MSA Pairing Transformer: protein interaction partner prediction with few-shot contrastive learning

Alex Hawkins-Hooker [1] Daniel Burkhardt Cerigo [2] Umberto Lupo [3 4] David T. Jones [5 6] Brooks Paige [1]

Abstract

We study the problem of pairing interacting protein sequences across protein families that are known to interact, in cases where multiple paralogs of each interaction partner can occur in a given species. We fine-tune the MSA Transformer to predict interaction partners by applying contrastive learning to embeddings of pairs of interacting domain sequences in scrambled single-chain multiple sequence alignments (MSAs). We demonstrate the effectiveness of our model across a set of bacterial interactions for which ground-truth pairings are known, finding that it is possible to achieve high pairing accuracy even within small sets of pairable sequences, unlike previous methods based on co-evolutionary statistics. Across a large dataset of prokaryotic interactions with experimentally determined complexes, paired MSAs generated by our model contain co-evolutionary signal that more strongly encodes interface contacts than MSAs paired by widely-used heuristic methods, suggesting the potential of our approach to improve the co-evolutionary analysis of protein-protein interactions.

1. Introduction

The maintenance of interaction specificity within protein-protein interactions conserved across species constrains sequence variation at sets of interacting residues. In principle, analysis of this evolutionary co-variation promises to help resolve interaction specificity, improving the ability to reconstruct protein interaction networks and transfer understanding of protein interactions across species. A particular challenge in the latter case is associated with the potential presence of multiple paralogs of each interaction partner in a given species, leading to ambiguity in which pairs of sequences are involved in a specific interaction. A variant of this problem arises notably in the context of structure prediction of multi-chain complexes (Ovchinnikov et al., 2014; Hopf et al., 2014; Bryant et al., 2022; Evans et al., 2022), where, in order to maximise the available co-evolutionary signal, it is desirable to produce 'paired' cross-chain MSAs in which each row consists of concatenated sequences which themselves interact. Previous approaches to this 'MSA pairing' problem have shown that simple statistical models of the co-evolutionary variation between interaction partners can successfully identify interacting pairs of sequences in the presence of multiple in-species paralogs, and have proposed iterative algorithms allowing the learning of these models to be bootstrapped from very small sets of known pairs (Gueudré et al., 2016; Bitbol et al., 2016; Bitbol, 2018; Lupo et al., 2024b). Recently, it has also been demonstrated that protein language models trained on MSAs or on entire genomes learn the hallmarks of interaction specificity, avoiding the need to build new statistical models for each interaction of interest, and allowing for accurate pairing within smaller families (Lupo et al., 2024a; Hwang et al., 2024; Malbranke & Bitbol, 2024). These approaches, however, rely on the pre-training tasks used to train the protein language model being aligned with the interaction partner prediction problem, so that meaningful signal for solving the latter can be extracted after pre-training.

In this paper, we propose to instead directly fine-tune protein language models to solve the MSA pairing problem. A central challenge is the absence of high-quality labelled datasets characterising protein interaction specificity across species. In order to circumvent this problem, we suggest leveraging the similarity between domain-domain interactions and chain-chain interactions. We propose the task of correctly distinguishing between interacting and non-interacting domain pairs within sets of homologous multi-domain proteins as a fine-tuning strategy designed to extract the knowledge of protein interaction partner specificity from pre-trained MSA-based language models. To solve this task, we use contrastive learning to fine-tune the MSA Transformer (Rao

[1]Centre for Artificial Intelligence, University College London [2]datavaluepeople [3]School of Life Sciences, Institute of Bioengineering, EPFL, Lausanne, Switzerland [4]SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland [5]Department of Computer Science, University College London [6]Institute of Structural and Molecular Biology, University College London. Correspondence to: Alex Hawkins-Hooker <ucabawk@ucl.ac.uk>.

Proceedings of the ICML 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery, Vienna, Austria. 2024. Copyright 2024 by the author(s).

et al., 2021) to correctly re-pair interacting domains within multi-domain MSAs whose rows have been scrambled. In contrast to prior work using the MSA Transformer without fine-tuning to resolve interaction specificity (Lupo et al., 2024b), fine-tuning in this way allows pairing to be performed efficiently in a single forward pass. We additionally exploit the MSA Transformer's ability for in-context learning, by allowing the model to condition its predictions for the scrambled rows on a set of correctly paired interacting sequences. We demonstrate the effectiveness of this strategy in producing accurately paired MSAs across a previously studied set of bacterial proteins for which ground-truth interaction partners are known. Furthermore, we show that paired MSAs produced by our model more strongly encode interface contacts than those paired with widely used sequence identity heuristics across a diverse set of prokaryotic complexes.

