Explainable Artificial Intelligence for Bioactivity Prediction: Unveiling the Challenges with Curated CDK2/4/6 Breast Cancer Dataset

Adam Sułek[1] [0000−0002−4248−1047], Jakub Klimczak[1] [0009−0008−9884−9286], Jakub Jończyk[2,1] [0000−0002−9731−908X], Tomasz Kosciolek[1] [0000−0002−9915−7387], Tomasz Danel[3,1,*] [0000−0001−6053−0028], and Barbara Pucelik[4,*] [0000−0002−0235−6532]

[1] Sano Centre for Computational Medicine, 30-054 Kraków, Poland
[2] Department of Medicinal Chemistry, Faculty of Pharmacy, Jagiellonian University Medical College, 30-688 Kraków, Poland
[3] Faculty of Chemistry, Jagiellonian University, 30-387 Kraków, Poland
[4] Łukasiewicz Research Network, Kraków Institute of Technology, 30-418 Kraków, Poland
tomasz.danel@uj.edu.pl, barbara.pucelik@kit.lukasiewicz.gov.pl

Abstract. In recent years, the interplay between machine learning (ML) and cheminformatics has driven advancements in bioactivity prediction. However, the challenge of model explainability remains a significant barrier to adopting these approaches in drug discovery. This study addresses critical shortcomings in existing modeling techniques by examining the assumptions of feature independence and contribution additivity that are the foundation of traditional explainability methods. We investigate fingerprint-based and molecular graph models within quantitative structure-activity relationship modeling. While these models demonstrate impressive predictive performance, they offer limited actionable insights for medicinal chemists. To assist researchers in developing useful and interpretable activity prediction models, we propose a new benchmark based on the pharmacophore concept, commonly used in preliminary compound filtering. Furthermore, we introduce PharmacoScore, a novel evaluation metric designed to assess whether ML-based explanations prioritize essential pharmacophore components over non-critical features. Our findings highlight a crucial misalignment between ML model explanations and established pharmacophore principles, revealing a pressing need for innovative interpretability strategies in cheminformatics. This work not only offers a valuable resource but also sets the stage for future research, enhancing the transparency of ML in drug discovery.

Keywords: explainable artificial intelligence · machine learning · cheminformatics · bioactivity prediction · pharmacophores.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_2

1 Introduction

Explainable artificial intelligence (XAI) is a rapidly expanding field, mostly due to the need for understanding predictions of machine learning (ML) models in high-risk applications such as medical image analysis, fraud detection, and autonomous vehicle decision-making. [1] In drug discovery, XAI methods help medicinal chemists identify novel therapeutic molecules with improved bioactivity, which is the most crucial property in the early stages of drug discovery. Bioactivity refers to the effect that a compound has on the organism, usually caused by the activation or suppression of the selected protein targets. Nowadays, ML models are employed to predict bioactivity from the chemical structure to expedite the search for new potent molecules and reduce the cost related to unsuccessful experimental trials. Explainability techniques provide additional insights about model predictions that can be leveraged by medicinal chemists to propose more effective molecular designs. [4]

In bioactivity prediction, simple ML models were classically used to predict the activity of small molecules within a series of compounds sharing the same core, e.g. by fitting a multiple linear regression to a few molecular descriptors. This process is called quantitative structure-activity relationship (QSAR) modeling. [4] Currently, more advanced ML models are utilized for predicting activity across more diverse compound libraries obtained from large databases. One of the predominant approaches is applying models like random forest (RF) or support vector machines (SVM) to predict activity from molecular descriptors or fingerprints, which are feature vector representations. The other approach uses graph neural networks (GNN) that work directly on chemical structures represented as molecular graphs. [13] Inspired by current advancements in natural language processing (NLP), text representations like SMILES are sometimes used. [9] All of these techniques are purely ligand-based and provide an alternative to more computationally demanding structure-based methods like molecular docking. However, the quality of predictions of such methods depends on the quality of the training data, and most of these models fail to generalize to novel chemical spaces, providing accurate predictions primarily for the compounds close to the training dataset. [7] That is also why XAI methods should be employed to better understand the knowledge gained by these models and avoid overfitting to certain chemical structures.

