nature portfolio

Peer Review File

Atlas of Predicted Protein Complex Structures Across Kingdoms

Corresponding Author: Dr Dacheng Ma

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The use of PDockQ with AlphaFold2.3 (or ColabFoldv3) is unfortunately not that good (see doi: 10.1093/bioinformatics/btad424). THerefore the entire conclusions on "reliable" predictions is flawed and needs to be redone.

Further it has also been observed that high ipTM scores are not sufficient to identify high-confident homomeric interactions, this needs to be benchmarked carefully first.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
In this article, the authors utilized AlphaFold2-based ColabFold to construct an atlas of around 1 million protein complexes, accompanied by their evaluated predictive confidence scores. They then applied this protocol to identify protein-protein interaction networks in bacteria, archaea, and human proteome. They also investigated human-virus protein interactions and proteins involved in fusion and fission events. Finally, the authors utilized this structural dataset to develop a deep-leaning model for predicting the protein surface binding sites and fine-tuned the proteinMPNN model to enhance protein design accuracy. Several computational studies were validated by their experiments.

This study demonstrates an effective way to utilize AlphaFold-predicted structural databases for expanding the existing protein-complex database, as well as facilitating the development of new models for evaluating protein-protein interactions. The authors also released their predicted protein complexes and codes, ensuring the transparency of the research. I have a few suggestions listed below:

1. The identification of PPIs in bacteria and archaea based on sequence co-localization is interesting, but can the authors illustrate how they address those protein pairs that are located far apart in the genomic sequence?
2. Since the code of AlphaFold3 has been released, it would be interesting to compare its predictions with the model's performance in this paper, as well as with other deep-learning structural complex prediction tools.
3. In the evaluation of Multimer-Surface and the finetuned ProteinMPNN, the authors utilized the predicted complexes. It is better to evaluate the model performance using an independent dataset composed of experimentally determined structures.
4. Extended Figure 3: the workflow in panel a is a little confusing, especially the distinction between the shaded and unshaded parts.
5. Figure 6e: The definition of accuracy and perplexity in this context is somewhat confusing. For example, what does the "percentage of correct amino acids recovered" refer to?

(Remarks on code availability)

The authors provided clear instructions for installing and running the code. I did not test the code due to its significant storage requirements.

Reviewer #3

(Remarks to the Author)
Qi et al. present a large-scale resource of protein-protein interaction (PPI) predictions across multiple kingdoms. Using ColabFold, they predicted 1.15 million PPIs, identifying 170,001 high-confidence interactions. While most large-scale PPI studies focus on human proteins, this work expands the scope to other kingdoms, addressing an essential gap in the field.

The study explores multiple fronts, including evolutionary analysis of domain fusion, detection of viral-human PPIs, and the use of predicted interactions for training machine learning classifiers—demonstrating broad applicability. However, the breadth of the study appears to come at the expense of depth in each specific area. My main concerns with this work are around comparison to previously published work and homology leakage of the classifiers.

Major:

- The paper does not include a comprehensive comparison to other core resources, despite many studies having already produced large-scale PPI prediction datasets. While some of this work is cited, it is not compared against it. For example, the Predictomes study by Schmid et al. (2024) and Zhang et al. (2024), Computing the Human Interactome, as well as the virulence factor study by Humphreys et al., Protein interactions in human pathogens revealed through deep learning, are highly relevant. Please include a comparison to quantify the novelty of predictions.

- Figure 1g: The tree visualization is not very informative, as the labels are difficult to read. Additionally, it is unclear how much novelty has been generated per species. A quantitative assessment of novelty at different taxonomic levels (species, phylum, and kingdom) would be critical to understanding the contribution of this resource.

- Figure 3a and 3j: These figures are difficult to interpret. I recommend reconsidering the analysis and exploring alternative visualization. A more intuitive approach, e.g. for j maybe some Hierarchical clustering, could help convey the key insights more easily.

- The application of fine-tuning ProteinMPNN is inconclusive. Structural similarity can be detected beyond 30% sequence identity, meaning sequence-based separation may not be sufficient. I recommend using Foldseek-Multimer or a PINDER-like approach to separate training and testing.

- The Multimer-Surface predictor is presented as an application of the database, but its evaluation has multiple issues. Currently, a random split is used, which may lead to homology leakage, making it unclear whether the model genuinely improves surface extraction. To prevent data leakage, I recommend splitting the test set based on interface similarity, ensuring that the system does not train and test on the same interface type. PINDER provides a well-structured benchmark dataset, and adopting their de-replication strategy could improve the potential leakage issue.
Additionally, a comparison against a PDB-trained method is necessary to determine whether the predicted database provides meaningful improvements over existing resources.

- The manuscript states: "All other data supporting the findings of this study are available from the corresponding author on reasonable request." To ensure full transparency and reproducibility, all necessary data should be made openly available. Currently, only the 170,000 high-quality predictions have been released—please consider also providing the remaining low-quality predictions to allow for further analysis.
Additionally, the code repository is missing essential scripts for the analysis and lacks documentation necessary for reproducing the study.

Minor:

- For others in the community trying to do large-scale PPI runs it would be great to have some technical information regarding runtime for MSA generation and inference of the models.

- Kim et al. (2024) predicted over 350,000 viral protein structures through BFVD. This could be as well referenced in the following sentence: "Recent research has predicted 67,715 newly identified viral protein structures, uncovering a multitude of structurally distinct viral proteins.".

- The code for in GitHub and has no license: https://github.com/Protein-Complex-Lab/Protein-Complex-Atlas

(Remarks on code availability)
The code repository includes many components of the study but still lacks some parts of the analysis. Additionally, it is missing a license file and a documentation to be more comprehensive would also be beneficial.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In this paper the authors provide an impressive set of predictions of PPIs. Unfortunately, this is also the major problem with the paper. It reads as a list of rather unrelated problems, all of potential interest, but none of them is explained in sufficient detail to provide interesting insights.

The most obvious part that is out of sight is the fine-tuning of ProteinMPNN. What does that add for our understanding of PPIs?

Another unclear part is what negative set is used for the homomer benchmark (Fig. S5) or how the virus-human dataset is constructed (Homology reduction to training set, etc).

I am convinced that this can become an interesting resource, but unfortunately, by trying to add too much into a single paper, the authors make it almost unreadable.

Availability;

It is unclear what is available where and under what license, see remarks below. Ideally the website www.biopredictnavigator.cn should provide an easy access to the data. However it is far from ready (https does not work), it contains still images of the proteins, the coloring scheme in the 3D viewer is not consistent with the labels (it is colored from N-C and not by pLDDT). etc.

Language:

The paper contains non-optimal expressions making it hard to read. It really needs a good rewrite of someone who can tell a story and not just staple sentences on top of each others.

One of these is "showed outperformed efficacy", I guess it should be "showed an efficacy" or something like that.

Another one is "Then we fine-tuned the released sourced model parameters using the distilled high-quality data". WHat do the authors mean here?

/Arne Elofsson

(Remarks on code availability)
I have not looked at the code, partly because I did not have time but also because the entire data structure was confusing.

The data availability is almost as confusing as the paper.

The github https://github.com/wensm77/Protein-Complex-Atlas

The site: https://www.biopredictnavigator.cn/ is not available only http://www.biopredictnavigator.cn/ (not secure giving a warning for most users)

The data that should be on https://www.modelscope.cn/collections/protein_complex_atlas-2ae5e7d4f4a343 is found at https://www.modelscope.cn/datasets/Yesir97/Protein-Complex-Atlas (I think), and this links to another GitHub site: https://github.com/Protein-Complex-Lab/Protein-Complex-Atlas, which in turn links to the HuggingFace site.

No doi numbers are provided, and it is not clear what license applies to what (many things seem to be cc-by-nc-nd, but not sure this is consistent with, for instance, colabfold, which is apache-2 (I think).

Reviewer #2

(Remarks to the Author)
The authors have addressed my questions, and I appreciate their efforts.

(Remarks on code availability)
I did not run through the code due to its significant storage requirement, but the README file has provided very detailed and clear instructions.

Reviewer #3

(Remarks to the Author)
I would like to thank the authors for considering all my comments.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The choice of using the "Commons CC BY-NC 4.0 license" makes this paper unsuitable for publication in a open access journal. The authors should change it to Apache 2.0 to allow commercial and non-commercial use of this method.

Given this serious shortcoming, I have not bothered about reading the rest of the paper.

(Remarks on code availability)

Response to Referees Letter

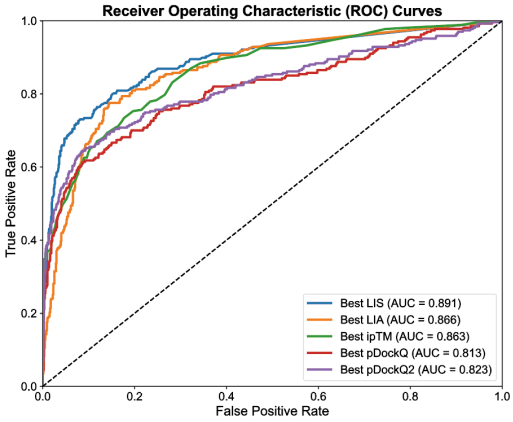
Reviewer #1 (Remarks to the Author):

The use of PDockQ with AlphaFold2.3 (or ColabFoldv3) is unfortunately not that good (see doi: 10.1093/bioinformatics/btad424). THerefore the entire conclusions on "reliable" predictions is flawed and needs to be redone.

Further it has also been observed that high ipTM scores are not sufficient to identify high-confident homomeric interactions, this needs to be benchmarked carefully first.

Re: We thank the reviewer for highlighting recent evidence that (i) pDockQ performance degrades for the AF2.3 (ColabFold v3) model, and (ii) high ipTM alone does not reliably distinguish true from false multimeric (especially homomeric) interactions. We have therefore re-evaluated and redone all downstream analyses using metrics whose discriminative performance we explicitly benchmarked in the revised manuscript, and we no longer rely on pDockQ/ipTM as a sufficiency criterion for “high-confidence” complexes.

Recently, multiple metrics have been developed to distinguish between true and false interaction in predicted protein complexes, including Predicted DockQ (pDockQ)[1], pDockQ2[2], Local Interaction Score (LIS)[3], Local Interaction Area (LIA)[3], AlphaFold Multimer-derived interface predicted Template Modeling (ipTM) score[4], and linear models specifically designed for homodimer identification[5].

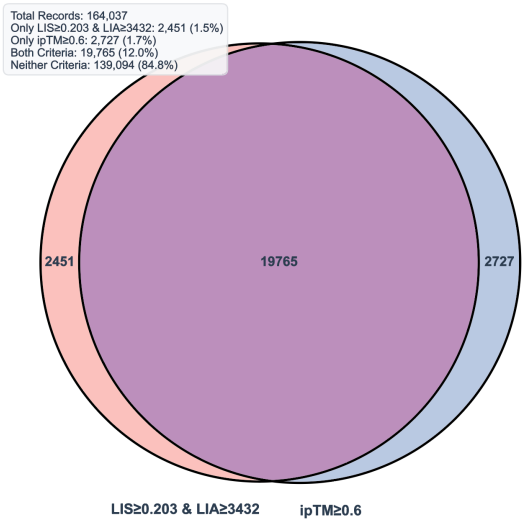


Response Figure 1. Receiver-operating-characteristic (ROC) curves comparing the discriminative power of five scores computed from the rank-1 model. Best Local Interaction Score (LIS), Best Local Interaction Area (LIA), interface-predicted TM-score (ipTM), pDockQ, and pDockQ2. Areas under the curve (AUC) are listed in the legend; Best LIS yields the highest AUC (0.891), followed by Best LIA (0.866) and Best ipTM (0.863). Source data was derived from previous research[3].

For heterodimer benchmark, a recent study has demonstrated that, when evaluating the

top-ranked model among 5 generated models for each predicted PPI complex, the Best LIS metric provides superior Receiver Operating Characteristic (ROC) performance for heterodimer identification compared to ipTM, pDockQ, and pDockQ2[3], indicating better discrimination between true positives and false positives. We confirmed the result as shown in Response Figure 1 by reanalyzing the results. Furthermore, as this research recommended, we confirmed that the combination of Best LIS and Best LIA metrics yielded the highest precision rate among evaluated scoring schemes (Supplementary Figure 4a). Meanwhile, we found that raising the threshold monotonically increases precision (blue bars) at the cost of decreased recall (green bars), illustrating how users can tune the score to prioritize either low false-positive or low false-negative rates depending on downstream applications (Supplementary Figure 4b).

Therefore, we adopted a threshold of Best LIS≥0.203 and Best LIA≥3432 as criteria to identify high-confidence heterodimeric PPIs for subsequent analyses and redone all the study in this revised manuscript.



Response Figure 2. Venn diagram showing the overlap between positives selected by the combined criterion Best LIS ≥ 0.203 & Best LIA ≥ 3432 and those selected by ipTM ≥ 0.6 across 164,037 predictions for heterodimers in 36 pathogens.

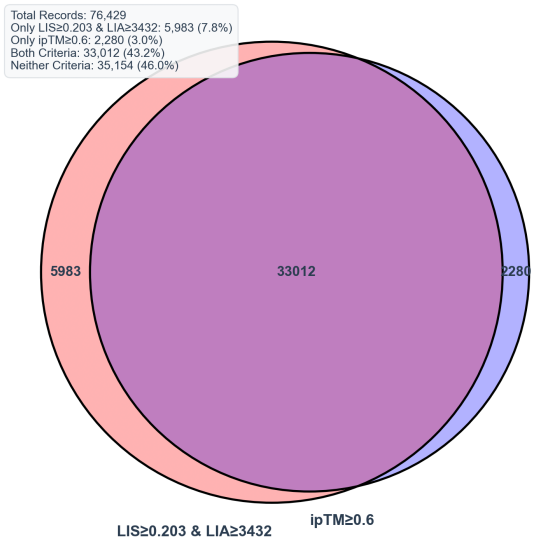
To compare Best LIS ≥ 0.203 plus Best LIA ≥ 3432 with the previously used ipTM ≥ 0.6 cutoff on heterodimers, as shown in Response Figure 2, we analyzed predictions for 36 pathogenic bacteria. The two criteria agreed on ~90% of the positive calls, with roughly 10% being discordant.

Secondly, to address the reviewer's concern about ipTM on homodimers, we constructed an evaluation set of 126 experimentally validated homodimers and 285 monomeric (non-dimerizing) proteins. For every structure we calculated eight scores: LIS, ipTM, pDockQ2, and the five probabilities introduced in the previous study[5]—dimer_proba_pae3, dimer_proba_pae4, dimer_proba_con3, dimer_proba_pae4_con3, and dimer_proba on the ipTM rank-1 model.

As shown in Supplementary Figure 5a, Receiver-operating-characteristic analysis showed that Best LIS provided the best overall discrimination (AUC = 0.922), followed closely by ipTM (AUC = 0.918) and pDockQ2 (AUC = 0.907).

Meanwhile, although the remaining classifiers displayed lower AUCs, dimer_proba_con3 achieved the high precision (0.844), with dimer_proba_pae4_con3 ranking second (0.830); the precision of pDockQ2 was 0.745.

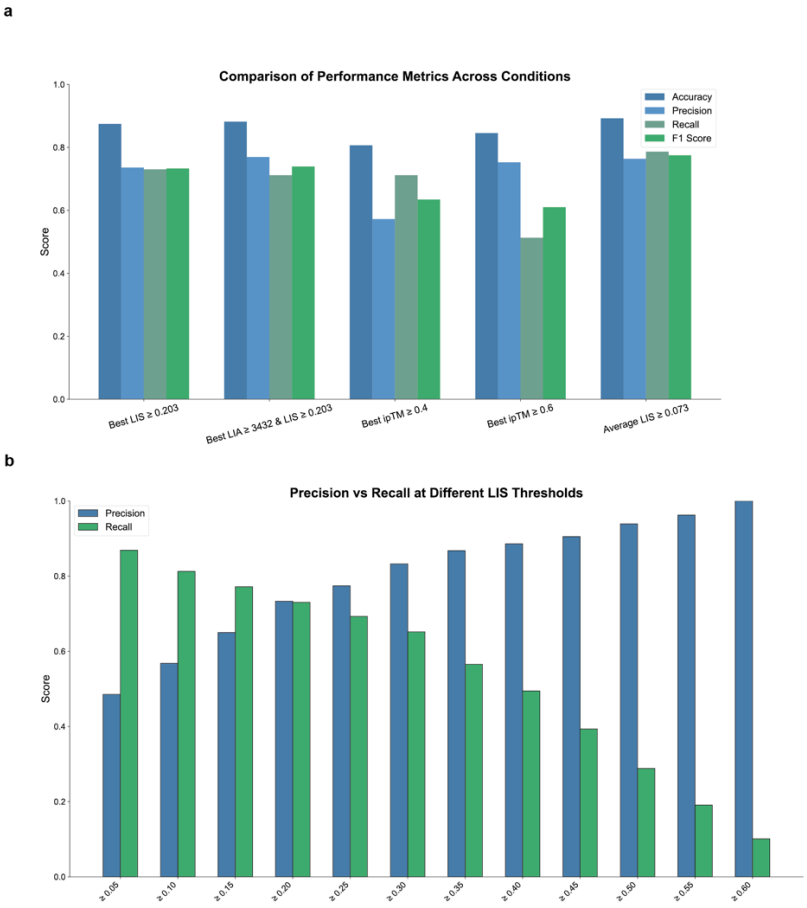
Since the Best LIS showed strongest overall discrimination, we further test the combination of the Best LIS with Best LIA as the metric for the homodimers. As shown in Supplementary Figure 5c, we found that such combinations keep similar discrimination ability while showed higher accuracy. On the basis of these analysis, we selected Best LIS and Best LIA combination (Best LIS≥0.203 and Best LIA≥3432) for subsequent prediction of homodimer formation.



