Deep learning and coevolution reveal proteome-wide protein-protein interactions

Ian Raymond Humphreys

A dissertation

submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

David Baker, Chair

Liangcai Gu

Sanjay Srivatsan

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Deep learning and coevolution reveal proteome-wide protein-protein interactions

Ian Raymond Humphreys

Chair of the Supervisory Committee:
Professor, David Baker
Department of Biochemistry

The total set of potential protein-protein interactions (PPI) within an organism's proteome guides a plethora of potential biological processes at an organism’s disposal. Understanding these PPIs is critical to our understanding of biological systems, however identifying interactions with high accuracy is challenging. Medium to high-throughput experimental techniques for identifying protein interactions result in high rates of false-negatives and false-positives. However, protein interactions are typically evolutionarily conserved resulting in co-varying mutations at the interface between complexes. Deep learning based protein structure prediction models capture coevolutionary information at significantly higher resolution than statistical methods and we exploit this coevolutionary signal to computationally predict protein-protein interactions with high accuracy based on gold-standard benchmarks. We create and apply bioinformatic and deep learning pipelines to rapidly predict proteome-wide protein-protein interactions in Bacteria and Eukaryotes to identify novel interactions and provide high resolution structural models to better understand their biological ramifications.

For my parents

ACKNOWLEDGEMENTS

I would like to begin by thanking my advisor, David Baker for the opportunity of a lifetime. You allowed me to pursue fascinating evolutionary questions in what may be one of the most non-natural labs in the world. Your tireless enthusiasm for science, diplomacy, and ability to steer hundreds of diverse projects, has been nothing short of inspiring.

To my mentor Qian Cong, thank you for being my teacher, my sounding board, and my strongest ally. I can't express how grateful I am to have worked with you and for your continued mentorship even after you started your own lab just several months after we began working together. I would never have been able to do this alone. For all the time you've invested and the expertise you've shared, I will be forever thankful. I'm truly honored to have been your "0th graduate student."

To Minkyung Baek, thank you for taking in the poor bioinformatician who only cared about native proteins in the land of design. Despite being the most popular person, an amazing mentor, and DL mastermind you always made time to help. Without your expertise and time, none of this would have been possible. One of my biggest regrets is that I didn't learn even more from you.

To the members of my thesis committee, Harmit Malik, Joseph Mougous, Liangcai Gu, Michelle Reniere, and Sanjay Srivatsan, thank you for your confidence in me, your guidance, and your insistence to dive deeper into the biology. Thank you to the many people who have taught me valuable skills and contributed to the work described below or to other projects I have been a part of, particularly Ivan Anishchenko, Jimin Pei, Yaxi Wang, and Jing Zhang. My research has heavily relied on the IPD's computational resources. With over 100 lab members and 200 cluster users, I'm amazed and grateful for Luki Goldschmidt, without whom I (and the rest of the IPD) would be

unable to function. Thanks also to Patrick Vecchiato, Kristina Herrera, and Zari Magness for ensuring everything runs like clockwork. Yang Hsia, thank you for the food and fun; graduate students are always on the prowl for free food, and you've certainly ensured there's an ample supply.

For introducing me to the world of research and changing the trajectory of my life and career, I would like to thank my undergraduate research advisor Jane Ishmael and the members of her lab: Xuemei Wan, Daphne Mattos, and Jeff Serrill. I'm honestly not sure what I would be doing today if Jane hadn't welcomed me into her lab. My friends outside of Seattle who understood the ups and downs of graduate school: Michelle Michelsen, Joe Kim, Courtney Armour, and Christopher Gaulke, what's life without a few group chats? To my MolES row-mates past and present, thank you for creating such a fun environment that I feel confident was the envy of many. And to everyone in the Baker lab and IPD, thank you, it's the large collective community which has made my experience so wonderful.

Next, I would like to thank some of the friends I've made during graduate school who have made: Han Altae-Tran, Will Chen, Matthias Gloegl, Naveen Jasti, Aditya Krishnakumar, Anna Lauko, Sanaa Mansoor, Amir Motmaen, Sam Pellock, Meerit Said, Thomas Schlichthaerle, Kiera Sumida, Doug Tischer, and Erin Yang, you've made Seattle a home. To my irreplaceable row-bros: day-one Baker lab buddy, Anna, and my kindred spirit, Kiera, thank you for years of constant laughter, banter, lunches, and happy hours. You two have made the time fly by. To my happenstance two-time neighbor and fellow pet enthusiast, Sanaa, I'm so glad we'll be in near-neighboring towns soon. To the future director of side chains, Meerit, your sass and strict schedule always brought levity and kept me on my toes. I feel incredibly fortunate to have shared graduate school with all of you. And to Doug, for a year of unforgettable adventures, magic, and joy—I can't wait to see what's next.

Finally, and most importantly, my family. Thank you to my parents, grandparents, siblings, aunts, uncles, and cousins, you've made me who I am today. To my dad, who has always blindly supported my academic endeavors despite thinking I was working in forestry for a few years. And to my mom, who is the most selfless person I know. I love you.

TABLE OF CONTENTS

Chapter 1. SEQUENCE, STRUCTURE, AND FUNCTION: STRUCTURAL AND FUNCTIONAL GENOMICS IN THE DEEP LEARNING ERA 1
Abstract 2
Introduction 2
Extracting protein evolutionary information from genomic sequences 3
Protein structure prediction from genomic sequences 5
Physics and homology-based structure modeling 6
First generation of deep learning for structure prediction: De novo modeling from MSAs 7
Second generation of deep learning for structure prediction: Highly accuracy prediction .9
Third Generation structure prediction networks: An atomistic view 11
Biological function of proteins using sequence and structure 13
Harnessing coevolving residues between proteins for interaction identification 13
Deep learning prediction of physically interacting proteins 15
Perspectives 16
References 19
Chapter 2. COMPUTED STRUCTURES OF CORE EUKARYOTIC PROTEIN COMPLEXES 26
Abstract 27
Main text 27
Complexes involved in DNA homologous recombination and repair 39
Complexes involved in translation and ribosome regulation 41
Complexes involving ubiquitin and small ubiquitin-like modifier (SUMO) ligases 42
Complexes involved in chromosome segregation 43
Complexes involved in molecule transport and membrane trafficking 44
GPI transamidase complex 47
Limitations of the current method 48
Conclusion 49
Methods 50
Acknowledgements 50
Funding 51
Author contributions 51
Competing interests 52
Data and materials availability 52
References 53
Supplemental Figures 61
Supplemental Tables 88
Chapter 3: PROTEIN INTERACTIONS IN HUMAN PATHOGENS REVEALED THROUGH DEEP LEARNING 93
Abstract 94
Introduction 94
Results 96

Computational pipeline for proteome-wide PPI identification 96
Experimental validation..... 101
Binary interactions 106
Multicomponent protein complexes..... 109
tRNA modification and sulfur transfers in the 2-thio modification complex of E. coli 111
A two-step nickel transfer in H. pylori urease complex 111
Interactors of the Sec translocon..... 112
Outer membrane β-barrel assembly machinery of P. aeruginosa and V. cholera 113
Discussion..... 114
Methods 116
Acknowledgments..... 116
Author contributions..... 117
Competing Interests..... 117
Data and materials availability..... 118
References..... 119
Supplemental Figures..... 125
Supplemental Tables..... 154

TABLE OF FIGURES

Chapter 1. SEQUENCE, STRUCTURE, AND FUNCTION: STRUCTURAL AND FUNCTIONAL GENOMICS IN THE DEEP LEARNING ERA 1

Chapter 2. COMPUTED STRUCTURES OF CORE EUKARYOTIC PROTEIN COMPLEXES 26

Figure 1. Evaluation of protein interaction and structure prediction accuracy 29

Figure 2. Protein complexes involved in transcription, translation, and DNA repair 35

Figure 3. Protein complexes involved in molecule transport, membrane translocation, and mitochondria 36

Figure 4. Protein complexes involved in metabolism, GPI anchor biosynthesis or including a protein of unknown function 37

Figure 5. Higher order protein complexes 38

Figure S1. Procedure for identifying the best hit to a query protein in the complete proteome of a different species 61

Figure S2. Diagram for paired multiple sequence alignments 62

Figure S3. Length (A) and depth (B) distributions of pMSA 63

Figure S4. A lighter-weight RoseTTAFold two-track (RF2t) model 64

Figure S5. Performance comparison between the two-track RoseTTAFold model trained with discontinuous crops and continuous crops 65

Figure S6. Distance matrix diagrams 66

Figure S7. Precision-recall curve for different modifications of RoseTTAFold scores evaluated on the gold-standard set (table S1) 67

Figure S8. Precision-recall curve for different modifications of RoseTTAFold scores evaluated on the literature-curated set (table S1) 68

Figure S9. Precision-recall curve for different modifications of AlphaFold scores evaluated on the gold-standard set (table S1). 69

Figure S10. Precision-recall curve for different modifications of AlphaFold scores evaluated on the literature-curated set (table S1) 70

Figure S11. Interface size (in number of amino acids) of experimentally determined complexes predicted or missed in our screen 71

Figure S12. Secondary structure of our models of protein complexes 72

Figure S13. Secondary structure of experimentally determined yeast protein complexes in the PDB 73

Figure S14. The percentage of protein complexes that are above different contact prediction accuracy cutoffs 74

Figure S15. Distribution of percentage of contacts in experimental structures that are predicted correctly in our models 75

Figure S16. Ski8 complex monomer models 76

Figure S17. Interfaces of Spo11 with Ski8 and Rec102 77

Figure S18. Model for Rad55–Rad57 binding to the Rad51–ssDNA filament 78

Figure S19. Structural analysis of the Rad33-Rad14 complex model 79

Figure S20. Structures of complexes involved in translation and ribosome regulation 80

Figure S21. Complexes involving SUMO and ubiquitin ligases 81

Figure S22. Predicted inter and intra-decamer interactions in the DASH/Dam1 complex 82
Figure S23. Predicted Vps2-Vps24 complex structure is consistent with unpublished mutagenesis data..... 83
Figure S24. Predicted GARP complex structure is consistent with negative stain data . 84
Figure S25. Predicted multivalent complex between Grh1-Sec23 85
Figure S26. Modeling of a SNARE complex consisting of Use1, Ufe1, Sec20, and Sec22 86
Figure S27. C-terminal GPI-T signal sequence recognition tunnel suggested by complex model of GPI transamidase 87
Chapter 3: PROTEIN INTERACTIONS IN HUMAN PATHOGENS REVEALED THROUGH DEEP LEARNING 93
Figure 1: Protein-protein interaction identification by coevolution and deep learning methods 97
Figure 2: Experimental validation of selected protein-protein interactions..... 103
Figure 3: Computed models of binary protein complexes 107
Figure 4: Computed models for multi-component protein complexes 110
Figure S1: Flowchart of reciprocal best hits filtering criterion for complete proteomes. 125
Figure S2: Representative distribution of monomeric MSA depth..... 126
Figure S3: Paired multiple sequence alignment schematic 127
Figure S4: Distribution of pMSA depth 128
Figure S5: Distribution of AlphaFold monomer model average pLDDT 129
Figure S6: PPI screening methodology performance 130
Figure S7: AlphaFold-multimer distance vs ipTM for PPI identification..... 131
Figure S8: Predicted interactions by genomic distance 132
Figure S9: Predicted unique interactions by genomic distance 133
Figure S10: Normalized fraction of predicted interactions by genomic distance..... 134
Figure S11: Experimentally validated pairs, models, and metadata..... 135
Figure S12: β-galactosidase activity of validated pairs by bacterial-two hybrid 136
Figure S13: PtsH-PtsN negative Co-IP pulldown..... 137
Figure S14: Uncropped western blot images for Figure 2: I 138
Figure S15: Uncropped western blot images for Figure 2: II 139
Figure S16: Uncropped western blot images for Figure 2: III 140
Figure S17: Uncropped western blot images for Figure 2: IV..... 141
Figure S18: Uncropped western blot images for Figure 2: V..... 142
Figure S19: Glucose-6-phosphate 1-dehydrogenase and OPXX cycle protein 143
Figure S20: tRNA 2-thiouridine synthesizing complex (Tus) and MnmA trimers 144
Figure S21: tRNA 2-thiouridine synthesizing complex (Tus) 145
Figure S22: Urease oligomeric assembly generation 146
Figure S23: Urease trimeric interactions..... 147
Figure S24: Sec translocon orthologous PDB validation and PpiD 148
Figure S25: Sec translocon interactions with CrgA..... 149

