# SPRINT Enables Interpretable and Ultra-Fast Virtual Screening against Thousands of Proteomes

**Andrew T. McNutt** *
University of Pittsburgh
anm329@pitt.edu

**Abhinav K. Adduri***
CMU, Arc Institute
abhinav.adduri@arcinstitute.org

**Caleb N. Ellington***
Carnegie Mellon University
cellingt@cs.cmu.edu

**Monica T. Dayao**
Carnegie Mellon University
mdayao@cs.cmu.edu

**Eric P. Xing**
CMU, MBZUAI, Petuum Inc.
epxing@cs.cmu.edu

**Hosein Mohimani**
Carnegie Mellon University
hoseinm@cs.cmu.edu

**David R. Koes**
University of Pittsburgh
dkoes@pitt.edu

## Abstract

Virtual screening of small molecules against protein targets can accelerate drug discovery and development by predicting drug-target interactions (DTIs). However, structure-based methods like molecular docking are too slow to allow for broad proteome-scale screens, limiting their application in screening for off-target effects or new molecular mechanisms. Recently, vector-based methods using protein language models (PLMs) have emerged as a complementary approach that bypasses explicit 3D structure modeling. Here, we develop SPRINT, a vector-based approach for screening entire chemical libraries against whole proteomes for DTIs and novel mechanisms of action. SPRINT improves on prior work by using a self-attention based architecture and structure-aware PLMs to learn drug-target co-embeddings for binder prediction, search, and retrieval. SPRINT achieves SOTA enrichment factors in virtual screening on LIT-PCBA and DTI classification benchmarks, while providing interpretability in the form of residue-level attention maps. In addition to being both accurate and interpretable, SPRINT is ultra-fast: querying the whole human proteome against the ENAMINE Real Database (6.7B drugs) for the 100 most likely binders per protein takes 16 minutes. SPRINT promises to enable virtual screening at an unprecedented scale, opening up new opportunities for *in silico* drug repurposing and development. SPRINT is available on the web as *ColabScreen*: `https://bit.ly/colab-screen`

## Introduction

Virtual screening has emerged as a powerful tool for predicting drug-target interactions (DTIs) and guiding experimental efforts, but conventional structure-based methods like molecular docking are often too slow for proteome-scale analyses [1]. This limitation hinders their application in crucial parts of the drug discovery process such as off-target prediction [2]. The need for scalable and interpretable virtual screening methods is particularly evident in, for example, antimicrobial drug discovery. The rapid emergence of antimicrobial-resistant pathogens poses a severe threat to public health [3], necessitating the development of new antibiotics with novel mechanisms of action to combat cross-resistances [4]. Effective antimicrobial virtual screening demands methods that are not

*These authors contributed equally

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

only fast and scalable but also interpretable, enabling researchers to: 1) identify new drug candidates with on-target effects across thousands of microbial proteomes and minimal off-target effects in humans, and 2) provide interpretations for predicted DTIs and potential mechanisms of action.

Recently, vector-based virtual screening has been proposed as an alternative to structure-based screening to efficiently predict DTIs, leveraging vector featurizations for molecules [5] and sequence models for protein targets [6, 7, 8]. One method, ConPLex [2], proposes co-embedding molecules and proteins into a shared vector space, where the distance between entities is proportional to interaction likelihood. This effectively reduces the task of computing a DTI to a dot product in the co-embedding space, enabling the screening of millions of molecules against the entire human proteome in 24 hours. However, ConPLex does not scale favorably when identifying DTIs across thousands of bacterial and fungal proteomes, and it cannot provide explanations of its DTI predictions. Similarly, DrugCLIP [9, 10] aligns the embeddings of protein pocket structures and ligands with contrastive learning such that similarity encodes the probability of interaction. They demonstrate state-of-the-art (SOTA) results on virtual screening benchmarks with a fraction of the compute time needed for other structure-based virtual screening methods. Their approach is restricted to structures in which binding pockets can be predicted through pocket-detection algorithms or homology-based approaches; however, [11] estimated that almost half of all structured domains may lack obvious pockets in their experimental structures.

In this work, we propose SPRINT (Structure-aware PRotein ligand INTeraction) for fast and accurate vector-based DTI predictions. SPRINT featurizes proteins using SaProt [8], a transformer model trained by augmenting the amino acid vocabulary with discrete structure-tokens [12]. Rather than featurizing proteins by averaging per-residue embeddings from protein sequence models, SPRINT uses a multi-head attention pooling scheme to learn a sequence-dependent aggregation. SPRINT is extremely fast: querying a single protein target against the ENAMINE REAL (6.7B drug) database and predicting its top-100 binders takes 7ms. Our main contributions are summarized as:

- We achieve a new SOTA on DTI (Table 1) and virtual screening benchmarks (Table 2).
- Enabling pan-proteome-scale DTI screens using vector store and retrieval, scaling to billions of molecules.
- Improving molecular property prediction using the molecule co-embeddings learned via predicting DTIs.
- Investigating attention weights and visualizing attention maps to interpret model predictions.

Our software is available on our GitHub repository: `https://github.com/abhinadduri/panspecies-dti` and is also available on the web as *ColabScreen*: `https://bit.ly/colab-screen`.

## Methods

To enable fast and accurate screens, we seek a co-embedded representation of drugs and protein targets where a simple similarity metric indicates binding likelihood. Let $D$ and $T$ denote the random variables representing drugs and targets, $f$ and $g$ denote the choice of frozen drug and target encoders, and $C_d$ and $C_t$ denote modality-specific neural networks that project drug and target embeddings, respectively, into a shared co-embedding space. Let $Y$ denote the random variable representing drug-target interaction, where $Y = 1$ denotes an interacting pair, and $Y = 0$ denotes a non-interacting pair. Denoting latent co-embeddings $Z_d = C_d(f(D))$ and $Z_t = C_t(g(T))$, our model is:

$$P(Y = 1|Z_d, Z_t) = \sigma\left(\alpha \frac{Z_d}{||Z_d||} \cdot \frac{Z_t}{||Z_t||}\right) \tag{1}$$

where $\sigma$ denotes the sigmoid activation function, and $\alpha$ is a constant scaling factor chosen to saturate the range of the sigmoid function, as unscaled cosine similarity ranges from $(-1, 1)$. In our implementation, we choose $\alpha = 5$. Our goal through training is to learn $C_d$ and $C_t$ that minimize binary cross-entropy loss against ground truth binding and non-binding pairs.

