the group contributions between the ridge regression and the SchNet for both the target properties. Regarding EGNN and ridge regression, the agreement between the group contributions from these models was much lower, especially concerning ΔH_f.

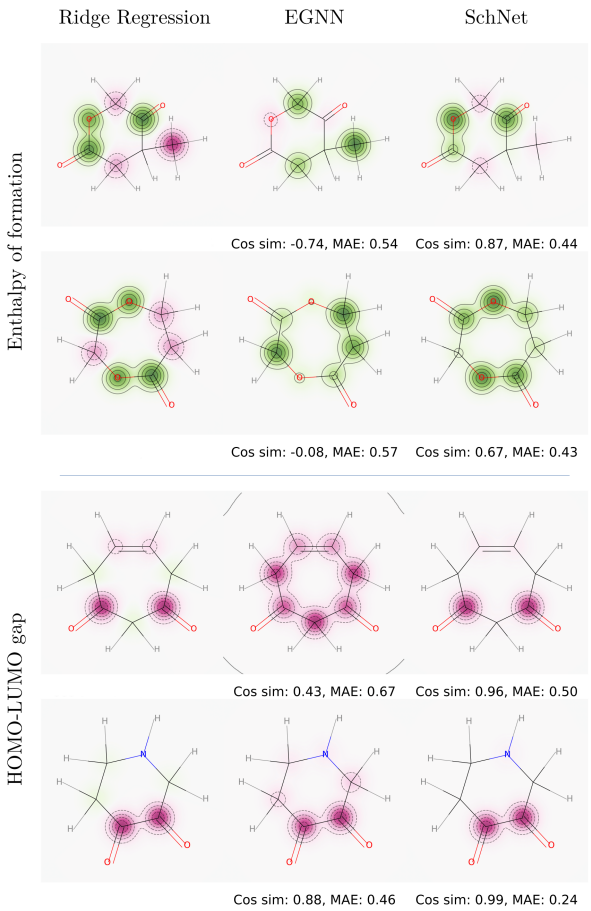

Figure 2. Example of heatmaps obtained with the group contributions for the ridge regression and the GNN models. In this case, Cos sim stands for cosine similarity. Green colors represent positive values, while pink colors represent negative values.

We also calculated the cosine similarity and MAE between contributions from the linear regression models and from the GNN models, as shown in Table 2. Again, the contributions from the ridge regression and SchNet models showed

high agreement, with a mean cosine similarity of 0.70 for ΔH_f and 0.62 for $\Delta\epsilon$, and a mean MAE of 0.25 for ΔH_f and 0.27 for $\Delta\epsilon$. In contrast, the EGNN models exhibited much lower cosine similarity values, particularly for ΔH_f. This indicates that although EGNN and SchNet have similar predictive accuracy, their underlying learned patterns differ. Since the ridge regression model is based on the approach proposed by Benson, the high agreement between SchNet and ridge regression suggests that these models share some chemical insights. However, the EGNN seems to be learning something different. This divergence invites further exploration into the distinct features identified by EGNN and whether these could provide new perspectives on chemical understanding.

Table 2. Average of the cosine similarity (CS) and MAE between the contributions of the ridge regression (R.R.) models and the GNN models. The distributions of the values can be seen in the supplementary materials.

Target	Model comparison	μ_{CS}	μ_{MAE} / ev
	Regular loss		
ΔH_f	EGNN vs. R.R.	-0.17	0.32
	SchNet vs. R.R.	0.70	0.25
$\Delta\epsilon$	EGNN vs. R.R.	0.51	0.23
	SchNet vs. R.R.	0.62	0.27
	Custom loss		
ΔH_f	EGNN vs. R.R.	0.99	0.52
	SchNet vs. R.R.	0.99	0.04
$\Delta\epsilon$	EGNN vs. R.R.	0.94	0.07
	SchNet vs. R.R.	0.96	0.07

Regarding the results obtained with the custom loss, it is evident that the agreement is much higher, with average values for cosine similarity reaching as high as 0.99 for ΔH_f and 0.96 for $\Delta\epsilon$. Yet, the high average MAE regarding EGNN when predicting ΔH_f further underscores the model's resistance to aligning with the group contributions derived from ridge regression. In other cases, the average MAE remained very low. Looking at the examples presented in Fig. 3, it is evident that the agreement is significantly higher across the models. These results are particularly significant because, while the models demonstrated similar accuracy for most cases, the incorporation of the custom loss function enabled the alignment of the model's learning with established chemical intuition. The parameter space of these models is vast, and due to the nature of the models, it is highly probable that multiple optima exist that yield accurate predictions. However, during training, a model may converge to an optimum that, although accurate, does not align well with specific domain knowledge. Our results indicate that it is possible to guide the models toward a region that better aligns with these concepts while maintaining high accuracy.

