Supplementary Information for **Atlas of Predicted Protein Complex Structures Across Kingdoms**

Xianzhi Qi[1†], Cheng Ye[1†], Jianqiang Liang[1†], Shimin Wen[2†], Yuanyuan Li[1†], Liupeng Li[1†], Kai Ding[1†], Yongfu Hao[1†], Junjie Fei[3†], Weian Mao[4,5†], Zhiyu Lin[2], Yichong Shen[1], Hongjie Zhu[2], Yayun Hu[1], Rui Zhang[1], Pengli Ji[1], Yafei Lu[1], Bonan Liu[2], Han Wang[2], Yuxuan Chen[2], Zhenguo Ma[1], Peiyuan Yang[5], Xinyu Xu[1], Junlong Wu[6,7], Youyuan Zhu[1], Qiaosha Zou[1], Wencheng Zhu[8], Shuya Li[9], Hongyi Xin[10], Daji Ergu[2], Jianyang Zeng[9], Zhi-Xiong Jim Xiao[3], Chunhua Shen[5*], Ying Cai[2*], Yong Yi[3*], Dacheng Ma[1,11,12*]

[1]Zhejiang Lab, Hangzhou, 311121, China.

[2]College of Computer Science and Artificial Intelligence, Southwest Minzu University, Chengdu, 610225, China.

[3]Center of Growth, Metabolism and Aging, Key Laboratory of Bio-Resource and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, 610064, China.

[4]Australian Institute for Machine Learning, The University of Adelaide, South Australia, 5005, Australia.

[5]Zhejiang University, Hangzhou, 310058, China.

[6]Department of Urology, Fudan University Shanghai Cancer Center, Shanghai, 200032, China.

[7]Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, 200032, China.

[8]Institute of Neuroscience, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, 200031, China.

[9]School of Engineering, Westlake University, Hangzhou, 310024, China.

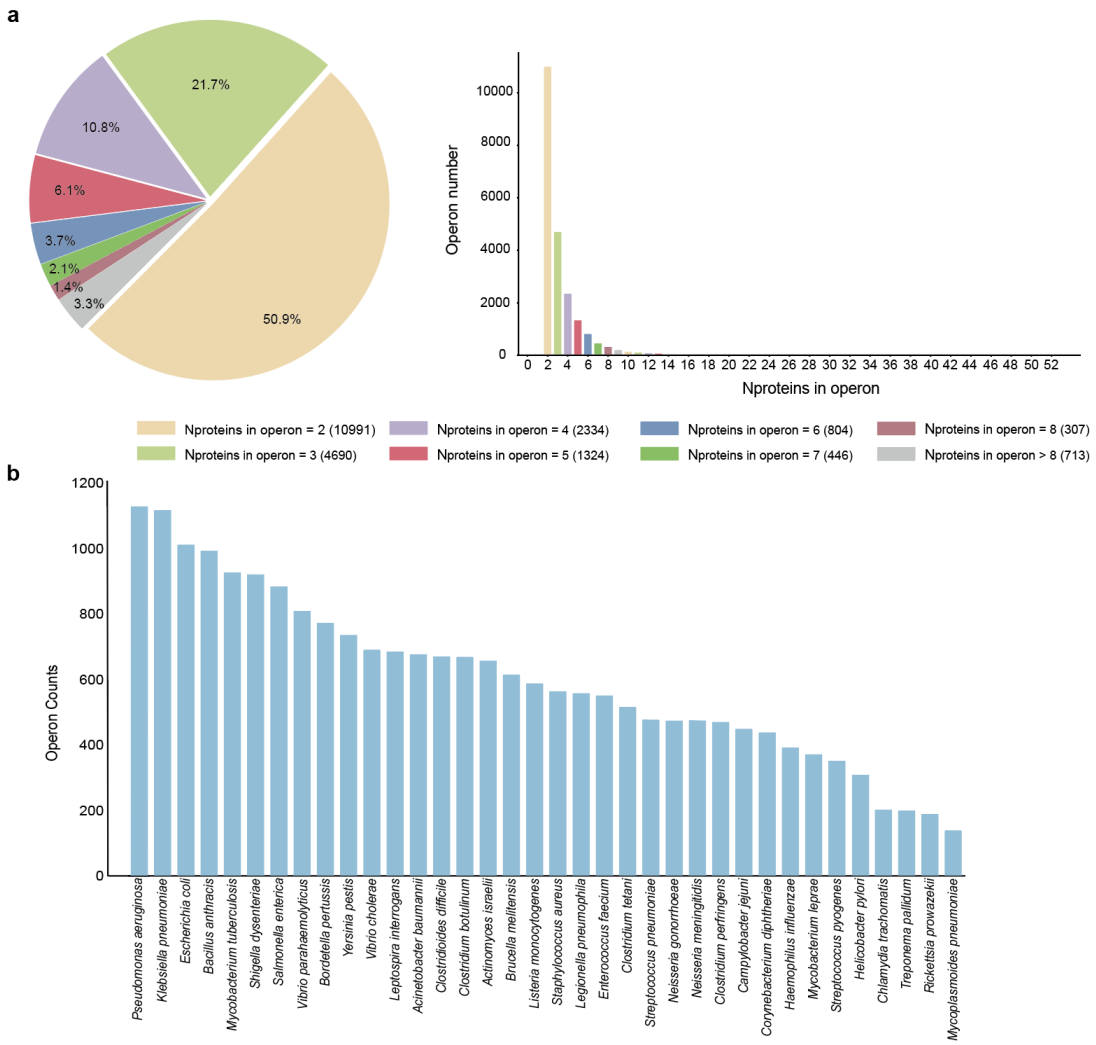
[10]Global Institute of Future Technology, Shanghai Jiao Tong University, Shanghai, 200240, China.

[11]School of Biotechnology, Jiangnan University, Wuxi, 214122, China.

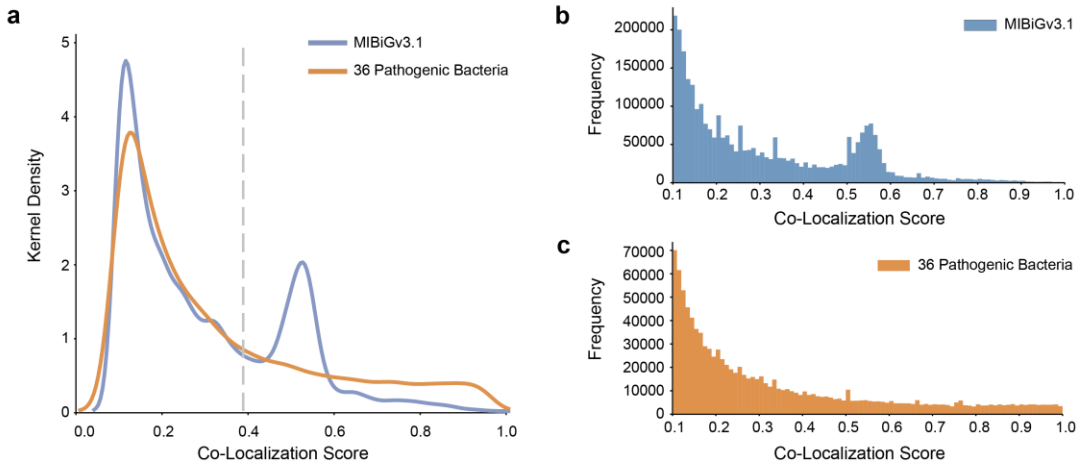
[12]Lead contact.

†These authors contributed equally.

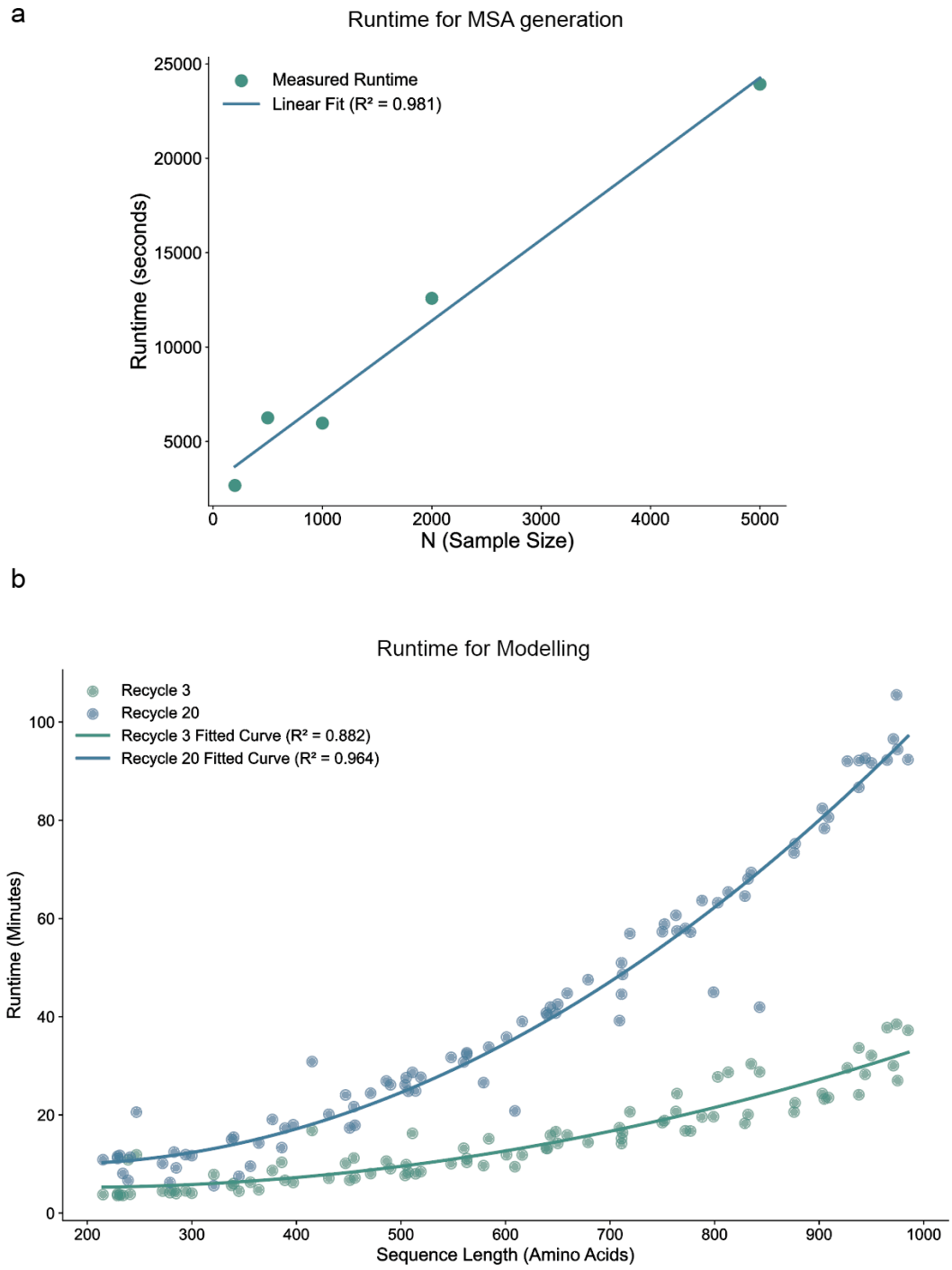
*To whom correspondence should be addressed: chunhuashen@zju.edu.cn (C. Shen), caiying@swun.edu.cn (Y. Cai), yy-yiyong@scu.edu.cn (Y. Yi), dacheng2023@126.com (D. Ma).



