

Machine Learning–Guided Drug Discovery: Predicting Protein–Ligand Binding Affinity with Graph Neural Networks

Elena Vasquez¹, Thomas Wright^{1,2}, Kenji Yamamoto², Prof. Laura Kingston¹

¹Department of Biochemistry, University of Oxford

²DeepMind, London, UK

Contact: elena.vasquez@bioch.ox.ac.uk | ICML 2025 – Poster #247

Introduction

Drug discovery is a lengthy and expensive process, with the average new drug requiring over 10 years and \$2.6 billion to develop. A critical bottleneck is the accurate prediction of protein–ligand binding affinity, which determines whether a candidate molecule will effectively interact with its target protein.

Traditional approaches:

- Molecular docking (AutoDock, Glide) – fast but inaccurate
- Molecular dynamics simulations – accurate but computationally prohibitive
- Classical QSAR models – limited to predefined descriptors

Our contribution: We propose **AffinityGNN**, a graph neural network that operates directly on the 3D molecular graph of the protein–ligand complex to predict binding free energy (ΔG) with near-experimental accuracy.

Methods

Graph Construction. We represent the protein–ligand complex as a heterogeneous graph $G = (V_P \cup V_L, E)$ where:

- V_P : protein residue nodes (C α atoms)
- V_L : ligand heavy atom nodes
- E : edges based on spatial proximity ($< 8\text{\AA}$)

Architecture. AffinityGNN consists of:

- Node feature encoder (atom type, charge, hybridization)
- 6 layers of message-passing with attention
- Graph-level readout with Set2Set pooling
- Prediction head: 3-layer MLP $\rightarrow \Delta G$

Message Passing:

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

where α_{ij} are attention weights incorporating edge features (distance, angle).

Training. We use a combined loss:

$$\mathcal{L} = \underbrace{\|\hat{y} - y\|_2^2}_{\text{MSE}} + \lambda \underbrace{(1 - \rho(\hat{y}, y))}_{\text{Correlation}}$$

Datasets

Training Data:

- PDBbind v2020 refined set (5,316 complexes)
- BindingDB kinase subset (12,400 complexes)
- Custom curated GPCR dataset (3,200 complexes)

Data Augmentation:

- Random rotation and translation of coordinates
- Gaussian noise on atomic positions ($\sigma = 0.1\text{\AA}$)
- Subgraph sampling for large complexes

Results

Benchmark Performance on PDBbind v2020 core set:

Method	RMSE	Pearson R
AutoDock Vina	2.41	0.604
RF-Score v3	1.82	0.713
OnionNet-2	1.54	0.782
PLIP-GNN	1.41	0.801
AffinityGNN	1.18	0.862

Virtual Screening on DUD-E:

Target	AUC-ROC	EF _{1%}
CDK2 (kinase)	0.94	42.3
COX-2 (enzyme)	0.91	38.7
EGFR (receptor)	0.93	45.1
HIV-RT (viral)	0.89	31.4
Average	0.92	39.4

Experimental Validation. We used AffinityGNN to screen 50,000 compounds against SARS-CoV-2 main protease (M^{pro}). The top 100 predictions were synthesized and tested:

- 23 compounds showed $\text{IC}_{50} < 10 \mu\text{M}$
- 4 compounds showed $\text{IC}_{50} < 100 \text{ nM}$
- Best hit: $\text{IC}_{50} = 28 \text{ nM}$ (comparable to nirmatrelvir)

Generalization Across Targets:

Protein Family	RMSE	N
Kinases	1.12	847
Proteases	1.21	523
GPCRs	1.35	312
Nuclear rec.	1.19	198

Ablation & Analysis

Component Contributions:

Variant	RMSE
Full model	1.18
– Attention mechanism	1.34
– Edge features	1.29
– Correlation loss	1.25
– Set2Set pooling	1.31
GCN baseline (no attention)	1.52

Attention Visualization. The learned attention weights highlight key interactions at the binding site, including hydrogen bonds, π -stacking, and hydrophobic contacts—consistent with known biochemistry.

Scaling Behavior. Performance improves log-linearly with training data:

Training Size	RMSE
1,000 complexes	1.72
5,000 complexes	1.38
10,000 complexes	1.24
20,916 complexes	1.18

Conclusions

- AffinityGNN achieves **state-of-the-art** binding affinity prediction (RMSE = 1.18 kcal/mol)
- Interpretable** attention mechanism reveals binding site interactions
- Practical impact:** identified 4 potent M^{pro} inhibitors from virtual screening
- Inference time: **0.3ms per complex** (GPU), enabling large-scale screening
- Code and models available at github.com/oxbiochem/affinitygnn

Key References

- Corso et al., “DiffDock,” *ICLR*, 2023.
- Stärk et al., “EquiBind,” *ICML*, 2022.
- Wang et al., “PDBbind v2020,” *JCIM*, 2020.
- Kipf & Welling, “GCN,” *ICLR*, 2017.
- Veličković et al., “GAT,” *ICLR*, 2018.
- Gilmer et al., “MPNN,” *ICML*, 2017.

Acknowledgements

This work was supported by the Wellcome Trust (Grant 203141/Z/16/Z), EPSRC Doctoral Training Partnership, and computing resources from JADE2 (EP/T022205/1). We thank the Oxford Structural Genomics Consortium for providing crystal structures and binding data.