For the drug encoder $f$, we use the Morgan fingerprint featurizer available in RDKit with bit length 2048 and radius 2 [13, 5]. For the target encoder $g$, we choose the structure-aware transformer model SaProt [8], a structure-aware protein language model that outputs per-residue embeddings, resulting

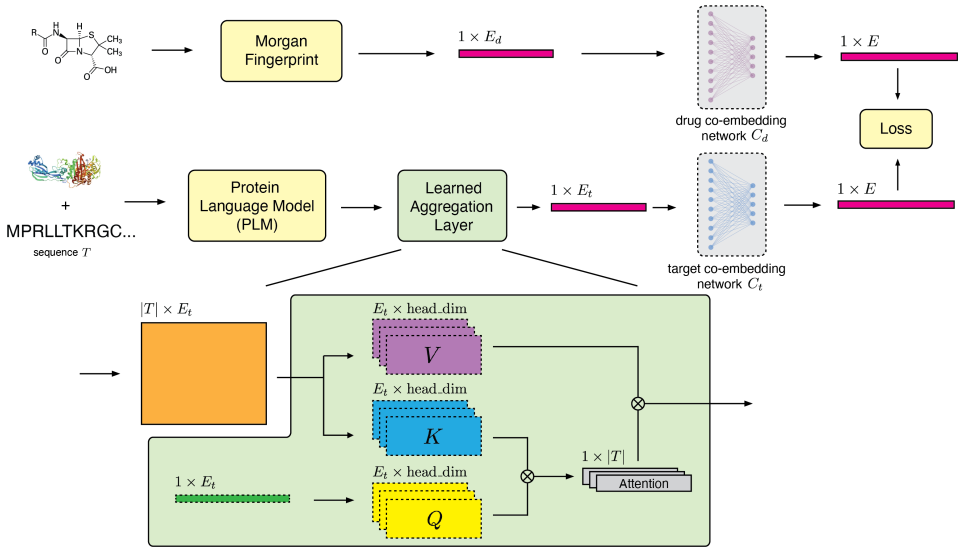



Figure 1: SPRINT learns protein representations via a multi-head attention pooling scheme. Then, SPRINT learns a shared co-embedding space between molecules and protein targets via modality-specific neural networks $C_d$ and $C_t$. The model is trained end-to-end via a binary cross entropy loss on binding and non-binding drug-target pairs, where the probability of interaction is computed as a sigmoid function of the cosine distance between the drug and target embeddings. The learnable parameters of the network are depicted with dashed borders.

Table 1: AUPR on test sets for DTI prediction with co-embedding models across benchmarks (mean $\pm$ std). Train, validation, and test splits for BIOSNAP, BindingDB, and DAVIS are taken from [2]. The MERGED dataset is split by homology (see Appendix C for more details). * indicates that we did not do contrastive training on DUD-E with the ConPLex model.

| Model | Backbone | Pooling | BIOSNAP | DAVIS | BindingDB | MERGED |
|---|---|---|---|---|---|---|
| ConPLex [2] | ProtBert | Avg | 0.883 ± 0.004 | 0.457 ± 0.037 | 0.616 ± 0.009 | 0.414 ± 0.004* |
| ConPLex-attn* | ProtBert | Attn | 0.904 ± 0.005 | 0.493 ± 0.014 | 0.672 ± 0.003 | 0.448 ± 0.018 |
| **SPRINT-xs** (10M) | SaProt | Attn | **0.936** ± 0.001 | **0.507** ± 0.005 | **0.718** ± 0.0004 | 0.481 ± 0.004 |
| **SPRINT-sm** (16M) | SaProt | Attn | 0.858 ± 0.001 | 0.446 ± 0.003 | 0.588 ± 0.0006 | **0.526** ± 0.002 |

in a $|T| \times E$ featurization for an input sequence $T$. SaProt optionally takes protein structure as an input to compute FoldSeek tokens for embedding [12]. We utilize AlphaFold2 [14] predicted structures to generate structure tokens when training the DTI model with SaProt. Unlike prior works, we employ multi-head attention pooling to aggregate these per-residue embeddings into a single vector representation of a protein (Fig. 1). This approach has two merits. First, we hypothesize that the model will be able to focus on information-rich residues due to the data-dependent nature of the attention scheme. Second, we can gain insights into the biological relevance of the attention patterns learned by the model by analyzing which residues are prioritized and how they may relate to known mechanisms of drug-target interaction. Further training details and hyperparameters are provided in Appendix C.

## Results

**Multi-head attention pooling improves DTI prediction.** A limitation of the ConPLex framework is that it averages the per-residue embeddings obtained from PLMs. As much of the relevant signal for DTIs is located in the binding pocket residues, average pooling is prone to noising the contact map information carefully learned by the PLMs through self-attention [15], particularly in the longer sequence length regime. Retraining the ConPLex model with an attention-based, learned aggregation function [16] achieves SOTA predictive scores for DTIs on most benchmarks (Table 1), even when using the same ProtBert model [6].

To see how the learned aggregation scales with the available training data, we trained a SPRINT model on a huge dataset of DTIs, which we refer to as "MERGED", [17] combining DTI data from

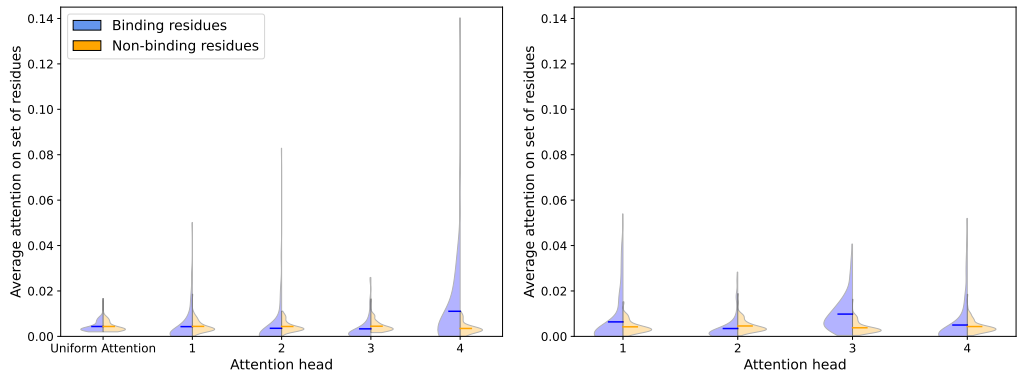
Table 2: Virtual Screening results on LIT-PCBA. SPRINT-ProtBert replaces SaProt with the ProtBert model, and SPRINT-Average replaces learned aggregation with average pooling and additional MLP layers. Parameter counts are shown in parentheses.

| | AUROC (%) | BEDROC (%) | EF | | |
|---|---|---|---|---|---|
| | | | 0.5% | 1% | 5% |
| Surflex [21] | 51.47 | - | - | 2.50 | - |
| Glide-SP [22] | 53.15 | 4.00 | 3.17 | 3.41 | 2.01 |
| Planet [23] | 57.31 | - | 4.64 | 3.87 | 2.43 |
| GNINA [24] | 60.93 | 5.40 | - | 4.63 | - |
| DeepDTA [25] | 56.27 | 2.53 | - | 1.47 | - |
| BigBind [26] | 60.80 | - | - | 3.82 | - |
| DrugCLIP [9] | 57.17 | 6.23 | 8.56 | 5.51 | 2.27 |
| SPRINT-Average (15.7M) | 67.49 | 7.80 | 7.23 | 6.26 | 3.71 |
| SPRINT-ProtBert (13.4M) | **73.4** | 11.9 | 11.68 | 10.19 | 5.27 |
| SPRINT-sm (16M) | **73.4** | **12.3** | **15.90** | **10.78** | **5.29** |

PubChem [18], BindingDB [19], and ChEMBL [20]. In total, the MERGED dataset is comprised of 9,067 unique protein targets and 3,529,822 unique ligands, accounting for 854,118 total positive interactions and 80,681,825 total negative interactions (training details can be found in Appendix C). Our largest model, SPRINT-sm, uses 3-layer MLPs to encode molecules and proteins after multi-head attention pooling, in contrast to SPRINT-xs's single-layer MLPs. SPRINT-sm exhibits overfitting on the BIOSNAP, BindingDB, and DAVIS datasets but significantly improves performance on the much larger MERGED dataset (Table 1), suggesting that there is value in scaling the SPRINT model size as we increase the amount of training data.

