Learning to Align Molecules and Proteins: A Geometry-Aware Approach to Binding Affinity

Mohammadsaleh Refahi
Drexel University
Philadelphia, PA
sr3622@drexel.edu

Bahrad A. Sokhansanj
Drexel University
Philadelphia, PA
bahrad@molhealtheng.com

James R. Brown
Drexel University
Philadelphia, PA
jb4633@drexel.edu

Gail Rosen
Drexel University
Philadelphia, PA
glr26@drexel.edu

Abstract

Accurate prediction of drug–target binding affinity can accelerate drug discovery by prioritizing promising compounds before costly wet-lab screening. While deep learning has advanced this task, most models fuse ligand and protein representations via simple concatenation and lack explicit geometric regularization, resulting in poor generalization across chemical space and time. We introduce **FIRM-DTI**, a lightweight framework that conditions molecular embeddings on protein embeddings through a feature-wise linear modulation (FiLM) layer and enforces metric structure with a triplet loss. An RBF regression head operating on embedding distances yields smooth, interpretable affinity predictions. Despite its modest size, FIRM-DTI achieves state-of-the-art performance on the Therapeutics Data Commons DTI-DG benchmark, as demonstrated by an extensive ablation study and out-of-domain evaluation. Our results underscore the value of conditioning and metric learning for robust drug–target affinity prediction.(Code available at https://github.com/EESI/Firm-DTI)

1 Introduction

The process of discovering a new therapeutic agent hinges on understanding how strongly a candidate molecule binds to a protein target. Binding affinity determines pharmacological potency and selectivity, but measuring it experimentally is laborious and expensive. Computational predictions of binding affinity therefore play an important role in virtual screening and lead optimization. Traditional quantitative structure–activity relationship models rely on handcrafted descriptors and are often restricted to narrow chemical series. Deep neural networks promise to learn richer representations from raw sequences and graphs, yet many state-of-the-art methods treat the two modalities independently and use ad-hoc interaction layers without explicit regularization. Furthermore, high-quality affinity data are scarce relative to the vast chemical and proteomic space, making generalization to novel drugs and targets challenging.

This work proposes a simple and effective architecture that addresses these limitations. Instead of concatenating ligand and protein features, we employ a FiLM conditioning layer to modulate molecular embeddings by protein embeddings, allowing the model to learn target-specific transformations. We additionally incorporate a triplet metric-learning objective that pulls interacting pairs together and pushes non-interacting pairs apart in the embedding space. A radial basis function (RBF) regression

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: 2nd Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences.

head then maps distances to continuous affinity values. As we show on a temporally split benchmark, this combination yields competitive performance with far fewer parameters than recent large models.

2 Background

Deep learning has become a cornerstone of drug–target interaction (DTI) and binding affinity prediction. Early sequence-based methods typically embedded drug SMILES strings and protein sequences with recurrent or convolutional networks, then combined the embeddings with a feed-forward layer. More recent representation learning approaches have introduced graph neural networks for molecular structures and large protein language models for amino acid sequences. For instance, MolE employs a disentangled attention transformer to produce atom-level graph embeddings Méndez-Lucio et al. [2024], while ESM leverages masked language modeling over massive protein corpora to yield residue-aware embeddings Lin et al. [2023]. ChemBERTaChithrananda et al. [2020] and rxnfpSchwaller et al. [2021] have further demonstrated the utility of pretrained transformers directly on SMILES strings for chemical property prediction.

Hybrid architectures often concatenate drug and protein embeddings and pass them through a multi-layer perceptron, but this strategy ignores conditional dependencies between ligands and targetsThafar et al. [2022]. Knowledge-based models attempt to mitigate this by integrating curated interaction networks or graph-derived embeddings (e.g., node2vec over protein–protein interaction graphs), yet such external information is costly to maintain and may not generalize across chemical spaceLam et al. [2023], Fattahi et al. [2019].