Figure S26: BAM complex orthologous PDB validation 150
Figure S27: BAM complex and SurA interactions 151
Figure S28: BepA putative orthologue identification and Bam/SurA interaction 152
Figure S29: Folding of TolC by BAM complex 153

TABLE OF TABLES

Chapter 2. COMPUTED STRUCTURES OF CORE EUKARYOTIC PROTEIN COMPLEXES 26
Table S1. Datasets we obtained from Yeast Interactome Database 88
Table S2: Extended annotations for modeled PPIs in Fig. 2. 89
Table S3: Extended annotations for modeled PPIs from Fig. 3. 90
Table S4: Extended annotations for modeled PPIs in Fig. 4. 91
Table S5: Extended annotations for modeled PPIs in Fig. 5 91
Chapter 3: PROTEIN INTERACTIONS IN HUMAN PATHOGENS REVEALED THROUGH DEEP LEARNING 93
Table S1: Proteome strains and Uniprot accessions 154
Table S2: Statistics for pathogens in our study 154
Table S3: RoseTTAFold2-Lite training setup 156
Table S4: Recall of filtering pipeline 157
Table S5: Predicted interactions in STRING 158
Table S6: RF2-Lite pilot-set and extended-set pairs by pathogen 159
Table S7: Metadata of experimentally validated interactions by B2H 160
Table S8: Metadata of experimentally validated interactions by Co-IP 161
Table S9: Uniprot annotations of interactions in Figure 3 162

Chapter 1. SEQUENCE, STRUCTURE, AND FUNCTION: STRUCTURAL AND FUNCTIONAL GENOMICS IN THE DEEP LEARNING ERA

Ian R. Humphreys

Abstract

Proteins are universally responsible for core cellular processes. Harnessing the biological information they provide is central to understanding biochemistry and many approaches are developed and deployed to this end. Evolutionary information encoded in multiple sequence alignments has been used for decades to identify functional properties of proteins, predict protein-protein interactions, and enhance protein modeling. However, the extensive sequencing boon of next-generation high-throughput technology and advances in deep learning have ushered in a new stage of computational biology wherein high-resolution structural information and petabytes of data are readily accessible. We review the state of computational biology with respect to protein structure, function, and the interplay of genomic sequences in the deep learning era while positing what we believe to be the next challenges of structure-augmented functional genomics.

Introduction

Billions of years of evolution have resulted in highly refined proteins that carry out cellular processes across all domains of life. Proteins are responsible for the most basic biological functions to complex signaling cascades involving numerous protein interactions and have the potential to be harnessed for therapeutic development. The advent of genomic sequencing has enabled the study of organisms at a nucleotide (or amino acid) resolution and technological advances have resulted in rich sequence data which can be mined using computational methods. The accessibility of high-resolution protein structure modeling (*1*, *2*) enables the rapid integration of structural and mutational data to better understand genetic diseases or biochemical conservation of functionally important regions. Additionally, the investigative scope

of protein interactions has broadened over the years from single protein studies to proteomes, and we look towards the prospect of interspecies or eventually community-wide interactomes becoming feasible in the near future. By integrating genomic sequencing and protein structure information, we enter a new era of functional genomics.

The essential relationship between amino acid sequence and structure has long been appreciated (*3*). Studies have demonstrated the relationship between amino acid substitutions and double mutations on protein structure and function (*4*, *5*), and initial codon conservation models began to account for amino acid coevolution (*5*). As evolutionary coupling models became more robust, coevolution information extracted from multiple sequence alignments (MSAs) could be leveraged for structure modeling (*6*, *7*), protein-protein interaction (PPI) identification (*8*–*10*), and now, evolutionary and coevolutionary features are extracted as an integral feature for second-generation DL-based structure prediction methods (*1*, *2*). These features are especially important for identifying PPIs and modeling hetero-oligomeric complexes (*11*–*17*) which can further elucidate protein function.

Here, we synthesize recent advances that enable the integration of high-resolution structural information into functional studies while reviewing key historical advances that have enabled these methods. Fundamentally, the reviewed work follows the central biochemistry dogma of sequence → structure → function through extensive cross-talk between genomics and structural biology which we believe will be amplified by the field moving forwards. We focus on work involving amino-acid sequence evolution, protein structure prediction, and the interplay of evolution on protein function primarily in the context of PPIs.

Extracting protein evolutionary information from genomic sequences

The proteins in an organism's proteome, encoded by genes, are typically sequenced as DNA and the codons are then translated into amino acids. Collections of evolutionarily or structurally related polypeptide chains constructed from homologs can be assembled into MSAs wherein each amino acid is aligned with the same position in each homologous protein. Consequently, MSAs provide rich evolutionary information such as amino acid conservation and covariation which can inform structure and function.

The rate of molecular evolution is inversely correlated with selective pressure (*18*) resulting in sites of functional and structural integrity that are highly conserved across homologous proteins (*19*). Early mutagenesis studies showed second-site reverse mutations occur in specific regions of the polypeptide sequence and double-mutations may be created to restore function. This suggests these forward and reverse mutations occur in close proximity within the folded 3D structure (*4*). Initial mathematical models accounting for the frequency of covarying amino acids across homologous protein sequences based on a background rate of random single and double mutations showed that co-mutating residue pairs are both infrequent and functionally important (*5*). Further analysis of residue conservation, divergence, and residue-residue covariation using residue-residue correlation matrices predicted structural features such as secondary structure, internal and surface residues, and functional sites (*20*). These observations lay the foundation for residue-residue covariation models to infer functional and structural features.

Borrowing from information theory, mutual information (MI) (*21*) applied to homologous protein sequences to measure residue covariation may be used to reveal functionally important interacting sites (*22*). However, identifying coevolving residues from noisy evolutionary data, where both beneficial and deleterious mutations may randomly arise across the protein, limits the inference potential of covarying residues from MI (*23*–*25*). To account for these background

entropic and phylogenetic noises, average product correction (APC) may be applied to compute a normalized MI between any *ij* residue pair in a protein by subtracting the average MI of position *i* multiplied by the average MI of position *j* divided by the average MI between all *ij* pairs of residues within the protein ($MI_{APC} = MI_{ij} - ((MI_{ix} * MI_{jx}) / MI_{protein})$) (*26*).

Even with the higher sensitivity identification of covarying residue pairs, MI_{APC} is unable to resolve direct from indirect covariation (eg., residues *A B* and *A C* coevolve, residues *B C* are indirectly correlated). Direct-coupling analysis (DCA) attempts to disentangle these correlations globally opposed to locally by applying an inverse residue-residue covariance matrix (*27*). Briefly, DCA computes the frequency of finding residue *A* in column *i* of the MSA ($f_i(A)$) and frequency of the residue pair *A,B* appearing in columns *ij* ($f_{ij}(AB)$) to infer "direct statistical couplings" which are measured by direct information (DI) for each pair of *ij* columns (*27*, *28*). Rather than using an inverse covariance matrix to approximate direct couplings, GREMLIN uses a pseudo-likelihood framework to optimize learning for residue contact predictions (*29*, *30*). Consequently, coevolving residue pairs predicted from GREMLIN are more accurate than those from DCA or MI which can be better informative for protein structure prediction. These sites of residue-residue covariation have been shown to reside in close 3D proximity in tertiary protein structures (*31*, *32*) and the strength of direct couplings is a predictor of this distance (*33*). Because of this rich evolutionary information, adequate sequence diversity in MSAs are a key determinant of the accuracy of current state-of-the art protein structure prediction methods.

Protein structure prediction from genomic sequences

Proteins are integral to biology and an understanding of their structure can inform functional mechanisms. For decades, the determination of protein structure from sequence has been a hallmark challenge in biology (*3*) and deservedly garnered intensive efforts. At its core, the

problem of protein folding stems from the massive conformational search space (*34*) which makes exhaustive sampling computationally intractable. The Critical Assessment of protein Structure Prediction (CASP) is a community-run biennial assessment of the accuracy of protein structure prediction methods in which groups blindly model novel proteins. As such, the majority of structure prediction advances are contextualized by recent CASP results. Below, we summarize the history and landmark advances in protein structure prediction in recent years.

Physics and homology-based structure modeling

Prior to the dominant performance of machine learning methods, protein structure prediction was carried out using physical-based algorithms to model atomic interactions in biomolecular structures. These algorithms such as Rosetta (*35*–*37*) rely on force fields and complex energy functions that incorporate numerous physical attributes such as van der Waals packing, hydrogen bonding, and desolvation to approximate local and global energy. As a consequence of amino acid sequences generally folding to their lowest energy state with few exceptions (*38*), minimizing this energy will result in the most likely stable conformation of a protein's tertiary structure. In the event that a closely related protein structure can be identified, template based homology modeling (e.g., RosettaCM) provides a starting point for conformational space (*39*); however, for more distant or novel proteins, *ab initio* prediction must traverse large energy landscapes (*40*).

Rosetta uses Monte Carlo based simulated annealing to search for the lowest-energy structure of the polypeptide chain followed by gradient minimization. The Rosetta method leverages short fragments of variable lengths between 3-19 residues with known structure to approximate local interactions. These fragments are then assembled via Monte Carlo to produce native-like models which take into account non-local interactions such as hydrophobic burial, backbone

hydrogen-bonding, and side-chain interactions. Selection of the *best* model out of approximately 20,000+, is based on minimizing a secondary all-atom energy function that predicts global energy based on known protein structures by incorporating hydrogen bonding, electrostatic interactions, van der Waals interactions and solvation (*41*).

In the event of nonexistent homologous templates during modeling, Rosetta has been shown to be effective (in the context of physical based methods) for small proteins (*37*), however, for larger proteins, experimental data was often necessary to augment predictions to obtain reasonable resolution (*42*), (*43*). These data augmentations may take the form of distance constraints that can be obtained from low resolution cryo-EM densities, NMR, or amino acid covariation information. As discussed previously, residues in contact with one another within proteins evolve together wherein substitutions at positions close in 3D space covary (*31*, *32*). Briefly, these amino acid covariations can be exploited to assess the likelihood that two residues interact by extracting evolutionary information from MSAs constructed using genomic sequences from homologous proteins and the strength of this covariation is correlated to the physical distance between residues (*33*). Because of this phenomena, accurate prediction of amino acid contacts from evolutionary data are readily minable and can be converted into angstrom distance constraints applied during protein modeling so long as MSAs contain approximately 5L (where L is the length of a polypeptide sequence in amino acids) (*30*, *44*). For both monomeric and oligomeric modeling, these evolutionary distance constraints are especially impactful during conformational space sampling and oligomeric state validation. While effective for their time, these classical physics-based modeling methods struggle with large proteins, higher order oligomeric complexes, and pale in comparison to the accuracy of DL methods.

First generation of deep learning for structure prediction: De novo modeling from MSAs

CASP13 marked the beginning of the deep learning for structure prediction era; where neural network-based methods (*45*–*49*) were able to predict fold-level accuracy of protein structure models. Of note, the best performing CASP13 model, 2018 version of AlphaFold (*50*), hereafter referred to as AlphaFold1.0 (AF1), significantly outperformed classical modeling methods (*51*). These deep convolutional residual neural networks learn residue-residue distance probability distributions from coevolutionary information extracted from MSAs to guide residue-residue contact predictions which can be used in 3D structure prediction by direct optimization.

AF1 is a neural network trained on PDB structures to predict the C_β atom distances between all *ij* residues of a protein. Leveraging an architecture composed of a 2D network with 220 residual blocks and dilated convolutions, AF1 extracts information from sequence and MSA features. The network predicts probability distributions for 0-22Å C_β-C_β distances and backbone torsion angles: φ and ψ. Gradient descent on protein-specific potential is applied to a spline fit to the negative log probabilities and summed across residue pairs to provide constraints for 3D structure modeling.

Expanding upon CASP13 advances (*45*, *46*, *50*, *52*) and AF1, transform-restrained Rosetta (trRosetta) (*53*) is a deep residual-convolutional neural network that predicts distance and contact predictions from sequence and MSA features by learning probability distributions over distances. In short, trRosetta further extends predictions to generate 6D orientation features (C_β-C_β, ω, θ_{12}, θ_{21}, φ_{12}, and φ_{21}) which exhaustively define the positions of backbone atoms of residue pairs to more accurately predict 3D structures. 3D structures are predicted from constrained minimization of distance and orientation-dependent potentials in PyRosetta (*54*). Collectively the integration of these additional orientation features aided in the accuracy improvements of trRosetta to outperform CASP13 methods (*53*).

