BOLTZ2ESI: ACCURATE ENZYME SPECIFICITY PREDICTION WITH CO-FOLDING FOUNDATION MODEL

Xiwei Cheng[1]*, **Seonghwan Seo**[2]*, **Jihang Chen**[1], **Songlin Jiang**[1], **Pengkang Guo**[3], **Wengong Jin**[1]
[1] Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
[2] Department of Chemistry, KAIST, Daejeon, South Korea
[3] École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{cheng.xiw,jih.chen,jiang.song,w.jin}@northeastern.edu,
shwan0106@kaist.ac.kr, pengkang.guo@epfl.ch

ABSTRACT

Predicting enzyme-substrate specificity is a fundamental challenge in biocatalysis and enzyme engineering. However, the scarcity of experimentally labeled data often limits the ability of existing methods to generalize across diverse enzyme families. Here, we present BOLTZ2ESI, an end-to-end framework based on the co-folding foundation model Boltz-2 for accurate enzyme-substrate interaction prediction. Our approach integrates a local spatial representation of the enzyme-substrate interface with corresponding enzyme evolutionary context provided by ESM3. On ESIBank benchmark, BOLTZ2ESI achieves state-of-the-art performance on four out of six evaluation sets, reaching an average AUROC of 0.7309. In particular, our approach shows significant improvements in data-scarce enzyme families, demonstrating robust generalization capabilities. Finally, we conduct extensive ablation studies to analyze the impact of each component on the enzyme-substrate interaction prediction.

1 INTRODUCTION

Enzymatic catalysis is a fundamental biological process, accelerating chemical reactions by several orders of magnitude compared to their spontaneous counterparts (Liu et al., 2024). This remarkable efficiency sustains complex metabolic networks and enables a wide range of industrial applications. Despite their critical importance, many enzymes remain poorly characterized; in particular, our understanding of **enzyme substrate specificity**—the ability to selectively act on target molecules—remains limited (Cui et al., 2025). As the demand for precision enzyme engineering grows, developing reliable methods for predicting substrate specificity is critical for elucidating natural biological functions and accelerating the discovery of novel biocatalysts.

To address this requirement, several deep learning approaches for predicting enzyme-substrate interaction (ESI) have emerged, moving from sequence-based modeling to 3D structure-aware modeling. Initial efforts primarily relied on 1D enzyme protein sequences and 2D substrate molecular graphs to predict interactions (Kroll et al., 2023a). However, these non-structural representations often struggle with generalization, as they fail to capture the precise spatial arrangement and chemical complementarity required for enzymatic catalysis. More recently, the field has shifted towards structure-based approaches (Liu et al., 2024; Cui et al., 2025), facilitated by the availability of high-resolution protein structure prediction tools (Jumper et al., 2021; Baek et al., 2021) and co-folding tools (Abramson et al., 2024; Passaro et al., 2025). Recent advancements further highlight that enforcing strict geometric and chemical constraints is essential for identifying viable catalytic designs (Lauko et al., 2025; Anishchenko et al., 2025). However, a significant challenge remains: while modern co-folding methods provide high-fidelity binding structures, the scarcity of experimentally labeled ESI data limits the ability of these models to fully capture the intricate nuances of catalytic affinity.

Here, we introduce BOLTZ2ESI, a framework for enzyme substrate specificity prediction that leverages the synergistic strengths of structural and evolutionary foundation models. Specifically, we

*Equal Contribution.

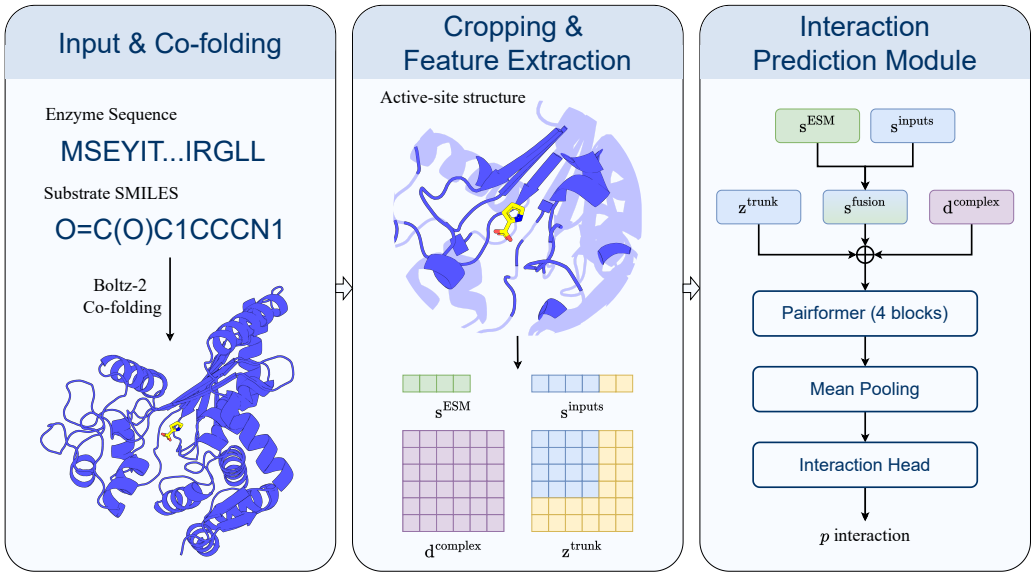

Figure 1: The overview architecture of BOLTZ2ESI. The model utilizes $\mathbf{s}^{\text{inputs}}$ (single representation) and $\mathbf{z}^{\text{trunk}}$ (pair representation), following the notation of AlphaFold3. Additional features include $\mathbf{s}^{\text{ESM}}$, protein ESM3 embedding of active site, and $\mathbf{d}^{\text{complex}}$, the distogram of the 3D binding structure.

