
A Multitask Protein Language Model For Protein Property and Protein-Protein Interactions

Sukanya Sasmal*, Robert Jiang, Sharmi Banerjee, Ahmed Ismail, Hoda Akl
Pete Bowry, Marcus Collins†, Karla-Luise Herpoldt, Hamid Mohamadi
Thanos Nikolakopoulos, Divyajyoti Panda, Gherman Uritskiy
Derric Williams, Mostafa Karimi, Bella Dubrov, Shang Shang
Amazon

Abstract

Protein language models (PLMs) are typically pre-trained on single sequences, yet many downstream applications require understanding protein-protein interactions (PPIs). We present ProPRINT, a 3B-parameter T5 encoder-decoder model jointly pre-trained on single-sequence inpainting and a novel pair inpainting objective that reconstructs masked residues of one protein conditioned on its interacting partner, using PPI data from StringDB (experimentally validated interactions) and SynthPPI (74.6M synthetic interacting domain pairs derived from the AlphaFold Protein Structure Database). For SynthPPI, we introduce a contact-aware masking strategy that preferentially masks binding-interface residues, integrating structural information into sequence-only models. ProPRINT-3B outperforms ESM2 models up to 3B and MINT on two of five PPI benchmarks and four of eight single-sequence benchmarks, showing that PPI-augmented pre-training has the potential to improve representation learning without sacrificing single-sequence representation quality.

1 Introduction

Protein language models (PLMs) learn rich representations of proteins from amino acid sequences Lin et al. [2023], Elnaggar et al. [2021], Chen et al. [2024], Hayes et al. [2025], Madani et al. [2020] and perform strongly on downstream structure, function, and property prediction. Structure-aware variants incorporate 3D tokens: ProstT5 Heinzinger et al. [2024] uses Foldseek Van Kempen et al. [2024] structure tokens and SaProt Su et al. [2023] combines structural tokens with amino acids. These models are typically pre-trained on millions of individual sequences with self-supervised objectives such as masked language modeling. However, proteins function through interaction networks, and protein-protein interactions (PPIs) underlie nearly all cellular processes, making PPI prediction central to drug discovery, protein design, and fundamental biology.

Efforts to model protein-protein interactions have taken complementary routes. MINT Ullanat et al. [2026] fine-tuned a pre-trained ESM2-650M model with a cross-attention module to produce interaction-aware embeddings, improving PPI prediction and mutational effect benchmarks over single-sequence baselines. To address the scarcity of experimental interaction data, La-Complexa Didi et al. [2026], a latent flow matching model for binder design, extracted interacting domain pairs from predicted multimeric structures in the AlphaFold Protein Structure Database (AFDB) Varadi et al. [2024] to build synthetic PPI data beyond the limited set of experimentally resolved complexes, showing that structure-derived synthetic data improves interaction prediction.

In this work, we investigate whether augmenting PLM pre-training with PPI data improves representation learning for downstream PPI tasks. We present ProPRINT (**P**rotein **P**roperty and **I**nteraction

*Corresponding author: sukanyaz@amazon.com

†Current affiliation: Astera Institute.

Prediction), a 3B-parameter T5 encoder-decoder model jointly pre-trained on two objectives: (i) a standard single-sequence inpainting task, and (ii) a pair inpainting task that reconstructs masked residues of one protein conditioned on its interacting partner, sourcing PPI data from both experimentally verified interactions and synthetic structural databases. We evaluate ProPRINT on five PPI benchmarks (binary interaction, mutational effect, and $\Delta\Delta G$ prediction) and eight single-sequence tasks. Our results show that multi-task pre-training with augmented PPI data and a novel contact-aware masking strategy outperforms single-sequence PLM baselines on two out of five PPI tasks and four out of eight single-sequence tasks.

Our key technical contributions are:

Multi-task pre-training objective. We introduce a multi-task pre-training framework that jointly trains on single-sequence inpainting and protein-protein pair inpainting objectives, integrating PPI understanding directly into the pre-training stage of a PLM without requiring architectural modifications or a separate fine-tuning step. To our knowledge, this is the first work where LLMs have been pre-trained jointly on single-sequence and PPI data.

Synthetic PPI data. We introduce SynthPPI, a large-scale synthetic PPI dataset of approximately 74.6 million interacting domain pairs derived from AFDB. This dataset can be used to augment limited experimental PPI data.

Contact-aware masking strategy. We propose a contact-aware masking strategy that preferentially masks residues at binding interfaces. Binding site information is obtained from AFDB when constructing SynthPPI, demonstrating a way to integrate multimodal information into sequence-only models. This is a novel masking strategy where the model directly learns binding site information during pre-training without any architectural changes. To our knowledge, this is the first work where this strategy has been used for masking during PLM pre-training.

2 Methods

2.1 Training Data

We trained ProPRINT on four different datasets - UniRef30 and BigDB for the inpainting task and StringDB and SynthPPI for the pair inpainting task. Dataset statistics are listed in table A1.

Inpainting Data. We used two single-sequence datasets for the inpainting task. UniRef30 Mirdita et al. [2022] is a clustered representation of UniRef100, grouping all UniProt Knowledgebase sequences at 30% sequence identity using MMseqs2; each cluster is represented by its seed sequence, yielding a non-redundant set of approximately 19.2 million representative sequences. BigDB was constructed as a deduplicated union of several large-scale protein sequence repositories: UniProt Archive (UniParc) uni [2025], UniRef Suzek et al. [2007], NCBI Reference Sequence (RefSeq) Pruitt et al. [2005], and MGnify Richardson et al. [2023], clustered at 85% sequence identity to reduce redundancy while preserving broader sequence diversity than UniRef30.

Pair Inpainting Data. We used two PPI datasets for the pair inpainting task. **StringDB** provides experimentally validated physical interactions: we used the physical interaction subset of STRING (version 2023) Szklarczyk et al. [2025], keeping only high-confidence pairs dominated by experimental evidence to minimize false positives. **SynthPPI** is a synthetic dataset of ~ 74.6 M interacting domain pairs derived from AFDB, where interacting domains are identified from predicted structures (fig. A2). SynthPPI is much more diverse than StringDB in unique sequences (table A1), with shorter sequences since they are drawn from domains. Full filtering and curation details are provided in appendix D.1.

2.2 Training Details and Objectives

We trained ProPRINT as a 3B-parameter T5 encoder-decoder using the Megatron-LM framework for 200B tokens on 32 NVIDIA H200 GPUs, with a training data composition of 34% inpainting, 33% StringDB, and 33% SynthPPI. Sequences were tokenized at single-residue granularity. The full run completed in 508K steps (~ 6 days). Complete architecture and optimization details are provided in appendix D.3.

The model is trained with the two training objectives – standard inpainting objective and a pair inpainting objective, both optimized with a cross-entropy loss.

Inpainting. In the first task, the model is trained to reconstruct residues masked in a single protein sequence. We follow the ESM tokenization convention, each masked residue is individually replaced with a mask token. The masking rate is sampled from a Beta-linear schedule with parameters $\alpha = 5$, $\beta = 8$, and a linear fraction of 0.5.

Pair inpainting. In the second task, two interacting protein sequences are concatenated with a special `<pair_sep>` token. Only the second sequence is masked using the same Beta-linear masking scheme, and the model learns to reconstruct the masked residues conditioned on the full first sequence. The model learns representations that capture inter-protein dependencies from this objective.