Recent studies show that residue-level protein language model embeddings contain rich contextual signals for drug–target interaction tasks, especially when combined with transfer learning strategies such as contrastive or task-specific pretraining NaderiAlizadeh and Singh [2025], Singh et al. [2023], Sledzieski et al. [2022]. However, most sequence-based DTI models still rely on simple concatenation of drug and protein representations, which limits their ability to capture robust geometric relationships in the latent space. Explicit geometry-aware designs remain underexplored: while protein PLMs enable similarity learning at the residue level, few models have incorporated metric-based objectives to enforce clustering of interacting pairs.

3 Methods

Our framework consists of three stages: featurization of molecules and proteins, conditioning via a FiLM layer, and distance-based affinity regression. Figure 1 summarizes the architecture.

3.1 Feature extraction

We obtain latent representations for ligands and proteins from pretrained experts. The MolE model treats a molecule as a graph where nodes represent atoms and edges represent bonds. Atom features include Daylight atomic invariants and the Morgan fingerprint, and positional information is encoded through relative bond distances. A disentangled attention transformer then produces an embedding $z_d \in \mathbb{R}^d$ for each molecule Méndez-Lucio et al. [2024]. For the target protein, we use the ESM2 language model which processes amino acid sequences and outputs a per-sequence embedding $z_t \in \mathbb{R}^d$ by averaging residue representations Lin et al. [2023].

3.2 FiLM conditioning

To incorporate target context, we apply a FiLM layer Perez et al. [2018]. Given drug embedding z_d and protein embedding z_t, the conditioned embedding is

$$\mathrm{FiLM}(z_d \mid z_t) = \gamma(z_t) \odot z_d + \beta(z_t),$$

where γ and β are learned linear functions of z_t and $\odot$ denotes element-wise multiplication. This transformation allows the model to scale and shift molecular features based on the target, capturing conditional interactions more flexibly than concatenation.

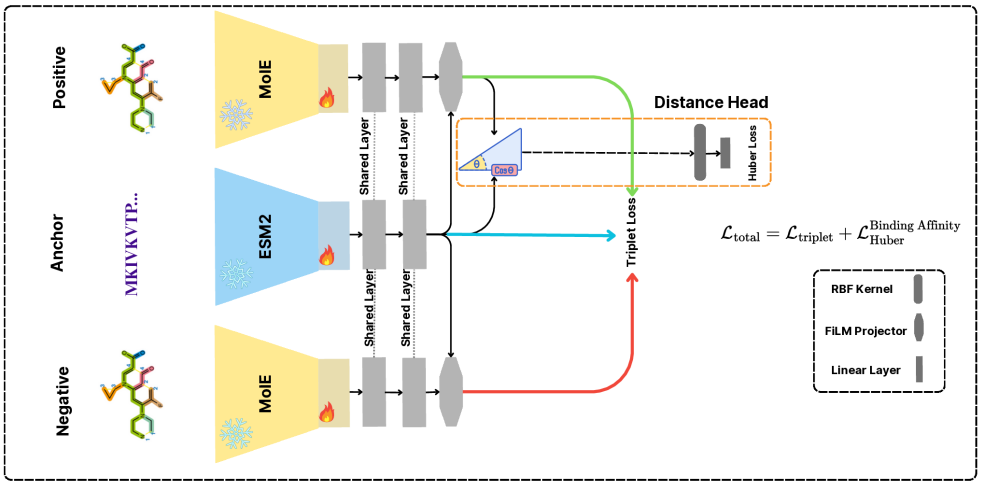


Figure 1: FIRM-DTI architecture. Drug embeddings (from MolE) and protein embeddings (from ESM2) are projected into a shared latent space via a FiLM conditioning layer. The model is trained with a triplet loss to enforce geometric alignment between interacting pairs, and an RBF-based prediction head is used to regress binding affinity values.