integrate Boltz-2 internal latent representations (Passaro et al., 2025) and ESM3 evolutionary features (Hayes et al., 2025) within a predicted 3D active site structure to capture high-fidelity geometric and functional context, as illustrated in Figure 1. By anchoring our model in these rich, pre-trained representations, BOLTZ2ESI incorporates evolutionary priors alongside the strict structural constraints necessary for catalysis. We demonstrate through a comprehensive benchmark that this integration outperforms current state-of-the-art methods, particularly across five diverse enzyme families. In addition, we explore how the individual representations of the enzyme protein, substrate ligand, and 3D binding contribute to overall ESI prediction performance.

2 RELATED WORK

Deep learning has significantly advanced enzyme engineering by enabling predictions across various scales of functionality. At the function level, contrastive learning approaches have been successfully applied to Enzyme Commission (EC) class prediction (Yu et al., 2023b; Sanderson et al., 2023; Kim et al., 2023). While informative, these methods often lack the specificity required to determine whether an enzyme will catalyze a particular substrate. Complementarily, recent models for predicting kinetic parameters, such as k_{cat}, offer quantitative estimates of activity (Li et al., 2022; Yu et al., 2023a; Kroll et al., 2023b; Boorla & Maranas, 2025; Nie et al., 2025). However, these models often rely on condition-dependent measurements that are difficult to standardize across assay conditions.

In contrast, ESI prediction modeling utilizes a more direct and transferable supervision signal by identifying whether a reaction occurs. Early ESI methods like ESP (Kroll et al., 2023a) relied on graph neural networks (GNNs) using 1D sequences and 2D molecular topologies, which limited their ability to capture the 3D spatial complementarity essential for catalysis. Addressing this, EZSpecificity (Cui et al., 2025) addressed this by incorporating structural information. Nonetheless, its reliance on structures derived from AlphaFill (Hekkelman et al., 2023) and conventional docking (Santos-Martins et al., 2021), which are based on homology search and local alignment, can lack the precision needed for fine-grained interaction modeling. While recent work (Lauko et al., 2025) has shown that filtering designs with geometric criteria from PLACER (Anishchenko et al., 2025) can improve catalytic success, such approaches depend on well-defined binding pockets that may not be available for uncharacterized enzymes. This underscores the need for an end-to-end framework capable of capturing high-fidelity structural context directly from foundation models.

3 METHODS

Boltz-2 (Passaro et al., 2025) represents a milestone in AI-based biomolecular interaction prediction, approaching the accuracy of free-energy perturbation (FEP) methods in predicting protein-ligand binding affinities while maintaining significantly higher inference efficiency. This capability underscores the potential of co-folding representations for deciphering complex biological interactions. However, the Boltz-2 affinity module primarily focuses on local spatial relationships within the active site, aiming to correlate structural geometry with the physical binding energy. While effective for general small molecule drug discovery task against a target protein, such a local-centric approach may overlook broader enzyme-specific features critical for enzyme engineering, such as global protein stability or distal allosteric effects. To address these limitations, we integrate evolutionary features atop the Boltz-2 to capture a more holistic representation of the enzyme-substrate interaction. The entire pipeline are detailed in Section A.

3.1 FEATURE PREPARATION

Structure Preparation via Induced-fit Docking. We utilize Boltz-2 (Passaro et al., 2025) for blind docking to capture enzyme-substrate spatial interactions while explicitly accounting for protein flexibility. For each enzyme-substrate pair, five structures are generated using Multiple Sequence Alignments (MSAs) and structure templates. The structure with the highest confidence score is selected to identify the putative active site, which serves as the structural foundation for subsequent modeling stages. To evaluate the impact of active site detection accuracy, we compares flexible docking approach and conventional rigid docking methods in Section 4.3.

Refined Structure Prediction on Active Site. Following the Boltz-2 affinity prediction pipeline, we define an active site by extracting 200 nearest residues to the ligand using Boltz-2 Affinity Cropper. We then generate five pocket-ligand structures using Boltz-2 and select one with the highest interface predicted TM-score (ipTM) to ensure a high-fidelity geometric context.