Contact-aware masking for SynthPPI data. For SynthPPI data, we explored a new masking strategy that distinguished between contact and non-contact residues in the second sequence. We defined a residue as a contact residue if it lies at the binding interface between the two domains (at least 15 residue pairs within 8 Å for the C β atoms in the predicted structure). We then applied two separate Beta-linear masking functions – (i) a Beta-linear distribution with $\alpha = 5$, $\beta = 5$ and a linear fraction of 0.5 for residues in the second sequence that formed contacts or direct protein-protein interactions, and (ii) a Beta-linear distribution with $\alpha = 2$, $\beta = 8$ and a linear fraction of 0.5 for the remaining residues of the second sequence. This contact-aware scheme allowed the model to see a higher masking rate at interaction interfaces while applying a lighter masking rate to non-contact regions. The residues at the binding site exhibit stronger correlations with the partner protein, and we hypothesize that preferentially masking these positions encourages the model to learn richer inter-protein dependencies.

2.3 Evaluation Details

We compared ProPRINT-3B against ESM-2 models spanning 8M to 3B parameters Lin et al. [2023] and MINT Ullanat et al. [2026] (tables 1 and 2). All models were evaluated using frozen encoders with identical MLP probing heads, ensuring that differences reflected representation quality rather than fine-tuning capacity. Sequence-level embeddings were obtained by mean-pooling residue-level representations, and a lightweight MLP probing head was trained on top to isolate representation quality. For protein-protein pairs, we extracted embeddings in two modes. In **inpainting mode**, each sequence in the pair is encoded independently and the resulting residue-level embeddings are concatenated before mean-pooling; this mode is also used for single-sequence models such as ESM2. In **pair inpainting mode**, the two sequences are concatenated with a `<pair_sep>` token and encoded jointly, and the residue-level embeddings of the resulting complex are mean-pooled together, allowing the encoder to model inter-protein context directly. Full embedding-extraction and probing protocols are provided in appendix E.3.

3 Results

We evaluated ProPRINT on five PPI benchmarks and eight single-sequence downstream tasks. The PPI benchmarks are as follows: binding free energy change prediction ($\Delta\Delta G$, SKEMPI Jankauskaitė et al. [2019]), mutational effect prediction (Mutational PPI Siwek et al. [2025]), and binary interaction prediction (Golden-PPI Bennett et al. [2024], Human-PPI, and Yeast-PPI Xu et al. [2022], Pan et al. [2010], Guo et al. [2008]). The single-sequence tasks cover protein function, property, and structure prediction. We report each benchmark with the metric used by its original source and provide full dataset descriptions, splits, and metric rationale in appendices E.1, E.2 and E.4.

ProPRINT achieves state-of-the-art on Golden-PPI and SKEMPI. On the five PPI benchmarks (table 1), ProPRINT-3B was best on Golden-PPI (AUROC 0.690, +16.2% over second-best MINT at 0.594) and SKEMPI (Pearson r 0.505, vs. MINT 0.488 and ESM2-3B 0.485). The Golden-PPI gain is notable given its controlled sequence-identity thresholds to prevent leakage, suggesting ProPRINT learned generalizable interaction features rather than memorizing sequence similarity. ESM2-3B led on Human-PPI (0.904), MINT on Yeast-PPI (0.656), and ESM2-150M on Mutational PPI (0.378).

No degradation on single-sequence tasks. On eight single-sequence benchmarks (table 2), ProPRINT-3B achieved the best result on four – Cloning classification (0.858), Fitness prediction (0.904), Material production (0.872), and Solubility prediction (0.769) – outperforming all ESM-2 variants and MINT despite allocating 66% of its training budget to PPI objectives, with ESM-2 variants leading on the remaining four. This suggests PPI-augmented pre-training acts as a beneficial regularizer for representation learning of single sequences rather than introducing negative transfer.

Table 1: Comparison of protein language models for five PPI tasks. Values reported as mean $\pm$ std. Bold indicates best result per column.

Model	Mode	Golden-PPI (AUROC $\uparrow$)	Human-PPI (Accuracy $\uparrow$)	Yeast-PPI (Accuracy $\uparrow$)	SKEMPI (Pearson r $\uparrow$)	Mutational PPI (AUPRC $\uparrow$)
ESM2-8M	IP	0.584 $\pm$.000	0.800 $\pm$.000	0.605 $\pm$.001	0.449 $\pm$.007	0.284 $\pm$.010
ESM2-35M	IP	0.582 $\pm$.000	0.860 $\pm$.000	0.564 $\pm$.001	0.457 $\pm$.009	0.339 $\pm$.011
ESM2-150M	IP	0.587 $\pm$.000	0.878 $\pm$.000	0.568 $\pm$.000	0.486 $\pm$.009	0.378 $\pm$.010
ESM2-650M	IP	0.585 $\pm$.000	0.870 $\pm$.001	0.653 $\pm$.001	0.478 $\pm$.006	0.353 $\pm$.010
ESM2-3B	IP	0.589 $\pm$.000	0.904 $\pm$.000	0.591 $\pm$.001	0.485 $\pm$.008	0.366 $\pm$.009
MINT	IP	0.594 $\pm$.000	0.878 $\pm$.000	0.656 $\pm$.001	0.488 $\pm$.010	0.357 $\pm$.011
MINT	PIP	0.593 $\pm$.000	0.728 $\pm$.000	0.639 $\pm$.001	0.467 $\pm$.020	0.278 $\pm$.008
ProPRINT-3B	IP	0.690 $\pm$.006	0.888 $\pm$.024	0.593 $\pm$.051	0.505 $\pm$.018	0.354 $\pm$.029
ProPRINT-3B	PIP	0.678 $\pm$.008	0.864 $\pm$.033	0.577 $\pm$.034	0.404 $\pm$.075	0.312 $\pm$.022

IP - inpainting, PIP - pair inpainting.

Table 2: Comparison of protein language models on single sequence downstream tasks. Values reported as mean $\pm$ std. Bold indicates best result per column.

Model	Cloning clf (ROC AUC $\uparrow$)	Enzyme cat. eff. (Pearson r $\uparrow$)	Fitness pred. (Spearman ρ $\uparrow$)	Fold pred. (Accuracy $\uparrow$)	Material prod. (ROC AUC $\uparrow$)	Metal ion bind. (Accuracy $\uparrow$)	Solubility pred. (Accuracy $\uparrow$)	Stability pred. (Spearman ρ $\uparrow$)
ESM2-8M	0.816 $\pm$.007	0.679 $\pm$.019	0.799 $\pm$.011	0.567 $\pm$.009	0.855 $\pm$.006	0.737 $\pm$.015	0.705 $\pm$.006	0.746 $\pm$.016
ESM2-35M	0.829 $\pm$.006	0.687 $\pm$.019	0.799 $\pm$.010	0.655 $\pm$.008	0.860 $\pm$.006	0.758 $\pm$.013	0.719 $\pm$.006	0.774 $\pm$.018
ESM2-150M	0.827 $\pm$.006	0.699 $\pm$.018	0.808 $\pm$.010	0.691 $\pm$.008	0.865 $\pm$.005	0.764 $\pm$.013	0.732 $\pm$.006	0.766 $\pm$.017
ESM2-650M	0.829 $\pm$.006	0.713 $\pm$.019	0.824 $\pm$.010	0.732 $\pm$.008	0.864 $\pm$.006	0.775 $\pm$.014	0.738 $\pm$.006	0.709 $\pm$.024
ESM2-3B	0.828 $\pm$.004	0.691 $\pm$.006	0.830 $\pm$.003	0.745 $\pm$.005	0.861 $\pm$.005	0.789 $\pm$.010	0.745 $\pm$.005	0.749 $\pm$.038
MINT	0.821 $\pm$.007	0.700 $\pm$.018	0.824 $\pm$.010	0.690 $\pm$.008	0.859 $\pm$.006	0.752 $\pm$.016	0.733 $\pm$.006	0.718 $\pm$.026
ProPRINT-3B	0.858 $\pm$.006	0.695 $\pm$.019	0.904 $\pm$.006	0.676 $\pm$.008	0.872 $\pm$.006	0.778 $\pm$.014	0.769 $\pm$.006	0.731 $\pm$.027

Inpainting mode yields stronger probing performance. Across both PPI and single-sequence tasks, inpainting-mode embeddings (independent per-sequence encoding) consistently outperformed pair inpainting mode under the MLP probe, consistent with MINT. The pair inpainting mode may require different downstream architectures (e.g., attention over the joint representation) to fully exploit its contextualized embeddings.

