Comparative analysis of model-agnostic explanation methods in materials science

Anna Przybyłowska, Joanna Neubauer, Monika Sztuder, Witold Taisner, Dariusz Brzezinski
Institute of Computing Science
Poznan University of Technology
ul. Piotrowo 2, 60-965 Poznan, Poland
{aprzybylowska,wtaisner,dbrzezinski}@cs.put.poznan.pl

Abstract

Machine learning models have become a powerful tool in materials science, accelerating the discovery process by accurately predicting molecular properties. However, the black-box nature of many of these models makes it difficult to understand or validate the reasoning behind their predictions. Consequently, explainable artificial intelligence (XAI) has emerged as a field of study aimed at making these models more transparent. This need for explainability is crucial in chemistry and materials science, as identified structure-property relationships can be utilized to guide the design of novel molecules. Yet, there is still a lack of XAI benchmarks focused on molecular data. As a preliminary step to address this gap, this study presents a comparative analysis of five selected XAI methods on molecular fingerprints, which are commonly used representations for property prediction tasks. Our results reveal significant discrepancies in the feature importance rankings generated by different XAI methods, demonstrating that the choice of explanation approach can introduce bias and alter scientific interpretation in the material discovery process.

1 Introduction

To overcome the limitations of time-consuming experimental work, material scientists increasingly integrate computational approaches into their discovery workflows. In particular, machine learning (ML) models help prioritize promising candidates for synthesis by rapidly predicting molecular properties. However, while complex ML models, such as graph neural networks (GNNs) [21], can achieve state-of-the-art results, their application in materials science is often limited by data scarcity. Consequently, researchers often need to rely on classic ML methods and tabular data representations. Among these representations, molecular fingerprints [26] are often used due to their simplicity and efficiency, offering high predictive accuracy that can even outperform GNN-based approaches [4]. However, a significant challenge remains, as the black box nature of many ML models prevents chemists from understanding the structure-property relationships learned by the model, which could help guide the discovery of novel molecules [28].

This lack of transparency of ML models underscores the need for explainable artificial intelligence (XAI) [5]. XAI methods can be categorized as offering *global* (model-level) or *local* (instance-level) explanations and can be further divided into either *factual* or *counterfactual*, based on whether they provide feature importance or counterexamples [5]. Despite the existence of numerous XAI benchmarks for tabular [1, 8, 9, 13] and graph [2, 10, 17] datasets, their findings do not necessarily extend to the chemical domain. General-purpose tabular XAI benchmarks lack the high dimensionality or structure typical of molecular descriptors, whereas graph-based benchmarks typically evaluate explanations based on abstract topological patterns rather than chemically meaningful substructures or are limited to a narrow analysis for only drug-like molecules.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4Mat: AI for Accelerated Materials Design.

Similarly, existing molecular XAI comparisons also present specific limitations – some lack comparative analysis of methods [7] or focus on a narrow subset or a single explanation approach [31, 27]. Others use only synthetic datasets [12], which may not capture the complexity of real-world tasks, or focus on just one predictive target [20]. Furthermore, other comparisons [18, 19] are model-specific and focus on explanations solely for GNNs. This demonstrates a need for a comparative XAI benchmark on tabular molecular representations.

This work addresses the aforementioned gap by comparing factual and counterfactual, local, model-agnostic, post-hoc XAI frameworks for their application to molecular property prediction tasks. Our experimental analysis is conducted at a model level, where we identify the features deemed most important by each XAI method. These features are then benchmarked against predictive faithfulness and, where possible, ground-truth faithfulness. An overview of the study is presented in Figure 1.

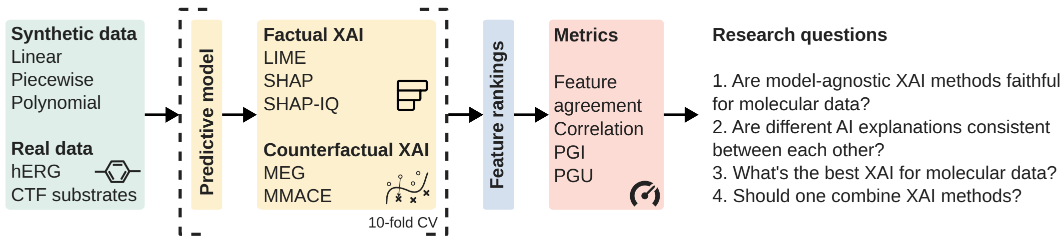


Figure 1: Schematic overview of the study.

2 Methods

2.1 Datasets

Our evaluation is conducted on two types of datasets: synthetic datasets with known ground-truth explanations and two real-world datasets where the influential features are not defined.

The synthetic prediction problems were generated from the QM9 dataset [34]. First, all molecules were represented as count-based ECFP fingerprints [24] from which we identified a subset of molecules with the highest fingerprint complexity, defined by the number of non-zero features. To mimic the problem of a lack of data often seen in real-world scenarios, we then sampled 200 instances from this subset for the final dataset. For this dataset, we defined three target functions, each based on six selected ECFP features, hence creating three synthetic predictive tasks with known ground-truth explanations. To ensure a comprehensive evaluation, the target functions included linear, piecewise linear, and polynomial functions. Further details regarding the representation of molecules and the definition of target functions can be found in Appendix A.1.

For the real-world case studies, we used the hERG dataset [3] and a dataset of Covalent Organic Framework (COF) substrates. The hERG dataset contains drug-like molecules with their corresponding pIC50 potency values for estimating the cardiotoxicity risk of a molecule. From the hERG dataset, we selected a subset of 200 molecules and represented them using ECFP fingerprints. Further details on the hERG dataset can be found in Appendix A.2. The COF dataset contains COF substrates with experimentally measured specific capacitance (F/g). COFs are a promising type of electrode material for supercapacitors due to their desirable characteristics such as high surface area and versatile electronic structure [25]. We represent these COF substrates using ECFP fingerprints and custom descriptors. A detailed description of all the features utilized for COFs can be found in Appendix A.3.

2.2 Selected XAI approaches

Our comparison focuses on model-agnostic, post-hoc XAI methods that are applicable to tabular molecular descriptors. To benchmark both domain-agnostic and chemistry-specific explainers, we selected a set of five methods, consisting of three general-purpose, factual approaches (SHAP, SHAP-IQ, LIME) and two counterfactual methods developed for molecular data (MMACE, MEG).

