A Case Study in Explainable AI for Drug-Drug Interaction Prediction: A SHAP-Based Approach

Anonymous

No Institute Given

Abstract. Drug-drug interactions (DDIs) can lead to serious adverse effects and compromised treatment efficacy. As artificial intelligence models gain traction in DDI prediction, the interpretability of these models becomes crucial. Despite advancements in predictive performance, most DDI models lack adequate explainability, limiting their clinical utility. This study aims to enhance the interpretability of a deep learning DDI prediction model by applying a post-hoc Explainable Artificial Intelligence (XAI) method and evaluating the model's reasoning in the context of domain knowledge. The analysis focused on a publicly available DDI model based on autoencoders and a deep neural network. Kernel SHAP was applied to the model using the original input space, which included structural, gene target, and gene ontology similarity matrices. DrugBank served as the reference for validating pharmacological relevance. Explanations were evaluated based on domain knowledge and visualized using SHAP tools. Several features highlighted by Kernel SHAP appeared to align with the underlying mechanisms of DDIs. Structural similarity features appeared frequently among the top contributors to the prediction. Often, high importance was assigned to identity-like structural similarity features, which raises concerns about potential shortcut learning. Gene target and ontology features were less consistent, sometimes reflecting mechanisms of interaction and other times appearing misleading due to limitations in similarity calculations. The study demonstrates that XAI methods, when supported by domain-informed interpretation, can uncover valuable insights into a model's internal reasoning. It also raises important concerns regarding feature design and the necessity for more rigorous validation.

Keywords: Drug Interactions · Artificial Intelligence · Interpretability.

1 Introduction

The rapid expansion of available medications, coupled with an aging population and rising rates of chronic illness, has led to increased polypharmacy and a heightened risk of drug-drug interactions (DDIs) [36]. DDIs occur when two or more drugs are administered simultaneously, resulting in one or more of them having an altered and often undesirable effect [4]. Butkiewicz et al. estimated the incidence of DDIs to be at least two interactions per prescription containing around six medications [4]. Moreover, around 2-4% of all hospital admissions are estimated to be caused by DDIs [36].

Developing new drugs requires rigorous testing for drug-drug and drug-food interactions that need to be verified and reported before a drug enters the market [9]. Traditional DDI testing involves laboratory assessment of pharmacokinetic interactions and, to a lesser extent, pharmacodynamic interactions to evaluate a drug's safety, followed by clinical trials involving human subjects before the drug is brought to the market [9]. However, these methods are slow, expensive, and do not guarantee capturing all the significant interactions [44]. Shortcomings of traditional DDI studies include design or analysis bias, the exclusion of pharmacodynamic interactions, and ethical concerns related to the potential risks posed to study volunteers [39]. Moreover, the incomplete knowledge of many drugs' exact mechanism of action further complicates identifying pharmacodynamic DDIs [30].

In healthcare, drug information resources are essential tools for identifying potential DDIs throughout the patient care journey. These resources typically rely on pharmacokinetics and pharmacodynamics data provided by drug manufacturers, limited clinical trial results, and post-marketing pharmacovigilance reports to make recommendations about potential interactions between two drugs [16]. However, these resources typically do not account for the combined effects of more than two drugs and may sometimes base interaction conclusions on extrapolation without direct verification [16]. As a result, our understanding of DDIs remains fragmented and incomplete.

The increasing availability of drug-related data and rapid advancements in Artificial Intelligence (AI) have shifted efforts toward building high-performance, complex computational models to support drug discovery and improve DDI prediction accuracy [40]. These models can integrate various drug properties, including chemical structure, therapeutic activity, and genomic characteristics [30]. This ability makes them particularly valuable for supporting pharmacodynamic DDI studies, which often require extensive drug pharmacology and biological systems knowledge [30]. Specifically, Deep Learning (DL) has been increasingly adopted due to its superior ability to capture the complexity, non-linearity, and ambiguity inherent in real-world pharmacological interactions, compared to traditional Machine Learning (ML) methods [40]. However, this increase in model complexity has come at the cost of interpretability, creating a major obstacle to clinical adoption where transparency and trust are critical [40]. As a result, enhancing model interpretability has become a key focus aiming to gain the confidence of healthcare stakeholders [40]. These efforts aim to bridge the gap between model development and real-world clinical implementation. Despite this, the use of Explainable Artificial Intelligence (XAI) techniques in the DDI prediction domain remains limited. Most studies rely on attention mechanisms, which offer only partial interpretability and remain debatable [21,34].

Hence, in a review by Vo et al. [40], the limited emphasis on interpretability in DDI prediction models was highlighted as a significant research gap. Moreover, given that most current DDI prediction models often rely on DL to capture the multi-level nature of drug interactions [2], they present an even stronger inherent lack of interpretability, which sets a noteworthy barrier to their adoption in clinical practice, where transparency and trust are essential [2,3]. Consequently, this study aims to enhance the interpretability of a DL-based DDI prediction model by applying a post-hoc XAI method and evaluating the model's reasoning in the context of domain knowledge.

Contributions

This paper presents an approach to explainable DDI prediction, with the following key contributions:

1. The integration of a high-performing DL model for DDI prediction with a robust post-hoc SHAP-based explainer, enabling feature-level interpretability of interaction classifications.
2. The incorporation of domain knowledge into the interpretation of SHAP outputs, providing context-aware and pharmacologically meaningful explanations to support decision-making in the DDI domain.