LIT-PCBA is a challenging, commonly used virtual screening benchmark that addresses biases in the previously used DUD-E dataset [27] to explicitly enable validation of machine learning models. To evaluate the performance of SPRINT models at virtual screening on LIT-PCBA in the zero-shot setting, we pre-trained the deeper SPRINT-sm (16M) model on the MERGED dataset after removing all protein sequences with $\geq 90\%$ sequence homology to the LIT-PCBA set using MMSeqs2 [28]. We see that the structure-aware SPRINT models significantly outperform competitor methods in AUROC, BEDROC (alpha = 0.85), and across all enrichment factor thresholds (Table 2). The SPRINT models outperform similarly sized models trained using ProtBert featurizations and multi-head attention pooling (SPRINT-ProtBert), and models trained using SaProt featurizations and average pooling (SPRINT-Average) demonstrating the importance of structure and self-attention.

**Structure-aware protein embeddings improve attention maps.** Following training on the MERGED dataset with either ProtBert or SaProt as the PLM backbone for SPRINT, we analyze the attention patterns learned on a set of single-chain protein-ligand binding structures. We find that the models with the greatest enrichment factors on LIT-PCBA, trained with increased negative sampling, have sparse attention which focuses on residues very distant from ligand interactions (Figures S2, S3, and S5). Therefore, we focus our attention analysis on SPRINT models trained with equal positive and negative sampling (LIT-PCBA results are provided in Table S2). All but one of the ProtBert attention heads attend less to the binding residues than the non-binding residues (Figure 2a). By introducing explicit knowledge of the protein's structure with SaProt, we increase the number of heads attending to the binding residues more than the non-binding residues on average (Figure 2b). We visualize the attention on the bound structure of a serine/threonine kinase (PDB ID: 2X4Z) in Figure 3 (additional visualizations provided in Appendix E). Both models have sparse attention maps, with only a handful of residues with non-trivial attention values per head. Attention head 2 of the SaProt model pulls out several residues near the binding site of the kinase while none of the Protbert heads have much, if any, attention on the residues near the binding site. Most of the residues selected by the ProtBert models are on the edges of the protein. We compare the attention patterns to a multiple-sequence alignment (MSA) of 497 human kinase domains from [29] and find that both models attend to non-conserved residues of the kinase which could identify the exact protein with a small residue fingerprint. While the learned aggregation layer allows for model interpretation, we find there is little biological relevance for the attention patterns of the model at its current scale.




(a) ProtBert (b) SaProt

Figure 2: Comparing the average attention weight of binding and non-binding residues on our set of 109 single-chain protein-ligand binding structures after training on the MERGED Dataset (Methodology detailed in Appendix D). We visualize the Protbert and SaProt models trained with equal positive and negative sampling. The horizontal line indicates the average across the proteins. Visualizations of the ProtBert and SaProt models trained with increased negative sampling are in Figure S2).

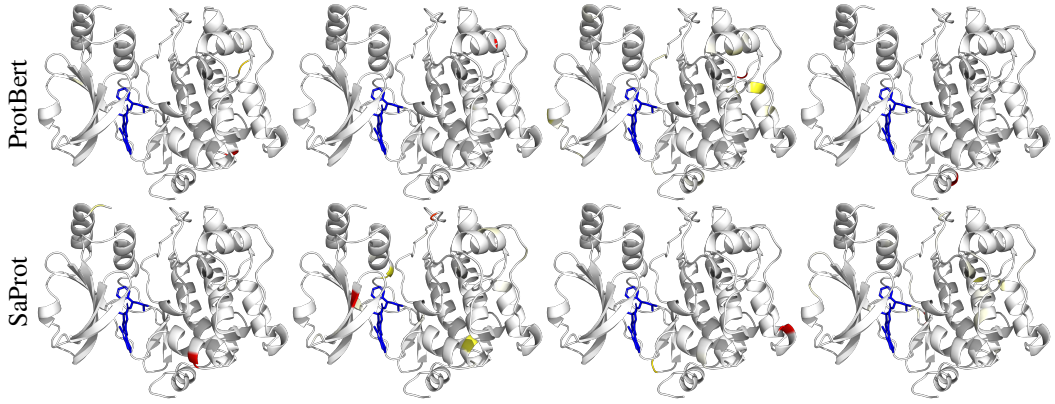



Figure 3: Analyzing the attention on PDB ID 2X4Z using ProtBert and SaProt models trained with equal ratio of positive and negative examples (identical models trained with different initial random seeds visualized in Figures S4 and S6; models trained with increased negative sampling visualized in Figures S3 and S5). Each column is a different attention head. Gradient from white to red indicates the attention weight, where white is no attention and red is max attention for that head. The ligand is shown in blue.

**SPRINT enables querying binding partners from 5132 proteomes.** To demonstrate the utility of SPRINT at the pan-species proteome scale, we constructed a dataset of 5,043 bacterial proteomes, 88 fungal proteomes, and the human proteome, containing 4,291,525 total protein sequences. To store and query the co-embeddings, we use the Chroma vector store[30], a tool developed for semantic search and retrieval-augmented generation in natural language processing. The scaling properties of this framework are highly favorable (Fig. 4): querying a ligand for the 100 most likely binders against the entirety of UniProt (60M sequences) takes 0.0001s, and querying all 2e6 molecules in CHEMBL for each of their 10 most likely binders against the 4.3M proteins in our multi-species dataset takes less than 4 hours. As a proof of concept, we co-visualized several antimicrobials and drugs with their known protein targets across microbial proteomes (Fig. S1).

**Pre-training to predict DTIs improves property prediction.** To benchmark the usefulness of DTI co-embeddings for marginal property prediction, e.g., predicting the properties of a compound only from its molecular graph, we computed SPRINT DTI co-embeddings for several drug-like compounds [31] and natural products [32]. Concatenating the SPRINT-

xs molecule co-embedding to a Morgan fingerprint consistently outperformed an equivalently sized neural network using only the Morgan fingerprint as input (Table S1). However, we observe that using only the SPRINT embedding in these tasks degrades performance, suggesting that SPRINT embeddings can synergistically enhance traditional fingerprints.

