Comparative Analysis of Model-Agnostic Explanation Methods in Materials Science

Anna Przybyłowska [1, †], Joanna Neubauer [1], Monika Sztuder [1], Witold Taisner [1], Agnieszka Hryniewicka [2], Julia Kolodziejczyk [2], Marta Eliza Plonska-Brzezinska [2], Dariusz Brzezinski [1]

[1] Institute of Computing Science, Poznan University of Technology, Piotrowo 2, 61-138 Poznań, Poland
[2] Department of Organic Chemistry, Faculty of Medicine with the Division of Dentistry and Division of Medical Education In English, Medical University of Bialystok, ul. A. Mickiewicza 2A, 15-222 Białystok, Poland
†corresponding author's email: anna.przybylowska@cs.put.poznan.pl

Machine learning models are increasingly deployed in materials science to predict molecular properties and accelerate the discovery process. However, the opaque nature typical for these models makes it difficult to trace the reasoning behind their predictions. Explainable Artificial Intelligence (XAI) aims to resolve this lack of transparency. This is particularly essential in chemistry and materials science, where structure-property relationships guide the design of novel molecules. Yet, the current lack of standardized XAI benchmark tailored to molecular data leaves researchers without clear guidance on which explainers to trust.

To address this gap, our recent study [1] introduces a comparative analysis of selected model-agnostic XAI techniques, including both factual [2, 3, 4] and counterfactual [5, 6] methods (Figure 1). We systematically evaluate these approaches using molecular fingerprints – a representation common in molecular property prediction – at both the local, instance-specific level and the global, model-wide level. We quantitatively benchmark the explanations generated by each method against predictive faithfulness and available ground-truth data. Furthermore, we complement this numerical evaluation with a qualitative assessment by chemical domain experts.

Our results demonstrate that while general-purpose XAI methods, particularly SHAP [2], can yield faithful chemical explanations, significant inconsistencies exist across different XAI algorithms, including a disagreement on direction of feature importance in two different SHAP implementations. We further observe that factual and counterfactual methods frequently diverge in their focus, and explanations that score high in quantitative predictive faithfulness can simultaneously lack plausibility when evaluated by domain experts. The observed discrepancies suggest that relying on a single explainer can introduce bias into scientific interpretations. To mitigate this risk, we demonstrate the benefits of aggregating explanations, highlighting robust rank aggregation [7] as a particularly effective approach. Furthermore, the divergence between human expert assessments and quantitative scores indicates that standard XAI evaluation metrics fail to fully capture chemical nuances, emphasizing the need to develop better-suited, domain-specific evaluation frameworks. Finally, we hypothesize that the main cause of the observed discrepancies in the generated explanations may be the failure of current XAI methods to account for inherent feature dependencies. This highlights a potential future direction for developing truly chemistry-aware explainers.

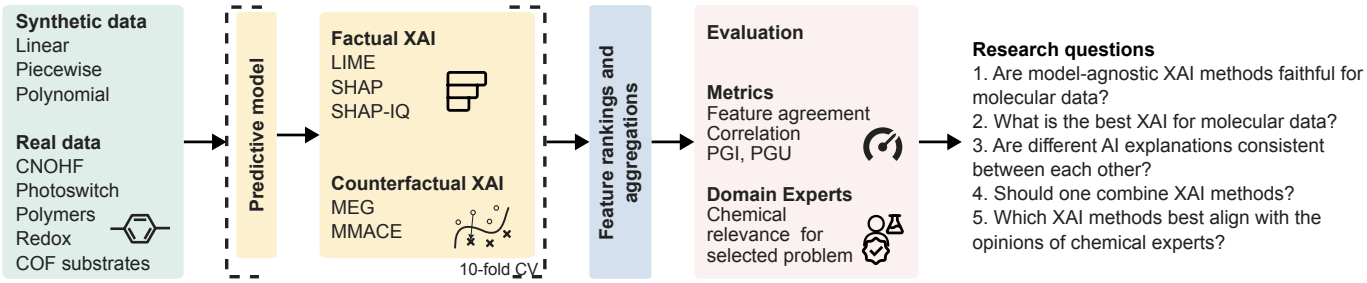

Figure 1: Overview of the study [1].

References

[1] A. Przybyłowska et al., "Comparative Analysis of Model-Agnostic Explanation Methods in Materials Science", *Advanced Intelligent Systems*, e70503 (2026).
[2] M. Lundberg and S.-I. Lee, " A Unified Approach to Interpreting Model Predictions," *NeurIPS* (2017).
[3] M. Muschalik et al., "shapiq: Shapley Interactions for Machine Learning," *NeurIPS* (2024).
[4] M. T. Ribeiro et al., ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *ACM SIGKDD* (2016).
[5] G. P. Wellawatte, A. Seshadri, and A. D. White, "Model Agnostic Generation of Counterfactual Explanations for Molecules," *Chemical Science* **13**, 3697 (2022).
[6] D. Numeroso and D. Bacciu, "MEG: Generating Molecular Counterfactual Explanations for Deep Graph Networks," *IJCNN* (2021).
[7] R. Kolde, S. Laur, P. Adler, and J. Vilo, "Robust Rank Aggregation for Gene List Integration and Meta-analysis," *Bioinformatics* **28** (4), 573 (2012).