2. MSA Pairing Transformer

2.1. Contrastive learning on interacting domains

To generate a set of ground-truth paired MSAs for training an interaction partner predictor, we collect protein monomers containing one or more pairs of interacting domains according to the CATH database (Sillitoe et al., 2021). For each protein, we construct domain-level MSAs, M_A and M_B, corresponding to interacting domains A and B in the original protein, then simulate a pairing task by permuting the rows of M_B, so that the domain-A homologue in a given row in M_A no longer necessarily corresponds to (i.e. interacts with) the domain-B homologue in the same row in the permuted $\widetilde{M_B}$. We train a single model, across many such pairs of scrambled domain-level MSAs, to correctly re-pair the interacting domain sequences.

To solve this simulated pairing task, we introduce the MSA Pairing Transformer (MPT), a variant of the MSA Transformer (Rao et al., 2021) fine-tuned with contrastive learning. We apply the InfoNCE loss (Radford et al., 2021; Oord et al., 2019) to sequence-level representations h_A and h_B produced by the model for each sequence $x_A \in M_A$ and each sequence $x_B \in \widetilde{M_B}$ (respectively). This loss encourages h_A and h_B to be close if the two domain sequences belong to the same protein chain, and are therefore interaction partners, and pushes them apart otherwise:

$$\mathcal{L}_{B|A}(M_A, \widetilde{M_B}) = -\sum_{ij} z_{ij} \log \frac{\exp(g(h_A^{(i)}, h_B^{(j)}))}{\sum_k \exp(g(h_A^{(i)}, h_B^{(k)}))}\,, \tag{1}$$

where i and j are row indices in the two MSAs, z_{ij} is equal to one if $x_A^{(i)}$ and $x_B^{(j)}$ interact and zero otherwise, and $g(\cdot,\cdot)$ is the cosine similarity.

To allow the MSA Pairing Transformer to condition on known sets of pairs, where available, we jointly embed the two MSAs M_A and $\widetilde{M_B}$ by concatenating the sequences in each row, and feeding the concatenated MSA through the model. During training, we randomly sample a number of correctly paired rows to pass to the model alongside the unpaired rows that result from concatenating M_A and the scrambled M_B.

2.2. Architecture modifications

The MSA Transformer applies a variant of axial attention over MSAs, in which the row attention matrices are shared across all rows (i.e. sequences) (Rao et al., 2021). To accommodate the use of concatenated MSAs, we modify this row attention operation to prevent attention between unpaired concatenated sequences in the same row, by ensuring the corresponding attention weights are masked to zero. Row attention between *paired* concatenated sequences and attention within sequences is still shared. Paired rows are additionally indicated to the model via the addition of a learned 'paired row' input embedding. Finally, to allow the extraction of sequence-level representations for each sequence in the two concatenated MSAs, we add start tokens to the start of all sequences in $\widetilde{M_B}$ and all sequences in M_A before concatenation. The final layer representations of these start tokens are used as domain-level representations h_A and h_B in the loss.

2.3. Dataset and training details

We constructed a dataset of monomers containing at least two interacting domains from the CATH database (Sillitoe et al., 2021), based on the topology-based splits of CATH 4.3 proteins created by Hsu et al. (2022). In total, we used 17,263 chains for training and 183 chains for validation, ensuring there was no overlap in topology code between domains in training and validation chains. For each of these monomers we downloaded a precomputed full chain MSA from OpenProteinSet (Ahdritz et al., 2023), from which domain-level MSAs were extracted by using CATH domain annotations to identify residue slices corresponding to individual domains. We further excluded from the training set chains whose MSAs had significant homology (hhsearch e-value < 0.01) with the MSAs for either interaction partner in the 6 bacterial interactions studied in Section 4.1.

3. Pairing interaction partners with the MSA Pairing Transformer

We apply the MSA Pairing Transformer to the MSA pairing problem. In this setting, given chain-level MSAs M_A and M_B containing homologues of two interacting chains A and B, the goal is to return a list of pairs of proteins from the two MSAs that interact with each other. The task is simplified by the fact that species annotations are typically available for all the sequences in each MSA, and it can be assumed that sequences only interact if they belong to the same species. The remaining problem is that there can be multiple homologues of each chain in a single species, leading to ambiguity in the within-species pairings.