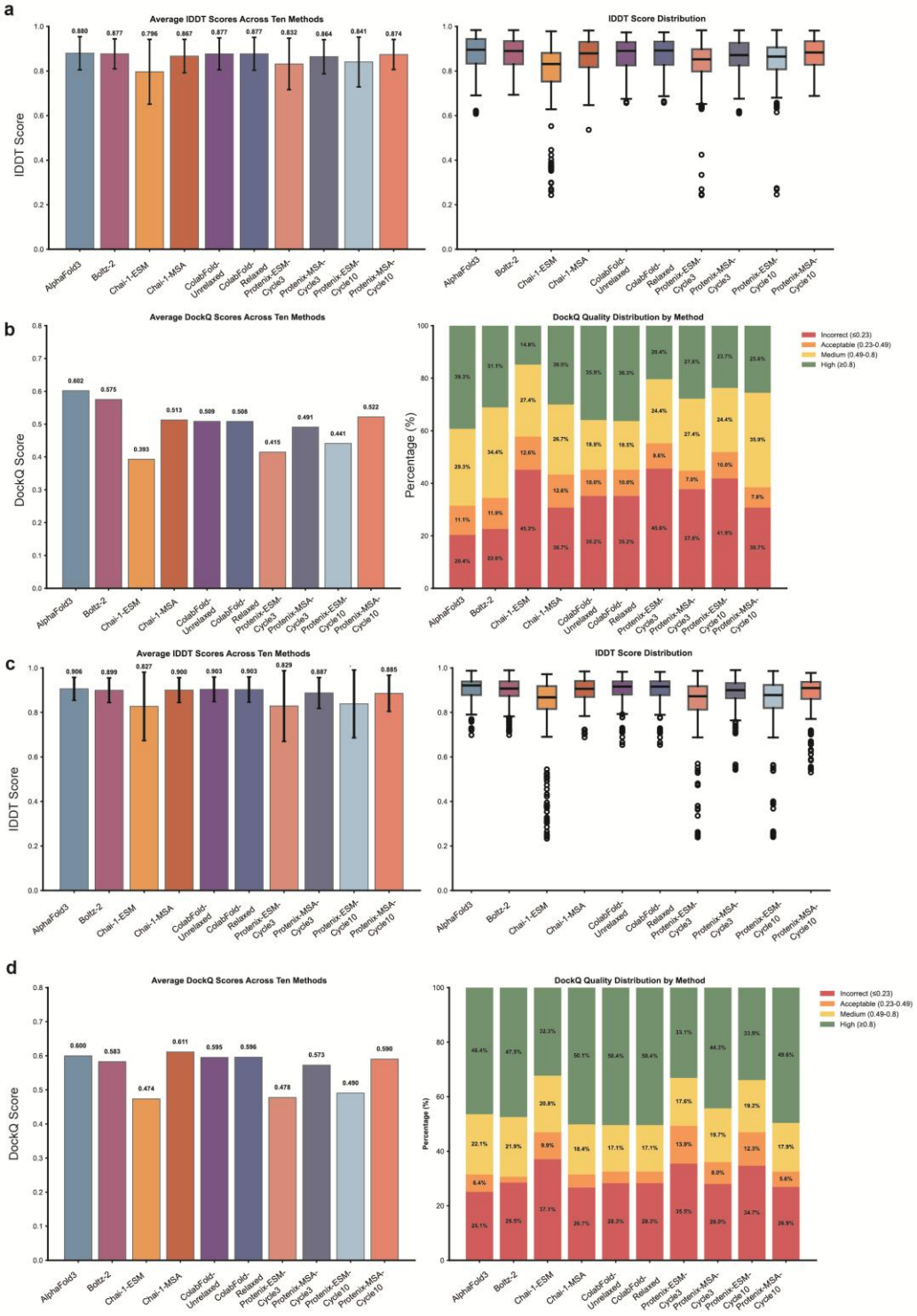
Supplementary Figure 1. Quantitative Analysis of Operons in each pathogenic bacterium. a Gene count distribution of predicted operons. The left panel shows a pie chart representing the distribution of operons based on the number of proteins in each operon across 36 bacterial genomes. The majority of predicted operons consist of 2 proteins (50.9%), followed by operons containing 3 proteins (21.7%) and 4 proteins (10.8%). The right panel displays a bar plot of operon cluster sizes, with the x-axis representing the operon gene count and the y-axis indicating the number of operons. The bar plot highlights that the most frequent operon sizes are small (2 to 3 proteins), with the number of clusters decreasing as the operon size increases. $N_{proteins}$ in operon, the number of proteins in each operon**. b** Operon count of 36 pathogenic bacteria. The bar plot illustrates the total number of operons (gene counts>1) predicted in 36 pathogenic bacterial species.



Supplementary Figure 2. Co-localization score distribution of the protein pairs in a batch of Biosynthetic Gene cluster (MIBiGv3.1)[1] and enumerated protein pairs in 36 pathogenic bacteria.



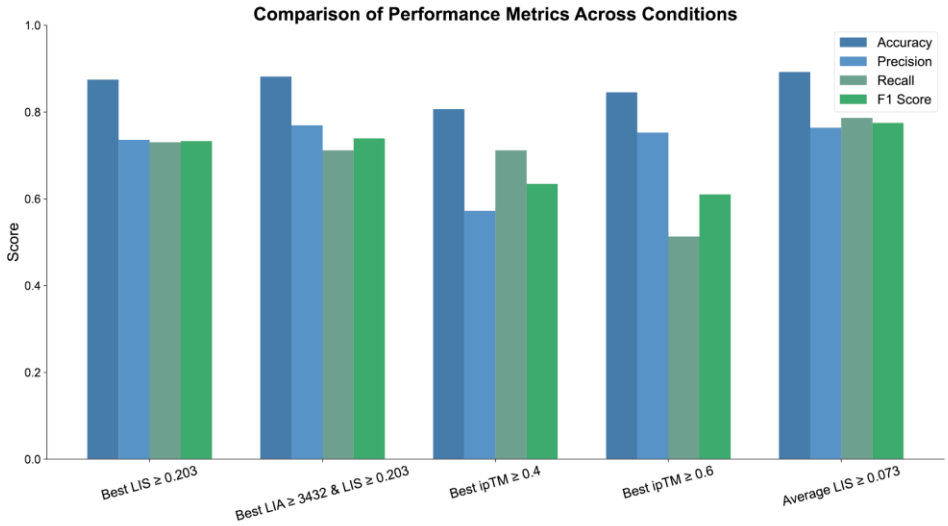
Supplementary Figure 3. Benchmarking of ColabFold Multimer Inference Runtime. a Runtime for Multiple Sequence Alignment (MSA) generation as a function of sample size (N). Measured runtimes are depicted as teal circles, and the linear fit to these data (R^2=0.981) is shown as a dark teal line. **b** Runtime 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.



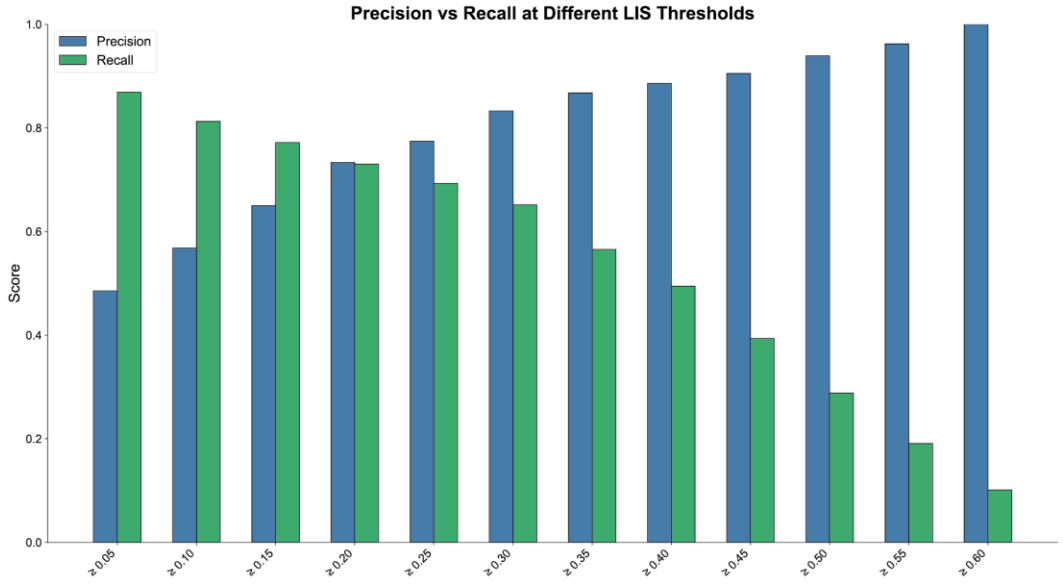
Supplementary Figure 4. Comparison of the prediction accuracy for AlphaFold3 with ColabFold, Chai-1, Protenix, and Boltz-2 for both heterodimers and homodimers. a Local Distance Difference Test (lDDT) scores were calculated between the predicted and experimental structures for 54 non-redundant heterodimers. 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.

