A PROTEOME-SCALE MASKED LANGUAGE MODEL FOR FAST PROTEIN-PROTEIN INTERACTION PREDICTION

Cyril Malbranke
School of Life Sciences
Institute of Bioengineering
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
cyril.malbranke@epfl.ch

Anne-Florence Bitbol
School of Life Sciences
Institute of Bioengineering
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
anne-florence.bitbol@epfl.ch

ABSTRACT

Protein-protein interactions (PPIs) play a central role in most biological processes. The large number of potential pairs of interacting proteins makes the determination of PPI networks challenging. High-throughput experimental methods to determine them remain prohibitive beyond some model species. Hence, computational methods are needed to screen these interactions. While some methods have relied on scoring each pair individually, we introduce ProteomeLM, a proteome-scale language model, which can rank all pairs in a single pass. Early results suggest that at least 70% of PPIs in *Escherichia coli* could be identified in the top 5% best ranked interactions. This method should be useful as a pre-screening tool, allowing to identify the most promising pairs for docking-based or experimental determination of PPI.

1 INTRODUCTION

Predicting protein-protein interactions (PPIs) is an important open question in biology. Experiments to determine these interactions are challenging, especially at a large scale. The existing data has allowed building experimentally verified datasets and databases such as BioGRID (Oughtred et al. (2021)), EcoCyc (Karp et al. (2018)) and IntAct (del Toro et al. (2022)). Building on this data, machine learning methods have been implemented to tackle this question *in silico*. Some of the most accurate methods, such as docking or multimer folding, rely extensively on protein complex structure determination (Bryant et al. (2022)). Other methods rely on coevolutionary signal. They are based on the idea that interacting proteins exert evolutionary constraints on one another. Some of these methods are based on the co-presence or absence of certain proteins across the proteome (Croce et al. (2019); Moi et al. (2020); Green et al. (2021)), while others score the residues that co-evolve across proteins through Potts models, also known as Direct Coupling Analysis methods (Cong et al. (2019)). Recently, deep learning was also extensively used to capture these coevolutionary signals through graph neural networks or protein language models Sledzieski et al. (2021); Hwang et al. (2023). For reviews of these methods, see Soleymani et al. (2022); Bernett et al. (2023).

BERT-based language models (LMs) are a family of deep learning models based on the attention mechanism that are trained with the masked language modeling (MLM) objective of reconstructing randomly masked amino acids in a protein sequence. This task allows the network to learn relationships between positions (amino-acid sites) in protein sequences. It favors the emergence of representations that are aware of some structural and functional properties of the protein. The network can then be fine-tuned for downstream tasks, including prediction of protein structure (Lin et al. (2023)), localization (Thumuluri et al. (2022)), activity or mutational effects (Rao et al. (2020)). It is tempting to also use these successful methods for PPI prediction (Lupo et al. (2024)). The last years have seen the release of BERT-based protein language models such as ProteinBERT (Brandes et al. (2022)) and Evolutionary Scale Modelling (ESM-2, Lin et al. (2023)), which will be extensively used here.

In this paper, we introduce ProteomeLM, a novel proteome-scale LM that builds upon ESM representations of the different proteins contained in a given proteome. This method aims at reconstruct-

