ProteomeLM: A Proteome-Scale Language Model Enables Accurate and Rapid Prediction of Protein-Protein Interactions and Gene Essentiality Across Taxa

Anonymous Authors[1]

Abstract

Language models trained on biological sequences are advancing inference tasks from the scale of single proteins to that of genomic neighborhoods. Here, we introduce ProteomeLM, a transformer-based language model that uniquely operates on entire proteomes from species spanning the tree of life. ProteomeLM is trained to reconstruct masked protein embeddings using the whole proteomic context, yielding contextualized protein representations that reflect proteome-scale functional constraints. Notably, ProteomeLM's attention coefficients encode protein-protein interactions (PPI), despite being trained without interaction labels. Furthermore, it enables interactome-wide PPI screening that is substantially more accurate, and orders of magnitude faster, than amino-acid coevolution-based methods. We further develop ProteomeLM-PPI, a supervised model that combines ProteomeLM embeddings and attention coefficients to achieve state-of-the-art PPI prediction across benchmarks and species. Finally, we introduce ProteomeLM-Ess, a supervised gene essentiality predictor that generalizes across diverse taxa. Our results demonstrate the potential of proteome-scale language models for addressing function and interactions at the organism level. Code and weights: `https://anonymous.4open.science/r/ProteomeLM-EB3F`.

1. Introduction

Protein language models trained with the masked language modeling (MLM) objective learn coevolution between amino acids, allowing them to capture protein structure (Rives et al., 2021; Rao et al., 2021; Lin et al., 2023; Elnaggar et al., 2022; Hayes et al., 2025). Genome language models extend this paradigm to nucleotide sequences (Nguyen et al., 2023; Dalla-Torre et al., 2025; Brixi et al., 2025; Hwang et al., 2024), but span at most a few megabases and cannot capture dependencies across entire genomes, especially in eukaryotes. Given the success of protein language models at capturing coevolution between amino acids, it is tantalizing to develop such models at the proteome scale, i.e. taking as input the ensemble of proteins encoded by a genome. Such models should capture coevolution between proteins, thereby generalizing over phylogenetic profiling methods (Pellegrini et al., 1999; Croce et al., 2019), making them well-suited to predict complete PPI networks and gene essentiality across diverse species. Moreover, such foundation models can also be used for other downstream applications where proteome context information is important, such as predicting gene essentiality (Gurumayum et al., 2021).

PPI are fundamental to most biological processes, including signal transduction, cellular metabolism, and immune responses. Knowing these interactions is critical for deciphering cellular processes and for developing therapeutic interventions. Sequence-based methods, in particular Direct Coupling Analysis (DCA) (Weigt et al., 2009), offer better scalability than structure-based approaches (Abramson et al., 2024; Bryant et al., 2022), but require paired multiple sequence alignments and train one model per candidate protein pair, limiting large-scale application. While such coevolution methods are often effective in bacteria and other well-represented clades, they struggle in eukaryotes or poorly sampled taxa (Cong et al., 2019; Humphreys et al., 2021; Zhang et al., 2024). Supervised sequence-based predictors (Sledzieski et al., 2021; Singh et al., 2022; Ko et al., 2024) improve accuracy, but rely on large labeled datasets.

In this paper, we introduce ProteomeLM, a transformer-based language model that uniquely reasons on entire proteomes spanning the tree of life (Figure 1). ProteomeLM leverages embeddings from ESM-Cambrian (ESM-C) (ESM Team et al., 2024), and integrates protein-level functional properties at the proteome scale via masked reconstruction. We show that its attention coefficients learn PPI in an un-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