2 Background

Several approaches have been developed for predicting DDIs, each differing in their underlying algorithms and prediction tasks [18]. Traditional ML methods typically rely on drug features to predict binary outcomes—i.e., whether an interaction exists between a given drug pair [18]. In contrast, DL techniques offer greater flexibility and are applied to a variety of input formats, including structured feature representations, similarity profile vectors, and graph-based models [18].

Similarity-based and graph-based approaches to DL for DDI prediction are becoming increasingly prevalent. Similarity-based approaches leverage structural and functional resemblances between drugs derived from various drug-related properties such as chemical structure, target gene similarity, Gene Ontology (GO) term similarity, and other biological or pharmacological attributes [18]. These similarities are encoded as feature vectors, commonly called similarity profiles, which serve as input to DL models. A structural similarity (SS) profile captures how similar a drug's molecular structure is to other drugs [18]. A target gene similarity (TS) profile quantifies the similarity between the target genes of different drugs, often estimated by their proximity within a functional interaction network of genes, proteins, and their interrelationships [24]. GO term similarity (GS) profiles describe functional similarities based on GO annotations, which encompass biological processes, molecular functions, and cellular components associated with drug target genes [37]. These profiles provide insights beyond direct gene interactions, offering a broader functional context for interpreting DDIs [38]. Notable models such as DeepDDI [33] and the model proposed by Lee et al. [24] exemplify this approach, using structured similarity-based feature vectors as input for DDI prediction.

Graph-based approaches are increasingly utilized to model and predict complex interactions within drug networks [28]. In these methods, data are structured as graphs, where nodes represent entities such as drug molecules, proteins, or other biologically relevant components, and edges denote the relationships or interactions between them [18,28]. DL approaches specifically designed for graph-structured data are known as Graph Neural Networks (GNNs) [46]. A widely used variant of GNNs in DDI prediction is the Graph Attention Network (GAT) [19,20,23]. It incorporates attention mechanisms to assign learnable weights to different nodes and edges based on their relevance to the prediction task [11]. As such, attention mechanisms provide a lens into the model's decision-making process and enhance the interpretability of graph-based DDI models [11]. However, the effectiveness of attention mechanisms as tools for XAI remains an area of active investigation, with ongoing debate regarding their reliability and interpretive consistency [21,34].

Although attention-based methods offer a degree of interpretability by highlighting relevant inputs, they are often sensitive to noise [34] and prone to human subjective interpretation [11] . Furthermore, empirical studies have shown that attention maps frequently misalign with gradient-based feature importance measures, casting doubt on their faithfulness as explanations [1]. Additionally, attention mechanisms often require substantial computational resources during training and can increase the risk of overfitting, which may compromise the generalizability of the model [11]. Consequently, relying solely on attention mechanisms for interpretability is controversial, and it is increasingly recommended that they be used in conjunction with complementary XAI techniques [15].

Despite these limitations, many DDI prediction studies continue to emphasize interpretability through attention [12,19,31] and other model-specific architectures, highlighting the need for broader experimentation with more robust and theoretically grounded XAI approaches. Some recent models, such as MI-DDI [8], STNN-DDI [45], and KnowDDI [41] aim to enhance interpretability through intrinsic model designs, often leveraging substructure-based representations, tensor factorization, or knowledge graph traversal strategies. However, like attention-based explanations, these approaches are typically model-specific and may lack generalizability. In contrast, model-agnostic XAI techniques, such as SHAP (Shapley Additive Explanations), are gaining momentum due to their formal, post-hoc interpretability framework, which is applicable across diverse model architectures. These methods enable researchers to systematically assess and validate the explanations produced by model-specific approaches, thereby supporting a more reliable and comprehensive understanding of DDI predictions.

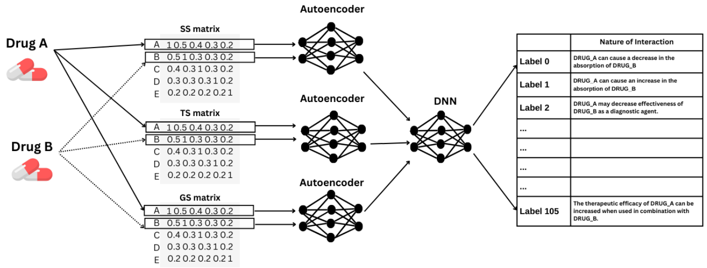
The SHAP method [27], based on Shapley values [35], is regarded as intuitive for humans to interpret [25]. SHAP generates explanations by calculating values that represent the contribution of each input feature to the model's final prediction [1,25]. Calculating Shapley values is typically computationally expensive, especially for DL models [32]. Therefore, the values are often approximated using efficient methods [32] like the Kernel SHAP Explainer [27], which approximates Shapley values by fitting a weighted linear regression model over a sampled subset of all possible coalitions [27]. The resulting regression coefficients approximate the Shapley values of each feature for the given instance [27].