Many XAI methods for molecules have been adopted from other domains, such as computer vision and NLP, where they proved to produce explanations that are effective and comprehensible for the users. However, the adaptation of these methods to the molecular domain is not always straightforward and can provide misleading insights if applied or interpreted incorrectly. For example, Local Interpretable Model-agnostic Explanations (LIME) [10] and SHapley Additive exPlanations (SHAP) [6] are explainability methods used for feature vector representations, but these techniques assume the features are independent, which is rarely held for typical molecular representations like calculated descriptors or fingerprints. [2,5] Additionally, LIME assumes that model behavior is locally linear, which may not be true for bioactivity prediction models

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_2

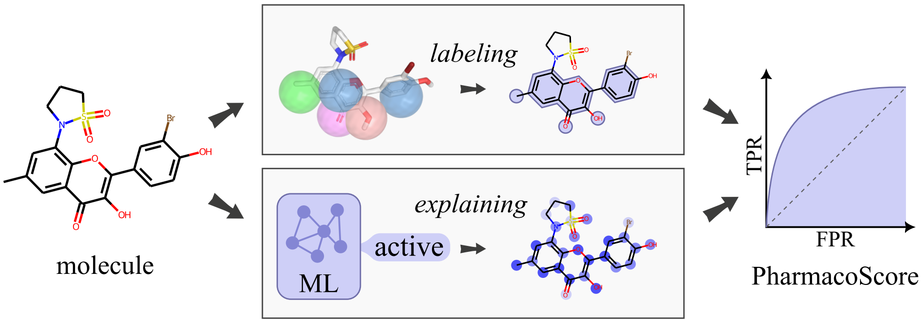

Fig. 1: Overview of our benchmark. ML methods predict compound activity, but only inherently interpretable models or post-hoc XAI techniques can identify the atoms that are significant for predictions. With our new PharmacoScore metric, we compare atom-based explanations against fragments that match a pharmacophore model used for data labeling.

due to so-called activity cliffs, which are pairs of similar molecules with huge differences in activity. SHAP assumes that predictions can be decomposed into a sum of effects attributable to each feature, but this may also be inaccurate for highly non-linear relations between compound structure and its activity. In the context of molecular graphs, saliency map techniques like Class Activation Mapping (CAM) [18] and Vanilla Gradients (VG) have been adopted from computer vision due to similarities in image and graph convolutions. [1] However, these methods are not particularly effective for elucidating how individual atoms contribute to the final prediction. For example, CAM explains predictions by multiplying atom activations after the last graph layer, which may not provide any meaningful signal due to the graph oversmoothing problem of GNNs.

We hypothesize that effective QSAR models, to accurately predict activity across diverse datasets of compounds, must be capable of identifying more general features that ensure binding to their macromolecular targets. One such category of features is a pharmacophore, which describes key points in three-dimensional space that will improve binding when a molecule matches that description. [15] Cyclin-dependent kinases (CDKs), a well-known class of anti-cancer drug targets, exemplify this by possessing a distinctive pharmacophore. For instance, CDK4/6 inhibitors, when combined with endocrine therapy, have become a milestone in managing hormone receptor-positive breast cancer (HR+BC). [3] However, the latest data revealed that the aberrant CDK2 plays a key role in driving resistance to CDK4/6 in HR+BC. Their ATP-competitive ligands typically contain a heterocyclic core with a hinge-binding motif and a hydrophobic component in the gatekeeper sub-pocket. [9]

To confirm that models can learn these high-level features, XAI methods can be applied since the models should prioritize these key features over more fine-grained structural modifications. We introduce a new dataset and an evaluation

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_2

metric based on pharmacophores to facilitate the testing of bioactivity prediction model generalization and the robustness of XAI methods (Figure 1). The dataset includes a primary collection of 2,131 molecules with experimentally measured CDK2 activity, along with 2,252 for CDK4 and 679 for CDK6, extended by an additional test set of generated decoy compounds. To simplify the complex problem of predicting activity, we propose a proxy task in which molecules are labeled by matching their structure to a pharmacophore hypothesis constructed from our activity dataset. Through this new benchmark, we make the following observations.

1. The evaluated XAI methods produce inconsistent explanations on our dataset that fail to highlight only the structures that should directly affect the predicted label.
2. Many commonly used ML models, such as RF, XGB, MLP, and GNN, face challenges in effectively learning the concept of a pharmacophore.
3. Activity-trained models share some common XAI patterns with pharmacophore-trained models, but they also reveal that activity data does not always align with the pharmacophore hypothesis.

2 Related work

The emergence of GNNs has prompted the creation of explainability methods for graph data. **GNN-Explainer** [16] is a model-agnostic approach designed to provide explanations for predictions made by GNNs. It does this by identifying a subgraph and node features that are most important for a given prediction. The approach formulates explanation generation as an optimization problem, where the goal is to find a compact subgraph that maximizes mutual information with the original model's prediction. **SubgraphX** [17] is an explainability method designed for GNNs that generates subgraph-based explanations in a self-interpretable manner. Instead of focusing on individual nodes or edges like traditional attribution-based methods, SubgraphX identifies the most influential subgraphs that contribute to the model's predictions. It employs a Monte Carlo Tree Search strategy to efficiently explore possible subgraphs and rank them based on their contribution to the final prediction. Unlike post-hoc methods that approximate model behavior, SubgraphX provides more stable and meaningful explanations by maintaining structural integrity within graphs. This makes it particularly effective for tasks in chemistry, biology, and social network analysis, where understanding group-wise interactions is critical.