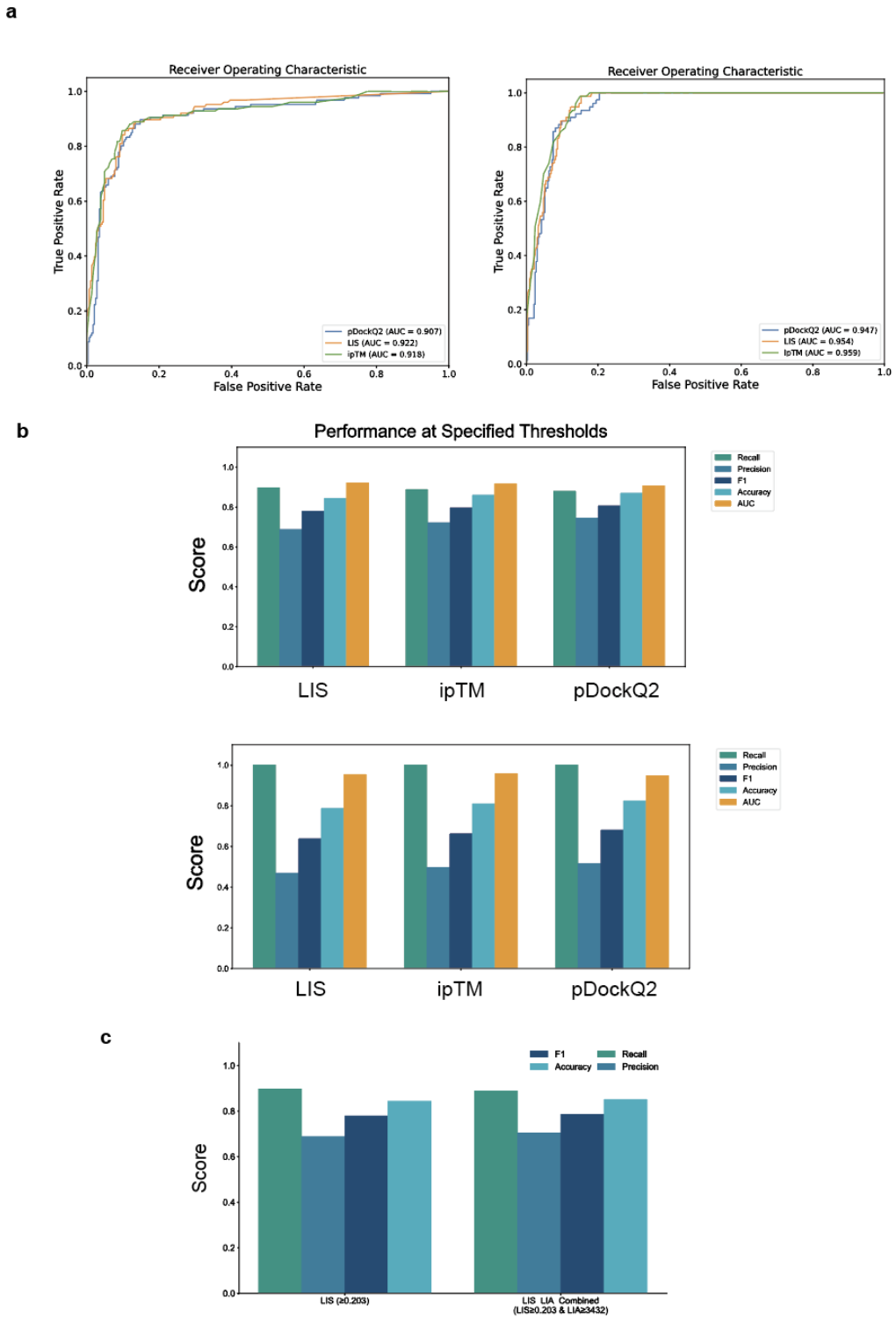
a



b

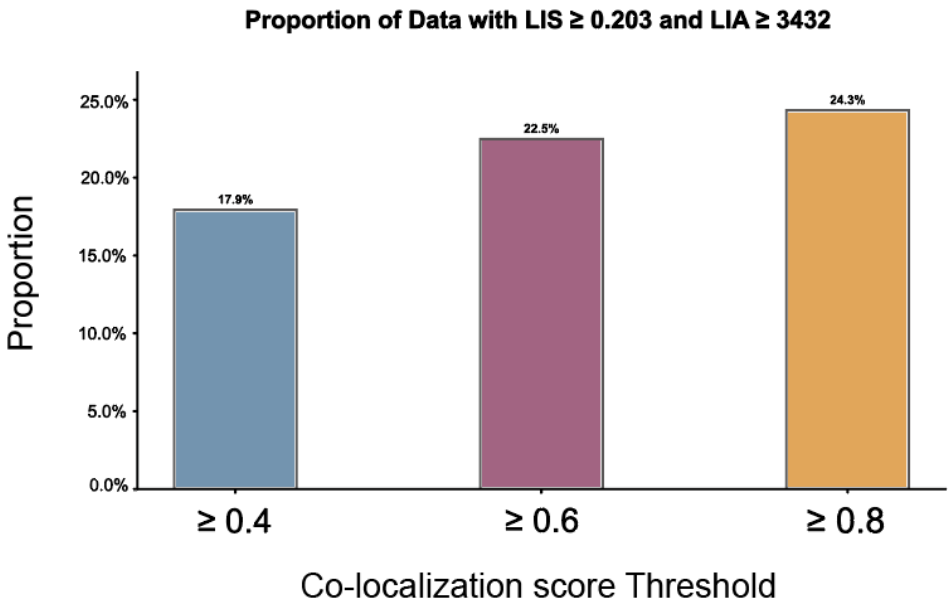


Supplementary Figure 5. Performance of scoring metrics for distinguishing true and false heterodimeric PPIs. **a** Summary of accuracy, precision, recall and F1-score obtained with representative decision rules. Columns from left to right: (i) Best LIS ≥ 0.203; (ii) combined filter Best LIA ≥ 3432 and 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 Best LIA with Best 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 6. 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) when the negative samples were chosen from proteins resolved by X-ray crystallography as monomer (approach 1, left). ROC curves for Best LIS (AUC = 0.954), ipTM (AUC = 0.959), and pDockQ2 (AUC = 0.947) were compared, with crystallographic dimers arising from fortuitous crystal contacts additionally treated as negative

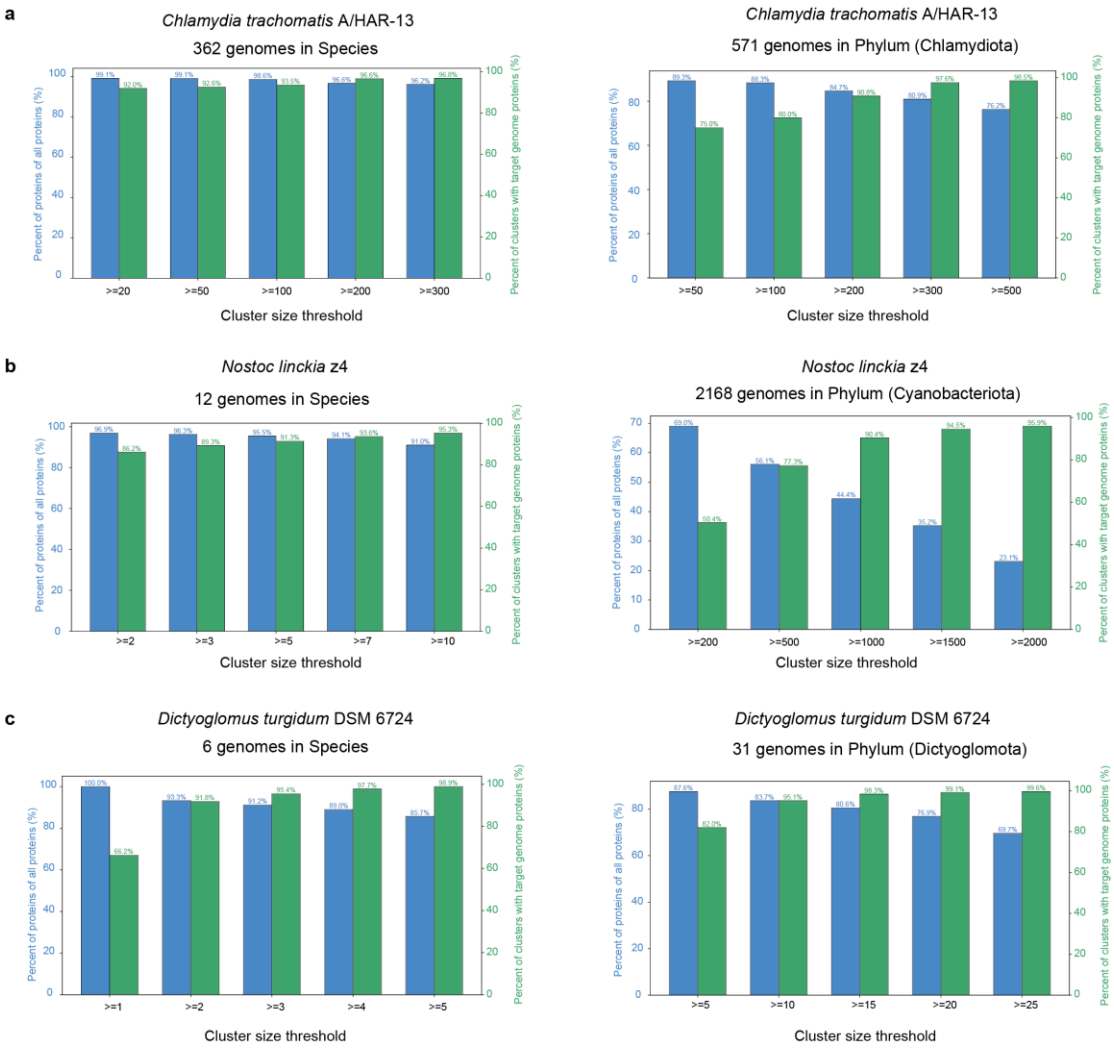
samples, as in the previous study (approach 2, right). All these data were analyzed on the ipTM rank-1 model. **b** Comparative bar plot illustrating multiple performance metrics (AUC, accuracy, precision, recall, and F1 score) of approach 1 (above) or 2 (below) with defined threshold (Best LIS as 0.203,ipTM as 0.6 or pDockQ as 0.23). **c** Bar-plot comparison of classification performance when using the Best LIS cut-off alone (Best LIS ≥ 0.203) versus a composite filter that requires both Best LIS ≥ 0.203 and Best LIA ≥ 3 432.



Supplementary Figure 7. Genomic features that enrich for high-confidence PPIs. Fraction of pairs that satisfy the structural criteria Best LIS ≥ 0.203 and Best LIA ≥ 3 432 under three co-localization score cut-offs. Enrichment rises from 17.9 % (≥ 0.4) to 24.3 % (≥ 0.8), showing that stronger co-localization scores preferentially capture true interactors.