## Discussion

Vector-based screens are extremely fast, enabling DTI prediction in regimes that would be impossible with structure-based approaches. We propose SPRINT, which improves on prior work using multi-head attention pooling that scales favorably as we increase the number of DTI training tokens. We also show that structure-aware PLMs like SaProt can confer huge performance gains in virtual screening. Interestingly, we find that the SPRINT models that perform best on the LIT-PCBA virtual screening benchmark, with increased negative sampling, have the least interpretable attention maps. We hypothesize that equal weighting of positive and negative drug-target pairs helps the model learn about residues that interact while increasing the amount of negatives dilutes the information of the positive examples. We demonstrate that SPRINT can perform virtual screening at pan-species proteome scales, e.g., for antimicrobials (Fig. S1). Lastly, we find that predicting DTIs via co-embedding is an effective pre-training strategy that enhances simple molecular property prediction (Table S1).

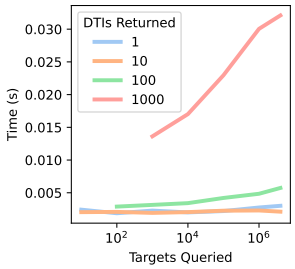



Figure 4: Times for predicting the top DTIs for a ligand using vector search.

We envision SPRINT as a useful benchmarking tool for protein and molecule encoders. Future work will evaluate other structure-aware PLMs, such as MULAN or S-PLM [33, 34], and pre-trained molecule encoders, like UniMol [35, 36], in the SPRINT framework for DTI prediction.

## References


[1] Jocelyn Sunseri and David Ryan Koes. Virtual screening with gnina 1.0. *Molecules*, 26(23):7369, 2021.

[2] 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, June 2023. Publisher: Proceedings of the National Academy of Sciences.

[3] World Health Organization. *Antimicrobial resistance: global report on surveillance*. World Health Organization, Geneva, 2014. Section: xxii, 232 p.

[4] Abigail Colclough, Jukka Corander, Samuel K Sheppard, Sion C Bayliss, and Michiel Vos. Patterns of cross-resistance and collateral sensitivity between clinical antibiotics and natural antimicrobials. *Evolutionary applications*, 12(5):878–887, 2019.

[5] H. 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, May 1965. Publisher: American Chemical Society.

[6] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, April 2022.

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

[8] Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein Language Modeling with Structure-aware Vocabulary, March 2024. Pages: 2023.10.01.560349 Section: New Results.

[9] Bowen Gao, Bo Qiang, Haichuan Tan, Yinjun Jia, Minsi Ren, Minsi Lu, Jingjing Liu, Wei-Ying Ma, and Yanyan Lan. Drugclip: Contrasive protein-molecule representation learning for virtual screening. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Yinjun Jia, Bowen Gao, Jiaxin Tan, Xin Hong, Wenyu Zhu, Haichuan Tan, Yuan Xiao, Yanwen Huang, Yue Jin, Yafei Yuan, et al. Deep contrastive learning enables genome-wide virtual screening. *bioRxiv*, pages 2024–09, 2024.

[11] Artur Meller, Michael Ward, Jonathan Borowsky, Meghana Kshirsagar, Jeffrey M Lotthammer, Felipe Oviedo, Juan Lavista Ferres, and Gregory R Bowman. Predicting locations of cryptic pockets from single protein structures using the pocketminer graph neural network. *Nature Communications*, 14(1):1177, 2023.

[12] Michel Van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature biotechnology*, 42(2):243–246, 2024.

[13] Rdkit: Open-source cheminformatics. https://www.rdkit.org.

[14] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[15] Zhidian Zhang, Hannah K Wayment-Steele, Garyk Brixi, Haobo Wang, Matteo Dal Peraro, Dorothee Kern, and Sergey Ovchinnikov. Protein language models learn evolutionary statistics of interacting sequence motifs. *bioRxiv*, pages 2024–01, 2024.

[16] Hugo Touvron, Matthieu Cord, Alaaeldin El-Nouby, Piotr Bojanowski, Armand Joulin, Gabriel Synnaeve, and Hervé Jégou. Augmenting convolutional networks with attention-based aggregation. *arXiv preprint arXiv:2112.13692*, 2021.

[17] Alex Golts, Vadim Ratner, Yoel Shoshan, Moshe Raboh, Sagi Polaczek, Michal Ozery-Flato, Daniel Shats, Liam Hazan, Sivan Ravid, and Efrat Hexter. A large dataset curation and benchmark for drug target interaction. *arXiv preprint arXiv:2401.17174*, 2024.

[18] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2023 update. *Nucleic acids research*, 51(D1):D1373–D1380, 2023.

[19] Michael K Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research*, 44(D1):D1045–D1053, 2016.

[20] Anna Gaulton, Anne Hersey, Michał Nowotka, A Patricia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J Bellis, Elena Cibrián-Uhalte, et al. The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954, 2017.

[21] Russell Spitzer and Ajay N Jain. Surflex-dock: Docking benchmarks and real-world application. *Journal of computer-aided molecular design*, 26:687–699, 2012.

[22] 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.

[23] X Zhang, H Gao, H Wang, Z Chen, Z Zhang, X Chen, Y Li, Y Qi, and R Wang. Planet: A multi-objective graph neural network model for protein-ligand binding affinity prediction. biorxiv. 2023.

[24] Andrew T McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. Gnina 1.0: molecular docking with deep learning. *Journal of cheminformatics*, 13(1):43, 2021.

[25] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[26] Michael Brocidiacono, Paul Francoeur, Rishal Aggarwal, Konstantin I Popov, David Ryan Koes, and Alexander Tropsha. Bigbind: learning from nonstructural data for structure-based virtual screening. *Journal of Chemical Information and Modeling*, 64(7):2488–2495, 2023.

[27] 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.

[28] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[29] Vivek Modi and Roland L Dunbrack Jr. A structurally-validated multiple sequence alignment of 497 human protein kinase domains. *Scientific reports*, 9(1):19790, 2019.

[30] Chroma - the open-source embedding database. https://www.trychroma.com/.

[31] Anjali Setiya, Vinod Jani, Uddhavesh Sonavane, and Rajendra Joshi. Moltoxpred: small molecule toxicity prediction using machine learning approach. *RSC advances*, 14(6):4201–4220, 2024.

[32] Barbara R Terlouw, Kai Blin, Jorge C Navarro-Muñoz, Nicole E Avalon, Marc G Chevrette, Susan Egbert, Sanghoon Lee, David Meijer, Michael JJ Recchia, Zachary L Reitz, et al. Mibig 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic acids research*, 51(D1):D603–D610, 2023.

[33] Daria Frolova, Marina Pak, Anna Litvin, Ilya Sharov, Dmitry Ivankov, and Ivan Oseledets. Mulan: Multimodal protein language model for sequence and structure encoding. *bioRxiv*, pages 2024–05, 2024.

[34] Duolin Wang, Mahdi Pourmirzaei, Usman L Abbas, Shuai Zeng, Negin Manshour, Farzaneh Esmaili, Biplab Poudel, Yuexu Jiang, Qing Shao, Jin Chen, and Dong Xu. S-plm: Structure-aware protein language model via contrastive learning between sequence and structure. *bioRxiv*, 2024.

[35] 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. 2023.

[36] Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, et al. Uni-mol2: Exploring molecular pretraining model at scale. *arXiv preprint arXiv:2406.14969*, 2024.

