SSEF 2026

CO 023

3DICE: Interpretable 3D Cross-Modal Learning for Drug–Target Interaction Prediction and Large-Scale Drug Discovery

Austin Liu Zi Rui

ABSTRACT

Drug–target interaction (DTI) prediction is a crucial step in modern drug discovery. Accurate and efficient predictions can substantially reduce costs and development time. Applications of deep learning methods for this purpose have been extensively studied in recent years, yielding instrumental contributions to this field. However, existing methods face issues pertaining to efficient learning of drug and target feature representations, which is detrimental to generalisability and performance in cold-start scenarios. Most approaches extract representations from SMILES strings for drugs and FASTA sequences for target proteins, which encode limited 3D structural information. Additionally, many models lack explainability, being black boxes that provide little physical insight into the underlying mechanisms behind such interactions. To address these limitations, we propose 3DICE, a novel framework leveraging cross-attention-based fusion and massively pre-trained 3D structural encoders for both drugs and proteins. UniMol and ESM-IF1 are employed to generate high-fidelity, 3D structure-aware embeddings which enable richer geometric and chemical understanding. Cross-modal fusion modules further augment representations to model intermolecular binding relationships. Importantly, this mechanism also provides intrinsic interpretability, highlighting and enabling qualitative analysis of most influential atoms or residues. Experiments conducted on two canonical benchmark datasets display the competitiveness of our model in real-world scenarios. 3DICE outperformed state-of-the-art models across multiple metrics on the DrugBank and KIBA datasets. Additional experiments provide a more rigorous analysis of interpretability than is typically reported in prior DTI studies, and we find that attention consistently highlights decision-critical regions which is not intrinsically class-specific. Source code is freely available at github.com/austinatose/3DICE.

Keywords: Drug–target interaction prediction, Structure-based representation, Multimodal fusion, Interpretable AI

1 Introduction

Drug–target interaction prediction is a crucial aspect of modern drug discovery [1]. Experimental screening of drug–target interactions (DTIs) is expensive, slow, and failure-prone [2], [3]. In-silico methods help to whittle down the vast search space to a manageable number of DTIs for experimental testing and verification [4], [5], [6]. Despite decades of research, accurate computational DTI prediction remains challenging. The many-to-many nature of DTIs, exacerbated by the chemical diversity of small molecules and structural diversity of proteins, complicates methods [7]. Sparse, noisy and biased experimental data also need to be carefully handled [8]. Several studies have explored docking and physics-based methods, which are accurate and explainable, but are computationally intensive and lack scalability [9], [10]. In recent years, machine learning techniques have emerged as an efficient data-driven alternative [11], [12].

A key component of computational DTI prediction is learning representations for both drugs and target proteins. Early approaches relied on character-level embeddings of Simplified Molecular Input Line System (SMILES) strings for drug molecules [13] and FASTA [14] or amino acid sequences for proteins. Early models such as DeepDTA [15] and DeepCPI [16] adopt this approach, and similar representations continue to be used by more recent methods like MolTrans [15]. While effective in capturing some chemical and sequential patterns, such representations neglect the 3D geometry that fundamentally governs binding.

There have been attempts to better capture molecular structure. Graph-based neural networks (GNNs) model geometry through atom-bond graphs. GraphDTA [17] and GraphCPI [18] use GNNs encoding drug molecular graphs, while TransformerCPI [19] uses Graph Convolutional Networks (GCNs). Graphormer [20] further embeds molecular graphs into vector representations for use in its transformer architecture. In contrast, explicit modelling of protein structures remains comparatively limited, with works citing complexity and demanding computational requirements [21]. Only a small number of works, such as GraphDTI [22], attempt to encode protein structures.

In parallel, large-scale pretrained models have also emerged as a promising alternative in extracting protein and drug features due to their strong generalisation capabilities [21], [23], [24], [25], [26]. On the drug side, sequence-based ChemBERTa [27] and graph-based MG-BERT [28] have been established. On the protein side, the ProtTrans [29] and Evolutionary Scale Modelling (ESM) [30] transformers have demonstrated strong performance on downstream tasks. The EviDTI [21] framework adopts this paradigm, with proteins processed with ProtTrans and drugs with MG-BERT, augmented with geometric representations from GeoGNN. However, current methods

still remain largely sequence or topology-driven and do not explicitly encode full 3D structural information that is pertinent in binding.

A model that stands out is UniMol [31], which introduces a large-scale 3D Molecular Representation Learning framework that allows it to learn structure-aware embeddings from conformers. UniMol has recently seen use in MocFormer [23], where its drug embeddings were paired with ESM-2's protein sequence embeddings. Complementary to this, Evolutionary Scale Modelling – Inverse Folding 1 (ESM-IF1) [32] is a unique pretrained protein structural encoder. Originally designed for predicting protein sequences based on 3D backbone coordinates, it has the novel ability to produce embeddings encoding rich geometric and spatial relationships between protein residues given the 3D protein structure. To the best of our knowledge, ESM-IF1 has not been previously explored in the context of DTI prediction. Despite its strong theoretical potential, its adoption may have been limited due to the scarcity of experimentally resolved protein structures and the substantial preprocessing required. However, recent advances in large-scale protein structure prediction, most notably AlphaFold [33], substantially mitigate this limitation by enabling training on predicted protein structures. These advances together demonstrate that new possibilities for high-fidelity representations for both drugs and proteins are now feasible through the coupling of 3D structure-awareness with the strong generalisation capabilities afforded by massive pretraining.

In a different vein, cross-modal fusion is an integral component of DTI modelling, as final prediction ultimately requires a joint representation of drug and protein features. While a simple concatenation of independently encoded and augmented representations can be used to form such a joint vector [11], [17], [19], [19], [21], [22], from a physical perspective, the interactions are inherently governed by specific local interactions between drugs and proteins, which makes explicit modelling of this qualitatively crucial. Many models have sought to do so. MolTrans utilises a Hadamard-product-based pooling; DrugBAN [34] a bilinear attention network; and MocFormer bilinear pooling. More recently, cross-attention-based [35], [36] approaches have been proposed, including ICAN [37], CAT-DTI [38], and CAIHGNN [39]. In addition to improving predictive performance, such explicit interaction modelling also facilitates interpretability, as learned attention weights or pairwise interaction maps can be analysed to identify the key binding sites for better physical understanding.

In this work, we propose 3DICE, a novel and interpretable framework for DTI prediction. 3DICE leverages massively pretrained 3D structural encoders, namely ESM-IF1 for proteins and UniMol for drugs, to produce high-fidelity, information-rich representations. To the best of our knowledge,

3DICE is the first DTI framework to jointly employ pretrained 3D structural encoders on both the drug and protein sides. Additionally, the framework is explicitly designed to model drug-protein interactions through a co-attention mechanism [40], which enables both improved predictive performance and intrinsic interpretability by highlighting regions critical to interactions. Extensive experiments demonstrate that 3DICE achieves competitive performance on multiple benchmarks compared to state-of-the-art methods. We further conduct novel targeted analyses to examine the interpretability and interaction patterns learned by the model, demonstrating that 3DICE is an effective and interpretable framework for DTI prediction.

The source code is freely available at github.com/austinatose/3DICE.