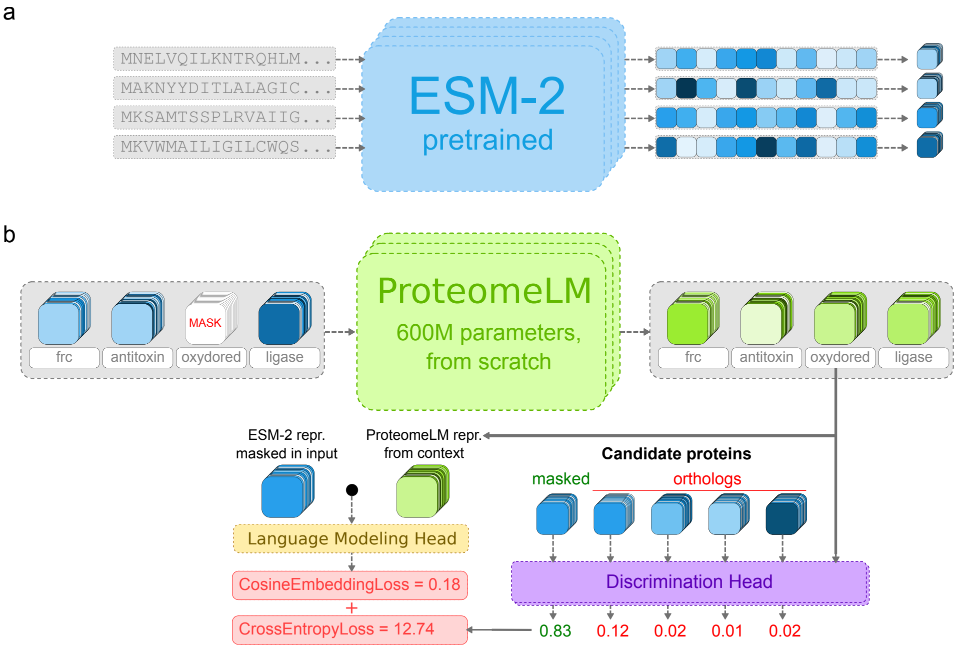

Figure 1: **Schematic of the ProteomeLM training pipeline. (a)** Input amino-acid sequences undergo feature extraction through the (pretrained) ESM-2 model (see section 2.1), yielding a fixed-dimensional embedding for each protein. **(b)** The extracted features serve as input to ProteomeLM, trained from scratch to predict the masked embeddings of proteins in the context of their proteome (see section 2.3). Proteins are annotated by their orthologous group (see section 2.2). ProteomeLM's training builds upon two objectives, (i) minimizing the difference between the input ESM-2 embedding and ProteomeLM's reconstructed embedding for each masked protein and (ii) identifying the right embedding among a list of embeddings corresponding to orthologs of the protein of interest.

ing the representation of masked proteins from the representation of the other proteins from the same proteome. This training process favors the emergence of representation and attention heads that are aware of the interactions, or at least interdependences, between different proteins. It also aims at contextualizing the representations of proteins.

First, we will present methods for data collection and preparation, and the architecture of the model and its training. Next, we will show our early results on the inference of PPIs in *Escherichia coli*. The model presented here focuses on bacteria, but we plan to extend it in the future.

2 Methods

2.1 Protein representation

The first challenge is to find a good fixed-dimensional protein representation, which is necessary to use them as inputs of a transformer. The raw representations of protein or gene sequences as one hot encoded sequences of amino acids or nucleotides have several weaknesses. They are high-dimensional and discrete, and their dimensionality depends on sequence length. Furthermore, the biological and physicochemical properties of protein sequences are encoded in a hardly tractable manner. For our purpose, a better representation must be found.

Protein language models trained with the MLM objective favor the emergence of rich representations that encompass various aspects of protein properties (Vig et al. (2021)). As these models are trained to predict the hidden amino acids in a protein sequence, the representations that emerge are aware of the statistical patterns across the protein, which often correlate with critical biological, physical or chemical properties. These emerging representations can, for example, be used as an efficient basis for protein folding (see ESM Fold by Lin et al. (2023)). We thus encoded every protein with

ESM-2 150M. Representations were reduced by averaging per-token representations over all tokens in a protein. Hence, each protein is represented by a 640-dimensional vector (Figure 1**a**).

