Challenges in Leveraging Functional Information to Evaluate Predicted Protein-Ligand Interactions

Joseph G. Wakim, Jose Manuel Marti, Jonathan E. Allen, Adam T. Zemla[†]

Lawrence Livermore National Laboratory, Livermore, CA, USA
[†]Corresponding Author: zemla1@llnl.gov

Abstract

Protein-ligand interactions (PLIs) are fundamental to the efficacy and toxicity of drugs, and predicting these interactions with computational models can accelerate drug development. Given an uncharacterized protein and its predicted structure, putative interactions with ligands can be identified based on structural alignment with known binding pockets. However, the accuracy of these predictions depends on the reliability of the protein model. Functional information offers an observable comparator for evaluating predicted PLIs. Yet, existing methods for embedding protein function cluster proteins inconsistently; for the same protein pairs, their relative distances in a functional latent space can vary depending on the embedding method used. To assess challenges in scoring protein function similarity, we evaluate similarity scores using benchmarks that label protein pairs based on shared attributes. For example, we consider benchmarks that label proteins based on shared localization or disease associations, where positive examples share the attribute and negative examples do not. For each benchmark, we quantify how well popular protein representations differentiate between the positive and negative groups. We then demonstrate an innovative framework for leveraging functional similarity scores to characterize drug selectivity and evaluate predicted PLIs. We show that our function-based evaluations remain limited by uncertainty in similarity scores. Overall, we demonstrate the critical need for more reliable similarity-scoring metrics and present a framework for their use in evaluating predicted PLIs during computational drug development.

While there are more than 250,000,000 known protein sequences, there are only around 230,000 experimentally solved protein structures, and the gap between sequences and experimental structures is continuing to grow [1, 2, 3]. To address this disparity, computational models such as Rosetta and AlphaFold2 have been developed to efficiently predict protein structures from sequence by capturing the principles that underlie protein folding [4, 5, 6]. By narrowing the data gap between sequence and structure, these models have enabled new applications like computational drug design [7, 8, 9], in which protein-ligand interactions (PLIs) may be predicted from modeled binding pockets [10, 11]. However, given the limited number of experimental structures and the growing reliance on computational tools, there is an urgent need for new methods to validate predictions.

Protein functions offer observable features with which to evaluate structural predictions. Since protein function depends heavily on structure, we expect proteins with similar domains to share similar functions. In computational drug design, we expect proteins targeted by a selective drug to share similar binding pockets and, therefore, share similar functions [10]. The Gene Ontology (GO) provides a controlled vocabulary for describing protein functions [12, 13]. Comparing the GO annotations of two proteins offers valuable insights into their functional similarities [14]. However, given the semantic complexity of GO terms and the subjectivity with which they are assigned, quantifying functional similarity remains difficult, and function-based evaluations are underutilized.

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

In this study, we survey methods for quantifying protein function similarity, highlighting their advantages and limitations when used to evaluate predicted PLIs. We begin by describing several methods for embedding protein function, revealing that functional similarity scores can vary dramatically with the choice of embedding method. To quantify this variability and rank embedding methods, we evaluate benchmarks that label protein pairs based on shared attributes, and we compare how well different embedding types cluster related proteins. We then introduce a novel framework for aggregating pairwise similarity scores and comparing sets of proteins. Using this framework, we identify selective drugs and evaluate predicted PLIs. Our results highlight inconsistencies across embedding methods, emphasizing the critical need for more robust functional similarity scoring methods.

Comparing Methods for Scoring Protein Similarity

We compare several methods for quantifying protein function similarity. We begin by representing proteins with embeddings from the following deep learning and foundation models: GO2Vec [14], TransformerGO [15], ProteInfer [16], ProtT5 [17], and Evolutionary Scale Modeling Cambrian (ESMC) [18]. For ProteInfer, we derive embeddings directly from the GO model, referred to as "ProteInfer (GO)." We also form concatenated embeddings from the GO, Enzyme Commission (EC) number, and Pfam models, collectively referred to as "ProteInfer (All)." GO2Vec and TransformerGO embed functions on a per-GO-term basis. Since proteins are represented by sets of GO terms, we use a variant of the modified Hausdorff distance (MHD) to quantify similarity between protein pairs (see Eq. 1) [14]. Meanwhile, ProteInfer, ProtT5, and ESMC generate one embedding per protein; for these embedding types, we use cosine similarity to compare protein pairs. In all cases, greater scores correspond to more similar proteins. In addition to using embeddings in their native dimensions, we separately fit principal component analysis (PCA) models for each embedding type to generate 32-dimensional compressed embeddings. We form an "Ensemble" representation by concatenating PCA-compressed ProteInfer (All), ProtT5, and ESMC embeddings. See B.1 in the Appendix for more details on protein similarity scoring methods.

The similarity score for a pair of proteins is highly dependent on the embedding method used (see Sec. C for details). However, without specific criteria, protein similarity is an abstract concept and it is difficult to determine which similarity scoring methods are reliable. Therefore, we turn to the `Geno-Prot` benchmark to compare similarity scoring methods [19]. We focus on seven benchmarks that group human proteins based on the following shared attributes: **(1)** Pfam domains, **(2)** subcel-

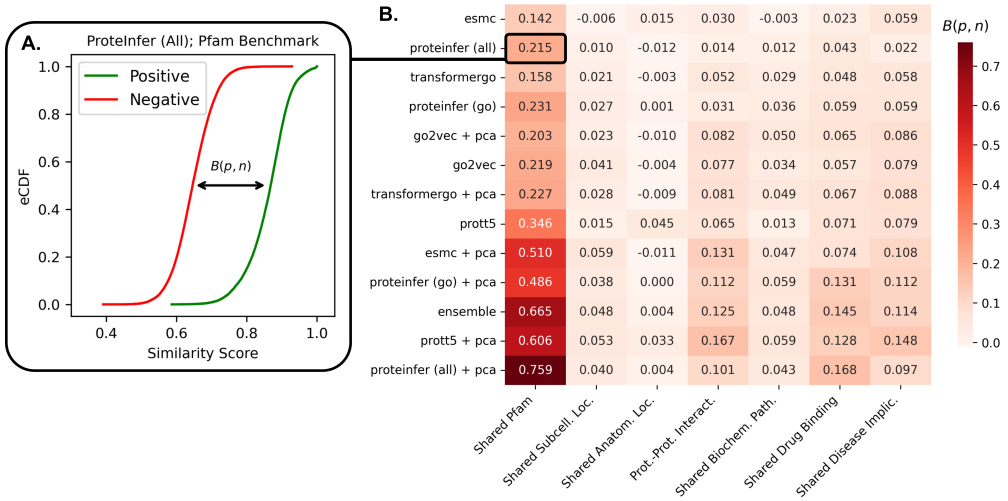

Figure 1: **Benchmark Scores. (A)** The benchmark score quantifies the difference in mean similarity scores between positive and negative groups. To demonstrate this, we plot the distributions of similarity scores assessed by "ProteInfer (All)" for pairs of proteins sharing a Pfam label and pairs that do not. The distributions are plotted as empirical cumulative distribution functions (eCDFs). **(B)** We report benchmark scores for all similarity scoring methods and seven `Geno-Prot` datasets. Rows are ordered by average performance across benchmarks, from least (top) to greatest (bottom).