Feature Extraction. Finally, we extract the internal single input representation $\mathbf{s}^{\text{inputs}}$ and trunk pair representations $\mathbf{z}^{\text{trunk}}$ from Boltz-2, and a distogram $\mathbf{d}$ derived from the selected active site structure. To further enrich the evolutionary context, we compute ESM3 representations (Hayes et al., 2025) for entire enzyme sequence and extract features of active-site residues $\mathbf{s}^{\text{esm}}$ as additional inputs for the ESI prediction module. The effects of this evolutionary context are analyzed in Section 4.2 and Section B.1. We note that the length of the input single representations and trunk pair representations are the same to the sum of the number of pocket residues and the number of ligand atoms.

3.2 ENZYME-SUBSTRATE INTERACTION PREDICTION MODULE.

Architecture. Given these features, we first fuse the input single representation $\mathbf{s}^{\text{inputs}}$ with the evolutionary features $\mathbf{s}^{\text{esm}}$ to form $\mathbf{s}^{\text{fusion}}$, as illustrated in Figure 1. This fused representation is then projected to the same dimension as the trunk pair representations $\mathbf{z}^{\text{trunk}}$ using a broadcasted linear projection applied row-wise and column-wise. Geometric context is then incorporated through an one-hot token-level distogram matrix $\mathbf{d}$ of the predicted active-site structure. The complete implementation is detailed in Algorithm 2.

Training Objective. As is typical for many real-world biological datasets, for enzyme-substrate interaction, positive samples are scarce while negative samples are abundant (see Table 2). To better leverage the limited positive supervisory signals, we sample each mini-batch with a fixed 1:10 positive-to-negative ratio. In addition, we employ focal loss (Lin et al., 2017) to further prioritize difficult samples and mitigate the label imbalance:

$$\mathcal{L}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t). \tag{1}$$

The hyperparameters used in this work are in Section C.2.

4 EXPERIMENTS

We validated BOLTZ2ESI using ESIBank (Cui et al., 2025), a public ESI dataset that aggregates large-scale data from ESIBank (Jeske et al., 2019) and six additional representative enzyme families:

Table 1: Summary of performance on ESIBank test set for enzyme-substrate interaction prediction. We report average AUROC across 4 splits for ESIBank and five enzyme families. Top 2 results are highlighted with **bold** and underlined, respectively. DUF indicates Domain of Unknown Function. We note that Boltz-2 indicates the pre-trained binary head of Boltz-2 affinity module, and additional results with fine-tuning are reported in Section B.1.

Model	ESIBank	Thiolase	Esterase	Phosphatase	Glycosyltransferase	DUF	Average
Boltz-2	0.5764	0.5307	0.5755	0.6219	0.7440	0.5468	0.5992
ESP	0.6778	0.6439	0.5091	0.5921	0.5878	0.4594	0.5784
EZSpecificity	**0.7198**	0.5970	**0.8035**	0.6741	0.7972	0.7067	0.7169
BOLTZ2ESI	0.7132	**0.6621**	0.7816	**0.6826**	**0.8214**	**0.7244**	**0.7309**

Thiolase, Esterase, Phosphatase, Glycosyltransferase, Nitrilases, and DUF (Domain of Unknown Function). First, we compare our framework against baseline models on ESIBank and enzyme families (Section 4.1). We then conduct ablation studies to evaluate: (1) the contribution of enzyme evolutionary features (Section 4.2) and (2) the impact of active site prediction accuracy (Section 4.3). In addition, we investigated whether Boltz-2's input featurizer provides a sufficiently generalized representation of substrate molecules by comparing it with auxiliary ligand representations from Uni-Mol (Zhou et al., 2023) and molecular fingerprints (Morgan, 1965) (Section 4.4). Finally, we compared BOLTZ2ESI against the Boltz-2's affinity module under various training configurations to assess transferability and architectural efficiency (Section B.1).

4.1 COMPARATIVE STUDY ON ESIBANK BENCHMARK

Evaluation Details. ESIBank includes four types of train-test splits: (1) random split; (2) unknown enzyme; (3) unknown substrate; and (4) unknown enzyme and substrate (Cui et al., 2025). In this work, we adopt the most stringent scenario: unknown enzyme and substrate, ensuring both enzymes and substrates in the test set are unseen during training. We evaluate model performance using four-fold cross-validation. We note that Nitrilase is excluded from evaluation, as certain splits for this protein class contain no positive samples in the test set. Details of ESIBank are described in Table 2.

Baselines. We evaluated BOLTZ2ESI against two representative enzyme-substrate interaction prediction models. **Boltz-2** (Passaro et al., 2025) trains an affinity module atop of co-folding representations on large-scale protein-ligand binding affinity database. **ESP** (Kroll et al., 2023a) utilizes a gradient boosting model built upon substrate representations from pre-trained GNNs and enzyme protein embeddings from ESM (Rives et al., 2021). **EZSpecificity** (Cui et al., 2025) is the current state-of-the-art model on ESIBank, which employs a cross-attention assisted SE(3)-equivariant GNN on predicted binding structures to integrate both sequence and structural information.