2.2 Dataset preparation and encoding

For now, we focus on bacterial proteomes. By proteome, we mean the set of protein-coding genes in a genome. We collected 17,551 complete annotated proteomes from OrthoDB (Zdobnov et al. (2021)). Each proteome is a list of protein sequences annotated by the orthologous group they belong to. We restricted to proteins in orthologous groups that are present in more than 1% of the genomes considered (i.e., in at least 170 genomes). An orthologous group contains descendants of an ancestral gene and is linked to functional annotations from Gene Ontology (Ashburner et al. (2000); The Gene Ontology Consortium et al. (2023)) that describe the localization and biological processes the protein is involved in. The definition of an orthologous group operates at a certain level of orthology. Here, we use orthologous groups defined at the Bacteria kingdom level.

Most natural language processing BERT models rely on positional encoding to encode the position of a token in the sequence. Here, we instead use a functional encoding based on the gene presence-absence matrix and the Gene Ontology (GO) terms associated to OrthoDB orthologous groups. Specifically, we perform dimensional reduction through singular vector decomposition (SVD) on the concatenation of these two matrices to get a 100-dimensional vector describing the functional properties of each protein. After encoding each protein through ESM-2, yielding a 640-dimensional vector, we also computed the mean and standard deviations of the ESM-2 representations for each orthologous group, which will serve as additional functional encoding for each orthologous group. We then end up with a functional description of each orthologous group based on 3 different vectors: (i) the SVD of the Gene Ontology terms and presence-absence matrices, (ii) the mean and (iii) the standard deviation of the ESM-2 representations.

2.3 Model architecture and training

We trained from scratch a language model so that it learns complex relationships between the ESM-2 representations of different proteins in a proteome. The core of the model is the pre-layer norm RoBERTa model (Liu et al. (2019); Ott et al. (2019)), as available in Hugging Face's transformer library (Wolf et al. (2020)). This model has 18 layers with 20 attention heads and an embedding dimension of 1280. In input, the functional encoding goes through an embedding module. The ESM-2 representations of each protein are normalized by the mean and standard deviation of their orthologous group, and also go through an embedding module.

Training employs the MLM objective (see Figure 1**b**). At each step, we randomly mask 15% of the protein representations in a proteome (the functional encoding remains unmasked). Two downstream heads are used in training. The *language modeling head* aims at fostering good reconstruction of the hidden normalized representation. It is trained through the minimization of the cosine embedding loss between the predicted embedding and the masked embedding. The *discrimination head* takes as input the last layer representation of masked proteins and the ESM-2 representations of candidate proteins. It scores how likely each of these sequences is to be the masked protein sequence. This head is trained by minimizing the cross-entropy loss when given the correct masked protein sequence and 9 other candidate sequences sampled randomly in the orthologous group of the protein of interest. The representations were normalized and the discrimination head was introduced to prevent the network from just producing a protein representation corresponding to the mean representation of the orthologous group it belongs to.

Training was done for approximately 7 days on 8 NVIDIA A100 GPUs with 80 GB of RAM each. We used Flash Attention (Dao et al. (2022)) to speed up training. A validation set of 10% of the proteomes was extracted at random to track the progress of the model. After 7 days, the average cosine similarity between predicted and masked embedding was 0.61 while the discrimination head given 10 candidate sequences was able to predict the right one in 82% of the case.