[37] Atanas G Atanasov, Sergey B Zotchev, Verena M Dirsch, and Claudiu T Supuran. Natural products in drug discovery: advances and opportunities. *Nature reviews Drug discovery*, 20(3):200–216, 2021.

[38] Akihiro Kishimoto, Hiroshi Kajino, Masataka Hirose, Junta Fuchiwaki, Indra Priyadarsini, Lisa Hamada, Hajime Shinohara, Daiju Nakano, and Seiji Takeda. MHG-GNN: Combination of Molecular Hypergraph Grammar with Graph Neural Network, September 2023. arXiv:2309.16374 [cs].

[39] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-Scale Chemical Language Representations Capture Molecular Structure and Properties, December 2022. arXiv:2106.09553 [cs, q-bio].

[40] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, Nina M Donghia, Craig R MacNair, Shawn French, Lindsey A Carfrae, Zohar Bloom-Ackermann, et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702, 2020.

[41] Mihir Mongia, Mustafa Guler, and Hosein Mohimani. An interpretable machine learning approach to identify mechanism of action of antibiotics. *Scientific Reports*, 12(1):10342, June 2022.

[42] Julian Cremer, Leonardo Medrano Sandonas, Alexandre Tkatchenko, Djork-Arné Clevert, and Gianni De Fabritiis. Equivariant graph neural networks for toxicity prediction. *Chemical Research in Toxicology*, 36(10):1561–1573, 2023.

[43] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.

[44] Lieyang Chen, Anthony Cruz, Steven Ramsey, Callum J Dickson, Jose S Duca, Viktor Hornak, David R Koes, and Tom Kurtzman. Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening. *PloS one*, 14(8):e0220113, 2019.

[45] Jochen Sieg, Florian Flachsenberg, and Matthias Rarey. In need of bias control: evaluating chemical data for machine learning in structure-based virtual screening. *Journal of chemical information and modeling*, 59(3):947–961, 2019.

[46] Michael M Mysinger, Michael Carchia, John J Irwin, and Brian K Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14):6582–6594, 2012.

[47] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[48] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[50] Zhihai Liu, Minyi Su, Li Han, Jie Liu, Qifan Yang, Yan Li, and Renxiao Wang. Forging the basis for developing protein–ligand interaction scoring functions. *Accounts of chemical research*, 50(2):302–309, 2017.

[51] Artem Gazizov, Anna Lian, Casper Alexander Goverde, Sergey Ovchinnikov, and Nicholas F Polizzi. Af2bind: Predicting ligand-binding sites using the pair representation of alphafold2. *bioRxiv*, pages 2023–10, 2023.

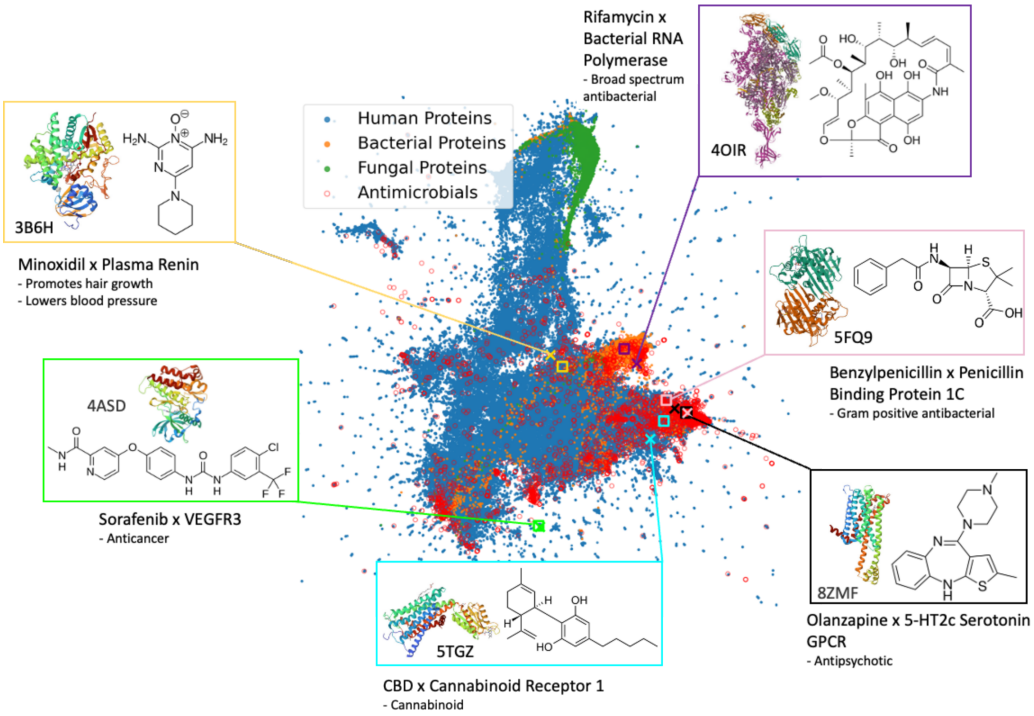



Figure S1: UMAP visualization of the binding co-embedding space of drug-like small molecules and their protein targets across bacterial, fungal, and human proteomes. We see that antimicrobial compounds co-localize with regions of the shared latent space that contain human, bacterial, and fungal proteomes.

## A SPRINT recovers known mechanisms of action

Our pan-species protein dataset is comprised of all predicted protein sequences in reference genomes from NCBI within the taxons bacteria, fungus, and human. Each taxon contained 3,379,854, 775,477, and 136,194 protein sequences respectively. We gathered a list of 3,112 natural products [32], as they are known to have a high prior likelihood for antimicrobial activity [37] and are out-of-distribution relative to our MERGED training dataset. We then co-visualized several antimicrobials and drugs with their known protein targets across microbial proteomes (Fig. S1), recovering several known mechanisms of action. The dataset is available to query at `https://bit.ly/colab-screen`.

## B SPRINT co-embeddings improve property prediction

Current antimicrobial and toxicity screening approaches are often formulated as molecular property predictors, framing antimicrobial activity and toxicity to humans as inherent properties of drug molecules [38, 39, 40, 41, 31, 42]. Our results demonstrate that augmenting Morgan fingerprints with SPRINT-xs ligand embeddings consistently outperformed an equivalently sized neural network using Morgan fingerprints alone, when evaluated on both an antibacterial activity dataset [32] and a toxicity dataset [31] (Table S1). We hypothesize that vectorizing the DTI space allows property prediction methods to leverage information about target neighborhoods around a drug, enhancing performance and offering mechanistic explanations for these properties based on likely binding partners. The embeddings from the deeper SPRINT-sm model consistently performed worse than those from the shallow SPRINT-xs model, suggesting that shallow transformations of the Morgan fingerprint work best in this setting. The standalone SPRINT embeddings achieved substantially lower performance than their concatenated counterparts, indicating that SPRINT embeddings may capture complementary molecular features to traditional fingerprints.