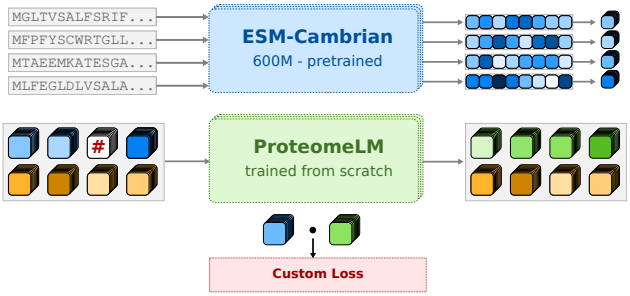

Figure 1. **ProteomeLM training.** Input amino-acid sequences are embedded through pretrained ESM-C, yielding a fixed-dimensional embedding for each protein. ProteomeLM is trained from scratch to predict the masked embeddings of proteins in the context of their proteome. Proteins are annotated by a functional encoding (orange) representing their orthologous group.

supervised way, that it screens whole interactomes orders of magnitude faster than DCA while substantially outperforming it, and that ProteomeLM embeddings improve supervised gene essentiality prediction over single-protein models.

2. Model

Architecture and training. ProteomeLM takes as input a proteome, i.e. the set of proteins encoded by a given genome, and aims to capture the functional and evolutionary signals present between proteins at the proteome level. Each protein is represented by an embedding generated by ESM-C (600M parameters) (ESM Team et al., 2024), obtained by averaging per-amino-acid embeddings to dimension 1152. We randomly mask 50% of the protein representations within a proteome, while their functional encodings are kept unmasked; the model is trained to reconstruct the original protein embeddings based on contextual signals from the rest of the proteome. The core of the model is the DistillBERT architecture with FlashAttention-2 (Dao, 2023). We trained four variants of ProteomeLM: XS (5.6M), S (36M), M (112M), and L (328M) parameters, on nearly 32,000 annotated proteomes from OrthoDB (Kuznetsov et al., 2022). Training remained stable across all model sizes, showing smooth convergence, and performance improved steadily from XS to M; the L model showed degraded performance, which we attribute to overfitting.

Functional encoding. ProteomeLM does not employ positional encoding along the genome, which sets it apart from existing genome language models. Instead, we propose a *functional encoding* based on OrthoDB orthology (Kuznetsov et al., 2022): for each protein, its orthologous group is represented by a hierarchical average of ESM-C embeddings computed bottom-up from leaf groups to the taxonomic root. During training, one level of this ancestral path is randomly sampled per epoch, exposing the model to functional descriptions at varying levels of specificity, and encouraging it to learn relationships between proteins that hold at multiple evolutionary scales. This functional description purely relies on statistical protein family and not on external annotations.

Polar loss. The standard MSE loss leads to a degenerate solution where the model simply reproduces the functional encoding. To address this, we decouple residual magnitude and direction and derive a custom *polar loss* via maximum likelihood:

$$\mathcal{L}_{\text{Polar}}(\hat{x}, x, \bar{x}) = (1 - \cos\theta) + \left(\|\hat{r}\| - \|r\|\right)^2, \quad (1)$$

where $r = x - \bar{x}$, $\hat{r} = \hat{x} - \bar{x}$, and θ is the angle between r and $\hat{r}$. The polar loss is minimized if and only if $\hat{x} = x$, and its angular gradient is independent of $\|\hat{r}\|$, preventing gradient collapse.

3. Results

3.1. Unsupervised PPI Recovery

We examine the attention coefficients of ProteomeLM (Vig et al., 2021; Rao et al., 2021) to assess whether they capture PPI without any interaction labels. We use the D-SCRIPT dataset (Sledzieski et al., 2021), derived from STRING (Szklarczyk et al., 2019) and focused on experimentally validated physical interactions, across six species. Figure 2 shows that many attention heads are predictive of interaction labels across all species: head 7 of layer 3 achieves an AUC of 0.92 in *E. coli*, while also performing strongly in other species. PPI are most accurately captured by central layers; in protein language models, central layers are known to capture more complex interactions than early layers (Vig et al., 2021), supporting the notion that higher-order interactions are essential for understanding PPI. This learning of PPI arises directly from the masked prediction training, which promotes the learning of dependencies between proteins in a proteome.