The selected factual methods can be divided into two categories. SHAP [11] and SHAP-IQ [14] are game-theoretic approaches that quantify feature importance by fairly distributing the model's prediction. While SHAP calculates the attribution of each feature individually, SHAP-IQ also analyzes feature interactions. In contrast, LIME [22] approximates individual predictions using a simple surrogate model and then derives the explanation from this approximation.

The counterfactual methods both focus on the exploration of the chemical space. MMACE [32] starts by exploring the local chemical space around a given instance using the STONED algorithm [15], and then provides a diverse set of counterfactuals from different chemical space clusters. On the other hand, MEG [16] employs reinforcement learning with a task-specific reward function to generate valid molecules as counterfactuals.

We note that features in ECFP fingerprints are dependent, which might hinder the performance of some factual methods. However, these methods were widely used in some prior work [23, 30, 29, 33] for analysis of the model's results, therefore, it is important to include them in this comparison.

2.3 Experimental evaluation

Our experimental evaluation (Figure 1) uses 10-fold cross-validation to compare the analyzed XAI methods. In each fold, we train a predictive model and then employ the selected XAI approaches to explain the model's predictions on the test set. We use two types of predictive models: the ground-truth function, applied to synthetic data, and a Random Forest (RF) model, applied to all datasets. Regarding the XAI methods, we compare LIME, SHAP, SHAP-IQ with first-order effects (SHAP-IQ-1), SHAP-IQ with second-order effects (SHAP-IQ-2), and the counterfactual explainers MEG and MMACE, each generating 25 counterfactuals per instance.

To perform the comparison, we derive a feature importance ranking from each XAI method. For the factual methods, we calculate the absolute average importance for each feature. For counterfactual approaches, the ranking is based on the frequency with which each feature was altered in the valid generated counterfactuals.

To evaluate the explanations, we use three metrics from the OpenXAI benchmark [1]: feature agreement (FA), important feature perturbation (PGI), and unimportant feature perturbation (PGU). While FA measures the XAI method's agreement with the known ground-truth function, PGI and PGU assess its faithfulness to the predictive model by analyzing changes in the model's prediction after perturbing the most and least important features, respectively. Furthermore, we analyze correlations between method rankings and evaluate the performance of an aggregated ranking. Further details on model and XAI method parameters, along with a description of the metrics can be found in Appendix B.

3 Results

The evaluation scores for the XAI methods applied to the ground-truth functions and Random Forest models are presented in Tables 1. 2a, and 2b, revealing interesting insights into the application of selected XAI methods to chemical tasks.

Table 1: Mean values and standard deviations (in parentheses, in units of the last significant digit of the mean value) of explainability metrics calculated for the analyzed XAI approaches explaining the ground-truth functions on the synthetic datasets: Linear, Piecewise Linear, and Polynomial. Best results highlighted in bold, second-best underlined.

	Linear			Piecewise Linear			Polynomial		
	FA ↑	PGI ↑	PGU ↓	FA ↑	PGI ↑	PGU ↓	FA ↑	PGI ↑	PGU ↓
LIME	**1.0(0)**	<u>20.6(7)</u>	**0.7(0)**	**1.0(0)**	<u>16.8(8)</u>	0.9(1)	**1.0(0)**	**18.6(9)**	**0.7(1)**
SHAP	**1.0(0)**	**20.7(7)**	**0.7(0)**	**1.0(0)**	**16.9(10)**	**0.6(1)**	**1.0(0)**	<u>18.5(8)</u>	**0.7(0)**
SHAP-IQ-1	**1.0(0)**	**20.7(7)**	**0.7(0)**	**1.0(0)**	**16.9(10)**	**0.6(1)**	**1.0(0)**	<u>18.5(10)</u>	**0.7(1)**
SHAP-IQ-2	**1.0(0)**	**20.7(6)**	5.5(3)	**1.0(0)**	<u>16.8(10)</u>	6.3(16)	**1.0(1)**	18.3(10)	8.7(16)
MMACE	<u>0.9(0)</u>	20.5(6)	2.2(4)	0.9(1)	<u>16.7(9)</u>	1.4(4)	<u>0.9(1)</u>	18.4(9)	2.0(7)
MEG	<u>0.9(1)</u>	20.4(5)	2.9(6)	0.8(1)	**16.9(9)**	2.0(5)	<u>0.8(0)</u>	18.3(9)	3.7(5)
Aggregated	**1.0(0)**	<u>20.6(8)</u>	<u>0.9(3)</u>	**1.0(0)**	<u>16.8(10)</u>	<u>0.7(1)</u>	**1.0(0)**	**18.6(9)**	<u>1.0(6)</u>

Table 2: Means and standard deviations (in parentheses, in units of the last significant digit of the mean value) of explainability metrics calculated for the XAI approaches explaining a Random Forest model on (a) the synthetic datasets: Linear, Piecewise Linear, and Polynomial, and (b) real datasets: hERG and COF. Best results highlighted in bold, second-best underlined.

(a) Results for synthetic datasets.

	Linear		Piecewise Linear		Polynomial	
	PGI ↑	PGU ↓	PGI ↑	PGU ↓	PGI ↑	PGU ↓
LIME	16.2(11)	1.2(2)	14.3(14)	1.4(3)	15.8(12)	1.2(2)
SHAP	16.2(12)	**1.1(1)**	14.2(14)	**1.1(2)**	15.5(12)	**1.0(2)**
SHAP-IQ-1	16.3(11)	1.2(5)	14.0(14)	1.6(12)	**16.0(12)**	1.2(3)
SHAP-IQ-2	16.3(12)	2.1(15)	14.1(15)	3.0(21)	15.7(11)	1.6(3)
MMACE	16.3(12)	1.6(2)	**14.4(16)**	2.3(7)	15.6(10)	2.0(5)
MEG	16.2(11)	1.7(3)	14.3(15)	2.2(7)	15.4(10)	2.5(5)
Aggregated	**16.5(12)**	1.2(2)	14.3(15)	1.3(3)	15.6(11)	1.2(2)

(b) Results for real-world datasets.