lular locations, **(3)** anatomical locations, **(4)** protein-protein interactions, **(5)** associations with biochemical pathways, **(6)** drug binding, and **(7)** disease implications. The `Geno-Prot` dataset for each attribute includes positive examples of protein pairs that share the attribute and negative examples of protein pairs that do not. Using each similarity scoring method, we separately generate similarity-score distributions for the positive and negative examples. Assuming that positive examples should typically have higher similarity scores than negative examples, we compute a "benchmark score" that quantifies the difference between the positive and negative distributions (see Sec. B.2 and Eq. 2). Accordingly, greater *benchmark scores* suggest a stronger ability to distinguish similar and dissimilar pairs. Fig. 1 plots the *benchmark scores* associated with each similarity scoring method. In general, PCA-compressed models tend to outperform their uncompressed counterparts. Our findings highlight a need for more robust methods of embedding proteins and scoring similarity.

Evaluating Predicted PLIs Based on Functional Information

Using modeled protein structures [4, 5], we predict interactions with ligands based on structural alignment to known binding pockets (see B.3 in the Appendix). We assume that a modeled protein with similar structure to a binding pocket in a known protein-ligand complex is capable of binding the same ligand [10]. However, the modeled structures carry uncertainty and require validation to increase confidence in the predicted interactions. Using models that embed proteins based on their functions, we evaluate predicted PLIs with orthogonal information.

Our evaluation framework involves three distributions of similarity scores: a "within-group" distribution that compares protein pairs reported to bind the same ligand, a "random" distribution that compares 100,000 random pairs of human proteins, and a "query" distribution that compares predicted and known targets of a ligand (see Fig. 2). First, we quantify ligand selectivity with a "selectivity score," defined as the mean difference between the *within-group* and *random* distributions (see B.4.1 and Eq. 3). Selective ligands target proteins sharing similar functions; their targets should tend to have higher similarity scores than random protein pairs, producing positive *selectivity scores*. These ligands are strong candidates for using functional information to evaluate predicted PLIs.

We then consider each "query protein" predicted to interact with a functionally selective ligand. We evaluate the prediction using the associated *within-group* and *query* distributions, defining a "misalignment score" by the difference in their means (see B.4.2 and Eq. 4). The *misalignment score* reflects how distinct the functions of the query protein are, relative to those of known targets. If the query protein has a high *misalignment score*, then its functions provide evidence against the predicted interaction. Since our metrics for selectivity and misalignment depend on the similarity scoring method used, we report average values derived from three function-specific methods: "GO2Vec + PCA," "TransformerGO + PCA," and "ProteInfer (All) + PCA." We weight each method's contribution to the average by its *benchmark scores* (see B.4.3).

Fig. 7 in the Appendix shows the distribution of *selectivity scores* among ligands in a dataset of reported PLIs from ChEMBL [20, 21]. We identify 14 ligands with strong selectivity for protein functions. We then use PDBspheres to predict 5,061 interactions between human proteins (with

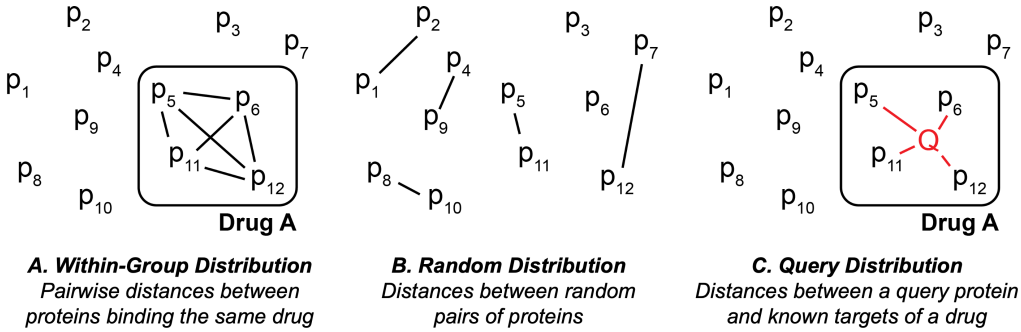

Figure 2: Schematic depicting pairwise similarity scores forming the **(A)** *within-group*, **(B)** *random*, and **(C)** *query* distributions for arbitrary drug A, proteins $p_1,\ p_2,\ \ldots,\ p_{12}$, and query protein Q.

available embeddings) and these ligands [10], and we evaluate these predicted interactions by calculating their *misalignment scores*. Fig. 3 illustrates the distribution of *misalignment scores* associated with our predicted PLIs. Among predicted interactions, the figure separates those that are also reported in ChEMBL from those that are novel. Novel interactions with low *misalignment scores* are supported by both structural and functional evidence and represent potential leads for future drug studies. For validation, we also iteratively evaluate reported interactions as if they were predictions using a leave-one-out approach (see B.4.4). As expected, reported interactions tend to be associated with low *misalignment scores*, supporting the validity of our framework.

Our framework for evaluating predicted PLIs would benefit from more reliable methods of scoring protein function similarity. When we apply several models for embedding protein function, we find that predicted PLIs are evaluated differently. For example, we find that *misalignment scores* obtained based on "GO2Vec + PCA" and "ProteInfer (GO) + PCA" embeddings have a correlation coefficient of just 0.52 (see Fig. 10). To circumvent inconsistencies in the scoring methods, we propose using an average *misalignment score* weighted by the performance of each model on the `Geno-Prot` benchmarks (see B.4.3). More reliable function-based evaluations of predicted PLIs would come from additional, robust function-specific representations of the proteins involved.

For ligands with strong selectivity, we expect the functions of their predicted protein targets to align with those of their known targets. However, more generally, a low *misalignment score* does not necessarily imply that the associated PLI will occur. For instance, our method is ill-suited for detecting low-quality predictions associated with promiscuous ligands. Since these ligands target a broad range of protein functions, the associated *within-group* distributions tend to be low, and even random query proteins can generate low *misalignment scores*; this could cause high false-negative rates when detecting low-quality interactions. Therefore, predicted PLIs should always be evaluated in the context of associated ligand selectivity.