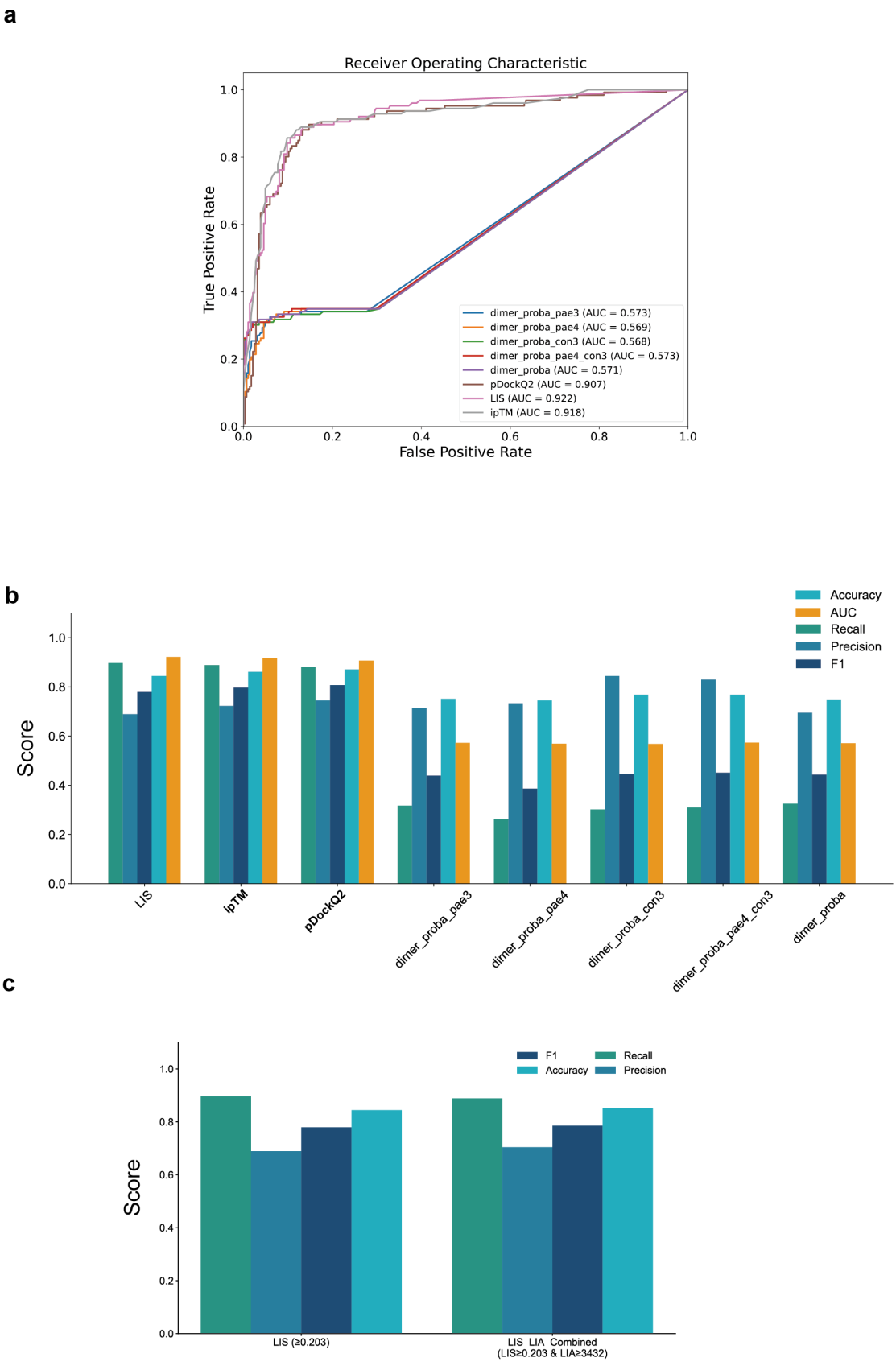
Response Figure 3. Venn diagram showing the overlap between positives selected by the combined criterion Best LIS ≥ 0.203 & Best LIA ≥ 3432 and those selected by ipTM ≥ 0.6 on homodimers in 36 pathogens.

Similarly, as shown in Response Figure 3, we also performed comparison of Best LIS ≥ 0.203 plus Best LIA ≥ 3432 with the previously used ipTM ≥ 0.6 cutoff on homodimers.

Collectively, we appreciate the reviewer's insightful comment. In response, we have (i) replaced the original metric, which performs sub-optimally in AF2.3 contexts, (ii) provided explicit benchmark results showing that the new metric pair performs better for both hetero- and homomeric interactions, and (iii) redone all downstream analyses using this more accurate confidence metric.



Supplementary Figure 4. Performance of scoring metrics for distinguishing true and false heterodimeric PPIs. **a** Summary of accuracy, precision, recall and F1-score obtained with representative decision rules. Columns from left to right: (i) Best LIS ≥ 0.203; (ii) combined filter Best LIA ≥ 3432 and Best LIS ≥ 0.203 (baseline cut-off used in this study); (iii) Best ipTM ≥ 0.40; (iv) Best ipTM ≥ 0.60; and (v) Average LIS ≥ 0.073. The combined LIA + LIS filter provides the highest overall precision, whereas Average LIS threshold yields the greatest recall. **b** Precision–recall trade-off for progressively stricter LIS thresholds (≥ 0.05 to ≥ 0.60). Raising the threshold monotonically increases precision (blue bars) at the cost of decreased recall (green bars), illustrating how users can tune the score to prioritize either low false-positive or low false-negative rates depending on downstream applications.



Supplementary Figure 5. Benchmark evaluation of homodimer prediction metrics. a Receiver operating characteristic (ROC) curves for various metrics assessing homodimer prediction performance, including Best LIS (Local Interaction Score, AUC = 0.922), ipTM (interface predicted Template Modeling, AUC = 0.918), pDockQ2

(AUC = 0.907), and several linear models-based classifiers (dimer_proba_pae3, dimer_proba_pae4, dimer_proba_con3, dimer_proba_pae4_con3, and dimer_proba), which yielded relatively lower AUC values (~0.57). Higher AUC indicates better discriminatory ability, with the diagonal line representing random performance (AUC = 0.5). All these data were analyzed on the ipTM rank-1 model. **b** Comparative bar plot illustrating multiple performance metrics (AUC, accuracy, precision, recall, and F1 score) across the evaluated classifiers. Notably, although the linear models-based classifiers exhibit lower AUC values, they achieve higher accuracy, suggesting effective performance when evaluated through a strict classification threshold. Conversely, LIS, ipTM and pDockQ2 metrics display superior overall discriminative power (high AUC). **c** Bar-plot comparison of classification performance when using the Best LIS cut-off alone (LIS ≥ 0.203) versus a composite filter that requires both LIS ≥ 0.203 and Best LIA ≥ 3 432.

Reviewer #2 (Remarks to the Author):

In this article, the authors utilized AlphaFold2-based ColabFold to construct an atlas of around 1 million protein complexes, accompanied by their evaluated predictive confidence scores. They then applied this protocol to identify protein-protein interaction networks in bacteria, archaea, and human proteome. They also investigated human-virus protein interactions and proteins involved in fusion and fission events. Finally, the authors utilized this structural dataset to develop a deep-leaning model for predicting the protein surface binding sites and fine-tuned the proteinMPNN model to enhance protein design accuracy. Several computational studies were validated by their experiments.

This study demonstrates an effective way to utilize AlphaFold-predicted structural databases for expanding the existing protein-complex database, as well as facilitating the development of new models for evaluating protein-protein interactions. The authors also released their predicted protein complexes and codes, ensuring the transparency of the research. I have a few suggestions listed below:

Re: We thank Reviewer #2 for the positive assessment of our large-scale proteome- and host–virus–spanning complex atlas and for the constructive suggestions. In revision we (i) extended bacterial/archaeal PPI discovery beyond local genomic proximity by integrating large-scale genomic co-occurrence with an MSA-free Light Chai-1 screen and confirmatory ColabFold modeling, identifying high-confidence long-range pairs (Fig. 6 and Supplementary Fig. 24) and benchmarking a fine-tuned sequence PPI classifier (Response Fig. 4); (ii) added an independent, non-redundant benchmark comparing AlphaFold3, ColabFold, Chai-1, Protenix, and Boltz-2 across homodimers and heterodimers (Supplementary Fig. 21), plus a single-diffusion-step acceleration (Supplementary Fig. 23-24); (iii) established leakage-prevented splits for MPBind (protein binding sites prediction) and ProteinMPNN via multimer/monomer clustering and interface similarity filtering, and evaluated on both predicted and newly deposited PDB complexes (updated Fig. 6 and Supplementary Fig. 22); (iv) clarified workflows (revised Supplementary Fig. 13) and standardized metric terminology (sequence recovery rate, test perplexity); and (v) completed open source: full ~1.1 M complex set and added missing scripts, licensing, and documentation. In sum, the revisions strengthen the study's quality and sophistication, ensuring it meets a higher scholarly standard.

1. The identification of PPIs in bacteria and archaea based on sequence co-localization is interesting, but can the authors illustrate how they address those protein pairs that are located far apart in the genomic sequence?

Re: Thank you for the reviewer's insightful suggestion. We fully agree that proteome-wide interaction mapping in bacteria is valuable; however, a single bacterial genome typically

encodes thousands of proteins, leading to millions of possible pairs, and exhaustive structural prediction remains computationally prohibitive.

A recent study introduced RoseTTAFold2-Lite, which uses multiple-sequence alignments (MSAs) to survey 78 million protein pairs across 19 pathogenic bacteria and confidently recovered 1,923 complexes for 3D structure predictions[6]. Interestingly, they also showed that the found pairs are mainly located nearby in genome. Despite its speed, the method still requires large-scale MSA generation.

The new Chai-1 framework removes this bottleneck by replacing MSAs with ESM embeddings, enabling MSA-free interaction prediction. Additionally, AlphaFold3-style diffusion models including Chai-1 can supply interaction scores in a single diffusion step without building full atomic models.

To push this concept further, we hypothesized that genomic co-occurrence can complement structure-based scoring, especially for interacting partners encoded far apart on the genome (Figure 6c). We therefore collected 21,676 reference prokaryotic genomes (~86 million proteins). As a proof of principle, we focused on *Campylobacter jejuni*.

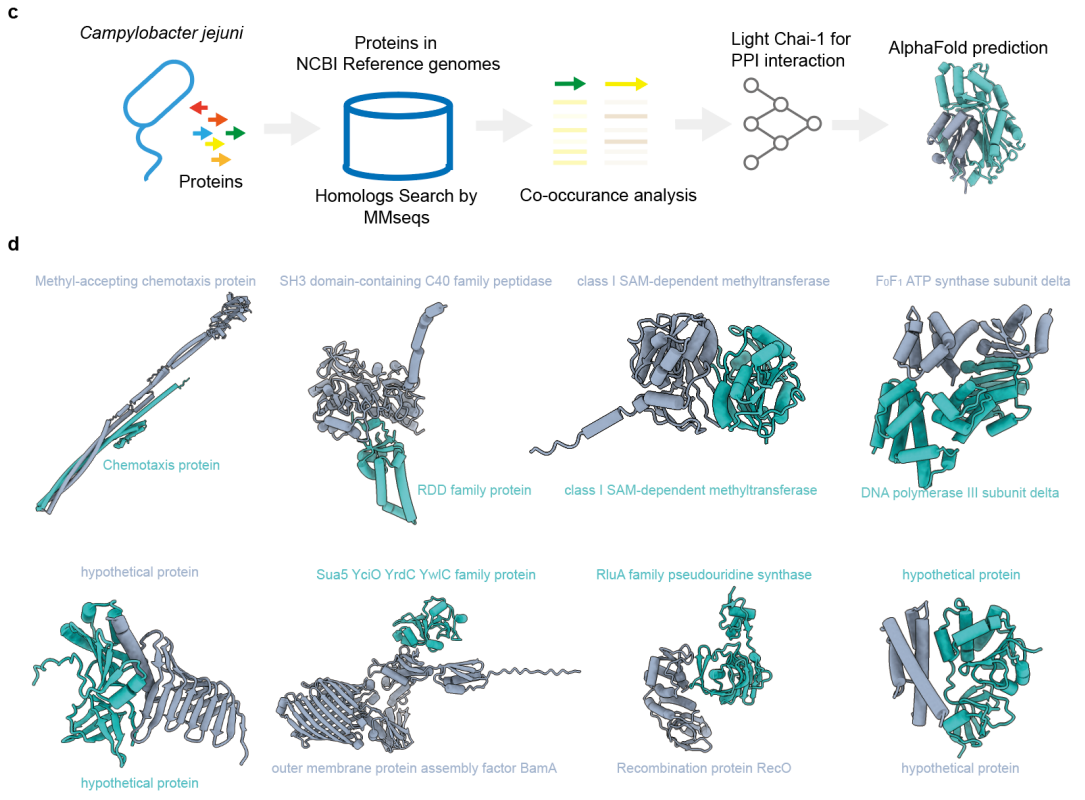
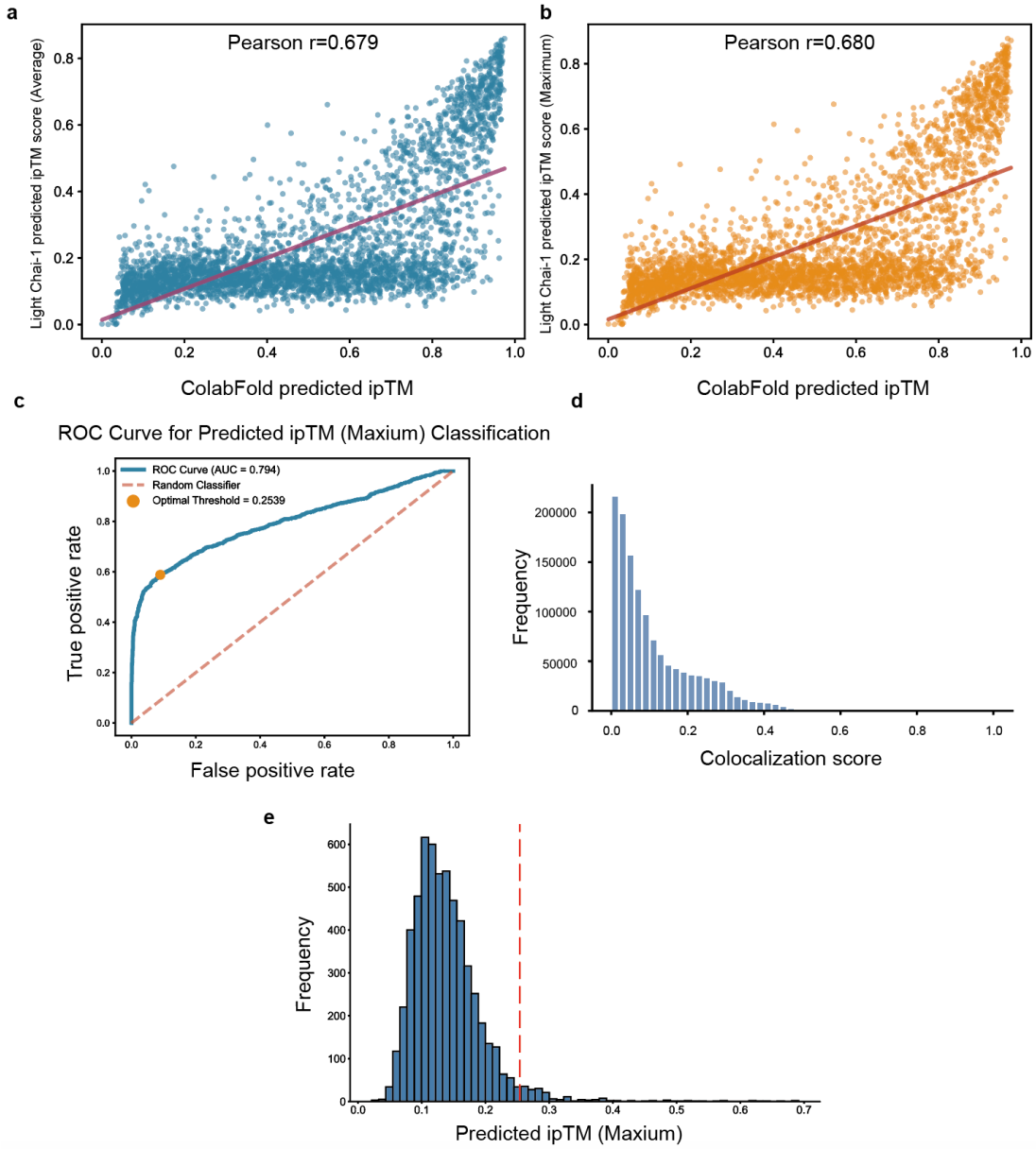


Figure 6. Interactive protein complexes located seperatively on genome of *Campylobacter jejuni*. c The pipeline to find protein pairs separated located in genome. The complete *C. jejuni* proteome was used as a query set to search 21 676 reference prokaryotic genomes (~8.6×10^7 proteins) with MMseqs2 (30–90 % pairwise identity). Genomic co-occurrence scores were then computed for all protein pairs. **d** Eight cases of protein complexes.



Supplementary Figure 24. Combining genomic co-occurrence and light Chai-1 for identifying protein pairs that are located far apart in the genomic sequence in *Campylobacter jejuni*. a and b Correlation between Light Chai-1 and ColabFold ipTM scores. Scatterplots show Pearson correlations using the mean (b; r = 0.679) or the maximum (c; r = 0.680) Light Chai-1 ipTM across five models; magenta lines indicate linear regressions. **c** Discriminatory power of Light Chai-1 ipTM. Receiver-operating-characteristic (ROC) curve, using ColabFold models with Best LIS≥0.203 and Best LIA≥3432 as true positives of predictions in b and c, yields an area under the curve (AUC) of 0.794. The optimal classification threshold (orange circle) is 0.2539. **d** Genome-wide distribution of co-occurrence scores for all *C. jejuni* protein pairs. **e** Distribution of maximum Light Chai-1 ipTM scores for the 5 780 filtered pairs; the red dashed line denotes the optimal threshold from panel c.