To apply the MPT to this task, we need a way to convert the chain-level embeddings returned by the model into pairing predictions. We assume that each sequence can interact with at most one other sequence within the same species and use this assumption to formulate an optimal matching problem, following previous work (Bitbol, 2018). For each pair of sequences $x_A^{(i)}$ and $x_B^{(j)}$ in a species S, we introduce an interaction score $I_{ij}^{(S)}$ computed from the corresponding embeddings, representing the log of the product of probabilities that the sequences interact:

$$I_{ij}^{(S)} = \log(p(y_B^{(i)} = j | x_A^{(i)}) p(y_A^{(j)} = i | x_B^{(j)})) \; , \qquad (2)$$

where $y_B^{(i)}$ denotes the index of the interaction partner of $x_A^{(i)}$ among the set of sequences x_B, and

$$p(y_B^{(i)} = j | x_A^{(i)}) = \frac{\exp(g(h_A^{(i)}, h_B^{(j)}))}{\sum_{\{k : x_B^{(k)} \in S\}} \exp(g(h_A^{(i)}, h_B^{(k)}))} \; . \qquad (3)$$

We then use the Hungarian algorithm to find the pairing that maximises the sum of the interaction scores across the set of paired sequences within each species. Interaction scores are computed for multiple species at a time, by passing unpaired MSAs comprising all sequences from the corresponding species through the model.

3.1. Iterative self-improvement via in-context learning

Previous work on pairing has made extensive use of iterative algorithms, in which the highest confidence predicted pairs in a given pairing round are treated as ground-truth pairs and used to update the pairing model for the next round (Bitbol et al., 2016; Gueudré et al., 2016). Inspired by the success of these approaches, we propose an iterative pairing algorithm (IPA) for the MSA Pairing Transformer which exploits its capacity for in-context learning.

To accommodate MSAs where the maximum number of pairable sequences N exceeds the maximum context size $M = 512$ encountered by the model during training, we partition unpaired sequences into $\frac{N}{M}$ partitions, in such a way that all unpaired sequences within a given species occur in the same partition. For each partition, we maintain a set of input pairs, initialised with a single pair of seed sequences known to interact, and iterate:

1. Score candidate pairs within each species by $I^{(S)}$, and predict within-species pairs given these scores.
2. Rank predicted pairs by decreasing value of $m^{(S)} \exp(I^{(S)})$, where the factor $m^{(S)}$ re-scales species-levels scores to make them more comparable across species.
3. Append the top K predicted pairs to the current set of input pairs, and return to step 1.

Iteration is terminated once all sequences in the partition are paired. The pairs predicted in each partition are concatenated to form the final set of predicted pairs. All results presented below are obtained with $K = 8$ and $m^{(S)} = \sqrt{n_A^{(S)} n_B^{(S)}}$, where $n_A^{(S)}$ is the number of paralogs in MSA A for species S. We explore the effect of normalisation in Appendix D.

4. Results

4.1. Predicting interaction specificity for bacterial interactions with known partners

To investigate whether our model could successfully transfer an understanding of interaction specificity from interacting domains within protein chains to interacting chains within complexes, we studied a set of bacterial interactions used in prior works on interaction partner prediction (Bitbol, 2018), for which ground-truth pairs are known due to the corresponding genes being co-located within operons. We first evaluate the accuracy of predictions in a ‘one-shot’ setting, in which the model is allowed to use a single known pair as input to guide its predictions for the remaining sequences. The accuracy of the model’s predicted pairings far surpassed that of a null baseline which predicted random pairings within each species, as well as the performance of a sequence-identity based ‘best hit’ heuristic similar to that employed in state-of-the-art structure prediction pipelines (Evans et al., 2022) (Figure 3). We additionally report results for different numbers of ground truth input pairs in Figure 5, demonstrating the ability of the model to condition on known pairs to improve its predictions.

We next explored whether the proposed iterative pairing strategy could achieve pairing accuracy competitive with previously proposed iterative pairing algorithms based on co-evolutionary signal. For each interaction, we evaluated performance given varying numbers N of total pairable sequences. For each $N \in \{64, 128, 256, 512, 1024\}$, we randomly subsampled species until the total number of pairable sequences was approximately equal to N, repeating the process 5 times. We compare the pairing accuracy of the iterative MPT algorithm with MI-IPA (Bitbol, 2018), a leading iterative pairing method, as well as the non-iterative ‘one-shot’ MPT. Iterating leads to a substantial increase in performance, indicating that the model is able to harness high-confidence predictions as in-context exemplars that can

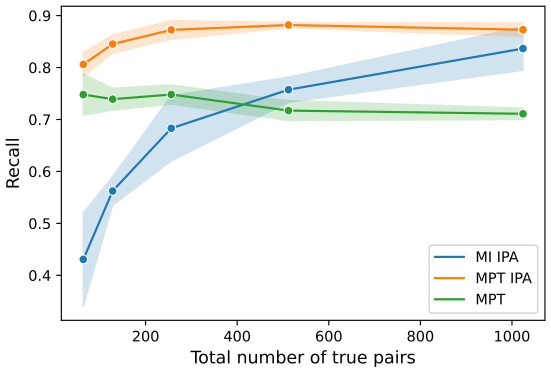

Figure 1. Pairing recall averaged across 6 bacterial interactions as a function of total number of pairable sequences. Since all sequences are paired, recall and precision values are the same.