The success of our method also depends on the quality and completeness of the reference data. When predicting interactions, we assume that the structural complexes in the Protein Data Bank (PDB) are accurate [2]. When evaluating those predictions, we assume that interactions reported by Heinzke *et al.* accurately represent the range of protein targets [20]. For models that embed proteins based on their GO terms, we also assume that protein annotations are accurate and comprehensive. If these assumptions are invalid, then we cannot reliably characterize the functions targeted by a ligand and use those functions in our evaluation. With more orthogonal descriptors of ligands and their targets, we can further improve our confidence in predicted interactions. Despite its limitations, this work introduces functional information as an additional descriptor for assessing PLIs, which enhances confidence over purely structural predictions.

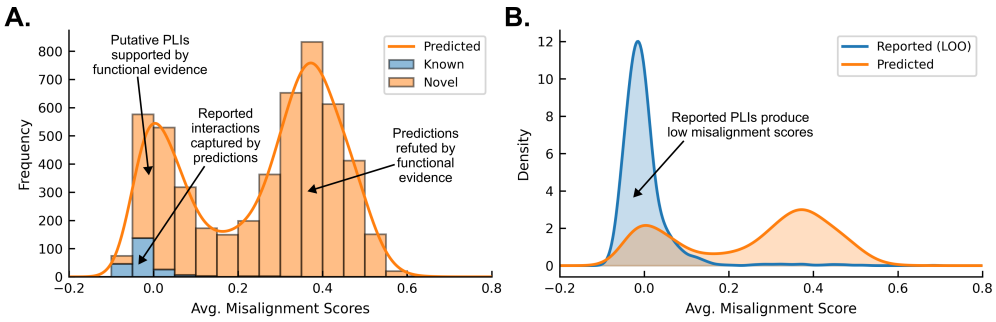

Figure 3: **Misalignment Scores. (A)** We plot the distribution of misalignment scores for predicted PLIs. Bar colors distinguish predictions that represent novel PLIs from those that capture reported interactions in ChEMBL. Our framework distinguishes putative PLIs with low misalignment scores, which may confer off-target effects or serve as candidates for drug repurposing, from those with high misalignment scores, which may be filtered based on functional information. **(B)** For comparison, we evaluate the distribution of misalignment scores associated with reported interactions using a leave-one-out (LOO) approach (see B.4.4). Since reported interactions have low misalignment scores, we gain some confidence in our method. Misalignment scores are computed as weighted averages of the values obtained from "GO2Vec + PCA," "TransformerGO + PCA," and "ProteInfer (All) + PCA" methods, with weights determined by each method's benchmark scores (see B.4.3).

In addition to evaluating predicted PLIs, our work can be used to indirectly evaluate protein structure models. For targets of functionally selective ligands, proteins with similar binding pockets should share similar functions. If we assume that reference structures are accurate, proteins are well-annotated, and similar binding pockets indeed enable similar interactions, then high *misalignment scores* could indicate errors in predicted protein structures. In this way, we can score predicted PLIs as a proxy for evaluating structural models. By offering orthogonal features for assessing structural models and predicted PLIs, functional information may promote more successful drug design and efficient drug discovery.

Summary

In this study, we develop a framework for leveraging functional information about proteins to assess ligand selectivity and evaluate predicted PLIs. We first identify 14 selective ligands that target proteins sharing similar functions, based on reported interactions. Then, using structural information, we predict 5,061 PLIs involving these ligands and human proteins. By our definition of selectivity, predicted targets of selective ligands should be functionally similar to known targets. Accordingly, we evaluate predicted PLIs involving selective ligands by comparing the functions of predicted and known targets. We finally distinguish sets of predicted PLIs that are corroborated and refuted by functional information. Our approach demonstrates the use of orthogonal evidence to predict and evaluate PLIs, mitigating biases that are introduced by a single modality.

However, our framework is constrained by a critical challenge: the evaluation of protein similarity. The assessed similarity of a protein pair depends strongly on how the proteins are embedded. We compare several models that generate protein embeddings, using benchmarks that group proteins by shared attributes. When quantifying selectivity and evaluating PLIs, we combine scores from multiple deep learning and foundation models, weighting each model's contribution according to its benchmark performance. Despite this provisional solution, our work demonstrates an open need for more reliable protein similarity scoring methods. Improved metrics for quantifying protein similarity will enhance detection of unlikely PLIs based on functional evidence. This would support applications like computational drug design where high-quality predictions are critical.

Acknowledgments

This manuscript has been authored by Lawrence Livermore National Security, LLC under Contract No. DE-AC52-07NA27344 with the U.S. Department of Energy. This material is based upon work supported by the Department of Energy, Office of Science, Office of Advance Scientific Computing Research. The United States Government retains, and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. Release Number: LLNL-CONF-2010834.

References

[1] The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, January 2025.

[2] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, January 2000.

[3] Robin Pearce and Yang Zhang. Toward the solution of the protein structure prediction problem. *Journal of Biological Chemistry*, 297(1), July 2021. Publisher: Elsevier.

[4] John Jumper and et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. Publisher: Nature Publishing Group.

[5] Mihaly Varadi and et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, January 2022.

[6] Kim T. Simons, Charles Kooperberg, Enoch Huang, and David Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions11edited by f. e. cohen. *Journal of Molecular Biology*, 268(1):209–225, 1997.

[7] Mark A. Lindsay. Target discovery. *Nature Reviews Drug Discovery*, 2(10):831–838, October 2003. Publisher: Nature Publishing Group.

[8] Jaeyoung Ha, Hankum Park, Jongmin Park, and Seung Bum Park. Recent advances in identifying protein targets in drug discovery. *Cell Chemical Biology*, 28(3):394–423, March 2021. Publisher: Elsevier.

[9] Caterina Vicidomini and Giovanni N. Roviello. Protein-Targeting Drug Discovery. *Biomolecules*, 13(11):1591, October 2023.

[10] Adam T Zemla, Jonathan E Allen, Dan Kirshner, and Felice C Lightstone. PDBspheres: a method for finding 3D similarities in local regions in proteins. *NAR Genomics and Bioinformatics*, 4(4):lqac078, December 2022.

[11] Xiliang Zheng, LinFeng Gan, Erkang Wang, and Jin Wang. Pocket-Based Drug Design: Exploring Pocket Space. *The AAPS Journal*, 15(1):228–241, November 2012.

[12] Michael Ashburner and et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. Publisher: Nature Publishing Group.

[13] The Gene Ontology Consortium and et al. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, May 2023.

[14] Xiaoshi Zhong, Rama Kaalia, and Jagath C. Rajapakse. GO2Vec: transforming GO terms and proteins to vector representations via graph embeddings. *BMC Genomics*, 20(9):918, February 2020.