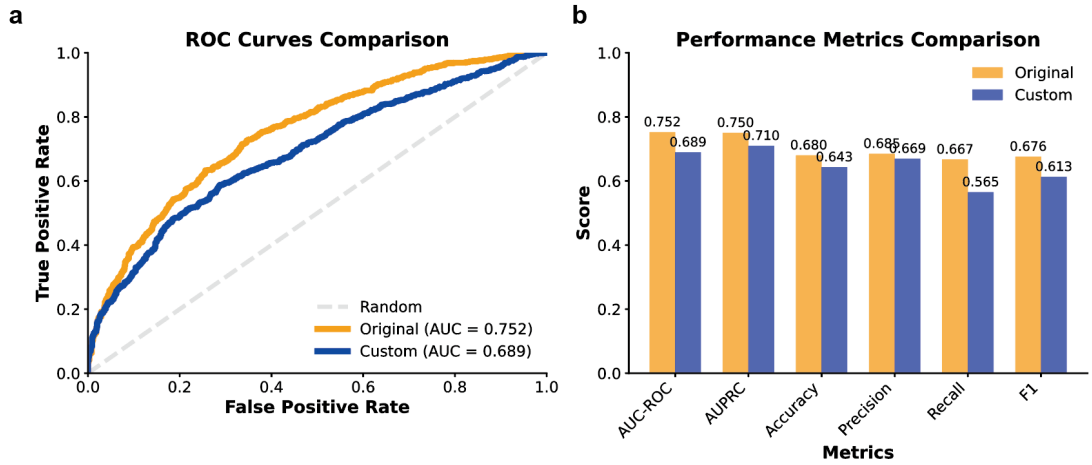
At first, we performed homology search, and each *C. jejuni* protein was queried against the database (30–90 % identity). Secondly, for co-occurrence analysis, protein pairs with a co-occurrence score ≥ 0.5 were retained (sequence length was also filtered to fit the prediction on RFX3090 GPU) for light Chai-1 score prediction (5780 pairs) (Supplementary Fig. 24).

Thirdly, after using Light Chai-1 prediction with the MSA-free model, a subset was predicted with ColabFold. Based on these small batch of prediction, we successfully identified 20 high-confidence complexes whose genes lie distantly on the genome (Figure 6d).

These preliminary results demonstrate that combining genomic context with rapid, MSA-independent structural scoring is a feasible strategy for uncovering long-range protein–protein interactions.

Additionally, we investigated whether a sequence-based PPI predictor could be used for high-throughput interaction screening. First, 1,000 protein pairs with a Best LIS ≥ 0.203 and Best LIA ≥ 3,432 were selected as a positive reference set, and 1,000 pairs that failed to meet either criterion were chosen as negatives. As shown in Response Figure 4, on this balanced benchmark the original MINT model reach ed an AUC-ROC of 0.65, indicating a modest level of discriminative power.

To enhance performance, we fine-tuned only the multi-layer perceptron (MLP) classifier module of MINT with 80,000 additional pairs drawn from our 1.1 million candidate interactions. The resulting model showed better discrimination on the test set with the AUC-ROC of 0.78; because the main embedding layers were left unchanged, these gains may not fully generalize to other highly divergent sequences.



Response Figure 4: Performance comparison between the original MINT model and the fine-tuned MINT model on the benchmark dataset selected from our predicted dataset. a Receiver-operating-characteristic (ROC) curves. The fine-tuned MINT model (blue) consistently dominates the original MINT model (orange) across the full range of false-positive rates, yielding a higher area under the ROC curve (AUC-ROC = 0.779 vs 0.655). The grey dashed line denotes random classification (AUC = 0.5). **b** Bar plot summarizing six classification metrics. Fine-tuning improves AUC-ROC, area under the precision–recall curve (AUPRC), overall accuracy and recall (0.797 vs 0.437), while slightly reducing precision (0.549 vs 0.831) and F1-score (0.573 vs 0.650), highlighting a trade-off between sensitivity and precision after model adaptation. Numeric values above the bars correspond to the exact scores.

We next ran the fine-tuned MINT model across all ~1.3 million possible protein-protein

pairs in *Campylobacter jejuni*. The model assigned a probability > 0.5 to 21,503 pairs, corresponding to 1.62 % of the total search space. To provide external validation, we generated independent structural predictions with AlphaFold 3 for the top 100 high-confidence pairs. Of these, 15 pairs were predicted as high-confidence interactions (AlphaFold3 ipTM>0.6), demonstrating that the sequence-based predictor can also recover interactions between proteins encoded far apart in the genome. We will further optimize the sequence-based PPI predictor work in the future work.

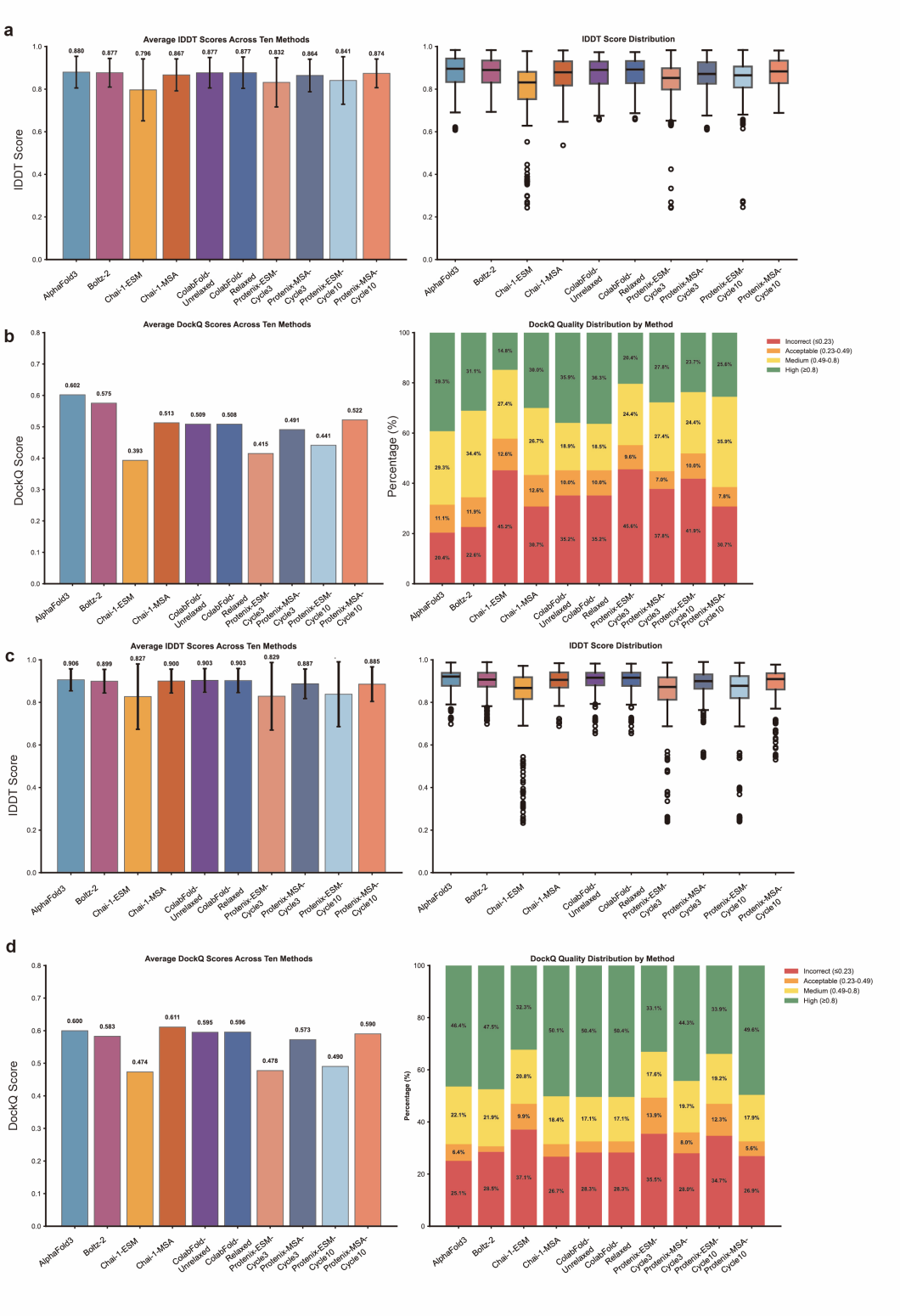
Altogether, we use two strategies to test whether we can predict protein pairs that are located far apart in the genomic sequence. Such protein-based PPI discrimination model would also apply for a wide range of organisms—including eukaryotes—and to the analysis of virus–host interaction networks.

2. Since the code of AlphaFold3 has been released, it would be interesting to compare its predictions with the model's performance in this paper, as well as with other deep-learning structural complex prediction tools.

Re: We thank the reviewer for encouraging a comparison with AlphaFold 3 and other contemporary complex-prediction frameworks. In the revised manuscript we therefore constructed an independent benchmark set of non-redundant protein homodimers and heterodimer for evaluation.

Briefly, heterodimeric and homodimeric complexes deposited in PDB, since 17 May 2023, were retrieved by Dockground server[7] filtered to a maximum resolution of 3.5 Å, an interface size of at least 12 contacting residues, and a buried surface area exceeding 800 Å^2. Redundancy was removed with Foldseek-multimer, yielding 54 heterodimers and 75 homodimers that serve as the definitive test set. Besides AlphaFold 3 and ColabFold-Multimer, the comparison now includes Protenix[8], Chai-1[9], and Boltz-2[10]. Because Protenix and Chai-1 accept either multiple-sequence alignments (MSAs) or single-sequence ESM embeddings, both input modes were assessed. Prediction quality was assessed using lDDT tool in Openstructure[11] to evaluate residue-level accuracy and DockQ[12] to measure overall chain–chain docking performance.

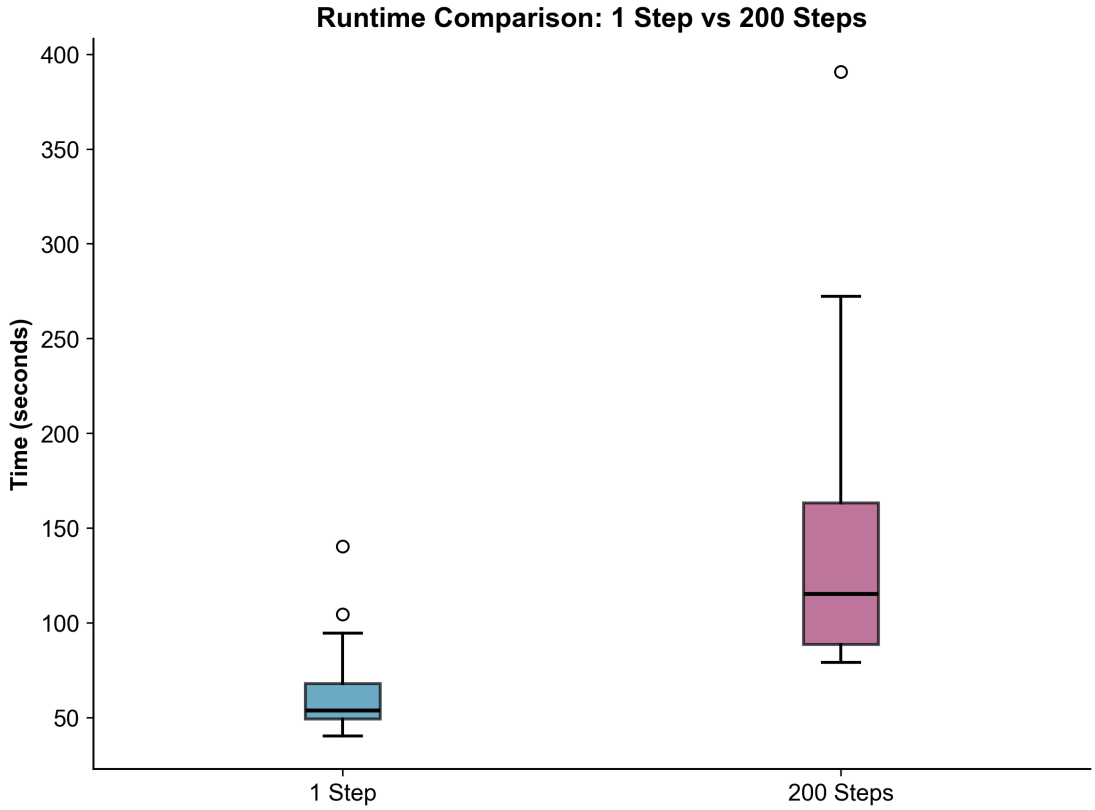
As shown in Supplementary Figure 21, the results underscore the value of evolutionary information: all three ESM-only tools performed poorly in both metrics, whereas the MSA-driven models achieved uniformly high lDDT scores. For homodimers, the MSA-based methods were broadly comparable. For heterodimers, AlphaFold 3 delivered the highest accuracy and Boltz-2 most closely approached that level, providing a clear hierarchy of current capabilities and a useful reference point for future studies.



Supplementary Figure 21. Comparison of the prediction accuracy for AlphaFold3 with ColabFold, Chai-1, Protenix, and Boltz-2 for both heterodimers and homodimers. a Local Distance Difference Test (lDDT) scores were calculated between the predicted and experimental structures for 54 non-redundant heterodimers. The bar chart (left) shows the mean ± s.e.m. for every method; The adjacent box-and-whisker plot visualizes the complete score distribution: the box spans the inter-quartile range (Q1–Q3), the horizontal black line marks the median, and open circles denote outliers. **b** DockQ scores for the same heterodimer set. Left, mean ± s.e.m.; right, stacked bars indicate

the fraction of predictions classified as High (DockQ ≥ 0.80, green), Medium (0.49 ≤ DockQ < 0.80, yellow), Acceptable (0.23 < DockQ < 0.49, orange) or Incorrect (DockQ ≤0.23, red). **c and d** Equivalent analyses for 75 non-redundant homodimers.

Finally, inspired by the Chai-1 architecture, we try to reduce ts diffusion steps from 200 to 1 in the ESM-only setting (as shown in Supplementary Fig. 23), transforming it into a rapid interaction-scoring tool that is 2.2-fold faster and well suited to large-scale screens of proteome-wide interactions or de-novo designs. All new analysis, together with the updated applications, have been incorporated into the revised manuscript.



Supplementary Figure 23. Comparison of the runtime for Chai-1 by using ESM with 1 or 200 steps in diffusion module.

3. In the evaluation of Multimer-Surface and the finetuned ProteinMPNN, the authors utilized the predicted complexes. It is better to evaluate the model performance using an independent dataset composed of experimentally determined structures.

Re: We appreciate the reviewer's suggestions. To provide a comprehensive assessment of our model, we constructed two evaluation sets: one based on independent, experimentally resolved protein structures and another derived from rigorously deduplicated, and predicted structure data.

For the protein binding-site prediction task, we first collected 71,667 high-quality

complexes, filtering by Best LIS ≥ 0.4 and pLDDT ≥ 80. To guarantee diversity and remove redundancy, we applied structural clustering with FoldSeek, eliminating any pair of protein monomers whose TM-score was ≥ 0.5. This procedure yielded a final, non-redundant training set of 9,609 dimeric complexes. Using this set, we further trained the MPBind model (https://github.com/jianlin-cheng/MPBind) for 45 epochs on an NVIDIA RTX 3090 GPU.

To prevent data leakage and enable a rigorous assessment, we built an independent validation set from predicted structures. We randomly selected partial of the remaining complexes as validation candidates. Subsequently, we conducted meticulous deduplication on these samples based on interface similarity and monomer similarity.

In the interface deduplication phase, we employed a stringent strategy that integrates structural alignment with interface site information which have been used in PINDER dataset before[13]. This strategy aims to identify and eliminate proteins that, despite differing overall structures, exhibit highly similar key functional interfaces. We first utilized FoldSeek to compute structural alignment information between all protein single chains in the candidate samples (command: foldseek easy-search query.pdb targetdb out.m8 tmp --format-output "query,target,evalue,bits,qstart,qend,tstart,tend" -s 9.5). For any pair of proteins (denoted as A and B), we examined the extent of overlap between their structural alignment regions and the pre-defined interface residues. Surface residues in these complexes were defined as those with a relative solvent accessibility greater than 5% that lost more than 1 Å² of absolute solvent accessibility upon protein-protein complex formation. If the overlap between the alignment region of protein A and the entire interface region of protein B exceeds 50%, or if the overlap between the alignment region of protein B and the entire interface region of protein A exceeds 50%, then the pair of proteins is deemed to exhibit interface similarity. Any complex containing protein chains determined to have interface similarity will be removed from the validation set candidate pool. Subsequently, in the monomer deduplication phase, we aimed to thoroughly exclude any single chains in the validation set complexes that display structural similarity with those in the training set. We employed FoldSeek to perform structural clustering of all monomers in the dataset, using the command: foldseek easy-cluster pdb res tmp -c 0.9.

Following this series of rigorous dual deduplication steps, we ultimately obtained a validation set comprising 205 monomers, thereby maximizing the reliability of the evaluation.

We further assessed MPBind on Pro_Test_315, an independent benchmark dataset[14]. As Figure 6b illustrates, the fine-tuned model (MPBind-PHC) achieved a marked performance boost on the duplicated Predicted test set and delivered higher AUC scores on Pro_Test_315. In addition, after we applied the same filtering protocol to eliminate potential leakage from Pro_Test_315—removing any cases that overlapped with the continued-training data—the model still showed a comparable improvement, confirming the robustness of the gain (Figure 6a and b).

For ProteinMPNN model, we also constructed a PDB-derived testing dataset and predicted structure-derived dataset for evaluation. In details, from the initial pool of high-confidence complexes, we retained 71,667 complexes that satisfied pLDDT ≥ 80 and Best LIS ≥ 0.4, and we supplemented these with 716 experimentally determined complexes from the PDB, yielding a total of 72,383 assemblies. Such PDB protein complexes were extracted by Dockground server[7] after Jan 1th 2023 and filtered as described above.