2 Research Methodology

2.1 Architecture

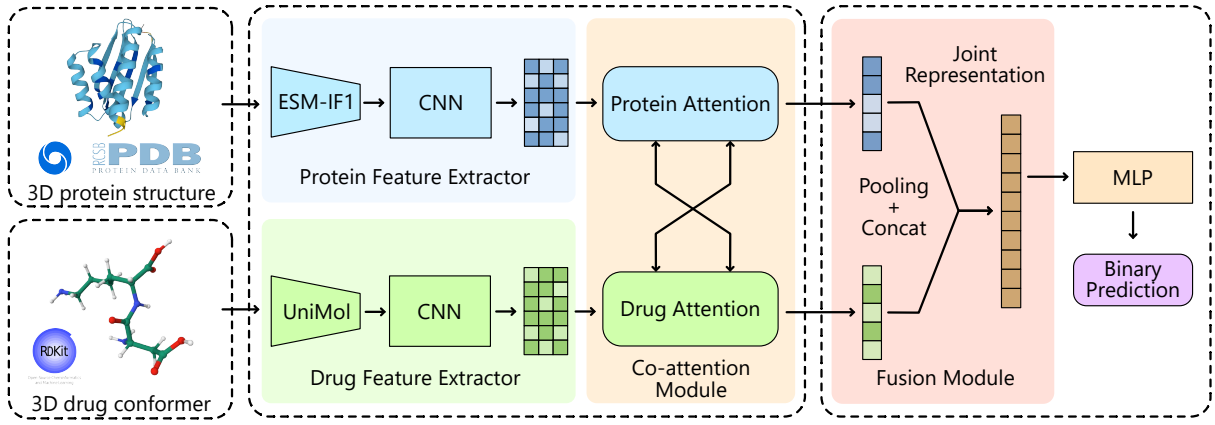

Figure 1: 3DICE Framework

Figure 1 shows the proposed 3DICE framework. The framework comprises five main components: a protein feature extractor, a drug feature extractor, a cross-attention module, a fusion module, followed by a readout multi-layer perceptron (MLP). In the protein feature extractor, the pre-trained model ESM-IF1 was used to encode 3D protein features given the 3D spatial structures. A 2-layer 1D CNN across the sequence dimension was used to extract local substructural patterns. In the drug feature extractor, the pre-trained model UniMol was used to capture 3D stereochemical information from drug conformers. A similar CNN to that on the protein side was implemented to augment drug features. Next, the enhanced protein and drug representations are fed into a co-attention module. By attending protein residues to drug atoms and vice versa, the module explicitly considers bidirectional cross-modal binding interactions. The embeddings are then pooled, reduced in dimension with a 1-layer MLP and concatenated to form a joint representation of each DTI pair. This is forwarded into a fully connected MLP to produce the final prediction. Detailed information about the framework is outlined in Appendix A.

2.2 Datasets

The datasets chosen were DrugBank [41] and KIBA [42]. A cold-start dataset was formulated based on DrugBank. Further details on the datasets are provided in Appendix B.

2.3 Benchmarks

3DICE was benchmarked against three models: MolTrans [15], DLM-DTI [26] and MocFormer [23]. Details of these models are provided in Appendix C. Details on the metrics used are provided in Appendix D.

2.4 Implementation and Hyperparameters

3DICE was implemented in Python 3.12.10, except for the protein-side embeddings, which were precomputed in Python 3.8.20. A batch size of 32, learning rate of 5e-5, dropout of 0.3 and weight decay of 0 was used for 3DICE. Training was performed on a single NVIDIA RTX 5090 GPU with 32GB of VRAM for 60 epochs. Some embeddings were computed on an Apple M1 Pro. More details on the hyperparameters are provided in Appendix E, while dependencies and package versions are listed on the GitHub repository.

3 Results and Discussion

3.1 3DICE achieves superior performance on benchmarks

The performance of 3DICE compared to other state-of-the-art models on the DrugBank and KIBA datasets is summarised in Table 1 and Table 2.

Table 1: Performance comparisons on the DrugBank dataset

DrugBank	ACC	PPV	TPR	F1	MCC	AUROC	AUPRC
MolTrans	0.76205 (0.00372)	0.73491 (0.00552)	0.81959 (0.00025)	0.76359 (0.00703)	0.52764 (0.00689)	0.84060 (0.00267)	0.83459 (0.00387)
DLM-DTI	0.80135 (0.00636)	0.78714 (0.00811)	0.82643 (0.00782)	0.80627 (0.00587)	0.60350 (0.01256)	0.88026 (0.00464)	0.87446 (0.00529)
MocFormer	0.82610 (0.00847)	0.79685 (0.01246)	**0.87493 (0.00494)**	0.83400 (0.00644)	0.65487 (0.01538)	0.89803 (0.00468)	0.88385 (0.00700)
3DICE	**0.84379 (0.00360)**	**0.85058 (0.00539)**	0.83427 (0.00497)	**0.84235 (0.00356)**	**0.68770 (0.00722)**	**0.91065 (0.00216)**	**0.90200 (0.00145)**

The best results are bolded, while the second-best results are underlined.
Five independent replications were performed (n = 5). Data is expressed as mean (std).

3DICE achieves excellent performance on the DrugBank dataset, outperforming the baselines across all reported metrics except for TPR.

Table 2: Performance comparisons on the KIBA dataset

KIBA	ACC	PPV	TPR	F1	MCC	AUROC	AUPRC
MolTrans	0.85213 (0.01237)	0.57484 (0.02922)	**0.85162 (0.01310)**	0.70741 (0.00338)	0.60182 (0.01230)	0.91127 (0.00153)	0.77603 (0.00312)
DLM-DTI	0.88274 (0.00052)	0.71417 (0.03087)	0.63091 (0.06324)	0.66764 (0.02406)	0.60014 (0.01739)	0.90367 (0.00105)	0.73328 (0.00307)
MocFormer	0.89958 (0.00361)	0.74267 (0.02323)	0.71261 (0.01599)	**0.72698 (0.00294)**	**0.66596 (0.00613)**	0.92744 (0.00201)	**0.80280 (0.00231)**
3DICE	**0.90012 (0.00082)**	**0.77473 (0.00864)**	0.65933 (0.00797)	0.71233 (0.00162)	0.65553 (0.00167)	**0.92744 (0.00130)**	0.79357 (0.00051)

The best results are bolded, while the second-best results are underlined.
Five independent replications were performed (n = 5). Data is expressed as mean (std).

It is competitive even on KIBA, which is particularly challenging due to the domination of the negative class. This suggests that incorporating more informative feature representations and detailed interaction modelling contributes to improved performance.

3.2 3DICE is competitive in cold-start scenarios

Table 3: Cold-start performance comparisons on the DrugBank dataset

Cold start	ACC	PPV	TPR	F1	MCC	AUROC	AUPRC
MolTrans	0.71015 (0.00238)	0.66208 (0.00116)	**0.85042 (0.01439)**	0.71533 (0.00838)	0.44288 (0.00317)	0.79196 (0.00091)	0.78525 (0.01008)
DLM-DTI	0.73460 (0.00569)	0.74732 (0.0102)	0.70407 (0.00817)	0.72498 (0.00464)	0.48696 (0.01160)	0.81927 (0.00663)	0.81724 (0.00427)
MocFormer	0.78975 (0.00737)	**0.79985 (0.00343)**	0.76933 (0.01848)	0.78422 (0.00984)	0.57817 (0.01418)	**0.86568 (0.00549)**	**0.86014 (0.00795)**
3DICE	**0.79611 (0.00429)**	0.79637 (0.01001)	0.79278 (0.02721)	**0.79423 (0.00875)**	**0.59270 (0.00903)**	0.86548 (0.00654)	0.85020 (0.00545)

The best results are bolded, while the second-best results are underlined.
Five independent replications were performed (n = 5). Data is expressed as mean (std).

In the cold-start scenario, expected performance degradations are observed for all models. Despite this, 3DICE still demonstrates competitive performance compared to state-of-the-art models, suggesting that its 3D structural encoders learn transferable stereochemical representations that generalise beyond seen compounds. This behaviour is particularly desirable in real-world drug discovery scenarios where predictions are often required for novel compounds lacking prior interaction data.