Compared to the Local Interpretable Model-Agnostic Explanations (LIME), Kernel SHAP offers local interpretability through Shapley value approximation for individual instances, and global insights by averaging these values across multiple instances or the entire dataset [1]. The SHAP package also includes advanced visualization tools that make the results more accessible to a broad audience, including healthcare stakeholders [32]. Kernel SHAP retains important properties such as local accuracy and consistency, derived from its foundation in Shapley values [1]. Additionally, Doumard et al. [10], in a quantitative comparison of several local interpretability methods, found that SHAP tends to assign higher importance to a few key features compared to LIME. This characteristic is particularly valuable in high-dimensional settings, where focusing on the most influential features can aid interpretation. The study also noted that LIME demonstrated less robustness when applied to more complex models [10].

In this study, we address these gaps of interpretability by selecting a high-performing DL model capable of reflecting real-world data complexity of DDIs and applying the model-agnostic Kernel SHAP to interpret predictions in terms of domain-specific features. This approach enables post-hoc explanation while preserving access to original input features. While multi-label models offer a more accurate reflection of real-world DDIs, by capturing interaction directionality and the possibility of multiple concurrent interaction types, binary classification remains more commonly adopted due to its relative simplicity [19,23,31]. Similarly, although incorporating biological features such as enzyme activity and target protein information can enhance the mechanistic understanding of DDIs, molecular structure-based inputs are still more widely used in practice [8,31,41]. This tendency toward simplified modeling strategies highlights the need for greater interpretability in complex DDI models. Enhancing the interpretability of more sophisticated models, those that capture drug-target interactions as well as the nature and directionality of interactions, could foster greater trust and encourage their adoption in DDI research.

Regarding the DDI predictor models, the one developed by Lee et al. [24] is recognized as one of the best performers (AUC = 99.61, AUPRC = 95.62) in the multi-class prediction experiment done by Luo et al. in their review [28], in which it was compared to existing DL and graph-based DDI prediction models on benchmark datasets. This model could also predict multiple labels, making it suitable for multi-label classification tasks as well [24]. Additionally, it integrates information from multiple levels of the biological system's hierarchy, including pairwise structural, target gene, and gene ontology similarities [24]. Moreover, many of the labels are directional, allowing for a distinction between the affected and the affecting drug.

The model's architecture, as shown in Figure 1, consists of autoencoders that learn lower-dimensional representations from the SS, TS, and GS profiles of the drug pair under study, which are extracted from the corresponding similarity matrices [24]. These representations are then passed to a feed-forward neural



	Nature of Interaction
Label 0	DRUG_A can cause a decrease in the absorption of DRUG_B
Label 1	DRUG_A can cause an increase in the absorption of DRUG_B
Label 2	DRUG_A may decrease effectiveness of DRUG_B as a diagnostic agent.
...	
...	
...	
...	
Label 105	The therapeutic efficacy of DRUG_A can be increased when used in combination with DRUG_B.

Fig. 1: Overview of the architecture of the model by Lee et al.

network for a multi-label classification task. The model is trained to predict among 106 classes, representing various kinds of drug interactions documented in the literature.

Each similarity profile contains similarity scores of a given drug against all 1,597 drugs in the dataset. SS scores are computed using the Tanimoto Coefficient, while TS scores are determined using the following equation [24]:

$$TS_{AB} = \frac{|\{(x, y) \in G_A \times G_B \mid d(x, y) \leq t_A\}|}{|\{(x, y) \in G_A \times G_B\}|}$$

$$t_A = \max\{d(x, y) \mid x, y \in G_A\}$$

Where G_A are the target genes for drug A, G_B are the target genes for drug B, $d(x, y)$ is the distance between genes x and y in the functional interaction network, and t_A is the maximum distance between any two target genes in G_A. In other words, the target gene similarity TS is calculated as the ratio of the number of gene pairs between G_A and G_B whose distance is less than t_A, over the total number of possible gene pairs between drugs A and B. Similarly, the GS scores are calculated based on the distance between the GO terms in the GO graph [24]

3 Preliminaries

The following notation is developed based on the work by Lee et al. [24], and presented to visualize the inner workings of the multi-label classifier implemented. Let a drug pair (d_i, d_j) be an input instance to the model. For each drug pair, three similarity profiles are extracted. This is done by selecting the corresponding rows for each drug from the respective similarity matrices and concatenating them into a single feature vector per profile. The resulting vectors $\mathbf{x}_{SS}$, $\mathbf{x}_{TS}$, and $\mathbf{x}_{GS}$ form the input representation for the model:

- $\mathbf{x}_{SS} \in \{0, 1\}^{2n}$: structural similarity profiles of drugs d_i and d_j, derived from the structural similarity matrix SS $\in \mathbb{R}^{n \times n}$,

- $\mathbf{x}_{TS} \in \{0,1\}^{2m}$: target gene similarity profiles of drugs d_i and d_j, derived from the target similarity matrix TS $\in \mathbb{R}^{m\times m}$,
- $\mathbf{x}_{GS} \in \{0,1\}^{2p}$: GO similarity profiles of drugs d_i and d_j, derived from the GO similarity matrix GS $\in \mathbb{R}^{p\times p}$.

Where n represents the number of drugs in the SS matrix, m represents the number of drugs in the TS matrix, and p represents the number of drugs in the GS matrix.