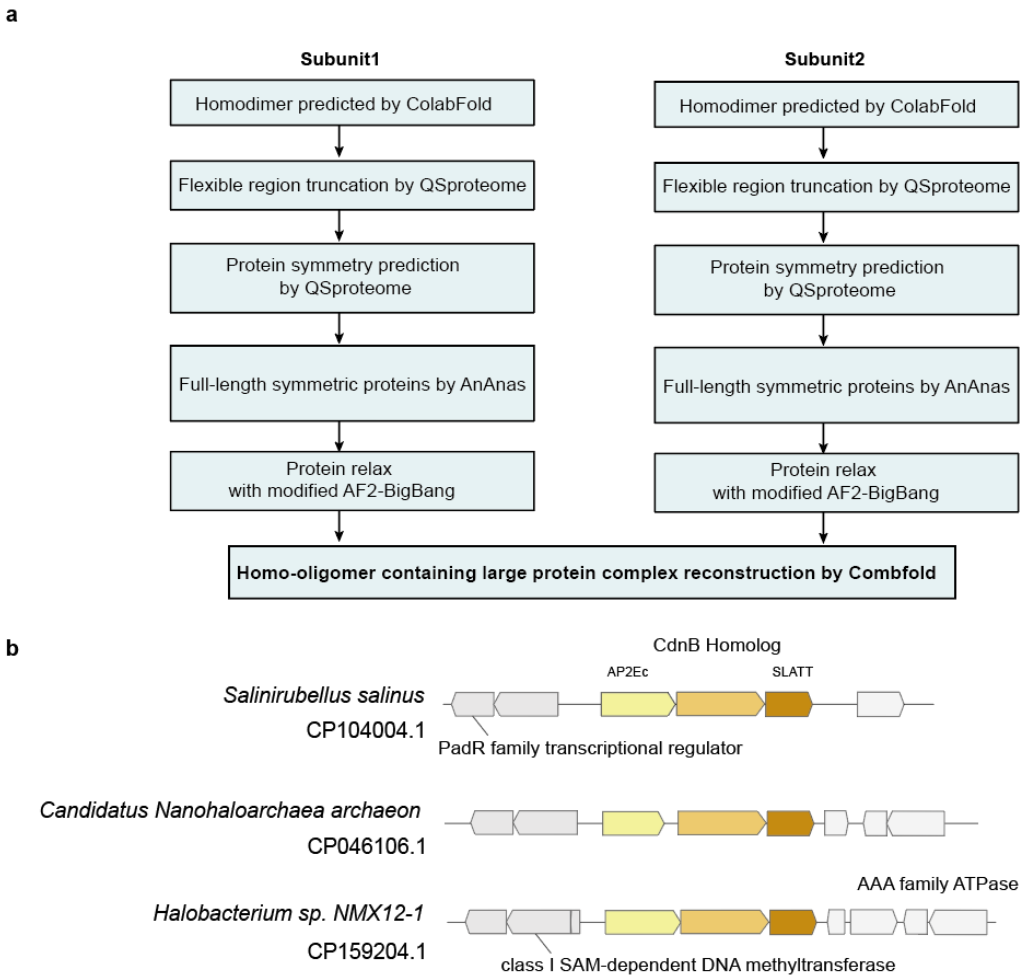
To remove redundancy, we first performed protein complex structural clustering with FoldSeek (command: foldseek easy-multimercluster pdb/ clu tmp --multimer-tm-threshold 0.65 --chain-tm-threshold 0.5 --interface-lddt-threshold 0.65). On the basis of these clusters, we further devised a stringent partitioning scheme that yields three mutually exclusive subsets. First, every experimentally determined PDB complex was assigned to the experimental structure test set, and any cluster sharing a member with PDB dataset was removed. Second, after excluding all PDB-derived clusters, we randomly sampled the remaining clusters to create the predicted structure-derived testing set. The unassigned clusters constituted the training set. To further guarantee that the validation and test sets are completely independent of the training set, we imposed an additional, stringent monomer-level deduplication by FoldSeek (command: foldseek easy-cluster pdb res tmp -c 0.9).

Following this rigorous, two-step deduplication and partitioning workflow, the final splits comprised 23, 459 complexes for training, 582 for testing set from predicted structures, and 137 from PDB. After fine-tuning ProteinMPNN, sequence recovery improved and perplexity fell on the predicted-structure test set, whereas performance on experimentally determined structures declined (Supplementary Fig. 22).

In summary, we evaluated two deep-learning models on both independent datasets derived from PDB or duplicated predicted structures. For binding-site prediction, our fine-tuning strategies improved performance on both predicted and experimentally determined structures. In the inverse protein-folding task, however, the model performed best on predicted backbone structures, underscoring that predicted structures cannot always serve as substitutes for experimentally resolved data.

4. Extended Figure 3: the workflow in panel a is a little confusing, especially the distinction between the shaded and unshaded parts.

Re: Thanks for reviewer's suggestions. We have corrected the workflow by adding the annotation and add the figure legend description to make it clear.



Supplementary Figure 13. Advanced Reconstruction of Protein Complexes and Gene Locus Analysis in the Genome. a We utilized QSproteome for symmetry prediction and subsequently relaxed the protein structures using AlphaFold2-bigbang. Symmetrical homo-oligomers of each subunit were assembled with CombFold. **b** AP2Ec domain-containing gene, CdnB homolog, and SLATT-domain containing gene arranged as an operon in multiple prokaryotic genomes.

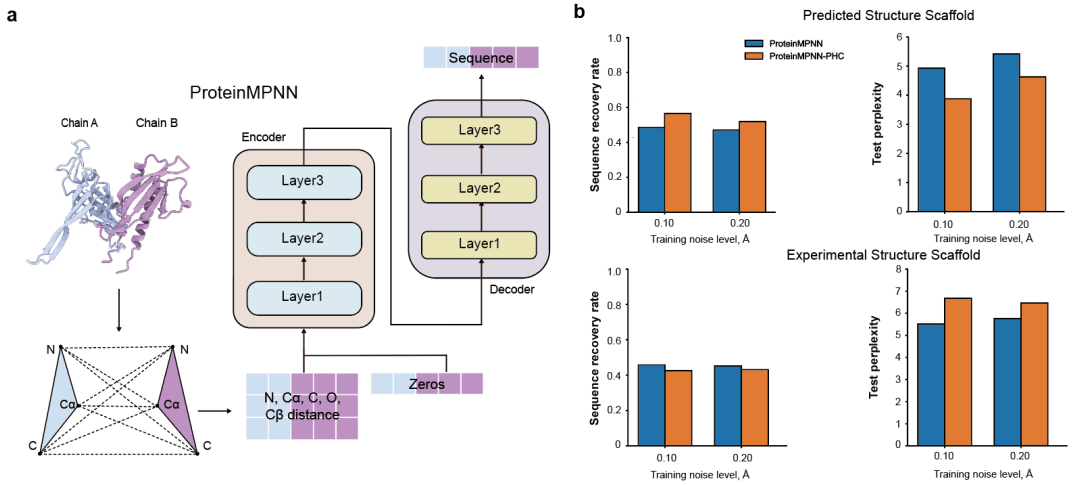
5. Figure 6e: The definition of accuracy and perplexity in this context is somewhat confusing. For example, what does the “percentage of correct amino acids recovered” refer to?

Re: Thank you for highlighting this point. We have revised the terminology to improve precision and consistency with the literature. To enhance clarity, we have incorporated established technical terms from related studies. Specifically, we now report as sequence recovery rate, defined as the fraction of target residues that are correctly reproduced, which directly reflects each model’s fidelity in designing sequences compatible with the reference structure. In addition, test perplexity is defined as the exponentiated categorical cross-entropy averaged over all residues in the sequence, providing a quantitative measure of

the predictive uncertainty. the revised paragraph is presented below.

"…Subsequently, the fine-tuned and baseline models were evaluated on test sets derived from either predicted or experimentally determined structures. Two canonical metrics were employed: sequence recovery rate, defined as the proportion of target residues accurately reproduced, which reflects the model's fidelity in designing sequences compatible with the target structure; and test perplexity, computed as the exponentiated categorical cross-entropy averaged across all residues, which quantifies the predictive uncertainty of the model…".

Additionally, we also revised the part in Supplementary Fig. 22 as below:



Supplementary Figure 22. Applications of ProteinMPNN models leveraging the large-scale protein complex structure dataset a Model architecture of ProteinMPNN, consisting of a 3-layer backbone encoder and a 3-layer sequence decoder. Distances between N, Cα, C, O, and virtual Cβ are encoded by the encoder. The encoded features are used together with partial sequences to iteratively generate amino acids in random decoding order through the decoder. **b** Performance of pretrained ProteinMPNN and finetuned ProteinMPNN-PHC on high-quality predicted heterologous complexes or Experimental Structural Scaffold.

Reviewer #2 (Remarks on code availability):

The authors provided clear instructions for installing and running the code. I did not test the code due to its significant storage requirements.

Re: Thanks for reviewer's comments. We further added the necessary scripts to the GitHub repository (https://github.com/wensm77/Protein-Complex-Atlas), along with clear labels to indicate where each script should be used. Additionally, we have included detailed documentation to guide users in reproducing the analysis. We hope these updates will facilitate a smoother and more transparent reproducibility process.

Reviewer #3 (Remarks to the Author):

Qi et al. present a large-scale resource of protein-protein interaction (PPI) predictions across multiple kingdoms. Using ColabFold, they predicted 1.15 million PPIs, identifying 170,001 high-confidence interactions. While most large-scale PPI studies focus on human proteins, this work expands the scope to other kingdoms, addressing an essential gap in the field.

The study explores multiple fronts, including evolutionary analysis of domain fusion, detection of viral-human PPIs, and the use of predicted interactions for training machine learning classifiers—demonstrating broad applicability. However, the breadth of the study appears to come at the expense of depth in each specific area. My main concerns with this work are around comparison to previously published work and homology leakage of the classifiers.

Re: We thank Reviewer #3 for the careful evaluation and constructive summary of our work. We agree that extending large-scale PPI prediction beyond humans addresses an important gap, and we appreciate the acknowledgement of the resource's breadth (multi-kingdom coverage, evolutionary analyses, viral–host PPIs, and downstream DL utility). We have added quantitative cross-dataset comparisons (Supp. Fig. 19), multi-level novelty/coverage analyses with improved phylogenetic visualization (Supp. Figs. 7-10), redesigned unclear figures (Figure 3), enforced stringent complex/monomer/interface deduplication for model training and evaluation (updated Fig. 6 and Suppl. Fig. 22), released the full ~1.1 M complex set with portal, completed & licensed the GitHub codebase (plus documentation), provided runtime scaling benchmarks (Supp. Fig. 3), and updated viral structure references, thereby strengthening novelty, rigor, and reproducibility.

Major:

- The paper does not include a comprehensive comparison to other core resources, despite many studies having already produced large-scale PPI prediction datasets. While some of this work is cited, it is not compared against it. For example, the Predictomes study by Schmid et al. (2024) and Zhang et al. (2024), Computing the Human Interactome, as well as the virulence factor study by Humphreys et al., Protein interactions in human pathogens revealed through deep learning, are highly relevant. Please include a comparison to quantify the novelty of predictions.

Re: Thank you for the reviewer's valuable suggestions. We agree that comparing multiple datasets will strengthen the study and further provides new insight. Accordingly, as shown in Supplementary Fig. 19, we have incorporated five additional datasets whose sizes range from 154,276 to 111 complexes. For each set, we calculated the number of complexes that overlap with our own and displayed these values above the corresponding bars. In

absolute terms, our collection is 7.4-fold larger than the next-largest resource. In breadth, it spans not only the human proteome but also microorganisms, plants, mouse, and human–virus pairs, providing a level of taxonomic coverage lacking in previous studies. Focusing on high-confidence models, we report 181,671 high confidence complexes—37,855 in the human interactome and 108,879 across 224 prokaryotic species. In particular, our operon-guided and genetic co-localization strategy in prokaryotes uncovered a wealth of previously uncharacterized assemblies, substantially expanding the landscape of protein complexes for functional and evolutionary analyses.

Especially, we agree that the study by Humphreys et al. is highly relevant to our work. While their analysis applied Light RosettaFold (RoseTTAFold2-Lite) to score 78 million protein pairs from 19 human bacterial pathogens, it yielded only 3,199 structural models. As shown in Response Figure 5, although they computed interaction scores for every pairwise combination, most of the predicted interactions involve proteins encoded in close genomic proximity. In contrast, we modelled 164,037 candidate interactions across 36 pathogenic bacteria by integrating operon context and co-localization mining. We then broadened the survey to 188 representative species, extracted 313,348 neighboring CDS pairs, and used ColabFold to generate their assemblies. Of these, 69,884 complexes (≈14 %) surpassed the high-confidence threshold—21.8-fold more than those reported by Humphreys et al. Moreover, our predictions cover not only pathogens but also diverse prokaryotes, plants, mouse, human, and human–virus pairs, providing a uniquely broad interactome. This comprehensive dataset enables integrative analyses ranging from evolutionary conservation (Figure 3h) and protein fusion/fission events (Figure 5) to host–virus interface mapping (Figure 4f). In addition, we have developed an MSA-free, time-saving interaction-scoring method built on the AlphaFold framework. Figure 6 illustrates how applying the Chai-1 diffusion-step-shrink strategy enabled us to predict high-confidence protein complexes in *Campylobacter jejuni.*

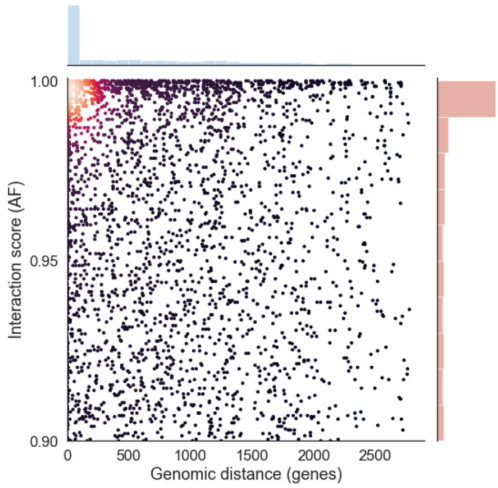
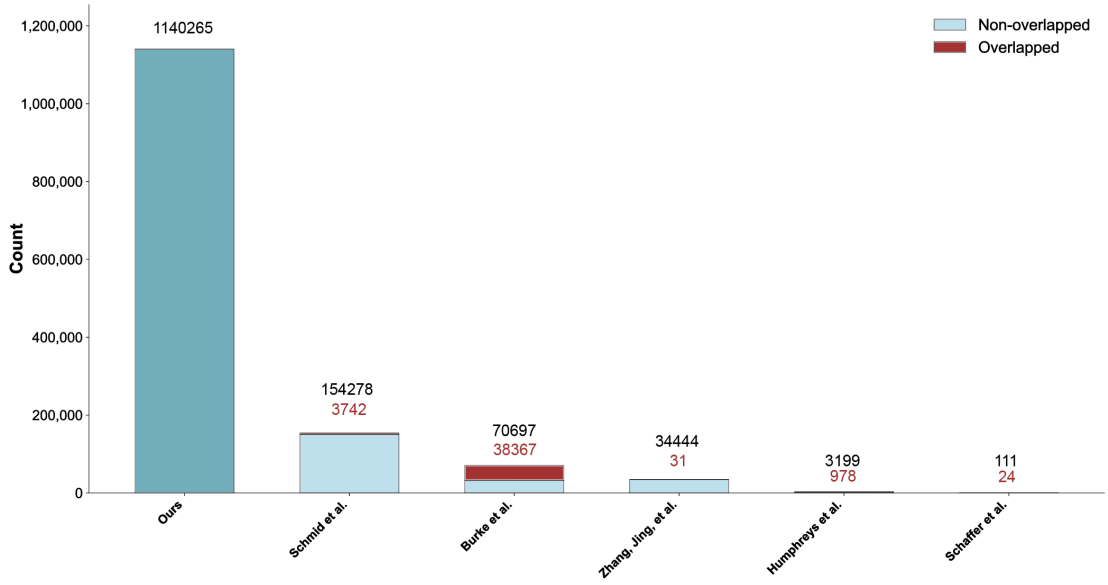


Figure S8: Predicted interactions by genomic distance

Response Figure 5. Predicted interactions are mainly located in close genomic distance. The figure was from reference 6 (Humphreys et al.).

Together, our dataset—spanning proteome-wide interactions in bacteria, archaea, human, mouse, plants, and human–virus pairs—greatly increases the coverage of predicted protein complexes. Especially, our locus-related filtering approach in prokaryote species enriched a large number of protein complexes, which would provide a more comprehensive reference dataset for various analyses.



Supplementary Figure 19. Comparison of the number of predicted large-scale protein complexes of multiple studies[6,15-18]. The total number of predicted complexes (black) and the number overlapping with our study (red) are shown above each bar.

- Figure 1g: The tree visualization is not very informative, as the labels are difficult to read. Additionally, it is unclear how much novelty has been generated per species. A quantitative assessment of novelty at different taxonomic levels (species, phylum, and kingdom) would be critical to understanding the contribution of this resource.

Re: We thank the reviewers for their helpful comments on the tree visualization. Fig. 1g now offers a concise overview of the evolutionary relationships among pathogenic bacteria, archaea, and other representative bacterial taxa. For additional detail, Supplementary Fig.7 provides, for each representative species, its phylogenetic tree, the total number of predicted proteins on the right side, and the corresponding accession numbers, except for the pathogenic bacteria which have been supplied in the supplementary Table 1.

