FragXsiteDTI: an interpretable transformer-based model for drug-target interaction prediction

Ali Khodabandeh Yalabadi*
Department of Industrial Engineering
University of Central Florida
Orlando, FL 32816
yalabadi@ucf.edu

Mehdi Yazdani-Jahromi*
Department of Computer Science
University of Central Florida
Orlando, FL 32816
yazdani@ucf.edu

Niloofar Yousefi
Department of Industrial Engineering
University of Central Florida
Orlando, FL 32816
niloofar.yousefi@ucf.edu

Aida Tayebi
Department of Industrial Engineering
University of Central Florida
Orlando, FL 32816
aida.tayebi@ucf.edu

Sina Abdidizaji
Department of Industrial Engineering
University of Central Florida
Orlando, FL 32816
sina.abdidizaji@ucf.edu

Ozlem Ozmen Garibay
Department of Industrial Engineering and
Department of Computer Science
University of Central Florida
Orlando, FL 32816
ozlem@ucf.edu

Abstract

Drug-Target Interaction (DTI) prediction is vital for drug discovery, yet challenges persist in achieving model interpretability and optimizing performance. We propose a novel transformer-based model, FragXsiteDTI, that aims to address these challenges in DTI prediction. Notably, FragXsiteDTI is the first DTI model to simultaneously leverage drug molecule fragments and protein pockets. Our information-rich representations for both proteins and drugs offer a detailed perspective on their interaction. Inspired by the Perceiver IO framework, our model features a learnable latent array, initially interacting with protein binding site embeddings using cross-attention and later refined through self-attention and used as a query to the drug fragments in the drug's cross-attention transformer block. This learnable query array serves as a mediator and enables seamless information translation, preserving critical nuances in drug-protein interactions. Our computational results on two benchmarking datasets demonstrate the superior predictive power of our model over several state-of-the-art models. We also show the interpretability of our model in terms of the critical components of both target proteins and drug molecules within drug-target pairs.

*These authors contributed equally.

The NeurIPS 2023 Workshop on New Frontiers of AI for Drug Discovery and Development (AI4D3 2023), New Orleans, LA, USA, 2023.

1 Introduction

Drug–target interaction (DTI), representing the binding relationship between a drug and its target, is pivotal for developing new drugs and/or repurposing existing ones. While computational approaches have been present for several decades, their increased effectiveness and prominence as alternatives to High-Throughput Screening (HTS) have mainly been realized with the advancements of Machine Learning (ML) algorithms, with deep learning models offering a significant improvement in the accuracy of DTI predictions. Deep Learning's advantage lies in its ability to automatically capture valuable latent features, enabling it to handle intricate patterns in molecular data effectively.

However, a notable constraint in utilizing deep learning models for drug discovery lies in the inherent lack of interpretability. While these models are powerful in predicting potential drug candidates, they often fall short in providing meaningful insights into why a particular compound exhibits certain properties or behaviors [22]. This challenge pertains to understanding the intricacies of communication between proteins and drugs, each characterized by its own unique language or representation. Understanding the mechanistic basis of drug efficacy or inefficacy is critical for optimizing drug design [31], refining candidate selection [1], and anticipating potential side effects or unforeseen interactions [17], all of which are fundamental to the drug development process.

To address this challenge and further enhance the performance of deep learning models in drug discovery, we need **information-rich representations** of both proteins and drugs. These representations should encapsulate the structural, chemical, and functional aspects of each component comprehensively. For proteins, this may involve encoding details about their 3D structures, amino acid sequence, and binding sites. Similarly, it may entail capturing information about molecular fragments, chemical properties, and pharmacological characteristics of drugs. Another crucial component is a **mediator that serves as a common language**, bridging the gap between the two linguistic worlds of drug and protein. Such a common language facilitates information translation from the protein language to the drug language and vice versa. This translation process enables the model to effectively understand and interpret the interactions between proteins and drugs. This ensures that the information conveyed by proteins and drugs is not lost in translation, allowing the model to capture the nuances of their interaction, leading to improved interpretability and performance in DTI prediction tasks.

Considering these challenges and leveraging the recent advancements in transformer-based models, we present FragXsiteDTI, an innovative transformer-based model that takes a new perspective on Drug-Target Interaction (DTI) prediction. Our approach provides a promising solution to both challenges by incorporating information-rich representations for both proteins and drugs alongside the integration of a learnable mediator that seamlessly connects these two distinct domains of drugs and proteins.

Our approach hinges on utilizing molecule fragments and protein pockets as the primary inputs to the model, a paradigm shift that enhances our understanding of the intricate interplay between drugs and their target proteins. Limited prior studies in the field have predominantly focused on using either protein pockets [32] or drug fragments for their DTI prediction models. However, to our knowledge, no existing research has harnessed protein pockets and drug fragments simultaneously as inputs in the context of DTI prediction. This finer granularity allows for a more precise analysis of which parts of the drug are critical for binding to specific binding sites of the protein targets. This information is pivotal in targeted and rationale drug design and generation. Instead of designing drugs as whole molecules and hoping to achieve effective interaction with a target protein, scientists can strategically design, combine or modify these fragments to optimize their ability to interact with specific proteins. Additionally, this detailed perspective goes beyond predictions to explanation by delving into the underlying mechanism of the casual-effect relationship. Fragments often represent functional groups or motifs within drugs that directly contribute to their pharmacological activity. This granular-level analysis can help answer fundamental questions, such as why particular functional groups or binding regions are critical for interaction or why some drug-protein pairs do not exhibit the desired interaction.