Second generation of deep learning for structure prediction: Highly accuracy prediction

Deepmind's AlphaFold2 (AF2) (*2*), presented remarkably accurate predictions at the CASP14 conference (*55*) with 3D structure predictions approaching experimental accuracy. This success inspired the rapid development of other high-accuracy methods such as RoseTTAFold (RF) (*1*). RF and AF2 extract 1D and 2D residue-pair features from MSAs and template information to predict 3D protein structures with high-accuracy. Despite their similarities, these two methods have several key differences which result in differing levels of accuracy.

RoseTTAFold is a three-track neural network in which information at the one-dimensional (1D) sequence level, the 2D distance map level, and the 3D coordinate level are successively transformed. During architectural exploration in development, and prior to the addition of a third structure track, a two-track model with biaxial attention in the 2D distance track outperformed prior methods by gradient based folding using pyrosetta (*54*). A third track (3D coordinates) was added to further improve prediction by providing tighter connections between features with SE(3)-equivariant transformer layers (*56*). RF achieves high accuracy structure prediction enabling molecular replacement and demonstrates through the addition of a 200 residue indexing gap, that a network trained on monomeric proteins can simultaneously fold multiple chains into high resolution complexes.

On the other hand, AlphaFold2 is an end-to-end two-track network consisting of a sequence track and a paired distance track which are transformed, integrated into the structure module, and then these features are recycled prior to final 3D structure prediction. To achieve its highly accurate predictions, AF2 integrates several key advances: 1) the novel message-passing Evoformer block, 2) invariant point attention (IPA), 3) frame-aligned point error (FAPE) structure loss function, 4) recycling of learned features, and 5) self-distillation training data. AF2 relies on

an input MSA and jointly embeds MSA and residue-residue pair features which are iteratively processed through the Evoformer which detects spatial and evolutionary relationships with triangle attention. While computationally expensive, triangle attention can learn and enforce triangle inequality to filter noise from coevolutionary signal. Within the Evoformer, the MSA representation updates the pair representation each block. IPA is used to enforce equivariance by updating single residue representations in 3D space to produce points in the frame of each residue. Frames are constructed using the position of three atoms: N, Cα, and C for backbones and atoms centered around the torsion bond for sidechains. These points are invariant to rotations and translations making each attention update to the rotation and translation of the backbone frame an equivariant attention operation. A new loss function, FAPE, was constructed to compare the predicted atom coordinates to the true atom coordinates based on the predicted local frames and true local frames, allowing local scoring regardless of global rotations greatly improving training. These feature representations extracted from both the MSA and paired-feature tracks are iteratively recycled through the network several times to refine predictions prior to entering the structure module which calculates rotation and translation for each residue. As a result of the high-resolution predictions, they were able to augment the training dataset with predicted structures from the network to substantially increase the training set depth. The structure predictions produced by these current generation tools are rapidly enabling biological insights (*57*–*59*).

ESMFold (*60*) integrates a pre-trained LLM (ESM-2) which predicts amino acid identity by masked token prediction to replace the MSA input for protein structure prediction. ESMFold largely borrows from AF2 using a simplified Evoformer for the folding black and an equivariant transformer with IPA. By replacing the MSA input with an LLM, ESMFold is orders of magnitude faster enabling the generation of large metagenomic database scale predictions, however the accuracy of a prediction still scales with the number of detectable homologues for a given

sequence and remains lower than that of AF2 or RF.

Following the release of AF2, people sought to understand what components contributed to such high performance. RoseTTAFold2 (*61*), an updated version of RF, explores components of AF2 within RF's three-track architecture: namely 1) recycling, 2) distillation, 3) FAPE loss, and 4) network depth. The merging of these components into RF2 results in comparable performance to AF2 and AF-multimer (*14*) (due to the addition of dimeric complexes from the PDB) demonstrating that despite large architectural differences, DL models may achieve similar accuracy. Importantly, RF2 retains biaxial structure attention rather than AF2's triangle attention which results in considerably faster inference and uses an SE(3)-transformer instead of IPA. RF was further extended to model protein-nucleic acids with RoseTTAFoldNA (*62*) (RFNA) offering the first *ab initio* deep learning method for predicting protein-nucleic acid complexes.

Third Generation structure prediction networks: An atomistic view

While many biological processes are mediated by proteins or protein-protein interactions, proteins interact with nucleic acids, small molecules, or metals and may have covalent modifications. The current generation of protein structure prediction tools have sought to incorporate these additional biomolecular interactions through an atomistic representation of molecules.

RoseTTAFold All-Atom (*63*) (RFAA) builds upon RFNA extension of protein-nucleic acid complexes to represent biomolecular assemblies with RF2 architecture. It combines sequence-based nucleic and amino acids with an atomic graph representation to handle all-atom context. The 1D track now includes 80 tokens (20 amino acids, 8 nucleic acids, 46 of the most common element types, and 6 gap/unknown tokens). In RF and AF2, proteins are

represented by Cα coordinates and a N-C-C frame, these are replaced by P in RFNA. For RFAA, this representation is encoded by heavy-atoms in the 3D track and an all-atom FAPE loss was introduced based on bonded neighbors. RFAA achieves high accuracy in protein-small molecule complex prediction while retaining comparable performance to AF2 in monomeric protein structure prediction and enabled modeling multicomponent biomolecular assemblies for the first time.

Shortly after RFAA, DeepMind announced AlphaFold3 (AF3) (*64*), an all-atom update to AF2 with additional improvements to oligomeric structure prediction and best-in-class biomolecular modeling. AF3 contains several key differences in architecture from AF2 in addition to tokenizing elements like in RFAA. 1) AF3 reduces the Evoformer mvΩodule from AF2 into a 4 block MSA module and 48 block pair module to increase speed and memory, 2) it removes equivariance, likely learning this by permuting coordinates, and 3) AF3 introduces a diffusion model in the structure module enabling generative predictions to sample larger energy landscapes rapidly after extracting MSA/pair information. Briefly, diffusion is a generative deep learning model which begins with random gaussian noise and continuously denoises a prediction (in this case, protein structure) for *t* time points. Much of the improved accuracy of AF3 is likely from greatly increased sampling space through generative modeling and prediction scoring. While other methods have attempted to expand upon sampling diversity by MSA truncation, MSA clustering, enabling dropout, or traversing latent space with a variational autoencoder (VAE) (*65*–*68*), AF3 handles MSA and pair features as well as recycling prior to diffusion allowing one to rapidly generate ensemble predictions. RFAA, while less accurate than AF3, currently remains the only open-source method for all-atom *ab initio* protein biomolecular protein complex prediction as AF3 is limited to a fixed number of server predictions and restricted element-type tokens.

Biological function of proteins using sequence and structure

The principal goal of protein structure determination is to glean biological insights about the protein of interest. However, prior to current deep learning approaches that make protein structure modeling accessible, experimental structure determination by X-ray crystallography or cryo-electron microscopy was arduous and *ab initio* structure modeling remained limited. As discussed, amino acid covariation has been used to identify structural and functional insights in monomeric proteins; yet, proteins rarely act alone thus, identifying and understanding a protein's potential interactions with other proteins has great functional potential. Classical large-scale screens have been conducted using two-hybrid, affinity-purification mass spectrometry (AP/MS), and other high-throughput experimental approaches to identify PPIs and have been applied to a variety of organisms (*69*, *70*). However, these methods suffer from biases which result in considerable discrepancies between screens along with high false-positives and false-negative rates (*71*, *72*), lack structural information, and in the case of AP/MS, cannot disentangle indirect from direct interactions. Towards this end, more readily accessible protein/genomic sequences have been used in place of poor to non-existent structural information or noisy high-throughput protein-protein interaction screens to add function to proteins.

Harnessing coevolving residues between proteins for interaction identification

Due to the expectation that interacting proteins evolve together, amino acid covariation can be leveraged to assess the likelihood that two proteins interact by extracting the coevolutionary information from pairing MSAs (pMSAs) constructed using genomic sequences from orthologous proteins (*8*–*10*). For example, for each pair of proteins in the proteome of an organism, each pMSA consists of two proteins *a* and *b* and contains only the orthologous

sequences where each relative of protein *a* and relative of protein *b* are from the same organism and evolutionary history. Using these paired alignments, sites of coevolving residues can be used to provide structural constraints to inform protein-protein interface sites (*8*, *9*). Yet pairing these orthologous proteins is challenging. Consequently, early works in predicting PPIs based on coevolving residues that relied on pMSAs were restricted to within protein family interaction partner predictions (*8*, *27*, *73*, *74*). Paralogous sequences complicate the evolutionary history of proteins and noise coevolutionary signal complicating pMSA curation. Thus, paralogs must either be excluded from coevolutionary studies, accounted for by modifications to methodology (*75*), or manually resolved.

Pairing sequences such as by genomic distance using uniprot accession IDs (taxonomy information) may yield rapid automated pMSAs, however they lack information about orthologous protein families which contributes to false-evolutionary noise. Exhaustive approaches that seek to reduce this noise through the identification of putative orthologues often necessitate full-proteome or large genomic databases. One such method for putative orthologue identification is by reciprocal best blast hits (rbh) (*76*) whereby for a pair of proteins *a* and *b* from the proteomes of organisms *A* and *B*, if the top hit of proteome *B* is blasted against protein *a* is protein *b* and the reciprocal blast (proteome *A* against protein *b*) results in the top hit of protein *a*, then the pair are putative orthologues. If the forward or reverse blast results in a differing top hit, then both pairs of proteins are deemed paralogues and are excluded. To reduce distant phylogenetically related sequences being selected by rbh, further filters on global sequence alignment quality and estimations of evolutionary distances can be applied to create reciprocal smallest distance (rsd) (*77*).

Given robustly paired protein alignments, coevolution has been shown to predict new PPIs at high accuracy and validate previously identified PPIs by different methods on a whole-proteome

scale. In a large-scale computational screen of the *Escherichia coli* and *Mycobacterium tuberculosis* proteomes, coevolutionary analyses identified hundreds of binary PPIs at accuracies above experimental screens (*9*). Similarly, further large coevolution screens have been conducted on the *E. coli* proteome using more automated approaches to pMSA generation with success (*10*). Linking these computationally identified PPIs, can further illuminate biochemical pathways and aid in annotating proteins of unknown functional capacities. However, statistical models have difficulty in capturing higher order covariation patterns and the biophysical information between residues which would likely improve interaction predictions.

Deep learning prediction of physically interacting proteins

Similar to monomeric protein structure prediction, deep learning has become a key component in the prediction of protein-protein interactions. Early work such as DPPI (*78*) used a sequence-based convolutional neural network to predict interaction scores trained on experimentally identified interaction pairs and immediately showed improved performance against prior models. However, the protein structure prediction method trRosetta (*53*, *79*) also showed promise for identifying coevolutionary contacts at the interface of physically interacting proteins to guide interaction prediction (*80*) and demonstrated one of many biological applications of DL structure prediction. Now, the current best-in-class protein structure prediction methods can also be used for predicting if two proteins are likely to form a biological complex with accuracy exceeding prior purely statistical methods (*11*–*13*, *17*).

To screen all potential protein-protein pairs in a proteome ($n \times (n\text{-}1) \div 2$) rapidly becomes a massive computational roadblock. For example, the human proteome which is composed of approximately 20,000 proteins would become around 200 million pairs of proteins to test for binary interactions alone. Thus, for large-scale computational screens of PPIs, the

computational methods must balance speed and performance. One such solution we have used to allow large-scale screens is the use of a light-weight pre-filtering method followed by a more accurate final pass. Using only the 1D MSA and 2D pairwise residue information tracks from RF, we created a light-weight version of RF (RF-2track), which significantly improved the ability to discriminate between true interacting protein pairs and noise with far superior performance to DCA (*11*, *12*). This lighter network sacrifices structure model quality and some discrimination accuracy compared to AlphaFold2 which may be used as a final stringent filter of protein interactions, resulting in high confidence predictions of PPIs (*11*, *12*) and models of protein complexes (*13*, *15*, *16*). Using RF and AF variants, PPI studies have been carried out in yeast (*11*), human mitochondria (*12*), oncogenic proteins (*13*), and pathogenic bacteria (*17*); each of which draws functional insights inaccessible to structural-free analyses opening the doors to a newly realized application of structure to functional genomics.

Protein-protein interactions aren't always heterologous, identical copies of proteins may form complexes called homodimers. Homodimers occur to stabilize chains, concentrate active sites, or otherwise function in consort with one another to carry out a biological function. AF was used to predict the oligomerization state of four proteomes revealing that ~45% of bacterial and archaeal proteins form homomers (*81*). These computationally predicted oligomeric states may aid in the experimental determination of protein structure or inform the necessary stoichiometry of proteins for given systems.