Results. As shown in Table 1, BOLTZ2ESI achieves state-of-the-art performance on 4 out of the 6 test sets: Thiolase, Phosphatase, Glycosyltransferase, and DUF, and remaining highly competitive on the ESIBank dataset. Notably, BOLTZ2ESI significantly outperforms existing baselines on the Thiolase and Glycosyltransferase families, where the labeled data is limited (550 and 1,019 positive samples, respectively). This underscores the robust generalization capabilities of BOLTZ2ESI, even in data-scarce enzyme families. Such results demonstrate that the model effectively transfers generalized protein-ligand interaction representations ($\mathbf{z}^{\text{trunk}}$), learned from large-scale complex structural databases, to the specialized domain of enzyme-substrate interactions.

4.2 IMPACT OF EVOLUTIONARY INFORMATION

Due to the substantial computational cost of training, all ablation studies were conducted on the first split of the ESIBank cross-validation. While Boltz-2 utilizes MSAs for co-folding, its affinity module employs a restricted MSA profile of the active site pocket to minimize computational overhead and prevent overfitting in standard ligand discovery tasks. However, enzyme engineering necessitates a global sequence perspective to accurately model protein stability, distal allosteric effects, and evolutionary plasticity. To evaluate the contribution of evolutionary context, we compared BOLTZ2ESI against an ablation baseline without ESM embedding ("w/o ESM"). As illustrated in Figure 2(a), removing ESM features causes the AUROC to drop sharply from 0.7164 to 0.6179.

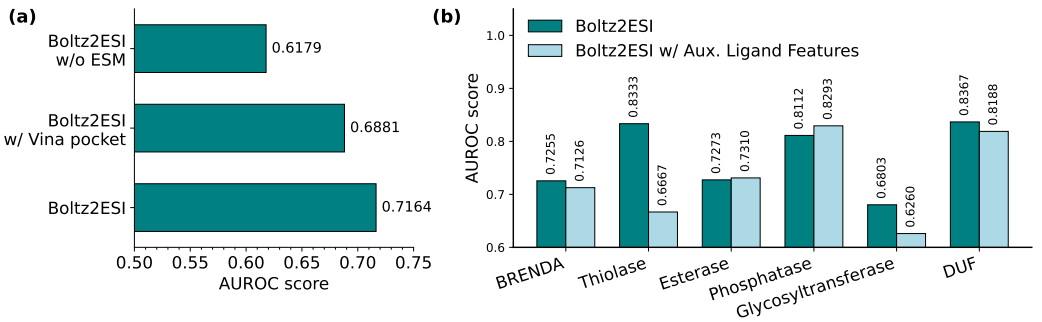

Figure 2: The results shows the performance of the variants of BOLTZ2ESI on the first split. **(a)** Ablation studies evaluating the impact of ESM embeddings and active site extraction strategies (Vina vs Boltz-2) on the ESIBank test set. **(b)** Model performance with auxiliary ligand features.

This performance collapse underscores that evolutionary information captures essential functional semantics and physicochemical properties that local structural features cannot provide.

4.3 IMPACT OF ACTIVE SITE PREDICTION ACCURACY

We further investigated how the fidelity of active site localization impacts model performance. Instead of extracting active site from the Boltz-2 predicted structure, we used conventional docking to obtain a binding structure, denoted as "w/ Vina pocket". The pocket is defined based on the docking pose from AutoDock-GPU (Santos-Martins et al., 2021), where unbound protein structure is predicted by AlphaFold2 (Jumper et al., 2021), with cofactors added by AlphaFill (Hekkelman et al., 2023). As shown in Figure 2(a), the model using the Boltz-2 pocket achieves an AUROC of 0.7164, significantly outperforming the model using the Vina pocket (0.6881). This improvement stems from the active site flexibility. Unlike traditional rigid-body docking, co-folding explicitly captures induced-fit effects and precise spatial arrangements. This higher fidelity ensures a physically consistent and optimized structural context, resulting in more robust interaction classification.

4.4 GENERALIZATION ABILITY OF BOLTZ-2 SUBSTRATE REPRESENTATION

Finally, we analyzed whether the intrinsic ligand representations learned by Boltz-2 are sufficient or if they benefit from auxiliary features. We trained a variant of BOLTZ2ESI that concatenates explicit ligand features: Uni-Mol representations (Zhou et al., 2023) and Morgan fingerprints (Morgan, 1965). Figure 2(b) presents the performance comparison across all enzyme families. Unexpectedly, incorporating these auxiliary features generally leads to a decline in performance or offers no significant gain. This observation suggests that the input featurizer of Boltz-2 already extracts a highly generalized and robust representation of substrate molecules. The addition of external ligand features likely introduces overfitting, hindering the model's ability to learn spatial relationship.

5 CONCLUSIONS

In this work, we introduced BOLTZ2ESI, a framework that leverages co-folding foundation models and evolutionary priors for accurate enzyme substrate specificity prediction. Our results demonstrate that BOLTZ2ESI outperforms current state-of-the-art methods across diverse enzyme families, particularly in data-scarce regimes. Through comprehensive ablation studies, we established that the integration of global evolutionary features from ESM3 and high-fidelity, flexible binding site predictions from Boltz-2 are critical for capturing the intricate nuances of enzymatic specificity.