We also introduce a transformer-based architecture inspired by the Perceiver IO framework, which facilitates the use of a mediator, enabling seamless communication between the distinct linguistic realms of proteins and drugs. This mediator is indeed a learnable latent array that undergoes a dynamic learning process. Initially, it is shaped based on the unique characteristics of proteins, allowing the model to focus on critical binding pockets of the protein. As the process unfolds, this

query array undergoes adjustments influenced by the self-attention block. Then, our end-to-end learning process allows the model to fine-tune its focus, aligning the latent query with essential drug-related information. This learnable latent query array, guided by both proteins and drugs, is at the heart of our model's ability to decipher intricate drug-protein interactions effectively.

Visualization of the proposed framework can be found in Figure 1. The computational results on two datasets demonstrate the predictive power of our FragXsiteDTI compared to several state-of-the-art models and across multiple evaluation metrics. Also, our model is the first and the only one providing an information-rich interpretation of the interaction in terms of the critical parts of the target protein and drug molecule in a drug-target pair.

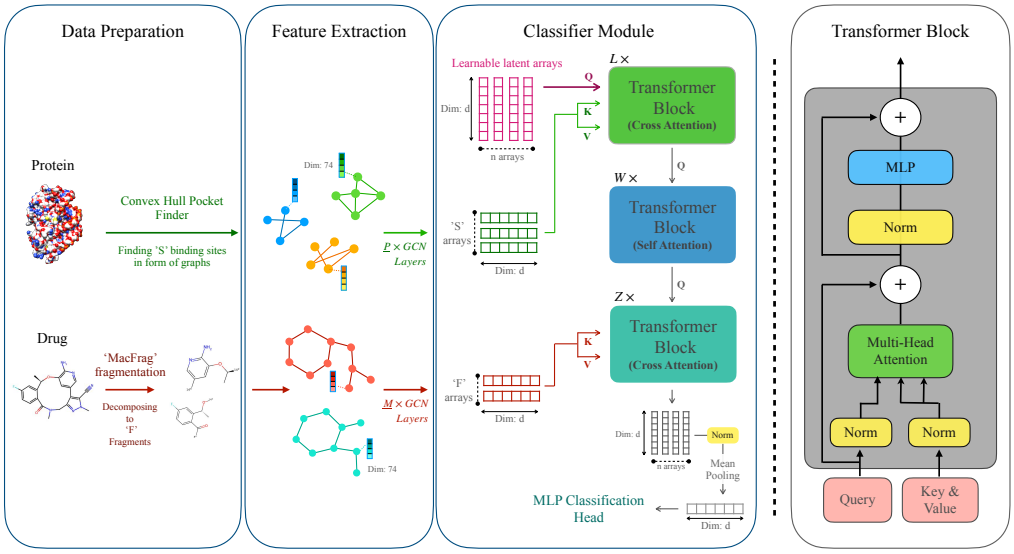


Figure 1: Our proposed framework, FragXsiteDTI, includes three main modules: (1) Preprocessing module, which consists of finding the binding sites of proteins, decomposing drugs' molecules into smaller fragments, and passing all of them to the next module in the form of graphs; (2) Feature extraction module, where we get graph representations of drugs' fragments and proteins' binding sites, and we create two multiple-layer graph convolutional neural networks to extract learnable embeddings; (3) Classifier module, where we introduce learnable latent arrays to first, find the most probable proteins' binding site(s) for interaction in a cross-attention transformer block(TB), second, pass through a self-attention TB to be prepared for finding the most probable drugs' fragment(s) in last cross-attention TB - Transformer Block (TB) at the right shows the details of TBs in our classifier module.

2 Related works

The utilization of deep learning methods has proven to be effective in addressing the complex challenge of predicting Drug-Target Interactions (DTI). These methods vary in their design and approaches for representing input data.

Initially, powerful models relied on one-dimensional data structures and employed raw SMILES strings and protein sequences as inputs [14, 9]. While one-dimensional representations suffice for small drug molecules, the larger and intricately interacting proteins often demand more comprehensive 3D representations.

Despite the limited availability of datasets containing 3D protein structures, recent developments in deep learning have incorporated these structures into their investigations [28, 18, 21]. The adoption of molecular graph representations for drugs, as seen in studies by Tsubaki et al. [25] and Nguyen et al. [13], inspired further research using graph convolutional networks to leverage protein 3D structures for DTI prediction, alongside graph representations of ligands [5, 8, 20, 11].

The introduction of Pocket Feature, an unsupervised autoencoder model proposed by Torng et al. [23], focused on learning representations from binding sites within target proteins. However, these studies encountered limitations, primarily due to the challenge of obtaining high-quality 3D protein structures through experiments. Consequently, there is a scarcity of datasets containing 3D structural information [34]. Zheng et al. [34] proposed an alternative approach by using 2D distance maps to represent proteins instead of direct 3D structure inputs.

With the emergence of Transformers and their demonstrated effectiveness, several studies sought to enhance DTI prediction using this architecture [2, 6]. Recent works have placed a significant emphasis on feature representation, and leveraged new capabilities like attention mechanisms and transformer blocks [32, 15, 30, 29].

Further details of this comprehensive literature review can be found in subsection 6.1.