3.3 Distance-based prediction head

After conditioning, we normalize the drug and protein embeddings and compute their cosine distance:

$$\text{dist}(\tilde{z}_d, \tilde{z}_t) = 1 - \frac{\tilde{z}_d \cdot \tilde{z}_t}{\|\tilde{z}_d\| \, \|\tilde{z}_t\|}.$$

This distance is passed through a set of radial basis functions with k centers μ_j evenly spaced in $[0, 2]$:

$$\phi_j = \exp\big(-((\text{dist}(\tilde{z}_d, \tilde{z}_t) - \mu_j)^2)/(2\sigma^2)\big), \quad j = 1, \ldots, k.$$

The final affinity prediction is computed via a linear layer $y_{\text{pred}} = W\phi + b$. This head enforces that similar embeddings yield similar predictions and provides a smooth mapping from distances to continuous affinities.

3.4 Training objective

Training requires both positive and negative drug–protein pairs. Positive examples are experimentally validated interactions with known binding affinities, while negative examples are randomly paired molecules and targets with no reported interaction. During training, we minimize a combination of a triplet loss and a Huber regression loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{triplet}} + \mathcal{L}_{\text{Huber}}^{\text{affinity}},$$

where

$$\mathcal{L}_{\text{triplet}} = \max\big(0, d(f(x_a), f(x_p)) - d(f(x_a), f(x_n)) + \alpha\big)$$

encourages the distance between an anchor x_a and a positive x_p to be smaller than the distance to a negative x_n by margin α, and the Huber loss with threshold δ stabilizes regression:

$$\mathcal{L}_{\text{Huber}}^{\text{affinity}} = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{if } |y - \hat{y}| \le \delta, \\ \delta\,|y - \hat{y}| - \frac{1}{2}\delta^2, & \text{otherwise.} \end{cases}$$

3.5 Dataset and evaluation protocol

We evaluate on the Drug–Target Interaction Domain Generalization (DTI-DG) benchmark introduced by the Therapeutics Data Commons Huang et al. [2021a]. The benchmark partitions binding affinity data derived from BindingDB by patent year, training models on interactions from 2013–2018 and

Table 1: Impact of model components on DTI-DG performance.

Model variant	PCC ↑
Full model	0.59
– without FiLM conditioning	0.55
– without triplet loss	0.32

testing on interactions from 2019–2021. This temporal split mimics the challenge of predicting affinities for newly patented drugs and targets. All models are trained for 25 epochs with identical hyperparameters, and performance is measured by Pearson correlation coefficient (PCC) between predicted and true affinities.

DTI-Datasets We evaluate on three standard DTI benchmarks: **DAVIS** Davis et al. [2011], **BindingDB** Liu et al. [2007], and **BIOSNAP** (ChG-Miner) Zitnik et al. [2018]. Following **MolTrans** Huang et al. [2021b], we binarize DAVIS and BindingDB using $K_d < 30$ (positive) and $K_d \geq 30$ (negative). BIOSNAP contains only positives; consistent with MolTrans, we generate negatives by randomly pairing proteins and compounds in equal number to the positives. Each dataset is split into 70% train, 10% validation, and 20% test.

4 Results

4.1 Ablation study

To understand the contributions of each component, we conduct an ablation study on DTI-DG. Table 1 reports the mean PCC over the test domains when removing the FiLM conditioning layer or the triplet loss. Eliminating the FiLM projector leads to a modest decline in performance, while omitting the triplet loss causes a severe drop, highlighting the importance of metric learning.