In parallel, intrinsically interpretable graph-based models are being proposed. For example, **ProGReST** [11] achieves interpretability by integrating prototype learning, soft decision trees, and GNNs. This architecture enables predictions to be explained through learned prototypical parts, which serve as reference points for molecular structures. Unlike post-hoc explainability methods, ProGReST is inherently interpretable, as it directly links its decision-making process to identifiable molecular substructures. Additionally, the model ensures interpretability through tree-based reasoning, where each decision node is associated with meaningful prototypes that can be analyzed and validated by experts.

Various reports highlight the use of XAI in drug discovery projects. Wong et al. [14] present a novel model that combines GNNs with explainable graph algorithms to uncover chemical substructures linked to antibiotic activity. By integrating explainability, the model allows researchers to interpret the key molecular features driving antibiotic properties, paving the way for a more rational and efficient approach to antibiotic discovery. The **EvoGradient** [12] model combines explainable deep learning with virtual evolution to predict and optimize antimicrobial peptides (AMPs) in drug discovery. Utilizing LSTM, Transformer, and gradient-based analysis, the model achieves high predictive performance in AMP classification and potency prediction. EvoGradient outperforms traditional models by identifying key amino acids linked to bioactivity. However, like attention-based models, its explanations are sparse, highlighting broad regions rather than specific antimicrobial features. The **BiLAT** model [9], leveraging BiLSTM and Transformer-based encoding layers, achieves high predictive performance for CDK activity, surpassing traditional ML methods in identifying the CDK2 hinge motif, though its explanations remain sparse, highlighting large molecular regions.

3 Methods

This section summarizes the ML modeling techniques and explainability methods employed in this study. Following this, we address the dataset construction, detailing how the pharmacophore hypothesis was created for labeling our dataset. Lastly, we outline the evaluation metrics established in our benchmark. The code and data are available at `https://github.com/AdamSulek/pharmacoscore-benchmark`.

3.1 Activity prediction models

We trained XGBoost (XGB), Random Forest (RF), and a multilayer perceptron (MLP) on the RDKit-generated ECFP fingerprint (radius 2, 2048-bit), and a graph neural network (GNN) with one-hot encoded atomic descriptors using PyTorch Geometric. Models were tuned using grid search and selected based on validation ROC AUC. The final evaluation was done on a separate test set.

3.2 Explainability methods

For model explainability, we employed SHAP explanations using the SHAP package for the RF, XGB, and MLP models. Additionally, we used vanilla gradient explanations for MLP and GNN models and Grad-CAM for GNN [8]. From each explanation for fingerprint-based models, we identified the top 5 most important features and then determined which atoms contributed to these bits per

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_2

molecule using atom fragment mappings. As a result, the top 5 bits often corresponded to more than 5 unique atoms. In contrast, for GNN models, we directly selected the top 5 most important nodes, which always corresponded to exactly 5 atoms. These identified atoms were then used for visualization, sparsity score calculations, pharmacophore-type matching, and PharmacoScore computation.

3.3 Dataset construction

This study employed three datasets profiling compound activity (IC_{50} and/or Ki) against CDK2 (1432 active, 699 inactive, 1 µM cutoff), CDK4 (1102 active, 1150 inactive, 1 µM cutoff), and CDK6 (307 active, 372 inactive, 100 nM cutoff). All compound data originated from the ChEMBL database. Data was split into training, validation, and test sets using Bemis-Murcko scaffold-based splitting to ensure structural dissimilarity between sets. This work introduces a new CDK pharmacophore-based benchmark dataset. To vary the dataset's difficulty, we used a five-element pharmacophore for CDK2 and three-element pharmacophores for CDK4 and CDK6. Molecules matching the pharmacophore were labeled '1', others '0', yielding 1057/1074 (pos/neg) for CDK2, 2071/181 for CDK4, and 638/41 for CDK6. We added a second test set to better assess machine learning models' understanding of pharmacophores. The updated dataset includes positive examples reused from the prior test set and newly generated negative examples, where we slightly altered the molecular structures while preserving their overall shapes. Our modifications involved random substitutions of hydrogen donors and acceptors with carbon atoms, removal of aromaticity in the rings, adjustments to the length of the linkers connecting the rings, and replacement of hydrophobic groups with polar ones.