	hERG		COF	
	PGI ↑	PGU ↓	PGI ↑	PGU ↓
LIME	0.59(22)	0.12(4)	91.4(144)	21.8(50)
SHAP	**0.60(22)**	**0.11(2)**	92.8(150)	**17.9(42)**
SHAP-IQ-1	**0.60(21)**	0.13(4)	89.8(140)	27.7(68)
SHAP-IQ-2	0.59(20)	0.29(19)	89.8(142)	45.9(98)
MMACE	0.58(19)	0.20(10)	89.9(148)	27.9(57)
MEG	0.57(19)	0.21(10)	86.4(138)	29.9(54)
Aggregated	0.59(21)	0.12(3)	**93.4(142)**	23.3(48)

First, as can be noticed in Table 1, the ground-truth faithfulness (FA) scores on the synthetic datasets were perfect for most XAI methods, indicating that they correctly identified important features. The only exceptions were counterfactual explanations, most likely because modifying all six ground-truth features was not always necessary to generate a valid counterfactual (for details on the validity of counterfactuals, see Appendix C.1).

Second, regarding predictive faithfulness (PGI and PGU, Tables 1, 2a, and 2b), while no single method proved consistently superior across all scenarios, the performance differences were most evident on the more complex real-world datasets, where SHAP achieves the best mean PGU results. Notably, there were discrepancies between SHAP and SHAP-IQ-1, both of which approximate first-order Shapley values. This suggests that the underlying implementation details can significantly impact explanations, and potentially the insights experts extract from them; exemplary inconsistencies are discussed in Appendix C.2.

Third, an analysis of ranking correlations (Figures 2 and 8) reveals that the methods form distinct clusters, which evolve with the complexity of the problem. For simple linear functions, we can observe two main groups: factual and counterfactual approaches. However, as the complexity of the problem increases, the methods become more distinct. These differences, combined with similar FA, PGI, and PGU results, emphasize that each XAI method focuses on some particular aspect of the model's behavior.

Finally, the usage of the aggregated ranking can be beneficial as it integrates diverse perspectives of the XAI methods. It consistently scores in the top-2 methods on most of the analyzed scenarios, highlighting its value as a reliable alternative to using a single explainer. It does not strongly align with any single method (Figures 2 and 8), but rather exhibits a similar, mild correlation to all of them, showcasing its role as a well-balanced compromise. The analysis of the properties of such aggregated explanations in the context of chemistry and materials science is part of our ongoing work.

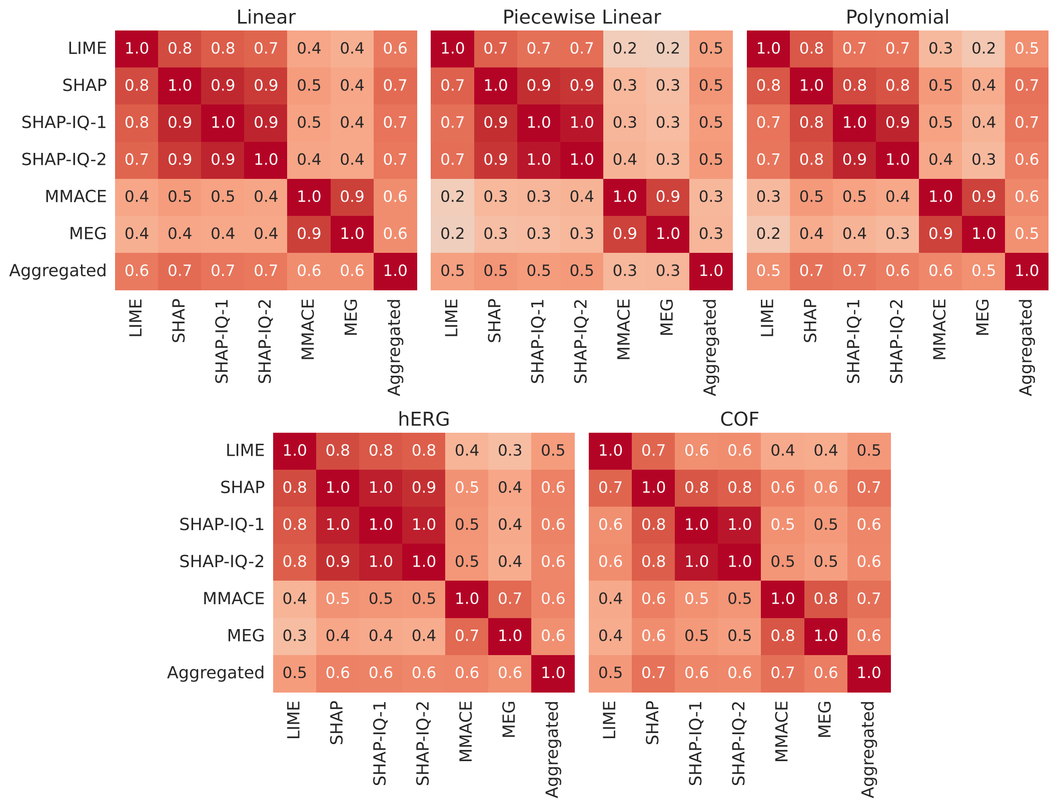


Figure 2: The mean correlation matrices between the feature rankings of XAI methods explaining RF models.

4 Discussion

In this work, we analyzed the properties and relations between selected XAI approaches, including factual (LIME, SHAP, SHAP-IQ-1, SHAP-IQ-2) and counterfactual (MMACE, MEG) methods. Results of experiments on synthetic and real data highlight inconsistencies between XAI methods. In the context of molecular property prediction, this means that, depending on the explanation method used, the chemical insights may differ.

However, to build a more complete understanding of the relationships between different XAI methods in the context of chemistry and materials science, it is necessary to expand the scope of our comparison. First, we plan to incorporate more complex and diverse synthetic target functions, as well as real-world molecular datasets. We also aim to add other predictive models to ensure the generalizability of our findings. Furthermore, the analysis will be extended by including other XAI methods, such as rule-based explainers, and will explore the integration of explainability and causality. Moreover, the promising performance of the aggregated feature ranking needs further investigation. Finally, we plan to evaluate the practicality of the XAI methods using surveys and direct interactions with chemists.