Perspectives

Monumental progress has been made within protein structure prediction which now now enables rapid integration of high resolution predicted protein models into genomics workflows. Efficient sequence and structure-based homology tools such as FoldSeek (*82*), and large public

databases of computed structures (*83*) are illuminating new protein families and identifying novel protein folds (*58*), which may be leveraged to find evolutionary links between very remote homologues (*59*). Structure-based models have also entered genomics, to deliver best-in-class variant effect prediction (*84*, *85*) improving our understanding of genetic disease.

An immense effort into modifying and extending the reach of AlphaFold beyond monomeric structure prediction tasks which have begun to address some of the great challenges in computational structural biology that remain post AF2 and RF2. For example in oligomeric complex prediction, AF2complex uses a residue-index gap modified AF2 and automated monomeric template searches to improve complex structure accuracy and provide a pMSA-free method of PPI identification (*86*) and others have developed a template-free graph traversal method to assemble large complexes with higher accuracy than 1-shot predictions and would otherwise exceed the limits of computational tractability (*87*). As previously mentioned, efforts have been made to modify AF and RF to model ensembles to identify the most accurate model or predict alternative conformational states (*67*, *68*). We see continued development in generative modeling through both post-hoc modifications and architectural changes like diffusion in AF3 as significant areas of focus for protein modeling moving forward.

Despite these great advances in recent years, we see that structure prediction methods remain limited by their reliance on MSAs (*60*, *64*) and have difficulty in discriminating between the preferred oligomeric-states further reducing model accuracy (*88*). Some classes of proteins such as antibodies or those with large flexible regions remain especially challenging for current methods as their evolutionary constraints are less pronounced due to rapidly evolving regions and backbone flexibility in general is an inherent issue for deterministic modeling.

Together, these will likely define forthcoming advances in protein structure prediction.

Concurrently, extended efforts in structural bioinformatics may look towards such advancing challenges by improving data-curation pipelines. For example, through MSA expansion and optimization, current methods are better able to predict protein structure (*7*, *89*–*91*), which in turn, likely improves higher order oligomer assembly or oligomeric state predictions. These improvements to the incorporation of genomic and structural data will enable further functional insights. Yet, we fear that the computational method development will likely become hindered by functional annotation of genes which ultimately require experimental characterization.

Recently with RFAA and AF3, deep learning models are beginning to reason over atomistic representations of proteins to model non-proteinaceous molecules in complex with proteins at reasonable accuracy. With much anticipation, we look towards the prospect of high resolution virtual drug screens with atomic resolution protein-small molecule structure prediction in the near future; which, in consort with resolved interactomes, will illuminate diseases and enable targeted therapeutic development. Furthermore, we hope efforts to improve deep learning based protein models will result in a reduced reliance on evolutionary coupling information from pMSAs giving rise to the possibility of computational host-pathogen PPI identification or interactions with decorations such as glycosylations present on many viral surface proteins. With protein structures being such data rich objects, pivotal to biology we believe this is still the early stages of the biological discoveries from the protein structure prediction revolution.

References

1. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).

2. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

3. C. B. Anfinsen, Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).

4. C. Yanofsky, V. Horn, D. Thorpe, PROTEIN STRUCTURE RELATIONSHIPS REVEALED BY MUTATIONAL ANALYSIS. *Science* **146**, 1593–1594 (1964).

5. W. M. Fitch, E. Markowitz, An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* **4**, 579–593 (1970).

6. A. R. Ortiz, A. Kolinski, P. Rotkiewicz, B. Ilkowski, J. Skolnick, Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins* **Suppl 3**, 177–185 (1999).

7. S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. A. Pavlopoulos, D. E. Kim, H. Kamisetty, N. C. Kyrpides, D. Baker, Protein structure determination using metagenome sequence data. *Science* **355**, 294–298 (2017).

8. S. Ovchinnikov, H. Kamisetty, D. Baker, Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030 (2014).

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

10. A. G. Green, H. Elhabashy, K. P. Brock, R. Maddamsetti, O. Kohlbacher, D. S. Marks, Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat. Commun.* **12**, 1396 (2021).

11. I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness, S. Banjade, S. R. Bagde, V. G. Stancheva, X.-H. Li, K. Liu, Z. Zheng, D. J. Barrero, U. Roy, J. Kuper, I. S. Fernández, B. Szakal, D. Branzei, J. Rizo, C. Kisker, E. C. Greene, S. Biggins, S. Keeney, E. A. Miller, J. Christopher Fromme, T. L. Hendrickson, Q. Cong, D. Baker, Computed structures of core eukaryotic protein complexes. [Preprint] (2021). https://doi.org/10.1126/science.abm4805.

12. J. Pei, J. Zhang, Q. Cong, Human mitochondrial protein complexes revealed by large-scale coevolution analysis and deep learning-based structure modeling, *bioRxiv* (2021). https://doi.org/10.1101/2021.09.14.460228.

13. J. Zhang, J. Pei, J. Durham, T. Bos, Q. Cong, Towards a structurally resolved cancer interactome, *bioRxiv* (2022). https://doi.org/10.1101/2022.01.21.477304.

14. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell,

J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer, *bioRxiv* (2021). https://doi.org/10.1101/2021.10.04.463034.

15. P. Bryant, G. Pozzati, A. Elofsson, Improved prediction of protein-protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).

16. D. F. Burke, P. Bryant, I. Barrio-Hernandez, D. Memon, G. Pozzati, A. Shenoy, W. Zhu, A. S. Dunham, P. Albanese, A. Keller, R. A. Scheltema, J. E. Bruce, A. Leitner, P. Kundrotas, P. Beltrao, A. Elofsson, Towards a structurally resolved human protein interaction network, *bioRxiv* (2021)p. 2021.11.08.467664.

17. I. R. Humphreys, J. Zhang, M. Baek, Y. Wang, A. Krishnakumar, J. Pei, I. Anishchenko, C. A. Tower, B. A. Jackson, T. Warrier, D. T. Hung, S. B. Peterson, J. D. Mougous, Q. Cong, D. Baker, Essential and virulence-related protein interactions of pathogens revealed through deep learning. *bioRxiv*, doi: 10.1101/2024.04.12.589144 (2024).

18. M. Kimura, The neutral theory of molecular evolution: a review of recent evidence. *Jpn. J. Genet.* **66**, 367–386 (1991).

19. W. M. Fitch, E. Margoliash, A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome c as a model case. *Biochem. Genet.* **1**, 65–71 (1967).

20. S. A. Benner, D. Gerloff, Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Adv. Enzyme Regul.* **31**, 121–181 (1991).

21. C. E. Shannon, *A mathematical theory of communication* (1948).

22. B. T. Korber, R. M. Farber, D. H. Wolpert, A. S. Lapedes, Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 7176–7180 (1993).

23. K. R. Wollenberg, W. R. Atchley, Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 3288–3291 (2000).

24. A. A. Fodor, R. W. Aldrich, Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* **56**, 211–221 (2004).

25. L. C. Martin, G. B. Gloor, S. D. Dunn, L. M. Wahl, Using information theory to search for co-evolving residues in proteins. *Bioinformatics* **21**, 4116–4124 (2005).

26. S. D. Dunn, L. M. Wahl, G. B. Gloor, Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* **24**, 333–340 (2008).

27. M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, T. Hwa, Identification of direct residue contacts in protein–protein interaction by message passing. [Preprint] (2009). https://doi.org/10.1073/pnas.0805923106.

28. F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E1293–301 (2011).

29. S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, C. J. Langmead, Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).

30. H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue-residue

contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15674–15679 (2013).

31. D. Altschuh, A. M. Lesk, A. C. Bloomer, A. Klug, Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *J. Mol. Biol.* **193**, 693–707 (1987).

32. U. Göbel, C. Sander, R. Schneider, A. Valencia, Correlated mutations and residue contacts in proteins. *Proteins* **18**, 309–317 (1994).

33. D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, C. Sander, Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766 (2011).

34. C. Levinthal, Are there pathways for protein folding? *J. Chim. Phys. Physicochim. Biol.* **65**, 44–45 (1968).

35. R. Bonneau, J. Tsai, I. Ruczinski, D. Chivian, C. Rohl, C. E. Strauss, D. Baker, Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins* **Suppl 5**, 119–126 (2001).

36. C. A. Rohl, C. E. M. Strauss, K. M. S. Misura, D. Baker, Protein structure prediction using Rosetta. *Methods Enzymol.* **383**, 66–93 (2004).

37. P. Bradley, K. M. S. Misura, D. Baker, Toward high-resolution de novo structure prediction for small proteins. *Science* **309**, 1868–1871 (2005).

38. D. Baker, J. L. Sohl, D. A. Agard, A protein-folding reaction under kinetic control. *Nature* **356**, 263–265 (1992).

39. Y. Song, F. DiMaio, R. Y.-R. Wang, D. Kim, C. Miles, T. Brunette, J. Thompson, D. Baker, High-resolution comparative modeling with RosettaCM. *Structure* **21**, 1735–1742 (2013).

40. K. T. Simons, I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, D. Baker, Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* **34**, 82–95 (1999).

41. R. Das, D. Baker, Macromolecular modeling with rosetta. *Annu. Rev. Biochem.* **77**, 363–382 (2008).

42. S. Raman, O. F. Lange, P. Rossi, M. Tyka, X. Wang, J. Aramini, G. Liu, T. A. Ramelot, A. Eletsky, T. Szyperski, M. A. Kennedy, J. Prestegard, G. T. Montelione, D. Baker, NMR structure determination for larger proteins using backbone-only data. *Science* **327**, 1014–1018 (2010).

43. O. F. Lange, P. Rossi, N. G. Sgourakis, Y. Song, H.-W. Lee, J. M. Aramini, A. Ertekin, R. Xiao, T. B. Acton, G. T. Montelione, D. Baker, Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 10873–10878 (2012).

44. S. Ovchinnikov, D. E. Kim, R. Y.-R. Wang, Y. Liu, F. DiMaio, D. Baker, Improved de novo structure prediction in CASP11 by incorporating coevolution information into Rosetta. *Proteins* **84 Suppl 1**, 67–75 (2016).

45. S. M. Kandathil, J. G. Greener, D. T. Jones, Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. *Proteins* **87**, 1092–1099 (2019).

46. J. Xu, Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 16856–16865 (2019).

47. J. Hou, T. Wu, R. Cao, J. Cheng, Protein tertiary structure modeling driven by deep learning and contact distance prediction in CASP13. *Proteins* **87**, 1165–1178 (2019).

48. W. Zheng, Y. Li, C. Zhang, R. Pearce, S. M. Mortuza, Y. Zhang, Deep-learning contact-map guided protein structure prediction in CASP13. *Proteins* **87**, 1149–1164 (2019).

49. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins* **87**, 1141–1148 (2019).

50. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

51. L. A. Abriata, G. E. Tamò, M. Dal Peraro, A further leap of improvement in tertiary structure prediction in CASP13 prompts new routes for future assessments. *Proteins* **87**, 1100–1112 (2019).

52. S. Wang, S. Sun, Z. Li, R. Zhang, J. Xu, Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. *PLoS Comput. Biol.* **13**, e1005324 (2017).

53. J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, D. Baker, Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 1496–1503 (2020).

54. S. Chaudhury, S. Lyskov, J. J. Gray, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).

55. A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, J. Moult, Critical assessment of methods of protein structure prediction (CASP)-Round XIV. *Proteins* **89**, 1607–1617 (2021).

56. F. B. Fuchs, D. E. Worrall, V. Fischer, M. Welling, SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks (2020). http://arxiv.org/abs/2006.10503.

57. K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Žídek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reiman, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).

58. J. Durairaj, A. M. Waterhouse, T. Mets, T. Brodiazhenko, M. Abdullah, G. Studer, G. Tauriello, M. Akdel, A. Andreeva, A. Bateman, T. Tenson, V. Hauryliuk, T. Schwede, J. Pereira, Uncovering new families and folds in the natural protein universe. *Nature* **622**, 646–653 (2023).

59. I. Barrio-Hernandez, J. Yeo, J. Jänes, M. Mirdita, C. L. M. Gilchrist, T. Wein, M. Varadi, S. Velankar, P. Beltrao, M. Steinegger, Clustering predicted structures at the scale of the known protein universe. *Nature* **622**, 637–645 (2023).

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

61. M. Baek, I. Anishchenko, I. R. Humphreys, Q. Cong, D. Baker, F. DiMaio, Efficient and accurate prediction of protein structure using RoseTTAFold2, *bioRxiv* (2023)p. 2023.05.24.542179.

62. M. Baek, R. McHugh, I. Anishchenko, H. Jiang, D. Baker, F. DiMaio, Accurate prediction of

protein–nucleic acid complexes using RoseTTAFoldNA. *Nat. Methods* **21**, 117–121 (2023).