Training efficiency. ProPRINT-3B’s 508K training steps are comparable to ESM2-3B (500K) but far fewer than MINT’s 1M fine-tuning steps on ESM2-650M. Validation loss was still decreasing across all data sources at the end of training (fig. A4), indicating further training would likely help.

Ablation studies. We conducted extensive ablations at the 1B-parameter scale to guide the design of ProPRINT-3B. These experiments show that multi-task pre-training consistently outperforms single-task inpainting and pair inpainting baselines, and that the contact-aware masking strategy improves performance over uniform masking. Moreover, a token-matched 3B single-sequence-only control confirms that pair inpainting data drives gains on 3 of 5 PPI tasks and 6 of 8 single-sequence tasks at final scale (appendix F.3). Full results and analysis are provided in appendix F.

4 Conclusion

We present ProPRINT, a 3B-parameter T5 encoder-decoder protein language model that integrates protein-protein interaction data directly into pre-training through a novel pair inpainting objective alongside standard single-sequence masked infilling. We curated SynthPPI – a dataset of 74.6 million interacting domain pairs extracted from predicted structures in the AlphaFold Protein Structure Database that serves as a scalable alternative to limited experimental PPI data – and introduce a contact-aware masking strategy that injects structural binding-site information into sequence-only models without architectural changes or explicit structure inputs. ProPRINT-3B achieved state-of-the-art performance on Golden-PPI and SKEMPI while leading on 4 of 8 single-sequence benchmarks, showing that PPI-augmented pre-training yields representations that are simultaneously stronger for interaction tasks and competitive for property prediction. Promising future directions for this work include exploring adaptive curriculum strategies that dynamically tune the single-sequence and PPI data mixing proportions per task, and extending contact-aware masking to multi-chain complexes and other structural modalities.

Acknowledgments

We thank Tommi Jaakkola for his guidance, and Kyle Serikawa for helping with the various approvals needed for this work.

References

- Uniprot: the universal protein knowledgebase in 2025. *Nucleic acids research*, 53(D1):D609–D617, 2025.
- Judith Bernett, David B Blumenthal, and Markus List. Cracking the black box of deep sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics*, 25(2):bbae076, 2024.
- Bo Chen, Xingyi Cheng, Pan Li, Yangli-ao Geng, Jing Gong, Shen Li, Zhilei Bei, Xu Tan, Boyan Wang, Xin Zeng, et al. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint arXiv:2401.06199*, 2024.
- Yao Cheng, Haobo Wang, Hua Xu, Yuan Liu, Bin Ma, Xuemin Chen, Xin Zeng, Xianghe Wang, Bo Wang, Carina Shiao, et al. Co-evolution-based prediction of metal-binding sites in proteomes by machine learning. *Nature chemical biology*, 19(5):548–555, 2023.
- UniProt Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212, 2015.
- Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling atomistic protein binder design with generative pretraining and test-time compute. In *The Fourteenth International Conference on Learning Representations*, 2026.
- Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghaliya Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: toward understanding the language of life through self-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):7112–7127, 2021.
- Yanzhi Guo, Lezheng Yu, Zhining Wen, and Menglong Li. Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Research*, 36(9):3025–3030, 05 2008. ISSN 0305-1048.
- Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025.
- 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, 2024.
- Justina Jankauskaitė, Brian Jiménez-García, Justas Dapkūnas, Juan Fernández-Recio, and Iain H Moal. Skempi 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics*, 35(3):462–469, 2019.
- Meghana Kshirsagar, Artur Meller, Ian R Humphreys, Samuel Sledzieski, Yixi Xu, Rahul Dodhia, Eric Horvitz, Bonnie Berger, Gregory R Bowman, Juan Lavista Ferres, et al. Rapid and accurate prediction of protein homo-oligomer symmetry using seq2symm. *Nature Communications*, 16(1):2017, 2025.
- Andy M Lau, Nicola Bordin, Shaun M Kandathil, Ian Sillitoe, Vaishali P Waman, Jude Wells, Christine A Orenge, and David T Jones. Exploring structural diversity across the protein universe with the encyclopedia of domains. *Science*, 386(6721):eadq4946, 2024.
- Feiran Li, Le Yuan, Hongzhong Lu, Gang Li, Yu Chen, Martin KM Engqvist, Eduard J Kerkhoven, and Jens Nielsen. Deep learning-based k cat prediction enables improved enzyme-constrained model reconstruction. *Nature catalysis*, 5(8):662–672, 2022.

- 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.
- Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.
- 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, May 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1.
- Tatsuya Niwa, Bao-Wen Ying, Kazuaki Saito, Wuzhang Jin, Shoji Takada, Takuya Ueda, and Hideki Taguchi. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of escherichia coli proteins. *Proceedings of the National Academy of Sciences*, 106(11):4201–4206, 2009.
- Xiao-Yong Pan, Ya-Nan Zhang, and Hong-Bin Shen. Large-scale prediction of human protein-protein interactions from amino acid sequence based on latent topic features. *Journal of proteome research*, 9(10):4992–5001, 2010.
- Kim D. Pruitt, Tatiana Tatusova, and Donna R. Maglott. Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(suppl_1) : D501 – –D504, 012005. ISSN 0305 – 1048.
- Lorna Richardson, Ben Allen, Germana Baldi, Martin Beracochea, Maxwell L Bileschi, Tony Burdett, Josephine Burgin, Juan Caballero-Pérez, Guy Cochrane, Lucy J Colwell, et al. Mgnify: the microbiome sequence data analysis resource in 2023. *Nucleic acids research*, 51(D1):D753–D759, 2023.
- Gabriel J Rocklin, Tamuka M Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K Mulligan, Aaron Chevalier, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017.
- Jane C Siwek, Alisa A Omelchenko, Prabal Chhibbar, Sanya Arshad, AnnaElaine Rosengart, Iliyan Nazarali, Akash Patel, Kiran Nazarali, Javad Rahimikollu, Jeremy S Tilstra, et al. Sliding window interaction grammar (swing): a generalized interaction language model for peptide and protein interactions. *Nature Methods*, 22(8):1707–1719, 2025.
- Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. *BioRxiv*, pages 2023–10, 2023.
- 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.
- Damian Szklarczyk, Katerina Nastou, Mikaela Koutrouli, Rebecca Kirsch, Farrokh Mehryary, Radja Hachilif, Dewei Hu, Matteo E Peluso, Qingyao Huang, Tao Fang, et al. The string database in 2025: protein networks with directionality of regulation. *Nucleic acids research*, 53(D1):D730–D737, 2025.
- Varun Ullanat, Bowen Jing, Samuel Sledzieski, and Bonnie Berger. Learning the language of protein-protein interactions. *Nature Communications*, 2026.
- 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.
- Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingsi Yeo, et al. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic acids research*, 52(D1):D368–D375, 2024.