3.4 Pharmacophore model preparation

The pharmacophore models were created with Maestro Schrödinger Release 2021-2, based on the structures of human protein complexes CDK2 (PDB ID: 2A4L), CDK4 (PDB ID: 9CSK) and CDK6 (PDB ID: 5L2I). Atom types, hydrogens, and charges were assigned using the Protein Preparation Wizard. OPLS4 and PROPKA optimized the complex's structure and hydrogen bond network. A starting set of pharmacophores was produced with the Phase module, using the ligand-receptor complex and the CDKs ATP-binding site cavity as a basis. For each instance, pharmacophores were generated using the automated E-Pharmacophore method, with a feature limit of ten. A subset of pharmacophores was created using the top features from the initial set of automatically generated pharmacophores. These features focused on ATP-binding pocket interactions, particularly Hinge Region contacts. The optimal pharmacophore model for discriminating activity from inactive compounds was determined through application of the Hypothesis Validation tool to a curated dataset. Phase considers a molecule aligned only if all pharmacophore points match corresponding features within a 2Å range and generates a single best-fit conformation for each ligand.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_2

3.5 Evaluation

The performance of bioactivity prediction models is conducted using standard classification metrics: ROC AUC, accuracy, and F1-score. The explanations are evaluated using the sparsity and fidelity metrics, defined as follows:

$$\text{Sparsity} = \frac{1}{N}\sum_{i=1}^{N}\frac{|m_i|}{n_i}, \quad \text{Fidelity} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{1}(y_i = f(x_i)) - \mathbb{1}(y_i = f(\bar{x_i}))\right),$$

where N is the number of testing examples, n_i is the number of atoms in the i-th example, $m_i \in \{0,1\}^{n_i}$ is an explanation for the i-th example, $|m_i|$ denotes the number of non-zero elements of m_i, x_i is the input representation of i-th molecule, y_i is its label, $\mathbb{1}$ is the indicator function, and $\bar{x_i}$ is the masked input where important atoms are removed.

We implement two additional metrics to measure the agreement between the ground-truth explanations and model explanations. Feature detection accuracy is defined as the percentage of testing examples in which a particular pharmacophoric feature is detected as important, i.e., at least one atom of this feature overlaps with the model explanation. We also introduce **PharmacoScore** (Ph-Score), an evaluation metric designed to test whether the whole pharmacophore is prioritized over other atoms in a molecule. It is defined as the ROC AUC between the explained atom importances and binary ground-truth atom labels.

4 Results

We trained ML models on our activity dataset. In the following sections, we first explore the quality of predictions. Second, we check if these models can distinguish decoys, which are molecules with small structural changes that break the given pharmacophoric structure. Next, we introduce a new evaluation metric based on pharmacophore alignment to test XAI methods on our dataset. Finally, we compare predictions of models trained on our proxy pharmacophore data with those trained on the original activity data.

4.1 Prediction of compound activity

Four ML models were trained on our dataset that was labeled by finding molecules that match the predefined pharmacophore. RF, XGB, and MLP are three models trained on ECFP fingerprints. GNN is a model trained on the molecular graph representation. All neural networks were trained for 20 epochs using the Adam optimizer. The best set of hyperparameters for each model was found using a random search with 32 randomly sampled hyperparameter sets.

Table 1 shows the performance of ML models in predicting both experimental activity and pharmacophore matching. The results indicate that models trained on the pharmacophore matching task achieve higher ROC AUC scores compared to those trained on the activity prediction task. This suggests that the pharmacophore task may be more straightforward or better defined, potentially due to

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_2

Table 1: Model performance measured on the testing set.

Target	Model	experimental activity			pharmacophoric labels		
		AUC	Accuracy	F1-score	AUC	Accuracy	F1-score
CDK2	RF	0.824	0.625	0.641	0.888	0.778	0.778
	XGB	0.813	0.611	0.648	0.900	0.775	0.783
	MLP	0.800	0.602	0.630	0.865	0.789	0.810
	GNN	0.713	0.637	0.628	0.826	0.724	0.734
CDK4	RF	0.918	0.858	0.886	0.939	0.933	0.961
	XGB	0.912	0.860	0.889	0.944	0.931	0.959
	MLP	0.892	0.824	0.856	0.930	0.931	0.961
	GNN	0.916	0.862	0.895	0.904	0.929	0.965
CDK6	RF	0.807	0.800	0.830	0.964	0.911	0.952
	XGB	0.803	0.785	0.820	0.878	0.933	0.963
	MLP	0.804	0.748	0.785	0.927	0.911	0.952
	GNN	0.769	0.704	0.733	0.919	0.975	0.983

the more homogeneous structural relationships inherent in pharmacophore data. In contrast, activity prediction may involve more complex factors, such as mixed or non-competitive enzyme inhibition, which could introduce greater variability and challenge model performance. Furthermore, methods utilizing fingerprint-based representations consistently outperform GNNs that rely on simple one-hot encoding of atom features. The models demonstrate strong generalizability across all metrics, ROC AUC, accuracy, and F1-score.

4.2 Model performance in decoy detection

Overly optimistic results from evaluating molecular property prediction models on random test sets may arise from the similarity of compounds within a given chemical series. To assess the generalizability of these models, a stratified sampling technique is sometimes employed, grouping compounds with identical chemical scaffolds within the same subset. This approach might also fail if the chemical series includes scaffold modifications.