3.2. Fast and Accurate Interactome Screening

Current large-scale interactome prediction workflows generally rely on a two-stage pipeline (Humphreys et al., 2021; Zhang et al., 2024): first, DCA-based methods (Weigt et al., 2009) identify promising candidate pairs; second, heavier structure-based methods like AlphaFold-Multimer (Evans et al., 2021) analyze those candidates. The first step is limited by the computational cost of MSA generation, and by the need to train one DCA model per candidate protein pair. The DCA pipeline of Zhang et al. (2024) required over 30 days on 50–100 GPUs to process the human proteome. In contrast, ProteomeLM inference takes under 10 minutes per proteome (Figure 3**A**) on a single RTX A6000 GPU;

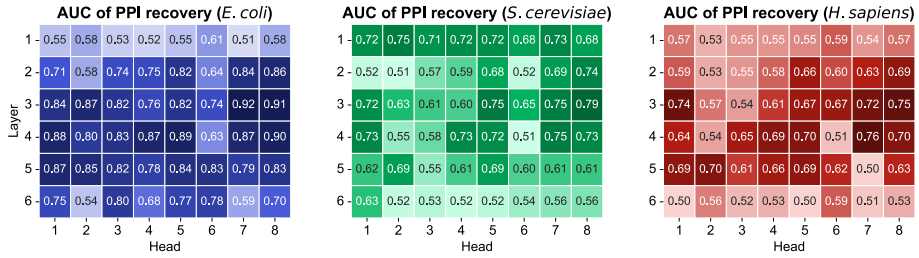

Figure 2. **Unsupervised detection of PPI using ProteomeLM attention coefficients.** AUC of each attention head in ProteomeLM-S for *E. coli*, *S. cerevisiae*, and *H. sapiens*

features are calculated for all possible protein pairs, without the need to train a separate model for each candidate pair, reducing compute by up to six orders of magnitude for inference alone.

We train a lightweight logistic regression on 48 ProteomeLM-S attention heads (Figure 3**B**), using only 100 positive and 1,000 negative pairs as supervision. In *H. sapiens*, ProteomeLM achieves an AUC of 0.83 versus 0.73 for DCA (Zhang et al., 2024); among the top 10M scored pairs, it recovers 50% of known PPI versus only 20% for DCA. Over 40% of the top 10,000 predictions align with known or suspected interactions in STRING (Szklarczyk et al., 2019), and nearly 10% correspond to high-confidence interactions. Across 19 human bacterial pathogens (102 million protein pairs), AUC values range from 0.87 to 0.92, confirming consistent generalization across highly diverse taxa. Thus, ProteomeLM has a very strong potential to reduce the burden on downstream structure-based modeling for precise PPI prediction.

3.3. Supervised PPI Prediction

We introduce ProteomeLM-PPI, a supervised PPI prediction network that employs both node-type features (ProteomeLM embeddings for individual proteins) and edge-type features (ProteomeLM attention coefficients for pairs of proteins). The supervised model relies on a modular neural network that processes individual protein embeddings and attention coefficients through distinct but integrated modules. To model the interaction between two proteins, the network combines their transformed representations by concatenating each of the two representations, their element-wise multiplication, and their absolute difference (see Figure 6A in supplementary). We trained and evaluated ProteomeLM-PPI on the D-SCRIPT dataset (Sledzieski et al., 2021) and the bias-controlled benchmark of Bernett et al. (2024), using the same splits as TUnA (Ko et al., 2024).

ProteomeLM-PPI outperforms state-of-the-art methods on *E. coli* and *S. cerevisiae*, and performs comparably on *D. melanogaster* and *C. elegans* (see Figure 6B in supplementary). In particular, it leads to an AUPR improvement of more than 0.1 (from 0.67 to 0.79) over TUnA (Ko et al., 2024) on *E. coli*, highlighting ProteomeLM's strong ability to capture PPI signals and to generalize from one species to others. On the dataset from Bernett et al. (2024), ProteomeLM-PPI consistently reaches or outperforms state-of-the-art methods (Figure 4), showing the robustness of ProteomeLM-PPI to biases of PPI prediction benchmarks. Ablation experiments confirm that ProteomeLM embeddings and attention coefficients are individually informative, and that their combination consistently yields the best predictive performance.