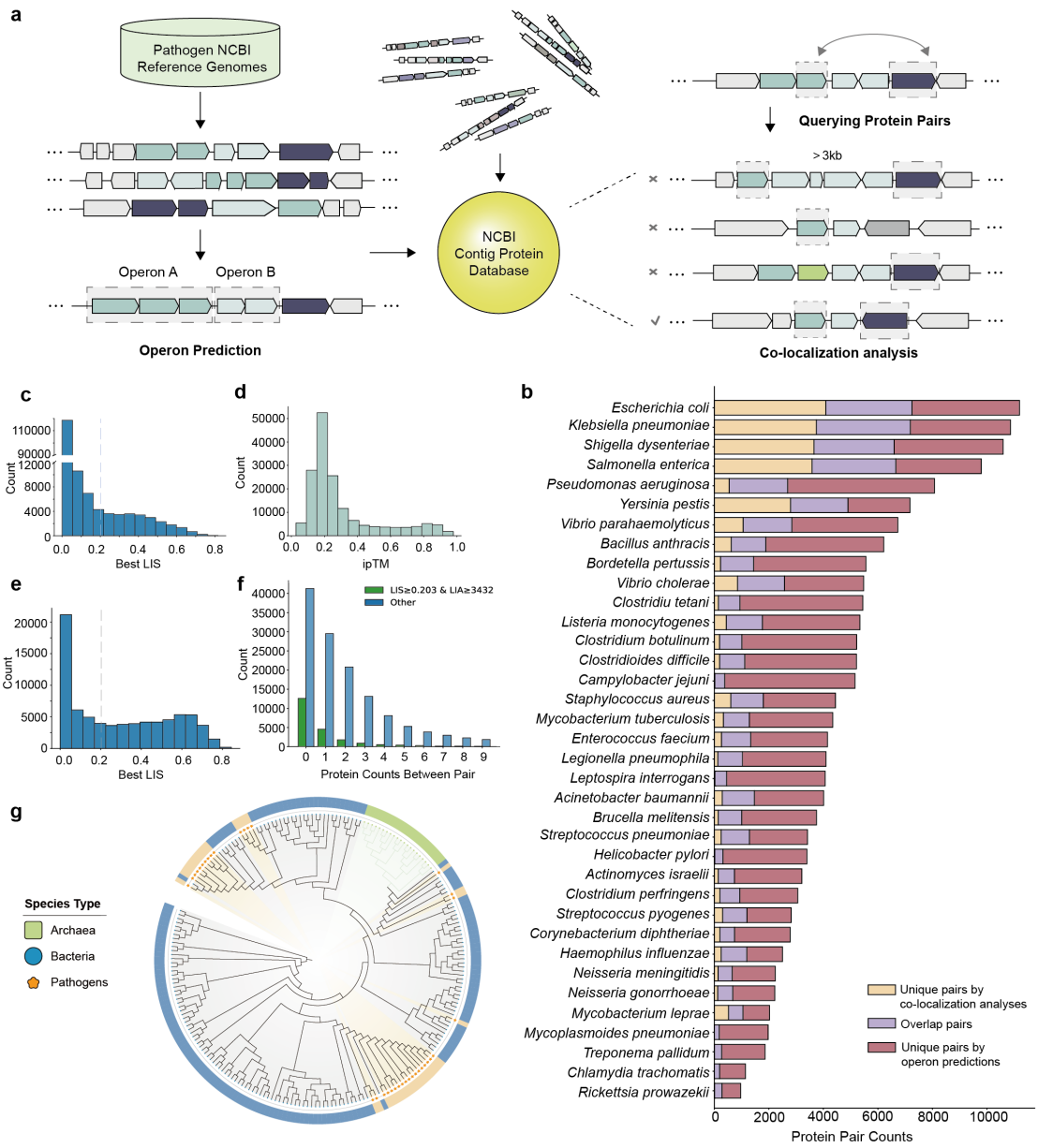


Figure 1. Identification of loci-linked protein complexes in bacteria and archaea. a A schematic representation of loci-linked protein pair identification through operon prediction and co-localization analysis. Operons were predicted from NCBI reference genomes, and homologs of potential protein pairs were queried against our collected NCBI Contig Protein Database. Co-localization analysis was used to identify protein pairs separated by less than 3 kb on the genome, with high co-localization score. **b** Distribution of protein pairs identified through the operon prediction and co-localization analysis across bacterial species. Yellow bars represent protein pairs identified exclusively through co-localization, dark purple bars indicate pairs located within operons but not identified by co-localization, and light purple bars represent overlapping pairs found by both methods. **c and d** Distribution of the Best Local Interaction Score (Best LIS) scores (c) and the interface predicted template modelling ipTM (ipTM) scores (d) for ColabFold multimer-predicted protein pairs across 36 pathogenic bacteria. **e** Distribution of Best LIS scores for homodimers of subunits from predicted protein pairs. **c and e** the dash lines denote Best LIS=0.203 as the threshold **f** Distribution of coding sequence (CDS) counts between high-confidence protein pairs. Blue represents pairs showed the CDS counts between all the predicted pairs across 36 pathogenic bacteria, while green represents

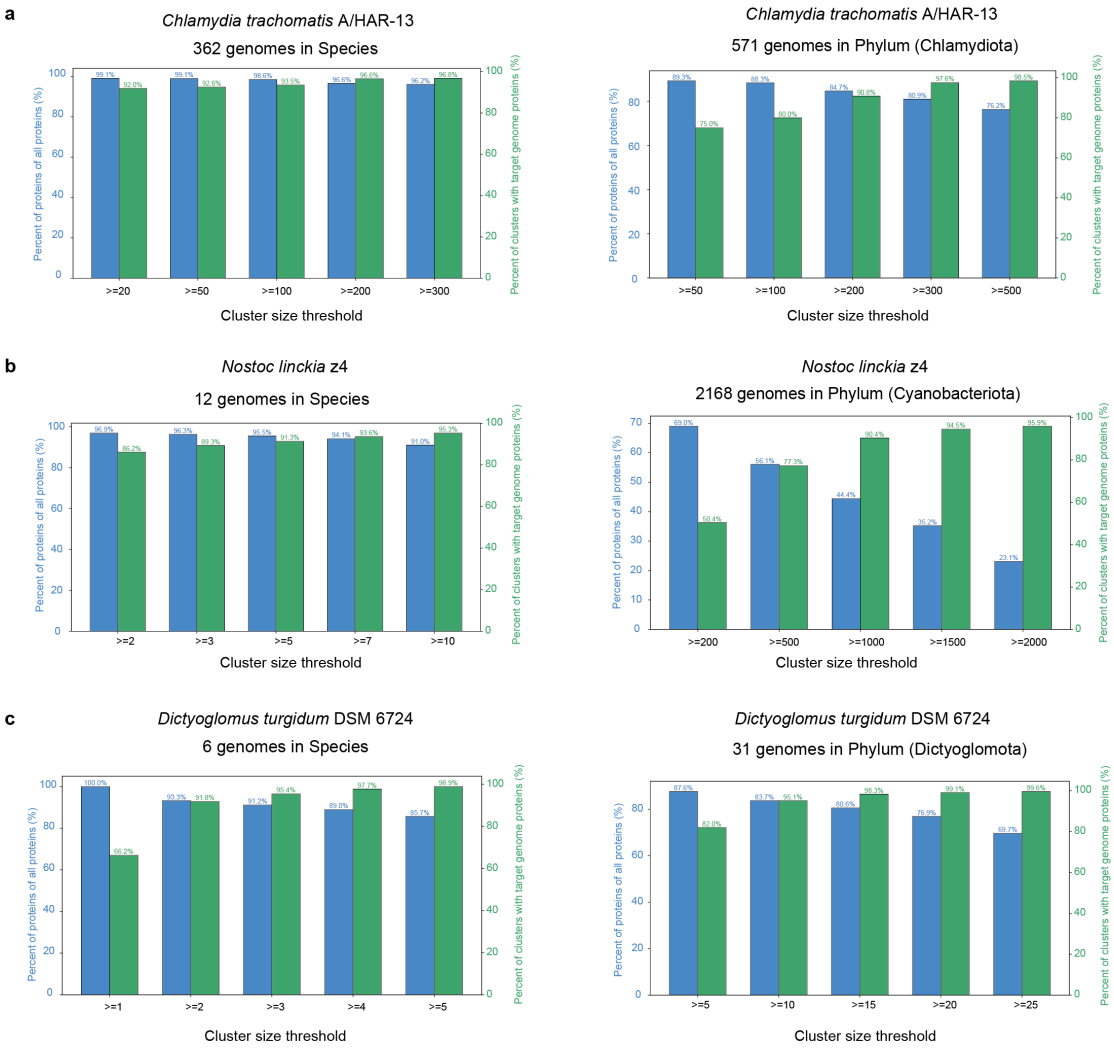
those identified as the high confidence pairs. **g** A phylogenetic tree of representative genomes from 188 phyla, and the 36 pathogenic bacteria used in this study. Pathogens are marked with orange stars, while archaea are represented in green.

Supplementary Figure 7. **Phylogenetic tree of study genomes.** Phylogenetic tree of representative genomes from 188 phyla together with the 36 pathogenic bacterial genomes (also shown in Fig. 1g). Branch coloring denotes domain affiliation (green, Archaea; blue, Bacteria); pathogenic taxa are flagged with orange stars. The protein number was shown on the right side.

Secondly, to quantitatively assess the novelty at different taxonomic levels, we asked what fraction of the entire proteome at different taxonomic levels is covered for by our selected proteins and their homologues, after choosing a representative genome for each phylum. Because many bacterial species have not yet been assigned a formal kingdom classification, we selected the species, phylum, and domain (super-kingdom) levels for analysis.

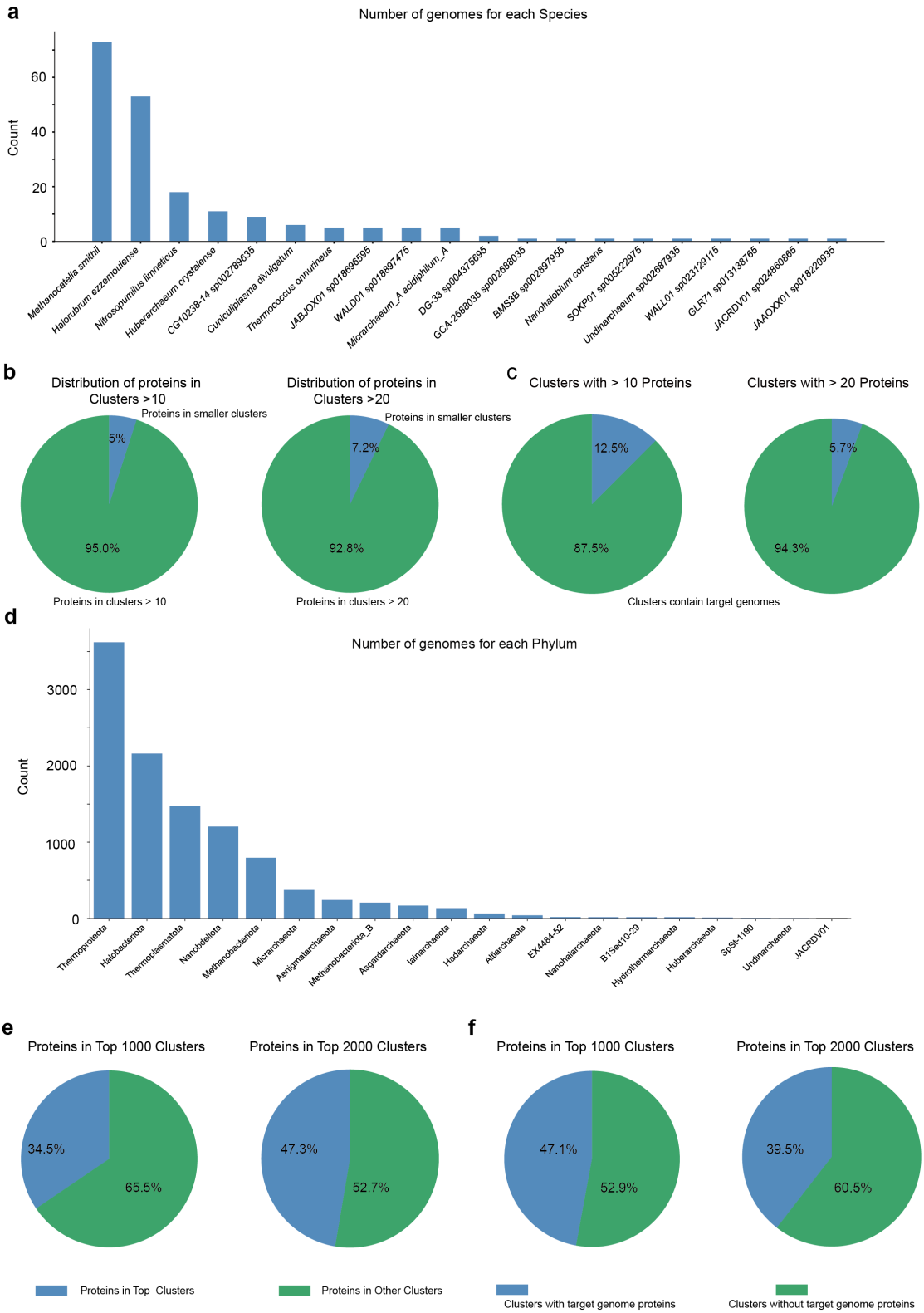
For species level, we collected genomes belong to this species from NCBI and further harvested all the proteins for performing protein clustering using MMseqs2 as threshold of 0.3. We chose three species from Bacteria domain and one from Archea domain. As expected, at the species level—whether in presentative species of Archaea or Bacteria—the vast majority of protein clusters, aside from a few small ones, contained proteins from the target genome and collectively covered more than 90 % of all proteins found in the sampled genomes of that species (Supplementary Fig. 8 and 9). When the analysis was expanded to the phylum level, coverage naturally declined but remained substantial. For instance, with a cluster threshold of ≥200 members, 84.7 % of all Chlamydiota proteins were included and >90 % of clusters still contained the C. trachomatis proteins. Similarly, across 2,168 Cyanobacteriota genomes, the ≥1,000-protein threshold retained 44.4 % of total proteins yet more than 90 % of these clusters that contained the N. linckia proteins (Supplementary Fig. 8 and 9). These results indicate that many high-abundance, conserved proteins are shared well beyond the species boundary. At the domain scale the same trend was evident (Supplementary Fig. 10).

Together, these data suggest that selecting a handful of representative species from each phylum is sufficient to recover most evolutionarily conserved proteins, whereas additional, more fine-grained predictions will be required to resolve proteins, for example, strain-specific gene products, that reside in the long tail of small clusters.



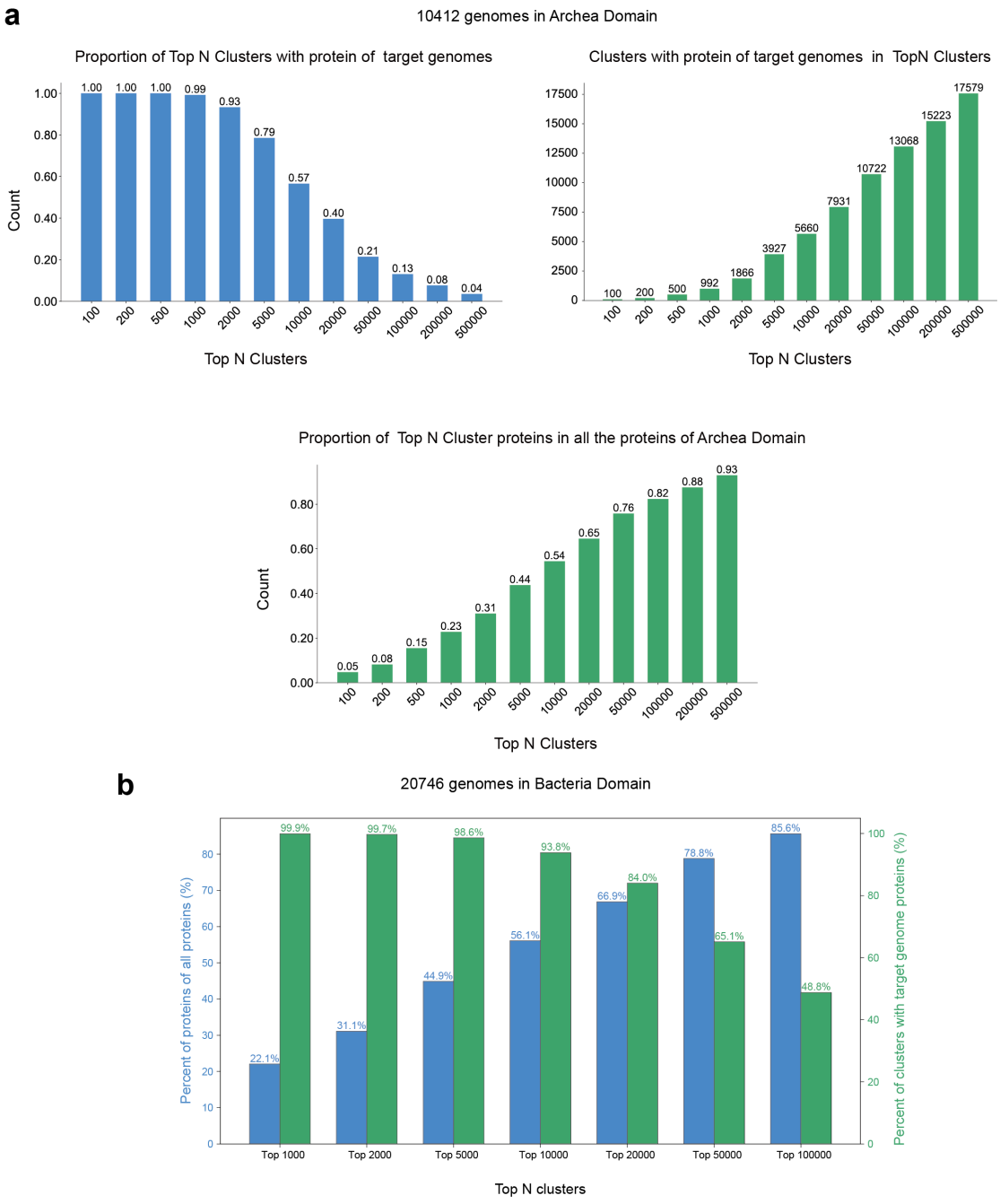
Supplementary Figure 6. Protein-cluster coverage for three representative bacterial genomes.