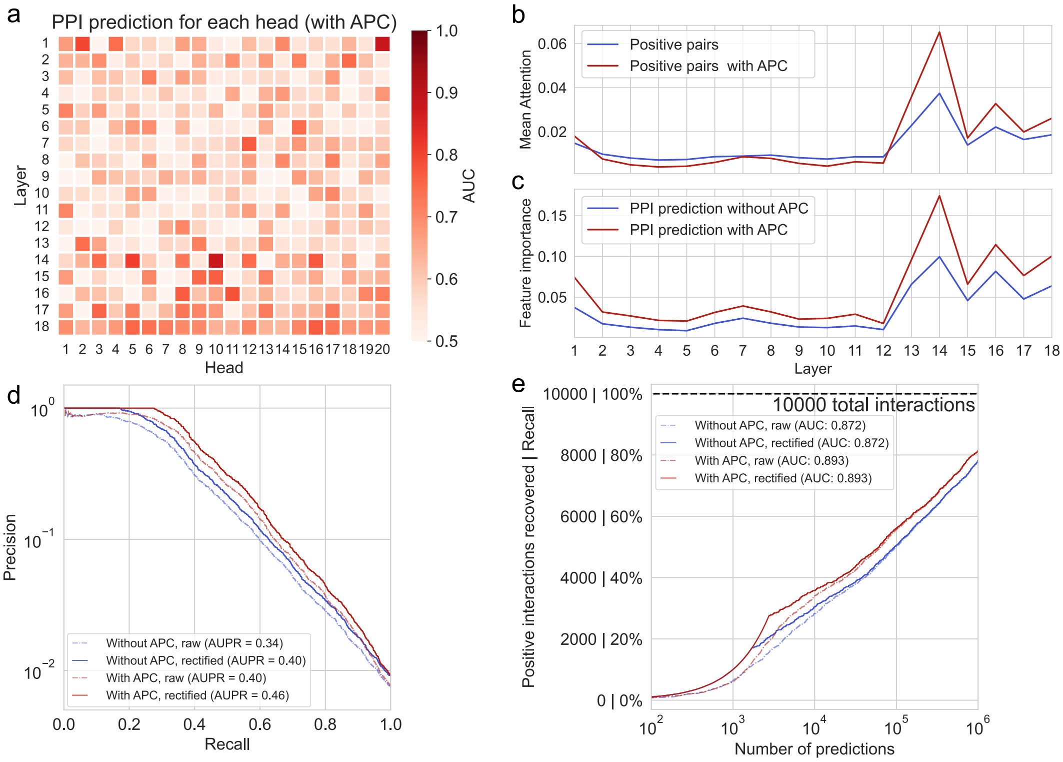

Figure 2: **Exploration of PPI prediction capabilities within the *E. coli* proteome using ProteomeLM's attention heads. (a)** AUC for PPI prediction for each attention head in each layer, after applying the APC. **(b)** Average attention within each layer for positive and random protein pairs, with and without APC, showing stronger attention for positive pairs. Late layers are particularly activated by interacting pairs. **(c)** Feature importance for PPI prediction with and without APC, across different layers. **(d)** Precision-recall curves illustrating the prediction quality for PPIs with APC and without APC (raw and rectified, see Appendix B). **(e)** The relationship between the number of predicted interactions and the recovered actual interactions, under the hypothesis of 10,000 total interactions, both with and without APC, for raw and rectified measurements.

3 Early results

3.1 PPI predictions on the *E. coli* proteome

We first explored the ability of our model to predict PPIs between proteins in the proteome of the model bacterial species *E. coli*. We took as a reference the dataset from Cong et al. (2019), which gathers verified PPIs from different sources as positive pairs and samples at random more than 200,000 pairs of proteins, which, given the sparsity of the interaction graph, are overwhelmingly negative. In this dataset, the proportion of positive pairs aims at reflecting reality. Among the proteins that we were able to process with our current model, we retrieved 1600 positive pairs of proteins and 200,000 negative pairs, involving 3200 proteins. We annotated and encoded these 3200 protein sequences using respectively OrthoDB and ESM-2. We then passed this processed *E. coli* proteome into our trained ProteomeLM model, and we collected the attention matrices for each attention head of each layer. Each of these matrices gives an attention score for each pair of protein. We performed the Average Product Correction (APC, see Appendix A) (Dunn et al. (2008); Ekeberg et al. (2013)) on each of these matrices to better highlight direct interactions.

We investigated each attention head's ability at predicting PPIs. Specifically, for each head of each layer, we evaluated the area under the receiver operating characteristic curve (AUC) using the strength of the attention as a predictor of PPIs. Figure 2**a** shows that some heads are excellent predictors, reaching AUCs up to 0.9. Figure 2**b** displays the mean value of the attention heads in each layer for positive PPI pairs and for random pairs. We observe that attention is much stronger between interacting proteins than in random pairs. In addition, attention heads that are most sensitive to interactions are in the first layer and in late layers of the network.