3.4. Gene Essentiality Prediction

While we focused on PPI prediction so far, ProteomeLM is a foundation model that can be used for diverse tasks where proteome-level information is important. Both protein sequence on the one hand, and genomic context and protein-protein interactions on the other hand, have been found to matter for predicting essentiality (Gurumayum et al., 2021). We introduce ProteomeLM-Ess, a two-layer fully connected classifier taking ProteomeLM embeddings as input, trained on the OGEE database (Gurumayum et al., 2021) (213,608 labeled genes across 91 species), with sequence-similarity-controlled splits (40% identity) to prevent homology-based shortcuts.

Classifiers based on ProteomeLM embeddings significantly outperform those based on ESM-C embeddings, demonstrating that the contextualized whole proteome-aware information present in ProteomeLM embeddings allows to better capture gene essentiality than protein-level information. The performance of ProteomeLM-Ess scales with ProteomeLM size; the best version (layer 8 embeddings from ProteomeLM-L) achieves an AUC of 0.93 (Figure 5). We held out the entire proteomes of *E. coli* and *S. cerevisiae* from training: 71% of experimentally essential genes (and only 2% of non-essential genes) in *E. coli* are correctly predicted as essential (Figure 7B). ProteomeLM-Ess further generalizes to the synthetic minimal cells JCVI-Syn1.0 and JCVI-Syn3A (absent from OGEE), achieving AUC of 0.88 and 0.83; JCVI-Syn3A, engineered to be close to minimal, constitutes a particularly stringent out-of-distribution test.

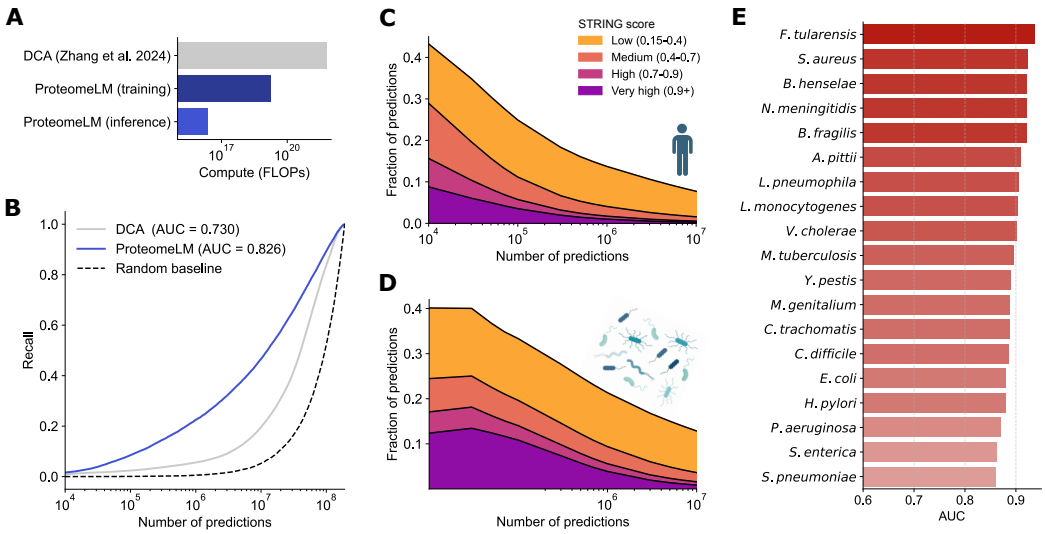

Figure 3. **Fast and high-precision screening of whole interactomes.** **(A)** Compute for ProteomeLM vs. DCA on the full human proteome. **(B)** *H. sapiens* interactome recall vs. number of predictions. **(C–D)** Fraction of top-scoring predictions matching STRING, for *H. sapiens* and 19 pathogens. **(E)** Per-species AUC across 19 pathogens.