For each species (left sub-panel) and its corresponding phylum (right sub-panel), bars show (blue, left axis) the percentage of all proteins captured by clusters of at least the indicated size and (green, right axis) the percentage of those clusters that contain a protein from the representative genome. Results are given for *Chlamydia trachomatis* (a), *Nostoc linckia* (b) and *Dictyoglomus turgidum* (c). Larger cut-offs are needed at the phylum level to achieve coverage comparable to that obtained within a species.



Supplementary Figure 7. Sequence-cluster coverage across taxonomic scales of Archeal domain. a Number of publicly available genomes for each species included in this study (top-ranked species on the x-axis). **b** Species-level clustering of *Methanocatella smithii*: pie charts show the share of proteins that fall into clusters larger than 10 or 20 sequences (green) versus smaller clusters (blue). **c** Corresponding fraction of sequence clusters (size > 10 or > 20) that contain at least one protein from the target selected genome. **d** Genome counts per phylum for all archaeal genomes analysed in this study. **e** Phylum-level clustering of thermoproteota: pie charts show the share of

proteins that fall into top 1000 or top 2000 clusters (blue) versus smaller clusters (green). **f** Corresponding fraction of sequence clusters (Top 1000 or 2000) that contain at least one protein from the target selected genome.



Supplementary Figure 8. Coverage of representative genome proteins at the domain level.

a Archaea (10 412 genomes). Top-left: fraction of the top N clusters that contain at least one protein from the representative genomes (blue).Top-right: the number of clusters containing at least one protein from the representative genomes (orange). Bottom: proportion of all archaeal proteins accounted for by the proteins in the top N clusters (green). Values are shown for increasing N (100 to 50 000). **b** Bacteria (20 746 genomes). Blue bars indicate the percentage of all bacterial proteins captured, while red bars give the fraction of clusters that include proteins from the representative genomes. The data show that relatively small subsets of large clusters already cover a substantial share of each domain's proteome, although the fraction of clusters containing target-genome proteins

declines as N increases.

- Figure 3a and 3j: These figures are difficult to interpret. I recommend reconsidering the analysis and exploring alternative visualization. A more intuitive approach, e.g. for j maybe some Hierarchical clustering, could help convey the key insights more easily.

Re: We thank the reviewer for this suggestion. We agree that an alternative visualization can convey the relationships more clearly.

For the original Figure 3a, we have replaced the network view with a hierarchical clustering heat-map in Figure 3c. Because the full set contains a large number of genes, we selected the 1,000 most highly connected genes. In the heat-map, blue cells denote high-confidence predicted interactions. The clustering reveals that, unlike in prokaryotes—where interacting genes tend to be genomically adjacent—human interactions are more widely dispersed, although a few small, locally connected modules can still be observed.

To further mine cross-kingdom protein-complex relationships, we applied Foldseek-Multimer clustering to complexes from human, Arabidopsis, mouse and representative prokaryotes. As shown in the new Figure 3h, the Sankey-style diagram summarizes the flow of complexes among clusters. We observed many shared structural classes between human and mouse, and 209 clusters that contain both prokaryotic and eukaryotic complexes. We hope these new visualizations help readers grasp the data structure and extract additional biological insights.

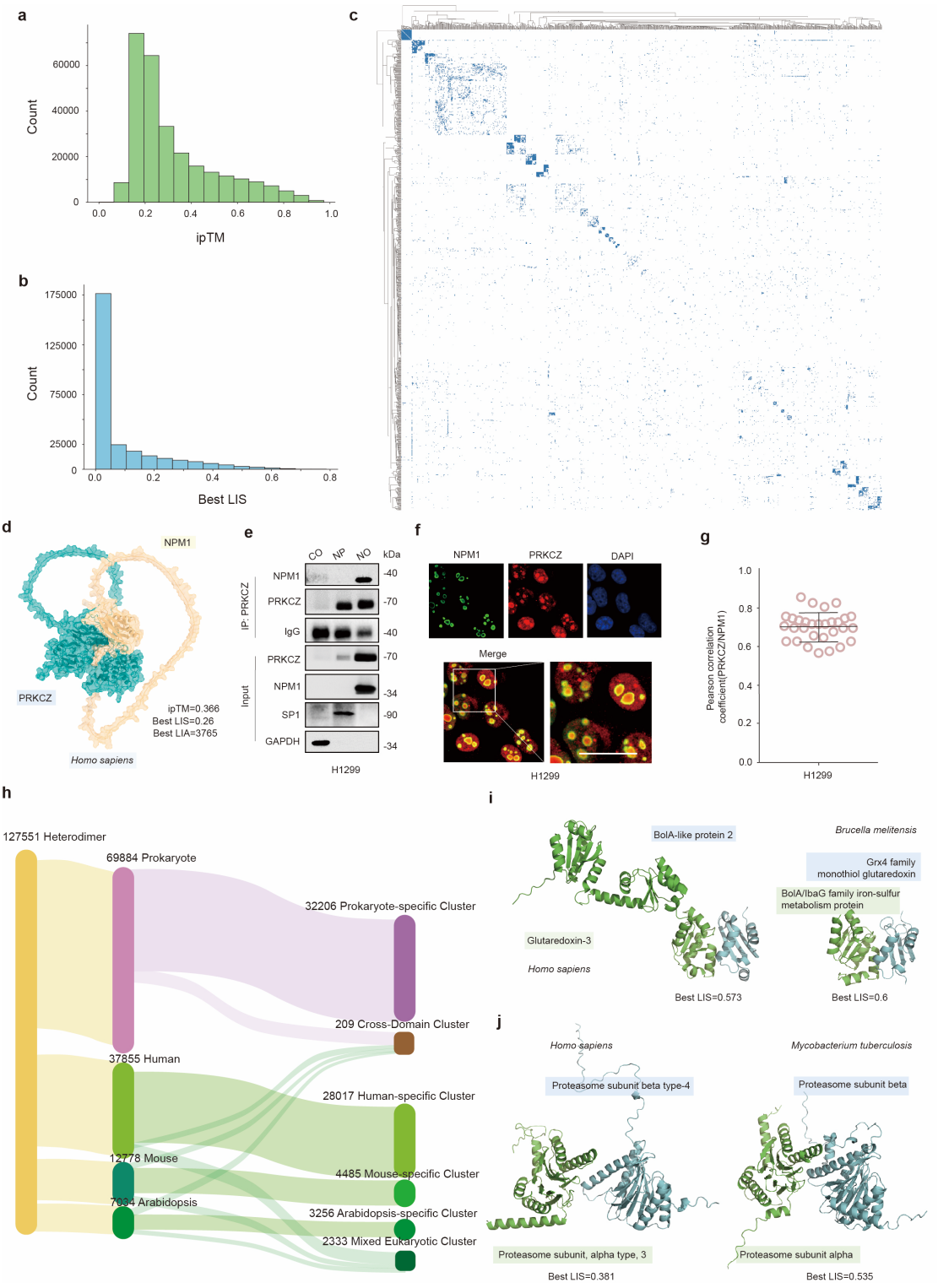


Figure 3. The Atlas of Human Protein Complex Structures. a and b Distribution of ipTM scores (a) or Best LIS score (b) for protein pairs selected from human. **c** Hierarchical clustering heat-map of the 1,000 most highly connected human high-confidence protein–protein interaction nodes. Blue pixels mark observed interactions. **d** Predicted complex structures of human PRKCZ and NPM1. **e** Immunoprecipitation and Western blot analyses of cell lysates from purified cytoplasm (CO), nucleoplasm (NP), or nucleolus (NO) of H1299 cells. **f and g** Immunofluorescence staining assays to assess PRKCZ and NPM1 in H1299 cells, including co-localization

quantification using Pearson's correlation coefficient from 30 cells across three independent experiments, scale bar = 25μm. **h** Sankey diagram of heterodimeric PPIs across taxa and cluster types. **i** Structural comparison between *Brucella melitensis* BolA/IbaG family iron-sulfur metabolism protein and Grx4 family monothiol glutaredoxin with human glutaredoxin-3 and human bolA-like protein 2. **j** Structural comparison between human and mycobacterial proteasome α/β heterodimers.

- The application of fine-tuning ProteinMPNN is inconclusive. Structural similarity can be detected beyond 30% sequence identity, meaning sequence-based separation may not be sufficient. I recommend using Foldseek-Multimer or a PINDER-like approach to separate training and testing.

Re: Thank you for raising this important point. To rigorously prevent data leakage, we used Foldseek-Multimer for complex-level deduplication and Foldseek for monomer-level deduplication.

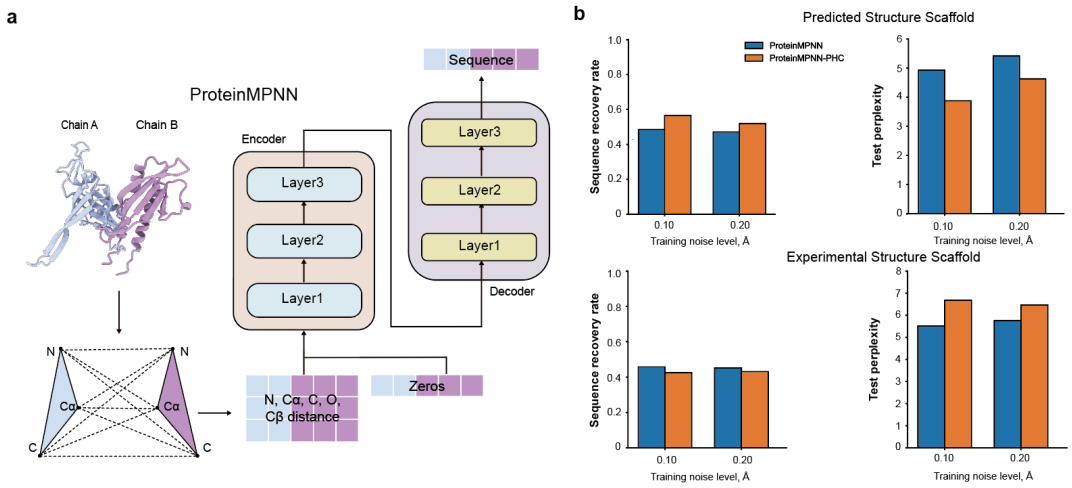
For ProteinMPNN model, we constructed a PDB-derived testing dataset and predicted structure-derived dataset for evaluation. In details, from the initial pool of high-confidence complexes, we retained 71,667 complexes that satisfied pLDDT ≥ 80 and Best LIS ≥ 0.4, and we supplemented these with 716 experimentally determined complexes from the PDB, yielding a total of 72,383 assemblies. Such PDB protein complexes were extracted by Dockground server[7] after Jan 1th 2023 and filtered to a maximum resolution of 3.5 Å, an interface size of at least 12 contacting residues, and a buried surface area exceeding 800 Å².

To remove redundancy, we first performed protein complex structural clustering with FoldSeek-Multimer (command: foldseek easy-multimercluster pdb/ clu tmp --multimer-tm-threshold 0.65 --chain-tm-threshold 0.5 --interface-lddt-threshold 0.65). On the basis of these clusters, we further devised a stringent partitioning scheme that yields three mutually exclusive subsets. First, every experimentally determined PDB complex was assigned to the experimental structure test set, and any cluster sharing a member with PDB dataset was removed. Second, after excluding all PDB-derived clusters, we randomly sampled the remaining clusters to create the predicted structure-derived testing set. The unassigned clusters constituted the training set. To further guarantee that the validation and test sets are completely independent of the training set, we imposed an additional, stringent monomer-level deduplication by FoldSeek (command: foldseek easy-cluster pdb res tmp -c 0.9).

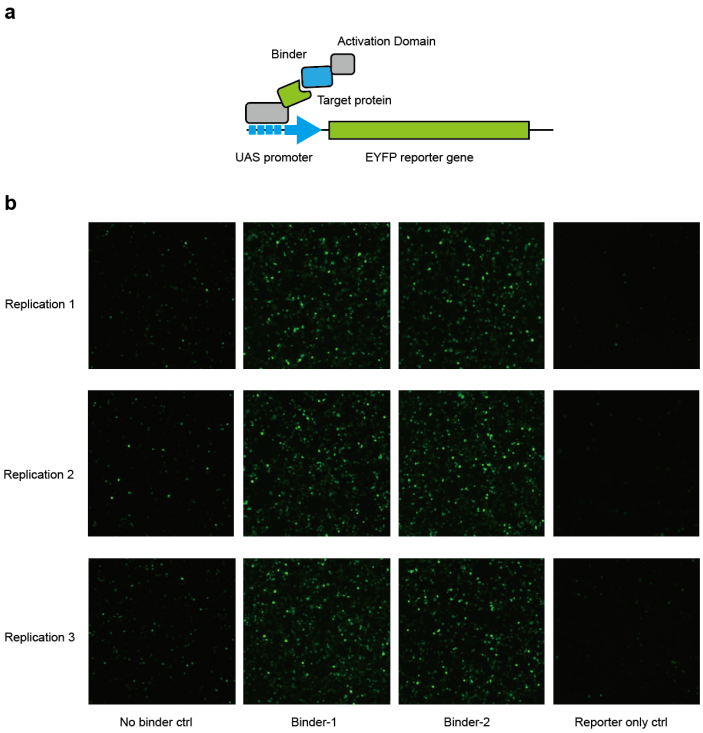
Following this rigorous, two-step deduplication and partitioning workflow, the final splits comprised 23, 459 complexes for training, 582 for testing set from predicted structures, and 137 from PDB. As shown in Supplementary Fig. 22, after fine-tuning ProteinMPNN, sequence recovery improved and perplexity fell on the predicted-structure test set, whereas performance on experimentally determined structures declined. Continued training of MPBind improved performance on both predicted and experimentally

determined structures, demonstrating its robustness across data sources. By contrast, ProteinMPNN showed gains only on predicted scaffolds, yet its weaker performance on experimental structures reminds us that predicted models cannot always replace experimentally resolved data.

Additionally, as shown in Response Figure 6, we were also tried to test the finetuned ProteinMPNN to design sequence from a predicted complex scaffold. We extract sequence from 8GAB, a resolved 3D structure with partial CTLA-4 with a binder[19]. Based on the re-predicted structure, we used our ProteinMPNN-PHC designed two new sequences. To test whether the new generated sequence can bind the CTLA-4, we engineered a reporter system by fusing GAL4 DNA binding domain with the CTLA-4 target region. Binder was fused with gene activation domain to drive the expression of EYFP reporter gene. As shown in Response Figure 6b, we observed weak but obvious EYFP expression. We will further develop such fast reporter system for identifying de novo design binder in the future work.



Supplementary Figure 22. Applications of ProteinMPNN models leveraging the large-scale protein complex structure dataset a Model architecture of ProteinMPNN, consisting of a 3-layer backbone encoder and a 3-layer sequence decoder. Distances between N, Cα, C, O, and virtual Cβ are encoded by the encoder. The encoded features are used together with partial sequences to iteratively generate amino acids in random decoding order through the decoder. **b** Performance of pretrained ProteinMPNN and finetuned ProteinMPNN-PHC on high-quality predicted heterologous complexes or Experimental Structural Scaffold.

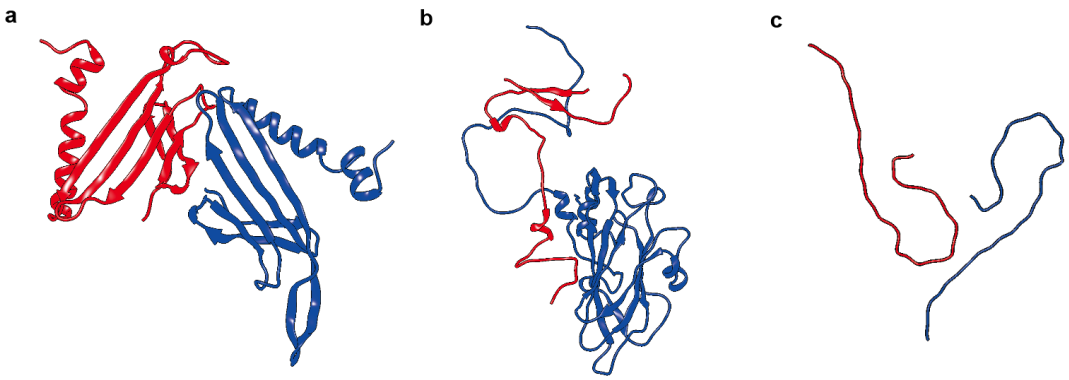


Response Figure 6 De novo design of a CTLA-4 binder and assessment of its binding affinity with a transcriptional activation assay. a Diagram of the reporter system. CTLA4 target region was fused with GAL4 DNA binding domain. Binder was fused with gene activation domain to drive the expression of EYFP reporter gene. **b** fluorescence intensity observed by microscopy. CTLA-4 and binder scaffold were derived from 8GAB[19].

- The Multimer-Surface predictor is presented as an application of the database, but its evaluation has multiple issues. Currently, a random split is used, which may lead to homology leakage, making it unclear whether the model genuinely improves surface extraction. To prevent data leakage, I recommend splitting the test set based on interface similarity, ensuring that the system does not train and test on the same interface type. PINDER provides a well-structured benchmark dataset, and adopting their de-replication strategy could improve the potential leakage issue.

Additionally, a comparison against a PDB-trained method is necessary to determine whether the predicted database provides meaningful improvements over existing resources.

Re: We thank the reviewer for this valuable suggestion. To eliminate any surface- or structure-related data leakage, we applied a rigorous two-stage deduplication pipeline. In addition, we created two independent test sets—one built from PDB complexes and the other from high-confidence predicted models—to compare performance on experimental versus predicted structures. Our binding-site predictor shows improved accuracy on both sets, with particularly pronounced gains on the predicted-structure benchmark.



Response Figure 7. Examples of structures from PINDER database[13].

