PPI-LLAMA2: *De Novo* GENERATION OF BINDING PROTEINS CONDITIONED ON TARGET SEQUENCE ALONE

Yinuo Zhang
Centre for Computational Biology
Duke-NUS Medical School
Singapore 169857
yzhang@u.duke.nus.edu

Phil He
Department of Biomedical Engineering
Duke University
Durham, NC 27701
phil.he@duke.edu

Ashley Hsu
Department of Biomedical Engineering
Duke University
Durham, NC 27701
ashley.hsu@duke.edu

Sophia Vincoff
Department of Biomedical Engineering
Duke University
Durham, NC 27701
sophia.vincoff@duke.edu

Pranam Chatterjee
Department of Biomedical Engineering
Duke University
Durham, NC 27701
pranam.chatterjee@duke.edu

ABSTRACT

The targeting of disease-driving proteins is a critical goal of biomedicine. However, many of these proteins do not possess accessible small molecule binding pockets and are oftentimes conformationally disordered, precluding binder design via structure-dependent methods. Here, we present **PPI-Llama2**, which tasks Meta's Llama2 autoregressive language model architecture to *de novo* generate protein binders conditioned directly on target sequences. Without relying on structural data and training only on protein-protein interaction (PPI) sequences, PPI-Llama2 effectively learns the evolutionary semantics of PPIs, enabling the generation of both novel and biologically-plausible binders. Comparative evaluations highlight PPI-Llama2's performance in generating binders for evolutionarily distant targets, performing strongly against structure-dependent methods like RFDiffusion. In total, our findings showcase PPI-Llama2's potential to aid therapeutic discovery for diseases driven by undruggable and disordered target proteins, and motivate further experimental screening efforts.

1 INTRODUCTION

Protein-protein interactions (PPIs) are foundational to numerous biological processes, such as signal transduction and enzyme activity regulation, as well as in the development and progression of various diseases, including cancer and neurodegenerative disorders (Xie et al., 2023). Proteins involved in these disease-driving pathways, however, oftentimes possess large and flat interfaces and are largely disordered, precluding design of standard small molecule-based therapeutics to these taregets (Xie et al., 2023; Behan et al., 2019). The design of protein-based binders, such as antibodies, nanobodies, miniproteins, and peptides, thus present exciting opportunities for therapeutic intervention (Chen et al., 2023). Nonetheless, traditional methods for binder generation, such as phage display and yeast two-hybrid systems, are labor-intensive and time-consuming, prompting the exploration of computational alternatives.

Current state-of-the-art computational methods for binder design, most notably RFDiffusion (Watson et al., 2023), rely on structural templates for binder scaffolding, precluding design to conformationally-disordered, disease-driving target proteins, such as transcription factors and fusion oncoproteins (Chen et al., 2023). In contrast, autoregressive protein language models (pLMs) such as ProtGPT2 have been trained on millions of protein sequences and demonstrate the ability to generate novel protein sequences without structural inputs (Ferruz et al., 2022), highlighting the semantic understanding of these models and motivating novel generation from their vocabulary space.

In this work, we present PPI-Llama2, leveraging Meta's Llama2 model (Touvron et al., 2023a;b) to develop an autoregresive pLM that generates binding proteins *de novo* conditioned directly on target sequence inputs. Distinct from GPT models, Llama2 offers systematic advantages, allowing customization of the model's architecture, training objectives, and procedures tailored to specific protein corpuses. Trained directly on PPI datasets without pre-training on extensive protein databases such as UniRef (Suzek et al., 2007), PPI-Llama2 captures sequence diversity and interaction patterns of naturally-occurring PPIs. By conditioning the sequence generation process solely on the target sequence during inference (Figure1), we direct PPI-Llama2's generative capabilities to produce novel, complementary binder candidates to diverse, non-homologous target substrates, facilitating rapid *in silico* screening without the requirement of structural information. As an extrinsic benchmark, we successfully confirm the binding capacity of our generated sequences via AlphaFold2-Multimer (Evans et al., 2021), and further establish their novelty and biomolecular and functional similarities to cognate targets. Overall, our results suggest that PPI-Llama2 grasps intricate relationships between protein sequence and function to generate stable and plausible binders, motivation utilization of the model for downstream therapeutic applications.