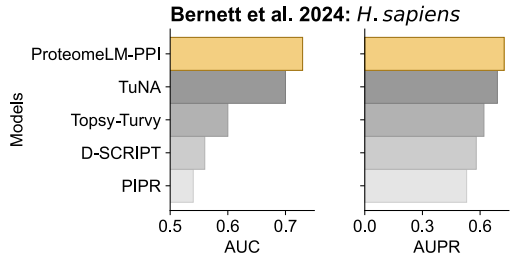

Figure 4. **Supervised prediction of PPI using ProteomeLM-PPI.** Performance on the bias-controlled benchmark of Bernett et al. (2024).

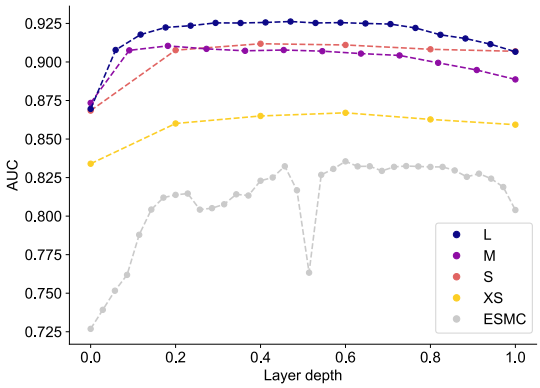

Figure 5. **Gene essentiality prediction with ProteomeLM-Ess.** AUC vs. normalized layer depth for ProteomeLM-Ess (all sizes) and ESM-C baseline.

4. Discussion

We introduced ProteomeLM, a transformer-based language model that learns contextualized protein representations from complete proteomes spanning the whole tree of life. Trained to reconstruct masked protein representations from the other proteins of a proteome, ProteomeLM learns dependencies between proteins that reflect functional and evolutionary constraints. As a foundation language model trained using the MLM objective, ProteomeLM can be used for many downstream tasks. The possibility of using the pretrained ProteomeLM, means that using it in inference for downstream tasks can be very computationally efficient. We demonstrated that ProteomeLM combines speed, interpretability, and accuracy across tasks ranging from PPI screening to gene essentiality prediction.

ProteomeLM can be used to map functional networks and co-expression clusters, predict protein complex membership across species, and study the evolution of these systems. For the determination of direct physical interactions, it can serve as a fast and accurate screening step to prioritize candidate protein pairs for computationally intensive analyses, including structural prediction approaches such as Boltz (Wohlwend et al., 2024; Passaro et al., 2025) and AlphaFold3 (Abramson et al., 2024). Beyond interaction prediction, ProteomeLM enables high-throughput *in silico* screens of gene essentiality and comparisons across species. ProteomeLM's performance remains stronger on prokaryotes than on eukaryotes, likely due to the relative scarcity of high-quality eukaryotic proteomes in the training dataset; expanding to more eukaryotic and metagenomic proteomes is a natural avenue for improvement.

In the future, the progress of long-context language models may enable ProteomeLM to directly operate at the amino-acid level across whole proteomes, paving the way to local-

ized cross-protein interactions and more granular modeling of functional dependencies. Another interesting perspective is to train ProteomeLM using embeddings from protein language models trained over additional modalities beyond sequences, such as structure (Hayes et al., 2025); exploiting the complementary information available in structure and sequences may help ProteomeLM infer more complex functional relationships. We expect that proteome-aware language models will become increasingly important in the coming years, enabling new ways of modeling system-level biological properties at the scale of proteomes and cells.

Acknowledgements

Anonymized.

References

J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O'Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Žídek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630 (8016):493–500, 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w.

J. Bernett, D. B. Blumenthal, and M. List. Cracking the black box of deep sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics*, 25(2), 2024.