[15] Ioan Ieremie, Rob M Ewing, and Mahesan Niranjan. Transformergo: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms. *Bioinformatics*, 38(8):2269–2277, 02 2022.

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

[17] Michael Heinzinger, Konstantin Weissenow, Joaquin Gomez Sanchez, Adrian Henkel, Milot Mirdita, Martin Steinegger, and Burkhard Rost. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 6(4):lqae150, 11 2024.

[18] ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning, 2024.

[19] Joseph G. Wakim, Vinayak Gupta, Jose Manuel Marti, Jonathan E. Allen, Brian Bartoldson, and Bhavya Kailkhura. Benchmarking biomolecular foundation models for cross-modal genomics-proteomics. In *NeurIPS 2025 Workshop on Multi-modal Foundation Models and Large Language Models for Life Sciences*, 2025.

[20] A. Lina Heinzke, Barbara Zdrazil, Paul D. Leeson, Robert J. Young, Axel Pahl, Herbert Waldmann, and Andrew R. Leach. A compound-target pairs dataset: differences between drugs, clinical candidates and other bioactive compounds. *Scientific Data*, 11(1):1160, October 2024. Publisher: Nature Publishing Group.

[21] Barbara Zdrazil, Eloy Felix, Fiona Hunter, Emma J Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, Harris Ioannidis, David Mendez Lopez, Juan F Mosquera, Maria Paula Magarinos, Nicolas Bosc, Ricardo Arcila, Tevfik Kizilören, Anna Gaulton, A Patrícia Bento, Melissa F Adasme, Peter Monecke, Gregory A Landrum, and Andrew R Leach. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, January 2024.

[22] Thin Nguyen, Hang Le, Thomas P Quinn, Tri Nguyen, Thuc Duy Le, and Svetha Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 10 2020.

[23] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35(2):309–318, 07 2018.

[24] Adiba Yaseen, Imran Amin, Naeem Akhter, Asa Ben-Hur, and Fayyaz Minhas. Insights into performance evaluation of compound–protein interaction prediction methods. *Bioinformatics*, 38(Supplement_2):ii75–ii81, 09 2022.

[25] Aurélien F.A. Moumbock, Jianyu Li, Pankaj Mishra, Mingjie Gao, and Stefan Günther. Current computational methods for predicting protein interactions of natural products. *Computational and Structural Biotechnology Journal*, 17:1367–1376, 2019.

[26] Kristy A. Carpenter and Russ B. Altman. Databases of ligand-binding pockets and protein-ligand interactions. *Computational and Structural Biotechnology Journal*, 23:1320–1338, 2024.

[27] Charlotte Sweeney, Nele P Quast, Fabian C Spoendlin, and Yee Whye Teh. Estimating protein flexibility via uncertainty quantification of structure prediction models. 2024. Paper in Machine Learning in Structural Biology Workshop.

[28] R Prabakaran and Y Bromberg. Quantifying uncertainty in protein representations across models and task. *bioRxiv*, 2025.

Technical Appendices for:

"Challenges in Leveraging Functional Information to Evaluate Predicted Protein-Ligand Interactions"

A Related Works

Nguyen *et al.* and Tsubaki *et al.* use machine learning models to predict PLIs based on protein sequences and ligand structures [22, 23]. One critical limitation in this approach is the lack of experimentally verified negative examples (*i.e.*, non-interacting protein-ligand pairs) [24]; while observed PLIs are typically reported, non-interacting protein-ligand pairs are often unreported in the literature. PDBspheres is an extensive collection of curated protein-ligand structures taken from the PDB. In this work, we use PDBspheres to predict PLIs from structural evidence, adopting the mechanistic hypothesis that similar protein binding pockets enable similar interactions [10]. Although our minimal model neglects additional environmental factors affecting PLIs, grounding predictions in mechanistic evidence reduces the dependence on negative examples.

Moumbock *et al.* discuss similar target-based methods for predicting interactions, and Carpenter and Altman review related databases of predicted PLIs derived from ligand-binding pockets [25, 26]. These works highlight the broad dependency on modeled protein structures to generate predictions during drug development. Given uncertainty in structural models, evaluating predicted PLIs is critical.

One approach for assessing uncertainty in predicted PLIs is to directly quantify confidence in the structural models. Sweeney *et al.* demonstrate that confidence metrics for structural models can be correlated with the flexibility of a protein's structure [27], which can affect a protein's ability to interact with a ligand. AlphaFold2 reports confidence in predicted protein structures using metrics generated by auxiliary prediction heads [4]. However, since these prediction heads are trained on experimental data, they are limited by artifacts, biases, and noise in the experiments.

Prabakaran and Bromberg report a model-agnostic framework for evaluating predicted protein structures based on their underlying embeddings [28]. The group generates a dataset containing both biological and synthetic, non-biological sequences. They find that uncertainty in predicted protein structures correlates with the quality of the underlying embeddings; sequences with embeddings that resemble the non-biological examples tend to produce structural models with higher predicted uncertainty. The work links the biological relevance of a protein sequence to the confidence of its structural model. While our work shares the use of protein embeddings to assess predictions, we focus on applications in drug discovery. Specifically, we evaluate predicted PLIs using protein functions, which are often obtained independently from structure and serve as "orthogonal" descriptors.

B Detailed Methods

We first introduce several methods for scoring protein function similarity, then show that the similarity scores generated by different methods tend to be inconsistent (see Sec. B.1). In response to these findings, we compare different similarity scoring methods using benchmarks provided by Ref. [19]. For each of seven attributes (*e.g.*, Pfam labels, localization, disease associations), the benchmarks group protein pairs into those sharing the attribute ("positive group") and those not sharing the attribute ("negative group"). Using the benchmarks, we quantify how well each similarity scoring method distinguishes protein pairs in the positive and negative groups (see Sec. B.2). Protein function offers an orthogonal comparator for evaluating predicted PLIs obtained from structural evidence; for selective ligands affecting a narrow range of protein functions, predicted targets should have functions that align with known targets. We predict PLIs based on structure-based evidence (see Sec. B.3), then demonstrate how functional similarity scores can be used to quantify confidence in the predictions (see Sec. B.4).

B.1 Generating Embeddings and Scoring Protein Similarity

To score the similarity of protein pairs, we represent the proteins with vector embeddings and quantify the similarities between them. We consider variants of three methods for embedding protein function: GO2Vec, TransformerGO, and ProteInfer [14, 15, 16]. By embedding proteins based on function alone, we can provide an unbiased evaluation of predicted PLIs obtained from structure-based methods. GO2Vec and TransformerGO embed protein functions on the basis of individual GO terms. ProteInfer generates embeddings on a per-protein basis, based on three types of functional annotations: GO terms, EC numbers, and Pfam labels. We consider two forms of ProteInfer embeddings: those generated solely by ProteInfer's GO model, and those concatenating embeddings from ProteInfer's GO, EC, and Pfam models.