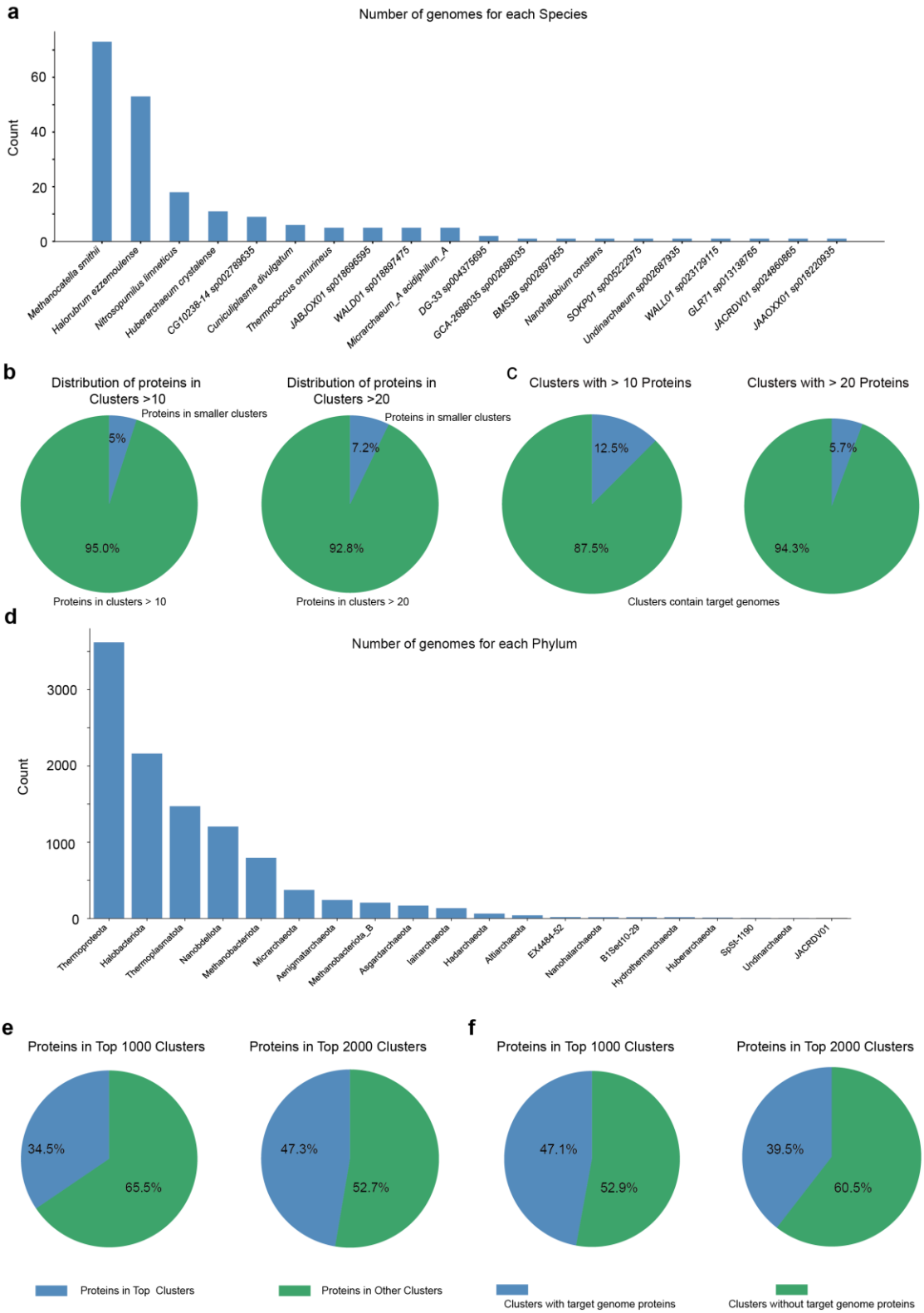

Tree scale: 1
Species Type
Archaea
Bacteria
Pathogens
Bacteria GCA 008933615.1
Bacteria GCF 001886815.1
Bacteria GCA 016783505.1
Bacteria GCA 020637395.1
Bacteria GCA 001304035.1
Bacteria GCA 020633655.1
Bacteria GCA 002085035.1
Bacteria GCA 011046415.1
Bacteria GCA 018674655.1
Bacteria GCF 009494135.1
Bacteria GCA 003477045.1
Bacteria GCA 022564135.1
Bacteria GCA 017997825.1
Bacteria GCA 002737115.1
Bacteria GCA 002449765.1
Bacteria GCA 002084765.1
Bacteria GCA 020348925.1
Bacteria GCA 009837395.1
Bacteria GCA 018672475.1
Bacteria GCA 002686955.1
Bacteria GCA 001777915.1
Bacteria GCA 001303705.1
Bacteria GCA 016703265.1
Bacteria GCA 016867725.1
Bacteria GCA 016215625.1
Bacteria GCA 023135015.1
Bacteria GCA 023132135.1
Bacteria GCA 016867715.1
Bacteria GCA 002686915.1
Bacteria GCA 003645115.1
Bacteria GCA 024653355.1
Bacteria GCA 023229285.1
Bacteria GCA 016927185.1
Bacteria GCA 016936235.1
Bacteria GCA 009839745.1
Bacteria GCA 011048995.1
Bacteria GCA 003647755.1
Bacteria GCA 003647715.1
Bacteria GCA 003220315.1
Bacteria GCF 025264625.1
Bacteria GCA 016212085.1
Bacteria GCA 002254365.1
Bacteria GCF 000299235.1
Bacteria GCA 002897695.1
Bacteria GCA 001873285.1
Bacteria GCA 001873295.1
Bacteria GCA 016931515.1
Bacteria GCA 002748425.1
Bacteria GCA 001790855.1
Bacteria GCA 023512255.1
Bacteria GCA 018669005.1
Bacteria GCA 018653085.1
Bacteria GCA 016193065.1
Bacteria GCA 001803565.1
Bacteria GCA 022562795.1
Bacteria GCF 900079095.1
Bacteria GCA 016192455.1
Bacteria GCA 023489065.1
Bacteria GCA 001797445.1
Bacteria GCA 028702995.1
Bacteria GCA 016202395.1
Bacteria GCA 001798135.1
Bacteria GCA 001799045.1
Bacteria GCA 016713535.1
Bacteria GCA 001797815.1
Bacteria GCA 012959805.1
Bacteria GCF 012843875.1
Bacteria GCA 002869005.1
Bacteria GCA 003695505.1
Bacteria GCA 020434085.1
Bacteria GCF 000196175.1
Bacteria GCA 020434345.1
Bacteria GCA 003598065.1
Bacteria GCA 017539605.1
Bacteria GCA 019637195.1
Bacteria GCA 020633895.1
Bacteria GCA 003648675.1
Bacteria GCA 016180645.1
Bacteria GCF 000194135.1
Bacteria GCF 000177635.2
Bacteria GCF 000487995.2
Bacteria GCA 002167555.2
Helicobacter pylori
Campylobacter jejuni
Bacteria GCA 012963765.1
Rickettsia prowazekii
Brucella melitensis
Bordetella pertussis
Neisseria meningitidis
Neisseria gonorrhoeae
Legionella pneumophila
Pseudomonas aeruginosa
Acinetobacter baumannii
Vibrio cholerae
Vibrio parahaemolyticus
Haemophilus influenzae
Yersinia pestis
Klebsiella pneumoniae
Salmonella enterica
Escherichia coli
Shigella dysenteriae
Bacteria GCA 013359265.1
Bacteria GCA 016215705.1
Bacteria GCA 026416715.1
Bacteria GCA 020848155.1
Bacteria GCA 001567115.1
Bacteria GCA 001730085.1
Bacteria GCA 002701385.1
Bacteria GCA 017640805.1
Bacteria GCF 000020225.1
Chlamydia trachomatis
Bacteria GCA 016932965.1
Bacteria GCA 030765805.1
Bacteria GCA 003599815.1
Bacteria GCA 013203085.1
Bacteria GCA 016178055.1
Bacteria GCA 012729785.1
Bacteria GCA 016933315.1
Bacteria GCA 030765405.1
Bacteria GCA 001873945.1
Bacteria GCA 016209155.1
Bacteria GCA 001871075.1
Bacteria GCA 031259185.1
Bacteria GCA 028712895.1
Bacteria GCA 017937435.1
Bacteria GCA 001778375.1
Bacteria GCA 013335425.1
Bacteria GCA 018000835.1
Bacteria GCA 018903055.1
Bacteria GCA 005223095.1
Bacteria GCA 018819665.1
Bacteria GCA 003574845.1
Bacteria GCA 018812925.1
Bacteria GCA 018812485.1
Bacteria GCA 016783345.1
Bacteria GCA 023818645.1
Bacteria GCA 018897555.1
Bacteria GCA 002780065.1
Bacteria GCA 001784175.1
Bacteria GCA 017999075.1
Leptospira interrogans
Treponema pallidum
Bacteria GCA 016699125.1
Bacteria GCA 013791745.1
Bacteria GCA 021162505.1
Bacteria GCA 025057615.1
Bacteria GCA 000999465.1
Mycoplasmoides pneumoniae
Bacteria GCA 002716355.1
Archaea GCF 009617975.1
Archaea GCA 002789275.1
Archaea GCA 002114285.1
Archaea GCA 002897955.1
Archaea GCA 013154095.1
Archaea GCF 000018365.1
Archaea GCF 000016525.1
Archaea GCA 024860865.1
Archaea GCA 004375695.1
Archaea GCF 900083515.1
Archaea GCA 005222975.1
Archaea GCA 022117465.1
Archaea GCA 018897475.1
Archaea GCA 023128115.1
Archaea GCA 002688035.1
Archaea GCF 016806735.1
Archaea GCA 002687935.1
Archaea GCA 002789635.1
Archaea GCA 013138765.1
Archaea GCA 018696595.1
Archaea GCA 018220935.1
Bacteria GCF 000158275.2
Bacteria GCA 017449135.1
Bacteria GCA 023228995.1
Bacteria GCA 001791795.1
Bacteria GCA 016190005.1
Bacteria GCA 002069885.1
Bacteria GCA 026394215.1
Bacteria GCA 022646745.1
Bacteria GCA 001771545.1
Bacteria GCA 014360465.1
Bacteria GCF 900465355.1
Bacteria GCF 002752635.1
Bacteria GCF 000381045.1
Bacteria GCA 011375045.1
Bacteria GCA 002898315.1
Bacteria GCF 000021645.1
Bacteria GCA 012839925.1
Bacteria GCF 003057965.1
Bacteria GCA 001871575.1
Bacteria GCA 024655865.1
Bacteria GCF 000020945.1
Bacteria GCA 018725125.1
Bacteria GCA 018725175.1
Bacteria GCA 030263615.1
Bacteria GCA 005882115.1
Bacteria GCF 002608075.1
Bacteria GCF 000724625.1
Actinomyces israelii
Corynebacterium diphtheriae
Mycobacterium leprae
Mycobacterium tuberculosis
Bacteria GCA 013043945.1
Bacteria GCA 001917855.1
Bacteria GCA 028685295.1
Bacteria GCA 012840065.1
Bacteria GCA 016190305.1
Bacteria GCF 019668485.1
Bacteria GCA 012522585.1
Clostridioides difficile
Clostridium perfringens
Clostridium botulinum
Clostridium tetani
Staphylococcus aureus
Listeria monocytogenes
Bacillus anthracis
Enterococcus faecium
Streptococcus pyogenes
Bacteria GCF 001457635.1
Streptococcus pneumoniae
Protein Counts

Supplementary Figure 8. **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 colouring denotes domain affiliation (green, Archaea; blue, Bacteria); pathogenic taxa are flagged with orange stars. The protein number was shown on the right side.



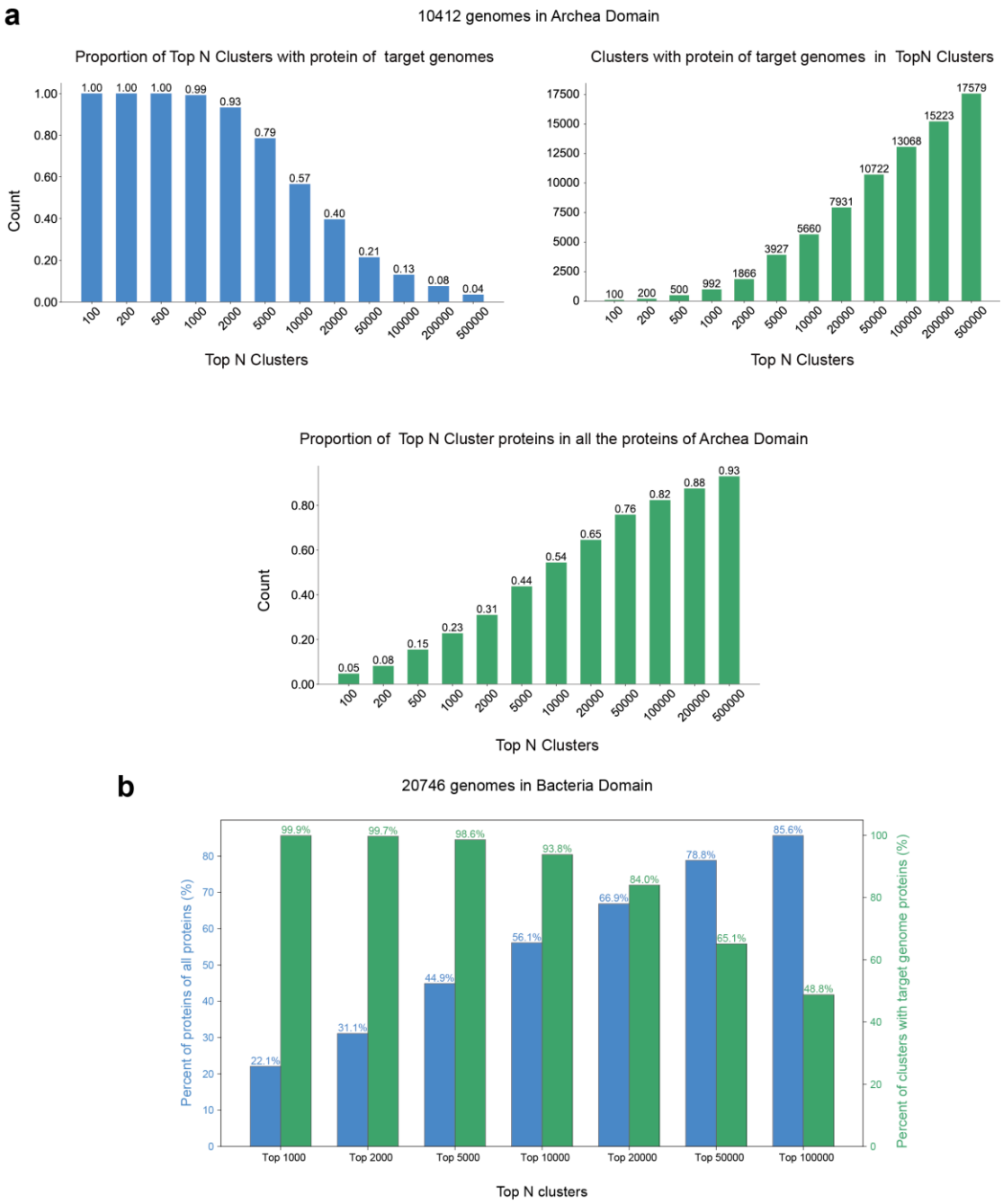
Supplementary Figure 9. Protein-cluster coverage for three representative bacterial genomes.

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.



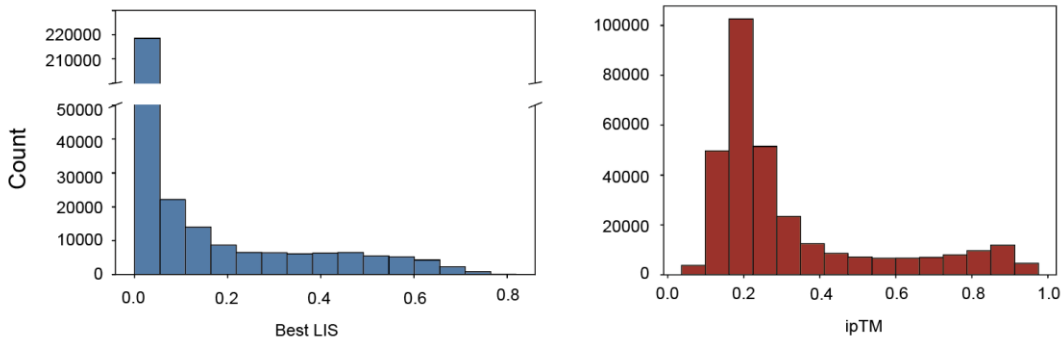
Supplementary Figure 10. Sequence-cluster coverage across taxonomic scales of Archeal domain. a Number of publicly available genomes for each species included in this study (top-ranked species on the x-axis). **b** Species-level clustering of *Methanocatella smithii*: pie charts show the share of proteins 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.