Each of these input vectors is independently passed through a dedicated autoencoder to obtain a lower-dimensional representation:

$$\mathbf{z}_{SS} = AE_{SS}(\mathbf{x}_{SS}) \in \mathbb{R}^{n'}, \quad \mathbf{z}_{TS} = AE_{TS}(\mathbf{x}_{TS}) \in \mathbb{R}^{m'}, \quad \mathbf{z}_{GS} = AE_{GS}(\mathbf{x}_{GS}) \in \mathbb{R}^{p'}$$

The resulting compressed vectors are then concatenated into a unified feature vector:

$$\mathbf{z}_{ij} = [\mathbf{z}_{SS} \,\|\, \mathbf{z}_{TS} \,\|\, \mathbf{z}_{GS}] \in \mathbb{R}^{d'}, \quad \text{where } d' = n' + m' + p'$$

This vector is input to a feed-forward neural network that performs a multi-label classification, predicting an interaction label vector:

$$\hat{\mathbf{y}}_{ij} = f_\theta(\mathbf{z}_{ij}), \quad \hat{\mathbf{y}}_{ij} \in \{0,1\}^L$$

where L is the number of possible interaction types, and θ denotes the model parameters. The goal is to approximate the true interaction labels $\mathbf{y}_{ij} \in \{0,1\}^L$.

To interpret the model predictions, we apply Kernel SHAP, a model-agnostic XAI technique. SHAP assigns an attribution value ϕ_k to each input feature x_k, reflecting its contribution to the prediction. SHAP uses a linear surrogate explanation model:

$$g(\mathbf{z}') = \phi_0 + \sum_{k=1}^{d} \phi_k z'_k$$

where $\mathbf{z}' \in \{0,1\}^d$ is a binary coalition vector indicating feature presence, and ϕ_k is the Shapley value for feature k. This provides a locally interpretable approximation of the model's behavior for a specific instance.

4 Methods

To achieve our objective, we first train the model, evaluate class distribution, and select a set of balanced interaction labels for analysis. We then choose representative background and test instances needed for Kernel SHAP. Finally, we perform a SHAP analysis and interpret the results from a domain-specific perspective.

The source code for the Lee et al. model [24] is obtained from its publicly available GitHub repository, which includes similarity matrices and a dataset mapping drug pairs to their corresponding interaction types (labels). The dataset consists of 188,258 drug pairs, while the similarity matrices comprise similarity scores of 1597 drugs. The model is trained and evaluated using the source code provided in the original repository. Training is performed using five-fold cross-validation, repeated according to the defined hyperparameters. For the purpose of SHAP analysis, the model trained on the first fold is selected as the representative model. Performance evaluation is conducted on this model using standard metrics, including accuracy, precision, and recall, to ensure its predictive quality prior to explanation.

Next, the label distribution within the dataset is examined using the Imbalance Ratio (IR), which measures the ratio between the number of instances in the most frequent class and that in the target class [6]. While there is no universally accepted threshold for IR, values below 10:1 are often considered acceptable in the literature, indicating that additional preprocessing may not be necessary [13,17]. To prevent the overestimation of Shapley values for underrepresented classes [7,26], only labels with an IR less than 10:1 will be selected for the analysis. Table 1) shows the six most frequent labels representing different types of DDI, which are prioritized for detailed analysis, as they are well-balanced and able to capture a significant portion of the underlying patterns in the data.

Table 1: Most frequent interaction types and their imbalance ratios (IR).

Label Number	Label Description	Percentage of Total (%)	IR
73	The risk or severity of adverse effects can be increased when Drug A is combined with Drug B	30.9%	1:1
68	The metabolism of Drug A can be decreased when combined with Drug B	14.0%	2:1
100	The serum concentration of Drug A can be increased when combined with Drug B	10.6%	2:1
43	Drug A may increase the hypotensive activities of Drug B	5.9%	5:1
104	The therapeutic efficacy of Drug A can be decreased when combined with Drug B	4.4%	7:1
99	The serum concentration of Drug A can be decreased when combined with Drug B	4.4%	7:1

Subsequently, background data is selected from the training folds, while the test data is drawn from the corresponding test fold. A stratified sampling strategy with equal allocation across balanced labels is applied to construct the background and test data sets. Specifically, eight instances are randomly sampled per label, resulting in 48 samples for both sets. This approach mitigates interclass imbalance within a subset of relatively balanced labels and supports meaningful cross-class comparisons of feature importance.

The SHAP analysis is performed using Kernel SHAP from the SHAP package, with minor adjustments to accommodate the model's input format. Since the model accepts three separate input matrices, SS, TS, and GS profiles, these are concatenated into a single feature vector per instance to conform to the expected input format of the Kernel Explainer. To ensure numerical stability, especially when dealing with extreme values, SHAP values are computed using the model's raw logits instead of probabilities obtained through sigmoid activation. Additionally, the L1 regularization (`l1_reg`) parameter is set to "aic" (Akaike Information Criterion), which regularizes the SHAP value estimation by penalizing model complexity while optimizing fit, an effective strategy for high-dimensional data.