3 Methodology

Our architecture is delineated into three principal modules:

1. **Data Preparation Module**: This module employs a fragmentation algorithm to dissect drugs and a simulation method to identify protein binding sites. Subsequently, a unique graph is constructed for each drug fragment and protein binding site. A comprehensive explanation is provided in Subsection 3.1.
2. **Feature Extraction**: Within this module, we utilize message-passing neural networks (MPNNs) to extract features from fragments and sites, encapsulating them into desired-sized embeddings. Further details are expounded in Subsection 3.2.
3. **Classifier Module**: Leveraging inspirations from the Transformer and Perceiver IO architectures, our approach systematically models drug-protein interactions in a tripartite manner. Initially, a learnable query is employed to attend to protein embeddings. This query undergoes refinement and subsequently interacts with drug fragment embeddings, culminating in a holistic interaction representation. Detailed insights are presented in 3.3.

A schematic representation of the entire architecture is illustrated in Figure 1.

3.1 Data Preparation

3.1.1 Extracting Fragments from The Ligand Molecule (MacFrag)

Creating high-quality fragment libraries by dividing organic compounds is a crucial aspect of drug discovery. Utilizing fragments of drug molecules can result in a more information-rich representation embedding. In this regard, we chose a recent paper in this domain called Macfrag [3]. This study introduces a novel approach for efficiently fragmenting molecules. MacFrag employs an adapted version of BRICS rules to cleave chemical bonds and introduces an efficient algorithm for swiftly extracting subgraphs, enabling rapid enumeration of the fragment space. Using this approach, the size of fragments can vary depending on the number of chosen building blocks, enabling flexibility. This method permits overlaps between fragments, ensuring that no critical information is overlooked. Moreover, based on their experiments, the fragments generated using this approach exhibit a closer adherence to the 'Rule of Three.'

3.1.2 Extracting Protein's Ligand-binding Pockets

We leverage the 3D configurations of proteins derived from the Protein Data Bank (PDB) files. PDB datasets comprise experimental measurements from Nuclear Magnetic Resonance (NMR), x-ray diffraction, cryo-electron microscopy, etc. associated with proteins. To identify protein binding sites, we employ the algorithm introduced by Saberi Fathi et al. [19]. This approach stands out for its simplicity in extracting protein binding sites from their 3D configurations. It operates on a simulation-based paradigm and can be applied before feeding the data to a deep-learning architecture. It simplifies the deep-learning models because they are no longer responsible for processing large molecules of protein. The algorithm calculates bounding box coordinates for each protein binding site. These coordinates subsequently simplify the entire protein structure to a selection of peptide segments.

3.1.3 Graph Construction

Post the extraction of protein binding sites and ligand fragmentation, we devised a representation technique. For proteins, atoms in binding sites are nodes in distinct graphs, with edges defined by inter-atomic distances below a threshold. For ligand fragments, edges are determined by atomic bonds. Atom features, encoded via one-hot vectors, encompass atom type, degree, implicit valence, charge, radical electrons, hybridization, aromaticity, and attached hydrogen count, yielding a 1×74 vector per node (detailed in Table 3 which can be found in the supplementary material). This methodology produces multiple graphs for both individual proteins and distinct ligands.

3.2 Feature Extraction

Subsequent to this Data Preparation stage, we employed message-passing neural networks (MPNNs) to encapsulate the constructed graphs into embeddings. This serves as the feature extraction component of our architecture. In the supplementary materials (Subsection 6.3), we delineate the descriptions and functionalities of the layers utilized. We combined two layers of TAGCN followed by a GAT layer; the model harnesses the strengths of both methods, potentially outperforming architectures that rely solely on TAGCN or GAT. The TAGCN layers adeptly capture varying local structures by considering different powers of the adjacency matrix, ensuring sensitivity to the graph's topology and enhancing local feature extraction [4]. Subsequently, the GAT layer introduces an attention-based pooling mechanism, allowing nodes to assign varying importance scores to their neighbors, emphasizing more relevant nodes and down-weighting less pertinent ones [27]. This selective attention mechanism leads to more discriminative graph embeddings, especially in graphs with varying node importance. Moreover, the combination ensures that the model recognizes diverse local structures while discerning node importance, offering a dual capability beneficial in complex graphs. Additionally, the GAT layer can mitigate the over-smoothing problem often seen in deep graph neural networks, ensuring distinct and informative node representations [10].

3.3 Classifier Module

We present a method to model drug-protein interactions, drawing inspiration from the Transformer architecture [26] and the Perceiver IO model [7]. Our approach is delineated into three distinct stages:

1. **Cross-Attention with Learnable Query:** We initiate with a learnable latent query array. This query attends to protein binding site embeddings through a cross-attention mechanism. The outcome of this stage is a weighted representation of protein binding sites based on the learnable query.
2. **Latent Query Processing:** The weighted representation from the first stage undergoes further refinement via a self-attention Transformer block. This processed query encapsulates the nuanced features of protein binding sites, preparing it for interaction modeling with drug fragments.
3. **Drug-Protein Interaction Modeling:** In this stage, the processed query from the preceding step acts as Q in a cross-attention Transformer block. Drug fragment embeddings serve as both K and V. This setup allows the model to focus on drug fragments that are most pertinent in relation to the protein binding sites, yielding a holistic representation of drug-protein interaction dynamics.

Our approach to modeling drug-protein interactions, while bearing foundational similarities with the Perceiver IO architecture [7], diverges in its handling of the latent queries and the source of keys and values for attention. Both methodologies employ learnable latent queries to attend to input data, facilitating the extraction of intricate patterns without the need for domain-specific architectures. However, in our method, we introduce a unique twist after the initial cross-attention between the latent query and protein binding site embeddings and subsequent latent query processing. Instead of relying on another set of learnable queries or expert-generated ones for outputs, as in Perceiver IO, we source the keys and values directly from the drug fragment space. This processed query then interacts with the drug fragment embeddings in a cross-attention Transformer block, with the query as Q and the drug fragment embeddings serving as both K and V. This design choice tailors our approach to more effectively capture the nuances of drug-protein interactions.