3.3 Attention is Not All You Need for Explanation

Many existing studies have rushed to claim that highly attended regions are relevant or important in binding [15], [37], [38]. However, prior literature has demonstrated that attention weights do not necessarily correspond to causal importance and can highlight spurious or inhibitory signals unrelated to the model's true decision process [43], [44]. This motivates us to assess faithfulness [45], [46], which refers to the extent to which highlighted areas genuinely influence the model's predictions, rather than merely providing post-hoc explanations. To do so, we perform a perturbation-based interpretability analysis. For each test sample, regions corresponding to highly attended drug atoms or protein residues are selectively masked, and the resulting changes in model outputs are measured. Detailed quantitative results and analysis of our faithfulness evaluation are presented in Appendix F. We also uncover interesting behaviour that is further investigated in Appendix G, where we perform a qualitative comparison of highly attended areas with experimentally critical structures and hypothesise how the model reasons. Overall, we conclude that while attention can reliably identify relevant, decision-critical regions, it should not be treated as a direct map of physical binding sites without additional consideration.

3.4 Ablation study

An ablation study was performed to justify architectural decisions made and is presented in Appendix H.

4 Limitations

The availability of 3D structures and associated computational demands remain long-standing challenges that hinder the proliferation of 3D structural encoders. Although this work demonstrates its feasibility and benefits, it is important to acknowledge the challenges that were encountered during the investigation process. In particular, the reliance on experimentally resolved or reliably predicted 3D structures resulted in the loss of a substantial portion of samples from the DrugBank and KIBA datasets due to missing structures or parsing failures. These limitations in 3D structure availability are expected to diminish with the evolution of protein-folding models. Additionally, precomputing and storing embeddings and full 3D structures introduces additional computational and storage overheads. While this is somewhat offset by the lightweight prediction head of our model (12.9 million parameters compared to MocFormer's 349 million), the practical cumbersomeness of dealing with 3D structures must nevertheless be acknowledged.

5 Conclusion

Drug–target interaction (DTI) prediction remains a crucial yet challenging task in modern day novel drug discovery, where accurate computational methods can significantly reduce experimental cost and time. In this work, we conceive and present 3DICE, an interpretable framework for effective DTI prediction that integrates massively pretrained 3D structural encoders with explicit cross-modal interaction modelling. Leveraging UniMol for drugs and ESM-IF1 for proteins allows us to extract rich structure-aware representations encoding detailed geometric information critical to molecular binding. These representations are further refined through a co-attention mechanism that explicitly models bidirectional cross-modal interactions between drug atoms and protein residues that fundamentally govern binding.

Extensive experiments across multiple canonical benchmarks demonstrate that 3DICE achieves competitive performance compared to state-of-the-art approaches, even in challenging imbalanced-data and cold-start scenarios that better reflect real-world drug-discovery settings. These results indicate that generalisation ability and robustness is indeed enhanced compared to traditional sequence-based methods.

Beyond predictive accuracy, this work places particular emphasis on interpretability. We conducted a novel perturbation-based faithfulness analysis to assess whether learned attention patterns genuinely influence model predictions. We obtain positive results showing that masking highly attended areas generally correlate to a greater decrease in prediction confidence, indicating that attention highlights decision-critical features for positive predictions. However, our key-site investigation shows that similar protein regions are often highlighted across both positive and negative samples, which suggests that attention does not necessarily encode class-specific binding evidence. Instead, it primarily serves only to select where the model focuses. Overall, we conclude that attention can provide meaningful mechanistic cues in some cases (suggesting potential binding sites), but it should be interpreted principally as identifying relevant, decision-critical sites, not as a direct map of physical binding sites or binding-promotive regions. This underscores the importance of combining qualitative attention analysis with metrics such as faithfulness when making claims about physical mechanisms in DTI models, since attention weights alone are insufficient to justify physical binding interpretations.

To conclude, 3DICE demonstrates that the novel coupling of pretrained 3D structural encoders with explicit interaction modelling yields an effective and interpretable approach to DTI prediction. This work paves the way for more efficient and interpretable DTI models, and such approaches open a myriad of possibilities for accelerating and optimising future drug discovery pipelines.

Declarations

Generative Artificial Intelligence

Generative Artificial Intelligence tools were not used in the authoring of this report beyond minor editorial tasks such as spelling and grammar checks. For code, GitHub Copilot and OpenAI Codex were used only to assist with debugging and to automate repetitive tasks.

Funding source

This study was self-funded.

Competing Interests

The author declares no conflict of interest in this publication.

References

[1] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, 'Prediction of drug–target interaction networks from the integration of chemical and genomic spaces', *Bioinformatics*, vol. 24, no. 13, pp. i232–i240, July 2008, doi: 10.1093/bioinformatics/btn162.

[2] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, 'Innovation in the pharmaceutical industry: New estimates of R&D costs', *J. Health Econ.*, vol. 47, pp. 20–33, May 2016, doi: 10.1016/j.jhealeco.2016.01.012.

[3] S. M. Paul *et al.*, 'How to improve R&D productivity: the pharmaceutical industry's grand challenge', *Nat. Rev. Drug Discov.*, vol. 9, no. 3, pp. 203–214, Mar. 2010, doi: 10.1038/nrd3078.

[4] Z. Nikraftar and M. R. Keyvanpour, 'A Comparative Analytical Review on Machine Learning Methods in DrugtargetInteractions Prediction', *Curr. Comput. Aided Drug Des.*, vol. 19, no. 5, pp. 325–355, Oct. 2023, doi: 10.2174/1573409919666230111164340.

[5] A. Vefghi, Z. Rahmati, and M. Akbari, 'Drug-Target Interaction/Affinity Prediction: Deep Learning Models and Advances Review', *Comput. Biol. Med.*, vol. 196, p. 110438, Sept. 2025, doi: 10.1016/j.compbiomed.2025.110438.

[6] X. Ru, L. Xu, W. Han, and Q. Zou, 'In silico methods for drug-target interaction prediction', *Cell Rep. Methods*, vol. 5, no. 10, p. 101184, Oct. 2025, doi: 10.1016/j.crmeth.2025.101184.

[7] A. Mayr *et al.*, 'Large-scale comparison of machine learning methods for drug target prediction on ChEMBL', *Chem. Sci.*, vol. 9, no. 24, pp. 5441–5451, 2018, doi: 10.1039/C8SC00148K.

[8] X. Xia, 'Bioinformatics and Drug Discovery', *Curr. Top. Med. Chem.*, vol. 17, no. 15, pp. 1709–1726, Apr. 2017, doi: 10.2174/1568026617666161116143440.

[9] G. M. Morris *et al.*, 'AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility', *J. Comput. Chem.*, vol. 30, no. 16, pp. 2785–2791, Dec. 2009, doi: 10.1002/jcc.21256.

[10] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath, 'Docking and scoring in virtual screening for drug discovery: methods and applications', *Nat. Rev. Drug Discov.*, vol. 3, no. 11, pp. 935–949, Nov. 2004, doi: 10.1038/nrd1549.

[11] H. Öztürk, A. Özgür, and E. Ozkirimli, 'DeepDTA: deep drug–target binding affinity prediction', *Bioinformatics*, vol. 34, no. 17, pp. i821–i829, Sept. 2018, doi: 10.1093/bioinformatics/bty593.

[12] R. Chen, X. Liu, S. Jin, J. Lin, and J. Liu, 'Machine Learning for Drug-Target Interaction Prediction', *Molecules*, vol. 23, no. 9, p. 2208, Aug. 2018, doi: 10.3390/molecules23092208.

[13] D. Weininger, 'SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules', *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, Feb. 1988, doi: 10.1021/ci00057a005.

[14] W. R. Pearson and D. J. Lipman, 'Improved tools for biological sequence comparison.', *Proc. Natl. Acad. Sci.*, vol. 85, no. 8, pp. 2444–2448, Apr. 1988, doi: 10.1073/pnas.85.8.2444.

