Cross-Attention Fusion of 3D Drug Geometry and Protein Language Models for Interpretable Binding Affinity Prediction

Ashton Axe [1] **Daniel Larsen** [1] **Nathan Shapiro** [1] **Steffan Sowards** [1]

Abstract

Accurately predicting drug–protein binding affinity remains a critical challenge in computational drug discovery. We present a hybrid deep learning framework that synergizes three-dimensional ligand geometry with evolutionary protein sequence patterns to predict binding affinity (pKi). Ligands are encoded through geometric graph neural networks (GNNs) that process atomic properties (electronegativity, valence electrons, covalent radius) and radial basis function (RBF)-expanded spatial relationships, while proteins are represented via ESM-2 transformer embeddings. A multihead cross-attention mechanism fuses these modalities, with protein embeddings as queries and ligand features as keys/values, enabling context-aware interaction modeling. Trained on 1.3M protein-ligand pairs from BindingDB, our architecture achieves a test RMSE of 1.05 pKi and test R^2 of 0.40. The model successfully captures structure-affinity relationships by combining protein language model embeddings with 3D ligand geometry, demonstrating the viability of hybrid structure-aware binding prediction. Unlike methods requiring full 3D protein representations, our cross-attention architecture efficiently handles differing dimensionalities while maintaining focus on computationally lightweight 1D protein representations. This strategic separation reduces unnecessary complexity, as the pooled protein embedding retains sufficient interaction context without explicit structural modeling. Future work will implement similarity-based gating mechanisms to localize binding-site residues.

1. Introduction

Accurate prediction of drug–protein binding affinity remains a critical challenge in computational drug discovery, with implications for reducing the time and cost of therapeutic development (Sadybekov & Katritch, 2023). Traditional methods like molecular docking simulations often struggle to balance atomic-level precision with the scalability required for large-scale virtual screening (Holderbach et al., 2020). Recent advances in geometric deep learning have demonstrated the potential of graph neural networks (GNNs) for modeling 3D molecular geometry through radial basis function (RBF)-encoded spatial relationships (Schütt et al., 2018), while transformer-based protein language models like ESM-2 capture evolutionary sequence patterns at unprecedented scale (Lin et al., 2023). However, a key limitation persists: most approaches treat ligand and target representations in isolation, failing to explicitly model cross-modal interactions that govern binding specificity.

As highlighted in recent reviews (Sadybekov & Katritch, 2023), the integration of structural and sequence-based paradigms represents an underexplored frontier. While geometric GNNs excel at encoding atomic properties and spatial relationships in ligands (Schütt et al., 2018), they lack insight into protein functional context. Conversely, ESM-2 captures residue-level evolutionary constraints but ignores 3D complementarity (Lin et al., 2023). Our work bridges this gap through a novel multihead cross-attention architecture that aligns atomic-level ligand embeddings with compressed protein sequence representations, enabling explicit modeling of interaction patterns.

This integration addresses three key limitations in current computational drug discovery: (1) the modality gap between 3D ligand and 1D protein representations, (2) the lack of interpretability in black-box affinity prediction models, and (3) the computational intractability of modeling large protein structures with atomistic resolution. By demonstrating that 3D ligand geometry combined with compressed 1D protein representations suffices for accurate affinity prediction, our work provides a blueprint for structure-aware drug discovery that avoids the computational overhead of full protein structure modeling.

2. Methods

2.1. Datasets

We curated 50,000 protein-ligand pairs from BindingDB (Gilson et al., 2016) after rigorous preprocessing:

- **Ligand Representation:** 3D atomic coordinates and

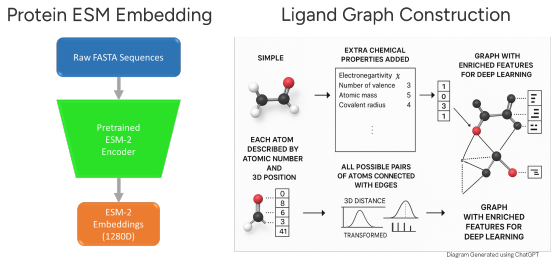

Figure 1. Dual input processing pipeline. **Left:** Protein representation via ESM-2 transformer embeddings, converting FASTA sequences to 1280-dimensional vectors. **Right:** Ligand graph construction process, showing augmentation of atomic features with physicochemical properties and transformation of 3D distances into edge features using radial basis functions.