Huilin Wang, Mingjun Wang, Hao Tan, Yuan Li, Ziding Zhang, and Jiangning Song. Predpcrys: Accurate prediction of sequence cloning, protein production, purification and crystallization propensity from protein sequences using multi-step heterogeneous feature fusion and selection. *PLOS ONE*, 9(8):1–17, 08 2014. doi: 10.1371/journal.pone.0105902. URL <https://doi.org/10.1371/journal.pone.0105902>.

Minghao Xu, Zuobai Zhang, Jiarui Lu, Zhaocheng Zhu, Yangtian Zhang, Ma Chang, Runcheng Liu, and Jian Tang. Peer: A comprehensive and multi-task benchmark for protein sequence understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35156–35173. Curran Associates, Inc., 2022.

A Impact Statement

This work aims to advance the understanding of protein-protein interactions (PPIs) through improved language model representations. Better modeling of PPIs has the potential to accelerate drug discovery, aid in the design of therapeutic proteins, and deepen our understanding of cellular signaling pathways. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

B Disclosure

Generative AI tools were used to assist with drafting this manuscript. All AI-generated content was reviewed and revised by the authors.

C Overview of work

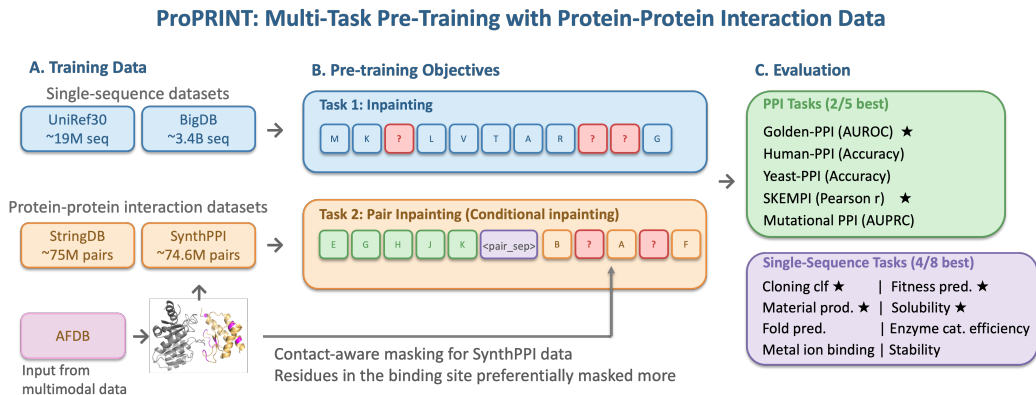


Figure A1: Overview of the ProPRINT model. The model is jointly pre-trained on two objectives: (i) single-sequence inpainting, and (ii) pair inpainting, where masked residues in one protein are infilled conditioned on its interacting partner.

D Training Data Details

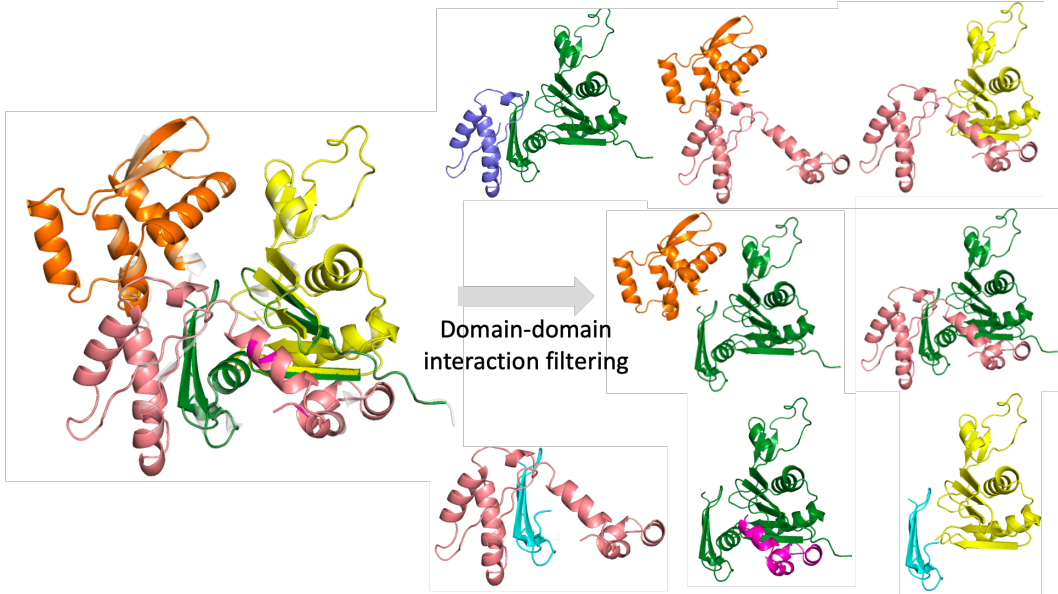


Figure A2: Curation of SynthPPI database: Cartoon representation of different domains present in the AlphaFold2 predicted structure AF-A0A059CIZ9-F1-model_v4 with each domain shown in a different color. We split the structure into multiple domain pairs which constitute our SynthPPI database. All pairs of domains are exhaustively sampled to identify interacting domains.

D.1 Pair Inpainting Data

StringDB. The STRING database Szklarczyk et al. [2025] compiles both direct physical protein-protein interactions and functional associations, drawing from diverse evidence sources including genomic neighborhood, gene fusion, co-occurrence, co-expression, text mining, curated databases, and experimental data. Each interaction is assigned a combined confidence score between 0 and 1000, aggregated across these evidence channels and corrected for the prior probability of observing an interaction by chance.

We used the physical interaction subset of STRING (version 2023), which retained only PPIs supported by direct evidence of physical binding. We applied two filters: (i) we retained only high-confidence pairs with a combined score greater than 700, and (ii) we required that the experimental evidence score was the dominant contributor to the combined score (`majority_experiments` filter). This strict filtering criterion was designed to minimize false-positive interactions, particularly those arising from text-mining co-occurrences.

SynthPPI. SynthPPI is a dataset derived from the AlphaFold Protein Structure Database (AFDB). This is a dataset of interacting protein-protein domains used as additional synthetic data for the pair inpainting task. We used only high and medium quality structures from the AFDB database as a starting point. Structures with average Predicted Local Distance Difference Test (Avg pLDDT) greater than equal to 70% are defined as high quality and structures with Avg pLDDT between 40-70% are defined as low quality. The following filtering criterion was used - i) Sequence of the predicted structure matches UniProt Consortium [2015] sequence ii) Sequence length is larger than 74 amino acids iii) Avg pLDDT $\geq 70\%$ iv) For predicted structures with Avg pLDDT in between 40-70%, include only those predicted as monomer by seq2symm Kshirsagar et al. [2025].