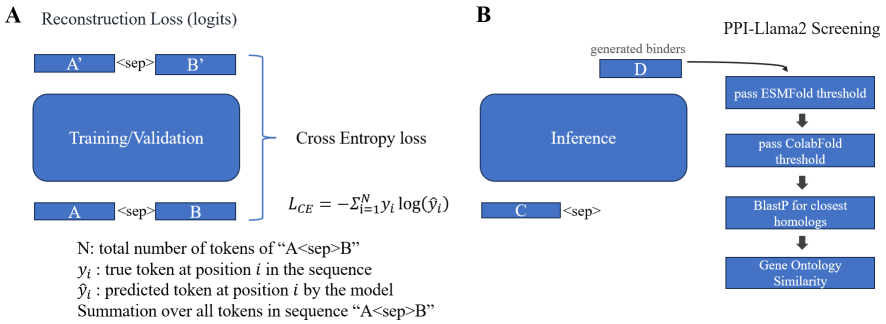
A
Reconstruction Loss (logits)
A'
<sep>
B'
Training/Validation
Cross Entropy loss
$L_{CE} = -\Sigma_{i=1}^{N} y_i \log(\hat{y}_i)$
A
<sep>
B
N: total number of tokens of "A<sep>B"
y_i : true token at position i in the sequence
$\hat{y}_i$: predicted token at position i by the model
Summation over all tokens in sequence "A<sep>B"
B
PPI-Llama2 Screening
generated binders
D
Inference
C
<sep>
pass ESMFold threshold
pass ColabFold threshold
BlastP for closest homologs
Gene Ontology Similarity

Figure 1: PPI-Llama2 Overview. A) Training procedure. B) Screening pipeline.

2 Methods

Data Preparation PPI datasets were curated from the IntACT (Hermjakob et al., 2004) and BioGrid (Oughtred et al., 2021) databases for training purposes. A byte pair encoding (BPE) (Shibata et al., 1999) tokenizer was trained on 569,786 proteins entries from Swiss-Prot (Boeckmann et al., 2003) (release date: 2023-09-13) with a vocabulary size of 32k, including the special token <sep>to indicate sequence separation. To ensure a robust training-validation split and prevent data leakage, we used MMseqs2 (Steinegger & Söding, 2017) with a sequence identity threshold of 0.3 for filtering PPI pairs and performed a cluster-based split, ensuring no sequence overlap between training and validation data. For benchmarking, we selected Docking Benchmark 5 (DB5) PPI pairs (Townshend et al., 2019) with less than 10% homology to the training set, identified through MMseqs2 comparison, to test our model's generalizability. Only 15 sequences passed that threshold and were held-out for benchmarking. For sequence evaluation by principal component analysis (PCA), we evaluated data distribution and sequence novelty by analyzing embeddings from natural, random (from ProtGPT2), and PPI-Llama2-generated sequences using the pre-trained, state-of-the-art ESM-2 model (esm2_t33_650M_UR50D) (Lin et al., 2023).

Hyperparameters and Training Details PPI-Llama2 consists of 342 million parameters. Training was performed on a NVIDIA 8xA100 GPU system for 3 days. Our training procedure adopted ESM-2's dynamic batching technique to maximize the utilization of GPU memory. Our model was configured via HuggingFace's Llama2 (Wolf et al., 2019) framework, incorporated with FlashAttention-2 (Dao, 2023) optimization for enhanced processing efficiency. We also integrated PyTorch Lightning frameworks (v2.1.4) (Falcon & The PyTorch Lightning team, 2019) to achieve distributed data parallel (DDP) training. The learning rate was set to 1e-4, and we followed Llama2's cosine scheduler (Touvron et al., 2023b) to adjust the learning rate. The training process involved using the <sep>token within PPI sequences to demarcate the transition between proteins, a method that was mirrored during inference to signal the start of binder generation. The model was trained with cross-entropy loss to regenerate the original PPI sequences, with pseudo-perplexity and accuracy meticulously tracked throughout the training period. A schematic of the training procedure can be found in (Figure 1A).

