LEARNING REPRESENTATIONS FOR DRUG TARGET FISHING

Anonymous authors
Paper under double-blind review

EXTENDED ABSTRACT

INTRODUCTION

Identifying the interactions a small molecule makes with different proteins is an important task in biology and a critical component of drug discovery. Retrieving the list of protein targets for a small molecule, a task often referred to as target fishing in the literature Galati et al. (2021), is especially challenging and important when little is known about the molecule and its biological activity is being explored. Before experimental testing, other methods must be used to narrow the sheer number of protein candidates. Recent machine learning based methods for biological representations have shown strong performance for related modalities and tasks, including joint protein and small molecule representation learning and binding affinity prediction Gao et al. (2024); Lee et al. (2022); Gorantla et al. (2024); Singh et al. (2023); McNutt et al. (2024). In this paper, we explore the application of several common protein and small molecule representations and learning methods to the task of target fishing. We develop a novel dataset designed to reflect practical scenarios for target fishing and compare the performance of different combinations of multimodal representations and contrastive learning techniques, including molecular docking as a domain specific baseline. We find in our preliminary work that although standard approaches to joint representation learning for proteins and small molecules may work to distinguish protein and small molecule binding affinities, they struggle to order protein targets for small molecules in their latent space and perform poorly on ranking protein targets unseen during training.

RELATED WORK

The success of machine learning representations for protein structure Abramson et al. (2024), protein sequence ESM Team (2024), and chemistry Chithrananda et al. (2020) have inspired the development of joint representations for proteins and small molecules Gao et al. (2024); Lee et al. (2022); Gorantla et al. (2024); Singh et al. (2023). By combining representations, these methods seek to capture the protein and small molecule interactions necessary for binding prediction. Often, these methods leverage contrastive learning objectives. For example, ConPLex Singh et al. (2023) uses triplet loss to learn drug-target interactions with binary labels and protein language model features. Other methods, including DrugCLIP Gao et al. (2024) and UniCLIP Lee et al. (2022), aim to predict binding affinities with the InfoNCEvan den Oord et al. (2018) objective. These methods also use binary labels and structural representations for proteins. Alternatively, models such as BALM Gorantla et al. (2024) attempt to directly predict binding affinities as a regression task.

However, binarizing affinity data results in the loss of important rank differences between targets for a given small molecule. Additionally, the InfoNCE objective when applied to protein and small molecule binding affinities, implicitly assumes all non-recorded affinity labels are non-binders, which may also inhibit protein ranking performance. Alternatively, enforcing the direct prediction of protein-small molecule binding affinities, may restrict ranking models too sharply, given that target fishing only requires proper retrieval of targets and not low absolute error.

DATASET

To investigate target fishing, we develop a new benchmark dataset of over 1 million unique protein-small molecule binding affinities, for 1,476 unique protein targets, and over 900,000 unique small molecules. To simulate practical applications of target fishing, where the investigated small

molecules are often very different from studied molecules and may interact with poorly studied proteins, we clustered and split our molecules between our training and validation set at 0.3 Tanimoto similarity. Likewise, we held out protein targets clustered and split at 30% sequence similarity.

METHODS

Informed by prior work, we investigate two sets of representation modalities for target fishing. Our two-modality ML method uses a molecule-level ChemBERTa Chithrananda et al. (2020) embedding and a protein-level ESM-C embedding. Our four-modality ML method also receives chemical descriptors and an atomic-level graph of the protein pocket which passes through a Graph Neural Network. All representations are then concatenated and fed through multi-layer perceptrons respectively to get the final molecule and protein embeddings.

We train our models with InfoNCE loss and a new "Hybrid" Margin Rank Loss. Hybrid Margin Rank Loss tries to balance learning by calculating a margin rank loss on the ranking of binding affinity across matched examples, using the experimental labels, with a margin rank loss that encourages ranking measured small molecule pairs higher than unmatched examples. Additional details are listed in the Appendix.

RESULTS

We report results on two holdout sets: an "easy" set where the proteins are in the model's train set but the ligands are held out, and a "hard" set where both the ligands *and* proteins are held out. We report mean average precision (mAP) and normalized discounted cumulative gain (NDCG), and compare our predictions to random rankings. We also compare to Smina Koes et al. (2013), a molecular docking algorithm, as a physics-based and non-machine learning baseline.