4.2 Out-of-domain prediction

Figure 2 compares our method against recent baselines on the DTI-DG test set. Despite using fewer parameters and no external knowledge (unlike the Otter-Knowledge ensemble Lam et al. [2023]), our model achieves a PCC of 0.59. This performance surpasses state-of-the-art sequence-based models such as PLM-SWE NaderiAlizadeh and Singh [2025] and TxGemma27B Wang et al. [2025], as well as network-based approaches like Otter-Knowledge Lam et al. [2023]. These results highlight that while some models rely on massive scale (e.g., 27B parameters in TxGemma) or external knowledge (e.g., Otter-Knowledge), our approach attains superior performance through conditioning and metric learning alone. As shown in Fig. 2b, the learned RBF head with $\sigma = 0.2$ produces a smooth mapping $\hat{y}(d) = \sum_j w_j \exp\left(-\frac{(d-\mu_j)^2}{2\sigma^2}\right)$ from cosine distance d to affinity, yielding $r = 0.91$ correlation and confirming that, unlike baselines in Fig. 2a, our model encodes binding affinity directly as a function of embedding geometry.

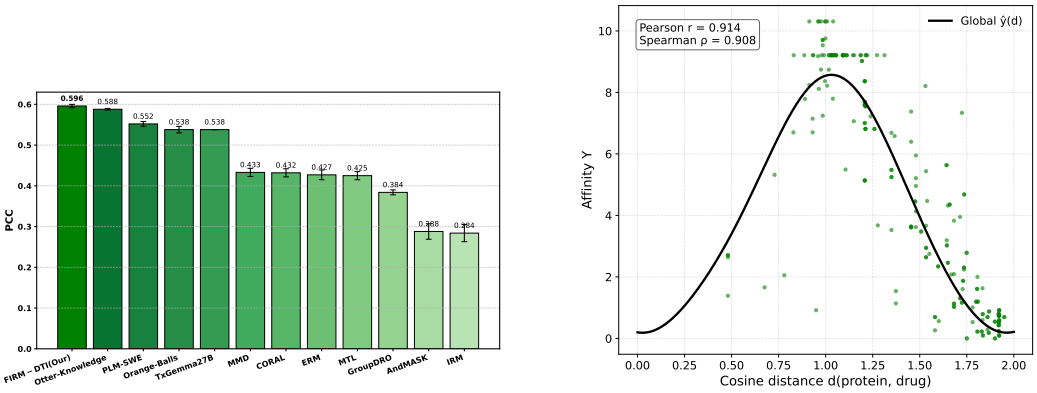
4.3 DTI Experiments

In addition to binding–affinity regression, we trained *drug–target interaction* (DTI) classifiers. Our setup follows Sledzieski et al. [2022], Singh et al. [2023] with two modifications. First, we replace the Huber regression objective with `BCEWithLogitsLoss` to model binary interaction labels directly. Second, we adopt a contrastive (triplet-style) sampling strategy for continual learning: within each mini-batch the *anchor* is the protein (target), the *positive* is a drug known to interact with that protein, and the *negative* is a different drug not known to interact with that protein (negatives can be resampled each epoch).

Table 2 compares FIRM-DTI with strong baselines on three standard DTI datasets. Across BIOSNAP and BindingDB our model consistently achieves the highest or comparable AUPR and AUROC scores, showing that the FiLM-conditioned, geometry-aware representation generalizes well beyond affinity regression. On the much smaller and more imbalanced DAVIS set, performance drops for

Table 2: Comparison on BIOSNAP, BindingDB, and DAVIS. Mean ± s.e.m. over 5 random seeds. Sledzieski et al. [2022].