Sequence Screening Pipeline In preliminary testing, we observed a positive correlation between ESMFold (Lin et al., 2023) predicted Local Distance Difference Test (pLDDT) and predicted TM score (pTM) scores with the likelihood of binding between two proteins, aligning with the recent publications that also use these metrics to assess binding sequences (McLean, 2024) (Supplementary Figure A). Although integrated predicted TM (ipTM) has been shown to be a more accurate metric for binding events (Bryant et al., 2022), the computational demands of AlphaFold2-Multimer Evans et al. (2021) led us to opt for ESMFold for the initial screening of candidate binders. We began by selecting generation parameters that produced outputs with stable pseudo-perplexity (standard deviation ≤ 50) within a specified parameter range (top-p: 0.7-0.95; temperature: 0.7-1; repetition_penalty: 1-5; length_penalty: 0.4-1.01). Subsequently, sequences were generated and screened using ESMFold, significantly accelerating the output screening process. Generated binders were then ranked by pLDDT changes and pTM scores to highlight the possible candidates for further screening. Final candidates were further filtered via ipTM scores by AlphaFold2-Multimer version 1.5.5 supported by ColabFold (Mirdita et al., 2022). For homology information, generated sequences were analyzed using BlastP (Madden, 2013) and Entrez (Schuler et al., 1996). Gene Ontology (GO) terms, particularly focusing on molecular function, were determined for these sequences and the target proteins via InterProScan (Paysan-Lafosse et al., 2023). Comparison of GO terms was performed using GOSemSim (v2.12.0) (Yu, 2020) to assess semantic similarities. A schematic of the screening pipeline can be found in Figure 1B.

Pseudo-Perplexity of PPI-Llama2 We adopted the transition scores from HuggingFace to calculate the pseudo-perplexity as a measurement of how stable the model's generation parameters behaved during inference.

$$\text{Pseudo-Perplexity}(a) = e^{-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{e^{l_{x_i}}}{\sum_j e^{l_j}}\right)}$$

This equation operates over all tokens i in the sequence, where N represents the total number of tokens. It calculates the log probability of the softmax-transformed logits (l), which are the model's transition scores, for each token position i in the sequence. In this context, $\mathbf{x}_i$ denotes the specific token at the position i in the sequence, as chosen or predicted by the model. $\mathbf{l}_{\mathbf{x}_i}$ is the logit corresponding to the token.

3 Results and Discussion

PCA Distribution Analysis We performed PCA to compare the latent space distribution of sequences generated by PPI-Llama2 to both natural sequences and random sequences as defined by ProtGPT2 Ferruz et al. (2022), using the ESM-2 model for embedding generation Lin et al. (2023). As illustrated in Figure 2, the distribution of PPI-Llama2-generated sequences occupies an intermediate position between that of natural and random sequences and leans towards the natural protein distribution (Figure 2). This positioning suggests that our model's outputs not only adhere to the distribution patterns characteristic of naturally-occurring proteins but also introduce a level of novelty that distinguishes them from randomly generated sequences. A closer alignment with natural sequences demonstrates the model's efficiency in learning the underlying patterns and functions that characterize naturally occurring proteins from the training set, which showed potential of our model to generate protein binders that are biologically plausible and functionally relevant.