We obtained 113 million structures from AFDB after filtering and extracted corresponding domain information using domain information from the TED database (version 4) Lau et al. [2024]. We defined two domains to be interacting if there are at least 15 residue pairs making a contact of 8 Å or less for the $C\beta$ atoms ($C\alpha$ for Glycine residue). We checked whether there is an interaction or not for each pair of domains in a structure (fig. A2). After processing, the database contained

about 415 million protein-protein interaction data. We further clustered the data based on 90% sequence similarity to obtain a clustered SynthPPI dataset used for pre-training containing 74,557,665 protein-protein interactions. Overall SynthPPI is much more diverse than StringDB in terms of unique sequences (table A1). The protein sequences are shorter in length than StringDB since they are obtained from domains.

D.2 Data holdout

To prevent data leakage and maintain fair evaluation, we first sampled random subsets (5-20K entries, proportional to each database’s size) from all four source datasets to form holdout sets for validation and test. For pair inpainting data, we removed from the training set any protein-protein pair in which either sequence appeared in the holdout set. These holdout samples were then pooled, and any exact sequence duplicates were filtered from the training data across all databases via sequence-level alignment.

Table A1: Summary of size and number of training tokens in the four pre-training datasets. PPI pairs is not applicable for single-sequence datasets.

Dataset	Task	Unique seq.	PPI pairs	Total tokens
UniRef30	Inpainting	~19.2M	–	~5.8B
BigDB	Inpainting	~3.4B	–	~781.9B
StringDB	Pair inpainting	~10.3M	~75.0M	~29.0B
SynthPPI	Pair inpainting	~106.8M	~74.6M	~17.6B

D.3 Training Details

We trained a 3B-parameter T5 model using the Megatron-LM framework for a total of 200B tokens on 32 NVIDIA H200 Tensor Core GPUs. The training data composition was 34% inpainting, 33% StringDB and 33% SynthPPI. Protein sequences were tokenized at single-residue granularity, where each amino acid corresponded to one token, and packed to a context length of 12,288 tokens.

We used a cosine learning rate schedule with a peak learning rate of 5×10^{-4} , a minimum learning rate of 5×10^{-6} , and 2,000 warmup steps. Optimization was performed with AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-8}$), a weight decay of 0.1, and gradient clipping at 1.0. In the decoder, we added QK normalization in the cross-attention layers to improve training stability, which enabled $2\text{--}2.5\times$ higher learning rates without divergence. We applied an attention dropout of 0.1 and a decoder dropout of 0.1. Training used BF16 mixed precision with a micro-batch size of 1 and a global batch size of 32. We evaluated every 2,000 steps. The full training run completed in 508626 steps and in approximately 6 days.

D.4 Ablation Study: 1B Model Training

For ablation studies, we trained smaller 1B-parameter variants of our model on a budget of 32B tokens for 67349 training steps using 1 node (8 NVIDIA H200 GPUs). We used FP16 mixed precision with a micro-batch size of 1 and a global batch size of 8. Sequences were packed to a context length of 59,392 tokens. The peak learning rate was set to 1×10^{-4} with a minimum learning rate of 1×10^{-5} , following the same cosine schedule and warmup strategy as the 3B model. We disabled both attention and decoder dropout in the 1B models, as we observed training instability when dropout was enabled at this scale. All other hyperparameters, including the optimizer configuration, gradient clipping, and training objectives, remained identical to the 3B setup described in the main text.

Table A2: Architecture details for the 3B and 1B model variants. Both models use a T5 encoder-decoder architecture with RoPE positional embeddings, RMSNorm, SwiGLU activations, and QK-LayerNorm.

HYPERPARAMETER	3B	1B
HIDDEN SIZE	2048	1024
NUM LAYERS	24	24
ATTENTION HEADS	32	16
KV CHANNELS	64	64
FFN HIDDEN SIZE	5504	2816
SEQUENCE LENGTH	2048	2048
VOCAB SIZE	255	255
POSITION EMBEDDING	RoPE	RoPE
ROTARY BASE	10000	10000
NORMALIZATION	RMSNORM	RMSNORM
ACTIVATION	SWIGLU	SWIGLU

E Evaluation data details

E.1 Evaluation Datasets

We hypothesize that augmenting PLM pre-training with PPI data results in better representation learning. We test our hypothesis by evaluating ProPRINT on eight single-sequence and five PPI tasks. The five PPI benchmarks span two task categories.

3-sequence PPI tasks. These tasks involved a wild-type (WT) sequence, a mutant sequence, and an interacting partner protein.

- **SKEMPI** Jankauskaitė et al. [2019]: A regression task predicting changes in protein-protein binding free energy ($\Delta\Delta G$) upon mutation. The dataset contained over 7,000 mutations across structurally resolved complexes from the PDB.
- **Mutational PPI** Siwek et al. [2025]: A binary classification task predicting whether a single-point mutation at a protein-protein interface disrupted interaction. The dataset contained 6,300 entries, of which approximately 76% were disrupted interactions. We evaluated with 5-fold cross-validation grouped by wild-type sequence and seed averaging, reporting AUPRC. This dataset is heavily imbalanced (76% disrupted interactions). AUPRC is more informative than AUROC for imbalanced datasets because it focuses on the precision-recall tradeoff for the minority class.

2-sequence PPI tasks. These binary tasks involved predicting whether two protein sequences interact or not. All tasks were evaluated with bootstrap aggregation.

- **Golden-PPI** Bennett et al. [2024]: A gold-standard binary PPI prediction benchmark with approximately 274K curated pairs across three intra-species splits and controlled sequence identity thresholds to prevent data leakage. The positive-negative ratio in this dataset is well-balanced and the original benchmark paper reports AUROC. We follow their convention. AUROC measures discrimination ability across all thresholds and is appropriate when classes are reasonably balanced. None of the pairs of sequences in this dataset was present in the training set for StringDB and SynthPPI.
- **Yeast-PPI** Xu et al. [2022], Guo et al. [2008]: A binary PPI prediction benchmark among *S. cerevisiae* proteins from TorchDrug, containing approximately 5,000 training pairs with fixed splits. We report accuracy for this task following TorchDrug convention.
- **Human-PPI** Xu et al. [2022], Pan et al. [2010]: A binary PPI prediction benchmark among human proteins from TorchDrug, containing over 35,000 training pairs with fixed splits. We report accuracy for this task following TorchDrug convention.

In addition to the PPI tasks, we evaluated ProPRINT on eight single-sequence downstream tasks spanning protein function, property and structure prediction (appendix E.4).

E.2 PPI Task Dataset Compositions

Table A3: Dataset splits and class distributions for the four binary PPI benchmarks. SKEMPI is a regression task and is excluded. Mutational PPI uses 5-fold cross-validation with no predefined splits.

Dataset	Train	Valid	Test	Positive	Negative
Golden-PPI	163,019	59,260	52,048	137,166	137,161
Human-PPI	26,319	234	180	18,520	17,701
Yeast-PPI	4,945	95	394	2,688	2,746
Mutational PPI	No predefined split - 6,306 total			4,833	1,473

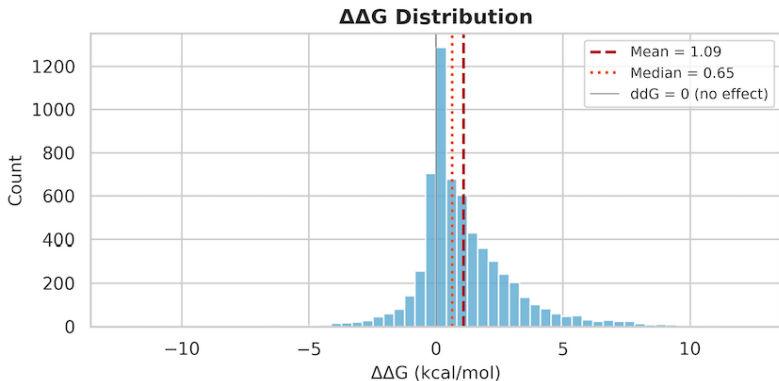


Figure A3: $\Delta\Delta G$ distribution for SKEMPI dataset.