We propose a more challenging test to learn if the evaluated models discover our pharmacophore hypothesis instead of memorizing molecular fragments in active compounds. Therefore, we create decoy molecules for each test molecule that fit our pharmacophore model. Decoys are used to introduce small structural modifications that should hinder binding to the biological target, removing the critical pharmacophoric features. This new testing set with generated close negatives is used to evaluate our models, and their results are shown in Table 2.

For most types of modifications generated, the accuracy of predictions is generally similar. However, in about 50% of CDK2 cases, models make errors when the distance between aromatic rings is altered. Although all models highly recognized both aromatic rings (Table 3), changing the length of the linker between the rings did not significantly affect the model predictions. As a result, the pre-

Table 2: Model performance tested on the decoy dataset.

	CDK2			CDK4			CDK6		
Model	AUC	Accuracy	F1-score	AUC	Accuracy	F1-score	AUC	Accuracy	F1-score
RF	0.715	0.621	0.482	0.661	0.434	0.586	0.826	0.448	0.618
XGB	0.732	0.631	0.504	0.687	0.484	0.602	0.852	0.630	0.702
MLP	0.635	0.505	0.436	0.568	0.418	0.583	0.580	0.459	0.623
GNN	0.585	0.514	0.416	0.532	0.533	0.612	0.530	0.525	0.650

dictions remained positive despite the absence of pharmacophore matching due to the incorrect distance in three-dimensional space.

The degraded performance on the testing set can often be attributed to the models making predictions for the molecules coming from outside the training distribution. This is not the case for our decoy dataset because all molecules are close analogs of positive examples in the original dataset, which is depicted in Figure 2. The mean Tanimoto distance between decoys and unmodified compounds is less than 0.33. The differences are subtle, often replacing only one atom or changing aromaticity of a ring, which is also depicted in the figure.

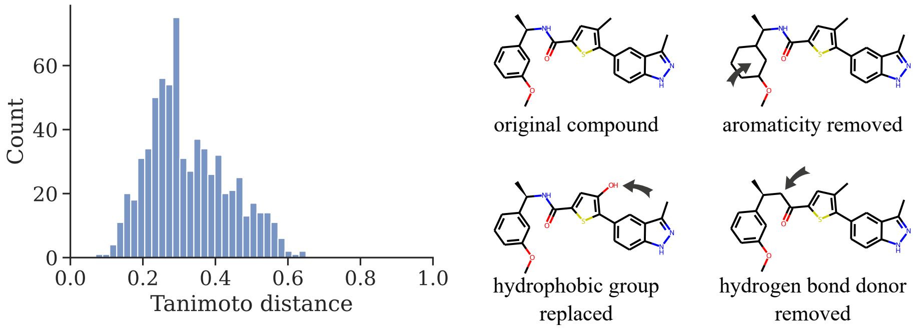

Fig. 2: Tanimoto distance between CDK2 pharmacophore-matching compounds and their decoys, with example decoys shown on the right.

4.3 Assessment of model explanations

By matching and aligning molecules with our pharmacophore hypothesis, we obtain ground-truth labels to evaluate XAI methods. In this experiment, we first test whether our models recover parts of the pharmacophore hypothesis and then employ our PharmacoScore metric to quantify the amount of the pharmacophore that is learned and correctly attributed by the XAI methods.

Table 3 shows the percentages of recovered pharmacophoric features by each explainability method. Additionally, sparsity is reported to account for different numbers of atoms highlighted by each method. A higher sparsity value—meaning more atoms are considered important in the explanation—correlates with a higher percentage of correctly covering all important atoms, while not penalizing models for incorrectly marking non-important atoms as important. The central aromatic group (aromatic 1), located near the HBA and HBD groups, was generally easier to identify across all models. Additionally, a high sparsity value is often linked to the presence of five or six atoms within a pharmacophore point, making it more likely that the model will highlight at least one of them.

Table 3: Ability of the models to detect specific CDK2 pharmacophoric features.

Method	Sparsity	HBA	HBD	Aromatic 1	Aromatic 2	Hydrophobic
RF+SHAP	22%	32%	36%	79%	41%	32%
XGB+SHAP	21%	22%	40%	78%	57%	24%
MLP+SHAP	23%	25%	41%	88%	47%	38%
MLP+VG	23%	25%	41%	88%	47%	38%
GNN+GradCAM	16%	18%	10%	52%	42%	47%
GNN+VG	16%	8%	6%	47%	18%	42%

Most XAI methods do not provide guidance on how to select important atoms. Instead, they assign each atom an importance weight. This allows us to rank all atoms and measure the model's ability to prioritize the pharmacophore correctly using our PharmacoScore. Table 4 presents the agreement between predicted atom attributions and the ground-truth pharmacophore label. Moreover, we report the fidelity of the XAI technique, which measures how crucial the highlighted structure is to the model prediction. Low fidelity scores in MLP and GNN may suggest that poor alignment with the true pharmacophore may result from errors in the explanation method.