For comparison, we also embed proteins using more holistic protein models, including ProtT5 and ESMC. These models encode proteins based on sequence, structure, function, and evolutionary lineage [17, 18]. For ProtT5, we load precomputed protein embeddings reported by UniProt [1]. ESMC generates embeddings for each amino acid in a protein sequence; to embed proteins using ESMC, we use the pre-trained 300-million-parameter model and perform mean pooling of the amino-acid embeddings.

The models we consider generate embeddings of varying length. We compare models before and after compressing the embeddings to a common dimension of 32, using PCA for dimensionality reduction. We fit separate PCA models to GO embeddings for GO2Vec and TransformerGO, and to protein embeddings for ProteInfer, ProtT5, and ESMC. Since ProteInfer (All), ProtT5, and ESMC all generate embeddings on a per-protein basis, we also test how concatenating their compressed embeddings affects protein similarity scores. We refer to the concatenated embeddings with the label "Ensemble."

Since proteins may be annotated with several GO terms, they may also be represented by sets of GO embeddings. For GO2Vec and TransformerGO, we score the functional similarity of each protein pair by computing a variant of the MHD between their sets of GO embeddings. For two proteins, p_1 and p_2, MHD is given by Eq. 1:

$$\text{MHD}(p_1, p_2) = \min \left\{ \frac{1}{N_{p_1}} \sum_{\mathbf{v}_i \in \mathbf{V}_{p_1}} \max_{\mathbf{v}_j \in \mathbf{V}_{p_2}} \cos(\mathbf{v}_i, \mathbf{v}_j), \frac{1}{N_{p_2}} \sum_{\mathbf{v}_j \in \mathbf{V}_{p_2}} \max_{\mathbf{v}_i \in \mathbf{V}_{p_1}} \cos(\mathbf{v}_j, \mathbf{v}_i) \right\} \quad (1)$$

where N_{p_i} and $\mathbf{V}_{p_i}$ denote the number and set of GO embeddings for protein p_i, respectively, $\mathbf{v}_i$ and $\mathbf{v}_j$ denote individual embeddings in the sets $\mathbf{V}_{p_1}$ and $\mathbf{V}_{p_2}$, respectively, and the function $\cos(\cdot)$ denotes cosine similarity between two embeddings. Since MHD is computed with cosine similarity, greater MHDs indicate greater similarity [14]. For ProteInfer, ProtT5, ESMC, and our Ensemble model, which all generate embeddings on a per-protein basis, we score functional similarity according to a cosine similarity metric. In all cases, greater similarity scores indicate more similar proteins. Table 1 summarizes the methods that we use for quantifying protein similarity.

Table 1: **Summary of protein embedding models and methods for scoring similarity.** In the table, "ProteInfer (GO)" refers to embeddings obtained from ProteInfer's GO model, while "ProteInfer (All)" refers to concatenating embeddings from ProteInfer's GO, EC, and Pfam models. The "+ PCA" suffix indicates embeddings compressed by PCA. "Ensemble" refers to the concatenation of compressed ProteInfer (All), ProtT5, and ESMC embeddings.

Model	Embed. Dim.	Basis	Scoring Function	Modality
Original Embeddings				
GO2Vec	128	GO term	MHD	Function
TransformerGO	64	GO term	MHD	Function
ProteInfer (GO)	1100	Protein	Cos. Sim.	Function
ProteInfer (All)	3300	Protein	Cos. Sim.	Function
ProtT5	1024	Protein	Cos. Sim.	Holistic
ESMC	960	Protein	Cos. Sim.	Holistic
PCA-Compressed Embeddings				
GO2Vec + PCA	32	GO term	MHD	Function
TransformerGO + PCA	32	GO term	MHD	Function
ProteInfer (GO) + PCA	32	Protein	Cos. Sim.	Function
ProteInfer (All) + PCA	32	Protein	Cos. Sim.	Function
ProtT5 + PCA	32	Protein	Cos. Sim.	Holistic
ESMC + PCA	32	Protein	Cos. Sim.	Holistic
Ensemble	96	Protein	Cos. Sim.	Holistic

B.2 Benchmarking Similarity Scores

We find that similarity scores are largely dependent on the embedding method used. To determine the most reliable method for scoring protein similarity, we use benchmark datasets that label protein pairs based on shared attributes. Specifically, the benchmarks assign binary labels to protein pairs according to: **(1)** shared Pfam domains, **(2)** shared subcellular localizations, **(3)** common anatomical localizations, **(4)** reported protein-protein interactions, **(5)** involvement in common biochemical pathways, **(6)** binding to common drugs, and **(7)** shared disease implications. For each embedding method and benchmark, we generate separate distributions of similarity scores for positive and negative examples. In all cases, protein pairs in the positive group should tend to have greater similarity scores than those in the negative group. We score the performance of each method on the benchmarks with the *benchmark score* $B(p, n)$, given by:

$$B(p, n) = \mu_p - \mu_n \tag{2}$$

where μ_p and μ_n are the mean similarity scores of the positive and negative groups, respectively. The *benchmark score* quantifies how well each method distinguishes similar and dissimilar proteins. Greater values indicate a stronger ability to distinguish positive and negative groups.

B.3 Structure-Based Predictions of PLIs

The PDB serves as a reference of experimentally verified PLIs and their structural complexes [2]. Zemla *et al.* introduce a tool called PDBspheres to identify protein binding pockets ("spheres") in these structural complexes. Using PDBspheres, they generate a database of experimentally observed spheres associated with clinically relevant ligands [10].

We predict the structures of human proteins using AlphaFold2 [4]. With modeled structures, we identify human proteins containing binding pockets that align with known spheres [10]. We assume proteins that align with a sphere may bind the associated ligand; aligned proteins provide structural evidence of interactions with the ligand. We predict PLIs based on structural evidence according to these alignments. Due to fluctuations in protein conformations and uncertainties in structural models, predicting PLIs from structure alone can produce erroneous results. Therefore, we propose evaluating predicted interactions using protein function.