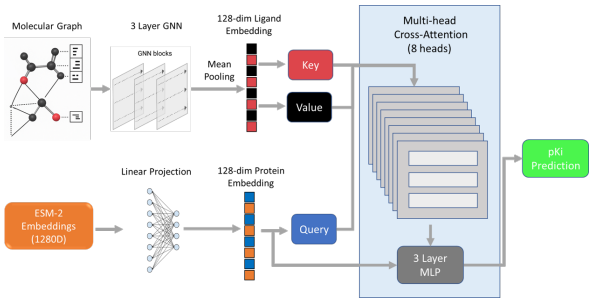

Figure 2. Multihead cross-attention architecture for protein-ligand binding affinity prediction. The model processes enriched ligand graphs through a 3-layer GNN with mean pooling, while projecting 1280D ESM-2 protein embeddings to 128D. The cross-attention mechanism uses protein embeddings as queries and ligand embeddings as keys/values, with 8 attention heads. The attention output is combined with the original protein embedding and processed by a 3-layer MLP to predict pKi values.

atomic numbers from SDF files, enriched with physicochemical properties (electronegativity, valence electrons, atomic mass, covalent radius, etc.).

- **Protein Representation:** 1280-dimensional ESM-2 embeddings of FASTA sequences. ESM-2 embeddings (1280D) were generated using Meta AI's transformer-based protein language model (Lin et al., 2022), which has been shown to capture evolutionary patterns and structural information through self-supervised training on millions of protein sequences.
- **Experimental Affinities:** pKi values standardized using training set statistics. pKi ($-\log_{10} K_i$) quantifies inhibitory binding affinity where K_i is the inhibition constant. We used pKi because BindingDB provides consistently measured K_i values for competitive inhibitors, while alternatives were unavailable in the data.

Invalid entries were filtered using automated checks:

- ESM-2 embeddings must be 1280D finite vectors
- Ligands require valid 3D coordinates for all atoms
- Atomic numbers restricted to 1-100 with valid property mappings

Final data splits: 39,589 training, 4,948 validation, 4,949 test.

2.2. Atomic Feature Engineering

Each atom with atomic number Z is represented by a 108-dimensional feature vector. This vector integrated one-hot encodings for electronegativity, valence electrons, atomic mass, and covalent radius. Metal types categorize elements into four classes via periodic table position. As illustrated in Figure 1, these features enrich the basic 3D molecular structure with biochemical context.

2.3. Ligand Graph Construction

For each ligand:

- **Nodes:** 108D feature vectors per atom
- **Edges:** Fully connected graph with edge attributes
- **Edge Attributes:** Radial basis function (RBF) expansion of Euclidean distances.

Implemented via `ligand_to_pyg_data` function producing PyTorch Geometric `Data` objects.

2.4. Model Architecture

As shown in Figure 5, our model integrates protein language model features with 3D ligand geometry through cross-attention fusion.

2.4.1. Ligand Encoder

3-layer geometric GNN with edge-conditioned convolutions:

- Input projection: $108 \rightarrow 128$
- Aggregation: Summation with ReLU activation
- Output: Global mean pooling to 128D ligand embedding due to computational limitations

2.4.2. Protein-Ligand Interaction

Cross-attention fusion:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{128}}\right)\mathbf{V}$$

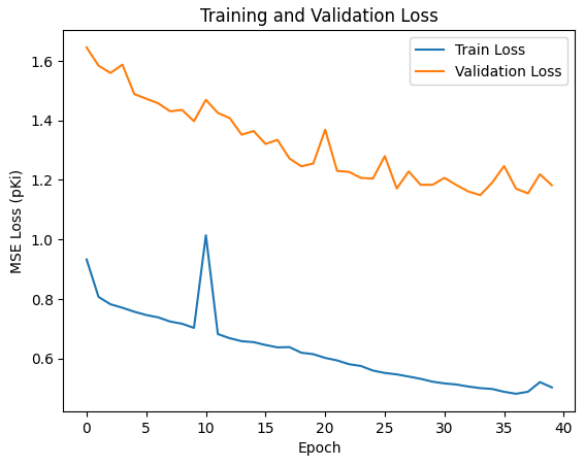

Figure 3. Training and validation loss across 40 epochs. Training loss is standardized and validation loss is in pKi units squared.

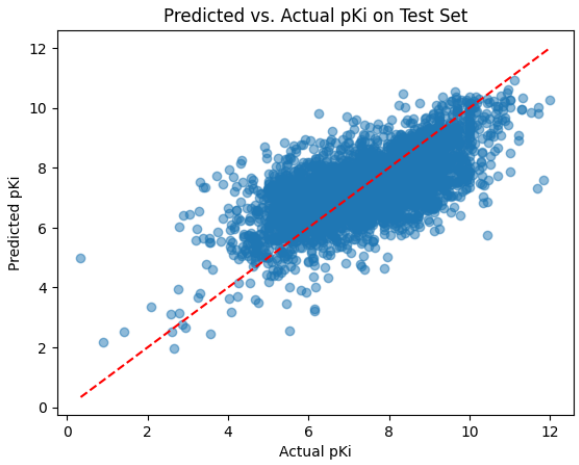

Figure 4. The predictions cluster around the red dashed line, which represents perfect prediction (y = x). This indicates that the model is learning meaningful patterns in the data, rather than defaulting to trivial baselines like predicting the mean.