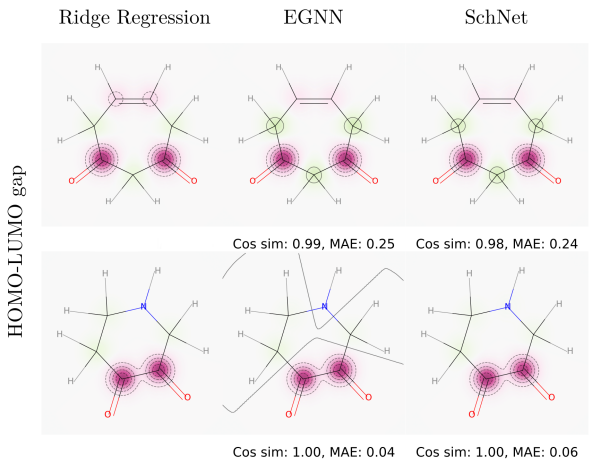

Figure 3. Example of heatmaps obtained with the group contributions for the ridge regression and the GNN models. Results obtained with custom loss. Again, Cos sim stands for cosine similarity. Green colors represent positive values, while pink colors represent negative values.

6. Conclusions

Our study demonstrates that modifying the aggregation step in GNN models significantly enhances their explainability with minimal impact on performance. This advancement is crucial for providing interpretable predictions while maintaining high accuracy. Furthermore, our analysis reveals a strong alignment between group contributions from both ridge regression and SchNet models, indicating that these models capture common underlying chemical principles. In contrast, the contributions from EGNN differ significantly from those of the ridge regression, suggesting that the models might be learning different patterns. This raises curiosity about the source of these differences and how we can benefit from them.

Additionally, the implementation of a custom loss function demonstrates that it is possible to align the model's learning process with specific chemical intuitions, potentially improving or at least keeping the model's accuracy. In this context, the use of group contributions offers a significant advantage, as it enables the incorporation of established group contribution values from the literature to enhance the model's alignment with chemical knowledge.

Overall, this study paves the way for the development of more interpretable GNN models in chemistry. Furthermore, the ability to align model learning with established chemical knowledge while keeping the same accuracy underscores the transformative potential of integrating advanced ML techniques with domain-specific expertise. Future work should investigate other models and datasets to further validate and expand upon these results.

Acknowledgements

This work was supported by the Young Talents Fellowship granted by the National Centre of Competence "Sustainable chemical processes through catalyst design" (NCCR Catalysis); by the Portuguese Recovery and Resilience Plan through project C645008882-00000055, Center for Responsible AI; and by national funds through FCT – Foundation for Science and Technology, I.P. (grant number UI/BD/153496/2022), within the scope of the project CISUC (UID/CEC/00326/2020).

References

Anderson, B., Hy, T.-S., and Kondor, R. Cormorant: Covariant molecular neural networks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Anstine, D. M. and Isayev, O. Generative Models as an Emerging Paradigm in the Chemical Sciences. *Journal of the American Chemical Society*, 145(16):8736–8750, April 2023. ISSN 0002-7863, 1520-5126. doi: 10.1021/jacs.2c13467.

Aouichaoui, A. R. N., Fan, F., Mansouri, S. S., Abildskov, J., and Sin, G. Combining Group-Contribution Concept and Graph Neural Networks Toward Interpretable Molecular Property Models. *Journal of Chemical Information and Modeling*, 63(3):725–744, February 2023. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.2c01091.

Benson, S. W. *Thermochemical Kinetics: Methods for the Estimation of Thermochemical Data and Rate Parameters*. Wiley, 1968. ISBN 978-0-471-06780-1.

Bihani, V., Mannan, S., Pratiush, U., Du, T., Chen, Z., Miret, S., Micoulaut, M., Smedskjaer, M. M., Ranu, S., and Krishnan, N. M. A. EGraFFBench: Evaluation of equivariant graph neural network force fields for atomistic simulations. *Digital Discovery*, 3(4):759–768, 2024. ISSN 2635-098X. doi: 10.1039/D4DD00027G.

Chen, L.-Y., Hsu, T.-W., Hsiung, T.-C., and Li, Y.-P. Deep Learning-Based Increment Theory for Formation Enthalpy Predictions. *The Journal of Physical Chemistry A*, 126(41):7548–7556, October 2022. ISSN 1089-5639, 1520-5215. doi: 10.1021/acs.jpca.2c04848.