PharmacoScore is a challenging metric, with standard models often struggling to accurately mark the pharmacophore atoms, despite being able to classify the entire molecule label. While the top 5 important fragments frequently highlight some of the pharmacophore atoms, standard models tend to mark atoms randomly, assigning high importance to non-relevant side fragments in the global explanations. As a simple baseline, we labeled all aromatic atoms as 1 and others as 0, achieving a score of 0.75 for a 5-point pharmacophore containing two aromatic groups. This high score reflects the prevalence of aromatic fragments, some of which include HBA or HBD atoms that contribute to the CDK2 pharmacophore. The 'all aromatic' baseline can serve as a reference point for future PharmacoScore evaluations. Notably, PharmacoScore performs better for 3-point pharmacophores, where approximately 50% of atoms are labeled as 1.

We visualize some of the explanations in Figure 3. The GNN models, which exhibited the lowest sparsity due to the selection of the top 5 nodes that accounted for an average of 16% of the molecule, were able to recognize hydrophobic fragments at a level close to 40%, demonstrating their effectiveness in identifying these key components. In models trained on fingerprints, the sparsity

DOI: 10.1007/978-3-031-97626-1_2

Table 4: Ability of the models to prioritize pharmacophore over less important molecular substructures. The model named "all aromatic" is a simple baseline where all aromatic atoms are marked as important.

	CDK2		CDK4		CDK6	
Method	Fidelity	Ph-Score	Fidelity	Ph-Score	Fidelity	Ph-Score
all aromatic	-	0.75	-	0.67	-	0.60
RF+SHAP	0.20	0.54	0.18	0.53	0.17	0.69
XGB+SHAP	0.30	0.53	0.53	0.51	0.82	0.62
MLP+SHAP	0.09	0.49	0.00	0.51	0.01	0.56
MLP+VG	0.05	0.51	0.01	0.53	0.01	0.64
GNN+GradCAM	0.14	0.41	0.35	0.49	0.27	0.54
GNN+VG	0.06	0.48	0.33	0.49	0.23	0.51

was higher, caused by the presence of several atoms within a single fingerprint feature. The models frequently predicted the NH group between aromatic rings as interacting with the HBD, which often resulted in correct predictions (e.g., CHEMBL482211). However, in some molecules (CHEMBL115220), this atom was not the actual interacting group. The models also struggled with molecules containing three aromatic rings (CHEMBL232735), failing to identify the key rings. The RF and XGB models highlighted all three aromatic rings, while the MLP model marked only two, ignoring one key aromatic ring. However, in many cases, fingerprint and SHAP produce similar XAI results.

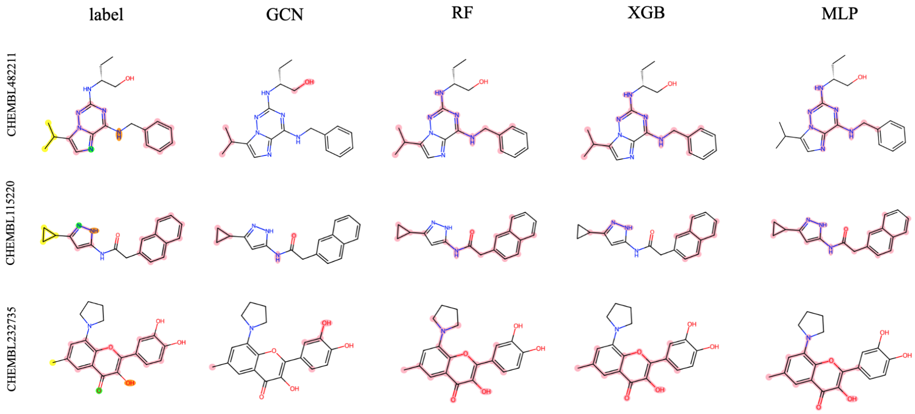

Fig. 3: Example explanations of XAI techniques applied in our proxy pharmacophore prediction problem. Pharmacophore atoms are highlighted: HBA in orange, HBD in green, hydrophobic in yellow, and aromatic in pink. Important atoms identified by Grad-CAM (GCN) and SHAP (RF, XGB, MLP) are shown in pink, highlighting key regions influencing model predictions.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97626-1_2

4.4 CDK2 inhibitors case study

To validate the usefulness of our benchmark in real-world scenarios, we compare the insights derived from the model trained on the experimental data with those from our proxy pharmacophore-based labels. Figure 4 illustrates the differences between explanations of XGB models trained on experimental and pharmacophoric labels. The model trained on experimental labels fails to identify HDA atoms in CHEMBL361833 and the aromatic ring in CHEMBL4297488. However, in some cases, predictions fully align, as observed for CHEMBL497854 and CHEMBL3655766. The overall consistency in highlighting similar structural features can be partly explained by the high precision of the pharmacophore hypothesis (73%) and the notable correlation between the label sets (accuracy 56% and precision 54% for compounds with matching labels).