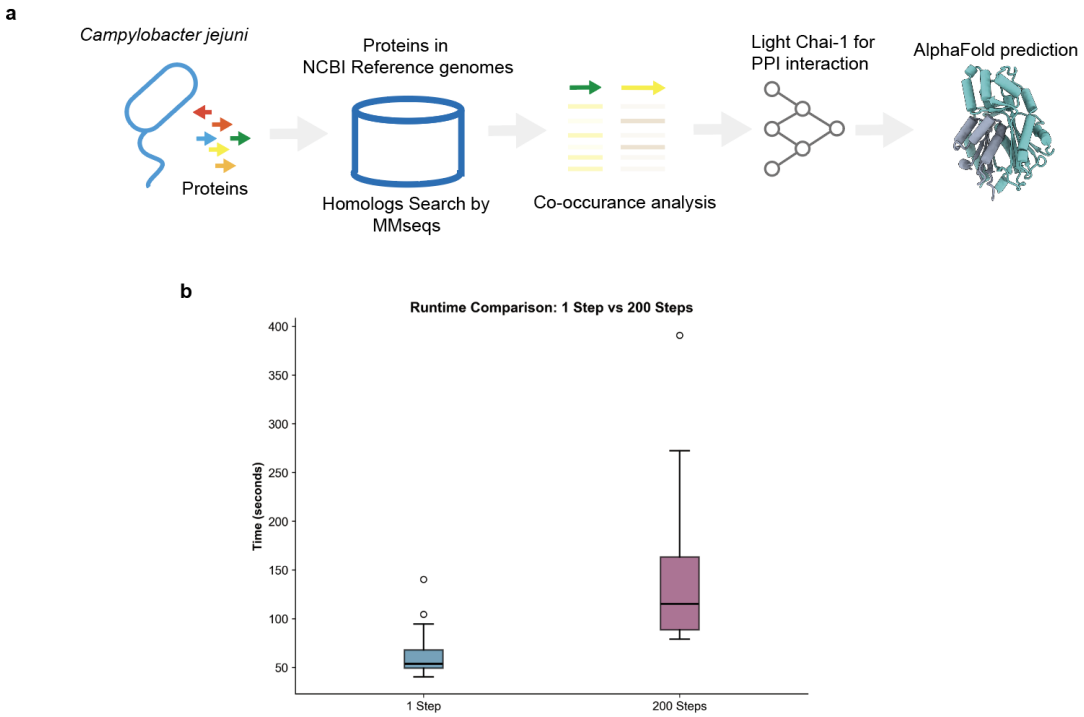


Supplementary Figure 11. Coverage of representative genome proteins at the domain level.

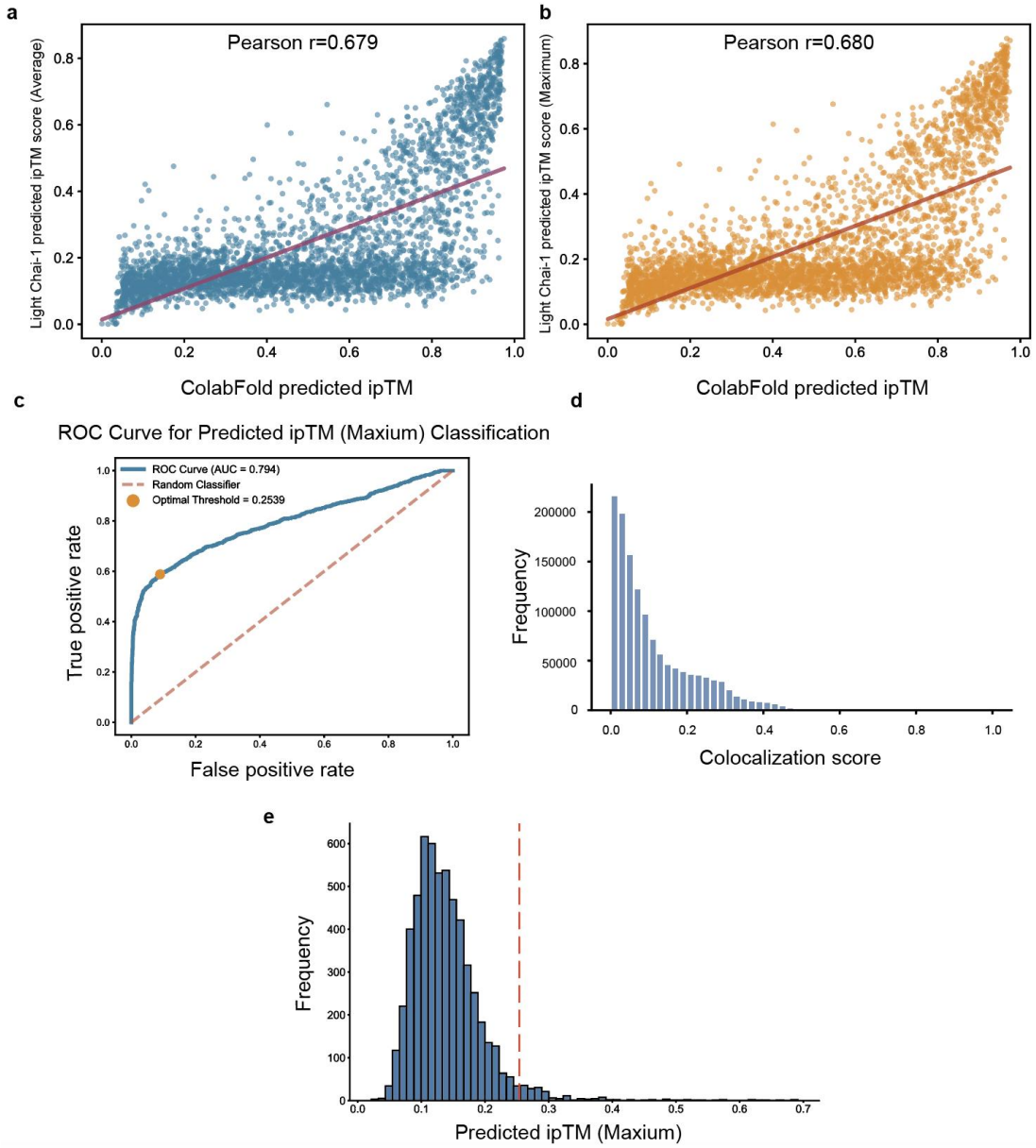
a Archaea (10 412 genomes). Top-left: fraction of the top N clusters that contain at least one protein from the representative genomes (blue).Top-right: the number of clusters containing at least one protein from the representative genomes (orange). Bottom: proportion of all archaeal proteins accounted 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.



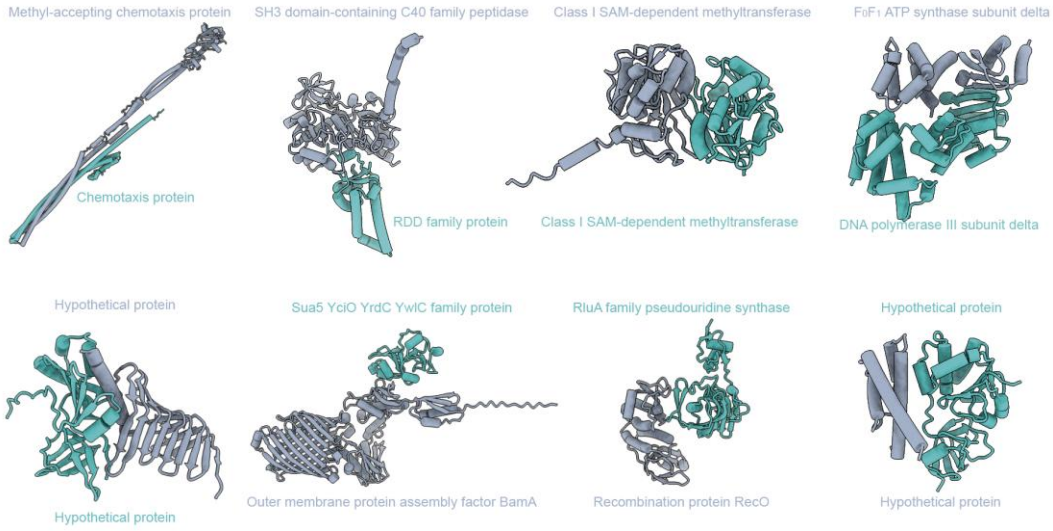
Supplementary Figure 12. Distribution of Best LIS and ipTM Scores for protein pairs in 188 representative bacterial and archaeal species. The histogram shows the distribution of ipTM and Best LIS scores for predicted archaeal and bacterial protein pairs. Out of a total of 313,348 predictions, 47,668 pairs (15.21%) showed score exceeding the best LIS above 0.203 and best LIA above 3432 thresholds.



Supplementary Figure 13. The pipeline to find protein pairs separated located in genome. a 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; pairs with a co-occurrence score ≥ 0.5 and lengths compatible with a single NVIDIA RFX3090 GPU run were retained. **b** Comparison of the runtime for Chai-1 by using ESM with 1 or 200 steps in diffusion module.

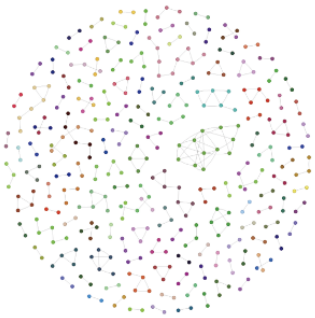


Supplementary Figure 14. Combining genomic co-occurrence and light Chai-1 for identifying protein pairs that are located far apart in the genomic sequence in *Campylobacter jejuni*. a and b Correlation between Light Chai-1 and ColabFold ipTM scores. Scatterplots show Pearson correlations using the mean (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.



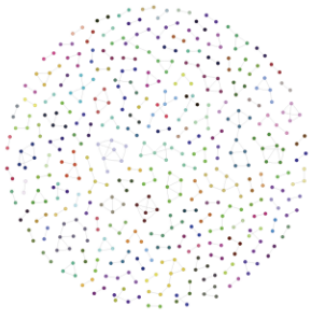
Supplementary Figure 15. Eight cases of high confidence predicted protein complexes.