Gani, R. Group contribution-based property estimation methods: Advances and perspectives. *Current Opinion in Chemical Engineering*, 23:184–196, March 2019. ISSN 22113398. doi: 10.1016/j.coche.2019.04.007.

Hoerl, A. E. and Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 42(1):80–86, February 2000. ISSN 0040-1706, 1537-2723. doi: 10.1080/00401706.2000.10485983.

Janet, J. P. and Kulik, H. J. *Machine Learning in Chemistry*. ACS In Focus. American Chemical Society, Washington, DC, USA, May 2020. ISBN 978-0-8412-9900-9. doi: 10.1021/acs.infocus.7e4001.

Jiménez-Luna, J., Grisoni, F., and Schneider, G. Drug discovery with explainable artificial intelligence. *Nature machine intelligence*, 2(10):573–584, October 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00236-4.

Kakkad, J., Jannu, J., Sharma, K., Aggarwal, C., and Medya, S. A Survey on Explainability of Graph Neural Networks. https://arxiv.org/abs/2306.01958, 2023.

Landrum, G., Tosco, P., Kelley, B., Rodriguez, R., Cosgrove, D., Vianello, R., sriniker, gedeck, Jones, G., NadineSchneider, Kawashima, E., Nealschneider, D., Dalke, A., Swain, M., Cole, B., Turk, S., Savelev, A., Vaucher, A., Wójcikowski, M., Take, I., Scalfani, V. F., Walker, R., Ujihara, K., Probst, D., guillaume godin, Pahl, A., Lehtivarjo, J., Berenger, F., jasondbiggs, and strets123. Rdkit/rdkit: 2024_03_1 (Q1 2024) Release. Zenodo, May 2024.

Li, Y., Zhou, J., Verma, S., and Chen, F. A Survey of Explainable Graph Neural Networks: Taxonomy and Evaluation Metrics. https://arxiv.org/abs/2207.12599, 2022.

Pinsky, E. and Klawansky, S. MAD (about median) vs. quantile-based alternatives for classical standard deviation, skewness, and kurtosis. *Frontiers in Applied Mathematics and Statistics*, 9:1206537, June 2023. ISSN 2297-4687. doi: 10.3389/fams.2023.1206537.

Ramakrishnan, R., Dral, P. O., Rupp, M., and Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, August 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22.

Rasmussen, M. H., Christensen, D. S., and Jensen, J. H. Do machines dream of atoms? Crippen's logP as a quantitative molecular benchmark for explainable AI heatmaps. https://chemrxiv.org/engage/chemrxiv/article-details/6388991244ccbc1731090a96, December 2022.

Rodríguez-Pérez, R. and Bajorath, J. Explainable Machine Learning for Property Predictions in Compound Optimization: Miniperspective. *Journal of Medicinal Chemistry*, 64(24):17744–17752, December 2021. ISSN 0022-2623, 1520-4804. doi: 10.1021/acs.jmedchem.1c01789.

Satorras, V. G., Hoogeboom, E., and Welling, M. E(n) Equivariant Graph Neural Networks. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9323–9332. PMLR, July 2021. doi: 10.48550/ARXIV.2102.09844.

Schütt, K. T., Kindermans, P.-J., Sauceda, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 992–1002, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4.

Walter, M., Webb, S. J., and Gillet, V. J. Interpreting Neural Network Models for Toxicity Prediction by Extracting Learned Chemical Features. *Journal of Chemical Information and Modeling*, 64(9):3670–3688, May 2024. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.4c00127.

Wellawatte, G. P., Gandhi, H. A., Seshadri, A., and White, A. D. A Perspective on Explanations of Molecular Prediction Models. *Journal of Chemical Theory and Computation*, 19(8):2149–2160, April 2023. ISSN 1549-9618, 1549-9626. doi: 10.1021/acs.jctc.2c01235.

Wu, Z., Wang, J., Du, H., Jiang, D., Kang, Y., Li, D., Pan, P., Deng, Y., Cao, D., Hsieh, C.-Y., and Hou, T. Chemistry-intuitive explanation of graph neural networks for molecular property prediction with substructure masking. *Nature Communications*, 14(1):2585, May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38192-3.

Yuan, H., Yu, H., Gui, S., and Ji, S. Explainability in Graph Neural Networks: A Taxonomic Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–19, 2022. ISSN 0162-8828, 2160-9292, 1939-3539. doi: 10.1109/TPAMI.2022.3204236.