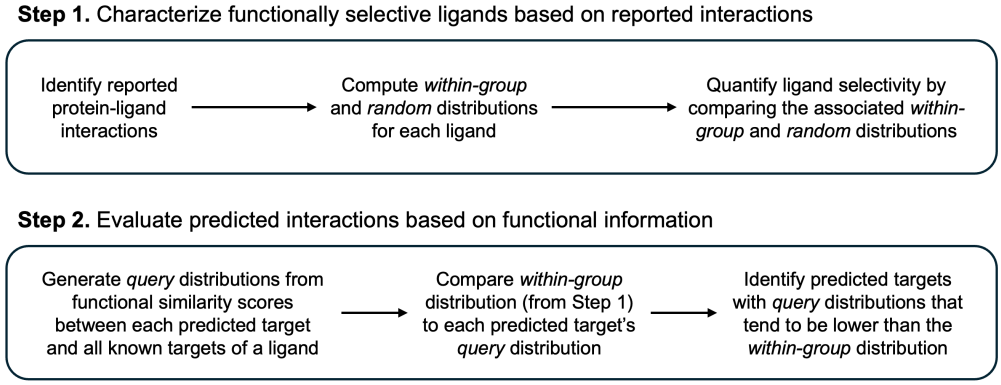

Figure 4: **Framework for Evaluating PLIs.** In Sec. B.3, we describe how AlphaFold2 and PDBspheres can be used to predict PLIs based on structural evidence [4, 5, 10]. For selective ligands, functional information offers an orthogonal comparator for evaluating the predicted interactions. The flowchart summarizes our use of functional information to evaluate predicted PLIs involving selective ligands.

B.4 Evaluating Predicted PLIs Using Functional Annotations

By aligning predicted protein structures to experimental binding pockets, we identify potential PLIs based on structural evidence [10]. However, due to uncertainties in modeled and measured protein structures, which propagate into the predicted interactions, there is a need to evaluate predictions based on orthogonal evidence. We propose the use of protein function information for this evaluation. For selective ligands, we expect the functions of predicted protein targets to align with those of known protein targets. Accordingly, we identify selective ligands (see Sec. B.4.1), then quantify how well the functions of predicted targets match those of known targets of the ligands (see Sec. B.4.2). Fig. 4 summarizes our framework for evaluating predicted PLIs.

B.4.1 Quantifying Functional Selectivity

We develop a framework for quantifying drug selectivity and evaluating predicted PLIs based on similarity scores between protein pairs. We first identify drugs that are associated with at least ten targets in ChEMBL for which embeddings are available [21, 20]. For each drug being evaluated, we use the reported targets to assess selectivity. We generate a *within-group* distribution of similarity scores associated with all pairs of proteins that bind the drug (see Fig. 2A). We then construct a *random* distribution of similarity scores between 100,000 random protein pairs (see Fig. 2B). We form empirical cumulative distribution functions for both distributions. For selective drugs targeting a narrow range of protein functions, the *within-group* distribution should typically be greater than the *random* distribution. Therefore, we define a *selectivity score* $S(l)$ of ligand l by the difference in the means of the *within-group* and *random* distributions, given by:

$$S(l) = \mu_w - \mu_r \tag{3}$$

where μ_w and μ_r are the means of the *within-group* and *random* distributions for the ligand, respectively. Greater *selectivity scores* indicate more selective drugs. Fig. 7 plots the distribution of *selectivity scores* for ligands, based on reported interactions in ChEMBL [20, 21].

B.4.2 Quantifying Misalignment

Using the scoring scheme defined above, we rank ligands by their selectivity. Predicted PLIs involving highly selective drugs are strong candidates for our function-based evaluation; by definition, selective drugs target proteins with a narrow range of functions, so we expect predicted protein targets to share similar functions with known targets of these ligands. Consider a predicted PLI between query protein Q and selective ligand L. First, we construct a *query* distribution from the similarity scores between Q and each known target of L (see Fig. 2C). We then score the misalignment of this *query* distribution with the *within-group* distribution associated with the ligand. We

define a *misalignment score* $M(q, w)$ for *query* distribution q and *within-group* distribution w by:

$$M(q, w) = \mu_w - \mu_q \tag{4}$$

where μ_q is the mean of the *query* distribution. Greater *misalignment scores* suggest more prominent inconsistencies between the *query* and *within-group* distributions. If the functions of Q are consistent with those of known targets for L, the two distributions should appear similar, and the *misalignment score* should be close to 0.

B.4.3 Aggregating Selectivity and Misalignment Scores Across Similarity Scoring Methods

Inconsistencies in functional similarity scores make the function-based evaluation of predicted PLIs sensitive to the choice of similarity scoring method. The uncertainty in functional similarity scores propagates as uncertainties in the *selectivity score* (defined by Eq. 3) and the *misalignment score* (defined by Eq. 4); as such, a function-based evaluation of predicted PLIs may be ambiguous. We propose quantifying ligand selectivity and evaluating predicted PLIs using average *selectivity scores* and *misalignment scores*, respectively, obtained from the following function-specific methods: **(1)** "GO2Vec + PCA," **(2)** "TransformerGO + PCA," and **(3)** "ProteInfer (All) + PCA." We weight our scores by the *benchmark scores* reported in Fig. 1, assigning proportional weights of 0.071, 0.076, and 0.173 to the three methods, respectively. These weights are normalized to a sum of one when computing weighted averages. By restricting our analysis to function-specific similarity scores, we promote an unbiased evaluation of predicted PLIs obtained from structural information.

B.4.4 Internal Validation: Misalignment Scores for Reported Interactions

To validate our framework for scoring predicted PLIs, we quantify misalignment for known interactions reported in ChEMBL, using a leave-one-out approach. Specifically, we iterate over each reported interaction, treating the interaction as if it were a prediction. We recompute the *within-group* distribution from all other known targets of the ligand, excluding the one we selected. We then generate a *query* distribution from pairwise similarity scores between the selected target and all other targets of the ligand. We compute the *misalignment score* associated with the *within-group* and *query* distributions, then iterate to the next known target of the ligand. Fig. 3B plots the distribution of *misalignment scores* associated with known targets of functionally selective ligands. As expected, reported PLIs are associated with low *misalignment scores*.

C Effects of Variability in Similarity Scores

To compare methods for scoring protein function similarity, we start by sampling 100,000 random protein pairs from the human proteome. For all pairs, we quantify similarity using each method listed in Tab. 1. Fig. 5 illustrates the correlations in similarity scores for all protein pairs between the scoring methods. For each protein pair, we then compute the median and interquartile range of similarity scores across the different scoring methods. We find that more similar protein pairs tend to be associated with less uncertainty in their similarity scores (see Fig. 6). As proteins become less alike, quantifying their similarity becomes more difficult.