E.3 Evaluation Details

E.3.1 Generation of Embeddings for PPI pairs

Embeddings were computed using a frozen encoder. We evaluated three protein language models: ESM2, MINT, and our model (ProPRINT). For ProPRINT, we used the inpainting and pair inpainting task mode when extracting embeddings for single sequence and protein-protein pairs respectively.

3-sequence PPI tasks. For single-sequence models (ESM2) or inpainting mode, each of the three sequences—wild-type, mutant, and binding partner—was encoded independently. Residue-level embeddings for wild-type (or mutant) and binding partner were concatenated and mean-pooled to obtain sequence-level representations for the protein-protein pairs.

For pair-inpainting mode, the wild-type or mutant sequence was concatenated with the interacting partner and encoded jointly. Residue-level embeddings for the protein complexes were mean-pooled together to obtain representations for the protein-protein pairs.

We further took a difference between the wild-type and mutant embeddings to isolate the effect of the mutation. This strategy reduced confounding effects from protein identity and focused the representation on the mutational effect.

2-sequence PPI tasks. For single-sequence models or inpainting mode, each protein in the pair was encoded independently. The embeddings were then concatenated and mean-pooled to obtain protein-protein pair embeddings. For pair-inpainting models, both sequences were concatenated and encoded jointly, followed by mean pooling.

E.3.2 Probing Head

We used a 2-layer MLP probing head (hidden dimension 640, ReLU activation) optimized with AdamW (learning rate 10^{-3} , weight decay 10^{-4}), following the setup from MINT Ullanat et al. [2026]. For datasets with predefined splits (Golden-PPI, Human-PPI, Yeast-PPI), we used the provided train/validation/test partitions. For datasets without predefined splits (SKEMPI, Mutational

PPI), we performed group 5-fold cross-validation, where groups were defined by PDB complex ID for SKEMPI and Interactor ID for Mutational PPI. In each fold, one-fifth of the groups were held out as the test split, another fifth as validation, and the remaining groups as training data. We tested each group only once. Each MLP was trained for 500 epochs per fold. For SKEMPI which is a regression task, we used mean-squared error loss, and for the other four PPI prediction tasks, we used binary cross-entropy loss.

All experiments were repeated across 20 random seeds, which controlled both the fold assignment and MLP weight initialization. We used predefined splits where available and group 5-fold cross-validation otherwise. We report metrics for predictions on the test set corresponding to the epoch with the highest validation metric. For cross-validation tasks, overall out-of-fold metrics were aggregated per seed with 95% t -distribution confidence intervals. For fixed-split tasks, metrics were aggregated across seeds with 95% bootstrapped confidence intervals.

E.4 Single-Sequence Downstream Tasks

Train-valid-test splits for all eight tasks were downloaded from <https://huggingface.co/biomap-research>.

Fitness Prediction Xu et al. [2022]. This task involves predicting the fitness landscape of the GB1 domain, assessing how sequence variants affected binding affinity and stability. The dataset captures combinatorial mutations and their measured fitness values. We report Spearman rank correlation coefficient (Spearman ρ).

Enzyme Catalytic Activity Li et al. [2022]. This task involves predicting k_{cat} values (enzymatic turnover numbers) representing the maximum chemical conversion rate for metabolic enzymes across diverse organisms. We report Pearson correlation coefficient (Pearson r).

Stability Prediction Rocklin et al. [2017]. This task involves predicting the protease concentration at which a protein retained its folded state, serving as a proxy for thermodynamic stability. We report Spearman rank correlation coefficient (Spearman ρ).

Solubility Prediction Niwa et al. [2009]. A binary classification task predicting whether a protein is soluble or insoluble, evaluated on a diverse set of proteins. We report accuracy.

Cloning Classification Wang et al. [2014]. Protein crystals for X-ray crystallography undergo a series of step during experimental structure prediction – including selection, expression, purification, and crystal structure determination. This task predicts whether a protein was stable at a given experimental stage, based on stage-specific labels. We report ROC AUC.

Material Production Wang et al. [2014]. This task predicts whether a protein sequence fails at the material production stage during experimental characterization. We report ROC AUC.

Metal Ion Binding Cheng et al. [2023]. Metal ion binding is one of the processes through which proteins function in cellular environments. This task predicts whether a protein bound metal ions. We report accuracy.

Fold Classification Chen et al. [2024]. A multi-class classification task assigning protein sequences to one of 1,195 known structural folds. We report accuracy.

E.5 Probing head for single-sequence tasks

We used a MLP as a linear probe to compare the learned protein representations. The per-residue embeddings were averaged, projected into 128 dimensions followed by ReLU activation and then passed to the output layer. Each MLP was trained for 2,000 epochs using AdamW. We selected test predictions from the epoch with the highest validation metric, and results were aggregated across 20 random seeds with 95% bootstrapped confidence intervals.

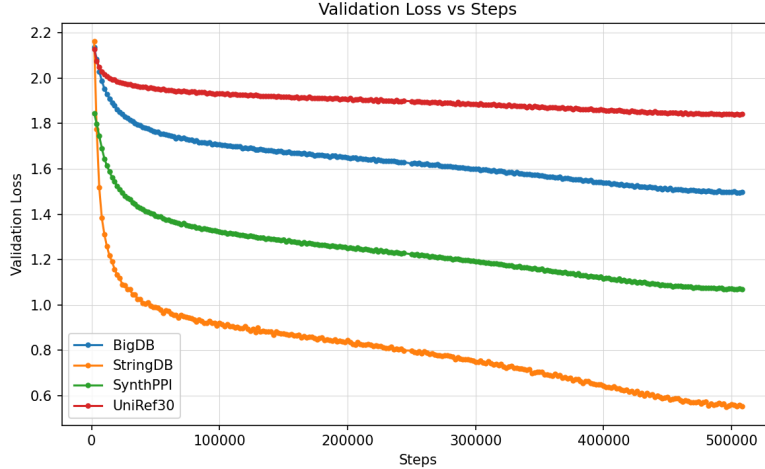


Figure A4: Validation loss curves for ProPRINT. Loss is still decreasing indicating model is under-trained.

F Ablation Studies

F.1 Effect of Data Mixing Proportions

We performed ablation studies with seven 1B models trained on 100B tokens but with different data proportions to understand the effect of different proportions of inpainting and pair inpainting data during pre-training. We evaluated the trained models on the 5 PPI tasks (table A4) and 8 single sequence tasks (table A5).

Multi-task training improved PPI benchmarks over single-task inpainting and pair inpainting baselines. Ablation experiments which included a mixture of the two tasks overall performed better for the PPI tasks compared to models trained with 100% inpainting or 100% pair inpainting data. There wasn't a single data composition that performed the best across all the 5 tasks, indicating that combining both PPI data sources StringDB and SynthPPI with single-sequence data yielded broadly strong representations. We made a similar observation for single sequence downstream tasks as well where no single data composition was a clear winner (table A5). We decided to train a larger 3B model on 34% Inpainting + 66% Pair inpainting (33% StringDB + 33% SynthPPI) data composition.

Table A4: 1B ablation experiments: PPI task metrics across the two task modes and for different task and data mixing proportions. Values reported as mean $\pm$ std over 20 seeds. Bold indicates best result per column across both modes.