Despite these advancements, there are still several room for improvement. First, our current structural representation for the enzyme protein is restricted to residue-level information. Atomistic-level detail is often critical for modeling the specific chemical environments required for catalysis; future iterations could incorporate all-atom representations to better reflect these local reaction dynamics. Second, while we currently utilize the Boltz-2 affinity prediction pipeline as a backbone, building the ESI module directly on top of the co-folding trunk could enable the model to leverage MSA and global protein structure information, potentially surpassing the performance gains provided by

external ESM embeddings. Furthermore, BOLTZ2ESI does not yet account for co-factors, which are important to predict the accurate co-folding structure and enzyme-substrate specificity. Future work will integrate atomistic details and enriched evolutionary context to accelerate next-generation biocatalyst design.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Ivan Anishchenko, Yakov Kipnis, Indrek Kalvet, Guangfeng Zhou, Rohith Krishna, Samuel J Pellock, Anna Lauko, Gyu Rie Lee, Linna An, Justas Dauparas, et al. Modeling protein–small molecule conformational ensembles with placer. *Proceedings of the National Academy of Sciences*, 122(45):e2427161122, 2025.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, August 2021. doi: 10.1126/science.abj8754. URL https://www.science.org/doi/10.1126/science.abj8754.

Veda Sheersh Boorla and Costas D Maranas. Catpred: a comprehensive framework for deep learning in vitro enzyme kinetic parameters. *Nature communications*, 16(1):2072, 2025.

Haiyang Cui, Yufeng Su, Tanner J Dean, Tianhao Yu, Zhengyi Zhang, Jian Peng, Diwakar Shukla, and Huimin Zhao. Enzyme specificity prediction using cross attention graph neural networks. *Nature*, pp. 1–3, 2025.

Thomas A Halgren, Robert B Murphy, Richard A Friesner, Hege S Beard, Leah L Frye, W Thomas Pollard, and Jay L Banks. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *Journal of medicinal chemistry*, 47(7):1750–1759, 2004.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.

Maarten L Hekkelman, Ida de Vries, Robbie P Joosten, and Anastassis Perrakis. Alphafill: enriching alphafold models with ligands and cofactors. *Nature Methods*, 20(2):205–213, 2023.

Xin Hong, Bowen Gao, Yinjun Jia, Wenyu Zhu, Qixuan Chen, Xiaohe Tian, Zhenyi Zhong, Jianhui Wang, and Yanyan Lan. How good is alphafold3 at ranking drug binding affinities? *bioRxiv*, pp. 2025–05, 2025.

Wei-Tse Hsu, Savva Grevtsev, Anna M Herz, Thomas Douglas, Aniket Magarkar, and Philip C Biggin. Can ai-predicted complexes teach machine learning to compute drug binding affinity? *Journal of Chemical Information and Modeling*, 65(24):13051–13056, 2025.

Lisa Jeske, Sandra Placzek, Ida Schomburg, Antje Chang, and Dietmar Schomburg. Brenda in 2019: a european elixir core data resource. *Nucleic acids research*, 47(D1):D542–D549, 2019.

Songlin Jiang, Yifan Chen, Ze Cao, and Wengong Jin. Flashaffinity: Bridging the accuracy-speed gap in protein-ligand binding affinity prediction. *bioRxiv*, pp. 2025–12, 2025.

Wengong Jin, Xun Chen, Amrita Vetticaden, Siranush Sarzikova, Raktima Raychowdhury, Caroline Uhler, and Nir Hacohen. Dsmbind: Se (3) denoising score matching for unsupervised binding energy prediction and nanobody design. *bioRxiv*, pp. 2023–12, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray

Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2.

Gi Bae Kim, Ji Yeon Kim, Jong An Lee, Charles J Norsigian, Bernhard O Palsson, and Sang Yup Lee. Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nature Communications*, 14(1):7370, 2023.

James King, Lewis Cornwall, Andrei Cristian Nica, James Day, Aaron Sim, Neil Dalchau, Lilly Wollman, and Joshua Meyers. On fine-tuning boltz-2 for protein-protein affinity prediction. *arXiv preprint arXiv:2512.06592*, 2025.

Alexander Kroll, Sahasra Ranjan, Martin KM Engqvist, and Martin J Lercher. A general model to predict small molecule substrates of enzymes based on machine and deep learning. *Nature communications*, 14(1):2787, 2023a.

Alexander Kroll, Yvan Rousset, Xiao-Pan Hu, Nina A Liebrand, and Martin J Lercher. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. *Nature communications*, 14(1):4139, 2023b.

Anna Lauko, Samuel J Pellock, Kiera H Sumida, Ivan Anishchenko, David Juergens, Woody Ahern, Jihun Jeung, Alexander F Shida, Andrew Hunt, Indrek Kalvet, et al. Computational design of serine hydrolases. *Science*, 388(6744):eadu2454, 2025.

Feiran Li, Le Yuan, Hongzhong Lu, Gang Li, Yu Chen, Martin KM Engqvist, Eduard J Kerkhoven, and Jens Nielsen. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature Catalysis*, 5(8):662–672, 2022.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Yong Liu, Chenqing Hua, Tao Zeng, Jiahua Rao, Zhongyue Zhang, Ruibo Wu, Connor W Coley, and Shuangjia Zheng. EnzymeCAGE: A Geometric Foundation Model for Enzyme Retrieval with Evolutionary Insights, December 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.12.15.628585.