Our framework involves quantifying ligand selectivity and evaluating predicted PLIs based on distributions of protein similarity scores (see B.4.1 and B.4.2 for details). Accordingly, the *selectivity scores* and *misalignment scores* we report are affected by uncertainty in protein similarity scores. To reduce the effect of uncertainty, we propose using average scores derived from three function-based protein similarity scoring methods (see B.4.3). In this section, we quantify *selectivity* and *misalignment scores* using individual embedding methods from Tab. 1. To do so, we separately generate distributions of protein similarity scores using each embedding method. We first focus on the *within-group* and *random* distributions, which are used to quantify selectivity. For the 10% most selective ligands (highlighted in Fig. 7), we plot these distributions in Fig. 8. For all ligands, Fig. 9 plots the correlation coefficients of *selectivity scores* calculated by the various methods, indicating relatively strong agreement. However, when considering the *within-group* and *query* distributions used in the evaluation of predicted interactions, we see less consistency. Fig. 10 plots the correlation coefficients of *misalignment scores* calculated by various methods for predicted targets of selective ligands. The results demonstrate that the evaluation of predicted interactions is sensitive to the choice of embedding method, highlighting the open need for more robust methods of quantifying protein similarity.

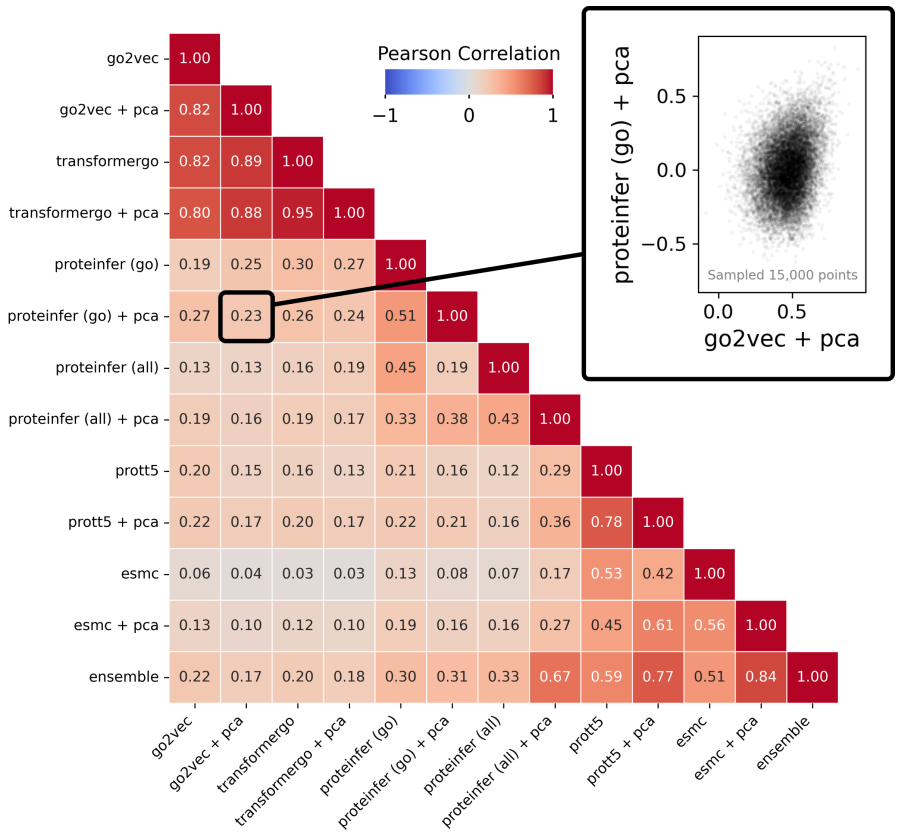

Figure 5: **Correlations in Similarity Scores.** For 100,000 random protein pairs, we evaluate similarity scores using all methods listed in Tab. 1, then evaluate the correlations in similarity scores between the methods. The inset scatter plot illustrates the weak correlation in similarity scores between two function-specific methods, based on 15,000 random protein pairs.

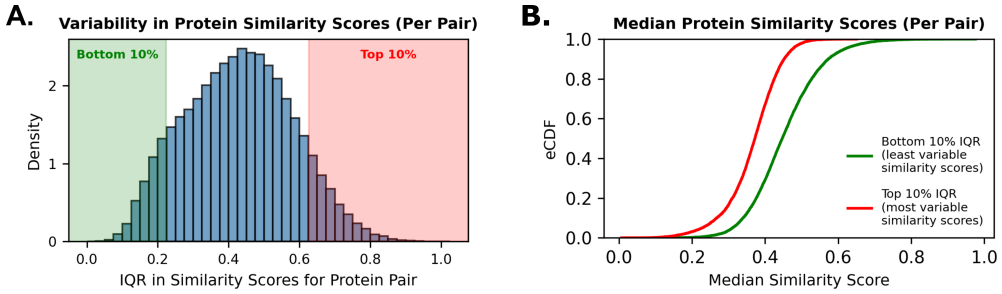

Figure 6: **Assessing Heteroscedasticity in Protein Similarity Scores.** For 100,000 random protein pairs, we evaluate similarity scores using the methods listed in Tab. 1. For each protein pair, we quantify the uncertainty in the similarity score based on its interquartile range (IQR) across scoring methods. **(A)** We plot the distribution of IQRs, highlighting the top 10% most uncertain pairs (in red) and bottom 10% least uncertain pairs (in green). **(B)** We then plot the distributions of median similarity scores (across scoring methods) for the protein pairs with the 10% most- and least-variable values. More similar protein pairs tend to have less uncertainty in their assessed similarity scores.

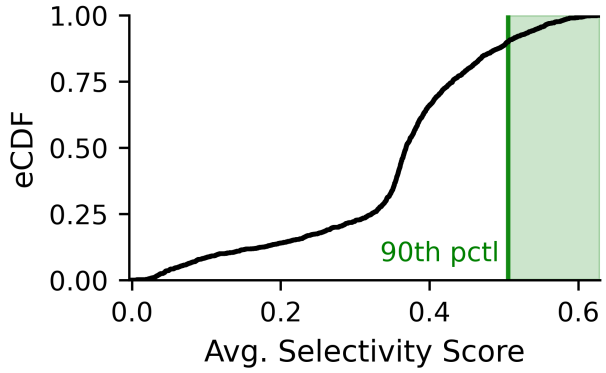

Figure 7: **Selectivity Scores.** Distribution of ligand selectivity scores, based on reported PLIs. The highlighted region captures the top 10% most selective ligands, based on average selectivity scores; these ligands are assessed further in Fig. 8.