Mycobacterium leprae

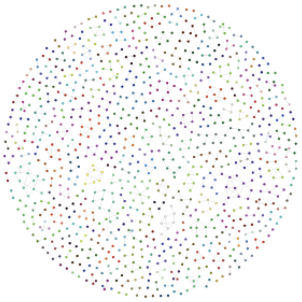


Neisseria gonorrhoeae

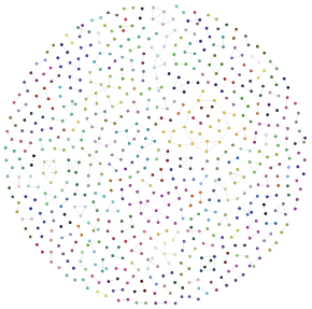
Haemophilus influenzae



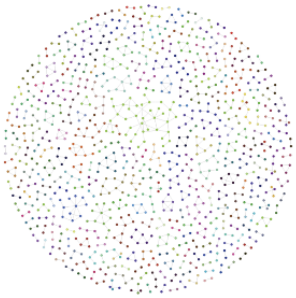
Vibrio cholerae



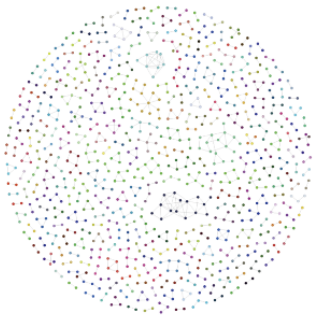
Enterococcus faecium



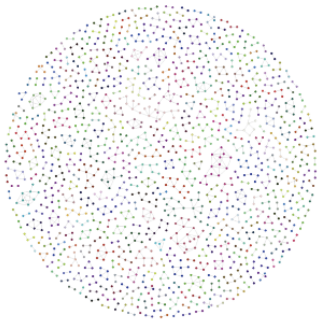
Clostridium botulinum



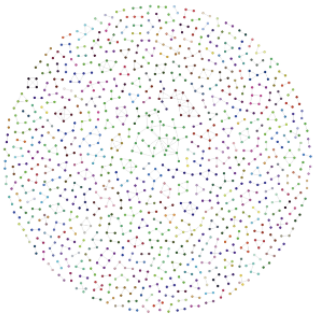
Actinomyces israelii



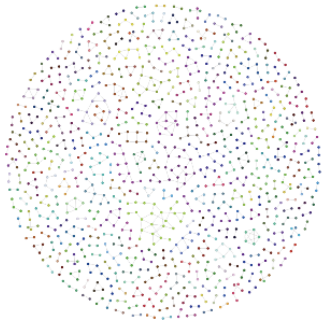
Salmonella enterica



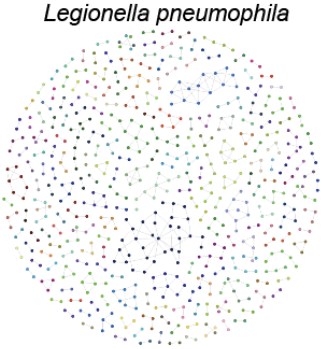
Bacillus anthracis



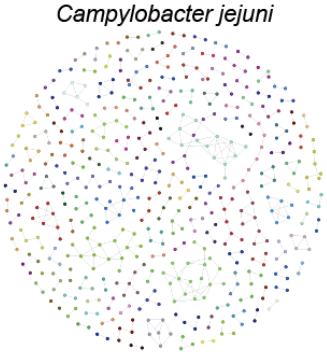
Clostridioides difficile

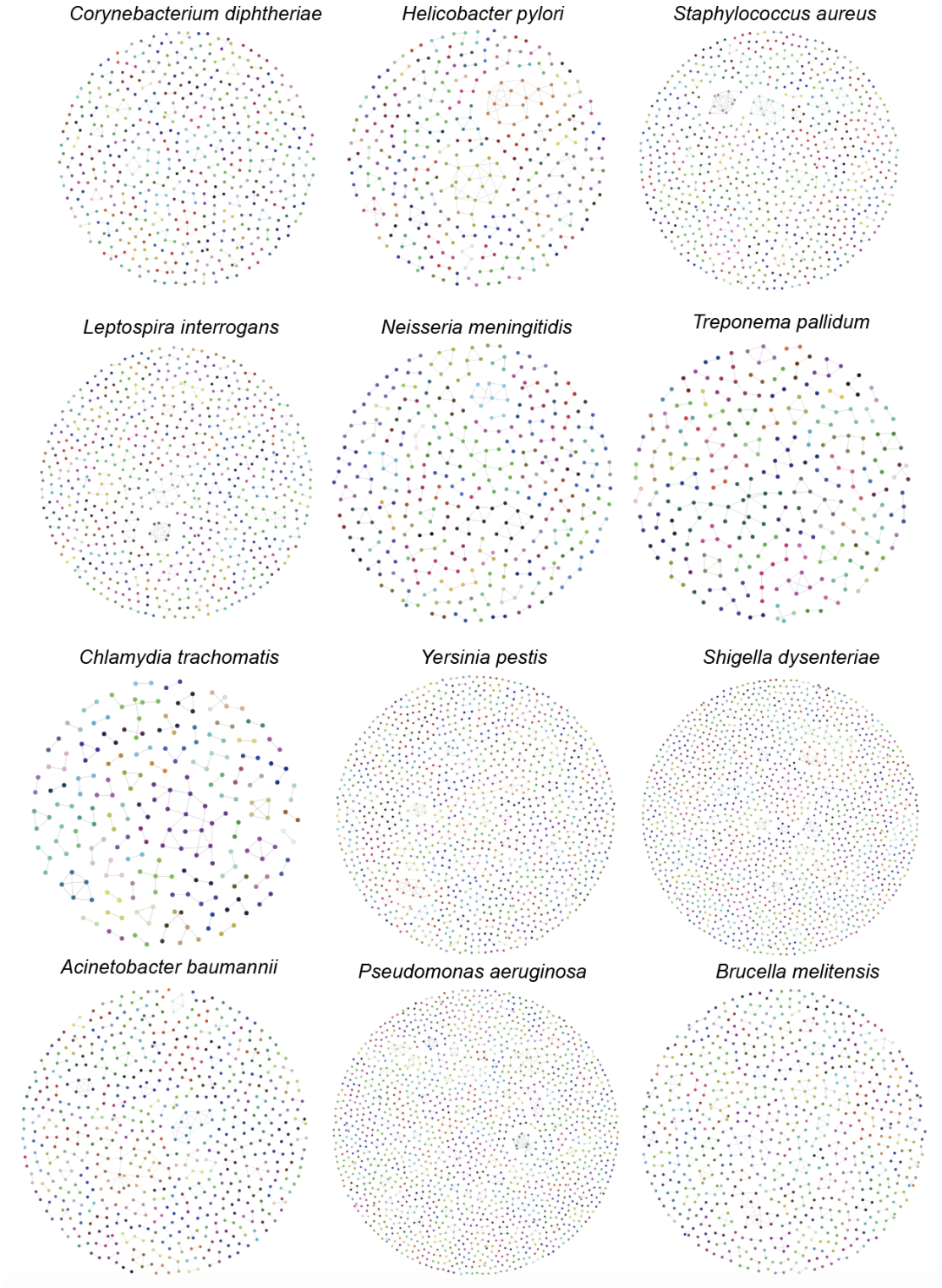


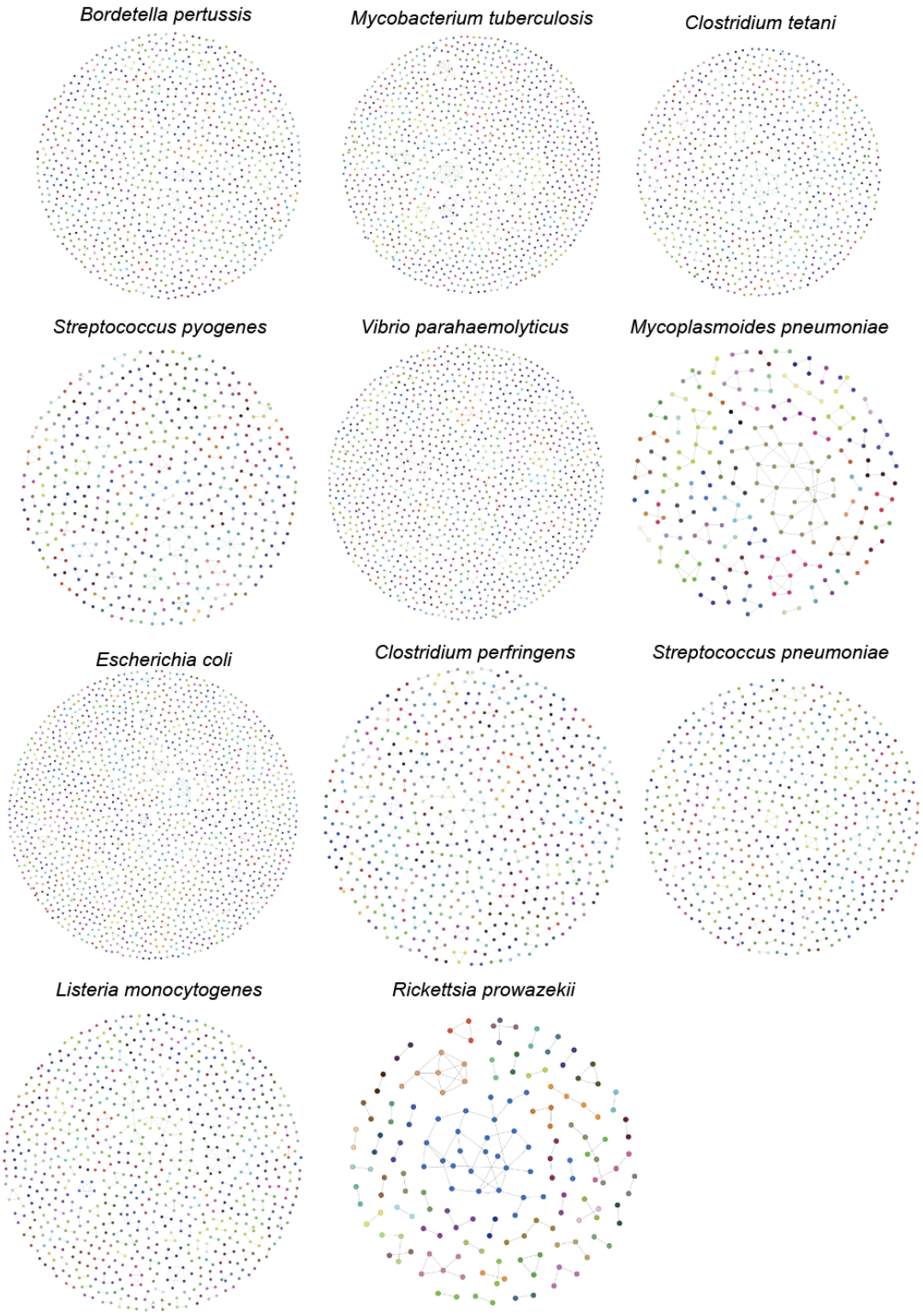
Legionella pneumophila



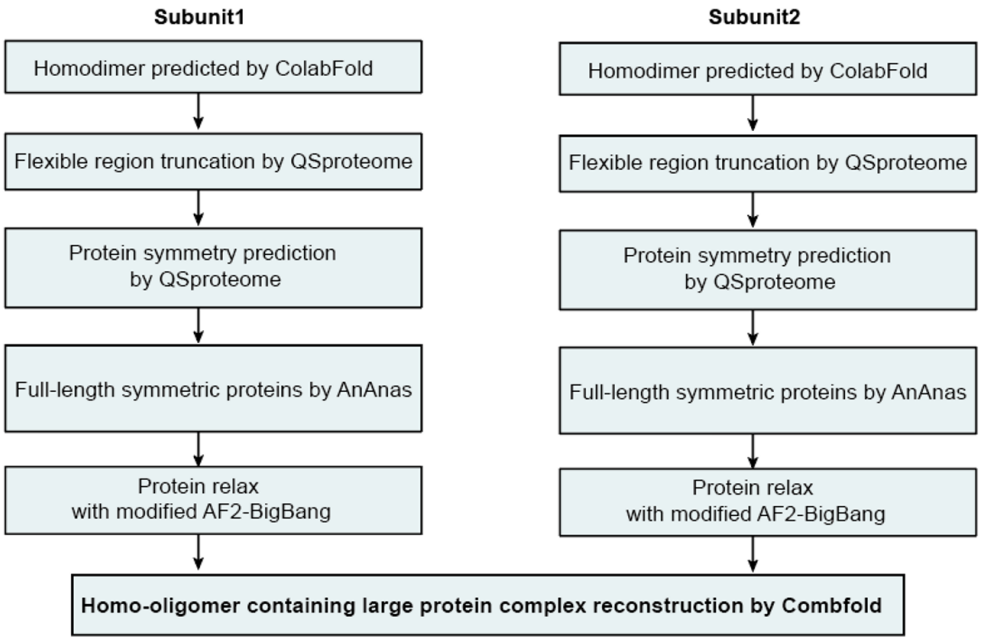
Campylobacter jejuni



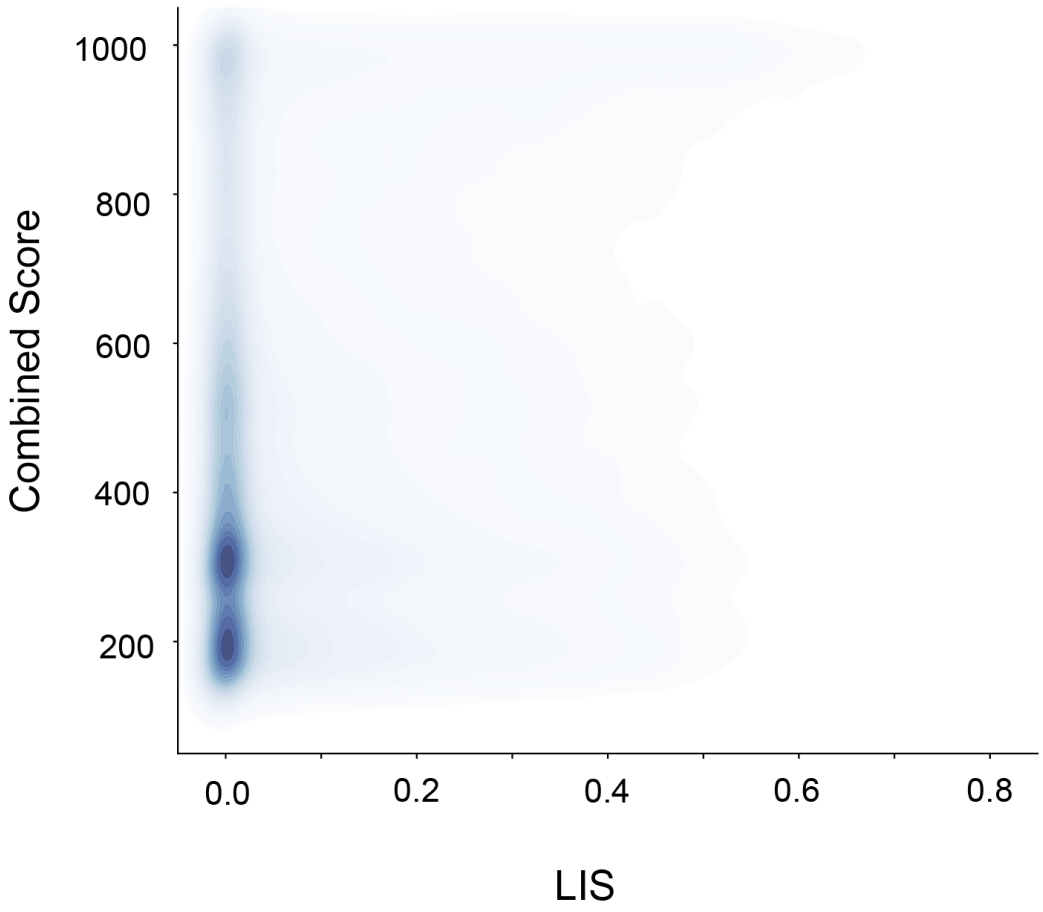

Corynebacterium diphtheriae
Helicobacter pylori
Staphylococcus aureus
Leptospira interrogans
Neisseria meningitidis
Treponema pallidum
Chlamydia trachomatis
Yersinia pestis
Shigella dysenteriae
Acinetobacter baumannii
Pseudomonas aeruginosa
Brucella melitensis



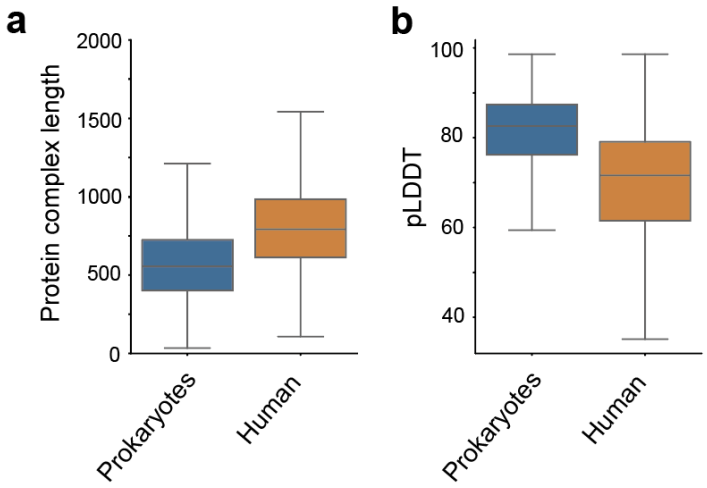
Supplementary Figure 16. Network of protein complexes in 35 species of pathogenic bacteria. This figure presents the network of protein complexes in 35 pathogenic bacteria, excluding Klebsiella pneumoniae (see Fig. 2a). Each node represents a protein, and nodes within the same connected subgraph are colored the same. edges indicate score above Best LIS of 0.203 and Best LIA of 3432. The name of the pathogenic bacteria to which each protein community graph belongs is labeled above each graph.



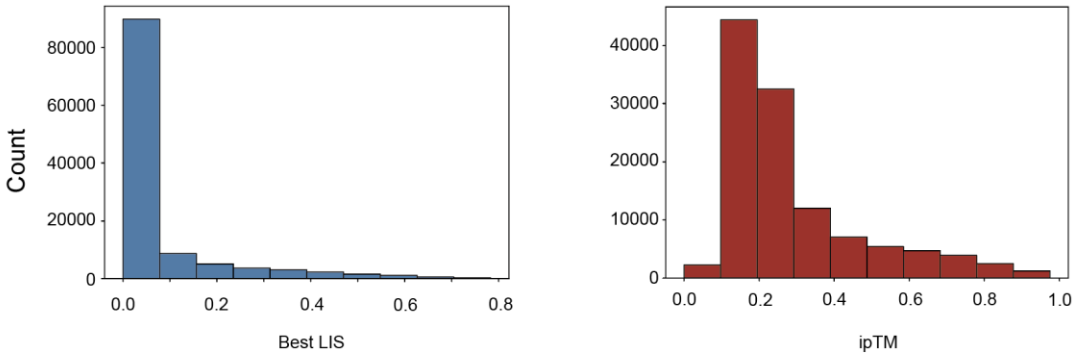