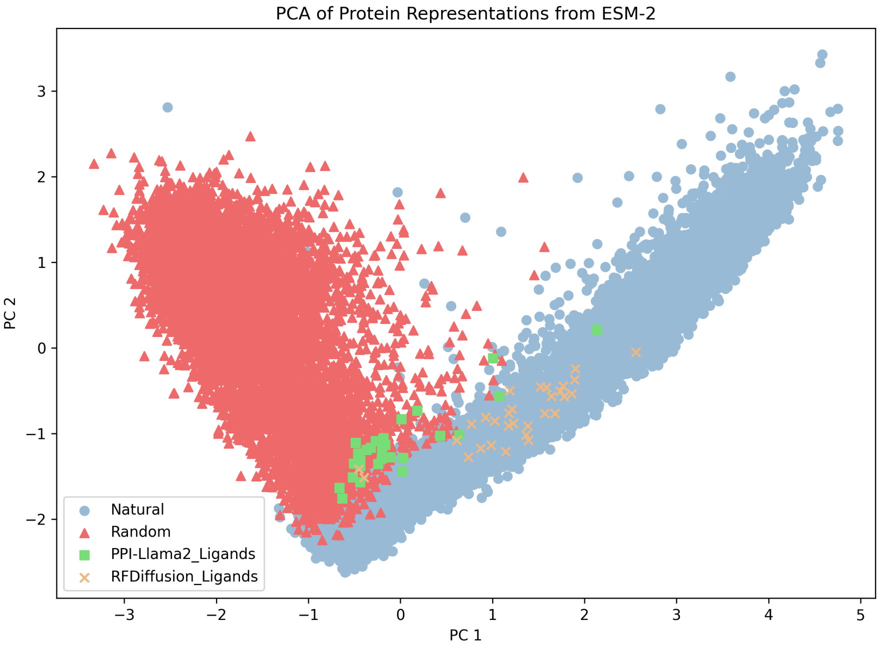

Figure 2: PCA of PPI-Llama2 and RFDiffusion generated protein distribution compared to natural and random proteins from ProtGPT2.

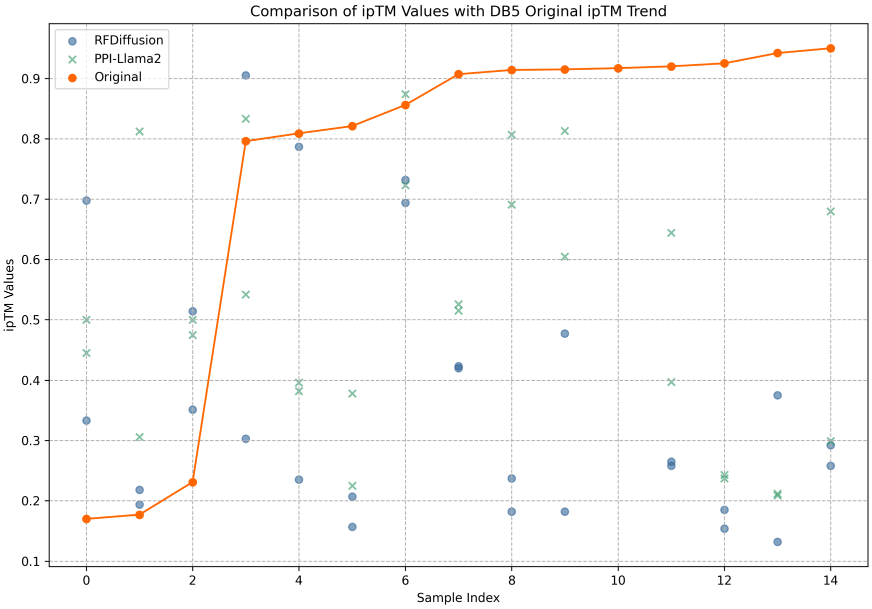

Figure 3: ipTM comparison among the original binder from DB5, the binder generated from RFDiffusion, and the binder generated from PPI-Llama2. ipTM are generated from AlphaFold2.

Binder Generation Analysis Given the complexity of protein-protein interactions (PPIs) documented in databases like IntAct and BioGRID, our analysis identified a protein that binds with over 4910 others, highlighting the extensive interconnectivity within PPI networks. To mitigate data leakage, we meticulously ensured that each protein appeared in only one dataset split, either for training or validation, as illustrated in Supplementary Figure B. For benchmarking, we utilized experimentally verified DB5 data (Townshend et al., 2019), selecting 15 sequences with less than 10% homology to those in our training set through MMseqs2. This preparatory step was crucial for assessing our model's capability to generate novel binders.