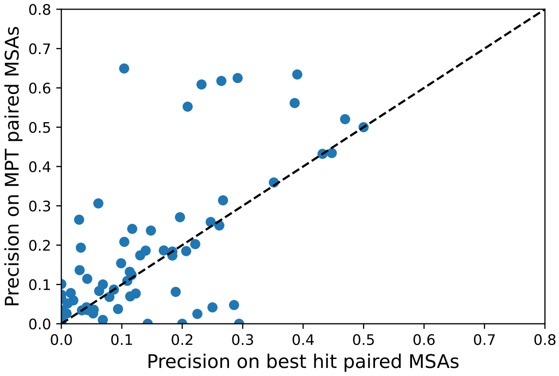

Figure 2. Precision of the top $0.2N$ interface contacts predicted by running GaussDCA on paired MSAs across a set of prokaryotic complexes with experimentally determined structures. N is the total number of interface contacts.

guide the pairing of more challenging sequences (Figure 1). The performance of the iterative MPT is almost independent of the total number of ground-truth pairs in the input MSAs to be paired by the model, in stark constrast to MI-IPA, which is only able to bootstrap its way to an accurate statistical model of the paired alignment when sufficiently large sets of pairable sequences are available, due to its inability to perform any kind of transfer learning across interactions.

4.2. Contact prediction on paired MSAs

As a further test of the success of pairing with the MSA Pairing Transformer, we studied the extent to which paired MSAs produced for prokaryotic complexes with known structures aided the prediction of interface contacts with co-evolutionary methods. We investigated a set of complexes studied in previous work (Green et al., 2021; Bryant et al., 2022), selecting for further study cases for which pure co-evolutionary analysis was able to correctly identify at least one correct interface contact within the top 10 predicted contacts (Green et al., 2021), despite the presence of paralogs making pairing non-trivial. For each complex, we generate chain-level MSAs with hhblits, then pair the MSAs using a modified version of the iterative MSA Pairing Transformer algorithm, as well as the best hit heuristic. For fair comparison we return only a single pair per species with both methods (Appendix C). Given paired MSAs, we run the co-evolution based contact prediction algorithm GaussDCA (Baldassi et al., 2014). Similar to Bryant et al. (2022), we evaluate the precision of the top interface contacts predicted by GaussDCA (Figure 2). In many cases, the paired MSAs returned by MPT lead to significant improvements in contact prediction performance, suggesting that they contain more correctly paired sequences from which accurate co-evolutionary inferences of interface structure can be made. We also evaluated performance for a set of eukaryotic complexes, for which we did not find evidence of improved contact prediction accuracy over the baseline method, although both methods perform substantially worse in the eukaryotic case, possibly due to overall weaker co-evolutionary signal or higher numbers of paralogs (Figure 4).

5. Discussion

Co-evolutionary signal is an important factor in the success of state-of-the-art methods in structure prediction and protein language modelling (Jumper et al., 2021; Lin et al., 2023; Abramson et al., 2024). The absence of large-scale datasets of known pairs of interacting sequences may therefore make it challenging to fully realise the potential of similar approaches for the study of protein-protein interactions. In this work, we investigate a strategy for exploiting known domain-domain interactions to allow the prediction of interaction partners within interacting protein families by fine-tuning the MSA Transformer. Interpretation of the extent to which our results indicate that the model has successfully learned to recognise generalisable sequence patterns encoding interaction specificity is made challenging by the fact that in many cases, the pretrained MSA Transformer may have seen examples of ‘fused’ proteins containing both partners in a given interaction in a single chain. Even in such cases, successfully transferring these patterns to unseen interaction partners is far from a trivial task, and an ability to do this accurately may be useful in structure prediction pipelines relying on paired MSAs. We have so far largely focussed on studying prokaryotic interactions, for which the possibility of producing ground-truth pairings based on genomic distance makes it easier to assess performance. In future work we will seek to better understand reasons for differences in performance of pairing algorithms across prokaryotic and eukaryotic interactions, and tailor our method to better handle the latter.

References

Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O., Willmore, L., Ballard, A. J., Bambrick, J., Bodenstein, S. W., Evans, D. A., Hung, C.-C., O'Neill, M., Reiman, D., Tunyasuvunakool, K., Wu, Z., Žemgulytė, A., Arvaniti, E., Beattie, C., Bertolli, O., Bridgland, A., Cherepanov, A., Congreve, M., Cowen-Rivers, A. I., Cowie, A., Figurnov, M., Fuchs, F. B., Gladman, H., Jain, R., Khan, Y. A., Low, C. M. R., Perlin, K., Potapenko, A., Savy, P., Singh, S., Stecula, A., Thillaisundaram, A., Tong, C., Yakneen, S., Zhong, E. D., Zielinski, M., Žídek, A., Bapst, V., Kohli, P., Jaderberg, M., Hassabis, D., and Jumper, J. M. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, pp. 1–3, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://www.nature.com/articles/s41586-024-07487-w. Publisher: Nature Publishing Group.