Table S1: F1 scores for MLP classification models applied to molecule embedding strategies (mean ± std). Models trained on Morgan fingerprints used a larger hidden size to match the size of models trained on fingerprints concatenated with embeddings.

| Featurization | Antibacterial Task | Toxicity Task |
|---|---|---|
| Morgan Fingerprint | 0.740 ± 0.027 | 0.720 ± 0.008 |
| SPRINT-xs Embedding | 0.687 ± 0.012 | 0.656 ± 0.015 |
| Morgan Fingerprint + SPRINT-xs Embedding | **0.749** ± 0.016 | **0.735** ± 0.006 |
| SPRINT-sm Embedding | 0.614 ± 0.027 | 0.631 ± 0.023 |
| Morgan Fingerprint + SPRINT-sm Embedding | 0.722 ± 0.027 | 0.701 ± 0.018 |

# C Training details

DTI models are trained using the same train/val/test splits as [2] for the DAVIS, BindingDB, and BIOSNAP datasets. All structure tokens for SaProt were computed on AlphaFold2 [14] generated structures. Structures were downloaded from the AlphaFold Protein Structure Database if they existed. When no precomputed structure was available, ColabFold [43] was run with 2 random seeds to generate 10 energy minimized structures, and the minimized structure with the highest pLDDT was used. Structure tokens were generated with Foldseek[12], masking the structure token if the residue pLDDT was less than 70.

We found that removing the contrastive training on the DUD-E dataset from our reproduced ConPLex model improved its DTI predictive performance. Prior work [44, 45] has shown that the DUD-E dataset [46] has hidden biases that encourage deep learning models to learn drug-only features rather than protein-ligand interactions for determining if a drug is active. Unlike ConPLex, which used DUD-E for contrastive training while using different datasets for binary training, we use the same dataset for binary training and simultaneous contrastive training. We utilize the InfoNCE loss [47, 48], rather than the triplet margin-distance loss employed by ConPLex, to leverage a larger number of negatives per positive example to further constrain the embeddings of the model.

We found that adding InfoNCE to our model decreased performance on all DTI datasets (Table S3), except for the smallest dataset. The InfoNCE loss seems to reduce the overfitting of the SPRINT-sm model on the smaller datasets, but further investigation is needed, as the InfoNCE models performed poorly on downstream tasks like virtual screening. Therefore, we only train our models with the binary cross entropy loss (2) after computing the probability of binding via the sigmoid of the cosine similarity between the protein and drug embeddings (3). Specifically, the loss $\mathcal{L}$ is written as

$$\mathcal{L} = \frac{1}{N}\sum_{i}^{N}\left[Y^i \log(\tilde{Y}^i) + (1 - Y^i)\log(1 - \tilde{Y}^i)\right] \tag{2}$$

$$\tilde{Y}^i = P(Y^i = 1 | Z_d^i, Z_t^i) = \sigma\left(\alpha \frac{Z_t^i}{\left\|Z_t^i\right\|} \cdot \frac{Z_d^i}{\left\|Z_d^i\right\|}\right) \tag{3}$$

where the protein, $T^i$, and drug, $D^i$, have been mapped to the SPRINT co-embedding space as $Z_t^i$ and $Z_d^i$, respectively. $Y_i \in \{0, 1\}$ is a ground-truth label with value 1 if $T^i$ and $D^i$ are binders or 0 if they are non-binders. The pre-sigmoid scalar value, $\alpha$, is used to expand the range of cosine-similarity to the domain of the sigmoid. We set $\alpha$ to 5.

ConPLex models are trained with the hyperparameters used in the original paper [2]. Our Attention Pooling models are trained with a learning rate of $1 \times 10^{-5}$ and a dropout value of 0.05 for 250 epochs, keeping all other hyperparameters the same as ConPLex training. The model checkpoint with the highest validation AUPR during training is evaluated on the test set (Table 1).

Table S2: LIT-PCBA evaluation ablation of negative sampling. '1:1' indicates equal sampling of positive and negative examples during training and '3:1' indicates the preferred model training with 3 negatives sampled for every positive example.

| Model | AUROC | BEDROC ($\alpha = 0.85$) | EF (0.5%) | EF (1%) | EF (5%) |
|---|---|---|---|---|---|
| SPRINT-ProtBert 1:1 | 71.53 | 7.78 | 6.81 | 5.87 | 3.86 |
| SPRINT-sm 1:1 | 72.71 | 10.16 | 10.31 | 8.86 | 4.73 |
| SPRINT-ProtBert 3:1 | **73.4** | 11.9 | 11.68 | 10.19 | 5.27 |
| SPRINT-sm 3:1 | **73.4** | **12.3** | **15.90** | **10.78** | **5.29** |

Table S3: AUPR values for DTI prediction with co-embedding models across benchmarks (mean $\pm$ std) ablating the use of InfoNCE. These preliminary results use a pre-sigmoid scalar value of 1 rather than 5.

| Model | Backbone | Pooling | BIOSNAP | DAVIS | BindingDB | MERGED |
|---|---|---|---|---|---|---|
| SPRINT-xs (10M) | SaProt | Attn | **0.910** ± 0.003 | 0.453 ± 0.007 | **0.677** ± 0.005 | 0.841 ± 0.003 |
| SPRINT-xs **+InfoNCE** | SaProt | Attn | 0.894 ± 0.002 | **0.499** ± 0.006 | 0.635 ± 0.010 | 0.816 ± 0.008 |
| SPRINT-sm (16M) | SaProt | Attn | 0.850 ± 0.002 | 0.425 ± 0.001 | 0.560 ± 0.008 | 0.844 ± 0.013 |
| SPRINT-sm **+InfoNCE** | SaProt | Attn | 0.900 ± 0.003 | 0.489 ± 0.004 | 0.605 ± 0.004 | **0.856** ± 0.007 |

To enable efficient training on the MERGED dataset, we featurize the unique proteins and molecules before training, storing their representations in memory-mapped files for quick retrieval using the Lightning Memory-Mapped Database Manager (LMDB) library. We use PyTorch Distributed Data Parallel (DDP) for huge improvements in training speed [49]. To address data imbalance in the binding data, for each epoch, we train using all of the drug-target binding pairs, and subsample an equivalent number of non-binding pairs without replacement. We observed that models trained with more negatives than positives (at a 3:1 ratio), achieved better virtual screening performance, but had less interpretable attention patterns (Tables 1, S2). Models are trained for 20 epochs. All other hyperparameters were kept the same.