63. R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. R. Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys, R. McHugh, D. Vafeados, X. Li, G. A. Sutherland, A. Hitchcock, C. N. Hunter, A. Kang, E. Brackenbrough, A. K. Bera, M. Baek, F. DiMaio, D. Baker, Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).

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

65. D. Del Alamo, D. Sala, H. S. Mchaourab, J. Meiler, Sampling alternative conformational states of transporters and receptors with AlphaFold2. *Elife* **11** (2022).

66. B. Wallner, AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics* **39**, btad573 (2023).

67. H. K. Wayment-Steele, A. Ojoawo, R. Otten, J. M. Apitz, W. Pitsawong, M. Hömberger, S. Ovchinnikov, L. Colwell, D. Kern, Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2023).

68. S. Mansoor, M. Baek, H. Park, G. R. Lee, D. Baker, Protein Ensemble Generation Through Variational Autoencoder Latent Space Sampling. *J. Chem. Theory Comput.*, doi: 10.1021/acs.jctc.3c01057 (2024).

69. S. V. Rajagopala, P. Sikorski, A. Kumar, R. Mosca, J. Vlasblom, R. Arnold, J. Franca-Koh, S. B. Pakala, S. Phanse, A. Ceol, R. Häuser, G. Siszler, S. Wuchty, A. Emili, M. Babu, P. Aloy, R. Pieper, P. Uetz, The binary protein-protein interaction landscape of Escherichia coli. [Preprint] (2014). https://doi.org/10.1038/nbt.2831.

70. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, J. M. Rothberg, A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. [Preprint] (2000). https://doi.org/10.1038/35001009.

71. A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, M. Gerstein, Bridging structural biology and genomics: assessing protein interaction data with known complexes. [Preprint] (2002). https://doi.org/10.1016/s0168-9525(02)02763-4.

72. J. Mackay, M. Sunde, J. Lowry, M. Crossley, J. Matthews, Protein interactions: is seeing believing? [Preprint] (2007). https://doi.org/10.1016/j.tibs.2007.09.006.

73. T. A. Hopf, C. P. I. Schärfe, J. P. G. L. M. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. J. J. Bonvin, D. S. Marks, Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3** (2014).

74. A.-F. Bitbol, R. S. Dwyer, L. J. Colwell, N. S. Wingreen, Inferring interaction partners from protein sequences. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 12180–12185 (2016).

75. A.-F. Bitbol, Inferring interaction partners from protein sequences using mutual information. *PLoS*

Comput. Biol. **14**, e1006401 (2018).

76. A. E. Hirsh, H. B. Fraser, Protein dispensability and rate of evolution. *Nature* **411**, 1046–1049 (2001).

77. D. P. Wall, H. B. Fraser, A. E. Hirsh, Detecting putative orthologs. *Bioinformatics* **19**, 1710–1711 (2003).

78. S. Hashemifar, B. Neyshabur, A. A. Khan, J. Xu, Predicting protein-protein interactions through sequence-based deep learning. *Bioinformatics* **34**, i802–i810 (2018).

79. I. Anishchenko, M. Baek, H. Park, N. Hiranuma, D. E. Kim, J. Dauparas, S. Mansoor, I. R. Humphreys, D. Baker, Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14. *Proteins* **89**, 1722–1733 (2021).

80. K. Macé, A. K. Vadakkepat, A. Redzej, N. Lukoyanova, C. Oomen, N. Braun, M. Ukleja, F. Lu, T. R. D. Costa, E. V. Orlova, D. Baker, Q. Cong, G. Waksman, Cryo-EM structure of a type IV secretion system. *Nature* **607**, 191–196 (2022).

81. H. Schweke, M. Pacesa, T. Levin, C. A. Goverde, P. Kumar, Y. Duhoo, L. J. Dornfeld, B. Dubreuil, S. Georgeon, S. Ovchinnikov, D. N. Woolfson, B. E. Correia, S. Dey, E. D. Levy, An atlas of protein homo-oligomerization across domains of life. *Cell* **187**, 999–1010.e15 (2024).

82. M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, J. Söding, M. Steinegger, Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).

83. M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

84. H. Gao, T. Hamp, J. Ede, J. G. Schraiber, J. McRae, M. Singer-Berk, Y. Yang, A. S. D. Dietrich, P. P. Fiziev, L. F. K. Kuderna, L. Sundaram, Y. Wu, A. Adhikari, Y. Field, C. Chen, S. Batzoglou, F. Aguet, G. Lemire, R. Reimers, D. Balick, M. C. Janiak, M. Kuhlwilm, J. D. Orkin, S. Manu, A. Valenzuela, J. Bergman, M. Rousselle, F. E. Silva, L. Agueda, J. Blanc, M. Gut, D. de Vries, I. Goodhead, R. A. Harris, M. Raveendran, A. Jensen, I. S. Chuma, J. E. Horvath, C. Hvilsom, D. Juan, P. Frandsen, F. R. de Melo, F. Bertuol, H. Byrne, I. Sampaio, I. Farias, J. V. do Amaral, M. Messias, M. N. F. da Silva, M. Trivedi, R. Rossi, T. Hrbek, N. Andriaholinirina, C. J. Rabarivola, A. Zaramody, C. J. Jolly, J. Phillips-Conroy, G. Wilkerson, C. Abee, J. H. Simmons, E. Fernandez-Duque, S. Kanthaswamy, F. Shiferaw, D. Wu, L. Zhou, Y. Shao, G. Zhang, J. D. Keyyu, S. Knauf, M. D. Le, E. Lizano, S. Merker, A. Navarro, T. Bataillon, T. Nadler, C. C. Khor, J. Lee, P. Tan, W. K. Lim, A. C. Kitchener, D. Zinner, I. Gut, A. Melin, K. Guschanski, M. H. Schierup, R. M. D. Beck, G. Umapathy, C. Roos, J. P. Boubli, M. Lek, S. Sunyaev, A. O'Donnell-Luria, H. L. Rehm, J. Xu, J. Rogers, T. Marques-Bonet, K. K.-H. Farh, The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).

85. J. Cheng, G. Novati, J. Pan, C. Bycroft, A. Žemgulytė, T. Applebaum, A. Pritzel, L. H. Wong, M. Zielinski, T. Sargeant, R. G. Schneider, A. W. Senior, J. Jumper, D. Hassabis, P. Kohli, Ž. Avsec, Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).

86. M. Gao, D. Nakajima An, J. M. Parks, J. Skolnick, AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* **13**, 1744 (2022).

87. P. Bryant, G. Pozzati, W. Zhu, A. Shenoy, P. Kundrotas, A. Elofsson, Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nat. Commun.* **13**, 1–14 (2022).

88. M. Kshirsagar, A. Meller, I. Humphreys, S. Sledzieski, Y. Xu, R. Dodhia, E. Horvitz, B. Berger, G. Bowman, J. L. Ferres, D. Baker, M. Baek, Rapid and accurate prediction of protein homo-oligomer symmetry with Seq2Symm. *Res Sq*, doi: 10.21203/rs.3.rs-4215086/v1 (2024).

89. Z. Peng, W. Wang, H. Wei, X. Li, J. Yang, Improved protein structure prediction with trRosettaX2, AlphaFold2, and optimized MSAs in CASP15. *Proteins* **91**, 1704–1711 (2023).

90. W. Zheng, Q. Wuyun, P. L. Freddolino, Y. Zhang, Integrating deep learning, threading alignments, and a multi-MSA strategy for high-quality protein monomer and complex structure prediction in CASP15. *Proteins* **91**, 1684–1703 (2023).

91. S. Lee, G. Kim, E. L. Karin, M. Mirdita, S. Park, R. Chikhi, A. Babaian, A. Kryshtafovych, M. Steinegger, Petabase-Scale Homology Search for Structure Prediction. *Cold Spring Harb. Perspect. Biol.* **16** (2024).

Chapter 2. COMPUTED STRUCTURES OF CORE EUKARYOTIC PROTEIN COMPLEXES

A version of this chapter has been previously published as:

Ian R. Humphreys[†], Jimin Pei[†], Minkyung Baek[†], Aditya Krishnakumar[†], Ivan Anishchenko, Sergey Ovchinnikov, Jing Zhang, Travis J. Ness, Sudeep Banjade, Saket R. Bagde, Viktoriya G. Stancheva, Xiao-Han Li, Kaixian Liu, Zhi Zheng, Daniel J. Barrero, Upasana Roy, Jochen Kuper, Israel S. Fernández, Barnabas Szakal, Dana Branzei, Josep Rizo, Caroline Kisker, Eric C. Greene, Sue Biggins, Scott Keeney, Elizabeth A. Miller, J. Christopher Fromme, Tamara L. Hendrickson, Qian Cong[#], and David Baker[#]. "Computed structures of core eukaryotic protein complexes." *Science 374 (6573), eabm4805.*

[†]These authors contributed equally to this work.

Abstract

Protein-protein interactions play critical roles in biology, but the structures of many eukaryotic protein complexes are unknown, and there are likely many interactions not yet identified. We take advantage of advances in proteome-wide amino acid coevolution analysis and deep-learning-based structure modeling to systematically identify and build accurate models of core eukaryotic protein complexes within the *Saccharomyces cerevisiae* proteome. We use a combination of RoseTTAFold and AlphaFold to screen through paired multiple sequence alignments for 8.3 million pairs of yeast proteins, identify 1,505 likely to interact, and build structure models for 106 previously unidentified assemblies and 806 that have not been structurally characterized. These complexes, which have as many as 5 subunits, play roles in almost all key processes in eukaryotic cells and provide broad insights into biological function.

Main text

Yeast two hybrid (Y2H), affinity-purification mass spectrometry (APMS), and other high-throughput experimental approaches have identified many pairs of interacting proteins in yeast and other organisms (*1*-*5*), but there are discrepancies between sets generated using the different methods and considerable false positive and false negative rates (*6*-*8*). Because residues at protein-protein interfaces are expected to coevolve, the likelihood that any two proteins interact can be assessed by identifying and aligning the ortholog sequences of the two proteins in many different species, joining them to create paired multiple sequence alignments (pMSA), and then determining the extent to which changes in the sequences of orthologs for the first protein covary with ortholog sequence changes for the second (*9*,*10*). Such amino acid coevolution has been used to guide modeling of complexes for cases in which the structures of

the partners are known (*11*,*12*), and to systematically identify pairs of interacting proteins in Prokaryotes with accuracy higher than experimental screens (*9*). Recent deep-learning-based advances in protein structure prediction (*13*,*14*) have the potential to increase the power of such approaches as they now enable accurate modeling not only of protein monomer structures but also protein complexes (*13*).

We set out to combine proteome wide coevolution-guided protein interaction identification with deep learning based protein structure modeling to systematically identify and determine the structures of eukaryotic protein assemblies (Fig. 1A). We faced several challenges in directly applying to eukaryotes the statistical methods we had found effective in identifying coevolving pairs in prokaryotes (8). First, far fewer genome sequences are available for eukaryotes than prokaryotes, and the average number of orthologous sequences (excluding nearly identical copies with > 95% sequence identity) is on the order of 10,000 for bacterial proteins, but 1,000 for eukaryotic proteins. Thus, multiple sequence alignments for pairs of eukaryotic proteins contain fewer diverse sequences, making it more difficult for statistical methods to distinguish true coevolutionary signal from the noise. Second, eukaryotes in general have a larger number of genes, making comprehensive pairwise analysis more computationally intensive, and increasing the background noise. Third, mRNA splicing in eukaryotes further increases the number of protein species, resulting in errors in gene predictions and complicating sequence alignments. Fourth, eukaryotes underwent several rounds of genome duplications in multiple lineages (*15*), and it can be difficult to distinguish orthologs from paralogs, which is important for detecting coevolutionary signal because the protein interactions of interest are likely to be conserved in orthologs in other species but less so in paralogs.