Ahdritz, G., Bouatta, N., Kadyan, S., Jarosch, L., Berenberg, D., Fisk, I., Watkins, A. M., Ra, S., Bonneau, R., and AlQuraishi, M. OpenProteinSet: Training data for structural biology at scale. *ArXiv*, pp. arXiv:2308.05326v1, August 2023. ISSN 2331-8422. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10441447/.

Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., and Pagnani, A. Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. *PLOS ONE*, 9(3):e92721, March 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0092721. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0092721. Publisher: Public Library of Science.

Bitbol, A.-F. Inferring interaction partners from protein sequences using mutual information. *PLOS Computational Biology*, 14(11):e1006401, November 2018. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1006401. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006401. Publisher: Public Library of Science.

Bitbol, A.-F., Dwyer, R. S., Colwell, L. J., and Wingreen, N. S. Inferring interaction partners from protein sequences. *Proceedings of the National Academy of Sciences*, 113(43):12180–12185, October 2016. doi: 10.1073/pnas.1606762113. URL https://www.pnas.org/doi/10.1073/pnas.1606762113. Publisher: Proceedings of the National Academy of Sciences.

Bryant, P., Pozzati, G., and Elofsson, A. 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.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. Protein complex prediction with AlphaFold-Multimer, March 2022. URL https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2. Pages: 2021.10.04.463034 Section: New Results.

Green, A. G., Elhabashy, H., Brock, K. P., Maddamsetti, R., Kohlbacher, O., and Marks, D. S. 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. Publisher: Nature Publishing Group.

Gueudré, T., Baldassi, C., Zamparo, M., Weigt, M., and Pagnani, A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. *Proceedings of the National Academy of Sciences*, 113(43):12186–12191, October 2016. doi: 10.1073/pnas.1607570113. URL https://www.pnas.org/doi/full/10.1073/pnas.1607570113. Publisher: Proceedings of the National Academy of Sciences.

Hopf, T. A., Schärfe, C. P. I., Rodrigues, J. P. G. L. M., Green, A. G., Kohlbacher, O., Sander, C., Bonvin, A. M. J. J., and Marks, D. S. Sequence co-evolution gives 3D contacts and structures of protein complexes. *eLife*, 3:e03430, September 2014. ISSN 2050-084X. doi: 10.7554/eLife.03430. URL https://doi.org/10.7554/eLife.03430. Publisher: eLife Sciences Publications, Ltd.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 8946–8970. PMLR, June 2022. URL https://proceedings.mlr.press/v162/hsu22a.html. ISSN: 2640-3498.

Hwang, Y., Cornman, A. L., Kellogg, E. H., Ovchinnikov, S., and Girguis, P. R. Genomic language model predicts protein co-regulation and function. *Nature Communications*, 15(1):2880, April 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-46947-9. URL https://www.nature.com/articles/s41467-024-46947-9. Publisher: Nature Publishing Group.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://www.nature.com/articles/s41586-021-03819-2. Publisher: Nature Publishing Group.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. 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.

Lupo, U., Sgarbossa, D., and Bitbol, A.-F. Pairing interacting protein sequences using masked language modeling. *Proceedings of the National Academy of Sciences*, 121(27):e2311887121, 2024a. doi: 10.1073/pnas.2311887121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2311887121.

Lupo, U., Sgarbossa, D., Milighetti, M., and Bitbol, A.-F. DiffPaSS – Differentiable and scalable pairing of biological sequences using soft scores. In *ICLR 2024 Workshop on Generative and Experimental Perspective for Biomolecular Design*, 2024b.

Malbranke, C. and Bitbol, A.-F. A proteome-scale masked language model for fast protein-protein interaction prediction. In *ICLR 2024 Workshop on Generative and Experimental Perspectives for Biomolecular Design*, 2024.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. ColabFold: making protein folding accessible to all. *Nature Methods*, 19(6): 679–682, June 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01488-1. URL https://www.nature.com/articles/s41592-022-01488-1. Publisher: Nature Publishing Group.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748 [cs, stat].

Ovchinnikov, S., Kamisetty, H., and Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife*, 3:e02030, May 2014. ISSN 2050-084X. doi: 10.7554/eLife.02030.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Rao, R. M., Liu, J., Verkuil, R., Meier, J., Canny, J., Abbeel, P., Sercu, T., and Rives, A. MSA Transformer. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8844–8856. PMLR, July 2021. URL https://proceedings.mlr.press/v139/rao21a.html. ISSN: 2640-3498.