SHAP values are then computed for all samples in the test set. Since the labels represent directed interactions between drug pairs, it is important to distinguish whether each feature originated from the similarity profile of drug A or drug B. Furthermore, since the similarity profiles are constructed from three distinct matrices (SS, TS, GS), each feature should be annotated with its corresponding matrix of origin. For example, a feature labeled "Drug A similarity to Mitemcinal – TS" indicates that the value represents the similarity of the first drug in the pair to mitemcinal, derived from the TS matrix. This approach allows for a more interpretable analysis of how specific components of each drug's profile contribute to the model's predictions.

For visualization and domain-specific interpretation, various tools from the SHAP package are employed. These visualizations, including beeswarm and waterfall plots, are used to extract both global and local explanations of the model's behavior. Example visualizations are made available through a dedicated GitHub repository[1], and the waterfall plots for all 48 drug pairs in the test dataset are shown in the additional materials. Where appropriate, the most influential features identified by SHAP are contextualized by mapping them to established pharmacological interactions, shared biological pathways, or known adverse effect profiles. This cross-referencing enhances the biological plausibility of the model's predictions. DrugBank version 5.1.13 [22] serves as the primary reference database for inferring drug properties and pathway associations, ensuring consistency and credibility in domain-specific interpretations.

5 Evaluation

Performance metrics are computed as averages across all cross-validation folds. The model achieves an average accuracy of 0.962, with macro precision of 0.941, macro recall of 0.948, micro precision of 0.964, and micro recall of 0.967. The model trained on the first fold and used for the SHAP analysis achieves comparable performance, with an accuracy of 0.963, macro precision of 0.949, macro recall of 0.947, micro precision of 0.965, and micro recall of 0.970. The training and test data from this fold are retrieved, and the background and test datasets are selected as previously described. SHAP plots are subsequently generated and analyzed to interpret the model's predictions.

Figure 2 presents stacked SHAP values for each instance based on its predicted label. This visualization illustrates the magnitude and direction of the

[1] `GitHub link removed to not compromise anonimity during the review process.`

top contributing features, identified by their highest mean absolute SHAP values, influencing the model's output. Each horizontal line corresponds to a specific feature, showing the distribution of its SHAP values across all instances. The position of each point along the x-axis indicates the direction and intensity of the feature's contribution, with positive values increasing the predicted probability of the corresponding label. Furthermore, the color gradient represents the original feature values, enabling the interpretation of how feature magnitude correlates with its impact on the prediction. Noticeably, most of these features represent structural similarities to drugs included in the test set. This visualization highlights the features with the greatest average influence and how consistently they impact different predictions. The presence of outliers in the SHAP distributions for specific features suggest that some features exert a disproportionately large effect on particular instances.

Example waterfall plots in Figure 3 illustrate the direction and magnitude of the top 10 features' contributions to the raw model output, providing instance-specific interpretability. The average model prediction is denoted as $E[f(x)]$, and the specific prediction for an instance is $f(x)$. Positive contributions are shown as red arrows pointing toward the prediction, with arrow lengths indicating the strength of each feature's impact. Feature values correspond to similarity scores ranging from 0 to 1, where 1 indicates maximum similarity. The feature naming follows the same convention as previous plots, identifying the source (drug A or B) and the matrix type.

Figure 3a presents a local explanation for the prediction involving the drug pair miglitol (Drug A) and iloperidone (Drug B), which is classified under label 104, indicating a potential decrease in the therapeutic efficacy of miglitol when co-administered with iloperidone. The explanation highlights that structural similarities to the same drugs involved in the instance are the top-ranked (first and third) contributing features, suggesting the model places significant weight on intra-instance structural resemblance. Notably, miglustat and voglibose, both structural analogs of miglitol, emerge as highly influential based on their similarity scores. Additional key contributors include target gene similarities between iloperidone and several antihistamines, including hydroxyzine, antazoline, and mequitazine, each showing a maximum similarity score of 1 to Drug B. Interestingly, histamine's role in glucose regulation is an emerging area of research [42], implying that the model may be uncovering biologically meaningful patterns not yet fully characterized in clinical literature.

Figure 3b presents the local explanation for the prediction of label 100 for the drug pair everolimus and carvedilol. The model assigns high SHAP values to features representing SS to everolimus and several structurally related compounds, including temsirolimus, sirolimus, and ridaforolimus, suggesting a firm reliance on class-based SS. One feature of interest is the TS between everolimus and alcaftadine, which receives a maximum similarity score of 1. However, this relationship lacks biological plausibility, as the two drugs target distinct gene pathways. Likewise, features representing low similarity (i.e., dissimilarity) were generally difficult to interpret and provided limited pharmacological insight.