To construct training dataset, we first collected 71,667 high-quality complexes, filtering by Best LIS ≥ 0.4 and pLDDT ≥ 80. To guarantee diversity and remove redundancy, we applied structural clustering with FoldSeek, eliminating any pair of protein monomers whose TM-score was ≥ 0.5. This procedure yielded a final, non-redundant training set of 9,609 dimeric complexes. A recent study introduced MPBind, which combines sequence, structural, and geometric features to deliver markedly improved predictions across diverse protein-binding sites. Using this set, we further trained the MPBind model (https://github.com/jianlin-cheng/MPBind)[20] for 45 epochs on an NVIDIA RTX 3090 GPU.

Next, we built an independent validation set from predicted structures. We randomly selected partial of the remaining complexes as validation candidates. Subsequently, we conducted meticulous deduplication on these samples based on interface similarity and monomer similarity.

In the interface deduplication phase, we employed a stringent strategy that integrates structural alignment with interface site information which have been used in PINDER dataset before[13]. This strategy aims to identify and eliminate proteins that, despite differing overall structures, exhibit highly similar key functional interfaces. We first utilized FoldSeek to compute structural alignment information between all protein single chains in the candidate samples (command: foldseek easy-search query.pdb targetdb out.m8 tmp --format-output "query,target,evalue,bits,qstart,qend,tstart,tend" -s 9.5). For any pair of proteins (denoted as A and B), we examined the extent of overlap between their structural alignment regions and the pre-defined interface residues. Surface residues in these complexes were defined as those with a relative solvent accessibility greater than 5% that lost more than 1 Å² of absolute solvent accessibility upon protein-protein complex formation. If the overlap between the alignment region of protein A and the entire interface region of protein B exceeds 50%, or if the overlap between the alignment region of protein B and the entire interface region of protein A exceeds 50%, then the pair of proteins is deemed to exhibit interface similarity. Any complex containing protein chains determined to have interface similarity will be removed from the validation set candidate pool. Subsequently, in the monomer deduplication phase, we aimed to thoroughly exclude any single chains in the validation set complexes that display structural similarity with those in the training set.

We employed FoldSeek to perform structural clustering of all monomers in the dataset, using the command: foldseek easy-cluster pdb res tmp -c 0.9.

Following this series of rigorous dual deduplication steps, we ultimately obtained a validation set comprising 205 monomers, thereby maximizing the reliability of the evaluation.

We further constructed the test dataset from experimental structures. As Response Figure 7, we first examined the PINDER protein-complex dataset. Although several entries are of high structural quality, many contain extensive disordered regions.Therefore, we selected Pro_Test_315, an independent protein experiment-resolved protein structure benchmark dataset[14]. As Figure 6b illustrates, the fine-tuned model (MPBind-PHC) achieved a marked performance boost on the duplicated Predicted test set and delivered higher AUC scores on Pro_Test_315. In addition, after we applied the same filtering protocol to eliminate potential leakage from Pro_Test_315—removing any cases that overlapped with the continued-training data—the model still showed a comparable improvement, confirming the robustness of the gain.

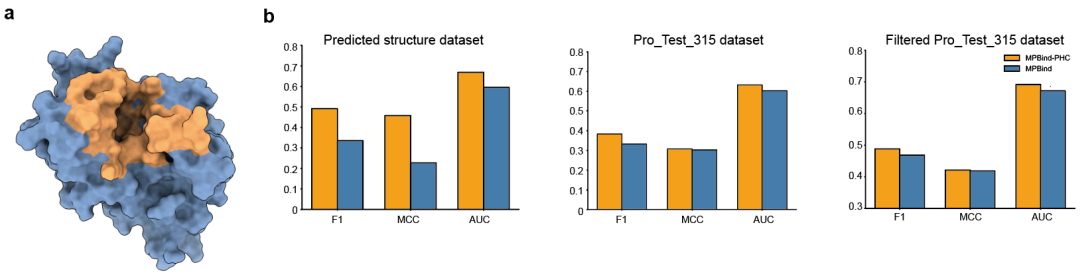


Figure 6. Applications of deep learning models leveraging the large-scale protein complex structure dataset. **a** Surface residues of a protein, where the orange regions represent surface residues, and the blue regions represent non-surface residues. **b** Performance comparison between MPBind and MPBind-PHC across different metrics. MPBind-PHC demonstrates better performance in AUC (Area under the curve, measuring classification ability), F1, and MCC (Matthews Correlation Coefficient, evaluating the quality of binary classifications) compared to MPBind.

In summary, we assessed our model on an independent PDB-derived set and on a rigorously deduplicated set of predicted structures. We also filtered at both structural and surface levels. Across all three evaluation splits, fine-tuned MPBind model which was trained on PDB based data boosted predictive ability, highlighting the value of large-scale PPI data for deep-learning models.

- The manuscript states: "All other data supporting the findings of this study are available from the corresponding author on reasonable request." To ensure full transparency and reproducibility, all necessary data should be made openly available. Currently, only the 170,000 high-quality predictions have been released—please consider also providing the remaining low-quality predictions to allow for further analysis.

Re: Thank you for this valuable suggestion. The full dataset—approximately 1.1 million

predicted structures—is now publicly available on ModelScope (https://www.modelscope.cn/collections/protein_complex_atlas-2ae5e7d4f4a343). We have also launched a companion website showcasing the subset of PPI structures (http://www.biopredictnavigator.cn). Together, these resources offer open access to both high- and lower-confidence models, facilitating comprehensive downstream analyses and applications.

Additionally, the code repository is missing essential scripts for the analysis and lacks documentation necessary for reproducing the study.

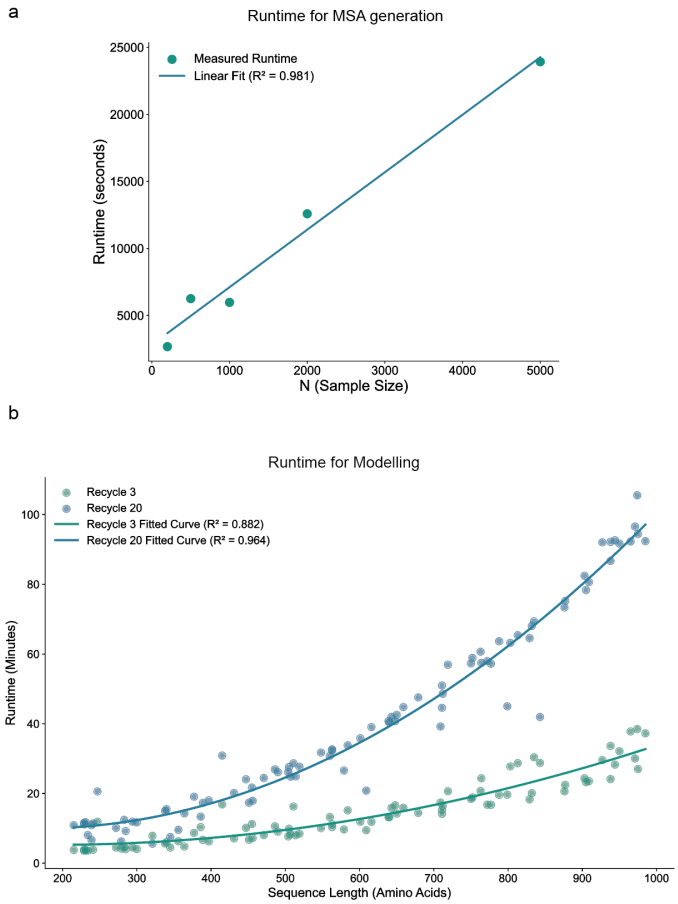
Re: Thank you for pointing out these issues. We added the missing analysis components including the figure generation used in the manuscript, Alphafold3 and related model evaluation, MPBind and proteinMPNN fine-tuning, data processing and evaluation, network generation, genomic co-occurrence analysis, and protein clash analysis. A readme document was supplied (https://github.com/wensm77/Protein-Complex-Atlas/blob/main/README.md). We hope these updates will facilitate a smoother and more transparent reproducibility process.

Minor:

- For others in the community trying to do large-scale PPI runs it would be great to have some technical information regarding runtime for MSA generation and inference of the models.

Re: Thank you for highlighting the importance of reporting runtimes for large-scale PPI analyses. In the revised manuscript we now provide detailed timing benchmarks:
Analyses were run on an x86_64 server (Intel Xeon Platinum 8358, 42 CPU cores, 512 GB RAM) equipped with four NVIDIA A40 GPUs. Supplementary Fig. 3 summarizes the wall-clock time required for MSA generation with different size of the input FASTA set. Supplementary Fig. 3 reports inference times for Colabfold multimer models at two recycle settings (3，20 cycles) and for protein complexes ranging from 200 to 1500 residues.

Briefly, MSA construction scales close to linearly with the number of sequences, whereas model inference shows an approximately quadratic relationship with the total residue count of the complex ($T \propto L^2$). For example, on a single A40 GPU three recycle of a 215-residue complex completes in ~227.45 s (3.79 min), while a 950-residue complex takes ~32 min.
We hope these benchmarks will help the community plan computational resources for high-throughput PPI prediction.



Supplementary Figure 3. Benchmarking of ColabFold Multimer Inference Runtime. a Runtime for Multiple Sequence Alignment (MSA) generation as a function of sample size (N). Measured runtimes are depicted as teal circles, and the linear fit to these data (R^2=0.981) is shown as a dark teal line. **b** Runtime for protein structure modelling as a function of sequence length in amino acids. Data points and fitted curves are shown for two different recycle settings: Recycle 3 (green circles, dark green line, R^2=0.882) and Recycle 20 (blue circles, dark blue line, R^2=0.964). Runtimes in (a) are in seconds, and in (b) are in minutes. Analyses were run on an x86_64 server (Intel Xeon Platinum 8358, 42 CPU cores, 512 GB RAM) equipped with four NVIDIA A40 GPUs.

- Kim et al. (2024) predicted over 350,000 viral protein structures through BFVD. This could be as well referenced in the following sentence: "Recent research has predicted 67,715 newly identified viral protein structures, uncovering a multitude of structurally distinct viral proteins.".

Re: We apologize for the oversight. The reference list and Introduction have been revised as follows:

"…Two recent studies have predicted 67,715 and more than 350,000 novel viral protein structures, respectively, thereby revealing an unprecedented diversity of viral protein architectures[16,17]….".

- The code for in GitHub and has no license: https://github.com/Protein-Complex-Lab/Protein-Complex-Atlas

Re: We apologize for the oversight. We have now added an MIT License to the GitHub repository (https://github.com/wensm77/Protein-Complex-Atlas).

Reviewer #3 (Remarks on code availability):

The code repository includes many components of the study but still lacks some parts of the analysis. Additionally, it is missing a license file and a documentation to be more comprehensive would also be beneficial.

Re: We acknowledged reviewers for raising this important problem. An MIT License has been added to the GitHub repository. We added the missing analysis components including the figure generation used in the manuscript, Alphafold3 and related model evaluation, MPBind and proteinMPNN fine-tuning, data processing and evaluation, network generation, genomic co-occurrence analysis, and protein clash analysis. A readme document was supplied (https://github.com/wensm77/Protein-Complex-Atlas/blob/main/README.md) to improve transparency and reproducibility.

Reference:

1. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications* **13**, 1265 (2022).
2. Zhu, W., Shenoy, A., Kundrotas, P. & Elofsson, A. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics* **39**(2023).
3. Kim, A.-R. *et al.* Enhanced Protein-Protein Interaction Discovery via AlphaFold-Multimer. *bioRxiv*, 2024.02.19.580970 (2024).
4. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.10.04.463034 (2021).
5. Schweke, H. *et al.* An atlas of protein homo-oligomerization across domains of life. *Cell* **187**, 999-1010.e15 (2024).
6. Humphreys, I.R. *et al.* Protein interactions in human pathogens revealed through deep learning. *Nature Microbiology* **9**, 2642-2652 (2024).
7. Douguet, D., Chen, H.-C., Tovchigrechko, A. & Vakser, I.A. Dockground resource for studying protein–protein interfaces. *Bioinformatics* **22**, 2612-2618 (2006).
8. Team, B.A.A.S. *et al.* Protenix - Advancing Structure Prediction Through a Comprehensive AlphaFold3 Reproduction. *bioRxiv*, 2025.01.08.631967 (2025).
9. Discovery, C. *et al.* Chai-1: Decoding the molecular interactions of life. *bioRxiv*, 2024.10.10.615955 (2024).
10. Passaro, S. *et al.* Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv*, 2025.06.14.659707 (2025).
11. Biasini, M. *et al.* OpenStructure: an integrated software framework for computational structural

biology. *Acta Crystallogr D Biol Crystallogr* **69**, 701-9 (2013).
12. Mirabello, C. & Wallner, B. DockQ v2: improved automatic quality measure for protein multimers, nucleic acids, and small molecules. *Bioinformatics* **40**(2024).
13. Kovtun, D. *et al.* PINDER: The protein interaction dataset and evaluation resource. *bioRxiv*, 2024.07.17.603980 (2024).
14. Fang, Y. *et al.* DeepProSite: structure-aware protein binding site prediction using ESMFold and pretrained language model. *Bioinformatics* **39**(2023).
15. Burke, D.F. *et al.* Towards a structurally resolved human protein interaction network. *Nature Structural & Molecular Biology* **30**, 216-225 (2023).
16. Schaffer, L.V. *et al.* Multimodal cell maps as a foundation for structural and functional genomics. *Nature* **642**, 222-231 (2025).
17. Schmid, E.W. & Walter, J.C. Predictomes, a classifier-curated database of AlphaFold-modeled protein-protein interactions. *Molecular Cell* **85**, 1216-1232.e5 (2025).
18. Zhang, J. *et al.* Computing the Human Interactome. *bioRxiv*, 2024.10.01.615885 (2024).
19. Yang, W. *et al.* Design of high-affinity binders to immune modulating receptors for cancer immunotherapy. *Nature Communications* **16**, 2001 (2025).
20. Wang, Y., Boadu, F. & Cheng, J. MPBind: Multitask Protein Binding Site Prediction by Protein Language Models and Equivariant Graph Neural Networks. *bioRxiv*, 2025.04.12.648527 (2025).

Response to Referees Letter

We thank the Reviewers for their careful assessment of our manuscript. In response to the key points raised—particularly by Reviewer #1 regarding focus, methodological clarity, and availability—we have substantially revised the paper to streamline the narrative and deepen the explanations. Specifically, we (i) removed the out-of-scope ProteinMPNN fine-tuning section, bacterial defense system identification sectionand reorganize the main text to ensure a more coherent and logically structured narrative;(ii) explicitly define the negative dataset used in the homomer benchmark (Fig. S6) and detailed construction of the virus–human dataset; (iii) improve accessibility by consolidating code/data links, enabling HTTPS on the website, correcting the 3D viewer's coloring scheme to match pLDDT labels, and providing the citation and a license description that clarifies compatibility of all components; and (iv) thoroughly edit the language of whole manuscript to improve readability, including rewriting ambiguous phrases. A point-by-point response follows, with reviewer remarks reproduced verbatim.

Reviewer #1 (Remarks to the Author):

In this paper the authors provide an impressive set of predictions of PPIs. Unfortunately, this is also the major problem with the paper. It reads as a list of rather unrelated problems, all of potential interest, but none of them is explained in sufficient detail to provide interesting insights.

Re: We thank the reviewer for recognizing the scope of our dataset and for pointing out that the previous draft read as a series of loosely connected analyses. We have substantially restructured the manuscript to achieve a tighter focus and clearer narrative centered on the systematic prediction of PPIs. Specifically, we removed the tangential ProteinMPNN fine-tuning and bacterial defense system identification subsections, which previously diluted the main message, and reorganized the Results to present a cohesive, logically progressive story.

The revised Results now comprise five streamlined figures (previously six), each accompanied by bridging text that clarifies transitions between analyses. We introduce the AlphaFold3-like model comparison and a rapid method for detecting locus-separated related complexes at the outset, establishing the methodological foundation before turning to large protein assemblies (revised Fig. 2). In Fig. 2, we replace a loose narrative with a reconstruction-focused analysis of multi-component and higher-order assemblies, organized around three aspects: proteome-scale community detection on predicted binary PPIs; a CopRS case study integrating homo- and heterodimeric interfaces with CombFold; and a systematic homodimer symmetry survey (QSproteome). This revision emphasizes principled reconstruction of representative architectures rather than a simple catalogue of complexes.

The protein binding site prediction analysis has been moved to the Supplementary Information to maintain narrative focus while preserving methodological completeness.

Together, these changes have greatly improved the logical flow, readability, and conceptual unity of the manuscript, while retaining sufficient technical depth to support meaningful biological insights.

The most obvious part that is out of sight is the fine-tuning of ProteinMPNN. What does that add for our understanding of PPIs?

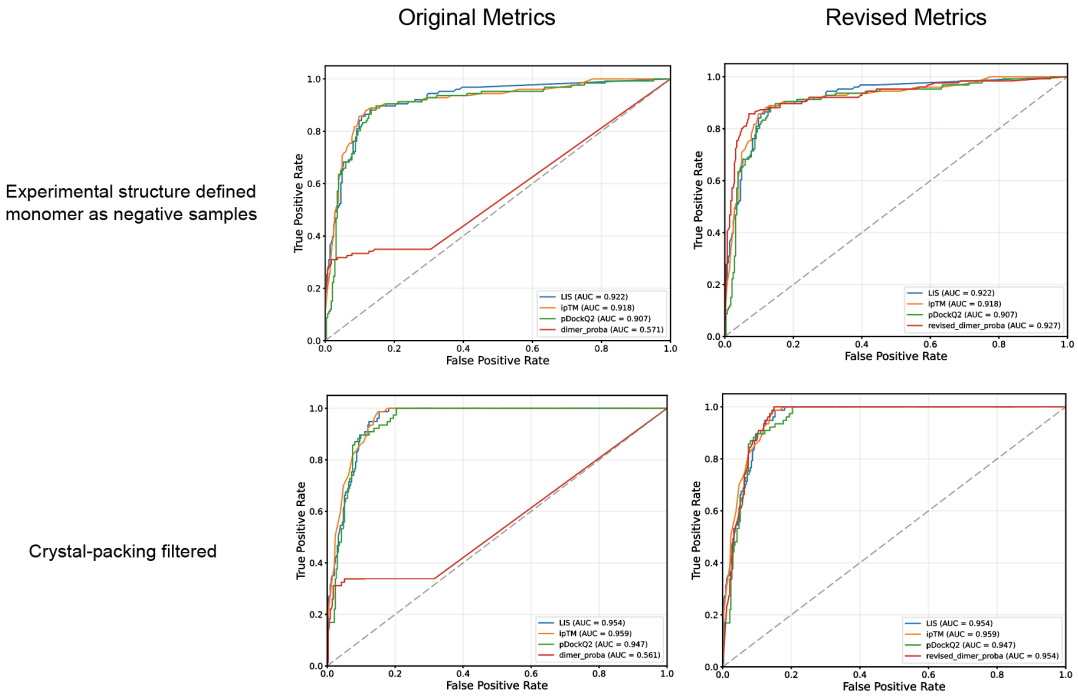
Response: We appreciate the reviewer's question. The ProteinMPNN fine-tuning component was tangential to our central aim—predicting PPI. To maintain a coherent narrative, we have removed this subsection from the manuscript.