In light of these observations, we extracted a set of 100 positive pairs and 10,000 random pairs to train a logistic regression based on either (i) the attention heads, or (ii) the attention heads corrected by APC. In Figure 2**c**, we used the coefficients learned by the regression to evaluate the importance of each layer. For this, we summed the coefficients over all attention heads of each layer. The results confirm that the most important layers are the first one and the late ones. We then tested the ability of our logistic regression to identify the unseen PPIs. We report the precision-recall curve in Figure 2**d**. Based on previous work (Launay et al. (2017)), we took 10,000 as an estimate of the total number of physical PPIs in *E. coli*. In Figure 2**e**, we plot the number of detected positive pairs versus the number of predicted pairs, both with and without APC. We also provide a rectified measurement based on the likelihood of a random pair to be actually positive (Appendix B).

These estimates show that up to 70% of the pairs could be in the top 200,000 predicted pairs. This result, combined with the light computational cost of processing one single proteome using the trained model, demonstrates the potential of our method for high-throughput PPI discovery. In particular, ProteomeLM could be used for pre-screening, identifying the most promising protein pairs to be studied by more precise but heavier computational methods such as protein docking (see Cong et al. (2019) for a similar approach with Direct Coupling Analysis methods) and then for experimental validation.

3.2 Effect of functional annotations on model precision

We also wanted to assess the importance of using Gene Ontology terms to annotate the data in our functional encoding. For this, we binned each pair according to the sum of the number of GO annotations of the two proteins involved, as richer functional context can make PPIs easier to infer. We also grouped each pair according to the number of shared GO annotations between the two members of the pair (recall that GO annotations comprise information such as localization or involvement in biological processes). The tables below display the ability of the model to detect positive pairs for various values of the sum of the number of GO annotations (on the left) and for various numbers of shared annotations in each pair (on the right):

Sum of GO terms	**Positive pairs**	**AUC**
$[1, 5]$	55	0.897
$[6, 10]$	380	0.882
$[11, 15]$	528	0.890
$[16, 20]$	428	0.905
> 20	240	0.917

Shared GO terms	**Positive pairs**	**AUC**
0	852	0.880
1	270	0.912
2	219	0.854
3 or 4	77	0.857
More than 5	215	0.998

Overall, these results suggest that ProteomeLM performs slightly better on pairs with rich functional annotations (the AUC is 0.917 with more than 21 GO terms versus 0.882 to 0.897 with 15 or fewer GO terms). Furthermore, the pairs sharing many GO terms are accurately identified as PPIs or non-PPIs. In particular, the method can single out almost all interactions when the proteins in the pair share more than 5 GO terms (AUC 0.998). This result is impressive given that among the 1,000 pairs we detected that shared 5 GO terms, only 215 of them were experimentally confirmed, which means that the network learned to distinguish a functional context linked to an interaction from one that is not.

4 Discussion

Our early results indicate that ProteomeLM is a promising framework to identify PPIs and may aid the discovery of novel interacting pairs, as a very fast pre-screening step before docking and experimental validation of PPIs. Indeed, our logistic regression based on the attention values of the network recovers a significant fraction of physically interacting pairs among its top predictions. In the future, ablation studies removing the GO annotations from the input should help better understand the role of functional annotations in PPI identification. Further work will focus on extending the model coverage to eukaryotes and to the full proteome of prokaryotes.

The final trained model will open many other applications. In particular, we will refine the PPI module to be able to identify different types of interactions beyond physical PPIs, such as co-regulation, or co-localization. The model will also be tested as a preliminary step towards prediction and discovery of biological networks and processes. Finally, other downstream tasks, such as prediction of functional properties such as enzyme activity, contextualized by the proteome, will be tested.

REFERENCES

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. ISSN 1546-1718. doi: 10.1038/75556. URL `https://www.nature.com/articles/ng0500_25`. Number: 1 Publisher: Nature Publishing Group.