Acknowledgments and Disclosure of Funding

This research was partly funded by the National Science Centre, Poland, grant numbers 2023/49/B/ST5/02403 and 2022/47/D/ST6/01770.

References

[1] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, Isha Puri, Marinka Zitnik, and Himabindu Lakkaraju. OpenXAI: Towards a transparent evaluation

of model explanations. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[2] Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *Scientific Data*, 10(1):144, Mar 2023.

[3] Issar Arab and Khaled Barakat. Toxtree: descriptor-based machine learning models for both herg and nav1.5 cardiotoxicity liability predictions, 2021.

[4] Delora Baptista, João Correia, Bruno Pereira, and Miguel Rocha. Evaluating molecular representations in machine learning models for drug response prediction and interpretability. *Journal of Integrative Bioinformatics*, 19(3):20220006, 2022.

[5] Francesco Bodria, Fosca Giannotti, Riccardo Guidotti, Francesca Naretto, Dino Pedreschi, and Salvatore Rinzivillo. Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37(5):1719–1778, Sep 2023.

[6] Darko Butina. Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750, 1999.

[7] Kevser Kübra Kırboğa, Sumra Abbasi, and Ecir Uğur Küçüksille. Explainability and white box in drug discovery. *Chem. Biol. Drug Des.*, 102(1):217–233, July 2023.

[8] Jun Rui Lee, Sadegh Emami, Michael David Hollins, Timothy C. H. Wong, Carlos Ignacio Villalobos Sánchez, Francesca Toni, Dekai Zhang, and Adam Dejl. Xai-units: Benchmarking explainability methods with unit tests. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, page 2892–2905, New York, NY, USA, 2025. Association for Computing Machinery.

[9] Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. In *Advances in Neural Information Processing Systems Datasets Track*, 2021.

[10] Antonio Longa, Steve Azzolin, Gabriele Santin, Giulia Cencetti, Pietro Lio, Bruno Lepri, and Andrea Passerini. Explaining the explainers in graph neural networks: a comparative study. *ACM Comput. Surv.*, 57(5), January 2025.

[11] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[12] Mariia Matveieva and Pavel Polishchuk. Benchmarks for interpretation of qsar models. *Journal of Cheminformatics*, 13(1):41, May 2021.

[13] Catarina Moreira, Yu-Liang Chou, Chihcheng Hsieh, Chun Ouyang, João Pereira, and Joaquim Jorge. Benchmarking instance-centric counterfactual algorithms for xai: From white box to black box. *ACM Comput. Surv.*, 57(6), February 2025.

[14] Maximilian Muschalik, Hubert Baniecki, Fabian Fumagalli, Patrick Kolpaczki, Barbara Hammer, and Eyke Hüllermeier. shapiq: Shapley interactions for machine learning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[15] AkshatKumar Nigam, Robert Pollice, Mario Krenn, Gabriel dos Passos Gomes, and Alán Aspuru-Guzik. Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (stoned) algorithm for molecules using selfies. *Chem. Sci.*, 12:7079–7090, 2021.

[16] Danilo Numeroso and Davide Bacciu. Meg: Generating molecular counterfactual explanations for deep graph networks. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021.

[17] Mario Alfonso Prado-Romero, Bardh Prenkaj, Giovanni Stilo, and Fosca Giannotti. A survey on graph counterfactual explanations: Definitions, methods, evaluation, and research challenges. *ACM Comput. Surv.*, 56(7), April 2024.

[18] Magdalena Proszewska, Tomasz Danel, and Dawid Rymarczyk. B-xaic dataset: Benchmarking explainable ai for graph neural networks using chemical data, 2025.

[19] Jiahua Rao, Shuangjia Zheng, Yutong Lu, and Yuedong Yang. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Patterns*, 3(12), Dec 2022.

[20] Maria H. Rasmussen, Diana S. Christensen, and Jan H. Jensen. Do machines dream of atoms? Crippen's logP as a quantitative molecular benchmark for explainable AI heatmaps. *SciPost Chem.*, 2:002, 2023.

[21] Patrick Reiser, Marlen Neubert, André Eberhard, Luca Torresi, Chen Zhou, Chen Shao, Houssam Metni, Clint van Hoesel, Henrik Schopmans, Timo Sommer, and Pascal Friederich. Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1):93, Nov 2022.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[23] Raquel Rodríguez-Pérez and Jürgen Bajorath. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *Journal of Medicinal Chemistry*, 63(16):8761–8777, Aug 2020.

[24] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010.

[25] Ahmed G. Saad, Ahmed Emad-Eldeen, Wael Z. Tawfik, and Ahmed G. El-Deen. Data-driven machine learning approach for predicting the capacitance of graphene-based supercapacitor electrodes. *Journal of Energy Storage*, 55:105411, 2022.

[26] Jie Shen and Christos A. Nicolaou. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies*, 32-33:29–36, 2019. Artificial Intelligence.

[27] Robert P. Sheridan. Interpretation of qsar models by coloring atoms according to changes in predicted activity: How robust is it? *Journal of Chemical Information and Modeling*, 59(4):1324–1337, 2019. PMID: 30779563.

[28] Lei Tao, Guang Chen, and Ying Li. Machine learning discovery of high-temperature polymers. *Patterns*, 2(4), Apr 2021.

[29] Sarveswara Rao Vangala, Navneet Bung, Sowmya Ramaswamy Krishnan, and Arijit Roy. An interpretable machine learning model for selectivity of small-molecules against homologous protein family. *Future Medicinal Chemistry*, 14(20):1441–1453, 2022. PMID: 36169035.

[30] Rafael F. Veríssimo, Pedro H. F. Matias, Mateus R. Barbosa, Flávio O. S. Neto, Brenno A. D. Neto, and Heibbe C. B. de Oliveira. Integrating machine learning and shap analysis to advance the rational design of benzothiadiazole derivatives with tailored photophysical properties. *Journal of Chemical Information and Modeling*, 65(15):7874–7886, 2025. PMID: 40300554.

[31] Geemi P. Wellawatte, Heta A. Gandhi, Aditi Seshadri, and Andrew D. White. A perspective on explanations of molecular prediction models. *Journal of Chemical Theory and Computation*, 19(8):2149–2160, Apr 2023.