In comparing the binder generation performance between PPI-Llama2 and RFDiffusion (Figure 3), we observed that RFDiffusion, which requires the target *structure* as input, struggled with structured targets that are evolutionarily distant, often yielding repeated tokens for binders in the 15-100 length range. Conversely, PPI-Llama2, which only requires the target *sequence* as input, demonstrated a superior ability to generate diverse targets with high AlphaFold2-Multimer ipTM scores, previously established as an indicative metric of binding affinity Bryant et al. (2022); Zhu et al. (2023). Example AlphaFold2-Multimer co-folds are shown in Supplementary Figures C and D.

To further validate the novelty and relevance of the generated binders, we conducted BlastP searches McGinnis & Madden (2004) of generated sequences, finding that many binders either lacked homologs or were distantly related (E-value $>$ 0.01) to segments of hypothetical proteins. This was particularly notable for newly added target proteins lacking Gene Ontology (GO) annotations (Paysan-Lafosse et al., 2023; Yu, 2020). For targets with GO labels, GO similarity analysis revealed that our generated proteins shared biochemical functionalities with their targets (Table 1), suggesting that PPI-Llama2 leverages evolutionary training data effectively to produce binders that are biologically plausible and functionally relevant for previously unseen proteins.

Target GO Term	Binder GO Term	Similarity Score
beta-lactamase activity (GO:0008800)	catalytic activity (GO:0003824)	0.454
metalloendopeptidase inhibitor activity (GO:0008191)	protein dimerization activity (GO:0046983)	0.072
NAD+-diphthamide ADP-ribosyltransferase activity (GO:0047286)	arginine deiminase activity (GO:0016990)	0.139
complement binding (GO:0001848)	antiporter activity (GO:0015297)	0.093
phosphatase activity (GO:0016791)	phosphatase activity (GO:0016791)	1
serine-type endopeptidase inhibitor activity (GO:0004867)	iron-sulfur cluster binding (GO:0051536)	0.072
cysteine-type peptidase activity (GO:0008234)	transmembrane transporter activity (GO:0022857)	0.134
RNA binding (GO:0003723)	carbohydrate binding (GO:0030246)	0.325

Table 1: The Binder GO Term is determined via the top hit (provided that its E-value $>$ 0.01) of BlastP (filtering out "hypothetical protein" hits) of the generated binder for the corresponding target protein.

4 CONCLUSION

In total, PPI-Llama2 establishes the ability of autoregressive language modeling architectures to learn meaningful sequence-function relationships from protein interaction data alone, without reliance on structural templates or pre-trained pLM embeddings. Notably, PPI-Llama2's performance in generating binders that closely resemble natural proteins suggests a high degree of biological relevance and functional potential, a critical feature for the practical application of generated binders. Our ongoing work is focused on continued model development and experimental validation via ELISA and SPR-based binding affinity characterization. As more PPI data becomes available, further training may enhance generalizability, especially to underrepresented protein families. High-throughput experimental screening of generated binders via cell surface display methods, for example, will provide real-world application of our model to novel, disease-relevant targets. In conclusion, PPI-Llama2 provides an innovative *in silico* approach to protein binder design *without* the requirement of target structure. With refinement and validation, this work may open new avenues for targeting previously undruggable, disordered proteins implicated in disease.

REFERENCES

Fiona M. Behan, Francesco Iorio, Gabriele Picco, Emanuel Gonçalves, Charlotte M. Beaver, Giorgia Migliardi, Rita Santos, Yanhua Rao, Francesco Sassi, Marika Pinnelli, Rizwan Ansari, Sarah Harper, David Adam Jackson, Rebecca McRae, Rachel Pooley, Piers Wilkinson, Dieudonne van der Meer, David Dow, Carolyn Buser-Doepner, Andrea Bertotti, Livio Trusolino, Euan A. Stronach, Julio Saez-Rodriguez, Kosuke Yusa, and Mathew J. Garnett. Prioritization of cancer therapeutic targets using crispr–cas9 screens. *Nature*, 568(7753):511–516, April 2019. ISSN 1476-4687. doi: 10.1038/s41586-019-1103-9. URL `http://dx.doi.org/10.1038/s41586-019-1103-9`.

Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, et al. The swiss-prot protein knowledgebase and its supplement trembl in 2003. *Nucleic acids research*, 31 (1):365–370, 2003.

Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using alphafold2. *Nature communications*, 13(1):1265, 2022.

Tianlai Chen, Lauren Hong, Vivian Yudistyra, Sophia Vincoff, and Pranam Chatterjee. Generative design of therapeutics that bind and modulate protein states. *Current Opinion in Biomedical Engineering*, pp. 100496, 2023.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, and Demis Hassabis. Protein complex prediction with alphafold-multimer. October 2021. doi: 10.1101/2021.10.04.463034. URL `http://dx.doi.org/10.1101/2021.10.04.463034`.

William Falcon and The PyTorch Lightning team. PyTorch Lightning, March 2019. URL `https://github.com/Lightning-AI/lightning`.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):4348, 2022.

Henning Hermjakob, Luisa Montecchi-Palazzi, Chris Lewington, Sugath Mudali, Samuel Kerrien, Sandra Orchard, Martin Vingron, Bernd Roechert, Peter Roepstorff, Alfonso Valencia, et al. Intact: an open source molecular interaction database. *Nucleic acids research*, 32(suppl_1):D452–D455, 2004.

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

Thomas Madden. The blast sequence analysis tool. *The NCBI handbook*, 2:425–436, 2013.

S. McGinnis and T. L. Madden. Blast: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*, 32(Web Server):W20–W25, July 2004. ISSN 1362-4962. doi: 10.1093/nar/gkh435. URL `http://dx.doi.org/10.1093/nar/gkh435`.

Thomas C McLean. Lazyaf, a pipeline for accessible medium-scale in silico prediction of protein-protein interactions. *bioRxiv*, pp. 2024–01, 2024.

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.

Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, et al. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200, 2021.

Typhaine Paysan-Lafosse, Matthias Blum, Sara Chuguransky, Tiago Grego, Beatriz Lázaro Pinto, Gustavo A Salazar, Maxwell L Bileschi, Peer Bork, Alan Bridge, Lucy Colwell, et al. Interpro in 2022. *Nucleic acids research*, 51(D1):D418–D427, 2023.

Gregory D Schuler, Jonathan A Epstein, Hitomi Ohkawa, and Jonathan A Kans. [10] entrez: Molecular biology database and retrieval system. In *Methods in enzymology*, volume 266, pp. 141–162. Elsevier, 1996.

Yusuxke Shibata, Takuya Kida, Shuichi Fukamachi, Masayuki Takeda, Ayumi Shinohara, Takeshi Shinohara, and Setsuo Arikawa. Byte pair encoding: A text compression scheme that accelerates pattern matching. 1999.

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.

Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Raphael Townshend, Rishi Bedi, Patricia Suriana, and Ron Dror. End-to-end learning on 3d protein structure for interface prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Xin Xie, Tingting Yu, Xiang Li, Nan Zhang, Leonard J. Foster, Cheng Peng, Wei Huang, and Gu He. Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. *Signal Transduction and Targeted Therapy*, 8(1), September 2023. ISSN 2059-3635. doi: 10.1038/s41392-023-01589-z. URL `http://dx.doi.org/10.1038/s41392-023-01589-z`.

Guangchuang Yu. Gene ontology semantic similarity analysis using gosemsim. *Stem Cell Transcriptional Networks: Methods and Protocols*, pp. 207–215, 2020.

Wensi Zhu, Aditi Shenoy, Petras Kundrotas, and Arne Elofsson. Evaluation of alphafold-multimer prediction on multi-chain protein complexes. *Bioinformatics*, 39(7), July 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad424. URL `http://dx.doi.org/10.1093/bioinformatics/btad424`.