Benchmark	Model	AUPR	AUROC
BIOSNAP	**FIRM-DTI (Our Result)**	**0.919 ± 0.016**	**0.910 ± 0.004**
	ConPLex Singh et al. [2023]	0.897 ± 0.001	–
	ProtBert + Morgan Sledzieski et al. [2022]	0.895 ± 0.004	0.873 ± 0.004
	MolTrans Huang et al. [2021b]	0.885 ± 0.005	0.876 ± 0.007
	GNN-CPI Tsubaki et al. [2019]	0.890 ± 0.004	0.879 ± 0.007
	DeepConv-DTI Lee et al. [2019]	0.889 ± 0.005	0.883 ± 0.002
BindingDB	**FIRM-DTI (Our Result)**	0.647 ± 0.003	**0.916 ± 0.001**
	ConPLex Singh et al. [2023]	0.628 ± 0.012	–
	ProtBert + Morgan Sledzieski et al. [2022]	**0.652 ± 0.005**	0.876 ± 0.007
	MolTrans Huang et al. [2021b]	0.598 ± 0.013	0.898 ± 0.009
	GNN-CPI Tsubaki et al. [2019]	0.578 ± 0.015	0.900 ± 0.004
	DeepConv-DTI Lee et al. [2019]	0.611 ± 0.015	0.908 ± 0.004
DAVIS	**FIRM-DTI (Our Result)**	0.460 ± 0.004	0.880 ± 0.001
	ConPLex Singh et al. [2023]	0.458 ± 0.016	–
	ProtBert+Morgan Sledzieski et al. [2022]	**0.511 ± 0.012**	**0.917 ± 0.003**
	MolTrans Huang et al. [2021b]	0.335 ± 0.017	0.889 ± 0.007
	GNN-CPI Tsubaki et al. [2019]	0.269 ± 0.020	0.840 ± 0.012
	DeepConv-DTI Lee et al. [2019]	0.299 ± 0.039	0.884 ± 0.008



(a) PCC on DTI-DG benchmark.

(b) Distance–affinity mapping.

Figure 2: (a) Our model outperforms recent baselines on DTI-DG despite fewer parameters and no external knowledge. (b) The learned RBF head produces a smooth mapping $\hat{y}(d)$ from cosine distance to affinity (r=0.91), demonstrating geometry-awareness.

all methods but FIRM-DTI remains competitive, underscoring its robustness under challenging data distributions.

5 Discussion

Our study shows that a relatively small model can achieve competitive accuracy on challenging drug–target affinity prediction tasks when equipped with the right inductive biases. The FiLM conditioning layer allows the network to learn target-specific transformations of molecular features, and the triplet loss enforces a geometric structure that aligns interacting pairs. Together, these components improve out-of-domain generalization without relying on large architectures or external knowledge bases. A limitation of our work is the reliance on pretrained experts; future research may explore joint training of molecular and protein embeddings or incorporate three-dimensional structural information. Additionally, investigating uncertainty estimation could further accelerate practical drug discovery workflows.

References

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

Fatemeh Fattahi, Mohammad S Refahi, and Behrouz Minaei-Bidgoli. Drug-target interaction prediction using edge2vec algorithm on the heterogeneous network via svm. In *2019 5th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pages 1–5. IEEE, 2019.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W Coley, Cao Xiao, Jimeng Sun, and Marinka Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021a.

Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021b.

Hoang Thanh Lam, Marco Luca Sbodio, Marcos Martínez Galindo, Mykhaylo Zayats, Raul Fernandez-Diaz, Victor Valls, Gabriele Picco, Cesar Berrospi Ramis, and Vanessa Lopez. Otter-knowledge: benchmarks of multimodal knowledge graph representation learning from different sources for drug discovery. *arXiv preprint arXiv:2306.12802*, 2023.

Ingoo Lee, Jongsoo Keum, and Hojung Nam. Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS computational biology*, 15(6): e1007129, 2019.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N Jorissen, and Michael K Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

Oscar Méndez-Lucio, Christos A Nicolaou, and Berton Earnshaw. Mole: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15(1):9431, 2024.

Navid NaderiAlizadeh and Rohit Singh. Aggregating residue-level protein language model embeddings with optimal transport. *Bioinformatics Advances*, 5(1):vbaf060, 2025.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Philippe Schwaller, Daniel Probst, Alain C Vaucher, Vishnu H Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152, 2021.

Rohit Singh, Samuel Sledzieski, Bryan Bryson, Lenore Cowen, and Bonnie Berger. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24):e2220778120, 2023.

Samuel Sledzieski, Rohit Singh, Lenore Cowen, and Bonnie Berger. Adapting protein language models for rapid dti prediction. *bioRxiv*, pages 2022–11, 2022.