[32] Geemi P. Wellawatte, Aditi Seshadri, and Andrew D. White. Model agnostic generation of counterfactual explanations for molecules. *Chem. Sci.*, 13:3697–3705, 2022.

[33] Leanne S. Whitmore, Anthe George, and Corey M. Hudson. Mapping chemical performance on molecular structures using locally interpretable explanations, 2016.

[34] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530, 2018.

A Datasets

A.1 Synthetic datasets

For the three synthetic datasets, we applied count-based ECFP fingerprints of length 128 and radius 2. We further filtered out near-constant columns, removing any feature where the most frequent value occurred in more than 85% of the samples.

Regarding the ground-truth functions, we defined them based on 6 selected ECFP features: ECFP10 (f_{10}), ECFP16 (f_{16}), ECFP30 (f_{30}), ECFP33 (f_{33}), ECFP81 (f_{81}), and ECFP123 (f_{123}). These features were selected for their diversity, enabling the creation of well-distributed and varied target property distributions. The full mathematical definitions of the functions are provided in Table 3.

Table 3: Definitions of ground-truth functions.

Function	Definition
Linear	$y = 8.5f_{30} + 10.5f_{123} - 3.5f_{10} + 3f_{16} - 2.5f_{81} + 5.5f_{33} + 30$
Piecewise Linear	$y = \begin{cases} 10.5f_{30} + 6.5f_{123} - 1.5f_{81} + 30 & \text{if } f_{10} < 1 \\ 5f_{30} + 13f_{123} - 2.5f_{16} + 30 & \text{if } 1 \leq f_{10} < 2 \\ -1.5f_{30} + 3.5f_{123} + 15.5f_{33} + 30 & \text{if } f_{10} \geq 2 \end{cases}$
Polynomial	$y = 2.5f_{81}^2 + 1.5f_{33}^2 + 7.5f_{30}f_{123} - 9.5f_{10} - 3.5f_{16} + 30$

Exemplary molecules and target distribution for each synthetic function can be seen in Figure 3.

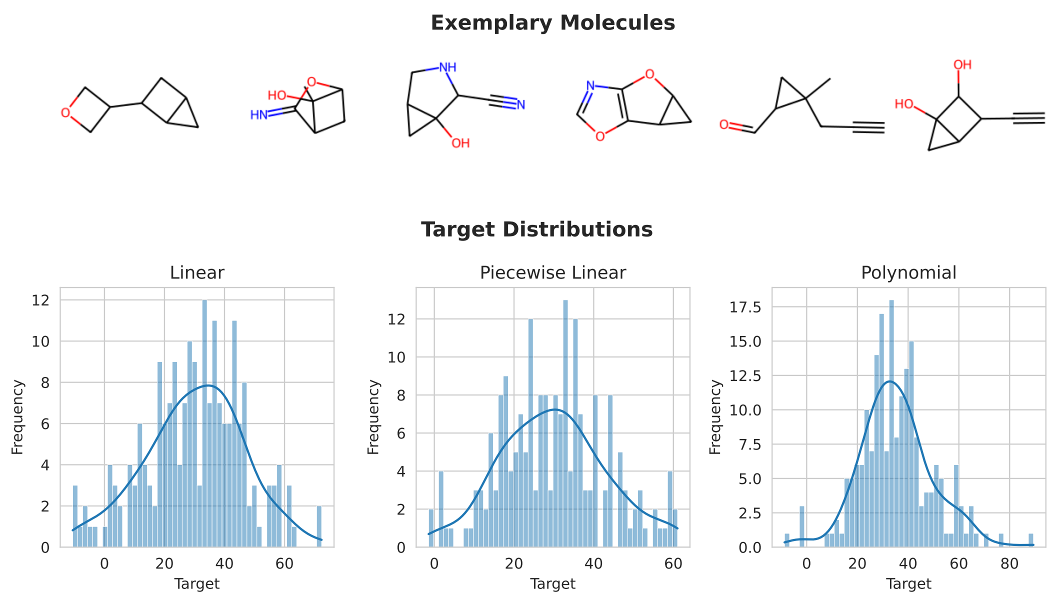


Figure 3: Exemplary molecules from a selected subset of QM9 dataset and target distribution for defined synthetic functions.

A.2 hERG dataset

For the hERG dataset, we first filtered out molecules based on their target value, resulting in a dataset of 8852 molecules. We then utilized Butina clustering [6] to select a diverse subset of 200 molecules. For this subset, we used count-based ECFP fingerprints with a length of 1024 and radius of 2.

Examples of molecules from the hERG subset and the target distribution in this subset are presented in Figure 4.

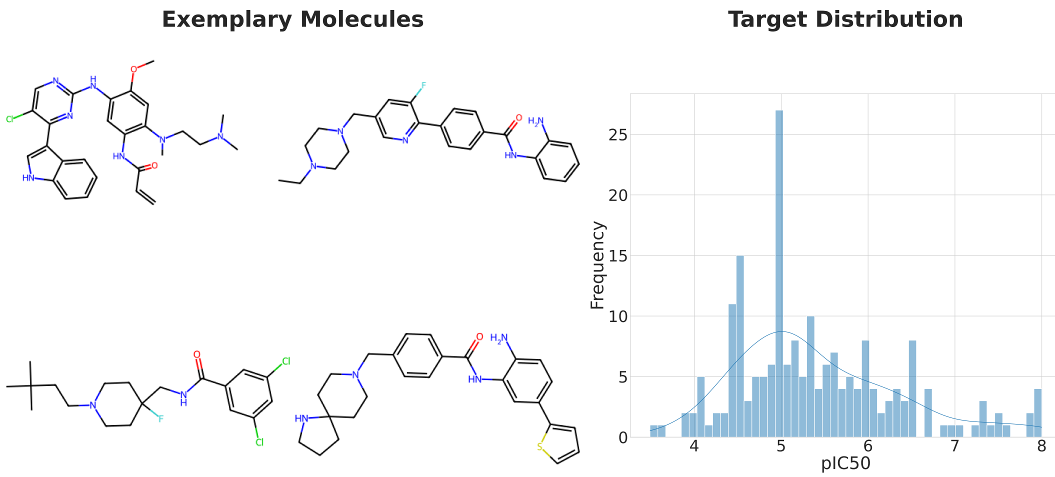


Figure 4: Exemplary molecules from a selected subset of hERG dataset and target distribution of this subset.