Mode	Task proportion	Golden-PPI (AUROC $\uparrow$)	Human-PPI (Accuracy $\uparrow$)	Yeast-PPI (Accuracy $\uparrow$)	SKEMPI (Pearson r $\uparrow$)	Mutational PPI (AUPRC $\uparrow$)
Inpainting	100% IP	0.673 $\pm$.006	0.829 $\pm$.028	0.578 $\pm$.033	0.478 $\pm$.019	0.300 $\pm$.013
	70% IP + 15% Str + 15% Syn	0.681 $\pm$.005	0.852 $\pm$.027	0.551 $\pm$.031	0.483 $\pm$.028	0.337 $\pm$.023
	50% IP + 50% Str	0.671 $\pm$.004	0.854 $\pm$.025	0.579 $\pm$.044	0.465 $\pm$.017	0.316 $\pm$.033
	50% IP + 50% Syn	0.668 $\pm$.006	0.856 $\pm$.031	0.554 $\pm$.037	0.478 $\pm$.020	0.349 $\pm$.023
	50% IP + 25% Str + 25% Syn	0.677 $\pm$.005	0.856 $\pm$.025	0.549 $\pm$.031	0.468 $\pm$.017	0.297 $\pm$.025
	34% IP + 33% Str + 33% Syn	0.684 $\pm$.004	0.870 $\pm$.025	0.561 $\pm$.036	0.476 $\pm$.037	0.333 $\pm$.022
	100% PIP (50% Str + 50% Syn)	0.669 $\pm$.006	0.844 $\pm$.025	0.556 $\pm$.036	0.470 $\pm$.017	0.289 $\pm$.024
Pair inpainting	100% IP	0.671 $\pm$.005	0.821 $\pm$.029	0.550 $\pm$.031	0.435 $\pm$.017	0.281 $\pm$.020
	70% IP + 15% Str + 15% Syn	0.671 $\pm$.003	0.871 $\pm$.025	0.543 $\pm$.029	0.441 $\pm$.054	0.302 $\pm$.019
	50% IP + 50% Str	0.668 $\pm$.005	0.866 $\pm$.023	0.565 $\pm$.041	0.437 $\pm$.019	0.369 $\pm$.029
	50% IP + 50% Syn	0.672 $\pm$.004	0.869 $\pm$.027	0.555 $\pm$.032	0.436 $\pm$.037	0.345 $\pm$.021
	50% IP + 25% Str + 25% Syn	0.679 $\pm$.006	0.875 $\pm$.023	0.558 $\pm$.042	0.449 $\pm$.025	0.330 $\pm$.023
	34% IP + 33% Str + 33% Syn	0.666 $\pm$.005	0.875 $\pm$.023	0.566 $\pm$.038	0.464 $\pm$.030	0.319 $\pm$.017
	100% PIP (50% Str + 50% Syn)	0.663 $\pm$.006	0.844 $\pm$.026	0.547 $\pm$.031	0.453 $\pm$.023	0.283 $\pm$.017

IP - inpainting, PIP - pair inpainting, Str - StringDB, Syn - SynthPPI.

Table A5: 1B ablation experiments: Single sequence evaluation metrics. Values reported as mean $\pm$ std over 20 seeds. Bold indicates best result per column.

Task proportion	Cloning clf (ROC AUC $\uparrow$)	Enzyme cat. eff. (Pearson r $\uparrow$)	Fitness pred. (Spearman ρ $\uparrow$)	Fold pred. (Accuracy $\uparrow$)	Material prod. (ROC AUC $\uparrow$)	Metal ion bind. (Accuracy $\uparrow$)	Solubility pred. (Accuracy $\uparrow$)	Stability pred. (Spearman ρ $\uparrow$)
100% IP	0.849 $\pm$.006	0.693 $\pm$.018	0.862 $\pm$.008	0.635 $\pm$.009	0.866 $\pm$.006	0.761 $\pm$.013	0.746 $\pm$.006	0.717 $\pm$.012
70% IP + 15% Str + 15% Syn	0.843 $\pm$.006	0.699 $\pm$.018	0.877 $\pm$.007	0.625 $\pm$.009	0.868 $\pm$.006	0.755 $\pm$.013	0.752 $\pm$.006	0.759 $\pm$.012
50% IP + 50% Str	0.843 $\pm$.006	0.695 $\pm$.018	0.860 $\pm$.009	0.597 $\pm$.009	0.864 $\pm$.006	0.759 $\pm$.014	0.750 $\pm$.006	0.721 $\pm$.012
50% IP + 50% Syn	0.835 $\pm$.006	0.692 $\pm$.019	0.863 $\pm$.008	0.645 $\pm$.008	0.866 $\pm$.006	0.766 $\pm$.014	0.741 $\pm$.007	0.685 $\pm$.020
50% IP + 25% Str + 25% Syn	0.841 $\pm$.006	0.691 $\pm$.019	0.874 $\pm$.007	0.628 $\pm$.009	0.865 $\pm$.005	0.753 $\pm$.014	0.754 $\pm$.006	0.735 $\pm$.016
34% IP + 33% Str + 33% Syn	0.841 $\pm$.006	0.697 $\pm$.019	0.871 $\pm$.008	0.607 $\pm$.009	0.865 $\pm$.005	0.760 $\pm$.013	0.747 $\pm$.006	0.712 $\pm$.013
100% PIP (50% Str + 50% Syn)	0.828 $\pm$.006	0.688 $\pm$.019	0.873 $\pm$.008	0.591 $\pm$.010	0.854 $\pm$.006	0.751 $\pm$.014	0.734 $\pm$.006	0.730 $\pm$.016

IP = inpainting, PIP = pair inpainting, Str = StringDB pair inpainting, Syn = SynthPPI pair inpainting.

F.2 Effect of Contact-Aware Masking

We compared two masking strategies for SynthPPI data for a 3B model trained with 100B tokens and under the same 50% Inpainting + 25% StringDB + 25% SynthPPI data mixing proportion (table A6). The training was completed in 127156 steps. The contact-aware strategy applied a separate Beta-linear masking function ($\alpha=5$, $\beta=5$) to residues at the binding interface and a lighter masking function ($\alpha=2$, $\beta=8$) to the remaining residues, while the uniform strategy applied the same Beta-linear masking ($\alpha=5$, $\beta=8$) across the entire second sequence.

Contact-aware masking consistently outperformed uniform masking across all five PPI benchmarks under both task modes. These results support our hypothesis that preferentially masking residues at binding interfaces helps the model learn stronger inter-protein dependencies, as there are stronger correlations between the two interacting proteins at the binding interface. Contact-aware masking also improved performance on most single sequence downstream tasks; full results are provided in table A7. This finding is further supported by the validation loss curves where contact-aware masking resulted in lower loss values, particularly for the SynthPPI dataset (fig. A5).

Table A6: Effect of contact-aware masking on PPI tasks across both task modes (50/25/25 data mix). Values reported as mean $\pm$ std over 20 seeds. Bold indicates better result per column within each mode.