G. Brixi, M. G. Durrant, J. Ku, M. Poli, G. Brockman, D. Chang, G. A. Gonzalez, S. H. King, D. B. Li, A. T. Merchant, M. Naghipourfar, E. Nguyen, C. Ricci-Tam, D. W. Romero, G. Sun, A. Taghibakshi, A. Vorontsov, B. Yang, M. Deng, L. Gorton, N. Nguyen, N. K. Wang, E. Adams, S. A. Baccus, S. Dillmann, S. Ermon, D. Guo, R. Ilango, K. Janik, A. X. Lu, R. Mehta, M. R. K. Mofrad, M. Y. Ng, J. Pannu, C. Ré, J. C. Schmok, J. S. John, J. Sullivan, K. Zhu, G. Zynda, D. Balsam, P. Collison, A. B. Costa, T. Hernandez-Boussard, E. Ho, M.-Y. Liu, T. McGrath, K. Powell, D. P. Burke, H. Goodarzi, P. D. Hsu, and B. L. Hie. Genome modeling and design across all domains of life with Evo 2. *bioRxiv*, page 2025.02.18.638918, 2025.

P. Bryant, G. Pozzati, and A. Elofsson. Improved prediction of protein-protein interactions using AlphaFold2. *Nature communications*, 13(1):1265, 2022.

Q. Cong, I. Anishchenko, S. Ovchinnikov, and D. Baker. Protein interaction networks revealed by proteome coevolution. *Science*, 365(6449):185–189, 2019.

G. Croce, T. Gueudré, M. V. Ruiz Cuevas, V. Keidel, M. Figliuzzi, H. Szurmant, and M. Weigt. A multi-scale coevolutionary approach to predict interactions between protein domains. *PLoS computational biology*, 15(10): e1006891, 2019.

H. Dalla-Torre, L. Gonzalez, J. Mendoza-Revilla, N. Lopez Carranza, A. H. Grzywaczewski, F. Oteri, C. Dallago, E. Trop, B. P. de Almeida, H. Sirelkhatim, G. Richard, M. Skwark, K. Beguir, M. Lopez, and T. Pierrot. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 22(2):287–297, 2025.

T. Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. *arXiv*, page 2307.08691, 2023.

A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, D. Bhowmik, and B. Rost. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3095381. URL https://ieeexplore.ieee.org/document/9477085.

ESM Team et al. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. *Evolutionary Scale Website*, 2024.

R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, and D. Hassabis. Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, page 2021.10.04.463034, 2021.

S. Gurumayum, P. Jiang, X. Hao, T. L. Campos, N. D. Young, P. K. Korhonen, R. B. Gasser, P. Bork, X.-M. Zhao, L.-j. He, and W.-H. Chen. OGEE v3: Online GEne Essentiality database with increased coverage of organisms and human cell lines. *Nucleic Acids Research*, 49(D1):D998–D1003, 2021.

T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert,

R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. A. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, and A. Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/10.1126/science.ads0018.

I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness, S. Banjade, S. R. Bagde, V. G. Stancheva, X.-H. Li, K. Liu, Z. Zheng, D. J. Barrero, U. Roy, J. Kuper, I. S. Fernández, B. Szakal, D. Branzei, J. Rizo, C. Kisker, E. C. Greene, S. Biggins, S. Keeney, E. A. Miller, J. C. Fromme, T. L. Hendrickson, Q. Cong, and D. Baker. Computed structures of core eukaryotic protein complexes. *Science*, 374: 1340, 2021.

Y. Hwang, A. L. Cornman, E. H. Kellogg, S. Ovchinnikov, and P. R. Girguis. Genomic language model predicts protein co-regulation and function. *Nature communications*, 15(1):2880, 2024.

Y. S. Ko, J. Parkinson, C. Liu, and W. Wang. TUnA: an uncertainty-aware transformer model for sequence-based protein–protein interaction prediction. *Briefings in Bioinformatics*, 25(5), 2024.