A.3 COF dataset

For the COF dataset, we first preprocessed the data by removing outliers based on the target variable, which resulted in a dataset of 102 molecules. For this dataset, we used a combination of count-based ECFP fingerprints and a set of custom expert-provided chemical descriptors.

We use ECFP fingerprints with a length of 1024 and radius of 2. A larger fingerprint length is specifically chosen to prevent bit collisions, as the molecules in the COF dataset generally exhibit higher similarity and are also larger than those in the QM9 dataset.

The custom set of descriptors included: radius, diameter, number of heteroatoms, number of rotatable bonds, number of H-bond acceptors and donors, topological polar surface area, molecular weight, and the percentages of oxygen, carbon, and nitrogen. The radius and diameter of a molecule are defined as the minimal length of the longest path between atoms and the maximum length of the shortest path, respectively.

While ECFP fingerprints encode the local substructures, the descriptors provide more global properties, combining information about molecular size, composition, and intermolecular interactions. Similarly to the synthetic dataset, we filtered out the near-constant columns.

Exemplary molecules along with target distribution for COF dataset are presented in Figure 5.

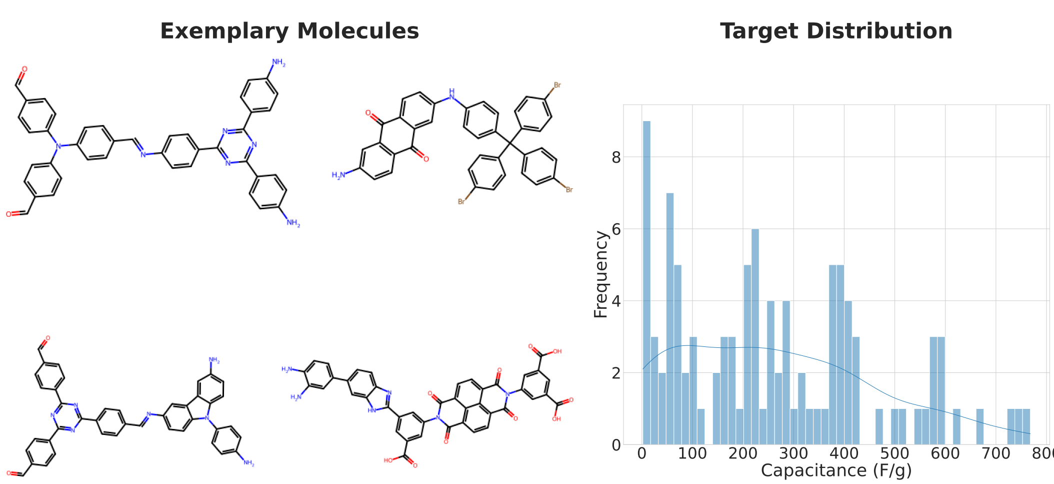


Figure 5: Exemplary molecules from COF dataset and distribution of capacitance in that dataset (target variable).

B Experimental details

B.1 Predictive models

For our comparative analysis, we employed two types of predictive models. First, for the synthetic datasets, we used the ground-truth functions, defined in Table 3, to provide a perfect, transparent model to which the explanations could be compared. Second, for all datasets, we trained Random Forest (RF) models. The RF model was trained to predict the synthetic target property and specific capacitance for the synthetic and COF datasets, respectively.

The RF models' hyperparameters were optimized within each cross-validation fold – the training data was split to find the best parameters, and then the final model was retrained on the entire, original training set for that fold.

The performance of RF models was evaluated using three metrics: root mean squared error (RMSE, Equation 1), symmetric mean absolute percentage error (SMAPE, Equation 2), and pairwise accuracy (PA, Equation 3). Pairwise accuracy was included in the evaluation because it is particularly relevant for applications in chemistry, where the predictive models are often used for initial screening of candidate molecules, making the correct relative ranking more important than precise value estimation.

$$\text{RMSE} = \sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{N}} \tag{1}$$

$$\text{SMAPE} = \frac{1}{n} \sum_i \frac{2|y_i - \hat{y}_i|}{|y_i| + |\hat{y}_i|} \tag{2}$$

$$\text{PA} = \frac{\sum_{i<j} \mathbb{I}(\text{sgn}(y_i - y_j) = \text{sgn}(\hat{y}_i - \hat{y}_j))}{\sum_{i<j} \mathbb{I}(y_i \neq y_j)} \tag{3}$$

The mean results of the RF models are presented in Table 4.

Table 4: Mean values and standard deviations (in parentheses) of predictive accuracy metrics calculated for the Random Forest models on the synthetic and real datasets. For COF dataset, RMSE is expressed in F/g.

	Dataset	RMSE ↓	SMAPE ↓	PA ↑
Synthetic	Linear	6.52(0.9)	0.29(0.1)	0.89(0.0)
	Piecewise Linear	8.05(3.1)	0.22(0.1)	0.82(0.1)
	Polynomial	7.81(1.4)	0.20(0.1)	0.83(0.0)
Real	hERG	0.85(0.1)	0.12(0.0)	0.67(0.1)
	COF	168.55(34.8)	0.59(0.1)	0.73(0.1)

B.2 XAI methods

LIME We employ the LIME explainer with the Ridge regression model as the local surrogate. The explainer is fitted on the training set of each fold.

SHAP We apply different SHAP variants for each model type – the TreeSHAP explainer for Random Forest models, and the KernelSHAP for the ground-truth functions. The KernelSHAP explainer is fitted on the training data for each fold.

SHAP-IQ Analogous to SHAP, we use TreeSHAP-IQ for Random Forest models and KernelSHAP-IQ for ground-truth functions. We analyze two configurations – SHAP-IQ-1, which computes standard first-order Shapley values, and SHAP-IQ-2, which calculates k-SII interaction indices with $k = 2$.

MMACE For the MMACE method, we first had to define a counterfactual for regression. Since this is not straightforward, we defined a counterfactual as an example that satisfied the following conditions: the absolute difference between the prediction for the original instance and the generated example must be greater than a specified `delta`, and the model's prediction for the example must fall on the opposite side of the training set's median.