Oscar Méndez-Lucio, Mazen Ahmad, Ehecatl Antonio del Rio-Chanona, and Jörg Kurt Wegner. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3(12):1033–1039, 2021.

Seokhyun Moon, Wonho Zhung, Soojung Yang, Jaechang Lim, and Woo Youn Kim. Pignet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science*, 13(13):3661–3673, 2022.

Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of chemical documentation*, 5(2):107–113, 1965.

Zhiwei Nie, Hongyu Zhang, Hao Jiang, Yutian Liu, Xiansong Huang, Fan Xu, Jie Fu, Zhixiang Ren, Yonghong Tian, Wen-Bin Zhang, et al. Omniesi: A unified framework for enzyme-substrate interaction prediction with progressive conditional deep learning. *arXiv preprint arXiv:2506.17963*, 2025.

Mohit Pandey, Mariia Radaeva, Hazem Mslati, Olivia Garland, Michael Fernandez, Martin Ester, and Artem Cherkasov. Ligand binding prediction using protein structure graphs and residual graph attention networks. *Molecules*, 27(16):5114, 2022.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, 2025.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Theo Sanderson, Maxwell L Bileschi, David Belanger, and Lucy J Colwell. Proteinfer, deep neural networks for protein functional inference. *Elife*, 12:e80942, 2023.

Diogo Santos-Martins, Leonardo Solis-Vasquez, Andreas F Tillack, Michel F Sanner, Andreas Koch, and Stefano Forli. Accelerating autodock4 with gpus and gradient-based local search. *Journal of chemical theory and computation*, 17(2):1060–1073, 2021.

Seonghwan Seo and Woo Youn Kim. Pharmaconet: deep learning-guided pharmacophore modeling for ultra-large-scale virtual screening. *Chemical Science*, 15(46):19473–19487, 2024.

Chao Shen, Xujun Zhang, Yafeng Deng, Junbo Gao, Dong Wang, Lei Xu, Peichen Pan, Tingjun Hou, and Yu Kang. Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer. *Journal of Medicinal Chemistry*, 65(15):10691–10706, 2022.

ByteDance AML AI4Science Team, Xinshi Chen, Yuxuan Zhang, Chan Lu, Wenzhi Ma, Jiaqi Guan, Chengyue Gong, Jincai Yang, Hanyu Zhang, Ke Zhang, et al. Protenix-advancing structure prediction through a comprehensive alphafold3 reproduction. *BioRxiv*, pp. 2025–01, 2025.

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

Izhar Wallach and Abraham Heifets. Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of chemical information and modeling*, 58(5):916–932, 2018.

Han Yu, Huaxiang Deng, Jiahui He, Jay D Keasling, and Xiaozhou Luo. Unikp: a unified framework for the prediction of enzyme kinetic parameters. *Nature communications*, 14(1):8211, 2023a.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023b.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The eleventh international conference on learning representations*, 2023.

A BOLTZ2ESI ARCHITECTURE

A.1 ENTIRE END-TO-END PIPELINE

Algorithm 1 described the detailed algorithm of the entire pipeline.

Algorithm 1 BOLTZ2ESI pipeline

Require: Enzyme protein sequence **seq**, Substrate ligand molecule **mol**

▷ *Stage 1: Predict a binding complex structure.*

$\{\mathbf{xyz}_i\} \leftarrow$ Boltz-2$(\{\mathbf{seq}, \mathbf{mol}\})$ ▷ Select one with top confidence score.

▷ *Stage 2: Extract active site pocket residues.*

$\mathcal{I}^{\text{residue}} \leftarrow$ AffinityCropper$(\{\mathbf{xyz}_i\})$ ▷ Extract pocket residue indices.

$\mathbf{seq}^* \leftarrow \mathbf{seq}[\mathcal{I}^{\text{residue}}]$

▷ *Stage 3: Predict active site structure and extract structural features.*

$\{\mathbf{s}_i^{\text{inputs}}\} \leftarrow$ InputFeatureEmbedder$(\{\mathbf{seq}^*, \mathbf{mol}\})$

$\{\mathbf{s}_i^{\text{trunk}}\}, \{\mathbf{z}_{ij}^{\text{trunk}}\} \leftarrow$ Trunk$(\{\mathbf{s}_i^{\text{inputs}}\})$

$\{\mathbf{xyz}_i\} \leftarrow$ SampleDiffusion$(\{\mathbf{s}_i^{\text{inputs}}\}, \{\mathbf{s}_i^{\text{trunk}}\}, \{\mathbf{z}_{ij}^{\text{trunk}}\})$ ▷ Select one with top ipTM score.