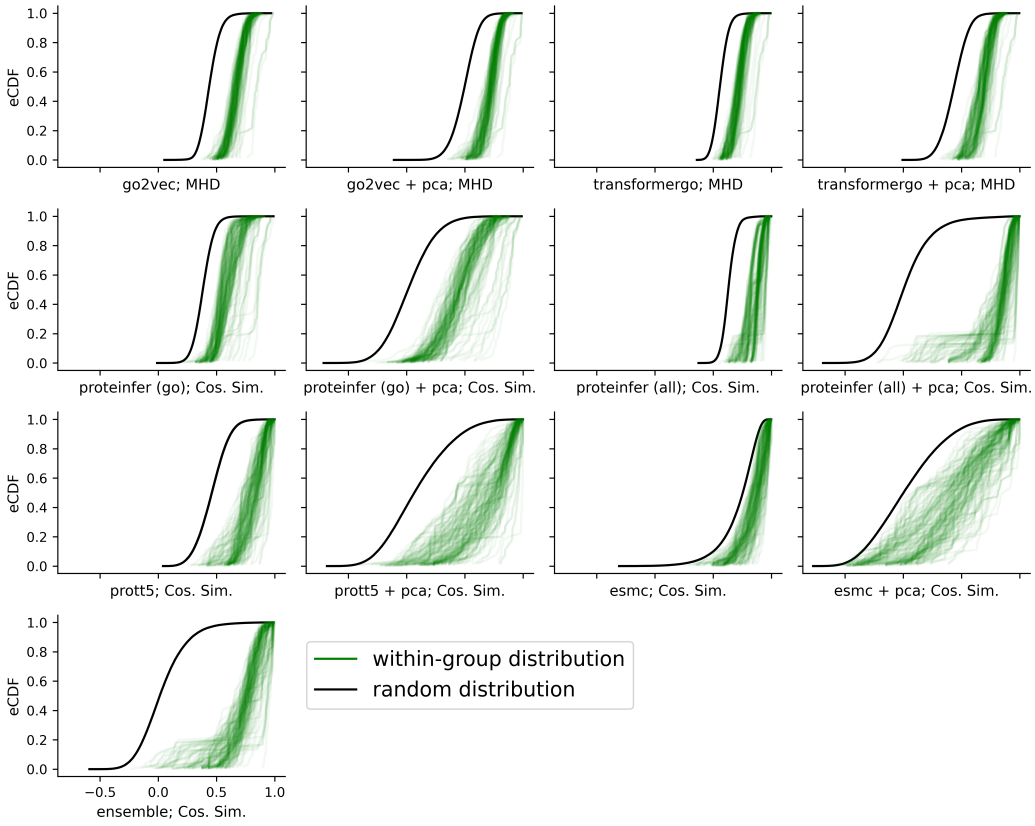

Figure 8: **Selectivity Evaluated by Each Similarity Scoring Method.** For the 10% most selective ligands identified based on reported interactions, we separately plot within-group distributions obtained from each similarity scoring method. For reference, we include the random distribution associated with the scoring methods. Recall that the mean difference between the within-group and random distributions produces the selectivity score. In all cases, the distributions of similarity scores between targets of selective ligands tend to be greater than random protein pairs.

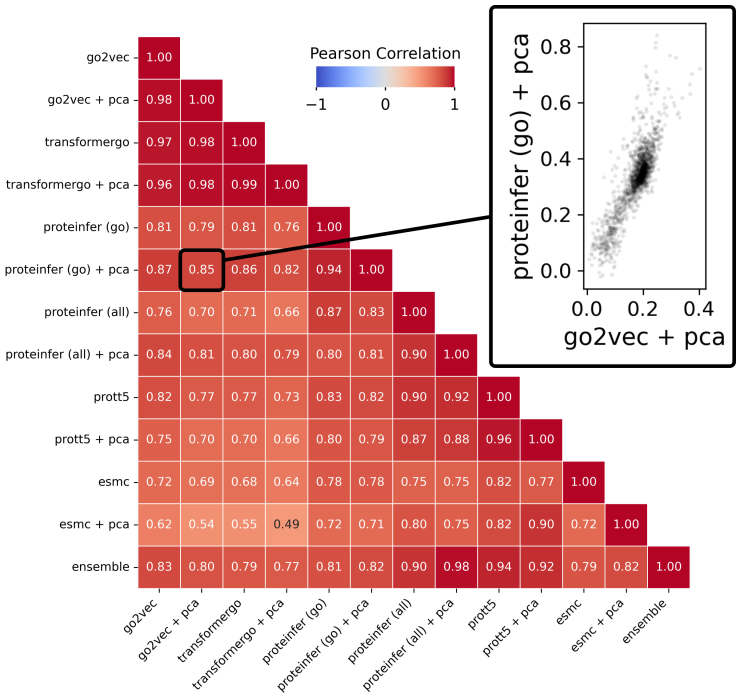

Figure 9: **Correlations in Selectivity Scores.** We evaluate selectivity scores for all ligands based on reported interactions in ChEMBL [20, 21], using each method in Tab. 1. We then calculate the correlation coefficient of selectivity scores for each pairing of embedding methods. The inset scatter plot exemplifies the correlation in selectivity scores obtained between different methods. In general, we see moderately strong correlations in selectivity scores between different scoring methods.

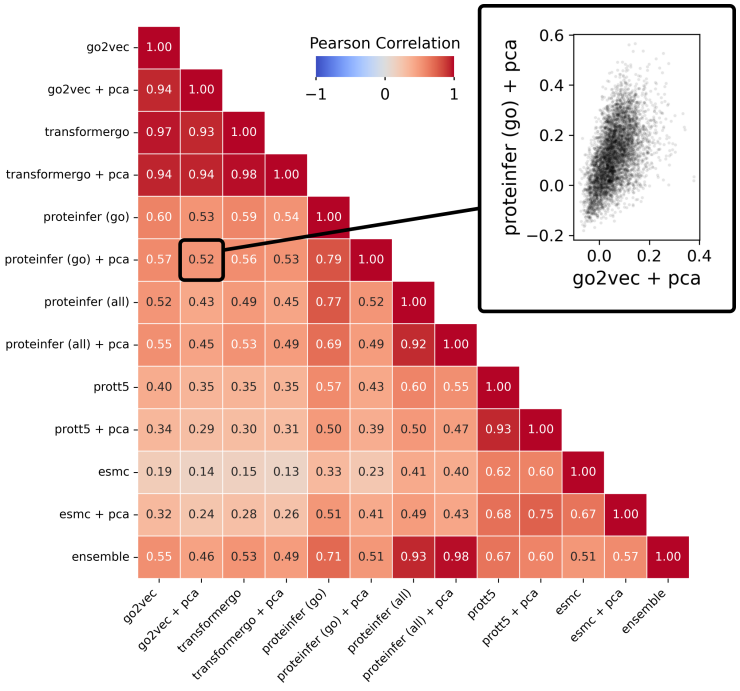

Figure 10: **Correlations in Misalignment Scores.** We evaluate misalignment scores for predicted interactions involving selective ligands using each method in Tab. 1. The inset scatter plot exemplifies the correlation in misalignment scores obtained between different scoring methods. Inconsistencies in the misalignment scores demonstrate the need for more robust similarity scoring methods.