To generate counterfactuals, MMACE requires the following parameters: the maximal number of token mutations (`#mutations`) to the original molecule, the token alphabet (`alphabet`), and the number of counterfactual examples to sample from the molecular neighborhood using the STONED algorithm (`#samples`). Following the search, MMACE clusters the found counterfactuals based on Tanimoto similarity and selected binary fingerprint (`fingerprint`) and returns the desired number of counterfactuals (`#counterfactuals`). The exact values of the aforementioned parameters are presented in Table 5.

Table 5: MMACE parameters.

Parameter	Value
`delta`	Synthetic datasets: 5.0 hERG dataset: 0.5 COF dataset: 50.0
`#mutations`	2
`alphabet`	Basic alphabet from MMACE updated with alphabet based on the training set
`#samples`	1000
`fingerprint`	ECFP4 fingerprint
`#counterfactuals`	25

MEG Since MEG is a reinforcement learning-based approach, we designed its optimization function $R(c, o, w)$ (Equation 4), where c is the candidate counterfactual, o is the original example, and w is a weighting parameter. This optimization function is a weighted sum of two components: a validity reward $V(c, o)$ and a similarity reward $S(c, o)$.

The validity reward $V(c, o)$ (Equation 5) encourages the generation of valid counterfactuals using the same definition as in MMACE. It evaluates if the change in prediction $d(c, o)$ is both large enough (relative to a target Δ, defined as a maximal value between the training set median and user-defined `delta`) and in correct optimization direction D_{opt} (guided by a sign agreement term A).

The similarity reward $S(c, o)$ focuses on promoting the candidates most similar to the original instance. It measures Tanimoto similarity between binary fingerprints of the counterfactual and the original molecule.

$$R(w, c, o) = (1 - w)V(c, o) + wS(c, o) \tag{4}$$

$$V(c, o) = \begin{cases} 1 & |d(c, o)| \geq \Delta \wedge \text{sgn}(d(c, o)) = D_{\text{opt}} \\ \tanh(A(d(c, o), D_{\text{opt}}) \cdot \frac{|d(c,o)|}{\Delta}) & \text{otherwise} \end{cases} \tag{5}$$

Furthermore, MEG requires the definition of parameters related to the learning process of the explainer (`learning rate`, `batch size`, `#epochs`), fingerprint used to compute the Tanimoto similarity (`fingerprint`), allowed modification types (`removals`, `atom additions`, `bond additions`, `bonds between rings`, `ring sizes`, `no modifications`), and maximal modification size (`max steps`). The values set for these parameters can be found in Table 6.

Implementation details We utilized open-source implementations of the aforementioned XAI methods, adapting them as necessary to fit our experimental framework. The links to the repositories are given in Table 7. Furthermore, the code and data used in this study are publicly available on GitHub at `https://github.com/JNeubau/XAI-ChemBenchmark.git`.

Table 6: MEG parameters.

Parameter	Value
`delta`	Synthetic datasets: 5.0 hERG dataset: 0.5 COF dataset: 50.0
w	0.1
`#counterfactuals`	25
`learning rate`	0.0001
`batch size`	32
`#epochs`	1000
`fingerprint`	ECFP4 fingerprint
`removals`	allowed
`atom additions`	allowed
`bond additions`	allowed
`bonds between rings`	allowed
`ring sizes`	5, 6
`no modifications`	not allowed
`max steps`	2

Table 7: Links to the utilized repositories.

Method	Repository
LIME	marcotcr/lime
SHAP	shap/shap
SHAP-IQ	mmschlk/shapiq
MMACE	ur-whitelab/exmol
MEG	danilonumeroso/meg

B.3 XAI evaluation metrics

This section provides a more detailed overview of the three selected XAI evaluation metrics, including the adjustments made for our analysis.

FA (Equation 6) measures ground-truth faithfulness as the overlap between top-k features of the explanation's ranking S_k^R and ground-truth S_k^{GT}. PGI (Equation 7) and PGU (Equation 8) are both predictive faithfulness metrics, measuring the change in the model's prediction $M(x)$ when perturbing the top-k most and least important features, respectively. Following the benchmark, we compute the area under the curve (AUC) for these metrics for consecutive values of k. However, for FA, we restrict the analysis to $k \geq 6$. This is because our ground-truth is composed of six features, for which precise ranking is difficult to determine for piecewise linear and polynomial functions.

$$\text{FA} = \frac{|S_k^{GT} \cap S_k^R|}{\max(k, |S^{GT}|)} \tag{6}$$

$$\text{PGI} = \mathbb{E}_{\mathbf{x}' \sim \text{perturb}(x, \text{top-k explanation features})}[|M(x) - M(\mathbf{x}')|] \tag{7}$$

$$\text{PGU} = \mathbb{E}_{\mathbf{x}' \sim \text{perturb}(x, \text{non top-k explanation features})}[|M(x) - M(\mathbf{x}')|] \tag{8}$$

Furthermore, the metrics are adjusted to handle interactions for SHAP-IQ. For PGI and PGU, features in an interaction are perturbed simultaneously. For ranking, a feature's first appearance, whether alone or in an interaction, determines its position.

C Additional results

C.1 Counterfactuals results – validity and similarity

For completeness of the results, we report the validity (Table 8) and similarity (Table 9) scores for the counterfactuals generated by MEG and MMACE. Notably, both methods demonstrated lower validity on real-world datasets compared to synthetic ones. This gap was especially pronounced for the COF dataset, emphasizing the complex nature of this predictive problem.

Moreover, the comparison reveals a trade-off: MMACE consistently generated a higher percentage of valid counterfactuals, but MEG's valid examples exhibited higher mean similarity to the original instances.

Table 8: Fractions of valid counterfactuals generated by MMACE and MEG methods.

		MMACE	MEG
RF model	Linear	0.96	0.88
	Piecewise Linear	0.97	0.94
	Polynomial	0.99	0.97
	hERG	0.95	0.82
	COF	0.65	0.34
Ground-truth model	Linear	0.94	0.88
	Piecewise Linear	0.99	0.96
	Polynomial	0.99	0.97

Table 9: Mean values and standard deviations (in parentheses) of similarity score of valid counterfactuals generated by MMACE and MEG.