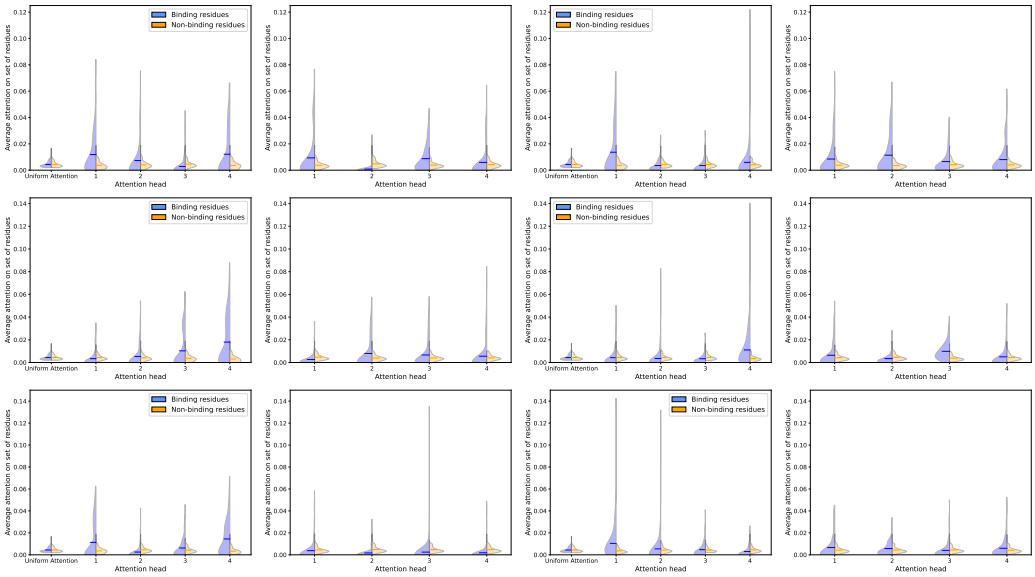
The MERGED dataset splits were determined by clustering protein sequences using MMSeqs2 [28] at 80% coverage threshold and 70% sequence identity, meaning two sequences appear in the same cluster if at least 80% of residues are aligned with at least 70% identity. Clusters were then assigned to splits by size, with smaller clusters preferentially assigned to test and validation sets until each contained approximately 10% of the total number of unique proteins. The remaining sequences were assigned to training. Drug-target interactions were then partitioned according to their protein assignments. The final training, validation, and test sets contained 79.5%, 10.3%, and 10.2% of total interactions, respectively.

Code and data to reproduce our DTI models are on our GitHub repository `https://github.com/abhinadduri/panspecies-dti`.

## D Investigating the learned aggregation layer

We investigate the attention pattern of our learned aggregation layer and compute the relative weighting of binding and non-binding residues. For this analysis, we use the intersection of the PDBbind refined v.2019 [50] dataset and the dataset created by [51]. The intersection of these datasets provides 109 single-chain, high-quality protein-ligand binding structures with annotated binding sites. Following the same protocol as [51], we determine binding residues based on a maximum heavy-atom distance of 5 Å between the residue and the ligand.

We first analyzed the attention patterns of each head in the learned aggregation layer to determine if any of the heads were selective for binding residues. We calculate the attention scores for each residue




(a) ProtBert (b) SaProt (c) ProtBert-1to1 (d) SaProt-1to1

Figure S2: Comparing the average attention weight of binding and non-binding residues on our set of 109 single-chain protein-ligand binding structures after training on the MERGED Dataset. The horizontal line indicates the average across the proteins. Each row is a different random seed and each column is a different PLM or different training regime, where '1to1' indicates that a 1:1 positive to negative sampling ratio was used during training.

in the protein and compute the mean attention value for the binding residues and the non-binding residues. Figure 2 shows the average weight of the binding residues and non-binding residues across the dataset for the ProtBert and SaProt models trained with equal positive and negative sampling. We find the models trained with equal positive and negative sampling have more interpretable attention maps despite a decreased performance on the LIT-PCBA benchmark compared to models trained with more negative samplings. We compare the attention of all the models in Figure S2, visualizing three otherwise identical models trained with different initial random seeds. Interestingly, the PLM with the worst performance on the LIT-PCBA benchmark, ProtBert, shows the most attention to binding site residues relative to its attention to non-binding site residues. The structure-aware PLM, SaProt, has two seeds that attend to binding residues more than non-binding residues across most of the attention heads and one seed that pays very little attention to the binding residues. The SaProt seed that has the least attention for binding residues as compared to non-binding residues performs the best on the LIT-PCBA benchmark. Across all PLMs, there is a large variance in the attention to binding residues as the models initial random seed is changed.

We visualize the learned aggregation layers attention heads on the protein-ligand structures for both ProtBert and SaProt models in Figure 3 with additional visualizations provided in Appendix E and on our github.

## E Structural visualizations

We visualize the attention patterns of the attention pooling layer on the protein-ligand bound structure of several PDB IDs. We compare the attention patterns of SPRINT-sm models with ProtBert and SaProt trained on the MERGED dataset. We see across these diverse proteins and ligands that on average, the SaProt model attends to residues closer to the ligand, while the ProtBert models often attend to residues far from the binding site.

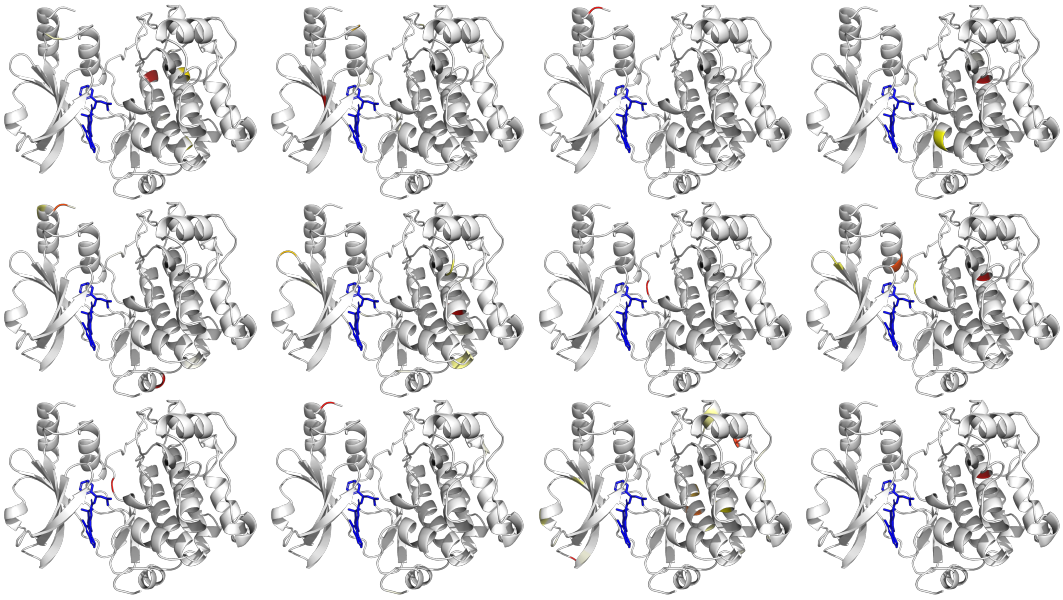



Figure S3: Analyzing the attention of the ProtBert model on PDB ID 2X4Z. Each row is the ProtBert model trained with different seed. Each column is a different attention head. Gradient from white to red indicates the attention weight, where white is no attention and red is max attention for that head. The ligand is shown in blue.