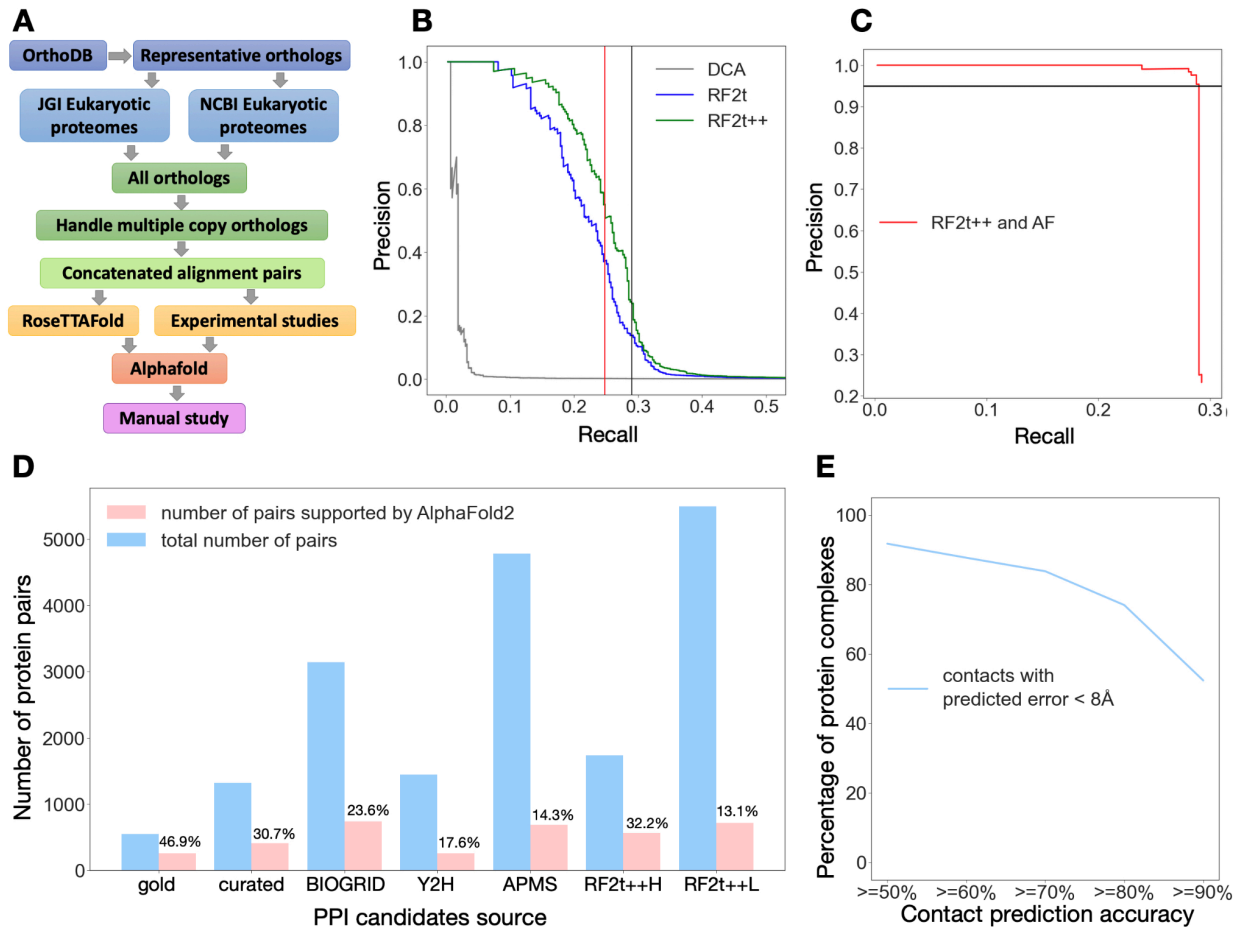


Figure 1. Evaluation of protein interaction and structure prediction accuracy
(**A**) The PPI screen pipeline. (**B**) Performance (precision at different levels of recall) of different methods in picking out gold standard PPIs from the set of 4.3 million pMSAs (Precision: number of true positives above a cutoff divided by the total number of pairs above this cutoff; recall: number of true positives above cutoff divided by the total number of true positives (gold standard PPIs). Pairs were ranked by the top coevolution score or contact probability between residue pairs. DCA: Direct coupling analysis. RF2t: top contact probability between residues of two proteins by RF 2-track model. RF2t++, optimized RF2t (see methods). RF2t++ predictions better than the cutoff shown in vertical black line (RF2t++L in Fig. 1C) were processed with AF; recall of gold standard PPIs at this cutoff is 29%; and precision is 23%. RF2t++ results with a more stringent cutoff (red vertical line) are also shown in Fig 1C (RF2t++H). (**C**) AF contact probability ranking of complexes selected by RF2t++ in panel (**B**); complexes with scores above the horizontal black line were selected for further analysis. (**D**) Number of high scoring (top contact probability > 0.67) AF predictions in PPI sets from different sources. (**E**) Distribution of percent of AF predicted inter-protein contacts with predicted error < 8Å found in contact (< 8Å) in closely-related experimental structures.

To mitigate the first three challenges, we chose to predict protein complexes for yeast *Saccharomyces cerevisiae* as the starting point because there are a large number of fungal genomes (*16*), the genome is relatively small (6,000 genes total), and there is relatively little

mRNA splicing (*17*). Furthermore, because the interactome of yeast has been extensively studied, there is a "gold standard" set (see supplemental Methods) of known interactions to evaluate the accuracy of predicted interactions and structures.

To distinguish orthologs from paralogs, we started from OrthoDB (*18*), a hierarchical catalog of orthologs across 1,271 Eukaryote genomes, and supplemented each orthologous group with sequences from 4,325 Eukaryote proteomes we assembled from NCBI (https://www.ncbi.nlm.nih.gov/genome) and JGI (*19*). Among these, 2,026 are fungal proteomes spanning 14 phyla (47 classes). We compared the sequences for each protein in each of the additional 4,325 proteomes against those of the most closely related species in the OrthoDB database, and used the reciprocal best hit criterion (*20*) to identify orthologs (fig. S1); these were then added to the corresponding orthologous group. A complication is that each species frequently contains multiple proteins belonging to the same orthologous group, leading to ambiguity in determining which protein should be included in pMSAs. These multiple copies may represent alternatively spliced forms of the same gene, parts of the same gene that were split into multiple pieces due to errors in gene prediction, or recent gene expansions specific to certain lineages. We dealt with these possibilities by keeping only the longest isoform of each gene, merging pieces of the same gene, and selecting the copy with the highest sequence identity to single-copy orthologs in other species. For 4,090 out of ~6,000 yeast proteins, we were able to assign a single-copy yeast protein to orthologs in other species, and we generated pMSAs for all 4,090 * 4,089 / 2 = 8,362,005 pairwise combinations of these proteins (fig. S2). We focused on 4,286,433 pairs with alignments containing over 200 sequences to increase prediction accuracy and less than 1,300 amino acids to accelerate computation (fig. S3).

In a first set of calculations, we found that even with the advantages of *S. cerevisiae* and improved ortholog identification, the statistical method (Direct Coupling Analysis, DCA) we had

used in our previous coevolution-guided protein-protein interaction (PPI) screen in Prokaryotes (*9*) (the more accurate GREMLIN (*11*) method is too slow for this) could not effectively distinguish a "gold standard" set of 768 yeast protein pairs known to interact (*5*) (http://interactome.dfci.harvard.edu/S_cerevisiae/) from the much larger set (768,000 pairs) of primarily non-interacting pairs (Fig. 1B, grey curve, area under the curve: 0.016). Progress required a more accurate and sensitive, but still rapidly computable, method to evaluate protein interactions based on pMSAs.

We explored the application of the deep learning based structure prediction methods, RoseTTAFold (RF) and AlphaFold (AF), to this problem. Even though RF was originally trained on monomeric protein sequences and structures, it can accurately predict the structures of protein complexes given pMSAs with a sufficient number of sequences (*13*). We found that a lighter-weight (10.7 million parameters) RF two-track model (figs. S4, S5) provided a good tradeoff between compute time and accuracy: the model requires 11 seconds (about 100 times faster than AF) to process a pMSA of 1,000 amino acids on a NVIDIA TITAN RTX graphic processing unit, and it can effectively distinguish gold standard PPIs amongst much larger sets of randomly paired proteins. The very short time required to analyze an individual pMSA made it possible to process all 4.3 million pMSAs. This method considerably outperformed DCA in distinguishing gold standard interactions from random pairs (Fig. 1B, blue curve, area under the curve: 0.219), using the highest predicted contact probability over all pairs of residues in the two proteins as a measure of the propensity for two proteins to interact (fig. S6). Performance was further improved (Fig. 1B, green curve, area under the curve: 0.248) by correcting overestimations of predicted contact probabilities between the C-terminal residues of the first protein and the N-terminal residues of the second protein, and of predicted interactions for a subset of proteins showing hub-like interactions with many other proteins (see Methods and figs. S7, S8). The much better performance of RF than DCA likely stems from the extensive

information on protein sequence-structure relationships embedded in the RF deep neural network; DCA by contrast operates solely on protein sequences with no underlying protein structure model.

We next explored whether AF residue-residue contact predictions could further distinguish interacting from non-interacting protein pairs. Like RF, AF was trained on monomeric protein structures, but given the good results with 2-track RF on protein complexes, and the higher accuracy of AF (also a 2-track network followed by a 3D structure module) on monomers, we reasoned that it might similarly have higher accuracy than RF on complexes; to enable modeling of protein complexes using AF, we modified the positional encoding in the AF script (see Methods). AF was too slow to be applied to the entire set of 4.3 million pMSAs (this would require 0.1-1 million GPU hours); instead we applied AF to the 5,495 protein pairs with the highest RF support (indicated by the black vertical line in Fig. 1B). Using the highest AF contact probability over all residue pairs as a measure of interaction strength, we found that the combination of RF followed by AF provided excellent performance (Fig. 1C and figs. S9, S11). Almost all the gold standard pairs were ranked higher than the negative controls, allowing selection of a set of 715 candidate PPIs with an expected precision of 95% at an AF contact probability cutoff of 0.67 (black horizontal line in Fig. 1C); we refer to this RF plus AF procedure as the *de novo* PPI screen, and the resulting set of predicted interactions, the *de novo* PPI set, below.

Due to the tradeoff between compute time and accuracy, and the necessity of setting a stringent threshold to avoid large numbers of false positives given the very large number of total pairs, we were concerned that some interacting proteins might not coevolve sufficiently to be identified robustly in our all-vs-all RF screen. Given the excellent performance of AF in distinguishing gold standard interactions amongst the RF filtered pairs, we also applied AF to pMSAs for PPIs

reported in the literature, including those identified in high throughput experimental screens. Similarly to our *de novo* PPI screen procedure, we considered protein pairs with AF contact probability larger than 0.67 to be confident interacting partners. We found that 47% of the gold standard PPIs were confidently predicted, with lower ratios (31% and 24%) for candidate PPIs from the literature (http://interactome.dfci.harvard.edu/S_cerevisiae/download/LC_multiple.txt) (*3*) or supported by low-throughput experiments according to BIOGRID (*21*) (Fig. 1D). The ratio of confidently predicted PPIs is even lower for protein pairs identified by Y2H (18%) or APMS (14%) screens (table S1), consistent with the known larger fraction of false positives in large-scale experimental screens (8,*22*). The fast RF 2-track model used in the *de novo* screen has comparable or better accuracy than the large-scale experimental screens when assessed in this way: with a high stringency RF cutoff (indicated by the red vertical line in Fig. 1B), the fraction of confidently predicted pairs among PPIs identified by RF is 32%, similar to the accuracy of low-throughput experiments; with a lower stringency cutoff (indicated by the black vertical line in Fig. 1B), this fraction becomes closer to that of the large-scale experimental screens but somewhat fewer true PPIs are missed than with the higher cutoff (Fig. 1D).

In total, we identified 715 likely interacting pairs from the "*de novo* RF → AF" screen, and 1,251 from the "pooled experimental sets → AF" screen, of which 461 overlap, resulting in a total of 1,505 PPIs (see figs. S11-S13 for interface size and secondary structure distributions for the predicted complex structures). Out of these, 699 have been structurally characterized, 700 have some supporting experimental data from literature and databases, and 106 are not to our knowledge previously described. To evaluate the accuracy of the predicted 3D structure of protein complexes, we used as a benchmark the 699 pairs with experimental structure in the Protein Data Bank (PDB). For 92% of these pairs, at least 50% of confident (predicted aligned error < 8 Å) AF-predicted contacts are present in the experimental structures (Fig. 1E, and fig. S14). The models do miss many contacts observed in the experimental structures however,

likely due to lower residue-residue co-evolution (fig. S15).

