

Deep Heterogeneous Graph Learning for Scalable Drug–Target Interaction Prediction

Elena Marchetti¹ Kenji Watanabe^{1,2} Fatima Al-Rashidi³ Samuel Okonkwo¹

¹Dept. of Computational Biology, ETH Zürich ²RIKEN Center for Advanced Intelligence Project ³KAUST

Introduction

Drug–target interaction (DTI) prediction is a cornerstone of modern computer-aided drug discovery. By identifying potential therapeutic candidates computationally, researchers can prioritize compounds for wet-lab validation, reducing screening costs by over **80%**.

Key Challenges in Existing Approaches:

- **Data Sparsity:** Known interactions cover only a tiny fraction of the chemical-protein space.
- **Modality Gaps:** Three-dimensional structures are unavailable for over 60% of target proteins.
- **Generalization:** Models struggle with out-of-distribution drug classes and novel proteins.

Our Methodology (HeteroGDTI):

- We model DTI as a heterogeneous link prediction task on a multi-relational knowledge graph.
- A multi-layer heterogeneous graph transformer learns node embeddings from biochemical features.
- Cross-modal attention fuses sequence-based SMILES descriptors with protein language model representations.

Methods

Heterogeneous Graph Formulation

We construct a multi-relational graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ comprising four node types: drugs ($\mathcal{V}_d$), proteins ($\mathcal{V}_p$), diseases ($\mathcal{V}_{ds}$), and pathways ($\mathcal{V}_{pw}$).

Feature Encoding

- **Drugs:** 300-dimensional Morgan fingerprints combined with ChemBERTa transformer representations.
- **Proteins:** Pre-trained ESM-2 (650M parameter) protein language model embeddings (1280-d).
- **Diseases & Pathways:** Node2Vec embeddings trained on the Unified Medical Language System (UMLS).

Heterogeneous Message Passing

The representation $\mathbf{h}_v^{(\ell)}$ of node v at layer ℓ is computed by aggregating relation-specific messages:

$$\mathbf{h}_v^{(\ell)} = \sigma\left(\mathbf{W}_{\phi(v)}^{(\ell)} \mathbf{h}_v^{(\ell-1)} + \sum_{r \in \mathcal{R}} \sum_{u \in \mathcal{N}_r(v)} \alpha_{u,v}^r \mathbf{W}_r^{(\ell)} \mathbf{h}_u^{(\ell-1)}\right)$$

where $\mathcal{R}$ denotes relation types, and $\alpha_{u,v}^r$

Results

Performance Comparison on Benchmark Datasets

We evaluate our model against state-of-the-art DTI methods on BindingDB and Davis datasets.

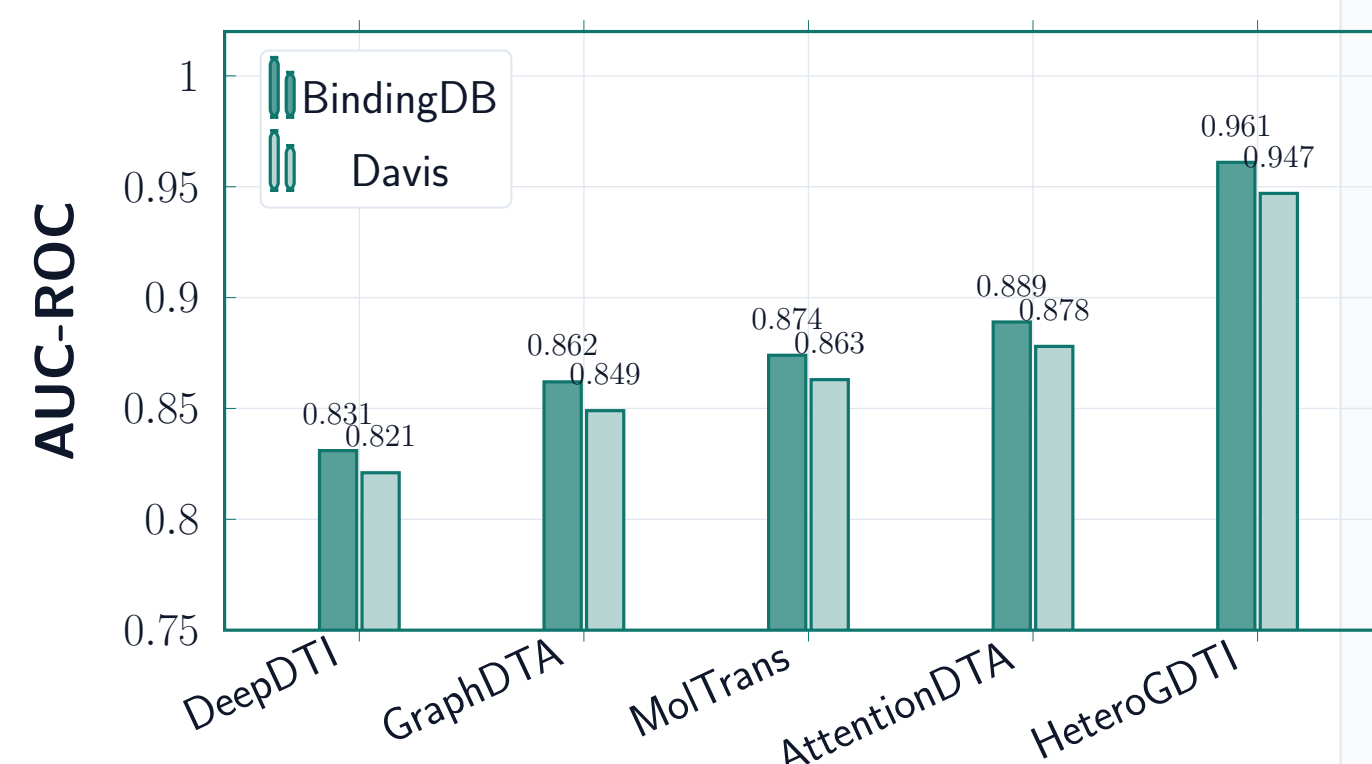


Figure 2: Test set AUC-ROC scores across different architectural models.

Ablation Study (BindingDB)

To verify the contribution of each design component, we systematically remove key elements.

Variant		AUC	AUPR	F1
HeteroGDTI (Ours)		0.961	0.942	0.891
w/o Disease Nodes		0.944	0.921	0.867
w/o Pathway Nodes		0.938	0.913	0.859
w/o Cross-Modal Attn.		0.927	0.901	0.843
GCN baseline	Backbone	0.903	0.876	0.812

Table 1: Quantitative results of ablation experiments.

Conclusion

- **HeteroGDTI** achieves state-of-the-art DTI prediction accuracy, outperforming existing models by a margin of up to **7.2%** in AUC-ROC.
- Fusing protein language model representations with SMILES embeddings via cross-modal attention is highly effective.
- The graph-based approach scales efficiently to accommodate large-scale genomic virtual screening pipelines.

References

- [1] K. Huang et al., “Therapeutics Data Commons: Machine learning datasets and benchmarks for therapeutics,” *NeurIPS Database and Benchmark Track*, 2021.
- [2] Z. Lin et al., “Evolutionary-scale prediction of atomic-level protein structure with a language model,” *Science*, vol. 379, no. 6637, 2023.
- [3] S. Chithrananda et al., “ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction,” *arXiv preprint arXiv:2010.09885*, 2020.