D. Kuznetsov, F. Tegenfeldt, M. Manni, M. Seppey, M. Berkeley, E. Kriventseva, and E. M. Zdobnov. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.*, 51 (D1):D445–D451, 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac998. URL https://doi.org/10.1093/nar/gkac998.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, A. Dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, and A. Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023.

E. Nguyen, M. Poli, M. Faizi, A. Thomas, M. Wornow, C. Birch-Sykes, S. Massaroli, A. Patel, C. Rabideau, Y. Bengio, S. Ermon, S. A. Baccus, and C. Ré. HyenaDNA: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

S. Passaro, G. Corso, J. Wohlwend, M. Reveiz, S. Thaler, V. R. Somnath, N. Getz, T. Portnoi, J. Roy, H. Stark, D. Kwabi-Addo, D. Beaini, T. Jaakkola, and R. Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, page 2025.06.14.659707, 2025.

M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 96(8):4285–4288, 1999.

R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives. MSA Transformer. *Proceedings of the 38th International Conference on Machine Learning*, 139:8844–8856, 2021.

A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA*, 118(15): e2016239118, 2021. ISSN 0027-8424. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/content/118/15/e2016239118.

R. Singh, K. Devkota, S. Sledzieski, B. Berger, and L. Cowen. Topsy-Turvy: integrating a global view into sequence-based PPI prediction. *Bioinformatics*, 38:i264–i272, 2022. Publisher: Oxford Academic.

S. Sledzieski, R. Singh, L. Cowen, and B. Berger. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems*, 12(10):969–982, 2021.

D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N. T. Doncheva, J. H. Morris, P. Bork, L. J. Jensen, and C. V. Mering. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, 47(D1): D607–D613, 2019.

J. Vig, A. Madani, L. R. Varshney, C. Xiong, R. Socher, and N. Rajani. BERTology meets biology: Interpreting attention in protein language models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YWtLZvLmud7.

M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl. Acad. Sci. USA*, 106(1):67–72, 2009.

J. Wohlwend, G. Corso, S. Passaro, N. Getz, M. Reveiz, K. Leidal, W. Swiderski, L. Atkinson, T. Portnoi, I. Chinn, J. Silterra, T. Jaakkola, and R. Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, page 2024.11.19.624167, 2024.

J. Zhang, I. R. Humphreys, J. Pei, J. Kim, C. Choi, R. Yuan, J. Durham, S. Liu, H.-J. Choi, M. Baek, D. Baker, and Q. Cong. Computing the human interactome. *bioRxiv*, page 2024.10.01.615885, 2024.