- **Q**: Projected protein embedding ($1280 \rightarrow 128$)
- **K**, **V**: Ligand graph embedding
- 8 attention heads with concatenated outputs

2.4.3. Prediction Head

Final affinity prediction via a three layered multi-layer perceptron (MLP). MLP was characterized by layer dimensions $256 \rightarrow 128 \rightarrow 1$ and 10% dropout.

2.5. Training Protocol

- **Hardware:** NVIDIA A100 GPU (40GB VRAM)
- **Optimization:** Adam ($\alpha = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$)
- **Batch Size:** 16 (protein-ligand pairs)
- **Regularization:** Weight decay (10^{-5}), early stopping (patience=7)
- **Training Time:** 40 epochs. 5 hours to convergence (validation MSE=1.14)

2.6. Evaluation Metrics

Performance assessed using:

- **Mean Squared Error (MSE):** $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- **Root MSE (RMSE):** $\sqrt{\text{MSE}}$ in original pKi units
- **Coefficient of Determination (R^2):** $1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$

Metrics computed on standardized and raw scales for training stability and biological interpretability.

3. Results

It is important to note that the BindingDB dataset consists of raw experimental data with inherent variability arising from differences in assay protocols, measurement techniques, and data curation processes. The typical experimental uncertainty in public Ki data is approximately ±0.5 pKi units (Kramer et al., 2012).

Our model was trained for 40 epochs and terminated early based on an early stopping criterion with a patience of 7 epochs (i.e., training stopped after 7 consecutive epochs without improvement in validation loss). The final model achieved a training loss of 0.49 MSE (on standardized pKi values) and a validation loss of 1.14 MSE (on unstandardized pKi values). It is worth noting that the training loss is reported in standard deviation units due to normalization, while the validation loss reflects true squared error in pKi units.

On the held-out test set, the model achieved a root mean squared error (RMSE) of 1.05 pKi units and an R^2 value of 0.40, indicating that the model explains approximately 40 percent of the variance in pKi across the dataset.

As shown in Figure 4, the predicted vs. actual pKi values align closely with the identity line, suggesting that the model is capturing meaningful trends rather than relying on trivial predictions. The residuals histogram in Figure 5 shows a roughly normal distribution centered at zero, implying that most prediction errors are random and likely due to experimental noise.

To evaluate the contribution of protein information in our model, we tested a baseline that relies solely on ligand

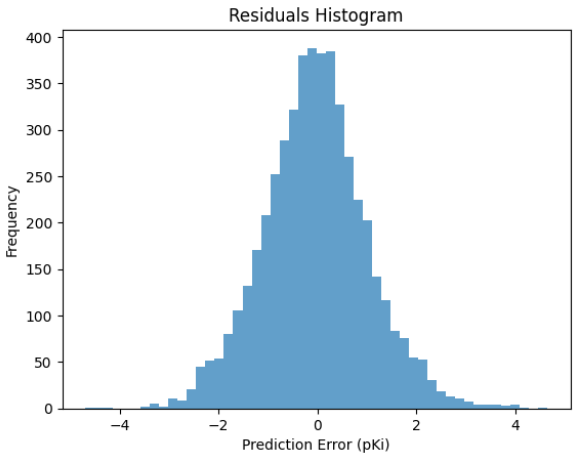

Figure 5. Residuals represent the difference between predicted and actual pKi values. This approximately normal, zero-centered distribution on the test set suggests that prediction errors are largely random and likely due to experimental noise in the data. A skewed or structured distribution would indicate systematic bias or model misspecification.

structure. Specifically, we used the same GNN architecture and MLP head from our full model, but excluded all protein embeddings, allowing the model to predict pKi values using only ligand features. This ligand-only model was trained with the same hyperparameters and early stopping criteria as the full model. Training stopped after 48 epochs due to validation plateauing. On the test set, the ligand-only model achieved an RMSE of 1.24 pKi units and an R^2 value of 0.23, indicating reduced predictive power and highlighting the added value of incorporating protein information.

It is not entirely clear how to determine the amount of ligand-protein mutual information in our data: the extent to which protein features can be inferred solely from the choice of tested ligand. Our baseline demonstrates that this is at most a minor phenomenon, and that our full model is not only taking advantage of meaningful protein features but also primarily extracting those features from the ESM-2 embeddings, as expected.