Judith Bernett, David B. Blumenthal, and Markus List. Cracking the black box of deep sequence-based protein-protein interaction prediction, January 2023. URL `https://www.biorxiv.org/content/10.1101/2023.01.18.524543v1`. Pages: 2023.01.18.524543 Section: New Results.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, April 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac020. URL `https://doi.org/10.1093/bioinformatics/btac020`.

Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, 13(1):1265, March 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28865-w. URL `https://www.nature.com/articles/s41467-022-28865-w`. Number: 1 Publisher: Nature Publishing Group.

Qian Cong, Ivan Anishchenko, Sergey Ovchinnikov, and David Baker. Protein interaction networks revealed by proteome coevolution. *Science*, 365(6449):185–189, July 2019. doi: 10.1126/science.aaw6718. URL `https://www.science.org/doi/full/10.1126/science.aaw6718`. Publisher: American Association for the Advancement of Science.

Giancarlo Croce, Thomas Gueudré, Maria Virginia Ruiz Cuevas, Victoria Keidel, Matteo Figliuzzi, Hendrik Szurmant, and Martin Weigt. A multi-scale coevolutionary approach to predict interactions between protein domains. *PLOS Computational Biology*, 15(10):e1006891, October 2019. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006891. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006891`. Publisher: Public Library of Science.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness, June 2022. URL `http://arxiv.org/abs/2205.14135`. arXiv:2205.14135 [cs].

Noemi del Toro, Anjali Shrivastava, Eliot Ragueneau, Birgit Meldal, Colin Combe, Elisabet Barrera, Livia Perfetto, Karyn How, Prashansa Ratan, Gautam Shirodkar, Odilia Lu, Bálint Mészáros, Xavier Watkins, Sangya Pundir, Luana Licata, Marta Iannuccelli, Matteo Pellegrini, Maria Jesus Martin, Simona Panni, Margaret Duesbury, Sylvain D Vallet, Juri Rappsilber, Sylvie Ricard-Blum, Gianni Cesareni, Lukasz Salwinski, Sandra Orchard, Pablo Porras, Kalpana Panneerselvam, and Henning Hermjakob. The IntAct database: efficient access to fine-grained molecular interaction data. *Nucleic Acids Research*, 50(D1):D648–D653, January 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1006. URL `https://doi.org/10.1093/nar/gkab1006`.

S.D. Dunn, L.M. Wahl, and G.B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, February 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btm604. URL `https://doi.org/10.1093/bioinformatics/btm604`.

Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1): 012707, January 2013. ISSN 1539-3755, 1550-2376. doi: 10.1103/PhysRevE.87.012707. URL `http://arxiv.org/abs/1211.1281`. arXiv:1211.1281 [cond-mat, physics:physics, q-bio].

Anna G. Green, Hadeer Elhabashy, Kelly P. Brock, Rohan Maddamsetti, Oliver Kohlbacher, and Debora S. Marks. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nature Communications*, 12(1):1396, March 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-21636-z. URL `https://www.nature.com/articles/s41467-021-21636-z`. Number: 1 Publisher: Nature Publishing Group.

Yunha Hwang, Andre L. Cornman, Elizabeth H. Kellogg, Sergey Ovchinnikov, and Peter R. Girguis. Genomic language model predicts protein co-regulation and function, October 2023. URL `https://www.biorxiv.org/content/10.1101/2023.04.07.536042v3`. Pages: 2023.04.07.536042 Section: New Results.

Peter D. Karp, Wai Kit Ong, Suzanne Paley, Richard Billington, Ron Caspi, Carol Fulcher, Anamika Kothari, Markus Krummenacker, Mario Latendresse, Peter E. Midford, Pallavi Subhraveti, Socorro Gama-Castro, Luis Muñiz-Rascado, César Bonavides-Martinez, Alberto Santos-Zavaleta, Amanda Mackie, Julio Collado-Vides, Ingrid M. Keseler, and Ian Paulsen. The EcoCyc Database. *EcoSal Plus*, 8(1):10.1128/ecosalplus.ESP–0006–2018, November 2018. doi: 10.1128/ecosalplus.esp-0006-2018. URL `https://journals.asm.org/doi/full/10.1128/ecosalplus.esp-0006-2018`. Publisher: American Society for Microbiology.