With these benchmark results providing confidence in the accuracy of the new complex interaction predictions and 3D models of the predicted complexes, we analyzed the structure models for the 806 complexes for which high resolution structural information was not available. We classified these models into groups based on their biological functions, and provide examples of complexes in each functional class in Figs. 2-4. A first set of complexes are involved in maintenance and processing of genetic information: DNA repair, mitosis and meiosis checkpoints, transcription, and translation (Fig. 2). A second set of complexes play roles in protein translocation, transport through the secretory pathway, the cytoskeleton and cell organelles (Fig. 3). A third set of complexes are involved in metabolism (Fig. 4). Examples of protein complexes in which proteins of unknown function are predicted to interact with well characterized ones are shown in Fig. 4: these interactions provide hints about the function of the uncharacterized proteins and could help identify new components of previously characterized assemblies. In cases where three or more proteins were predicted to mutually interact, we generated models of the full assemblies by using as input an sequence alignment for the entire complex (see Methods). Examples of these larger assemblies are shown in Fig. 5; in most cases the pairwise interactions are quite similar to those for the independently built binary complexes, but simultaneous modeling of the full complex has the advantage of allowing conformational changes that could accompany full assembly.

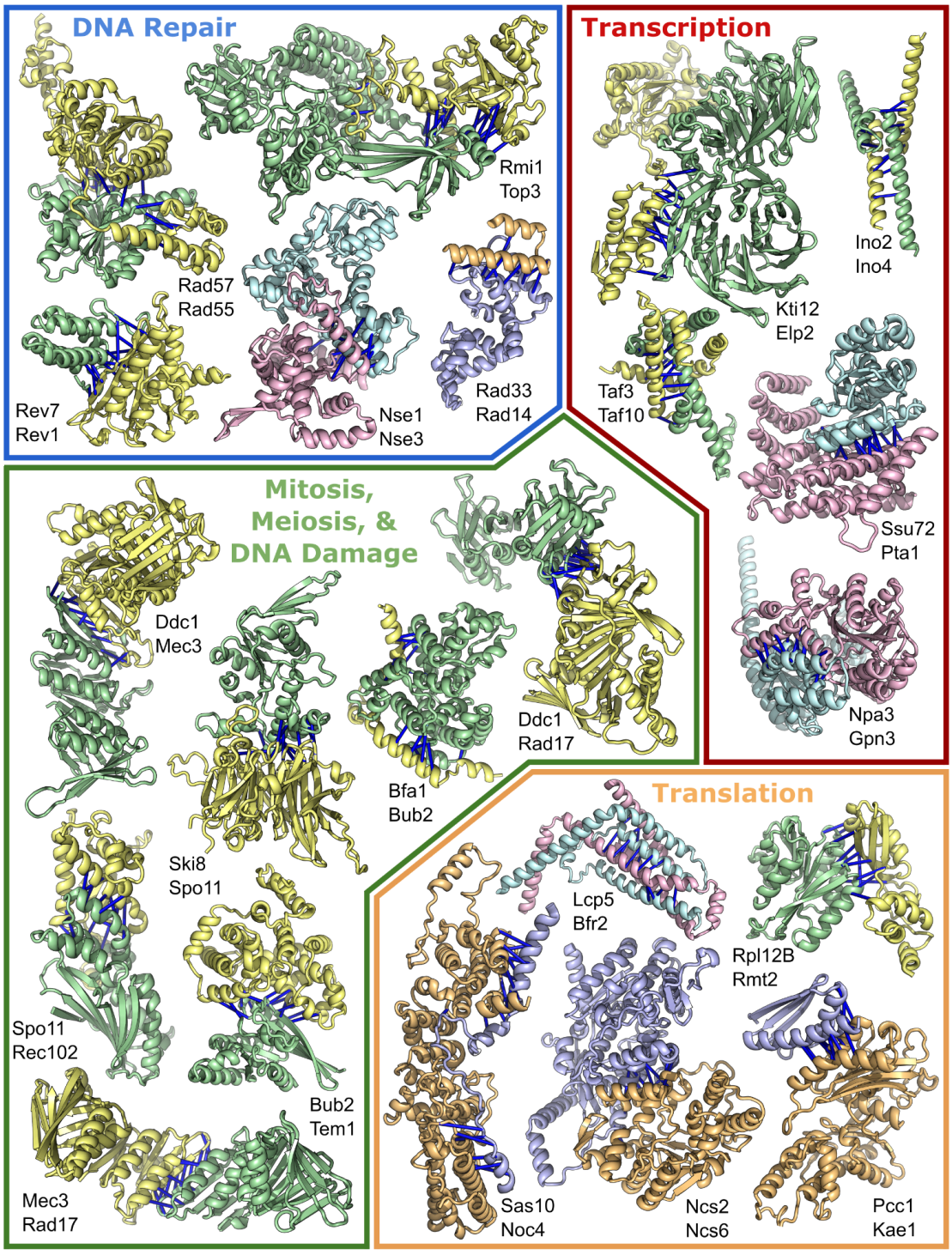


Figure 2. Protein complexes involved in transcription, translation, and DNA repair
Top predicted residue-residue contacts are indicated with bars. Pair color indicates the method of identification: pairs from the "pooled experimental sets → AF" screen are yellow and green, pairs from the "*de novo* RF → AF" screen are in blue and light-orange; and pairs present in both datasets are teal and pink. Full names of these proteins are in table S2.

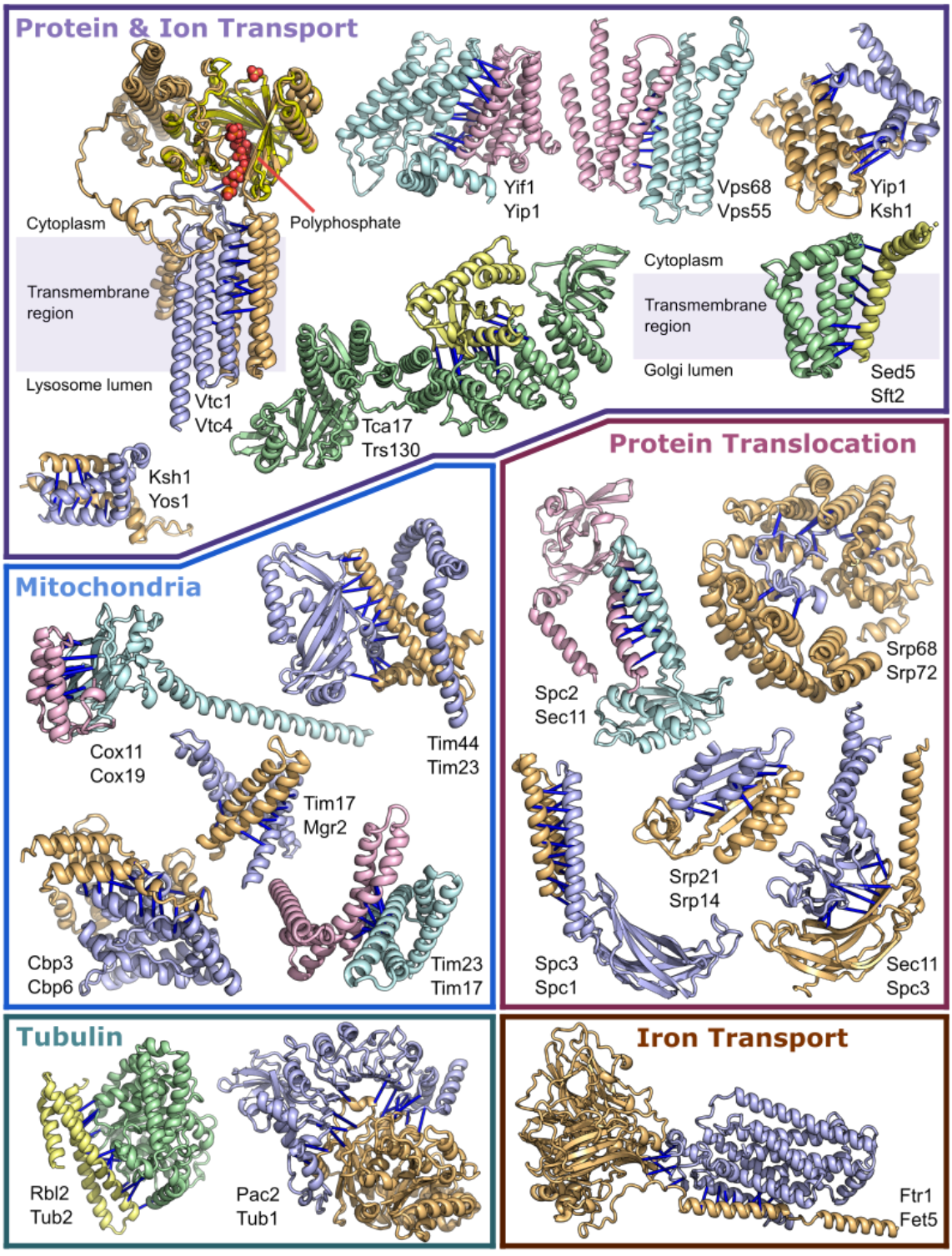


Figure 3. Protein complexes involved in molecule transport, membrane translocation, and mitochondria

Bars and coloring as in Fig 2. Full names for proteins are in table S3. Membrane spanning regions are annotated on Vtc1-Vtc4 and Sed5-Sft2. Top left: model of Vtc1-Vtc4 complex, with superimposed crystal structure (PDB: 3G3Q, Chain A) of the VTC4 (bright yellow) with phosphate bound (red balls).

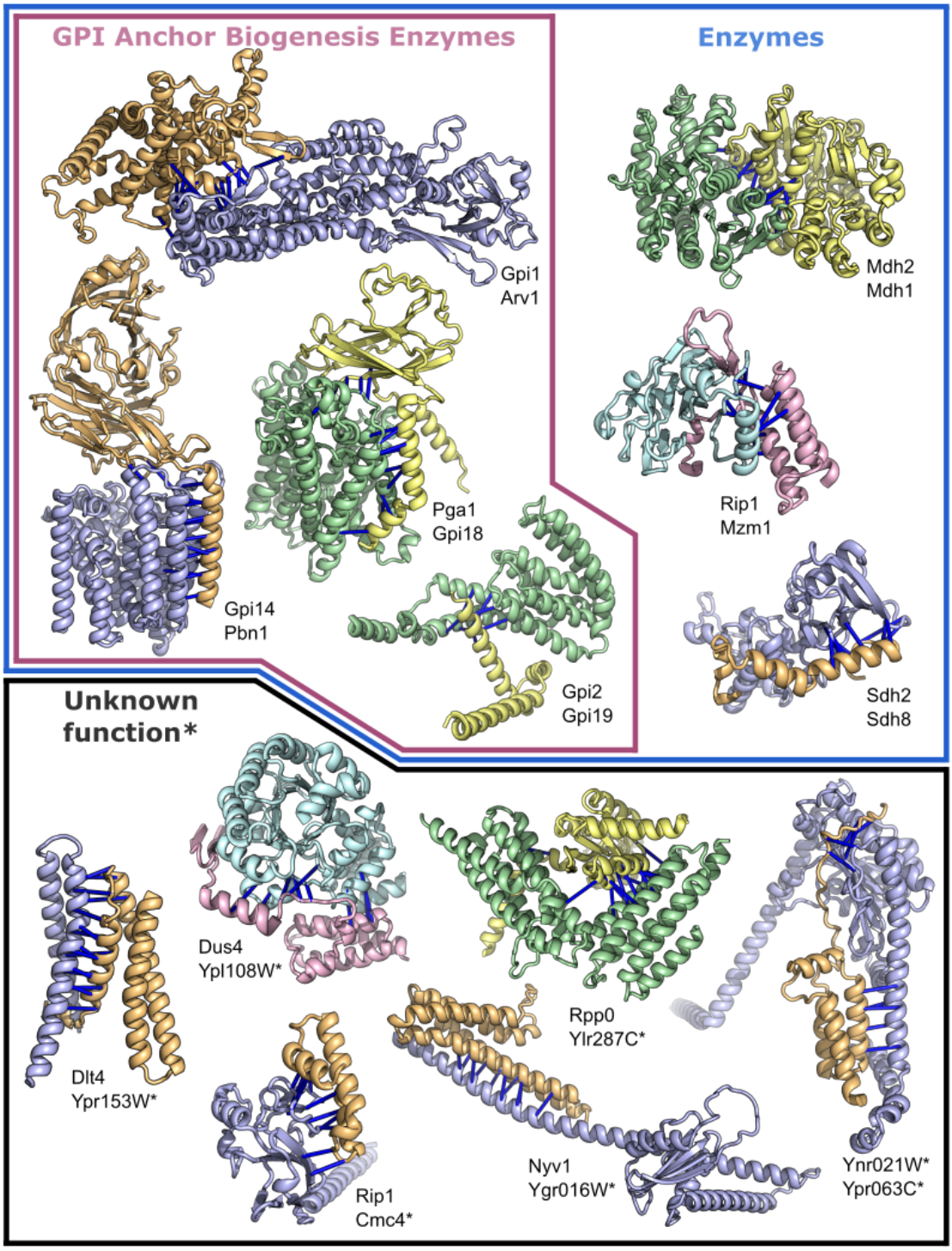


Figure 4. Protein complexes involved in metabolism, GPI anchor biosynthesis or including a protein of unknown function

Coloring is as in Fig. 2-3. Proteins annotated in the Uniprot database as uncharacterized proteins are denoted with an asterisk. Full names for these proteins are in table S4.