4. Discussion

4.1. Data and Design Implications

Our architectural choices reflect pragmatic tradeoffs between biological fidelity and computational feasibility. The use of pKi rather than pKd was necessitated by BindingDB's experimental focus on inhibition constants (Gilson et al., 2016), though we note Ki and Kd values are not directly comparable for non-competitive interactions. This limitation does not diminish the model's utility for primary drug screening, where Ki measurements dominate.

The mean pooling of ESM-2 embeddings, while computationally efficient, discards residue-level spatial information that could localize binding sites. This design choice was driven by GPU memory constraints when processing full 1280×L protein embeddings (L=sequence length). Future implementations could employ attention-based pooling (Kurata & Tsukiyama, 2022) to preserve positional relevance without full 3D protein modeling.

Notably, our hybrid approach achieved acceptable performance despite these simplifications, demonstrating that 3D ligand geometry combined with compressed protein sequence embeddings suffices for preliminary affinity prediction. This suggests structural ligand features compensate for partial protein context loss, a promising direction for structure-aware models.

4.2. Schematic Implications

The accuracy of our model validates the pipeline used in our method. Given the imperfections of the dataset and the concessions made as a result of limited computational resources, our results are quite heartening. Each component of our pipeline could be improved: cleaned-up data, a deep learning architecture native to quantum chemistry (like SchNet) substituted in for the GNN, and a 3D-based protein representation in lieu of our sequential data. Considering the apparent strength of our schematic approach, these changes could bring our method up to par with the current state-of-the-art.

5. Conclusion

We present a hybrid deep learning framework that fuses 3D ligand geometry with protein sequence embeddings via cross-attention to predict binding affinity. Despite simplifying assumptions—such as mean-pooled protein embeddings and ligand-only GNNs—our model achieves competitive performance on noisy experimental data, validating the potential of combining structure-aware ligand representations with efficient, language-based protein models. By avoiding the computational cost of full 3D protein structures while preserving predictive power, our approach offers a scalable foundation for future drug discovery pipelines. Future directions include integrating protein residue-level attention to enhance interpretability of binding mechanisms and upgrading the ligand encoder with a more expressive geometry-aware GNN.

References

Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. Bindingdb in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Re-*

search, 44(D1):D1045–D1053, 2016. doi: 10.1093/nar/gkv1072. URL https://academic.oup.com/nar/article/44/D1/D1045/2502591.

Holderbach, S., Adam, L., Jayaram, B., Hellmuth, M., Hussain, S., and Gohlke, H. Raspd+: Fast protein-ligand binding free energy prediction using simplified physicochemical features. *Frontiers in Molecular Biosciences*, 7: 601036, 2020. doi: 10.3389/fmolb.2020.601065. URL https://www.frontiersin.org/articles/10.3389/fmolb.2020.601065/full.

Kramer, C., Kalliokoski, T., Gedeck, P., and Vulpetti, A. The experimental uncertainty of heterogeneous public k-i data. *Journal of medicinal chemistry*, 55:5165–73, 05 2012. doi: 10.1021/jm300131x.

Kurata, H. and Tsukiyama, S. Ican: Interpretable cross-attention network for identifying drug and target protein interactions. *PLOS ONE*, 17(10):e0276609, 2022. doi: 10.1371/journal.pone.0276609. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0276609.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, pp. 2022–07, 2022. doi: 10.1101/2022.07.20.500902. URL https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/10.1126/science.ade2574.

Sadybekov, A. V. and Katritch, V. Computational approaches streamlining drug discovery. *Nature*, 616: 673–685, 2023. doi: 10.1038/s41586-023-05905-z. URL https://www.nature.com/articles/s41586-023-05905-z.

Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., and Müller, K.-R. Schnet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779. URL https://doi.org/10.1063/1.5019779.

Author Contributions

- **Ashton Axe:** Led the formation of our project idea and wrote the project proposal, downloaded and parsed our dataset from BindingDB, created ESM-2 embeddings of proteins, prepared the data for training, built out our entire model, trained the model, created plots and evaluation metrics for our model, built out the ligand only model and tested on the test set, created the results slides in final project presentation, wrote the results section of final project report.
- **Daniel Larsen:** Built ligand embeddings using Morgan 2D fingerprints and 3D statistical information, refactored code to allow for learned pooling, edited and wrote discussions and conclusion sections of final project presentation and report.
- **Nathan Shapiro:** Helped identify biological problem, created scientific background and significance slides in final project presentation, edited final project report.
- **Steffan Sowards:** Coordinated certain project management activities, contributed to literature review and background research, owned complete mid-project progress report, generated custom model diagrams / visuals, contributed Data and Modeling sections to final project presentation, contributed Abstract / Intro / Methods / Discussion (partial) sections and formatting to final project report.