Guillaume Launay, Nicoletta Ceres, and Juliette Martin. Non-interacting proteins may resemble interacting proteins: prevalence and implications. *Scientific Reports*, 7(1):40419, January 2017. ISSN 2045-2322. doi: 10.1038/srep40419. URL `https://www.nature.com/articles/srep40419`. Number: 1 Publisher: Nature Publishing Group.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. doi: 10.1126/science.ade2574. URL `https://www.science.org/doi/10.1126/science.ade2574`. Publisher: American Association for the Advancement of Science.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL `http://arxiv.org/abs/1907.11692`. arXiv:1907.11692 [cs].

Umberto Lupo, Damiano Sgarbossa, and Anne-Florence Bitbol. Pairing interacting protein sequences using masked language modeling, January 2024. URL `https://www.biorxiv.org/content/10.1101/2023.08.14.553209v3`. Pages: 2023.08.14.553209 Section: New Results.

David Moi, Laurent Kilchoer, Pablo S Aguilar, and Christophe Dessimoz. Scalable phylogenetic profiling using minhash uncovers likely eukaryotic sexual reproduction genes. *PLoS computational biology*, 16(7):e1007553, 2020.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling, April 2019. URL `http://arxiv.org/abs/1904.01038`. arXiv:1904.01038 [cs].

Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, Sonam Dolma, Jasmin Coulombe-Huntington, Andrew Chatr-aryamontri, Kara Dolinski, and Mike Tyers. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200, 2021. ISSN 1469-896X. doi: 10.1002/pro.3978. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3978`. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.3978.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners, December 2020. URL https://www.biorxiv.org/content/10.1101/2020.12.15.422761v1. Pages: 2020.12.15.422761 Section: New Results.

Samuel Sledzieski, Rohit Singh, Lenore Cowen, and Bonnie 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.e6, October 2021. ISSN 2405-4712. doi: 10.1016/j.cels.2021.08.010. URL https://www.cell.com/cell-systems/abstract/S2405-4712(21)00333-1. Publisher: Elsevier.

Farzan Soleymani, Eric Paquet, Herna Viktor, Wojtek Michalowski, and Davide Spinello. Protein–protein interaction prediction with deep learning: A comprehensive review. *Computational and Structural Biotechnology Journal*, 20:5316–5341, January 2022. ISSN 2001-0370. doi: 10.1016/j.csbj.2022.08.070. URL https://www.sciencedirect.com/science/article/pii/S2001037022004044.

The Gene Ontology Consortium, Suzi A Aleksander, James Balhoff, Seth Carbon, J Michael Cherry, Harold J Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L Harris, David P Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W Sternberg, Paul D Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A Siegele, Rex L Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N Erdol, Matthew C Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C C Saverimuttu, Renzhi Su, Kate E Thurlow, Ruth C Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E Dolan, Harold J Drabkin, David P Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L De Pons, Melinda R Dwinell, G Thomas Hayman, Mary L Kaldunski, Anne E Kwitek, Stanley J F Laulederkind, Marek A Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A Aleksander, J Michael Cherry, Stacia R Engel, Kalpana Karra, Stuart R Miyasato, Robert S Nash, Marek S Skrzypek, Shuai Weng, Edith D Wong, Erika Bakker, Tanya Z Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The Gene Ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, May 2023. ISSN 1943-2631. doi: 10.1093/genetics/iyad031. URL https://doi.org/10.1093/genetics/iyad031.

Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, 50(W1):W228–W234, July 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac278. URL https://doi.org/10.1093/nar/gkac278.

Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. BERTology Meets Biology: Interpreting Attention in Protein Language Models, March 2021. URL http://arxiv.org/abs/2006.15222. arXiv:2006.15222 [cs, q-bio].

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-Art Natural Language Processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Evgeny M Zdobnov, Dmitry Kuznetsov, Fredrik Tegenfeldt, Mosè Manni, Matthew Berkeley, and Evgenia V Kriventseva. OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 49(D1):D389–D393, January 2021. ISSN 0305-1048. doi: 10.1093/nar/gkaa1009. URL https://doi.org/10.1093/nar/gkaa1009.

A AVERAGE PRODUCT CORRECTION

The Average Product Correction (APC) is commonly used to improve the prediction of residue-residue contacts from multiple sequence alignments and also r prediction of protein-protein interactions. Here, the APC is applied to the raw scores obtained from attention maps to reduce the influence of some proteins that are strongly activating with numerous proteins due to their central role in some biological process(es) more than due to direct protein-protein interactions. The attention head score between protein i and j, denoted as S_{ij}^{APC}, is computed from the raw score, S_{ij}, using the following formula:

$$S_{ij}^{\text{APC}} = S_{ij} - \frac{S_{i\cdot}S_{\cdot j}}{S_{\cdot\cdot}}$$

where $S_{i\cdot}$ and $S_{\cdot j}$ are the average scores of residue i with all other residues and residue j with all other residues, respectively, and $S_{\cdot\cdot}$ is the average score over all pairs of residues. The averages are computed as:

$$S_{i\cdot} = \frac{1}{L-1} \sum_{k \neq i} S_{ik}$$

$$S_{\cdot j} = \frac{1}{L-1} \sum_{k \neq j} S_{kj}$$

$$S_{\cdot\cdot} = \frac{1}{L(L-1)} \sum_{k \neq l} S_{kl}$$

where L is the number of proteins in the proteome.

B RECTIFIED PREDICTION SCORE

Let x be a pair of proteins, $y \in \{0, 1\}$ the variable that defines if this pair is in interaction, and z the variable that defines if this pairs is labelled as an interaction or not. Let $f(x)$ be the variable associated to the classifier f that defines whether this pair of proteins is classified as interacting.

Let N be the total number of pairs, P the total number of interactions. We build a training set that contains Q labelled interactions and R randomly drawn pairs among the remaining pairs (these pairs can thus be positive or negative). Q and R are supposed to be small compared to N. We supposed that the labelled interactions are randomly drawn among the interactions, which means that $p(z|x, y) = p(z|y) = p(z|f(x), y)$.

In this set up, the number of interactions labelled negatively is $R\frac{P-Q}{N}$. And in total, we have in the set $Q + R\frac{P-Q}{N}$ interactions which rewrite as $(1 + r)Q$ with $r = \frac{R}{N}\frac{P-Q}{Q}$. We then have $p(z = 1|y = 1) = \frac{Q}{(1+r)Q} = \frac{1}{1+r}$.

We can then express the precision (for the prediction of $Z = 1$) by:

$$p(z = 1|f(x) = 1) = p(z = 1|y = 1)p(y = 1|f(x) = 1) = \frac{1}{1 + r} p(y = 1|f(x) = 1)$$

This means that we can correct the precision for the prediction of $Y = 1$ by:

$$\text{precision}(f, y) = (1 + r)\text{precision}(f, Z)\,.$$

We can also express the recall such that:

$$\begin{aligned}
\text{recall}(f, y) &= p(f(x) = 1|y = 1) \\
&= p(y = 1|f(x) = 1)\frac{p(f(x) = 1)}{p(y = 1)} \\
&= (1 + r)p(z = 1|f(x) = 1)\frac{p(f(x) = 1)}{p(z = 1)}\frac{p(z = 1)}{p(y = 1)} \\
&= (1 + r)\text{recall}(f, z)\frac{Q}{N}\frac{N}{(1 + r)Q} \\
&= \text{recall}(f, z)
\end{aligned}$$