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J., and Orengo, C. A. CATH: increased structural coverage of functional space. *Nucleic Acids Research*, 49(D1):D266–D273, January 2021. ISSN 1362-4962. doi: 10.1093/nar/gkaa1079.

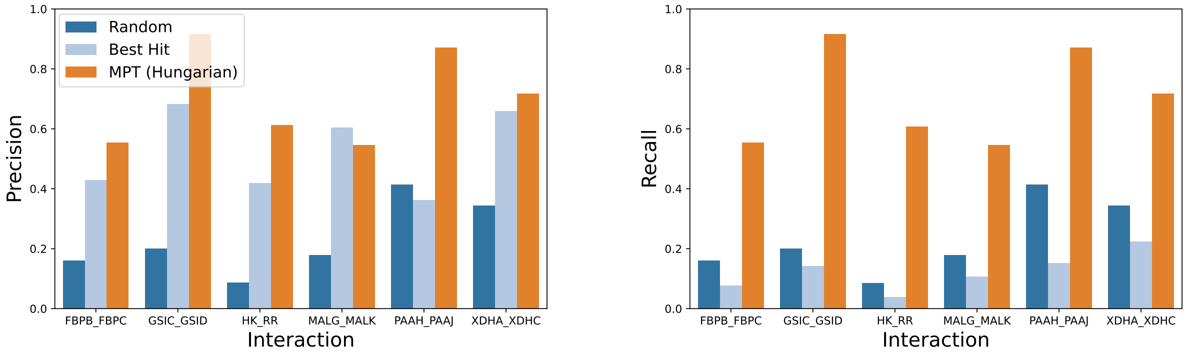

Figure 3. Precision (left) and recall (right) of prediction of interaction partners on a set of 6 bacterial interactions with known ground truth from Bitbol (2018). Precision and recall numbers are identical for methods which produce a complete pairing (MPT and Random).

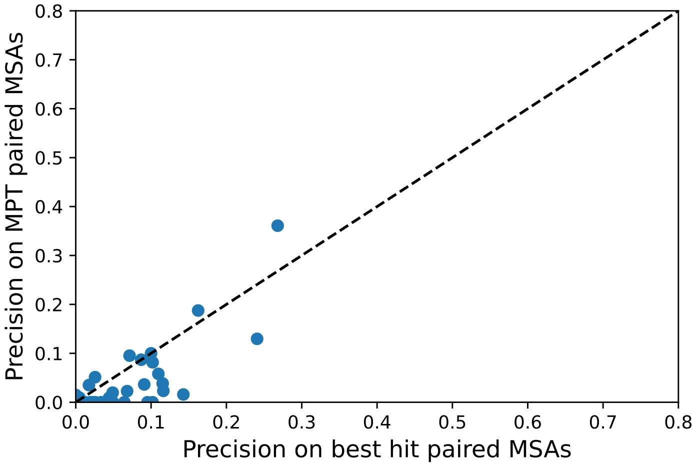

Figure 4. Precision of the top $0.2N$ interface contacts predicted by running GaussDCA on paired MSAs across a set of eukaryotic complexes with experimentally determined structures. N is the total number of interface contacts.

A. Model

Our model is based on the MSA Transformer, a masked language model trained to reconstruct masked tokens within (single-chain) MSAs. To adapt the MSA Transformer for the problem of interaction partner prediction, we propose a number of modifications which allow the model to more naturally handle multi-chain MSAs, and to be fine-tuned as an interaction partner predictor. In particular, we use attention masking and row embeddings to allow the model to distinguish between paired and unpaired rows, thereby endowing the model with the ability to condition its pairing predictions on known or previously predicted sets of pairs. These modifications are described in detail in the following sections.

A.1. MSA Transformer

The MSA Transformer is a protein language model which operates directly on multiple sequence alignments rather than individual sequences. An input multiple sequence alignment is represented as a one-hot tensor $M \in \mathbb{R}^{N \times L \times K}$, where N is the number of rows in the MSA, L the number of columns, and K the size of the amino acid alphabet. The model stacks