[15] K. Huang, C. Xiao, L. M. Glass, and J. Sun, 'MolTrans: Molecular Interaction Transformer for drug–target interaction prediction', *Bioinformatics*, vol. 37, no. 6, pp. 830–836, Mar. 2021, doi: 10.1093/bioinformatics/btaa880.

[16] F. Wan *et al.*, 'DeepCPI: A Deep Learning-Based Framework for Large-Scale *in Silico* Drug Screening', *Genomics Proteomics Bioinformatics*, vol. 17, no. 5, pp. 478–495, Oct. 2019, doi: 10.1016/j.gpb.2019.04.003.

[17] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh, 'GraphDTA: predicting drug–target binding affinity with graph neural networks', *Bioinformatics*, vol. 37, no. 8, pp. 1140–1147, May 2021, doi: 10.1093/bioinformatics/btaa921.

[18] Z. Quan, Y. Guo, X. Lin, Z.-J. Wang, and X. Zeng, 'GraphCPI: Graph Neural Representation Learning for Compound-Protein Interaction', in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA: IEEE, Nov. 2019, pp. 717–722. doi: 10.1109/BIBM47256.2019.8983267.

[19] L. Chen *et al.*, 'TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments', *Bioinformatics*, vol. 36, no. 16, pp. 4406–4414, Aug. 2020, doi: 10.1093/bioinformatics/btaa524.

[20] M. Gao *et al.*, 'GraphormerDTI: A graph transformer-based approach for drug-target interaction prediction', *Comput. Biol. Med.*, vol. 173, p. 108339, May 2024, doi: 10.1016/j.compbiomed.2024.108339.

[21] Y. Zhao *et al.*, 'Evidential deep learning-based drug-target interaction prediction', *Nat. Commun.*, vol. 16, no. 1, p. 6915, July 2025, doi: 10.1038/s41467-025-62235-6.

[22] G. Liu *et al.*, 'GraphDTI: A robust deep learning predictor of drug-target interactions from multiple heterogeneous data', *J. Cheminformatics*, vol. 13, no. 1, p. 58, Aug. 2021, doi: 10.1186/s13321-021-00540-0.

[23] Y.-L. Zhang, W.-T. Wang, J.-H. Guan, D. K. Jain, T.-Y. Wang, and S. K. Roy, 'MocFormer: A Two-Stage Pre-training-Driven Transformer for Drug–Target Interactions Prediction', *Int. J. Comput. Intell. Syst.*, vol. 17, no. 1, p. 165, June 2024, doi: 10.1007/s44196-024-00561-1.

[24] L. Zhao, H. Wang, and S. Shi, 'PocketDTA: an advanced multimodal architecture for enhanced prediction of drug−target affinity from 3D structural data of target binding pockets', *Bioinformatics*, vol. 40, no. 10, p. btae594, Oct. 2024, doi: 10.1093/bioinformatics/btae594.

[25] Y. Sun, Y. Y. Li, C. K. Leung, and P. Hu, 'iNGNN-DTI: prediction of drug–target interaction with interpretable nested graph neural network and pretrained molecule models', *Bioinformatics*, vol. 40, no. 3, p. btae135, Mar. 2024, doi: 10.1093/bioinformatics/btae135.

[26] J. Lee, D. W. Jun, I. Song, and Y. Kim, 'DLM-DTI: a dual language model for the prediction of drug-target interaction with hint-based learning', *J. Cheminformatics*, vol. 16, no. 1, p. 14, Feb. 2024, doi: 10.1186/s13321-024-00808-1.

[27] S. Chithrananda, G. Grand, and B. Ramsundar, 'ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction', 2020, *arXiv*. doi: 10.48550/ARXIV.2010.09885.

[28] X.-C. Zhang *et al.*, 'MG-BERT: leveraging unsupervised atomic representation learning for molecular property prediction', *Brief. Bioinform.*, vol. 22, no. 6, p. bbab152, Nov. 2021, doi: 10.1093/bib/bbab152.

[29] A. Elnaggar *et al.*, 'ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning', *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 7112–7127, Oct. 2022, doi: 10.1109/TPAMI.2021.3095381.

[30] Z. Lin *et al.*, 'Evolutionary-scale prediction of atomic-level protein structure with a language model', *Science*, vol. 379, no. 6637, pp. 1123–1130, Mar. 2023, doi: 10.1126/science.ade2574.

[31] G. Zhou *et al.*, 'Uni-Mol: A Universal 3D Molecular Representation Learning Framework', May 2022, doi: doi:10.26434/chemrxiv-2022-jjm0j.

[32] C. Hsu *et al.*, 'Learning inverse folding from millions of predicted structures', Apr. 10, 2022, *Systems Biology*. doi: 10.1101/2022.04.10.487779.

[33] J. Jumper *et al.*, 'Highly accurate protein structure prediction with AlphaFold', *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/s41586-021-03819-2.

[34] P. Bai, F. Miljković, B. John, and H. Lu, 'Interpretable bilinear attention network with domain adaptation improves drug-target prediction', Jan. 19, 2023, *arXiv*: arXiv:2208.02194. doi: 10.48550/arXiv.2208.02194.

[35] C.-F. Chen, Q. Fan, and R. Panda, 'CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification', Aug. 22, 2021, *arXiv*: arXiv:2103.14899. doi: 10.48550/arXiv.2103.14899.

[36] H. Lin *et al.*, 'CAT: Cross Attention in Vision Transformer', June 10, 2021, *arXiv*: arXiv:2106.05786. doi: 10.48550/arXiv.2106.05786.

[37] H. Kurata and S. Tsukiyama, 'ICAN: Interpretable cross-attention network for identifying drug and target protein interactions', *PLOS ONE*, vol. 17, no. 10, p. e0276609, Oct. 2022, doi: 10.1371/journal.pone.0276609.

[38] X. Zeng, W. Chen, and B. Lei, 'CAT-DTI: cross-attention and Transformer network with domain adaptation for drug-target interaction prediction', *BMC Bioinformatics*, vol. 25, no. 1, p. 141, Apr. 2024, doi: 10.1186/s12859-024-05753-2.

[39] K. Che, Q. Ning, R. Li, X. Wei, H. Li, and S. Guo, 'Cross Attention and Intra-layer Attention in Heterogeneous Graph Neural Networks for Drug-Target Interaction Prediction', *IEEE Trans. Comput. Biol. Bioinforma.*, pp. 1–12, 2025, doi: 10.1109/TCBBIO.2025.3578584.

[40] J. Lu, J. Yang, D. Batra, and D. Parikh, 'Hierarchical Question-Image Co-Attention for Visual Question Answering', 2016, *arXiv*. doi: 10.48550/ARXIV.1606.00061.

[41] D. S. Wishart *et al.*, 'DrugBank 5.0: a major update to the DrugBank database for 2018', *Nucleic Acids Res.*, vol. 46, no. D1, pp. D1074–D1082, Jan. 2018, doi: 10.1093/nar/gkx1037.

[42] J. Tang *et al.*, 'Making Sense of Large-Scale Kinase Inhibitor Bioactivity Data Sets: A Comparative and Integrative Analysis', *J. Chem. Inf. Model.*, vol. 54, no. 3, pp. 735–743, Mar. 2014, doi: 10.1021/ci400709d.

[43] S. Serrano and N. A. Smith, 'Is Attention Interpretable?', June 09, 2019, *arXiv*. arXiv:1906.03731. doi: 10.48550/arXiv.1906.03731.

[44] S. Jain and B. C. Wallace, 'Attention is not Explanation', 2019, *arXiv*. doi: 10.48550/ARXIV.1902.10186.