Supplementary Figure 17. Advanced Reconstruction of Protein Complexes. We utilized QSproteome for symmetry prediction and subsequently relaxed the protein structures using AlphaFold2-bigbang. Symmetrical homo-oligomers were assembled with CombFold.



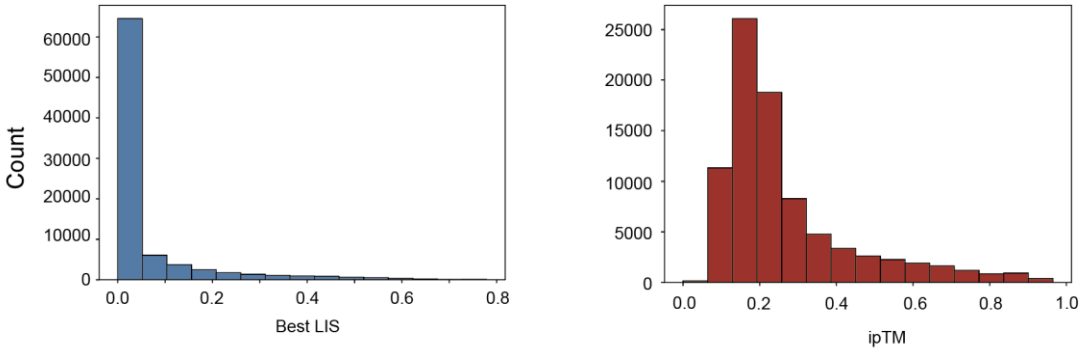
Supplementary Figure 18. Joint distribution of LIS and integrated interaction confidence scores. Kernel-density estimation (KDE) was used to visualise 180,612 protein–protein interaction pairs for which both the LIS (x-axis) and a STRING-derived combined confidence score (y-axis) were available. Density was estimated with a Gaussian kernel (band-width adjustment = 0.8) and is shown as filled contours (30 levels; darker blue indicates higher point density). The majority of interactions cluster at very low Best LIS values (< 0.05) and moderate combined scores (150 – 400), whereas high-confidence STRING interactions (scores > 700) span the entire LIS range. Overall, the two measures display only a weak but statistically significant association (Pearson $r = 0.155$, $p < 0.001$).



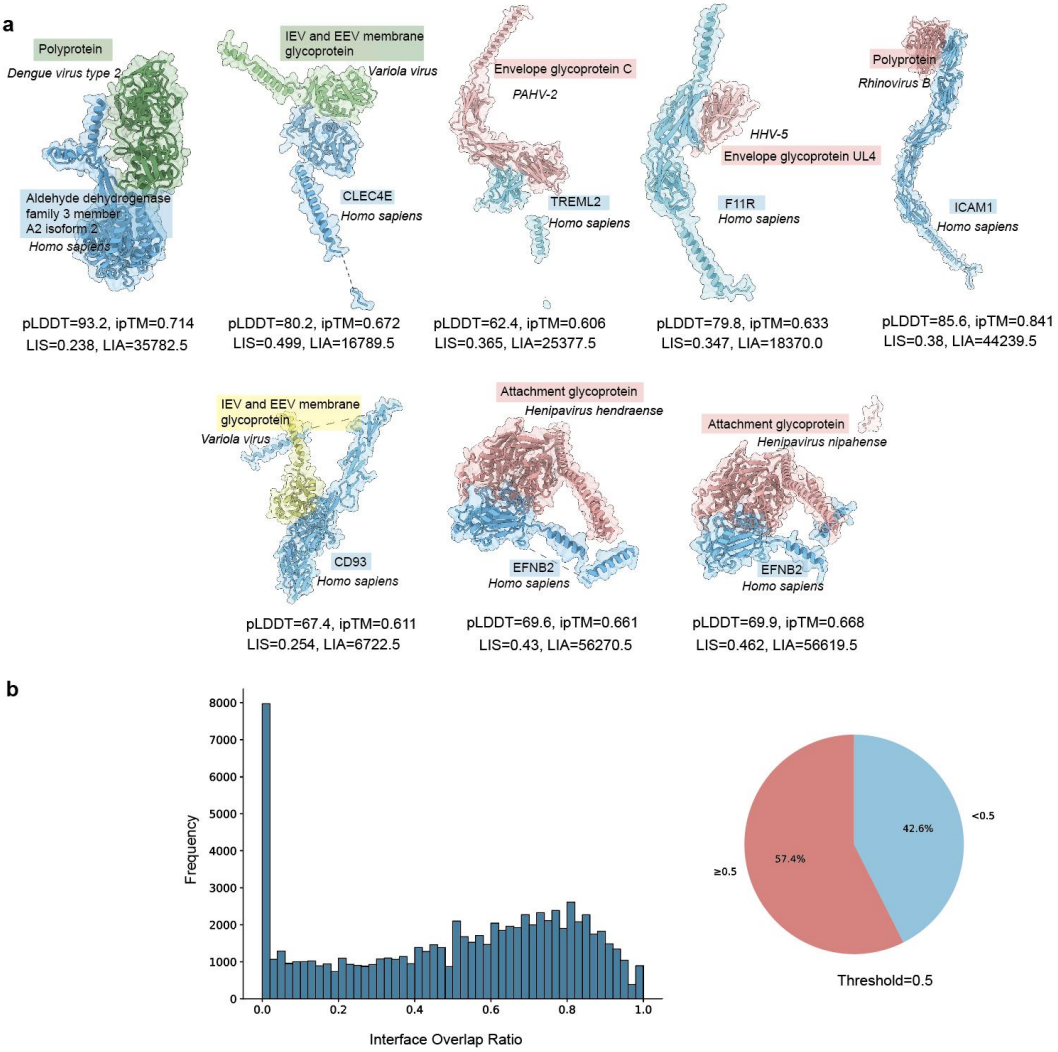
Supplementary Figure 19. Length and predicted reliability comparison between prokaryotes and human. a Length analysis of subunits of potential PPIs in human or prokaryotic species. **b** The pLDDT score distribution human or prokaryotic species.



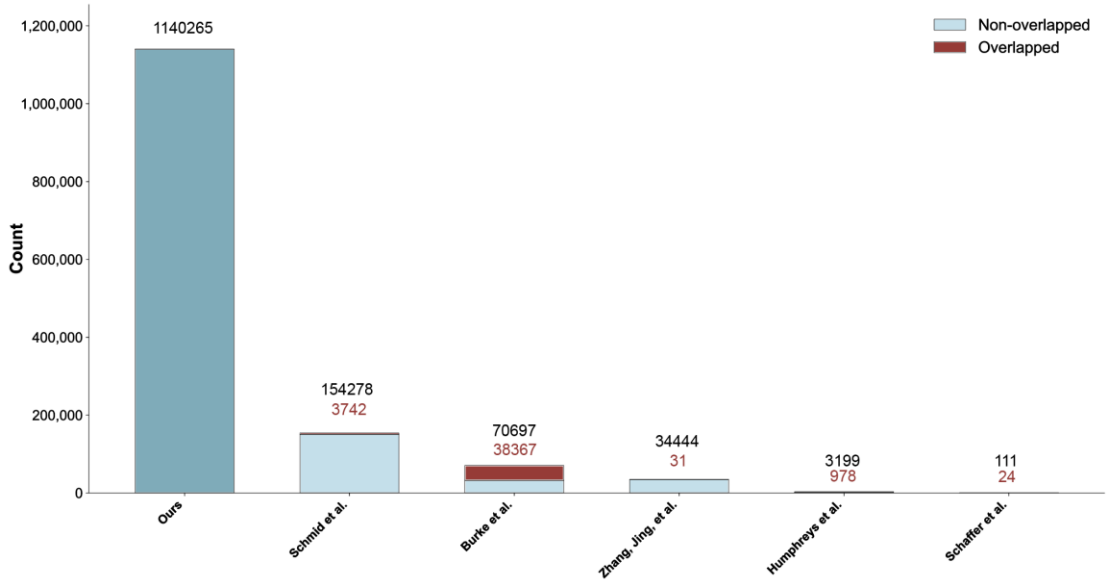
Supplementary Figure 20. Distribution of Best LIS and ipTM Scores for predicted mice protein pairs. The histogram shows the distribution of Best LIS and ipTM scores for predicted mice protein pairs. Out of a total of 116,060 predictions, 12,778 pairs were identified as high-confidence predictions, meeting the defined thresholds (Best LIS >= 0.203 and Best LIA > =3432).