$\{\mathbf{d}_{ij}\} \leftarrow$ ToDistogram$(\{\mathbf{xyz}_i\})$

▷ *Stage 4: Extract evolutionary features.*

$\{\mathbf{s}_i^{\text{esm}}\} \leftarrow$ ESM3$(\mathbf{seq})[\mathcal{I}^{\text{residue}}]$

▷ *Stage 5: Run enzyme-substrate interaction module.*

$p^{\text{interaction}} \leftarrow$ ESIModule$(\{\mathbf{s}_i^{\text{inputs}}\}, \{\mathbf{s}_i^{\text{esm}}\}, \{\mathbf{z}_{ij}^{\text{trunk}}\}, \{\mathbf{d}_{ij}\})$

return $p^{\text{interaction}}$

A.2 ESIMODULE ARCHITECTURE

Algorithm 2 described the detailed algorithm of ESIModule.

Algorithm 2 Enzyme-Substrate Interaction Prediction Module

Require: Input single representations $\mathbf{s}_i^{\text{inputs}}$, ESM3 evolutionary representations $\mathbf{s}^{\text{esm}}$,
Trunk pair representations $\mathbf{z}_{ij}^{\text{trunk}}$, Distogram $\mathbf{d}_{ij}$, Protein mask $\mathbf{m}_i^{\text{prot}}$

$\mathbf{m}_{ij}^{\text{pair}} \leftarrow 1 - \mathbf{m}_i^{\text{prot}} \cdot \mathbf{m}_j^{\text{prot}}$

$\mathbf{s}_i^{\text{fusion}} \leftarrow \text{concat}(\mathbf{s}_i^{\text{inputs}}, \mathbf{s}_i^{\text{esm}})$

$\mathbf{s}_i^{\text{fusion}} \leftarrow \text{MLP}(\mathbf{s}_i^{\text{fusion}})$

$\mathbf{s}_i^{\text{fusion}} \leftarrow \mathbf{m}_i^{\text{prot}} \cdot \mathbf{s}_i^{\text{fusion}} + (1 - \mathbf{m}_i^{\text{prot}}) \cdot \mathbf{s}_i^{\text{inputs}}$

$\mathbf{z}_{ij} \leftarrow \text{LinearNoBias}(\text{LayerNorm}(\mathbf{z}_{ij}^{\text{trunk}}))$

$\mathbf{z}_{ij} \leftarrow \mathbf{z}_{ij} + \text{LinearNoBias}(\mathbf{s}_i^{\text{fusion}}) + \text{LinearNoBias}(\mathbf{s}_j^{\text{fusion}})$

$\mathbf{z}_{ij} \leftarrow \text{PairConditioning}(\mathbf{z}_{ij}, \text{LinearNoBias}(\mathbf{d}_{ij}))$

$\{\mathbf{z}_{ij}\} \leftarrow \text{PairFormerStack}(\{\mathbf{z}_{ij}\}, \text{mask} = \{\mathbf{m}_{ij}^{\text{pair}}\})$

$\mathbf{g} \leftarrow \text{Mean}_{i,j}(\{\mathbf{z}_{ij} \mid i \neq j, \mathbf{m}_{ij}^{\text{pair}} = 1\})$

$p^{\text{interaction}} \leftarrow \text{Softmax}(\text{MLP}(\mathbf{g}))$ $p^{\text{interaction}} \in [0, 1]$

return $p^{\text{interaction}}$

B Additional Results

B.1 Compared to Boltz-2 Affinity Module

The Boltz-2 affinity module was constructed on the co-folding framework to utilize generalized biomolecular interaction patterns. In this section, we investigated whether these interaction patterns–trained on broad protein-ligand binding affinity databases–could be directly transferred to the specialized domain of enzyme-substrate catalysis without incorporating external information, such as ESM features.

To assess this, we first established a baseline using the pre-trained binary classification head of Boltz-2 affinity module ("Boltz-2"). We subsequently evaluated the performance achieved by finetuning the pre-trained module on the ESIBank dataset ("Boltz-2 w/ Finetuning"). As illustrated in Figure 3, these experiments demonstrate that evolutionary information is essential in ESI prediction, as the inclusion of ESM features in the full Boltz2ESI model yielded a substantial increase in AUROC from 0.6054 to 0.7164 on ESIBank test set.

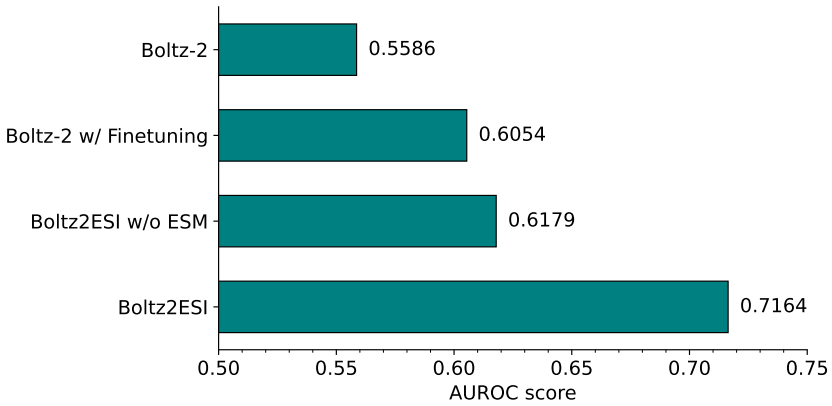

Figure 3: Performance comparison to Boltz-2 affinity module.