Self-Attention Mechanism assigns weights to different segments of an input sequence when formulating an output sequence. Defined as Equation 1.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is their dimensionality. This mechanism adeptly captures long-range dependencies without the constraints of recurrent layers.

Cross-Attention Mechanism the queries come from one sequence (or representation), while the keys and values come from another. This allows the model to focus on relevant parts of the second sequence based on the information from the first, effectively bridging information between two distinct sources.

Transformer Block is a composite of attention layers, feed-forward networks, and normalization organized in a layered fashion.

4 Experiments

4.1 Datasets

We establish the effectiveness of our proposed model through a series of comparative experiments. In these experiments, we evaluate the performance of FragXsiteDTI alongside several state-of-the-art methods. To do so, we employ two benchmark datasets that offer essential 3D structural information of target proteins, a crucial requirement for our model.

Human and *C.elegans* These datasets were constructed by amalgamating a collection of exceptionally trustworthy and dependable negative drug-protein samples through a systematic in silico screening technique, which was then combined with the known positive samples [12]. The human dataset comprises 3,369 positive interactions involving 1,052 distinct compounds and 852 unique proteins; the *C. elegans* dataset encompasses 4,000 positive interactions, encompassing 1,434 distinct compounds and 2,504 unique proteins. To facilitate a direct comparison, we adopted identical train, validation, and test partitions (80%, 10%, 10%) as those employed in the recent studies [34, 32].

4.2 Implementation and evaluation

Experimentation strategies. For our implementations, we utilized PyTorch 1.8.2, a long-time support version. The supplementary material contains all the hyperparameters employed for our model (Table 4). The experimentation was conducted on an Nvidia RTX 3090 GPU with 24 GB of memory.

Evaluation metrics. We conducted evaluations of our models using various metrics, such as the Area Under the Receiver Operating Characteristic Curve (AUC), precision, recall, and the F1 score.

4.3 Comparison on target datasets

4.3.1 Human

We conducted a comprehensive comparison of our model against recently developed deep learning-based approaches, including GraphDTA [13], CPI–GNN [25], DrugVQA [34], AttentionSiteDTI [32], DeepDTA [14], DeepConv-DTI [9], MolTrans [6], TransformerCPI [2], as well as CSDTI [15]. In Section 6.1, you can find a brief overview of each of these models.
The performance of these models is summarized in Table 1. Notably, our proposed model demonstrates superior prediction performance compared to all these models. It achieves competitive results with AttentionSiteDTI, which currently holds the top performance among them. Deep learning models prove to be highly effective in extracting essential features governing the intricate interactions within drug-target pairs. Building upon this foundation, our model further enhances accuracy, highlighting the quality of features extracted and learned during the end-to-end training process of fragXsiteDTI.

Table 1: Human Dataset Comparision

	AUC	Precision	Recall	F1 Score
GraphDTA	0.960	0.882	0.912	0.897
GCN	0.956	0.862	0.928	0.894
CPI–GNN	0.970	0.923	0.918	0.920
DrugVQA	0.979	0.954	0.961	0.957
DeepDTA	0.972	0.938	0.935	0.936
DeepConv-DTI	0.967	0.939	0.907	0.923
MolTrans	0.974	0.955	0.933	0.944
TransformerCPI	0.97	0.911	0.937	0.924
AttentionSiteDTI	0.991	0.951	**0.975**	0.963
CSDTI	0.982	0.937	0 .946	0.941
AMMVF-DTI	0.986	0.976	0.938	0.957
FragXsiteDTI(Ours)	**0.991**	**0.977**	0.952	**0.964**

4.3.2 C.elegans

For this dataset, we compared our model with deep learning models that had high performance based on the selected metrics. These models include MDL-CPI [30], CPI-GNN [25], graph convolutional network (GCN), GraphDTA [13], TransformerCPI [2], and AMMVF-DTI [29]. The performance of these models is summarized in Table 2. Similar to the Human dataset, our model outperforms the existing good models in all prediction metrics.

Table 2: *C.elegans* Dataset Comparision

	AUC	Precision	Recall	F1 Score
MDL-CPI	0.975	0.943	0.923	0.933
CPI-GNN	0.978	0.938	0.929	0.933
GCN	0.975	0.921	0.927	0.924
GraphDTA	0.974	0.927	0.912	0.919
TransformerCPI	0.988	0.952	0.953	0.952
AMMVF-DTI	0.990	0.962	0.96	0.961
FragXsiteDTI(Ours)	**0.992**	**0.971**	**0.971**	**0.971**