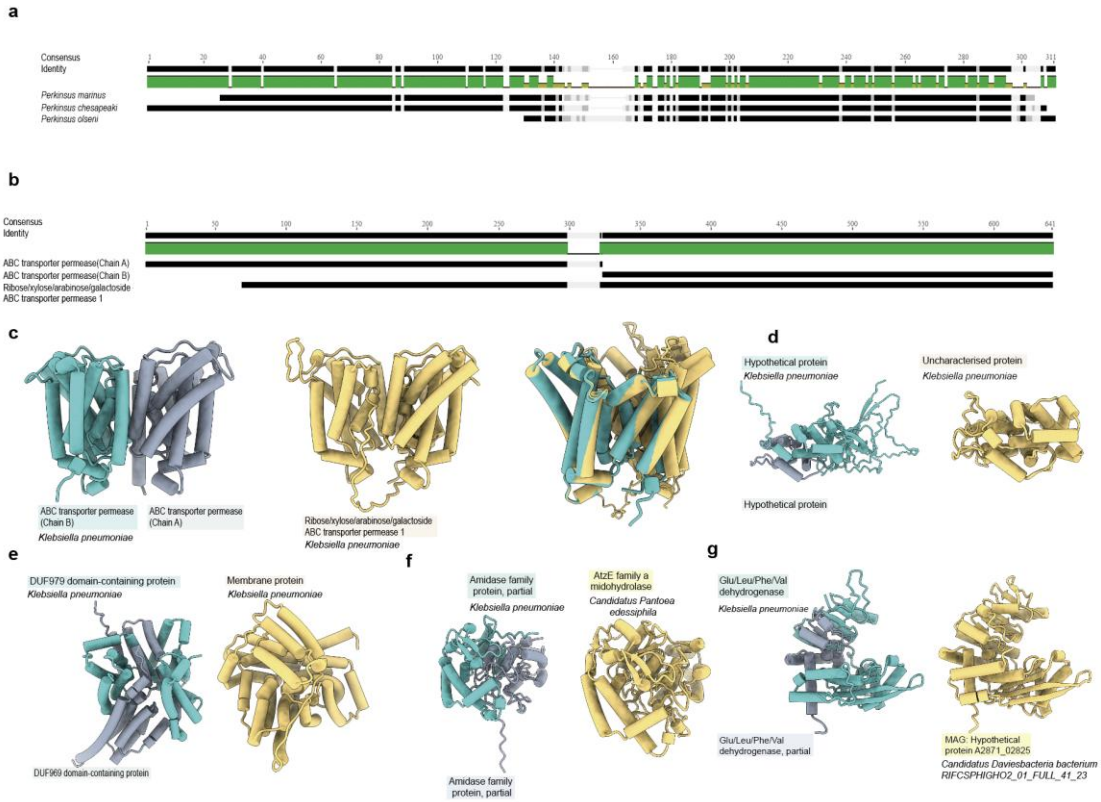
Supplementary Figure 21. Distribution of Best LIS and ipTM Scores for predicted *Arabidopsis Thaliana* protein pairs. Out of a total of 84,498 predictions, 7034 pairs were identified as high-confidence predictions, meeting the defined thresholds (Best LIS >= 0.203 and Best LIA > =3432).



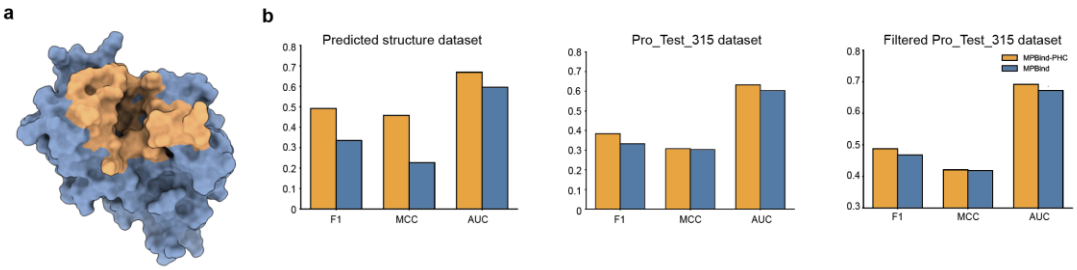
Supplementary Figure 22. Analysis of interactions between viral and human proteins. a Prediction results of interactions between 8 different human receptor proteins and viral proteins, with regions having pLDDT values less than 40 not shown. **b** The proportion of overlapping surface binding sites when the viral protein was predicted to interact with a single subunit of the human dimer.



Supplementary Figure 23. Comparison of the number of predicted large-scale protein complexes of multiple studies[2-6]. The total number of predicted complexes (black) and the number overlapping with our study (red) are shown above each bar.



Supplementary Figure 24. Sequences and structures alignment of monomers and protein complexes. a Sequence alignment by Geneious Prime for homologs of monomer protein in Fig. 5d. **b and c** Sequence and Structure alignment of ABC transporter permease complex and monomers in *Klebsienlla pneumoniae*. **d-g** Comparison of monomer and protein complex.



Supplementary Figure 25. Applications of deep learning models leveraging the large-scale protein complex structure dataset for protein binding site prediction. a Surface residue 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.

Supplementary Table 1 Information of 36 species of pathogenic bacteria

Pathogen Bacteria Name	Genbank Assembly Accession	Protein Counts	Predicted Operon Counts
Salmonella enterica	GCF_000006945.2	4554	882
Staphylococcus aureus	GCF_000013425.1	2892	562
Klebsiella pneumoniae	GCF_000240185.1	5779	1115
Streptococcus pyogenes	GCF_900475035.1	1630	351
Vibrio cholerae	GCF_008369605.1	3826	689
Streptococcus pneumoniae	GCF_001457635.1	2192	476
Neisseria gonorrhoeae	GCF_013030075.1	2166	473
Pseudomonas aeruginosa	GCF_000006765.1	5571	1124
Enterococcus faecium	GCF_009734005.1	2936	549
Neisseria meningitidis	GCF_022869645.1	2421	473
Mycobacterium tuberculosis	GCF_000195955.2	4031	925
Legionella pneumophila	GCF_001941585.1	3071	556
Clostridium perfringens	GCF_016027375.1	2842	469
Campylobacter jejuni	GCF_000009085.1	1643	448
Clostridioides difficile	GCF_018885085.1	3591	668
Clostridium botulinum	GCF_000063585.1	3671	667
Escherichia coli	GCF_008727195.1	5598	1010
Shigella dysenteriae	GCF_022354085.1	4888	919
Clostridium tetani	GCF_000007625.1	2788	514
Mycobacterium leprae	GCF_003253775.1	2923	371
Brucella melitensis	GCF_000007125.1	3170	613
Rickettsia prowazekii	GCF_000277165.1	852	187
Leptospira interrogans	GCF_002073495.2	3742	683
Chlamydia trachomatis	GCF_000008725.1	887	198
Mycoplasmoides pneumoniae	GCF_900660465.1	741	137
Actinomyces israelii	GCF_030253495.1	3243	655
Acinetobacter baumannii	GCF_008632635.1	3642	675
Haemophilus influenzae	GCF_000931575.1	1752	391
Bacillus anthracis	GCF_000008445.1	5627	991
Corynebacterium diphtheriae	GCF_001457455.1	2195	437
Treponema pallidum	GCF_000246755.1	980	197
Helicobacter pylori	GCF_017821535.1	1513	306
Vibrio parahaemolyticus	GCF_000196095.1	4548	807
Yersinia pestis	GCF_000222975.1	4144	734
Listeria monocytogenes	GCF_000196035.1	2867	586
Bordetella pertussis	GCF_004008975.1	3907	771

Supplementary Table 2 Information for the used extracellular domain sequences (3xFlag shown in red, 3xHA tag shown in blue)

E3-HA

MERPLKVVVEAGHNVTLPHLSGSHQTGHVTWLVETSDYGSASPDNFIFSGQKLCQFTDRTMVWPYYNLH
FNCENYDLNLFWLKVENSAIYNVKNTVNASETNIYYDLRVVQIFPPKCIITSKYLTNDYCHITINCTNSDYP
NKVVFNNVSRWYYGYGKGSPTLPNYFITNFNVSGITKSFNHTYPFNELCYPYDVPDYAYPYDVPDYAYPY
DVPDYA

UL44-HA

MGSAPAPFRCKRPDVVALFGGRVAMGCRVPRPTADFRLRIWRVAAPRAEAGRPEPAEPGAVLVNVTAPPPD
GELVYDSAPGATGARVRWAEGAGPDARPRVYALEGTFPTLRLVIQELTLARQGWYFWARGPADQPLRYGT
WTRVRMLRRPSLSIRPHAVLEGEPFGATCVAANYYPGDRAAFRWFEGGDEVAAPDRVRTRVDAQRDGFS
ATSTLTSEARAGLAPPRNLTCEFTWHRDAVSFSRRNATGAPTVLPRPTIEMEFGGDEAVCTAACVPEGVAL
QWLLGADPAPAGNAVASGGPCPGRPGLARLRSALPLSREHSEYTCRLVGYPATVPVLEHHGRREPPPRYPY
DVPDYAYPYDVPDYAYPYDVPDYA

HLA-DRA-Flag

MIKEEHVIIQAEFYLNPDQSGEFMFDFDGDEIFHVDMAKKETVWRLEEFGRFASFEAQGALANIAVDKAN
LEIMTKRSNYTPITNVPPEVTVLTNSPVELREPNVLICFIDKFTPPVVNVTWLRNGKPVTTGVSETVFLPRE
DHLFRKFHYLPFLPSTEDVYDCRVEHWGLDEPLLKHWEFDGSDYKDHDGDYKDHDIDYKDDDDK

CD4-Flag

MDYKDHDGDYKDHDIDYKDDDDKKKVVLGKKGDTVELTCTASQKKSIQFHWKNSNQIKILGNQGSFLT
KGPSKLNDRADSRRSLWDQGNFPLIIKNLKIEDSDTYICEVEDQKEEVQLLVFGLTANSDTHLLQGQSLTLT
LESPPGSSPSVQCRSPRGKNIQGGKTLSVSQLELQDSGTWTCTVLQNQKKVEFKIDIVVLAFQKASSIVYK
KEGEQVEFSFPLAFTVEKLTGSGELWWQAERASSSKSWITFDLKNKEVSVKRVTQDPKLQMGKKLPLHLT
LPQALPQYAGSGNLTLALEAKTGKLHQEVNLVVMRATQLQKNLTCEVWGPTSPKLMLSLKLENKEAKVS
KREKAVWVLNPEAGMWQCLLSDSGQVLLESNIKVLPTWSTPVQP

Supplementary References:

1. Terlouw, B.R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res* **51**, D603-D610 (2023).
2. Burke, D.F. *et al.* Towards a structurally resolved human protein interaction network. *Nature Structural & Molecular Biology* **30**, 216-225 (2023).
3. Humphreys, I.R. *et al.* Protein interactions in human pathogens revealed through deep learning. *Nature Microbiology* **9**, 2642-2652 (2024).
4. Schaffer, L.V. *et al.* Multimodal cell maps as a foundation for structural and functional genomics. *Nature* **642**, 222-231 (2025).
5. Schmid, E.W. & Walter, J.C. Predictomes, a classifier-curated database of AlphaFold-modeled protein-protein interactions. *Molecular Cell* **85**, 1216-1232.e5 (2025).
6. Zhang, J. *et al.* Computing the Human Interactome. *bioRxiv*, 2024.10.01.615885 (2024).