[45] V. Petsiuk, A. Das, and K. Saenko, 'RISE: Randomized Input Sampling for Explanation of Black-box Models', Sept. 25, 2018, *arXiv*. arXiv:1806.07421. doi: 10.48550/arXiv.1806.07421.

[46] Y. Liu, H. Li, Y. Guo, C. Kong, J. Li, and S. Wang, 'Rethinking Attention-Model Explainability through Faithfulness Violation Test', 2022, *arXiv*. doi: 10.48550/ARXIV.2201.12114.

[47] H. M. Berman, 'The Protein Data Bank', *Nucleic Acids Res.*, vol. 28, no. 1, pp. 235–242, Jan. 2000, doi: 10.1093/nar/28.1.235.

[48] J. Fleming *et al.*, 'AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities', *J. Mol. Biol.*, vol. 437, no. 15, p. 168967, Aug. 2025, doi: 10.1016/j.jmb.2025.168967.

[49] N. Schneider, R. A. Sayle, and G. A. Landrum, 'Get Your Atoms in Order—An Open-Source Implementation of a Novel and Robust Molecular Canonicalization Algorithm', *J. Chem. Inf. Model.*, vol. 55, no. 10, pp. 2111–2120, Oct. 2015, doi: 10.1021/acs.jcim.5b00543.

[50] T. A. Halgren, 'Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94', *J. Comput. Chem.*, vol. 17, no. 5–6, pp. 490–519, Apr. 1996, doi: 10.1002/(SICI)1096-987X(199604)17:5/6%3C490::AID-JCC1%3E3.0.CO;2-P.

[51] M. Wen *et al.*, 'Deep-Learning-Based Drug–Target Interaction Prediction', *J. Proteome Res.*, vol. 16, no. 4, pp. 1401–1409, Apr. 2017, doi: 10.1021/acs.jproteome.6b00618.

[52] H. A. AL-Ghulikah, S. A. El-Sebaey, A. K. A. Bass, and M. S. El-Zoghbi, 'New Pyrimidine-5-Carbonitriles as COX-2 Inhibitors: Design, Synthesis, Anticancer Screening, Molecular Docking, and In Silico ADME Profile Studies', *Molecules*, vol. 27, no. 21, p. 7485, Jan. 2022, doi: 10.3390/molecules27217485.

[53] P. K. Deb, R. P. Mailabaram, B. Al-Jaidi, and M. J. Saadh, 'Molecular Basis of Binding Interactions of NSAIDs and Computer-Aided Drug Design Approaches in the Pursuit of the Development of Cyclooxygenase-2 (COX-2) Selective Inhibitors', in *Nonsteroidal Anti-Inflammatory Drugs*, A. G. A. Al-kaf, Ed., InTech, 2017. doi: 10.5772/intechopen.68318.

[54] S. W. Rowlinson *et al.*, 'A Novel Mechanism of Cyclooxygenase-2 Inhibition Involving Interactions with Ser-530 and Tyr-385', *J. Biol. Chem.*, vol. 278, no. 46, pp. 45763–45769, Nov. 2003, doi: 10.1074/jbc.M305481200.

Appendix A: Architectural Details

Protein Feature Encoder

The frozen pretrained model Evolutionary Scale Modelling – Inverse Folding 1 (ESM-IF1) was used which produced a $L_p \times 512$ feature representation.

$$\boldsymbol{X} \in \mathbb{R}^{L_p \times 512}$$

We apply a 2-layer 1D convolutional neural network (CNN) across the sequence dimension to capture local substructural patterns.

$$\mathrm{Block}(\boldsymbol{A}) = \boldsymbol{A} + \mathrm{Dropout}\Big(\mathrm{ReLU}\big(\mathrm{Conv1D}(\boldsymbol{A})\big)\Big)$$

$$\mathrm{CNN}(\boldsymbol{A}) = \mathrm{Block}_2\big(\mathrm{Block}_1(\boldsymbol{A})\big)$$

$$\boldsymbol{X}^{(1)} = \mathrm{CNN}(\boldsymbol{X})$$

Drug Feature Encoder

The frozen pretrained model UniMol was used which produced a $L_d \times 512$ feature representation.

$$\boldsymbol{Y} \in \mathbb{R}^{L_d \times 512}$$

A similar CNN was applied:

$$\boldsymbol{Y}^{(1)} = \mathrm{CNN}(\boldsymbol{Y})$$

Co-attention Module

A co-attention module was implemented to explicitly model interactions between proteins and drugs via bidirectional cross-attention. Each cross-attention block once again follows a pre-norm Transformer-style architecture with residual connections, dropout, and a position-wise feed-forward network.

Multi-head attention was implemented for this component. For each head $i \in \{1, \dots, H\}$ we first form:

$$\mathrm{Head}_i(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i) = \mathrm{Softmax}\left(\frac{(\boldsymbol{Q}_i \boldsymbol{W}_i^Q)(\boldsymbol{K}\boldsymbol{W}_i^K)^\top}{\sqrt{\frac{d}{H}}}\right)(\boldsymbol{V}\boldsymbol{W}_i^V)$$

We then define the MultiHeadAttention function:

$$\mathrm{MultiHeadAttention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \mathrm{concat}\big(\mathrm{Head}_i(\boldsymbol{Q}_i, \boldsymbol{K}_i, \boldsymbol{V}_i), \dots, \mathrm{Head}_{\mathrm{H}}(\boldsymbol{Q}_H, \boldsymbol{K}_H, \boldsymbol{V}_H)\big) \cdot \boldsymbol{W}_0$$

Next, the feed-forward network:

$$\mathrm{FFN}(\boldsymbol{A}) = \mathrm{ReLU}(\boldsymbol{A}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2$$

On the protein side:

$$\boldsymbol{O}_p = \boldsymbol{X}^{(1)} + \mathrm{Dropout}\left(\mathrm{MultiHeadAttention}\left(\mathrm{LN}_p^{(1)}\left(\boldsymbol{X}^{(1)}\right), \boldsymbol{Y}^{(1)}, \boldsymbol{Y}^{(1)}\right)\right)$$

$$\boldsymbol{O}_p^{(1)} = \boldsymbol{O}_p + \mathrm{Dropout}\left(\mathrm{FFN}_p\left(\mathrm{LN}_p^{(2)}\left(\boldsymbol{O}_p\right)\right)\right)$$

And similarly for drugs:

$$\boldsymbol{O}_d = \boldsymbol{Y}^{(1)} + \mathrm{Dropout}\left(\mathrm{MultiHeadAttention}\left(\mathrm{LN}_d^{(1)}\left(\boldsymbol{Y}^{(1)}\right), \boldsymbol{X}^{(1)}, \boldsymbol{X}^{(1)}\right)\right)$$

$$\boldsymbol{O}_d^{(1)} = \boldsymbol{O}_d + \mathrm{Dropout}\left(\mathrm{FFN}_d\left(\mathrm{LN}_d^{(2)}\left(\boldsymbol{O}_d\right)\right)\right)$$

$\boldsymbol{O}_p^{(1)}$ and $\boldsymbol{O}_d^{(1)}$ represent attended protein and drug features. Padding masks are applied during attention to account for different lengths.

