Deep Learning Approaches to improve Drug-Target Binding Affinity Prediction in Precision Oncology

Supervisor: Dr. Sushant Kumar
Princess Margaret Cancer Centre, UHN
sushant.kumar@uhn.ca

Mohammed Balkhair
King Abdulaziz University
MohammedBalkhair02@gmail.com

Aug 23, 2025

Abstract

This work presents a comprehensive benchmarking study of graph neural network (GNN) architectures for drug-target binding affinity (DTA) prediction, with a focus on integrating structural and language-based protein and ligand representations. We address previous methodological limitations—particularly unfair dataset splits and limited structural coverage—by constructing a fair evaluation protocol across three standard DTA datasets. In addition, we explore the impact of architectural enhancements such as graph attention and cross-branch fusion, and we benchmark the generative structure prediction model Boltz-2 for biological perturbation studies. Through mutation and truncation experiments on EGFR, we evaluate Boltz-2's capacity to detect functional regions and inform drug sensitivity.

1 Project Objectives and Goals

The primary objective of this project is to improve the robustness and fairness of deep learning models in drug-target binding affinity prediction by integrating protein structural dynamics and chemical language representations. Specifically, the study aims to (1) develop a fair evaluation benchmark across models using consistent data subsets, (2) incorporate chemical and protein language model embeddings to enrich graph inputs, (3) explore the use of attention-based architectures to model interactions more effectively, and (4) benchmark the utility of Boltz-2 in structural biology-inspired affinity analysis.

2 Introduction

The accurate prediction of drug-target binding affinity is a critical task in computational drug discovery and precision oncology [2,3]. Graph neural networks (GNNs) allow the encoding of molecular graphs for both proteins and small molecules, providing structural and relational context. However, prior studies often suffer from flawed benchmarking setups, inconsistent data coverage, or inadequate architectural ablation.

To address these issues, we build on work from the CCG Lab [1] to implement a fair evaluation framework that standardizes data splits, allows for structural and language-based node features, and systematically benchmarks multiple GNN variants. Additionally, we evaluate Boltz-2 [8], a generative structure prediction model, for studying functional regions through mutations and truncations.

3 Materials and Methods

3.1 Datasets

We utilize the Davis [5], KIBA [6], and PDBbind [7] datasets. Each provides protein-drug affinity data.

Dataset	Protein	Compound	Total Binding Entities
Davis	442	68	30,056
Kiba	229	2,111	118,254
PDBbind	3,889	12,639	19,443

Figure 1: Dataset statistics: number of unique proteins, compounds, and interactions across Davis, KIBA, and PDBbind.

3.2 Graph Construction

Drug molecules are encoded as graphs using molecule raw data or SELFIES tokenization with ChemLLM embeddings. [4]. Protein graphs are built with different kind of tools like AlphaFold2-derived contact maps or embeddings from protein language models (PLMs) to leverage node feature and edge weights.

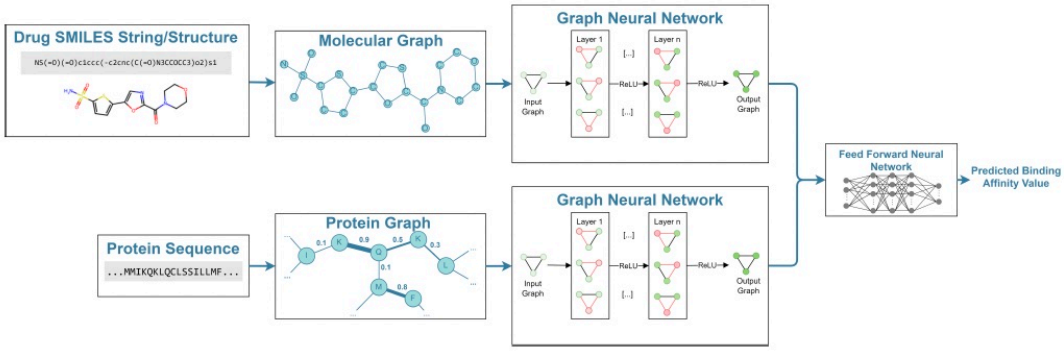


Figure 2: Graph construction for protein and drug branches.

3.3 Architectures Evaluated

We benchmark the following models: GCN (Graph Convolutional Network) with static features, GAT (Graph Attention Network) with learned neighbor weighting, a fusion model combining ChemLLM and PLM embeddings, and a cross-attention architecture linking protein and drug branches.

3.4 Training Strategy

To ensure fair evaluation, we filtered all datasets to include only proteins with AlphaFold2 structural coverage. This ensures that both structure-aware and language-model-based variants are trained on identical subsets.

4 Preliminary Results or Findings

Initial experiments indicate that structure-aware GNNs perform well when trained on properly curated datasets. However, when enhanced with graph attention and fusion layers, models using chemical and protein language model embeddings are able to outperform traditional structural models. The Boltz-2 truncation experiments on EGFR successfully identify regions critical to drug binding, confirming the model's utility in studying structural perturbations. However, absolute binding affinity predictions from Boltz-2 remain less reliable than relative ranking among drug candidates.