Mode	Masking strategy	Golden-PPI (AUROC $\uparrow$)	Human-PPI (Accuracy $\uparrow$)	Yeast-PPI (Accuracy $\uparrow$)	SKEMPI (Pearson r $\uparrow$)	Mutational PPI (AUPRC $\uparrow$)
Inpainting	Contact-aware	0.672 $\pm$.004	0.876 $\pm$.023	0.566 $\pm$.048	0.481 $\pm$.014	0.342 $\pm$.031
	Uniform	0.668 $\pm$.007	0.841 $\pm$.027	0.545 $\pm$.032	0.466 $\pm$.045	0.329 $\pm$.038
Pair inpainting	Contact-aware	0.669 $\pm$.005	0.870 $\pm$.023	0.594 $\pm$.028	0.438 $\pm$.033	0.304 $\pm$.017
	Uniform	0.663 $\pm$.005	0.820 $\pm$.029	0.582 $\pm$.033	0.426 $\pm$.031	0.283 $\pm$.020

Table A7: Effect of contact-aware masking on single sequence downstream tasks (inpainting embeddings, 50 Inpainting/25 StrinDB/25 SynthPPI data mix, 3B model trained on 100B tokens). Values reported as mean $\pm$ std over 20 seeds. Bold indicates better result per column.

Masking strategy	Cloning clf (ROC AUC $\uparrow$)	Enzyme cat. eff. (Pearson r $\uparrow$)	Fitness pred. (Spearman ρ $\uparrow$)	Fold pred. (Accuracy $\uparrow$)	Material prod. (ROC AUC $\uparrow$)	Metal ion bind. (Accuracy $\uparrow$)	Solubility pred. (Accuracy $\uparrow$)	Stability pred. (Spearman ρ $\uparrow$)
Contact-aware	0.841 $\pm$.006	0.696 $\pm$.018	0.853 $\pm$.009	0.641 $\pm$.008	0.870 $\pm$.005	0.766 $\pm$.014	0.750 $\pm$.006	0.732 $\pm$.021
Uniform	0.841 $\pm$.006	0.694 $\pm$.019	0.859 $\pm$.008	0.630 $\pm$.009	0.863 $\pm$.006	0.748 $\pm$.014	0.746 $\pm$.006	0.726 $\pm$.020

F.3 Contribution of PPI Data at Final Scale

To directly isolate the contribution of PPI data at final scale, we trained a token-matched 3B single-sequence-only control: an identical T5 encoder-decoder architecture with an identical training budget (200B tokens, 508K steps), but trained on 100% single-sequence inpainting data only (no pair inpainting, no StringDB, no SynthPPI). Comparing this control against ProPRINT-3B isolates the effect of the pair inpainting data, since the two models differ only in their training data composition.

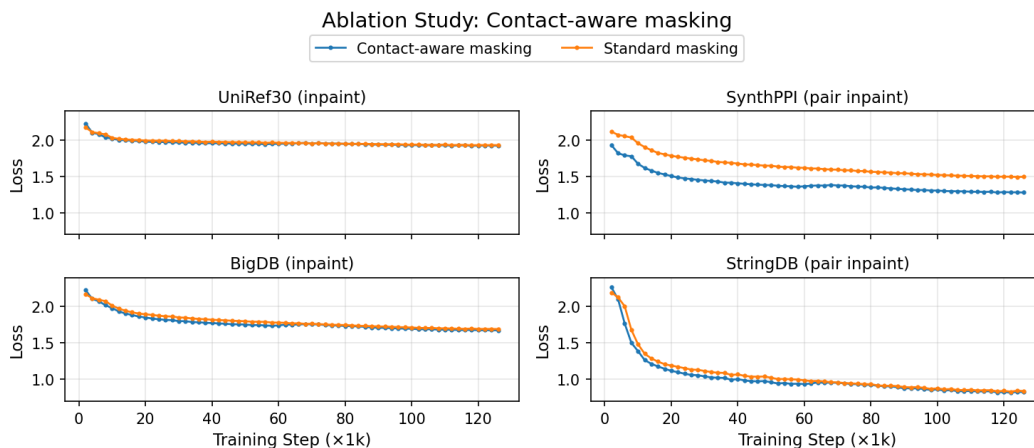


Figure A5: Validation loss curves comparing contact-aware masking versus uniform masking across four validation sets. Contact-aware masking achieves lower loss on all datasets, with the largest gains on SynthPPI pair inpainting, consistent with the hypothesis that preferentially masking binding-interface residues results in better representations for protein-protein pairs.

Results for the PPI tasks and single-sequence tasks are shown in tables A8 and A9 respectively, where Δ denotes the pair inpainting contribution (ProPRINT-3B minus control).

Table A8: Contribution of PPI data on PPI tasks: token-matched 3B single-sequence-only control (100% inpainting) versus ProPRINT-3B. Values reported as mean $\pm$ std over 20 seeds. Δ is the pair inpainting contribution (ProPRINT-3B – control).

Task	Metric	3B Control (inpainting only)	ProPRINT-3B	Δ (pair inpainting)
Golden-PPI	AUROC	0.676 $\pm$.005	0.690 $\pm$.006	+0.014
Human-PPI	Accuracy	0.849 $\pm$.026	0.888 $\pm$.024	+0.039
Yeast-PPI	Accuracy	0.603 $\pm$.029	0.593 $\pm$.051	−0.010
SKEMPI	Pearson r	0.482 $\pm$.017	0.505 $\pm$.018	+0.023
Mutational PPI	AUPRC	0.368 $\pm$.027	0.354 $\pm$.029	−0.014

Table A9: Contribution of PPI data on single-sequence tasks: token-matched 3B single-sequence-only control (100% inpainting) versus ProPRINT-3B. Values reported as mean $\pm$ std over 20 seeds. Δ is the pair inpainting contribution (ProPRINT-3B – control).

Task	Metric	3B Control (IP only)	ProPRINT-3B	Δ (pair inpainting)
Cloning clf	ROC-AUC	0.852 $\pm$.006	0.858 $\pm$.006	+0.007
Enzyme cat. eff.	Pearson r	0.651 $\pm$.022	0.695 $\pm$.019	+0.044
Fitness pred.	Spearman ρ	0.867 $\pm$.008	0.904 $\pm$.006	+0.037
Fold pred.	Accuracy	0.656 $\pm$.009	0.676 $\pm$.008	+0.020
Material prod.	ROC-AUC	0.874 $\pm$.005	0.872 $\pm$.006	−0.001
Metal ion bind.	Accuracy	0.768 $\pm$.016	0.778 $\pm$.014	+0.010
Solubility pred.	Accuracy	0.759 $\pm$.006	0.769 $\pm$.006	+0.010
Stability pred.	Spearman ρ	0.745 $\pm$.024	0.731 $\pm$.027	−0.014

Pair inpainting provides gains on 3 of 5 PPI tasks at final scale. Golden-PPI (+0.014), Human-PPI (+0.039), and SKEMPI (+0.023) all improve with PPI data. SKEMPI is the cleanest attribution: the control (0.482) sits below ESM2-3B (0.485), while pair inpainting lifts ProPRINT to 0.505, above all baselines in table 1.

PPI data acts as a beneficial regularizer for single-sequence tasks. Six of eight single-sequence tasks show positive contributions from pair inpainting, with the strongest gains on enzyme catalytic efficiency (+0.044) and fitness prediction (+0.037).

Yeast-PPI and Mutational PPI show no significant difference. Both are within noise across the control and ProPRINT-3B.

G Limitations

ProPRINT currently operates on sequence pairs; extending to multi-chain complexes and higher-order assemblies is a natural next step. The contact-aware masking relies on predicted structures from AFDB, inheriting any biases in AlphaFold’s structure predictions. Our evaluation used frozen-encoder probing; fine-tuning may improve performance. Finally, the model has not converged after 500k steps and that scaling compute would likely yield further improvements.