Fusion Module

Mean pooling was applied to the attended protein and drug representation matrices to get a fixed-length representation vector. Mean pooling was applied, then were each passed through a fully connected layer that reduced their dimensions from 512 to 256:

$$\boldsymbol{P} = \mathrm{MeanPool}\left(\boldsymbol{O}_p^{(1)}\boldsymbol{W}_p + \boldsymbol{b}_p\right), \boldsymbol{D} = \mathrm{MeanPool}\left(\boldsymbol{O}_d^{(1)}\boldsymbol{W}_d + \boldsymbol{b}_d\right)$$

where $\boldsymbol{W}_p, \boldsymbol{W}_d \in \mathbb{R}^{256\times512}$. Finally, a concatenation forms the final joint representation vector.

$$\boldsymbol{Z} = [\boldsymbol{P} \parallel \boldsymbol{D}] \in \mathbb{R}^{512}$$

Readout MLP

The fused representation is forwarded to a fully connected multilayer perceptron (MLP) that outputs two logits corresponding to the positive and negative interaction. The larger of the two logits was taken to be the predicted class.

$$\boldsymbol{R} = \mathrm{MLP}(\boldsymbol{Z}) \in \mathbb{R}^2$$

Appendix B: Datasets

Details

The DrugBank, KIBA and Davis datasets provide drug and protein identification codes and canonical string representations (SMILES [13] and FASTA [14]), which are customarily used in literature. However, in this study, 3D structural information is required to generate the embeddings used. Protein UniProt IDs were cross-referenced with the Protein Data Bank (PDB) [47]. Experimentally obtained structures with the highest fidelity selected based on resolution were identified and fetched in mmCIF format. When experimental structures were unavailable, structures predicted by AlphaFold [33] were used. These were obtained from AlphaFoldDB [48]. Drug 3D conformers were generated from SMILES strings using RDKit's ETKDG algorithm followed by MMFF94 force-field optimisation [49], [50]. The resultant molecular coordinates were used for downstream tasks.

When SMILES strings could not be parsed or if 3D protein information was unavailable from both PDB and AlphaFoldDB, the affected drug–protein pairs were removed. 3DICE itself does not face character length, size or structural complexity limits. However, the baseline models used for benchmarking may have such constraints. To ensure a fair comparison, the datasets were further pruned down to the smallest common subset that is supported by all the models. This involved removing drugs with inorganic compounds or very small molecule compounds, and the maximum length of drugs and proteins was limited to 200 and 700 respectively. 8:1:1 splits were performed for the train, validation, and test datasets. We additionally constructed a cold-start scenario to assess the model's ability to predict novel DTIs in realistic drug discovery settings. 10% of drugs from the DrugBank dataset were randomly selected, and all DTI pairs associated with these drugs were used as the test set.

DrugBank

DrugBank 5.1.13 [41], published on January 2, 2025, was used in this study. In total, 13,311 pairs of known DTIs were collected, comprising 5,088 drugs and 3,251 targets. These curated interactions were treated as positive samples. The negative set was generated through random disruption of the positive set as is conventional in literature [15], [21], [23], [51]. Specifically, for each positive DTI pair, we selected the target and reassigned it to a different drug that is known to interact with another target, thereby forming a synthetic non-interacting drug–target pair. Negative samples were produced in a 1:1 ratio with the positive set, yielding 26,622 total pairs. In the cold-start case, there were 2,673 test samples and 23,717 training samples.

KIBA

The KIBA [42] dataset tabulates experimental values for drug–target binding affinities, represented by the KIBA score. The threshold of 12.1 was used to curate a binary classification dataset from continuous affinity values [11], [21]. 12,765 positive and 55,306 negative pairs were collected, comprising 133 drugs and 2,106 targets.

Appendix C: Benchmarks

MolTrans

MolTrans [15] operates on SMILES strings of drugs and amino acid sequences of proteins directly with a transformer model to encode feature embeddings. A Hadamard-product-based interaction matrix is constructed, and predictions are subsequently made by CNNs and FCNs.

DLM-DTI

DLM-DTI [26] is a dual language model for the prediction of DTIs with hint-based learning. It comprises of the drug encoder, the target encoder, and the interaction prediction head. In the teacher model for target encoding, knowledge is distilled from the ProtBERT transformer. In the drug encoder, molecular representation vectors are obtained from SMILES sequences by employing the ChemBERTa encoder. Joint representation is achieved through concatenation and the prediction is made by a Fully Connected Layer.

MocFormer

MocFormer [23] is a proposed two-stage pretraining-driven transformer. UniMol and ESM-2 are used to encode drug and protein features respectively. The representations are passed through multiple multi-head self-attention stacks for further feature extraction. Bilinear pooling is performed to model cross interactions, obtaining a interaction map along with a joint representation forwarded through a readout Fully Connected Layer.

Appendix D: Performance Metrics

Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives respectively. The following metrics were used and labelled as follows:

Accuracy (ACC) measures the overall proportion of correctly classified samples:

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}}$$

Precision (PPV) measures the proportion of predicted positive samples that are truly positive:

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall (TPR) measures the proportion of true positive samples that are correctly identified:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

F1-score is the harmonic mean of precision and recall:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}}$$

Matthews Correlation Coefficient (MCC) provides a balanced measure of classification quality that accounts for all four confusion matrix terms:

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

Area Under Receiver Operating Curve (AUROC) measures the model's ability to discriminate between positive and negative samples across varying decision thresholds.

Area Under Precision-Recall Curve (AUPRC) measures the trade-off between precision and recall across thresholds.

Appendix E: Hyperparameters

Parameter	Value
Epoch	60
Batch	32
Optimiser	Adam
Learning rate	5e-5
Weight decay	0
Dropout	0.3
Activation Function	ReLU
CNN Kernel Size	3
CNN Layer Count	2
CNN Dimensions	[512, 512]
Attention head number	4
Feed-forward Dimension	2048
Pooling method	Mean Pooling
Reduction MLP Dimension	256
Readout MLP Dimensions	[1024, 1024, 512, 2]

Appendix F: Faithfulness Test

In our test, we will examine how the behaviour of the model changes as portions of the input drug or protein are masked. We will be comparing selective deletion, where positions in the drug and protein representations are masked in descending priority based on attention weights, and random deletion.

To measure the change in model output, we define the margin between the positive and negative logit, m.

$$m = z_+ - z_-$$

An increase in m corresponds to a lean toward the positive class, and vice versa. The selective and random deletion cases will be compared to the baseline altered model, which motivates us to define the change in margin:

$$\Delta m = m' - m_0 = (z'_+ - z'_-) - (z_{+_0} - z_{-_0})$$

A positive Δm means the model gravitated towards a positive prediction for all samples, and vice versa for a negative Δm.

In our experiment, we investigate the change in the logit margin Δm as the fraction of the input masked is increased. Below we present the results of our experiment, performed with 512 positive and negative samples each.