5 Challenges and Obstacles Encountered

Several challenges emerged throughout this project. First, inconsistencies in dataset preprocessing and missing structural coverage created bias in earlier benchmarks. Identifying and correcting these issues required detailed dataset inspection and filtering. Training deep learning models on large datasets with multiple architectures proved computationally intensive, demanding multi-day GPU usage. Furthermore, integrating embeddings from language models into GNN architectures required custom tokenizer support and model freezing strategies to balance performance and complexity.

6 Results

Structure-aware GCN models consistently outperformed their LLM-only counterparts when trained with fair splits. However, introducing GAT layers and fusion embeddings allowed language-model-based approaches to match or exceed structure-aware models, particularly in hybrid cross-attention configurations.

6.1 Boltz-2 Mutation and Truncation Study

We used Boltz-2 to simulate truncation experiments on EGFR, progressively removing 20 residues at a time. Figure 3 shows a notable drop in predicted binding probability between residues 790–810 for several ligands.

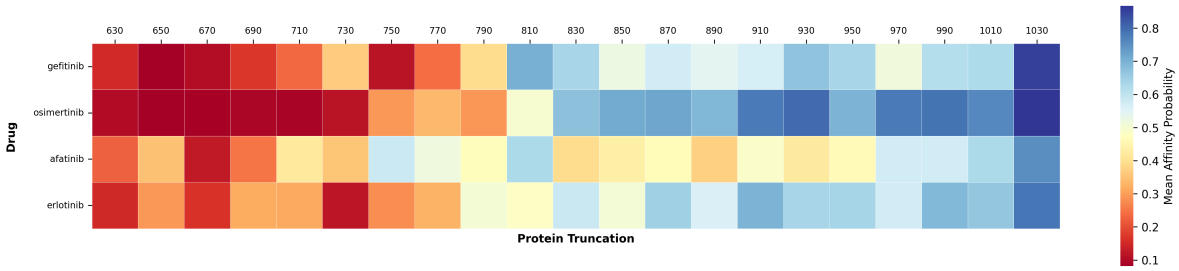


Figure 3: Binding probability predictions from Boltz-2 for truncated EGFR.

Next, residue-level mutations were introduced into the sensitive region. Boltz-2 was able to detect residues that are sensitive to changes and how it affects on ligand binding affinity (Figure 4).

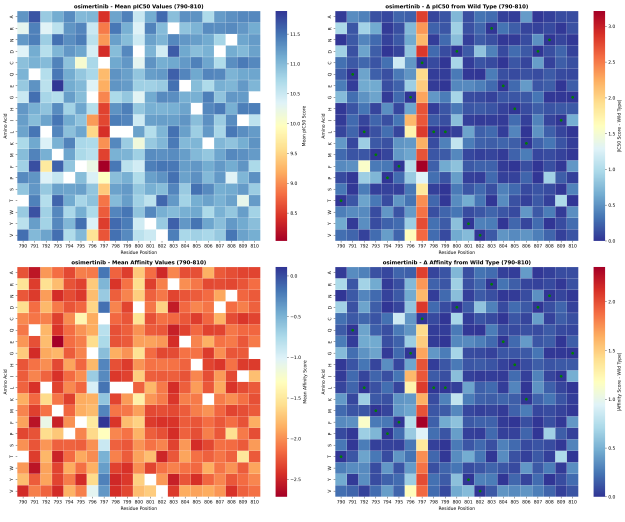


Figure 4: Residue-wise mutation analysis for EGFR.

7 Next Steps and Future Work

The next stage of the project will involve expanding the Boltz-2 mutation analysis to include additional cancer-relevant proteins and integrating $\Delta\Delta G$ prediction tools to assess protein stability under mutations. Additionally, cross-attention models will be trained and evaluated on larger datasets such as KIBA and PDBbind. Final findings will be prepared for peer-reviewe, and results will be presented at upcoming academic events, including KAUST and UHN showcases.

8 Conclusion

We present a robust benchmarking pipeline for GNN-based DTA prediction, highlighting key architectural tradeoffs and dataset biases. Our experiments with Boltz-2 demonstrate its potential as a mutation evaluation tool in early-stage target discovery.

References

[1] J. Yaacoub et al., "Manuscript for Method Paper," Princess Margaret Cancer Centre, 2023.

[2] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: deep drug–target binding affinity prediction," Bioinformatics, vol. 34, no. 17, pp. i821–i829, 2018.

[3] T. Nguyen et al., "GraphDTA: Predicting drug–target binding affinity with graph neural networks," Bioinformatics, vol. 37, no. 8, pp. 1140–1147, 2021.

[4] S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction," arXiv:2010.09885, 2020.

[5] M. I. Davis et al., "Comprehensive analysis of kinase inhibitor selectivity," Nat. Biotechnol., vol. 29, no. 11, pp. 1046–1051, 2011.

[6] J. Tang et al., "Making sense of large-scale kinase inhibitor bioactivity datasets," J. Chem. Inf. Model., vol. 54, no. 3, pp. 735–743, 2014.

[7] R. Wang et al., "The PDBbind database: Collection of binding affinities," J. Med. Chem., vol. 47, no. 12, pp. 2977–2980, 2004.

[8] S. Passaro et al., "Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction," bioRxiv, 2025, doi:10.1101/2025.06.14.659707.