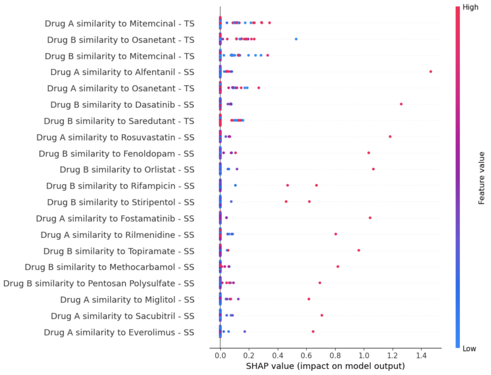
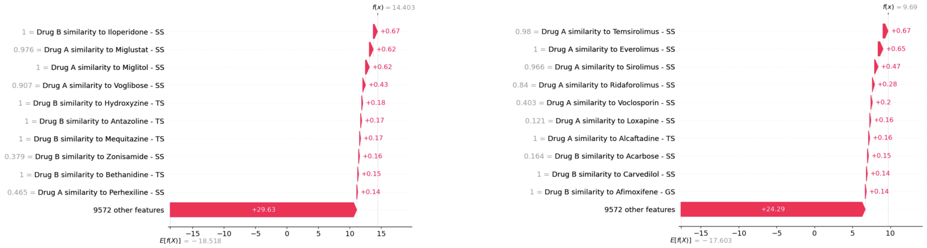


Fig. 2: SHAP Explanation of top features driving predicted labels (Beeswarm view). SHAP values are computed per sample with respect to the model's predicted label.



(a) miglitol–iloperidone

(b) everolimus–carvedilol

Fig. 3: Waterfall plots for instances miglitol vs. iloperidone (left), predicted label 104 (the therapeutic efficacy of Drug A can be decreased when combined with Drug B), and everolimus vs carvedilol (right) predicted label 100 (the serum concentration of Drug A can be increased when combined with Drug B).

6 Discussion

Overall, the SHAP visualizations reveal that features reflecting structural identity or near-identity of either drug in the test instance frequently exert the strongest influence on the model's predictions. In contrast, TS and GS features are identified as important contributors in some cases, but their interpretability is often less direct or consistent. These features are derived from the proximity of drug target genes within functional interaction networks or GO-based semantic spaces. However, they do not encode information about the directionality or pharmacological nature of a drug's action on a gene, limiting their utility for mechanistic interpretation. For example, two drugs may interact with the same CYP450 enzyme but exert opposite effects, one acting as an inducer and the other as an inhibitor, while still receiving a high similarity score due to their network-based proximity. Furthermore, a single drug can have divergent effects on different CYP450 isoforms. For instance, teriflunomide inhibits CYP2C8 but induces CYP1A2 [5]. Such opposing effects, which could lead to markedly different interaction outcomes depending on the co-administered drug [5], are not captured by current similarity metrics.

Including identity-related features in the training process is of particular concern, as it may provide the model with a shortcut, enabling it to encode information about a drug's identity or position within the feature matrix. Geirhos et al. [14] observed that DNNs are prone to "shortcut learning", where models exploit spurious but predictive correlations instead of capturing the underlying causal mechanisms. This behavior, influenced by "inductive biases" inherent in the model architecture or training data [14], may lead the model to prioritize easily learnable identity patterns over the more complex biological interactions that underlie DDIs. Although drug pairs are separated between the training and test sets, there is no guarantee that individual drugs, or their structurally similar analogs, are entirely excluded from the training data. Consequently, features with high SS scores can act as implicit identifiers, enabling the model to recognize previously encountered drugs during inference. This behavior compromises the model's capacity to generalize to unseen drug combinations, a concern also noted in [29].

Future research can build upon this study by applying XAI techniques to this model, architecturally similar models, or other DDI prediction frameworks. XAI methods better suited for interpreting large feature spaces could be explored to gain a more comprehensive understanding of such complex, high-dimensional models. Additionally, in-depth ablation studies may offer valuable insights into model behavior. For example, masking identity-like features (e.g., Tanimoto similarities > 0.85) or shuffling similarity matrices can help assess the model's reliance on individual features or positional bias. Investigating feature importance across a broader range of labels, particularly after addressing class imbalance, may reveal deeper patterns in the model's decision-making. Finally, refining the computation of target gene and GO term similarities to incorporate the directionality of drug effects, such as induction or inhibition of genes or enzymes, could provide more biologically informative features, thereby improving the model's ability to learn the underlying mechanisms of DDIs.

This study faced several limitations that could affect the confidence in its findings. While Kernel SHAP demonstrated stability and consistency, it has known limitations. It assumes local linearity and feature independence, which can result in inflated SHAP values for correlated features and underestimation for others [10]. Local linearity may also fall short of capturing the true behavior of the model. Additionally, Kernel SHAP is computationally intensive, requiring significant time and resources. The limited dataset size may have affected the accuracy of the Akaike Information Criterion [43], especially since the corrected formula for small sample sizes is not yet available in the SHAP package. Moreover, the visualizations presented only the most important features, a small subset of the total feature set. However, the strong presence of identity-like features also supported the decision to limit the number of displayed features while still drawing meaningful conclusions.

7 Conclusion

This case study aims to enhance the interpretability of a high-dimensional DDI prediction model by applying Kernel SHAP, complemented by domain-informed analysis. It offers valuable insights into the model's decision-making behavior by visualizing the most influential features and linking them to known pharmacological mechanisms. Despite the limited sample size, the analysis revealed consistent patterns, particularly the model's frequent attribution of high importance to identity features or those reflecting high structural similarity between drugs. This tendency suggests a reliance on shortcut learning, where identity-like features are used to guide or reinforce predictions rather than uncover the underlying bio-logical mechanisms. Although the model exhibited some ability to detect biologically relevant relationships, the interpretability of target gene and Gene Ontology-based similarity features remains limited. This is likely due to how these similarity measures are computed, as they may fail to capture the directional, context-dependent nature of drug-target interactions.