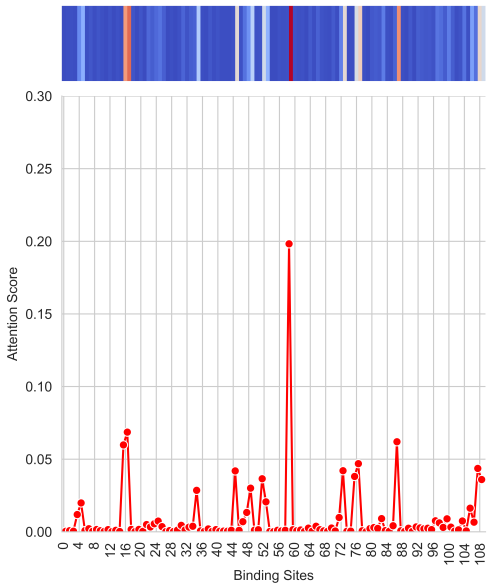
4.4 Interpretation

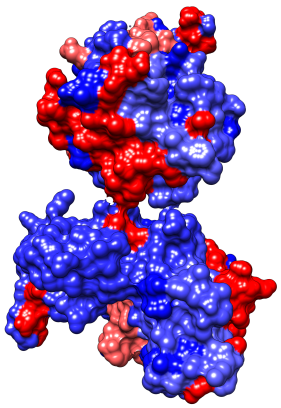
4.4.1 Protein

In this study, we leverage the attention mechanism to enhance the model's ability to predict the likelihood of specific protein binding sites interacting with a given ligand. This likelihood is quantified through the attention matrix computed within the model. The visualization of this attention mechanism can be observed in Figure 2, where it is presented as a heatmap for the protein with PDB code of 4BHN when interacting with a drug characterized by the molecular formula of $C_{21}H_{23}Cl_2N_5O_2$. This figure also includes the projection of the heatmap onto the protein structure.

4.4.2 Drug

We also have an attention matrix for each drug molecule that determines which fragments of that drug have the highest probabilities for interaction with the particular protein. These candidate fragments can explain the chemical properties that caused the interaction or be used for designing and generating new drugs. Figure 3 demonstrate an example of certain drug ($C_{21}H_{23}Cl_2N_5O_2$) that binds with the target protein (4BHN).





PDB code: 4BHN

Figure 2: (Left) a heatmap and line plot that represent the cross-attention mechanism weights for every binding site with the latent array. These weights signify the likelihood of each computed binding site on the protein becoming active when interacting with the specific ligand ($C_{21}H_{23}Cl_2N_5O_2$). (Right) a heatmap that projects the cross-attention weights onto the protein with the PDB code 4BHN. This visualization illustrates our model's interpretability with respect to the proteins. The protein's visualization was generated using UCSF Chimera software [16].

5 Conclusion

In this work, we introduced a groundbreaking method for modeling drug-protein interactions, drawing from the robustness of both the Transformer and Perceiver IO architectures. Our empirical evaluations on Human and *C. elegans* datasets not only underscore the superiority of our approach in terms of predictive accuracy but also highlight its unparalleled interpretability. One of the standout features of our method is its capability to pinpoint which fragment of a drug interacts with specific regions of a protein. This granularity is invaluable, offering researchers a detailed map of interaction hotspots, which can guide drug modifications and optimizations. The use of attention modules doesn't merely serve as a mechanism for improved performance; it acts as a window into the model's decision-making process. By visualizing attention scores, we can discern the importance the model assigns to different fragments, shedding light on potential pharmacologically active regions. With the insights provided by our model, drug designers can make informed decisions, potentially reducing the trial-and-error nature of drug discovery. By knowing which fragments are likely to interact with target proteins, drug modifications can be more strategic and purpose-driven. Given the model's precision in understanding drug-protein dynamics, there's potential for tailoring drug designs to individual protein structures, paving the way for more personalized therapeutic interventions in the future. While our current evaluations are on specific datasets, the foundational architecture suggests potential scalability to other organisms and broader drug-protein interaction landscapes, making it a versatile tool in the bioinformatics toolkit. As the pharmaceutical industry and medical research communities continue their quest for more effective and targeted drugs, our method stands out as a beacon, promising to play a transformative role in the future landscape of drug discovery and design.

Data, Code and more results availability There are more datasets and recent results in our arXiv version `https://arxiv.org/abs/2311.02326`. Also, all datasets, instructions, and codes for our experiments are publicly available at `https://github.com/yazdanimehdi/FragXsiteDTI`.

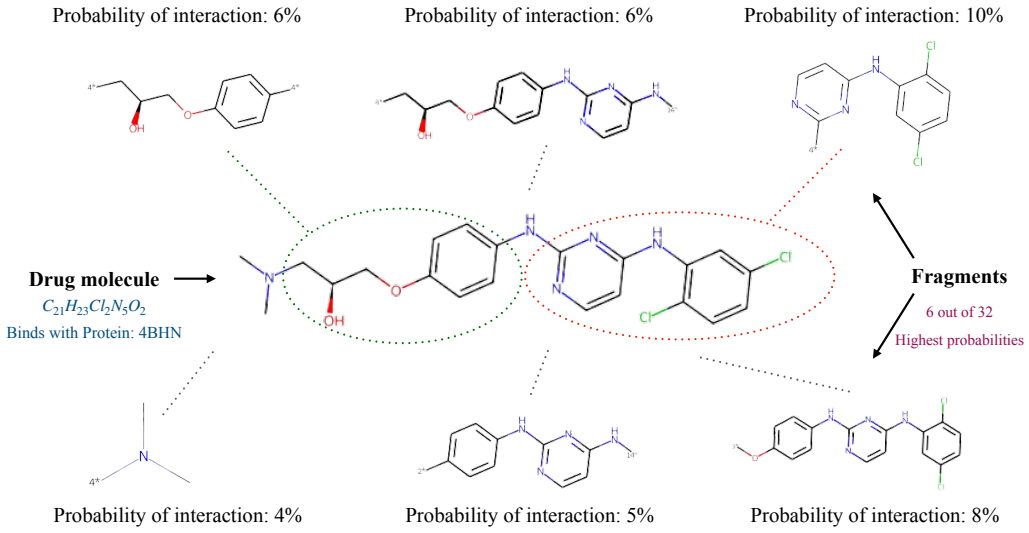


Figure 3: This figure demonstrates the interpretability of our model regarding the drug molecules. It showcases the specific drug molecule that binds to protein 4BHN, along with its six fragments exhibiting the highest interaction probabilities among a total of 32 fragments. These probabilities are normalized, and 10% is the highest. Note that since these fragments can be repeated in other fragments, these probabilities can be summed up to higher numbers for each fragment.