Maha A Thafar, Mona Alshahrani, Somayah Albaradei, Takashi Gojobori, Magbubah Essack, and Xin Gao. Affinity2vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning. *Scientific reports*, 12(1):4751, 2022.

Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 2019.

Eric Wang, Samuel Schmidgall, Paul F Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*, 2025.

Marinka Zitnik, Rok Sosic, and Jure Leskovec. Biosnap datasets: Stanford biomedical network dataset collection. *Note: http://snap. stanford. edu/biodata Cited by*, 5(1), 2018.

Technical Appendices

Training details

We trained our model on a single NVIDIA H100 GPU with 80 GiB of memory. The total number of learnable parameters is $\approx 1.2 \times 10^8$, of which the vast majority reside in the pretrained encoders. For the protein encoder we use the ESM2 model `esm2_t12_35M_UR50D` from Facebook AI's repository[1]; the molecular encoder is MolE, trained on the GuacaMol dataset[2]. Protein sequences are truncated or padded to a maximum length of 1,500 residues.

We optimize with the AdamW algorithm ($\epsilon = 10^{-6}$) and use a cosine learning rate schedule with linear warmup. The main hyperparameters are summarized in Table 4.

The sensitivity of performance to the Huber loss threshold δ is shown in Table 3.

Table 3: Effect of Huber loss threshold δ on PCC (DTI-DG).

δ	**PCC**
0.75	0.56
0.50	0.59
0.25	0.58

Table 4: Training hyperparameters.

Hyperparameter	**Value**
Batch size	24 drug–protein pairs
Learning rate	5×10^{-5} (cosine decay, 500 warmup steps)
Optimizer	AdamW, $\epsilon = 10^{-6}$
Weight decay	0.1
Triplet margin α	0.9
Max protein length	1500 residues
Epochs	25
Total parameters	$\approx 120M$
GPU used	NVIDIA H100 80 GiB
Training time	$\sim$29 hours per run

[1] `https://huggingface.co/facebook/esm2_t12_35M_UR50D`

[2] `https://codeocean.com/capsule/2105466/tree/v1`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We carefully reviewed the abstract and introduction, and they do not contain any claims that are not justified in the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We explicitly acknowledge limitations in Section 6, including reliance on pretrained experts, the absence of 3D structural information, and the need for future work on uncertainty estimation.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide all necessary assumptions and a complete derivation of the loss functions, including the triplet and Huber objectives.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification:We use publicly available datasets and describe all model architectures, hyperparameters, and training procedures in detail. The code will be released upon acceptance to further support reproducibility.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used are publicly available and cited appropriately. We will release the code and detailed instructions for reproducing the main experimental results in the supplementary material and a public repository upon acceptance.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper provides detailed descriptions of data splits, optimizer settings, training schedules, and hyperparameters in Appendix. These include the batch size, learning rate, number of epochs, and evaluation metrics used across all benchmarks.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For datasets with official splits, we follow the provided train/test partitions without additional resampling.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the GPU type, memory configuration, and training time estimates in Appendix 6.1 These include hardware specifications such as H100 GPUs and per-task training durations to support reproducibility.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We outline positive impacts for accelerating drug discovery and potential risks related to misuse or overreliance on computational predictions.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The datasets and models used in this work are not considered high risk for misuse; they are based on publicly available genomic data and do not involve sensitive personal or dual-use information.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external datasets and models used in this work are properly cited, and we rely exclusively on publicly released assets with appropriate licenses, as noted in the references and Appendix A.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This work does not introduce any new datasets, models, or software assets beyond those already publicly available and cited.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This research does not involve crowdsourcing or human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects research was conducted; thus, IRB approval was not required.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: Large language models were not used as part of the core methodology or experimental components of this research. Any use of LLMs was limited to minor editing and formatting support and does not impact the scientific contributions.