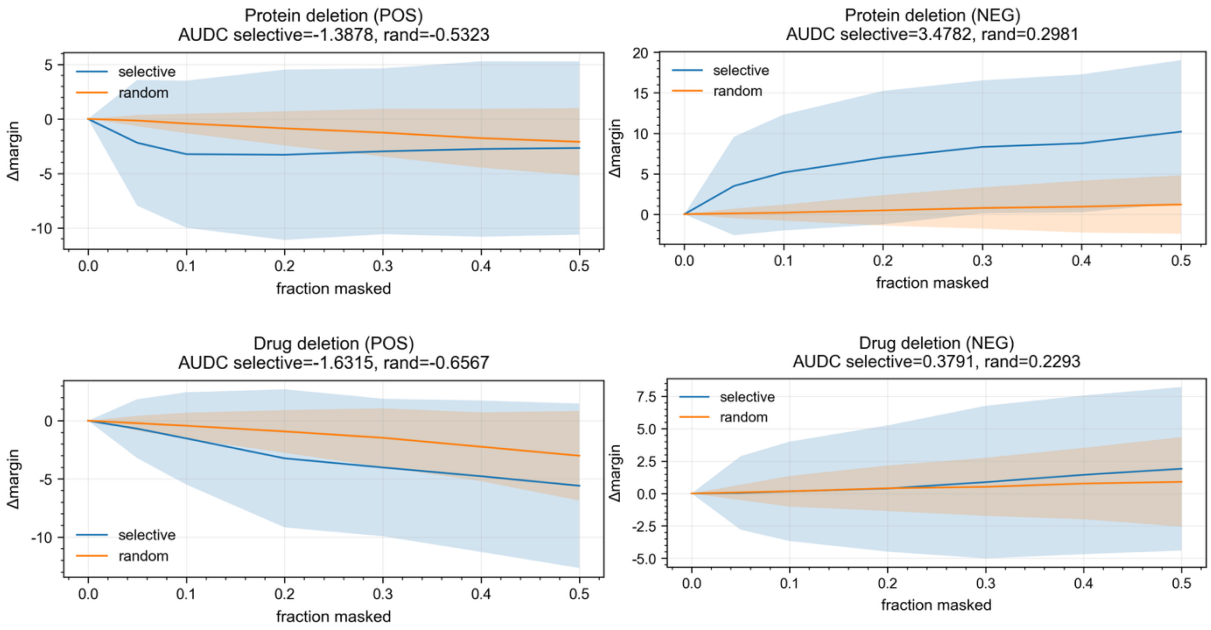

Figure 2: Deletion Curves from Faithfulness Test

In Figure 2, we plot Δm against the masked fraction, which we term the **Deletion Curve**. The Area Under the Deletion Curve (AUDC) is listed for selective and random deletion.

In positive samples, for both protein and drug deletion, as fraction masked increases, Δm becomes more negative, with it decreasing faster in the selective case than the random case, hence yielding a more negative AUDC. This is rather expected as masking positive binding signals highlighted by attention leads to a decreased confidence in predictions, shifting them towards the negative class. In the protein case, it is interesting to note that the selective curve decreases greatly initially, then slowly converges with the random curve as the masking fraction increases. This may be explained by the observation that only a small number of resides carry high attention weight and predictive influence, as further illustrated in Appendix G (Key Site Investigation).

For negative samples though, we observe more nuanced behaviour. Under drug deletion, we observe little change in the logit margin for both random and selective deletion, yielding similar slightly positive AUDCs. A conservative interpretation is that drug-side features contribute minor discriminative inhibitory evidence for negative predictions in this setting.

In contrast, when proteins are masked, we observe that as fraction masked increases, Δm in the selective case increases more rapidly than in the random case, yielding a more positive AUDC. This indicates that selectively removing residues emphasised by the model also reduces its confidence in the negative class. It is tempting to infer that these residues somehow convey negative binding information, but such a conclusion cannot be drawn from deletion curves alone, and requires examining whether the same sites are highlighted across labels and how their contribution is determined downstream. We therefore explore the mechanism underlying this behaviour in the following section.

Appendix G: Key Site Investigation

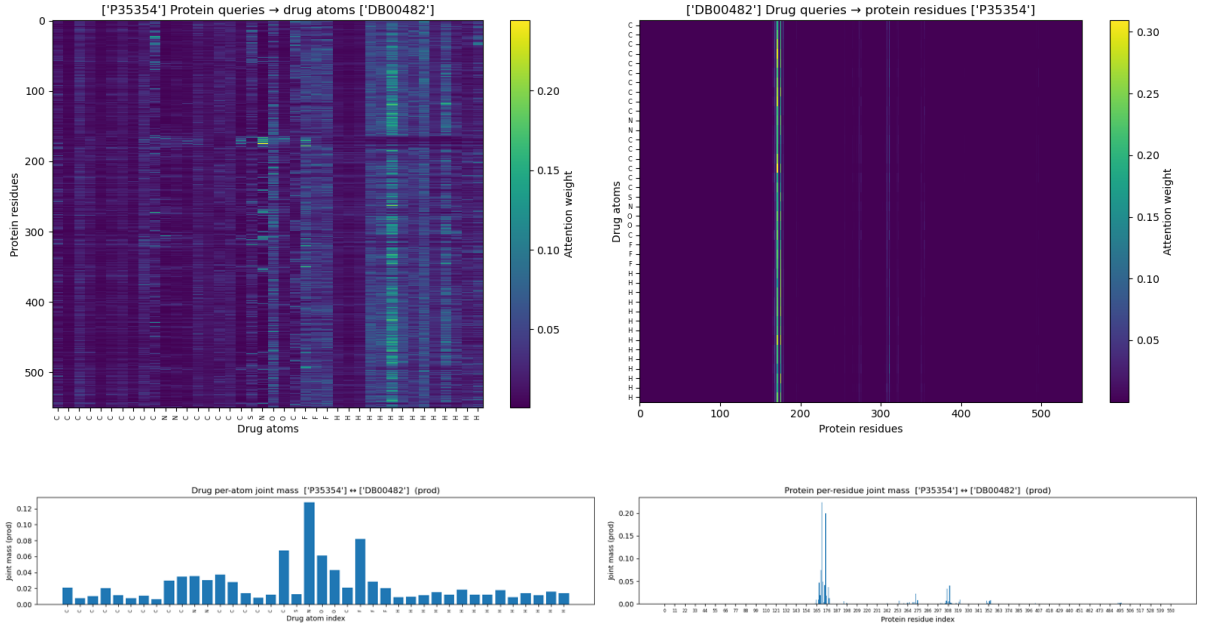

Figure 3: Attention weights in positive sample

We shall investigate a well-known DTI case: the interaction between the non-steroidal anti-inflammatory drug (NSAID) **celecoxib** (DrugBank ID: DB00482) and prostaglandin-endoperoxide synthase 2, commonly referred to as **PTGS2** or **COX-2** (UniProt ID: P35354), where celecoxib operates as a cyclooxygenase inhibitor. We feed this pair into the model and examine the two cross-attention weight matrices produced by the co-attention module: the protein-drug matrix where each residue attends over atoms, and the drug-protein matrix where each atom attends over residues.

We observe that protein residue weights are concentrated in highly specific locations, while drug weights are comparatively more spread out across the molecular map. Notably, on the protein-drug map, when examining the rows corresponding to the protein residues strongly highlighted by the drug-protein map, the attention is more localised. This indicates to a greater degree that those specific atoms are extra important because they exhibit low promiscuity with the non-attended protein residues. For example, N20 is sharply conditioned to a certain residue, while an atom such as H32 appears globally salient but lacks specificity. This asymmetry where the model's focus on protein residues is much more specific compared to that on drug atoms is interesting and we shall keep it in mind for the subsequent analysis.

For easy comparison, we compute a joint cross-attention mass by taking the elementwise product of the two cross-attention matrices then summing along each drug row or protein column to obtain

a global importance score for each atom or residue. This joint mass allows us to favour mutual coupling which provides more physical insight into the detailed interactions.

Note that the structural residue indices produced by the model do not necessarily correspond to the canonical FASTA indices used in literature, since experimentally resolved structures may omit flexible segments or loops. Therefore, the below analysis is performed after aligning the resolved residue sequence to the FASTA sequence and remapping each structural index to its FASTA index for proper comparison.

Comparing to literature, we observe the largest peak in residue attention mass at around structural index 175, which corresponds to the vicinity of Gln192. On the drug side, N20 corresponds to the nitrogen on celecoxib's sulfonamide group. This structure is repeatedly reported to engage the polar pocket region that includes Gln192, consistent with both crystallographic analyses and multiple independent redocking studies [52], [53], [54]. This supports the view that the co-attention module recovered a **chemically meaningful** anchoring motif for this interaction. However, other widely discussed residues such as Leu352 and Arg513 [52], [53] are not among the top peaks in our joint-mass scoring. We also observe weaker peaks around FASTA indices 257, 292, and 335, which are not classic celecoxib pocket residues. Importantly, when we inspect the rows corresponding to these residues, no single drug atom is strongly singled out, suggesting that these other residues may be encoding global protein features rather than localised chemical contacts. To learn more, we repeat the analysis on a negative sample, with results presented in Figure 4.