6 Supplementary Material

6.1 Literature review

Deep learning-based approaches have emerged as effective solutions for addressing the challenging problem of Drug-Target Interaction (DTI) prediction. These approaches exhibit variations in both their architectural design and their strategies for representing input data.

In DeepDTA [14], a Convolutional Neural Network (CNN) is employed to analyze both the raw SMILES string and the protein sequence, allowing for the extraction of local residue patterns. The primary objective is to predict binding affinity values. To transform this into a binary classification problem, a Sigmoid activation function can be introduced at the end of the model. DeepConv-DTI [9] employs CNN along with a global max-pooling layer to capture local patterns of varying lengths within the protein sequence. Additionally, it applies a fully connected layer to the drug fingerprint Extended-Connectivity Fingerprint 4 (ECFP4).

Notably, while small drug molecules can be efficiently represented in one-dimensional space, proteins, due to their larger size and intricate interactions, often necessitate more comprehensive 3D representations. Despite the limited availability of datasets containing 3D protein structures, recent deep learning literature has increasingly incorporated these structures into their investigations. For instance, AtomNet [28] pioneered the utilization of 3D protein structures as input for a 3D Convolutional Neural Network (CNN), enabling the prediction of drug-target binding using a binary classifier. Ragoza et al. [18] proposed a CNN scoring function that harnessed the 3D representation of protein-ligand complexes to discern critical features crucial for binding prediction, surpassing the performance of the AutoDock Vina score. Pafnucy [21], another significant advancement, employed 3D CNNs to predict binding affinity values for drug-target pairs. Their approach involved representing inputs as a 3D grid, treating both protein and ligand atoms similarly. By applying regularization techniques, their designed network focused on capturing the general properties governing interactions between proteins and ligands.

However, these studies face several limitations, primarily stemming from the considerable difficulty in experimentally acquiring high-quality 3D protein structures. Consequently, there is a scarcity of datasets containing 3D structural information [34]. Moreover, most studies that employ 3D structural

data predominantly rely on CNNs, which exhibit sensitivity to structural orientations and are computationally demanding.

CPI-GNN [25], and GraphDTA [13] employ graph convolutional network (GCN) for molecular graph representation of drugs, and improve the prediction accuracy in the result. Some other studies [5, 8] introduced the use of GCN approaches to take advantage of protein 3D structures as input for DTI prediction. Some studies have extended the application of GCNs to protein-ligand complexes. A notable example is GraphBAR [20], which stands as the pioneering 3D graph CNN utilizing a regression approach for predicting drug-target binding affinities. Instead of relying on 3D voxelized grids, GraphBAR employs graphs to represent protein-ligand complexes. These graphs manifest in the form of multiple adjacency matrices, with entries calculated based on distances and feature matrices encapsulating molecular properties of atoms. Additionally, GraphBAR augments its model with data derived from docking simulations.

Lim et al. [11] introduced a graph convolutional network model complemented by a distance-aware graph attention mechanism. This model extracts interaction features directly from the 3D structures of drug-target complexes generated through docking software. While their model demonstrated improved performance compared to both docking simulations and various deep learning-based models, it exhibited limitations, including reduced explainability and the introduction of additional docking errors into the deep learning model.

Pocket Feature, proposed as an unsupervised autoencoder model by Torng et al. [23], specializes in learning representations from binding sites within target proteins. The model employs 3D graph representations for protein pockets and 2D graph representations for drugs. It trains a GCN model to extract features from these graph representations and drug SMILEs. Notably, Pocket Feature outperforms 3D CNNs, as demonstrated in reference [18], and also surpasses docking simulation models such as AutoDock Vina [24], RF-Score [12], and NNScore [12].

Zheng et al. [34] drew attention to the inefficiency of employing direct 3D structure inputs. Instead, they adopted 2D distance maps to represent proteins. In doing so, they reformulated the DTI prediction problem as a classical visual question-answering (VQA) task. In this paradigm, given a distance map of a protein, the model determines whether a given drug interacts with the target protein. Although their model outperformed several state-of-the-art models, it was primarily tailored to classification, predicting interactions between drug-target pairs, without any interpretability.

With the emergence of Transformers and their proven power, TransformerCPI [2] utilizes sequence representation of drugs and proteins. After a customized encoder layer, this model employs a transformer decoder to construct an interaction sequence and predict the interaction based on that. MolTrans [6] combines the transformer's capability with sub-structures of drugs and targets. This model decomposes drugs and proteins into smaller structures based on their graphs and sequences respectively. Then, an interaction module is responsible for explicable prediction.

In more recent studies, Yazdani et al. [32] exploited 3D protein binding sites along with graph attention embeddings for both drug and protein. Their model finds protein binding sites by a simple docking-based model proposed by Fathi et al. [19]. By using a self-attention module and sentence-level relation from NLP literature, this model determines the most probable candidate binding site for interaction resulting in interpretability as well as high accuracy. Furthermore, they demonstrated that the 3D structure of protein binding sites carries an immense amount of information related to binding affinity and it can improve the performance of benchmark models just by adding this information to the classifier [33].