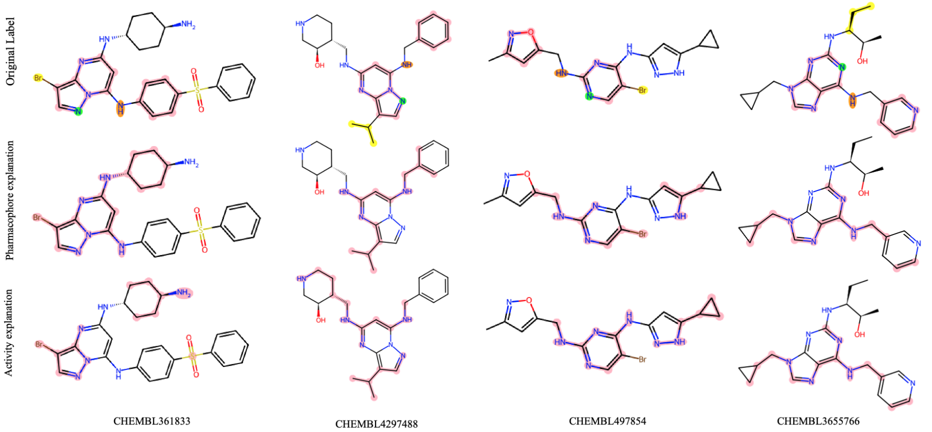

Fig. 4: Comparison of explainability-based pharmacophore classification model and activity classification model against the reference label. The plot illustrates how both models align with the ground truth, highlighting differences in their ability to capture key pharmacophore features.

Interestingly, some misclassification of the activity model can be attributed to finding pharmacophoric features, which is shown in Figure 5. In the HBD modification, new fragments are highlighted as important, but fragments common with the original molecule prediction are still visible, hence the false positive (FP) prediction. In the HBA modification, the key pharmacophore fragment was removed, so the model does not highlight this fragment as important. However, the presence of remaining important fragments still results in a positive prediction despite the lack of pharmacophore matching. In the example of hydrophobic modification, the original prediction was positive, and the hydrophobic point was not crucial for the positive classification. Therefore, after modifying this site, the prediction remained positive. However, correct prediction of the distance between aromatic rings is crucial to solving the pharmacophore matching issue. The modification of ring distance shows that the model does not recognize the correct distance, still highlighting both aromatic rings as important despite the reduction of the distance and the lack of pharmacophore matching.

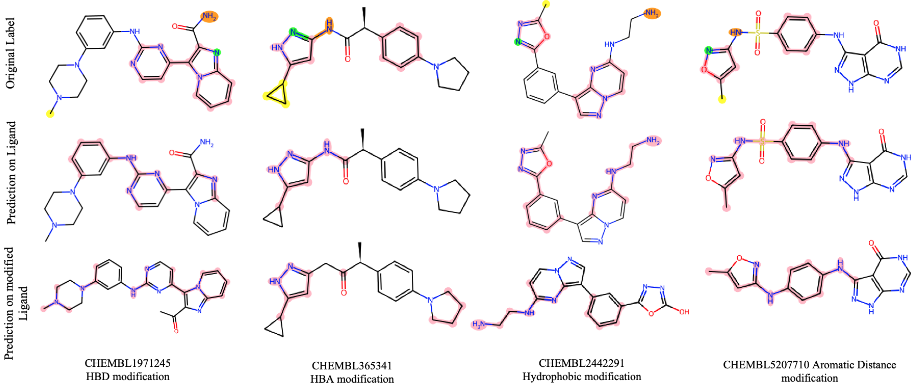

Fig. 5: Visualization of FP cases by the XGB model, showing positive true labels, positive predictions, and modification-based predictions. SHAP explanations highlighting key atoms in the model's predictions.

4.5 Limitations

Although the proposed benchmark spotlights crucial problems with the commonly used QSAR models and XAI techniques, it has limitations that should be considered when drawing conclusions in more general drug discovery setups.

Multiple pharmacophores. The proposed proxy activity labeling considers only one pharmacophore per target. In reality, molecules can have different binding modes that affect their target. Moreover, some molecules can bind allosterically to a different part of the target protein, making the pharmacophore constructed on a set of typical orthosteric ligands uninformative. Nevertheless, the labels produced using only one pharmacophore prove to be challenging for some models, and more complex models should be used only after this benchmark is solved.

3D prediction models. This benchmark does not consider 3D models because they require one particular molecular conformation as input. However, the conformation that matches our pharmacophore might differ from the lowest-energy conformation that can be computed using force-field methods. It is possible that models that use 3D descriptors might be better at understanding high-level pharmacophoric features, which we leave as future work.