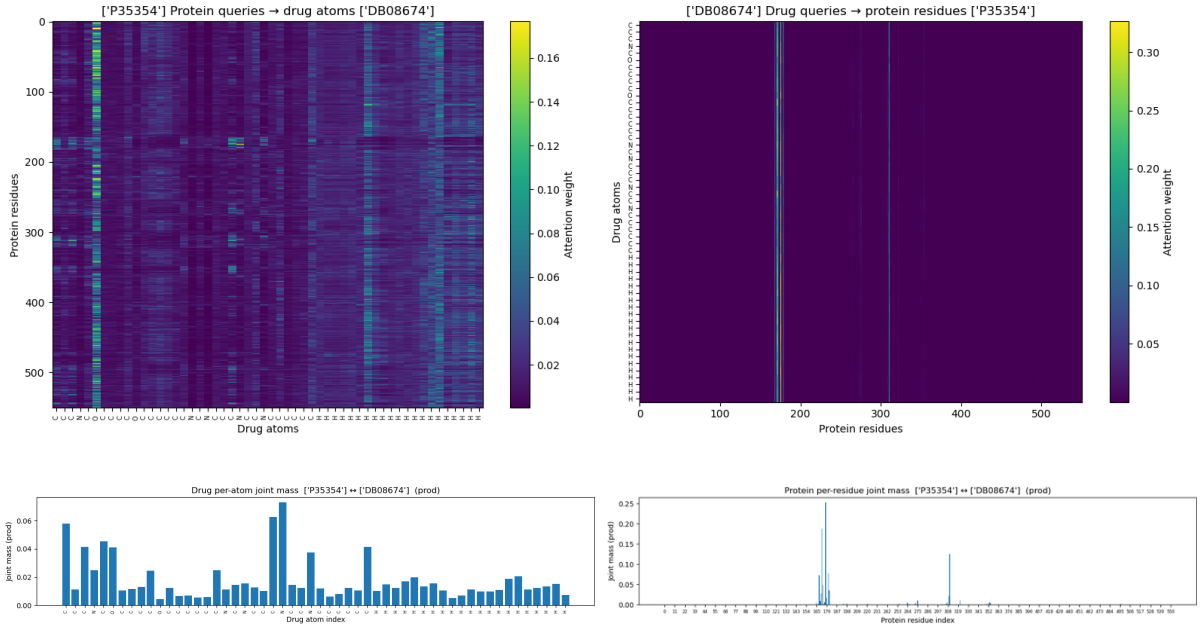

Figure 4: Attention weights in negative sample

Interestingly, similar residue regions on the protein are highlighted. Together with the previously highlighted difference between the protein and drug maps, this suggests a potential mechanism of how the model discriminates: it first selects critical sites on the protein, then considers if chemical interactions may occur downstream given a more generalised drug context. This also provides a possible explanation for the faithfulness test; discriminative ideas appear to be performed primarily from protein-side evidence, which is why masking protein residues causes a much larger drop in confidence than masking drug atoms, especially in the negative case.

We now return to Appendix F, where we previously discussed the negative protein deletion curve. Under this new hypothesis of how the model reasons, we can conclude that that protein sites **do not intrinsically encode** specific regions with positive or negative effects on binding. Attention only serves to select where the model focuses on, while the downstream classifier determines how the information from those sites contributes to the decision.

Nevertheless, the results of the faithfulness test do indicate that masking these residues degrades performance, indicating that they carry predictive signal used by the model. Thus, in this setting, it is important to note that attention appears to prioritise residues that are **important for the decision** (which may include true binding residues, structural determinants, or broader signatures), rather than residues that can be unambiguously interpreted as sites with positive binding characteristics.

Finally, these results motivate caution. Future work may want to explore the physical importance and characteristics of such residues. ICAN has also reported that on average only 2 of the top 30 ranked highlighted sites correspond to experimentally verified binding residues on average [37]. Our findings are consistent with this: attention can yield genuine mechanistic cues (Gln192's interaction with sulfonamide) while also highlighting residues that are predictive for reasons other than direct contact (proven through faithfulness). Thus, while attention provides useful hypotheses and qualitative physical insight, works should avoid decisive claims that highly attended regions correspond to actual binding sites without additional structural or experimental validation, and post-hoc explanations of binding mechanisms based on attention weights only must be taken with a grain of salt.

Appendix H: Ablation Study

In our ablation study, we evaluate the effectiveness of the CNN and co-attention (CA) components. For the CNN variants, the CNN encoder was replaced with a self-attention (SA) module comparable to those used in state-of-the-art transformer-based DTI models.

Table 4: Ablation study on the DrugBank dataset

DrugBank	ACC	PPV	TPR	F1	MCC	AUROC	AUPRC
3DICE	0.81675	0.80544	0.83569	0.82022	0.63394	0.88903	0.87353
w/o CA	(0.00370)	(0.01086)	(0.00615)	(0.00279)	(0.00941)	(0.00249)	(0.00201)
3DICE	0.82689	0.81352	0.84833	0.83066	0.65442	0.89563	0.87650
Drug SA	(0.00414)	(0.01093)	(0.01206)	(0.00302)	(0.00788)	(0.00081)	(0.00572)
3DICE	0.83072	0.81690	**0.85390**	0.83446	0.66306	0.90426	0.89361
Prot SA	(0.00358)	(0.01718)	**(0.03092)**	(0.00702)	(0.00760)	(0.00580)	(0.00623)
3DICE	**0.84379**	**0.85058**	0.83427	**0.84235**	**0.68770**	**0.91065**	**0.90200**
(base)	**(0.00360)**	**(0.00539)**	(0.00497)	**(0.00356)**	**(0.00722)**	**(0.00216)**	**(0.00145)**

The best results are bolded, while the second-best results are underlined.
Five independent replications were performed (n = 5). Data is expressed as mean (std).

Because 3DICE (Prot SA) achieved performance closest to 3DICE (base), we conducted additional experiments on the remaining datasets to more thoroughly compare their performance.

Table 5: Ablation study on the KIBA dataset

KIBA	ACC	PPV	TPR	F1	MCC	AUROC	AUPRC
3DICE	0.89739	**0.77767**	**0.66038**	**0.71268**	0.65440	**0.92780**	**0.80590**
Prot SA	(0.00542)	**(0.01936)**	**(0.03926)**	**(0.01460)**	(0.00883)	**(0.00124)**	**(0.01017)**
3DICE	**0.90012**	0.77473	0.65933	0.71233	**0.65553**	0.92744	0.79357
(base)	**(0.00082)**	(0.00864)	(0.00797)	(0.00162)	**(0.00167)**	(0.00130)	(0.00051)

The best results are bolded, while the second-best results are underlined.
Five independent replications were performed (n = 5). Data is expressed as mean (std).

Table 6: Ablation study on the DrugBank dataset in the cold-start case

Cold start	ACC	PPV	TPR	F1	MCC	AUROC	AUPRC
3DICE	**0.79835**	**0.80991**	0.77714	0.79274	**0.59760**	0.86301	**0.85665**
Prot SA	**(0.00212)**	**(0.01775)**	(0.03302)	(0.00870)	**(0.00297)**	(0.00184)	**(0.00035)**
3DICE	0.79611	0.79637	**0.79278**	**0.79423**	0.59270	**0.86548**	0.85020
(base)	(0.00429)	(0.01001)	**(0.02721)**	**(0.00875)**	(0.00903)	**(0.00654)**	(0.00545)

The best results are bolded, while the second-best results are underlined.
Five independent replications were performed (n = 5). Data is expressed as mean (std).

Although 3DICE (Prot SA) may generalise better in some cases, 3DICE (base) performed better and more consistently overall and was therefore chosen as the presented model.