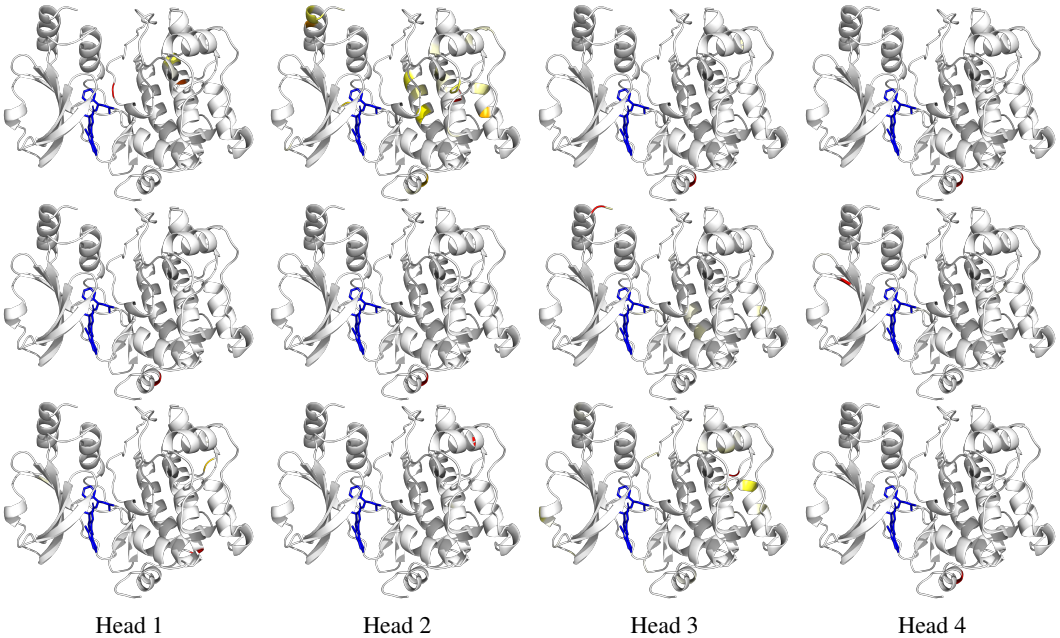



Figure S4: Analyzing the attention of the ProtBert-1to1 model (trained with 1:1 positive to negative ratio) on PDB ID 2X4Z. Each row is the ProtBert model trained with different seed. Each column is a different attention head. Gradient from white to red indicates the attention weight, where white is no attention and red is max attention for that head. The ligand is shown in blue.

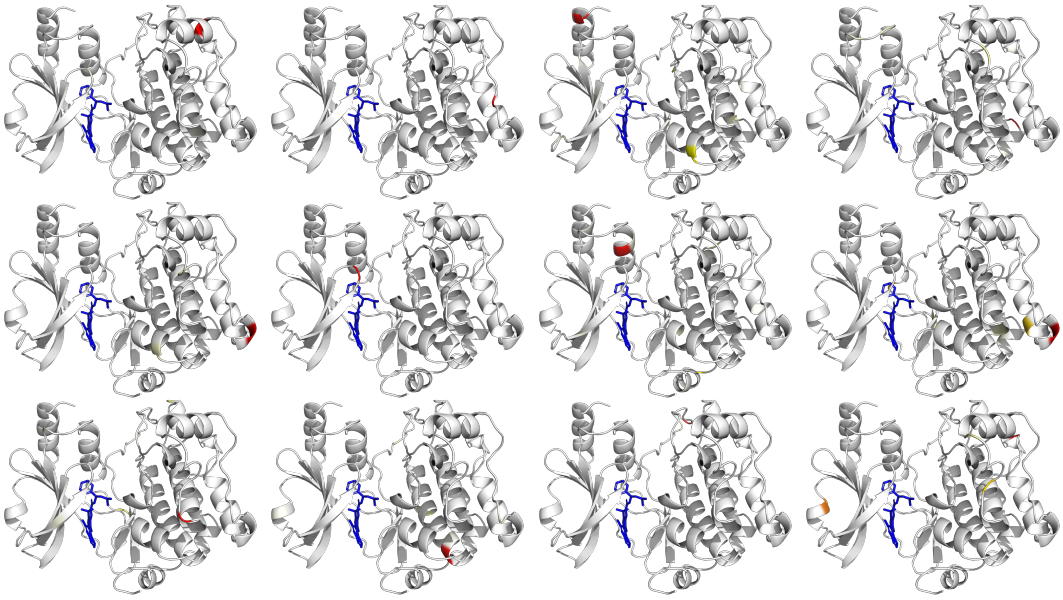



Figure S5: Analyzing the attention of the SaProt model on PDB ID 2X4Z. Each row is the SaProt model trained with different seed. Each column is a different attention head. Gradient from white to red indicates the attention weight, where white is no attention and red is max attention for that head. The ligand is shown in blue.

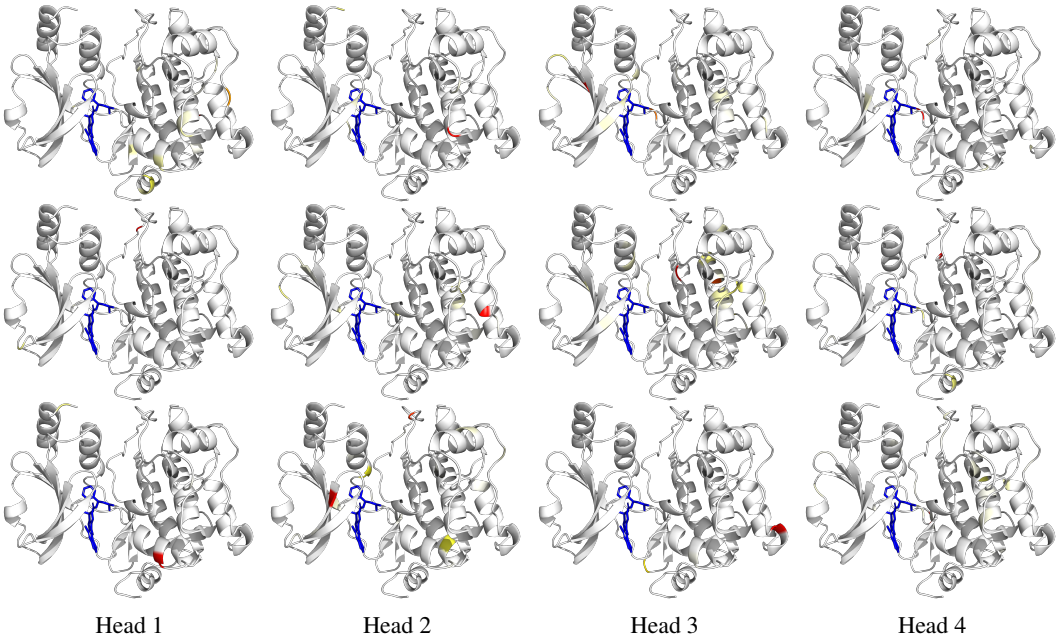



Figure S6: Analyzing the attention of the SaProt-1to1 model (trained with 1:1 positive to negative ratio) on PDB ID 2X4Z. Each row is the SaProt model trained with different seed. Each column is a different attention head. Gradient from white to red indicates the attention weight, where white is no attention and red is max attention for that head. The ligand is shown in blue.