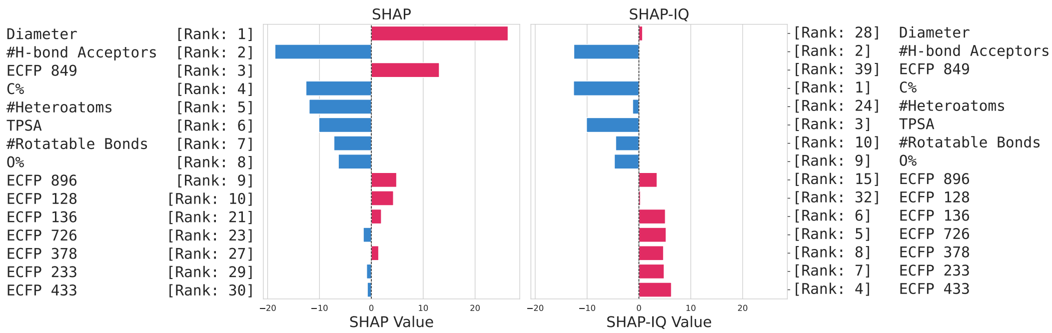
		MMACE	MEG
RF model	Linear	0.28(0.14)	0.32(0.10)
	Piecewise Linear	0.30(0.14)	0.34(0.10)
	Polynomial	0.30(0.14)	0.37(0.09)
	hERG	0.37(0.18)	0.62(0.11)
	COF	0.42(0.18)	0.62(0.10)
Ground-truth model	Linear	0.31(0.16)	0.34(0.11)
	Piecewise Linear	0.32(0.16)	0.36(0.10)
	Polynomial	0.31(0.14)	0.37(0.09)

C.2 Exemplary inconsistencies between SHAP and SHAP-IQ rankings

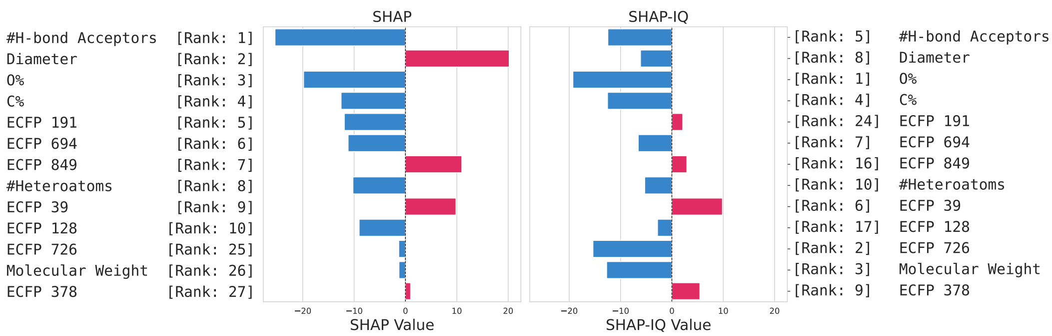
The unexpectedly low ranking correlation between SHAP and SHAP-IQ-1, as well as differences in evaluation metrics, led us to investigate their disagreements on specific examples. Figures 6 and 7 show representative instances from the COF dataset, where SHAP and SHAP-IQ top-10 feature rankings exhibited significant discrepancies. Notably, these differences did not involve only the feature order. The methods even disagreed on the direction of a feature's impact, as can be observed in the case of ECFP features in Figure 6a (for the molecule presented in Figure 7a) and Diameter feature in Figure 6b (for the molecule presented in Figure 7b).

C.3 Additional results: XAI methods for ground-truth functions

Figure 8 shows the ranking correlations of the selected XAI methods for ground-truth functions.

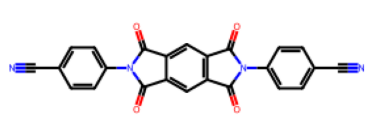


(a) Rankings for molecule $C_{24}H_{10}N_4O_4$

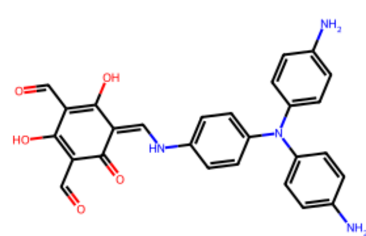


(b) Rankings for molecule $C_{27}H_{22}N_4O_5$

Figure 6: Representative instances from the COF dataset showcasing the discrepancies in the top-10 rankings of SHAP and SHAP-IQ.



(a) Molecule $C_{24}H_{10}N_4O_4$



(b) Molecule $C_{27}H_{22}N_4O_5$

Figure 7: Molecules from the COF dataset for which the discrepancies in the top-10 rankings of SHAP and SHAP-IQ are presented in Fig 6.

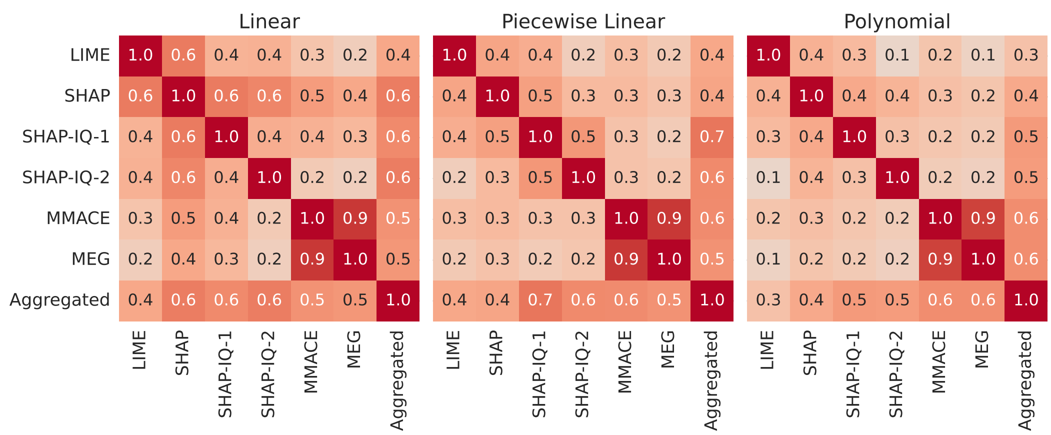


Figure 8: The mean correlation matrices between the rankings of XAI methods explaining ground-truth functions.