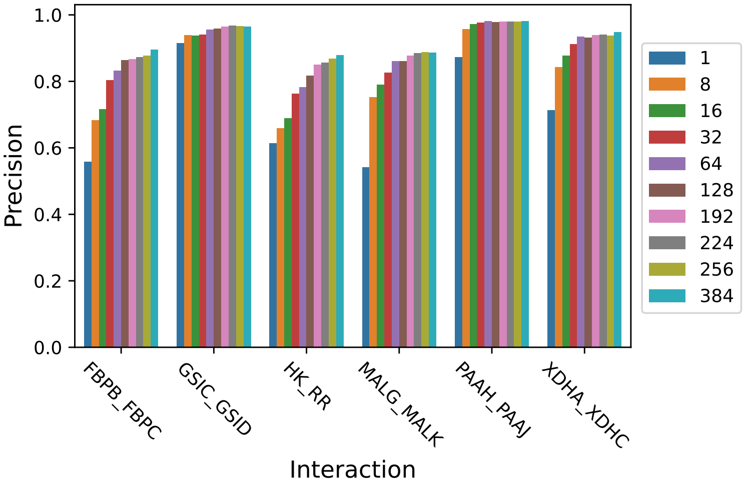

Figure 5. Few-shot pairing using the MSA Pairing Transformer. Where available, a set of ground-truth exemplars can be passed as in-context exemplars to guide the pairing of remaining sequences. To demonstrate the effectiveness of few-shot pairing, we report the precision of predicted pairs as a function of the number n of (randomly sampled) correctly paired sequences passed as input to the model. A complete pairing of the remaining sequences is predicted by using the Hungarian algorithm to assign pairs given interaction scores returned by the model. Evaluation performed on the 6 bacterial interactions from Bitbol (2018).

multiple layers, each performing axial attention over rows and columns, to produce an updated set of hidden representations of each amino acid in the MSA $H \in \mathbb{R}^{N \times L \times D}$. Importantly, the row attention within each layer is shared across all rows, motivated by the the fact that each sequence in an MSA satisfies a closely related set of constraints reflecting a common underlying structural graph. The pre-softmax shared row attention weights within each attention head are thus calculated by aggregating interaction scores between each pair of columns over all rows:

$$a_{ij} = \frac{\sum_{r=1}^{N} (Q\mathbf{h}_{ri})^T (K\mathbf{h}_{rj})}{\sqrt{Nd}} = \frac{z_{ij}}{\sqrt{Nd}} \tag{4}$$

Here r indexes a row, and i and j represent the columns between whose embeddings attention is to be computed, and d is the head dimension. The denominator multiplies the standard scaled dot product attention normalization factor by an additional factor of $\sqrt{N}$ to account for differences across MSAs in the number of rows N over which interaction scores are summed.

A.1.1. Chain-based Attention Masking

In our multi-chain setting, the one-hot input MSA M is formed by concatenating the rows of multiple sub-MSAs $M_A, M_B, ...$ corresponding to the chains $A, B, ...$ of a protein complex. In the remainder we will focus on the case of a pair of MSAs representing a dimeric complex for simplicity. To preserve the semantics of row attention, we choose to prevent row attention across chain boundaries for rows containing unpaired (non-interacting) sequences.

To achieve this, we first partition the N rows of an input MSA into a set $M^{(p)}$ of $N_p = |M^{(p)}|$ paired rows and a set $M^{(u)}$ of $N_u = |M^{(u)}|$ unpaired rows. Thus each row in $M^{(p)}$ is a concatenated pair of sequences x_A, x_B, where x_A and x_B are assumed paired. Rows in $M^{(u)}$ are still formed from concatenated sequences x_A and x_B, but crucially these sequences are no longer assumed to be (correctly) paired. In typical settings we will often start with one paired row, corresponding to a pair of 'seed' sequences known or hypothesised to interact in a particular species. Other sources of paired rows are discussed below.

Given an MSA whose rows are partitioned in this way, we then use attention masking to control the row attention operation in each attention head so that paired rows are allowed to perform both intra- and inter-chain attention, while unpaired rows can perform only intra-chain attention.

Thus, for unpaired rows:

$$a_{ij}^{(u)} = \begin{cases} a_{ij} \text{ , if i and j are in the same chain} \\ -\infty \text{ otherwise} \end{cases} \tag{5}$$

where a_{ij} is the standard MSA Transformer pre-softmax shared row attention weight between columns i and j computed over the whole MSA (Equation 4). For paired rows,

$$a_{ij}^{(p)} = \begin{cases} a_{ij} \text{ , if i and j are in the same chain} \\ \frac{\sum_{r \in X^{(p)}} (Q\mathbf{h}_{ri})^T (K\mathbf{h}_{rj})}{\sqrt{N_p d}} \text{ otherwise} \end{cases} \tag{6}$$

A.1.2. Chain break and row encoding

To make the additional structure introduced into the MSAs above explicitly available to the model we modify the input encoding employed by the MSA Transformer in two ways. First, we add a special chain-start token at the start of each chain, rather than just at the start of each row as in the original MSA Transformer. This allows the model to reason about positions relative to chain breaks. Second, we add embeddings to paired rows to help the model distinguish them from unpaired rows. Whereas in the original MSA Transformer each row's index within the set of all rows in the MSA is embedded, we now embed the index within the set of paired rows for paired rows only.