Another unclear part is what negative set is used for the homomer benchmark (Fig. S5) or how the virus-human dataset is constructed (Homology reduction to training set, etc).

Re: We thank the reviewer for pointing out the lack of clarity and for the helpful suggestions. For the homomer benchmark, we used the dataset curated in prior work[1]. One entry has since been removed from the PDB database; after excluding this and restricting sequences to <1200 residues, the evaluation set comprised 411 proteins. Following the prior study[1], we considered two ways to distinguish true homodimers from monomers. Approach 1 relies on the experimental annotation of the biological assembly (monomer vs dimer). Approach 2 follows the analyses that some crystallographic dimers arise from fortuitous crystal contacts; in this case, those entries are treated as monomers. As shown in Supplementary Figure 6a and Letter Figure 1, under Approach 1, three metrics showed strong discriminative power: pDockQ2 (AUC = 0.907), best LIS (AUC = 0.922), and ipTM (AUC = 0.918). When we ran the authors' public GitHub script without modification (https://github.com/HugoSchweke/QSproteome_protocol), the reported metric dimer_proba displayed weak discrimination (AUC = 0.571), which is inconsistent with their publication[1]. However, after removing a prior distance-based filtering step in the script, we observed performance comparable to other metrics (AUC = 0.927). We hypothesized that the uploaded script is incomplete and therefore we did not include the modified result in the main text. Consistent with Approach 2, when a subset of assemblies annotated as dimers in the experimental structures was reclassified as monomers due to likely crystal-packing fortuitous interactions as provided in the supplementary information of previous research, Best LIS, ipTM and pDockQ along with the revised dimer_proba metric achieved higher discriminative performance (dimer_proba AUC=0.954, Best LIS AUC=0.954). Additionally, we showed the label for each PDB entries of two approaches in the source data file.

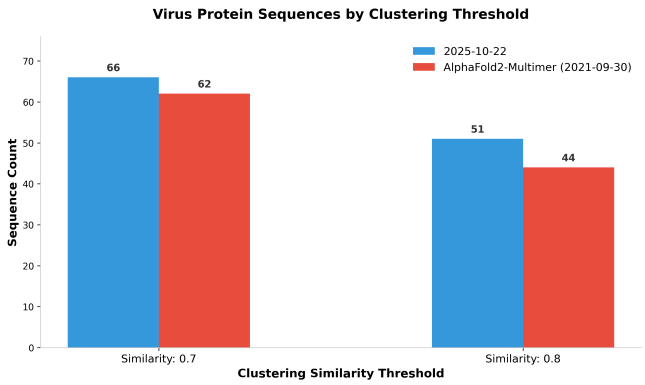
Such analysis was described in the main text as:"… To evaluate the metrics for identifying true homodimers in ColabFold modeling, we constructed a benchmark dataset comprising 411 proteins from the PDB, curated in previous studies[1] to remove redundancy

and eliminate overlap with the AlphaFold2 training set. In the beginning, proteins resolved as monomers by X-ray crystallography were used as the negative dataset. Similarly, Best LIS showed best ROC performance when we analyzed the top-ranked model of each prediction (Supplementary Fig. 6a). Following previous research[1], we also reclassified certain assemblies from dimers to monomers to correct for likely crystal-packing artifacts. This adjustment improved the overall classification performance, as evidenced by an increased AUC (Supplementary Fig. 6a and b).… ”



Letter Figure 1: ROC curve of multiple metrics for discriminating true homodimers.

Secondly, we constructed the virus–human interaction set from two previously published prediction datasets lacking comprehensive 3D structures: HVIDB (with both experimentally verified and predicted PPIs) and P-HIPSTer. Given computational budget constraint (P-HIPSTer >280k interactions; HVIDB 48,643 PPIs), we retained HVIDB entries and further sampled from P-HIPSTer. After removing duplicates, we obtained 81,235 unique virus–host PPI candidates covering 3,531 viral proteins and 8,532 host proteins for downstream analyses. We did not filter predictions using the AlphaFold2-Multimer training set so as to provide a comprehensive assessment of model outputs. To evaluate the novelty of the predicted high-confidence virus–human PPIs, however, we performed sequence-homology searches against the AlphaFold-Multimer training dataset (PDB entries by 2021-09-30) and against the current whole PDB database. As shown in Letter Figure 2, using a 70% sequence-identity threshold, 62 of predicted interactions exhibited detectable similarity to entries in the AlphaFold-Multimer training set, in which both interacting chains matched homologs present in solved structures.



Letter Figure 2: Detected homolog structures in PDB database

We added corresponding description in main text as: "…wo curated databases of predicted human–virus protein–protein interactions, HVIDB and P-HIPSTer, provide precise resources for predictions. Due to computational constraints (P-HIPSTer >280k interactions; HVIDB 48,643 PPIs), we retained HVIDB entries and subsampled P-HIPSTer. After removing duplicates, we totally collected 81,235 unique PPI candidates covering 3,531 virus proteins and 8,532 host proteins (Fig. 4a). Novelty was evaluated against the AF-Multimer training set, and 62 virus–human PPIs (at a 70% sequence identity threshold) were found to match training-set structures for both chains.....".

I am convinced that this can become an interesting resource, but unfortunately, by trying to add too much into a single paper, the authors make it almost unreadable.

Re: We appreciate this insight and agree that breadth had come at the expense of clarity in the earlier draft. We have now tightened the scope around a single central objective, reorganized the Results into one step-by-step workflow, and deleted tangential components. Concretely, we (i) removed the ProteinMPNN and bacterial defense system identification subsection, (ii) reduced main figures from 6 to 5, (iii) standardized terminology and writing. We believe these changes substantially improve readability while preserving the substance of the resource.

Availability;

It is unclear what is available where and under what license, see remarks below. Ideally the website www.biopredictnavigator.cn should provide an easy access to the data. However it is far from ready (https does not work), it contains still images of the proteins, the coloring scheme in the 3D viewer is not consistent with the labels (it is colored from N-C and not by pLDDT). etc.

Re: We thank the reviewer for highlighting the shortcomings in data availability, licensing, and website configuration. In response, we have clarified what resources are available, where they can be accessed, and under which licenses they are distributed.

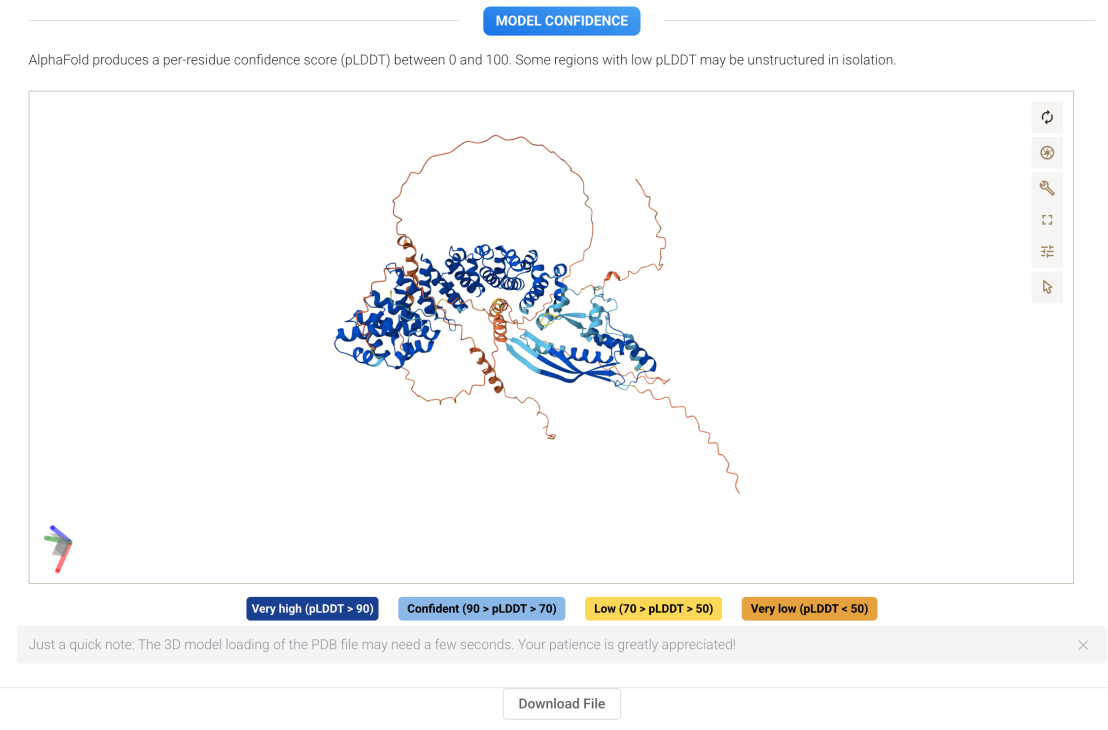
Specifically, ColabFold's source code is released under the MIT license, and AlphaFold2's source code under the Apache 2.0 license. In line with Reviewer 1's suggestions for clearer licensing, and consistent with the Apache 2.0 license of AlphaFold2, we now explicitly release our own datasets and derived predictions for academic use under the Creative Commons CC BY-NC 4.0 license.

For reference, two other ColabFold-derived prediction datasets are distributed under the CC BY 4.0 license (https://viro3d.cvr.gla.ac.uk/about and https://bfvd.steineggerlab.workers.dev/). We have added the following notice to both the ModelScope page and the website:

"All data are freely available for academic use under the Creative Commons CC BY-NC 4.0 license. If you use this resource, please cite: Xianzhi Qi et al. 'Atlas of Predicted Protein Complex Structures Across Kingdoms,' DOI: [to be inserted]."

Once the paper is accepted and a DOI is assigned, we will update the page accordingly. Additionally, our own code is released under the MIT license, which is clearly indicated in the GitHub repository.

We have also addressed the technical issues noted by the reviewer as follows:
The HTTPS failure during the initial review was due to a certificate configuration issue, which has now been fixed. HTTPS is now fully enabled and enforced via HSTS (https://www.biopredictnavigator.cn/). Still images of proteins have been removed from the website. As shown in Letter Figure 3, the coloring scheme has been corrected to reflect pLDDT values using the standard AlphaFold palette, accompanied by an on-screen legend (0–50 = red, 50–70 = orange, 70–90 = yellow, 90–100 = blue).



Letter Figure 3: pLDDT values with a color legend

Language:

The paper contains non-optimal expressions making it hard to read. It really needs a good rewrite of someone who can tell a story and not just staple sentences on top of each others.

One of these is "showed outperformed efficacy", I guess it should be "showed an efficacy" or something like that.

Another one is "Then we fine-tuned the released sourced model parameters using the distilled high-quality data". WHat do the authors mean here?

/Arne Elofsson

Re: We thank the reviewer for pointing out the issues related to language expression and narrative flow. In response, we have thoroughly revised the manuscript to enhance its clarity and coherence in three main aspects: (1) optimizing sentence structures by adopting more precise terminology, (2) improving the logical flow through the addition of appropriate transitional phrases, and (3) correcting ambiguous or redundant expressions.

Regarding the first question: "showed outperformed efficacy" in our manuscript. We agree that this expression was non-optimal and grammatically awkward, which hindered readability. This has been improved as: "… MPBind, a recent developed protein binding site prediction model, was based on protein language models and equivariant graph neural networks…".In response to the second question, we note that the corresponding content was part of the ProteinMPNN section and has therefore been removed from the main text.

Then, we conducted a comprehensive revision of the language and grammar of the entire article.

(1) Enhancing sentence structure by adopting precise terminology
To strengthen the scientific tone and improve readability, we carefully revised numerous expressions throughout the manuscript. For instance, mine in "Here, we developed an approach to mine potential protein-protein interacting pairs" was replaced with identify, and "neighboring protein pairs" was refined to adjacent protein pairs for greater precision. Similarly, "multiple metrics have been developed to distinguish between true and false interactions in predicted protein complexes" was rewritten as "To distinguish true interactions from artifacts, we assessed each predicted complex using multiple established confidence metrics," improving both clarity and conciseness. Other examples include revising "but was accompanied by a reduction in the recall rate" to "but concurrently resulted in a reduction in the recall rate," and "we asked what fraction of the entire proteome" to "we examined what proportion of the entire proteome." Phrasing such as "analysis was expanded to the phylum level" was also refined to "analysis was extended to the phylum level." Additionally, "These results indicate that many high-abundance, conserved proteins

are shared well beyond the species boundary" was polished to "These results indicate that numerous conserved proteins are shared well beyond the species boundaries." These are representative examples, and similar targeted revisions were applied extensively across the manuscript to enhance linguistic precision, consistency, and overall fluency.

(2) Improving logical flow through added transitional phrases

To enhance the logical flow and ensure a more coherent narrative, we incorporated additional transitional phrases and logical connectors throughout the manuscript. For instance, in the paragraph corresponding to Figure 2, we introduced the sentence "Having delineated communities largely from heterodimeric edges, we next asked whether their constituent subunits also self-associate, thereby assembling into higher-order complexes." Likewise, we added "Building on complexes inferred from heterodimeric interactions, we next examined whether homodimerization could nucleate higher-order assemblies, especially in virulence modules." These additions help establish smoother conceptual transitions between analytical steps. Furthermore, we incorporated conjunctions such as "consequently," "notably," and "furthermore" to make the context less segmented and more cohesive. Although these modifications are relatively minor, we believe they collectively enhance the manuscript's readability and logical continuity.

(3) Correcting ambiguous or redundant expressions

Terminology was standardized and refined for greater accuracy and fluency. For instance, "AlphaFold-based structural modeling method" was revised to "AlphaFold-based structure-prediction pipeline" to better reflect methodological scope, and "encompassing a diverse spectrum of biological kingdoms including viruses, mammals, plant, archaea, and bacteria" was simplified to "spanning viruses, archaea, bacteria, plants, and mammals" for conciseness and parallelism.

Scientific phrasing was modernized to align with current conventions—"The rapid development of high-accuracy structural modeling" was updated to "The rapid progress of high-accuracy structure prediction", while "full-atom models" was refined to "especially all-atom models" for terminological precision. Additionally, statements were restructured for clarity and emphasis: "The low high-confidence PPI ratio among the experimental approaches applied may be due to experimental false positives and undetected interactions" was rewritten as "High-throughput screening identified numerous candidate protein–protein interactions; nevertheless, only a minority received high-confidence predictions from AlphaFold-Multimer", providing a clearer and more balanced scientific interpretation.

Overall, we have thoroughly revised the entire manuscript to reduce non-optimal expressions and enhance its readability.

Reviewer #1 (Remarks on code availability):

I have not looked at the code, partly because I did not have time but also because the entire data structure was confusing.

The data availability is almost as confusing as the paper.

The github https://github.com/wensm77/Protein-Complex-Atlas

The site: https://www.biopredictnavigator.cn/ is not available
only http://www.biopredictnavigator.cn/ (not secure giving a warning for most users)

The data that should be
on https://www.modelscope.cn/collections/protein_complex_atlas-2ae5e7d4f4a343 is found at https://www.modelscope.cn/datasets/Yesir97/Protein-Complex-Atlas (I think), and this links to another GitHub site: https://github.com/Protein-Complex-Lab/Protein-Complex-Atlas, which in turn links to the HuggingFace site.

No doi numbers are provided, and it is not clear what license applies to what (many things seem to be cc-by-nc-nd, but not sure this is consistent with, for instance, colabfold, which is apache-2 (I think).

Re: We thank the reviewer for pointing out this issue. In response to the comments regarding code availability and repository organization, we have removed the redundant GitHub links and retained a single repository—https://github.com/wensm77/Protein-Complex-Atlas, which is now cited in the Code Availability section as:

“The code for this manuscript is available at https://github.com/wensm77/Protein-Complex-Atlas.”

We have also corrected the TLS configuration so that the website now serves securely over HTTPS with automatic HTTPS redirection and without any browser warnings (https://www.biopredictnavigator.cn/).

ColabFold’s source code is MIT-licensed, and AlphaFold2’s code is distributed under the Apache 2.0 license. In line with Reviewer 1’s suggestion for clear licensing, and considering the Apache 2.0 license of AlphaFold2, we now explicitly release our derived prediction datasets for academic use under the Creative Commons CC BY-NC 4.0 license. The following notice has been added to the ModelScope page and the project website:

“All data are freely available for academic use under the Creative Commons CC BY-NC 4.0 license. If you use this resource, please cite: Xianzhi Qi et al. ‘Atlas of Predicted Protein Complex Structures Across Kingdoms,’ DOI: [to be inserted].”

Once the paper is accepted and a DOI is assigned, we will update the page accordingly. Additionally, our own code is released under the MIT license, which is clearly indicated in the GitHub repository.

1. Schweke, H. *et al.* An atlas of protein homo-oligomerization across domains of life. *Cell* **187**, 999-1010.e15 (2024).