A. Supplementary Figures

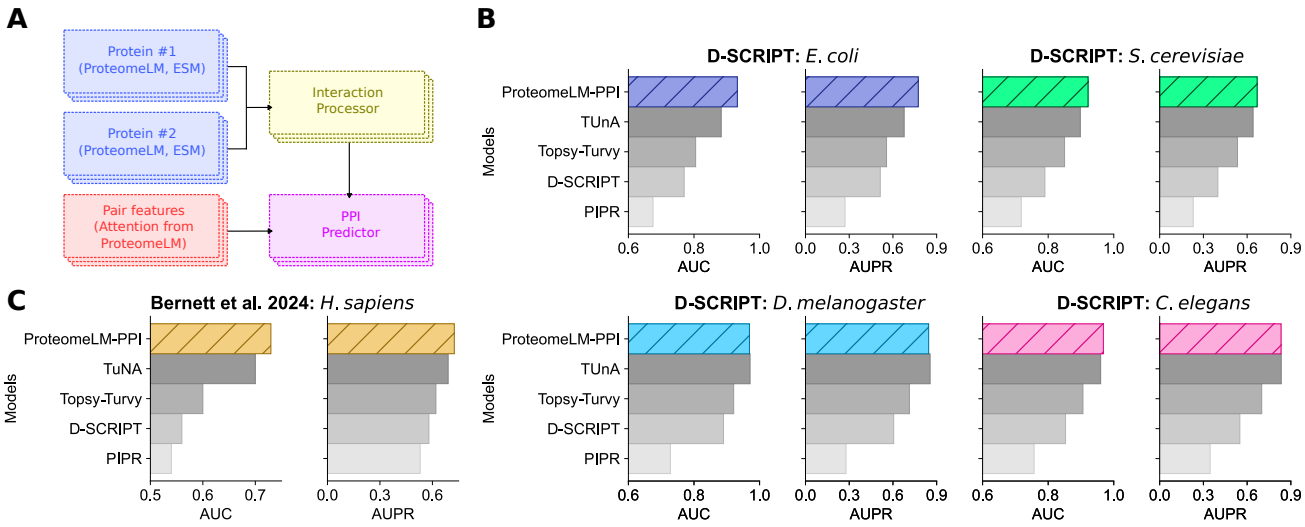

Figure 6. **Supervised prediction of PPI using ProteomeLM-PPI. (A)** Architecture: node features (ProteomeLM and ESM-C embeddings) and edge features (attention coefficients) feed a modular network. **(B)** Cross-species generalization on D-SCRIPT (Sledzieski et al., 2021). **(C)** Performance on the bias-controlled benchmark of Bernett et al. (2024).

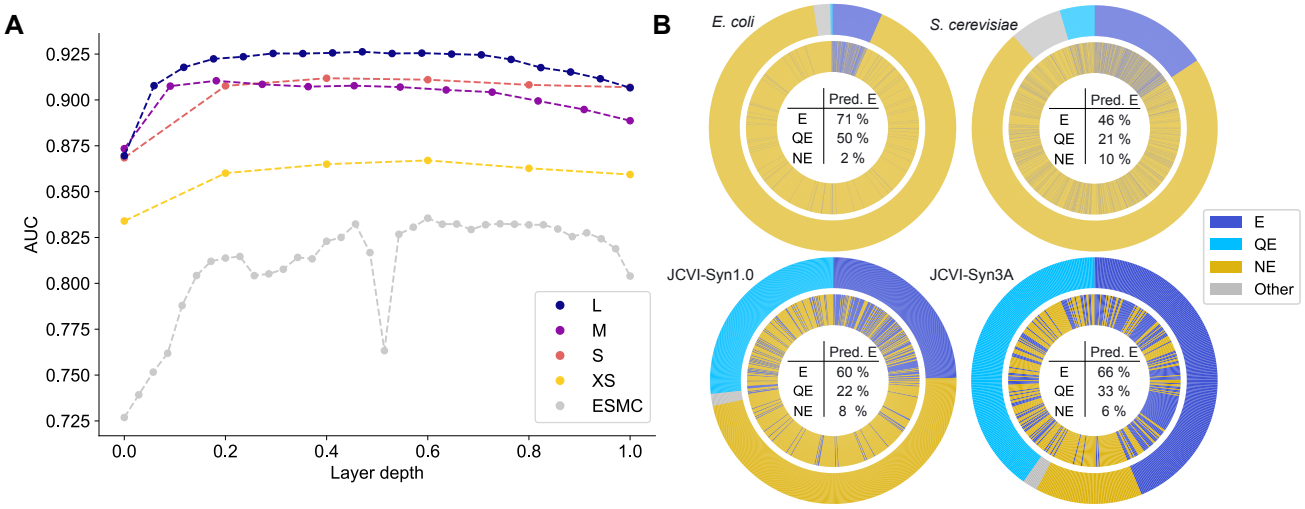

Figure 7. **Gene essentiality prediction with ProteomeLM-Ess. (A)** AUC vs. normalized layer depth for ProteomeLM-Ess (all sizes) and ESM-C baseline. **(B)** Predicted vs. experimental essentiality labels for *E. coli*, *S. cerevisiae*, JCVI-Syn1.0, and JCVI-Syn3A (held out from training).