CSDTI [15] is another model that focuses on the representation of drugs and proteins to improve interpretability and gain better performance in prediction. They employed a drug molecule aggregator and the multiscale 1D convolution-based protein encoder in order to extract the best representation and a cross-attention block to learn the relation between them. Focusing on representation, the MDL-CPI model [30] considers proteins as words and employs a hybrid network architecture based on BERT and CNN to extract the feature representations. Following this work, AMMVF-DTI [29] is a model that utilizes local and global information of proteins and ligands in the form of node-level and graph-level embeddings, respectively. This model tried to improve performance with attention mechanisms and interactive information.

Table 3: Encoding atom features in each extracted fragment and binding site

Type	Encoding
Atom type	C, N, O, S, F, Si, P, Cl, Br, Mg, Na, Ca, Fe, As, Al, I, B, V, K, Tl, Yb, Sb, Sn, Ag, Pd, Co, Se, Ti, Zn, H, Li, Ge, Cu, Au, Ni, Cd, In, Mn, Zr, Cr, Pt, Hg, Pb (one-hot encoding)
Degree of atom	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 (one-hot encoding)
Implicit valence of an atom	0, 1, 2, 3, 4, 5, 6 (one-hot encoding)
Formal charge for an atom	Value
Number of radical electrons for an atom	Value
Hybridization of an atom	SP, SP2, SP3, SP3D, SP3D2 (one-hot encoding)
Is aromatic	0 or 1
Number of hydrogen attached	0, 1, 2, 3, 4 (one-hot encoding)

Table 4: Hyperparameters

Hyperparameters	Value
GCN Protein	TAG(74, 74), TAG(74, 128), GAT(128, 256)
GCN Drug	TAG(74, 74), TAG(74, 128), GAT(128, 256)
Optimizer	AdamW
Weight Decay	0.03
Epochs	100
Dropout	0.05
Drop Path	0.05
Scheduler	0.98
Dim	256
Depth SA	2
Depth CA Input	2
Depth CA Output	2
Latent Size	200
Num Head	4
Learning Rate	1e-4

6.2 Tables

6.3 MPNN Networks

Topology Adaptive Graph Convolutional Networks (TAGCN) is a graph representation learning approach that captures the local structures of a graph by applying polynomial filters to its adjacency matrix. Unlike traditional Graph Convolutional Networks (GCNs) that use a fixed neighborhood size, TAGCN considers varying neighborhood sizes by leveraging different powers of the adjacency matrix. Specifically, the method operates as demonstrated in Equation 2.

$$H^{(l+1)} = \sigma\left(\sum_{k=0}^{K} \Theta_k^{(l)} A^k H^{(l)}\right) \tag{2}$$

where $H^{(l)}$ is the feature matrix at layer l, A is the adjacency matrix, $\Theta_k^{(l)}$ is the trainable weight matrix for the k-th power of the adjacency matrix at layer l, K is the maximum power considered, and σ is a non-linear activation function [4].

Graph Attention Networks (GAT) introduce an attention mechanism to graph neural networks, allowing nodes to weigh their neighbors differently during aggregation. Instead of uniformly aggregating information from neighbors like traditional GCNs, GAT computes attention coefficients that capture the importance of each neighbor node. The main operation in GAT is described by Equation

3.

$$h_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} W^{(l)} h_j^{(l)}\right) \tag{3}$$

where $h_i^{(l)}$ is the feature of node i at layer l, $W^{(l)}$ is a weight matrix, $\alpha_{ij}^{(l)}$ is the attention coefficient between nodes i and j at layer l, and σ is a non-linear activation function. The attention coefficients are computed using a shared attention mechanism across all node pairs, making the model's capacity invariant to the graph size [27].

6.4 Training procedure

Algorithm 1 shows the detailed steps of the training process happening in the classifier module (Fig. 1).

Algorithm 1 FragXsiteDTI Training Procedure

N: Number of epochs
L : Learn-able latent arrays
G_d: List of graphs of drug fragments in data instance
G_p: List of graphs of protein binding sites in data instance
P : Protein graph convolution network
P_P: Protein Pooling Layer
D : Drug graph convolution network
P_D: Drug pooling Layer
MLP : Multi-layer perceptron classification layers
CA : Cross-Attention Transformer block
SA : Self-Attention Transformer block
F_p: List of feature vectors for protein binding sites in data instance
F_d: List of feature vectors for drug fragments in data instance
y_i: Label of the data instance $= 0, 1$
$p(y_i)$: Predicted label
while Epoch $< N$ **do**
  for $(G_d, G_p, label)$ in data **do**
    $F_p, F_d = [\,], [\,]$
    for $graph$ in G_p **do**
      $F_p \leftarrow F_p \cup [P_P(P(graph)]$
    end for
    for $graph$ in G_d **do**
      $F_d \leftarrow F_d \cup [P_D(D(graph)]$
    end for
    $L \leftarrow CA(L, F_p)$
    $L \leftarrow SA(L)$
    $L \leftarrow CA(L, F_d)$
    $p(y_i) \leftarrow MLP(Mean\_Pooling(Norm(L)))$
    Update L, P, P_P, D, P_D, CA, SA, MLP by descending
    $-\nabla y_i \log(p(y_i)) + (1 - y_i) \log(1 - p(y_i))$
  end for
end while

References

[1] 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.