A Supplementary Figures

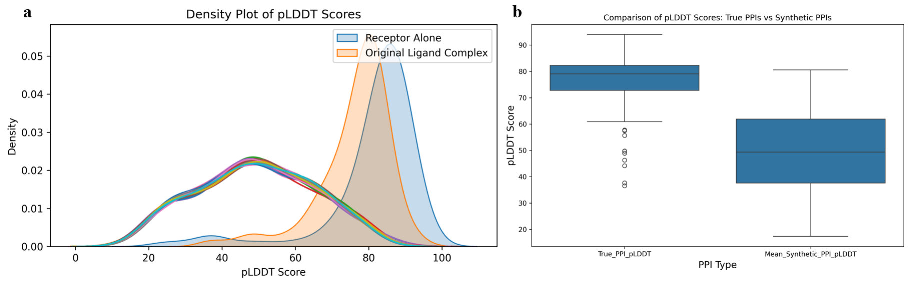

Figure A: Evaluation of ESMFold for binder screening. We first generated 50 synthetic binders drawn from uniform distribution of amino acids to each of the DB5 target proteins. Then we performed ESMFold on (1) target proteins; (2) target+experimental binders (original PPI); (3) target+synthetic binders(synthetic PPIs). And retrieved their pLDDTs accordingly. a) The uncolored pLDDT distributions represent receptor sequences binding with random binders, and the colored distributions represent receptor pLDDT alone and the original complex pLDDT. Patterns indicate that binding with random synthetic binders decrease the ESMFold pLDDT compared with binding with their natural binders. b) In general, synthetic PPIs with random binders tend to have lower pLDDT. Our evaluation verified that ESMFold can discriminate natural, verified binders from random, synthetic binders.

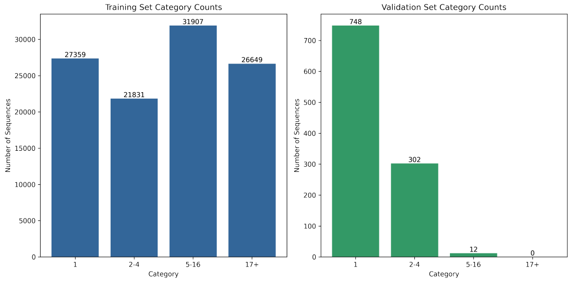

Figure B: Evaluation of ESMFold for binder screening. We ensure that there are no overlapping partner sequences within PPIs in the train-validation data splits. As such, each PPI is unique. The x-axis represents the number of alternate sequences to which the query sequence binds, and the y-axis represents the total counts for each category. The 17+ bin highlights the "super" sequences that participate in complex networks and bind to $\geq$17 other target proteins.

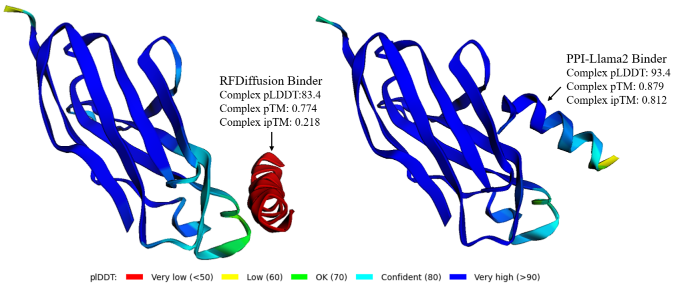

Figure C: Example visualization of binders generated for target 2I25 chain A by RFDiffusion and PPI-Llama2.

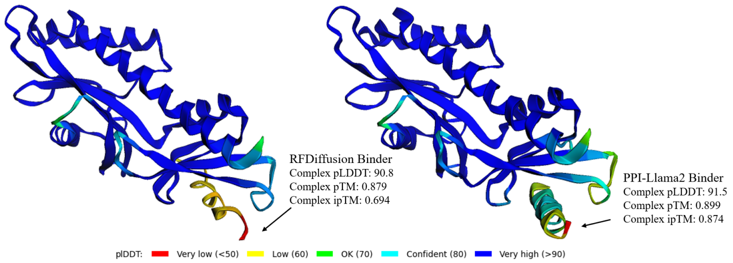

Figure D: Example visualization of binders generated for target 1ZHH chain B by RFDiffusion and PPI-Llama2.