C Experimental Details

C.1 Dataset Details

Table 2 summarizes the data distribution for each protein family within the ESIBank dataset across four independent cross-validation splits (Unknown enzyme; unknown substrate scenario). For brevity, only the training and test set sizes are reported, with validation set statistics omitted.

Table 2: The four-fold cross validation splits for enzyme families in ESIBank

			Split 1		Split 2		Split 3		Split 4	
Family	**Class**	**Entire**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**	**Train**	**Test**
Thiolase	Positive	550	355	12	321	26	270	27	294	20
	Negative	545	305	1	284	13	335	12	300	22
Esterase	Positive	3076	1735	103	1704	95	1848	68	1636	117
	Negative	10940	6185	339	6144	347	6072	374	6212	325
Phosphatase	Positive	5424	3152	143	3180	165	2953	223	2954	166
	Negative	30546	17184	959	17032	966	17383	879	17095	1004
Glycosyltransferase	Positive	1019	570	35	574	36	527	37	628	26
	Negative	3995	2283	134	2220	127	2270	122	2207	141
Nitrilase	Positive	85	57	2	57	0	49	5	31	9
	Negative	599	349	19	307	28	357	16	333	19
DUF	Positive	274	175	7	181	7	151	13	111	18
	Negative	2463	1398	77	1392	77	1422	71	1329	98

C.2 Training Details

We trained the proposed model for a total of 10 epochs using the AdamW optimizer with learning of 1×10^{-4} and a weight decay of 1×10^{-5}. Similar to AlphaFold3 (Abramson et al., 2024), learning rate is linearly increased from 0 over the first 500 warm-up steps, followed by a dacay factor of 0.95 every 1000 steps. To train model with multiple data sources, we simply applied dataset weights proportional to the size of each respective dataset. Each training batch was constructed with a ratio of 4 positive samples to 40 negative samples (1 positive sample and 10 negative sample for each GPU). To further address this class imbalance, we trained the model using Focal Loss with hyperparameters $\alpha = 0.8, \gamma = 1.0$. For feature extraction, we followed the default configuration of Boltz-2 official repository[1]. The training procedure was completed within 2 days on 4 NVIDIA A100-80GB GPUs.

Table 3: Training hyperparameters

Hyperparameter	**Value**
Max epoch	10
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-5}
LR scheduler	AlphaFold3
Dataset weights	$\propto$ dataset size
Batch size	4 (positive) + 40 (negative)
Focal loss weights (α, γ)	0.8, 1.0
Num recycles (full)	3
Num diffusion samples (full)	5
Num recycles (active site)	5
Num diffusion samples (active site)	5

[1] https://github.com/jwohlwend/boltz/blob/main/docs/prediction.md

D Additional Related Work

Structure-based Protein-Ligand Interaction Prediction. The prediction of protein-ligand binding affinities has evolved from empirical energy scoring functions (Trott & Olson, 2010; Halgren et al., 2004) to data-driven deep learning architectures. Early learning-based approaches primarily integrated 2D ligand molecular graphs with binding pocket information (Pandey et al., 2022). However, these methods often suffer from limited generalization, tending to memorize structural biases inherent in training datasets rather than learning the underlying physical principles of protein-ligand interactions (Wallach & Heifets, 2018; Seo & Kim, 2024).

To address these limitations, recent advancements have incorporated 3D binding structures to better capture the spatial relationship between protein and ligand (Moon et al., 2022; Jiang et al., 2025). Notably, Méndez-Lucio et al. (2021) and Shen et al. (2022) developed residue-atom distance likelihood potentials derived from statistical distributions, and Jin et al. (2023) introduced SE(3)-equivariant denoising score matching to learn energy landscapes directly from unlabeled crystal structures, thereby mitigating the scarcity of experimentally labeled structural data. More recently, Passaro et al. (2025) constructed a binding affinity prediction module atop a co-folding framework to leverage generalized biomolecular interaction patterns, achieving accuracy competitive with free-energy perturbation (FEP) methods.

Co-folding representation Recent advances in co-folding foundation models have enabled the learning of rich internal representations that encode geometric and chemical interaction patterns, extending their utility beyond structure prediction. Beyond the affinity module in Boltz-2, several recent works have explicitly repurposed co-folding representations for downstream scoring and ranking tasks. AlphaRank (Hong et al., 2025) leverages PairFormer outputs from Protenix (Team et al., 2025) within a lightweight ranking network to compare protein-ligand affinities, demonstrating that co-folding embeddings capture transferable signals useful for discrimination. Boltz-2-PPI (King et al., 2025) showed that incorporating sequence-based embeddings alongside co-folding representations provides complementary information, leading to improved protein–protein interaction affinity prediction. Hsu et al. (2025) suggests that performance gains from using co-folding models to train binding affinity scoring functions critically depend on the quality of the predicted complex structures, highlighting both the promise and limitations of co-folding–based supervision.