[2] Lifan Chen, Xiaoqin Tan, Dingyan Wang, Feisheng Zhong, Xiaohong Liu, Tianbiao Yang, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, and Mingyue Zheng. Transformercpi: improving compound–protein

interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics*, 36(16):4406–4414, 2020.

[3] Yanyan Diao, Feng Hu, Zihao Shen, and Honglin Li. Macfrag: segmenting large-scale molecules to obtain diverse fragments with high qualities. *Bioinformatics*, 39(1):btad012, 2023.

[4] Jian Du, Shanghang Zhang, Guanhang Wu, José MF Moura, and Soummya Kar. Topology adaptive graph convolutional networks. *arXiv preprint arXiv:1710.10370*, 2017.

[5] Joseph Gomes, Bharath Ramsundar, Evan N Feinberg, and Vijay S Pande. Atomic convolutional networks for predicting protein-ligand binding affinity. *arXiv preprint arXiv:1703.10603*, 2017.

[6] Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021.

[7] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.

[8] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

[9] Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6):e1007129, 2019.

[10] Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[11] Jaechang Lim, Seongok Ryu, Kyubyong Park, Yo Joong Choe, Jiyeon Ham, and Woo Youn Kim. Predicting drug–target interaction using a novel graph neural network with 3d structure-embedded graph representation. *Journal of chemical information and modeling*, 59(9):3981–3988, 2019.

[12] Hui Liu, Jianjiang Sun, Jihong Guan, Jie Zheng, and Shuigeng Zhou. Improving compound–protein interaction prediction by building up highly credible negative samples. *Bioinformatics*, 31(12):i221–i229, 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv256.

[13] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: Predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

[14] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[15] Yaohua Pan, Yijia Zhang, Jing Zhang, and Mingyu Lu. Csdti: an interpretable cross-attention network with gnn-based drug molecule aggregation for drug-target interaction prediction. *Applied Intelligence*, pages 1–14, 2023.

[16] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.

[17] António J Preto, Pedro Matos-Filipe, Joana Mourão, and Irina S Moreira. Synpred: Prediction of drug combination effects in cancer using full-agreement synergy metrics and deep learning. 2021.

[18] Matthew Ragoza, Joshua Hochuli, Elisa Idrobo, Jocelyn Sunseri, and David Ryan Koes. Protein–ligand scoring with convolutional neural networks. *Journal of chemical information and modeling*, 57(4):942–957, 2017.

[19] Seyed Majid Saberi Fathi and Jack A. Tuszynski. A simple method for finding a protein's ligand-binding pockets. *BMC Structural Biology*, 14(1):18, 2014. doi: 10.1186/1472-6807-14-18. URL `https://doi.org/10.1186/1472-6807-14-18`.

[20] Jeongtae Son and Dongsup Kim. Development of a graph convolutional neural network model for efficient prediction of protein-ligand binding affinities. *PloS one*, 16(4):e0249404, 2021.

[21] Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018.

[22] Yuxuan Tang. Deep learning in drug discovery: applications and limitations. *Frontiers in Computing and Intelligent Systems*, 3(2):118–123, 2023.

[23] Wen Torng and Russ B. Altman. Graph Convolutional Neural Networks for Predicting Drug-Target Interactions. *Journal of Chemical Information and Modeling*, 59(10):4131–4149, 2019. ISSN 1549-9596. doi: 10.1021/acs.jcim.9b00628.

[24] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2): 455–461, 2010.

[25] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 2019.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[27] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[28] Izhar Wallach, Michael Dzamba, and Abraham Heifets. Atomnet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv:1510.02855*, 2015.

[29] Lu Wang, Yifeng Zhou, and Qu Chen. Ammvf-dti: A novel model predicting drug–target interactions based on attention mechanism and multi-view fusion. *International Journal of Molecular Sciences*, 24(18): 14142, 2023.

[30] Lesong Wei, Wentao Long, and Leyi Wei. Mdl-cpi: Multi-view deep learning model for compound-protein interaction prediction. *Methods*, 204:418–427, 2022.

[31] Jiannan Yang, Zhen Li, William Wu, Shi Yu, Qian Chu, and Qingpeng Zhang. Deep learning can identify explainable reasoning paths of mechanism of drug action for drug repurposing from multilayer biological network. 2022.

[32] Mehdi Yazdani-Jahromi, Niloofar Yousefi, Aida Tayebi, Elayaraja Kolanthai, Craig J Neal, Sudipta Seal, and Ozlem Ozmen Garibay. Attentionsitedti: an interpretable graph-based model for drug-target interaction prediction using nlp sentence-level relation classification. *Briefings in Bioinformatics*, 23(4):bbac272, 2022.

[33] Niloofar Yousefi, Mehdi Yazdani-Jahromi, Aida Tayebi, Elayaraja Kolanthai, Craig J Neal, Tanumoy Banerjee, Agnivo Gosai, Ganesh Balasubramanian, Sudipta Seal, and Ozlem Ozmen Garibay. BindingSite-AugmentedDTA: enabling a next-generation pipeline for interpretable prediction models in drug repurposing. *Briefings in Bioinformatics*, 24(3):bbad136, 04 2023. ISSN 1477-4054. doi: 10.1093/bib/bbad136. URL `https://doi.org/10.1093/bib/bbad136`.

[34] Shuangjia Zheng, Yongjian Li, Sheng Chen, Jun Xu, and Yuedong Yang. Predicting drug–protein interaction using quasi-visual question answering system. *Nature Machine Intelligence*, 2(2):134–140, 2020.