5 Conclusions

In this study, we introduced a novel benchmark dataset alongside a new metric, PharmacoScore, to evaluate the explainability of ML models in cheminformatics. Our findings reveal that commonly used classification and regression models, despite their strong performance in traditional tasks, face significant challenges in aligning with the interpretability requirements of pharmacophore-based modeling. Pharmacophores present a notable difficulty for ML models, as even minor structural modifications—such as those introduced by decoys—lead to a significant drop in performance, with ROC AUC scores decreasing about 0.2. This decline highlights the critical need for models that can reliably predict bioactivity for compounds with small modifications, a capability essential for hit-to-lead optimization in drug discovery. Moreover, our results demonstrate that typical explainability methods are not well-suited for bioactivity prediction tasks, as they often fail to accurately identify pharmacophoric interaction sites. This underscores a fundamental gap in the applicability of current ML approaches to cheminformatics, where interpretability is as crucial as predictive accuracy. By providing a benchmark and a robust evaluation framework, we aim to inspire the development of more interpretable and chemically meaningful models, ultimately advancing the field of computational drug discovery.

Acknowledgements This research has been conducted as part of the FIRST TEAM FENG project No FENG.02.02-IP.05-0029/23, funded by the Foundation for Polish Science (FNP) under the European Funds for a Modern Economy (FENG) program, co-financed by the European Union. The publication was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857533. This publication is supported by Sano project carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017108.

References

1. Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J.M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., Herrera, F.: Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. Information fusion **99**, 101805 (2023)
2. Garreau, D., Luxburg, U.: Explaining the explainer: A first theoretical analysis of lime. In: International conference on artificial intelligence and statistics. pp. 1287–1296. PMLR (2020)

3. Guven, D.C., Sahin, T.K.: The association between her2-low status and survival in patients with metastatic breast cancer treated with cyclin-dependent kinases 4 and 6 inhibitors: a systematic review and meta-analysis. Breast Cancer Research and Treatment **204**(3), 443–452 (2024)
4. Jiménez-Luna, J., Grisoni, F., Schneider, G.: Drug discovery with explainable artificial intelligence. Nature Machine Intelligence **2**(10), 573–584 (2020)
5. Li, Z.: Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost. Computers, Environment and Urban Systems **96**, 101845 (2022)
6. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
7. Muratov, E.N., Bajorath, J., Sheridan, R.P., Tetko, I.V., Filimonov, D., Poroikov, V., Oprea, T.I., Baskin, I.I., Varnek, A., Roitberg, A., et al.: Qsar without borders. Chemical Society Reviews **49**(11), 3525–3564 (2020)
8. Pope, P.E., Kolouri, S., Rostami, M., Martin, C.E., Hoffmann, H.: Explainability methods for graph convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10772–10781 (2019)
9. Qian, X., Dai, X., Luo, L., Lin, M., Xu, Y., Zhao, Y., Huang, D., Qiu, H., Liang, L., Liu, H., et al.: An interpretable multitask framework bilat enables accurate prediction of cyclin-dependent protein kinase inhibitors. Journal of Chemical Information and Modeling **63**(11), 3350–3368 (2023)
10. Ribeiro, M.T., Singh, S., Guestrin, C.: " 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. pp. 1135–1144 (2016)
11. Rymarczyk, D., Dobrowolski, D., Danel, T.: Progrest: Prototypical graph regression soft trees for molecular property prediction. In: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM). pp. 379–387. SIAM (2023)
12. Wang, B., Lin, P., Zhong, Y., Tan, X., Shen, Y., Huang, Y., Jin, K., Zhang, Y., Zhan, Y., Shen, D., et al.: Explainable deep learning and virtual evolution identifies antimicrobial peptides with activity against multidrug-resistant human pathogens. Nature Microbiology pp. 1–16 (2025)
13. Wojtuch, A., Danel, T., Podlewska, S., Maziarka, Ł.: Extended study on atomic featurization in graph neural networks for molecular property prediction. Journal of Cheminformatics **15**(1), 81 (2023)
14. Wong, F., Zheng, E.J., Valeri, J.A., Donghia, N.M., Anahtar, M.N., Omori, S., Li, A., Cubillos-Ruiz, A., Krishnan, A., Jin, W., et al.: Discovery of a structural class of antibiotics with explainable deep learning. Nature **626**(7997), 177–185 (2024)
15. Yang, S.Y.: Pharmacophore modeling and applications in drug discovery: challenges and recent advances. Drug discovery today **15**(11-12), 444–450 (2010)
16. Ying, Z., Bourgeois, D., You, J., Zitnik, M., Leskovec, J.: Gnnexplainer: Generating explanations for graph neural networks. Advances in neural information processing systems **32** (2019)
17. Yuan, H., Yu, H., Wang, J., Li, K., Ji, S.: On explainability of graph neural networks via subgraph explorations. In: International conference on machine learning. pp. 12241–12252. PMLR (2021)
18. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)