Acknowledgments. Removed to not compromise anonimity during the review process.

Disclosure of Interests. No conflicts of interest or external funding were associated with the research.

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 (Nov 2023)

2. Alizadehsani, R., Oyelere, S.S., Hussain, S., Jagatheesaperumal, S.K., Calixto, R.R., Rahouti, M., Roshanzamir, M., De Albuquerque, V.H.C.: Explainable Artificial Intelligence for Drug Discovery and Development: A Comprehensive Survey. IEEE Access **12**, 35796–35812 (2024)
3. Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V.I.: Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making **20**(1), 1–9 (Dec 2020)
4. Butkiewicz, M., Restrepo, N.A., Haines, J.L., Crawford, D.C.: DRUG-DRUG INTERACTION PROFILES OF MEDICATION REGIMENS EXTRACTED FROM A DE-IDENTIFIED ELECTRONIC MEDICAL RECORDS SYSTEM. AMIA Summits on Translational Science Proceedings **2016**, 33–40 (Jul 2016)
5. Cada, D.J., Demaris, K., Levien, T.L., Baker, D.E.: Teriflunomide. Hospital Pharmacy **48**(3), 231–240 (Mar 2013)
6. Charte, F., Rivera, A.J., del Jesus, M.J., Herrera, F.: Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing **163**, 3–16 (Sep 2015)
7. Chen, Y., Calabrese, R., Martin-Barragan, B.: Interpretable machine learning for imbalanced credit scoring datasets. European Journal of Operational Research **312**(1), 357–372 (Jan 2024)
8. Cheng, Z., Wang, Z., Tang, X., Hu, X., Yang, F., Yan, X.: A Multi-View Feature-Based Interpretable Deep Learning Framework for Drug-Drug Interaction Prediction. Interdisciplinary Sciences: Computational Life Sciences (Feb 2025)
9. Cole, S., Kerwash, E., Andersson, A.: A summary of the current drug interaction guidance from the European Medicines Agency and considerations of future updates. Drug Metabolism and Pharmacokinetics **35**(1), 2–11 (Feb 2020)
10. Doumard, E., Aligon, J., Escriva, E., Excoffier, J.B., Monsarrat, P., Soulé-Dupuy, C.: A quantitative approach for the comparison of additive local explanation methods. Information Systems **114**, 102162 (Mar 2023)
11. El Houda Dehimi, N., Tolba, Z.: Attention Mechanisms in Deep Learning : Towards Explainable Artificial Intelligence. In: 2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS). pp. 1–7 (Apr 2024)
12. Gao, S., Xie, J., Zhao, Y.: A Multi-Source drug combination and Omnidirectional feature fusion approach for predicting Drug-Drug interaction events. Journal of Biomedical Informatics **162**, 104772 (Feb 2025)
13. García-Pedrajas, N.: Partial random under/oversampling for multilabel problems. Knowledge-Based Systems **302**, 112355 (Oct 2024)
14. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence **2**(11), 665–673 (Nov 2020)
15. Gonçalves, T., Rio-Torto, I., Teixeira, L.F., Cardoso, J.S.: A Survey on Attention Mechanisms for Medical Applications: are we Moving Toward Better Algorithms? IEEE Access **10**, 98909–98935 (2022)
16. Grannell, L.: Drug interaction resources: mind the gaps. Australian Prescriber **43**(1), 18–23 (Feb 2020)
17. Hamid, M.H.A., Yusoff, M., Mohamed, A.: Survey on Highly Imbalanced Multi-class Data. International Journal of Advanced Computer Science and Applications (IJACSA) **13**(6) (2022)
18. Han, K., Cao, P., Wang, Y., Xie, F., Ma, J., Yu, M., Wang, J., Xu, Y., Zhang, Y., Wan, J.: A Review of Approaches for Predicting Drug–Drug Interactions Based on Machine Learning. Frontiers in Pharmacology **12**, 814858 (Jan 2022)