Table 1: Model Metrics on Held Out *Ligands Only* (Easy)

Model	Loss Function	mAP ↑	NDCG ↑
ChemBERTa/ESM-C	InfoNCE	0.46	0.57
GNN/ESM-C/ChemBERTa/Fingerprint	InfoNCE	0.48	0.59
GNN/ESM-C/ChemBERTa/Fingerprint	Hybrid Margin Rank	0.39	0.52
Random	-	0.05	0.21

Table 2: Model Metrics on Held Out *Ligands and Proteins* (Hard)

Model	Loss Function	mAP ↑	NDCG ↑
ChemBERTa/ESM-C	InfoNCE	0.11	0.28
GNN/ESM-C/ChemBERTa/Fingerprint	InfoNCE	0.11	0.28
GNN/ESM-C/ChemBERTa/Fingerprint	Hybrid Margin Rank	0.11	0.28
Smina	-	0.23	0.38
Random	-	0.07	0.23

On the easy split, the ML methods significantly outperform the random baseline. However, on the hard split, the ML methods barely outperform the random baseline, and significantly underperform Smina, indicating that the ML methods do not generalize to unseen protein-ligand complexes. These results highlight the need for new architectures and loss functions for improved performance in the target fishing problem.

MEANINGFULNESS STATEMENT

A meaningful representation of life is one that captures complex interactions and dependencies across biological modalities. We believe that building a representation of proteins and small-molecule ligands that captures their interactions (including interactions that may not be experimentally measured) will be a significant boost in understanding selectivity, off-target interactions, and drug repurposing. However, in this paper, we find that simply using common machine learning

strategies such as contrastive loss and ranking loss do not capture the complexity of these interactions, and that this area of research should be explored with more detail.

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*, pp. 1–3, 2024.

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

ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning, 2024. URL https://evolutionaryscale.ai/blog/esm-cambrian.

Salvatore Galati, Miriana Di Stefano, Elisa Martinelli, Giulio Poli, and Tiziano Tuccinardi. Recent advances in in silico target fishing. *Molecules*, 26(17):5124, 2021.

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.

Rohan Gorantla, Aryo Pradipta Gema, Ian Xi Yang, Álvaro Serrano-Morrás, Benjamin Suutari, Jordi Juárez Jiménez, and Antonia SJS Mey. Learning binding affinities via fine-tuning of protein and ligand language models. *bioRxiv*, pp. 2024–11, 2024.

David Ryan Koes, Matthew P Baumgartner, and Carlos J Camacho. Lessons learned in empirical scoring with smina from the csar 2011 benchmarking exercise. *Journal of chemical information and modeling*, 53(8):1893–1904, 2013.

Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. *Advances in Neural Information Processing Systems*, 35:1008–1019, 2022.

Andrew T McNutt, Abhinav K Adduri, Caleb N Ellington, Monica T Dayao, Eric P Xing, Hosein Mohimani, and David R Koes. Sprint enables interpretable and ultra-fast virtual screening against thousands of proteomes. *arXiv preprint arXiv:2411.15418*, 2024.

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.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

A APPENDIX

A.1 DATASET CURATION AND PREPROCESSING

To complete ligand preprocessing, we first merged all entries between ChEMBL and BindingDB. We dropped all ligands that had a molecular weight below 100 Daltons or above 1,000 Daltons, or any ligands with SMILES string entries that could not be processed by RDKit. We then filtered binding affinity measurements from each dataset to reduce the number of mislabeled samples in our training data; see APPENDIX for more information. After processing we took each remaining ligand SMILES entry and used Schrodinger's LigPrep tool to generate relevant tautomers and ionization states for each input, at physiological pH. Each of these newly generated ligands was assigned the same binding protein and affinity label as its source ligand SMILES.

To complete protein preprocessing, we sampled up to 50 holoprotein structures for each protein, as identified by its UniProt ID. We attempted to repair any malformed PDB files by replacing

missing residues or heavy atoms, addeed hydrogens at physiological pH with PropKA, and centered the binding pocket of the protein based on the bound ligand in the structure at the origin. Any PDB structures failing preprocessing were discarded, leaving approximately 35,000 PDBs.

A.2 HYBRID MARGIN RANK LOSS

In a batch, we calculate the total loss as the sum of the single measurement loss and the pair measurment loss: $L_s = \frac{1}{n^2}(\sum_n^i \sum_n^j 1(f(m_i, p_i) - f(m_i, p_j)) + margin_s)$ in contrast to the more common margin rank loss applied on our measured pairs: $L_p = \frac{1}{n^2}(\sum_n^i \sum_n^j max(0, y - (f(m_i, p_i) - f(m_j, p_j)) + margin_p)$ In which n is the batch size, f is the similarity function between the learned molecule and protein embedding (m_i, p_i) and proteins and molecules in which the index are the same, are pairs that have had an experimental measurement. $margin_s$ and $margin_p$ are two parameters that need to be tuned. We chose a larger margin for the L_s to avoid collapse of the embedding space.