A.1.3. Loss Details

The output of the MSA Pairing Transformer is a set of embeddings $H \in \mathbb{R}^{N, L_A + L_B, D}$. We supervise these embeddings in two ways. First, we employ the masked language modelling objective used by the original MSA Transformer: within each MSA a random set of amino acids are masked, and the model is tasked with predicting them. Second, we introduce an explicit contrastive pairing los. The final loss is a weighted combination of a masked language modelling loss and a symmetrised contrastive loss:

$$\mathcal{L}(M_A, \widetilde{M_B}) = 0.7\mathcal{L}_{MLM}(M_A, \widetilde{M_B}) + 0.15(\mathcal{L}_{B|A}(M_A, \widetilde{M_B}) + \mathcal{L}_{A|B}(M_A, \widetilde{M_B})) \tag{7}$$

B. Best hit baseline

In our implementation, the 'best hit' heuristic sorts all sequences in each species in the MSA for each chain by their sequence identity to the seed sequence, then pairs the 'best hit' within a species in M_A to the 'best hit' within the same species in M_B. Variants of this procedure are used in state-of-the-art complex prediction pipelines (Abramson et al., 2024; Evans et al., 2022; Mirdita et al., 2022).

C. MSA pairing for contact prediction

Given the large size of the unpaired MSAs used to evaluate complex contact prediction, we apply a more efficient version of the iterative pairing algorithm to pair this set of MSAs. Instead of running iteration over each of $\frac{N}{M}$ partitions, we run iteration within a single species-based partition, then use the predicted pairs from this partition as input pairs to guide the pairing of the remaining unpaired sequences. Given that co-evolution based contact prediction can be affected by the size of the input MSA, we ensure that the size of the paired MSAs returned by the best hit baseline and the MPT are the same by returning only the pair with the highest interaction score within each species.

D. Analysis of pairing confidence

The iterative pairing scheme proposed above relies on associating a confidence score to predicted pairs. To demonstrate the effectiveness of the proposed confidence score, we sort all predicted pairs by confidence and calculate the precision of predicted pairings within the top K ranked pairs for all values of K (Figure D). High confidence pairings are heavily enriched for true pairs, while normalising by number of paralogs (as described in Section 3.1) helps improve the ranking of pairs across species.

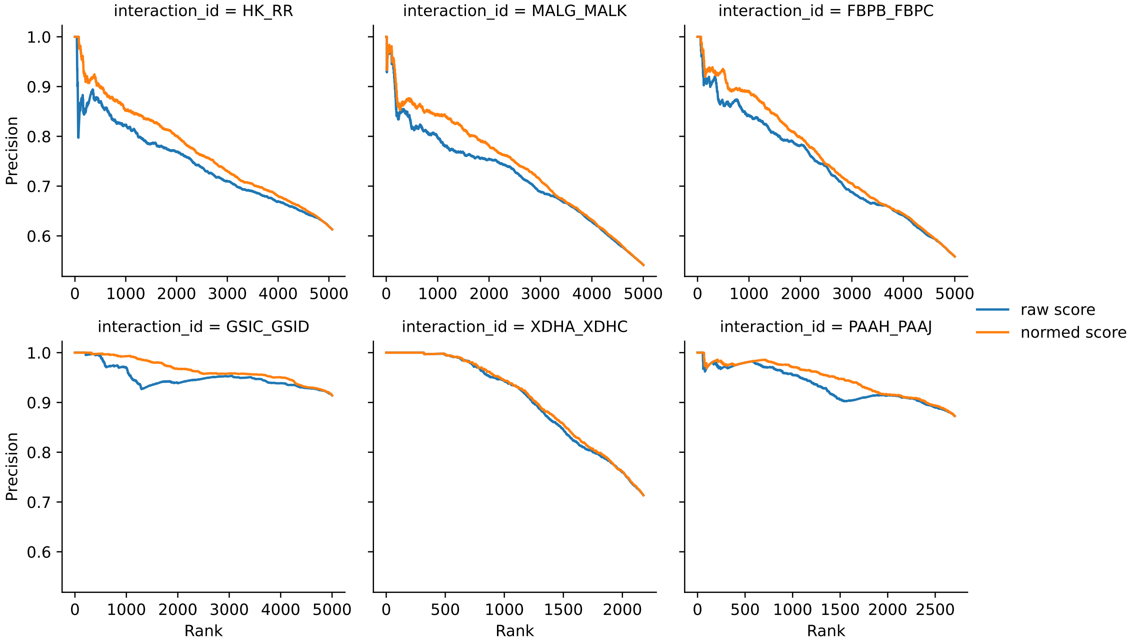

Figure 6. Precision of top K predicted pairs after sorting by interaction score, for all values of K. Results reported across the 6 bacterial interactions from Bitbol (2018).