19. He, J., Sun, Y., Ling, J.: A Molecular Fragment Representation Learning Framework for Drug–Drug Interaction Prediction. Interdisciplinary Sciences: Computational Life Sciences (Oct 2024)
20. Hong, Y., Luo, P., Jin, S., Liu, X.: LaGAT: link-aware graph attention network for drug-drug interaction prediction. Bioinformatics (Oxford, England) **38**(24), 5406–5412 (Dec 2022)
21. Jain, S., Wallace, B.C.: Attention is not Explanation. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 3543–3556. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019)
22. Knox, C., Wilson, M., Klinger, C., Franklin, M., Oler, E., Wilson, A., Pon, A., Cox, J., Chin, N.E., Strawbridge, S., Garcia-Patino, M., Kruger, R., Sivakumaran, A., Sanford, S., Doshi, R., Khetarpal, N., Fatokun, O., Doucet, D., Zubkowski, A., Rayat, D., Jackson, H., Harford, K., Anjum, A., Zakir, M., Wang, F., Tian, S., Lee, B., Liigand, J., Peters, H., Wang, R.Q., Nguyen, T., So, D., Sharp, M., da Silva, R., Gabriel, C., Scantlebury, J., Jasinski, M., Ackerman, D., Jewison, T., Sajed, T., Gautam, V., Wishart, D.: DrugBank 6.0: the DrugBank Knowledgebase for 2024. Nucleic Acids Research **52**(D1), D1265–D1275 (Nov 2023)
23. Kundi, I., Sheikh, S., Malik, F., Bhatti, K.: DDI-KGAT: A Graph Attention Network on Biomedical Knowledge Graph for the Prediction of Drug-Drug Interactions. IEEE ACCESS **12**, 162028–162039 (2024)
24. Lee, G., Park, C., Ahn, J.: Novel deep learning model for more accurate prediction of drug-drug interaction effects. BMC Bioinformatics **20**(1), 415 (Aug 2019)
25. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy **23**(1), 18 (Dec 2020)
26. Liu, M., Ning, Y., Yuan, H., Ong, M.E.H., Liu, N.: Balanced background and explanation data are needed in explaining deep learning models with SHAP: An empirical study on clinical decision making (Jun 2022), arXiv:2206.04050 [cs]
27. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc. (2017)
28. Luo, H., Yin, W., Wang, J., Zhang, G., Liang, W., Luo, J., Yan, C.: Drug-drug interactions prediction based on deep learning and knowledge graph: A review. iScience **27**(3), 109148 (Mar 2024)
29. Manica, M., Oskooei, A., Born, J., Subramanian, V., Sáez-Rodríguez, J., Rodríguez Martínez, M.: Toward Explainable Anticancer Compound Sensitivity Prediction via Multimodal Attention-Based Convolutional Encoders. Molecular Pharmaceutics **16**(12), 4797–4806 (Dec 2019)
30. Niu, J., Straubinger, R.M., Mager, D.E.: Pharmacodynamic Drug-Drug Interactions. Clinical pharmacology and therapeutics **105**(6), 1395–1406 (Jun 2019)
31. Niu D, Zhang L, Zhang B, Zhang Q, Li Z: DAS-DDI: A dual-view framework with drug association and drug structure for drug-drug interaction prediction. J Biomed Inform **156**, 104672 (2024)
32. Ponce-Bobadilla, A.V., Schmitt, V., Maier, C.S., Mensing, S., Stodtmann, S.: Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. Clinical and Translational Science **17**(11), e70056 (2024)

33. Ryu, J.Y., Kim, H.U., Lee, S.Y.: Deep learning improves prediction of drug–drug and drug–food interactions. Proceedings of the National Academy of Sciences of the United States of America **115**(18), E4304–E4311 (May 2018)
34. Serrano, S., Smith, N.A.: Is Attention Interpretable? In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2931–2951. Association for Computational Linguistics, Florence, Italy (Jul 2019)
35. Shapley, L.S.: 17. A Value for n-Person Games, pp. 307–318. Princeton University Press, Princeton (1953)
36. Sánchez-Valle, J., Correia, R.B., Camacho-Artacho, M., Lepore, R., Mattos, M.M., Rocha, L.M., Valencia, A.: Prevalence and differences in the co-administration of drugs known to interact: an analysis of three distinct and large populations. BMC Medicine **22**(1), 166 (Apr 2024)
37. The Gene Ontology Consortium: Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research **45**(D1), D331–D338 (Jan 2017), epub 2016 Nov 29
38. Thomas, P.D.: The Gene Ontology and the meaning of biological function. Methods in molecular biology (Clifton, N.J.) **1446**, 15–24 (2017)
39. Tornio, A., Filppula, A.M., Niemi, M., Backman, J.T.: Clinical Studies on Drug–Drug Interactions Involving Metabolism and Transport: Methodology, Pitfalls, and Interpretation. Clinical Pharmacology and Therapeutics **105**(6), 1345–1361 (Jun 2019)
40. Vo, T.H., Nguyen, N.T.K., Kha, Q.H., Le, N.Q.K.: On the road to explainable AI in drug-drug interactions prediction: A systematic review. Computational and Structural Biotechnology Journal **20**, 2112–2123 (Jan 2022)
41. Wang, Y., Yang, Z., Yao, Q.: Accurate and interpretable drug-drug interaction prediction enabled by knowledge subgraph learning. Communications Medicine **4**(1), 59 (Mar 2024)
42. Wang, Y., Fang, F., Liu, X.: Targeting histamine in metabolic syndrome: Insights and therapeutic potential. Life Sciences **358**, 123172 (Dec 2024)
43. Yafune, A., , Mamoru, N., , Ishiguro, M.: A Note on Sample Size Determination for Akaike Information Criterion (AIC) Approach to Clinical Data Analysis. Communications in Statistics - Theory and Methods **34**(12), 2331–2343 (Dec 2005)
44. Yu, H., Mao, K.T., Shi, J.Y., Huang, H., Chen, Z., Dong, K., Yiu, S.M.: Predicting and understanding comprehensive drug-drug interactions via semi-nonnegative matrix factorization. BMC Systems Biology **12**(1), 14 (Apr 2018)
45. Yu, H., Zhao, S., Shi, J.: STNN-DDI: a Substructure-aware Tensor Neural Network to predict Drug–Drug Interactions. Briefings in Bioinformatics **23**(4), bbac209 (Jul 2022)
46. Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., Wang, L., Li, C., Sun, M.: Graph neural networks: A review of methods and applications. AI Open **1**, 57–81 (Jan 2020)