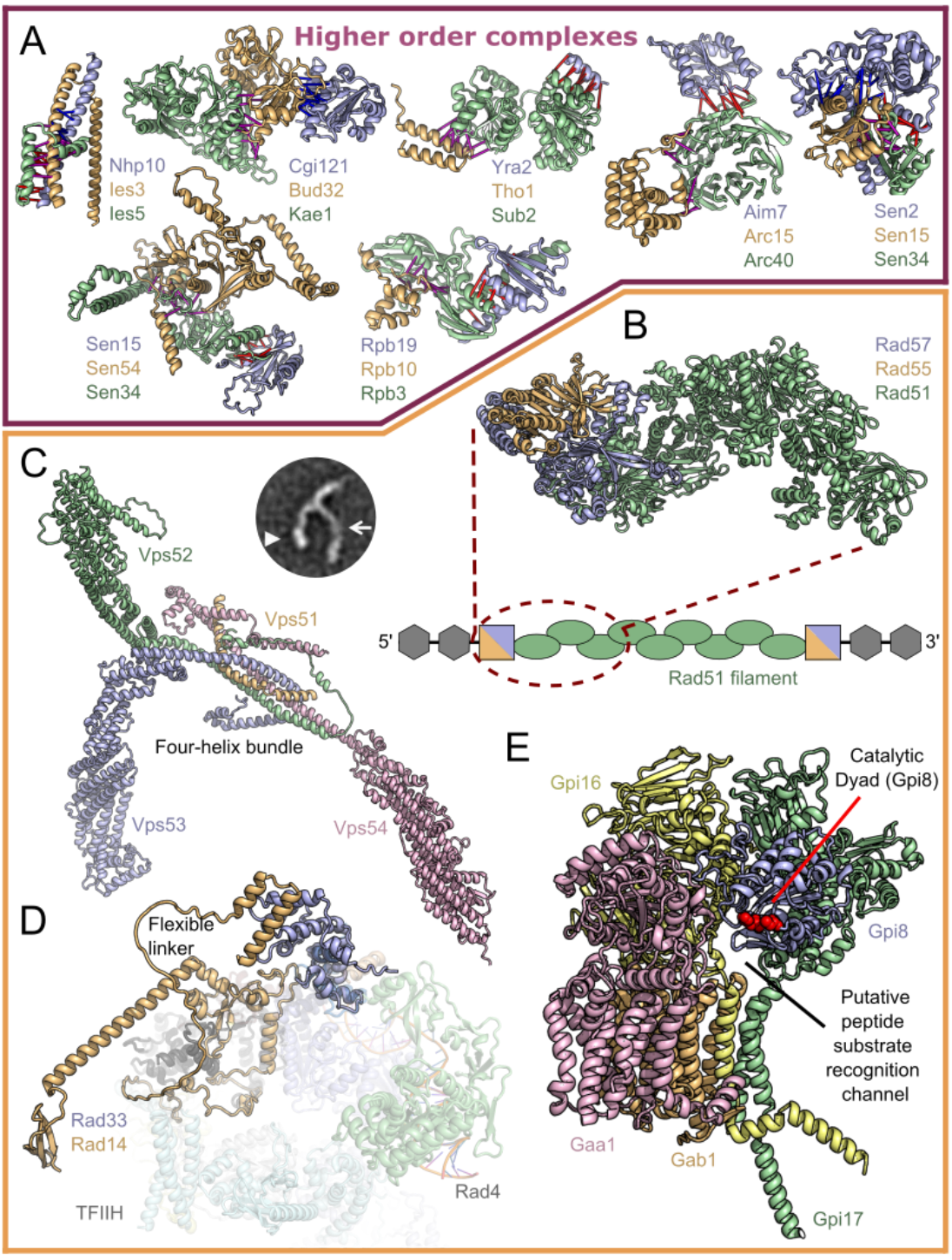


Figure 5. Higher order protein complexes
(**A**) Top predicted residue-residue contacts for trimers are indicated with bars. Bar color corresponds to the interacting protein pair; protein 1:2 are blue, 1:3 are red, 2:3 are purple. Full names of each protein within the complex are in table S5. (**B**) Model of Rad55-Rad57-Rad51 and cartoon depiction of placement of this complex in the larger Rad51 filament. Additional information in fig. S18. (**C**) GARP complex model constructed by predicting structure of central hetero-oligomeric helical bundle, and superimposing models of individual components onto this. 2D class average of GARP complex with minor adaptation (77);

reprinted by permission from Springer Nature Customer Service Center GmbH: Springer Nature, *Nature Structural and Molecular Biology*, CATCHR, HOPS and CORVET tethering complexes share a similar architecture, H-T Chou, D. Dukovski, M.G. Chambers, K.M. Reinisch, and T. Walz, 2016). Alternative GARP models are in fig. S24. (**D**) Rad33-Rad14 complex model superimposed onto previously determined TFIIH/Rad4-Rad23-Rad33 complex structure (7k04). See fig. S19 for additional details. (**E**) GPI-T pentamer model highlighting a possible peptide substrate recognition channel adjacent to the catalytic dyad. See fig. S27 for additional details.

It is not possible to analyze the functional implications of all of the new complexes in a single paper. Instead, as an illustration of the insights which can be gained from these, we focus on a few selected examples in the following sections. To enable broader study of the functional implications of the full set of models, we have made them available at https://modelarchive.org/doi/10.5452/ma-bak-cepc and additional information is provided in the supplemental Excel file.

Complexes involved in DNA homologous recombination and repair

The homologous recombination required for accurate chromosome segregation during meiosis is initiated by DNA double-strand breaks made by Spo11 (*23*). Spo11 is essential for sexual reproduction in most Eukaryotes (*24*,*25*), but mechanistic insight has been limited by a deficit of high-resolution structural information. We predict the structures of complexes of Spo11 with its essential partners Ski8 and Rec102 (Fig. 2 and figs. S16, S17). The predicted Spo11–Ski8 structure is supported by crosslinking and mutagenesis data (*26*,*27*). Our model resembles a previous model based on the Ski3–Ski8 complex, with Ski8 contacting a sequence in Ski3 that is similar to the sequence $QREIF_{380}$ in Spo11 (*27*,*28*) (fig. S17A), but suggests a more extensive interaction surface than previously appreciated, involving an insertion in Ski8 that is present in *Saccharomyces* species but not in *Schizosaccharomyces pombe* and *Sordaria macrospora*, where Ski8 is also required for meiosis (*29*,*30*) (fig. S17B,C). Rec102 was proposed to be a remote homolog of the transducer domain of the Top6B subunit of archaeal topoisomerase VI

(*31*), which couples ATP-dependent dimerization of Top6B subunits to DNA cleavage by Top6A subunits (*32*). Our predicted Rec102–Spo11 complex resembles the Top6A–Top6B interface: a four-helix bundle consisting of two C-terminal helices from Rec102 and two helices from Spo11 (the first helix of the winged helix domain (WHD) plus a more N-terminally located helix) (fig. S17D). Alanine substitutions in this portion of Rec102 disrupt interaction with Spo11 and block meiotic recombination *in vivo (27)*. The model clarifies the Spo11 portion of this interface, which was not well structured in previous homology models (*27*,*31*). Both Rec102 and Top6B have long, helical arms that feed into the Spo11 interface; our model predicts a different angle for this arm and contains a kink that corresponds to a conserved sequence motif $EYPMVF_{192}$ in *Saccharomyces* that is missing in both archaeal TopoVI and mammals (fig. S17D,E). Mutations in this region can suppress *rec104* conditional alleles (*33*), suggesting that this part of Rec102 is important for integrating Rec104 function into the Spo11 core complex.

The highly conserved Rad51 protein central to DNA repair carries out key reactions during homologous recombination, and mutations in human paralogs are associated with Fanconi anemia and multiple types of cancer (*34*). Rad51 paralogs can be positive regulators of Rad51 activity (*35*); in yeast the Rad51 paralogs Rad55 and Rad57 form a stable homodimer that accelerates assembly of Rad51 filaments on single–stranded DNA (ssDNA) during homologous recombination through a transient interaction with Rad51 (*36*). The lack of structural data for the Rad55–Rad57 complex and its interface with Rad51 has limited mechanistic understanding of this process. We generated a model of the trimeric Rad55–Rad57–Rad51 complex, which in combination with the known Rad51 filament structure (*37*), suggests that Rad55–Rad57 binds at the 5’ end of the Rad51 filament where it could promote growth of the Rad51 filament in a directional manner (Fig. 5B and fig. S18).

Nucleotide excision repair (NER) requires a search for lesions in DNA that is mediated by a

conserved complex containing Rad4 (XPC), Rad23 (HR23B) and Rad33 (Centrin2) in yeast. The Rad4–Rad23–Rad33 complex is essential for global genome NER and is the major player in initial damage recognition (*38*). Rad14 (XPA) is recruited at a later stage and activates the helicase Rad3 (XPD) subunit of the general transcription and DNA repair factor IIH complex (TFIIH, consisting of Rad3, Ssl2, Ssl1, Tfb1, Tfb2, Tfb4, and Tfb5) through the release of the TFIIK (CAK) complex following interactions with the TFIIH subunits Tfb5 (p8) and Ssl2 (XPB), and double stranded DNA (*39*). The structures of Rad14 that are currently available only comprise the extended DNA binding domain and lack the N- and C-terminus, where the latter interacts with Tfb5. We generated a model of the complex between full length Rad14 and Rad33 that resolves much of the current structural ambiguity in this system (Fig. 2 and fig S19B), shedding light on how Rad14 may be recruited to the Rad4–Rad23–Rad33 complex. Placing this model into a cryo EM map comprising XPA (Rad14) and TFIIH bound to DNA (39) suggests how the Rad14 C-terminus, which fits into previously unmodeled density, interacts with TFIIH. The long central helix observed in the Centrin2 (Rad33) structure (*40*) is kinked about 90° in our Rad33-Rad14 complex model (fig. S19B); both conformations are feasible and are compatible for the interaction with Rad14. In a recent cryo EM structure of the TFIIH/Rad4–Rad23–Rad33 initial recognition complex (*41*), only the C- terminal part of Rad33 was determined. Superposition of Rad33 in the Rad33-Rad14 complex model onto this structure (Fig. 5D) shows how Rad14 can interact with the Rad4–Rad23–Rad33 recognition complex (*38*,*42*) while maintaining the TFIIH interaction, bridging the steps of initial damage recognition and damage verification. Our model suggests that Rad14 and Rad4 can be present at the same time in the repair cascade; crosstalk between these important proteins could modulate downstream events.

Complexes involved in translation and ribosome regulation

Throughout evolution the eukaryotic machinery for protein production has expanded in size and complexity (*43*), which facilitated the development of sophisticated mechanisms for the regulation of gene expression at the post-transcriptional level (*44*) and increased integration with the cellular environment (*45*). The expanded complexity of the eukaryotic translational machinery came at the cost of a highly complex process for ribosome maturation (*46*). We generate models of complexes which had not been structurally characterized previously that involve components of the translation apparatus (Fig. 2 and fig. S20). Two complexes, Rpl12B–Rmt2 and Rpl7A–Fpr4, involving enzymes that introduce protein modifications such as arginine methylations or proline isomerizations (*47*), provide insight into mechanisms that expand the chemical diversity of ribosomal proteins at functional sites (*48*) and possibly regulate translation (*49*). A complex between components of the U3 ribosome-maturation factor and a protein involved in the regulation of glycerol, Lcp5–Sgd1 (*50*), could play a role in coupling translation with metabolism. A complex between eIF2B, an auxiliary factor for eIF2 recycling after GTP hydrolysis, and transcriptional factor regulator Dig2 could help couple translation and transcription: the delivery of the first aminoacyl-tRNA (Met-tRNA$_i^{Met}$) is a key event in eukaryotic translation regulation by the GTPase eIF2 (*51*) and targeting eIF2 via its nucleotide exchanger eIF2B is a basal mechanism of translation regulation. This possible cross-talk between ribosome-maturation pathways and metabolic sensors, and translation initiation regulators such as eIF2 with transcription factors suggests exciting new avenues to further map the highly integrated nature of translation within eukaryotic cells.

Complexes involving ubiquitin and small ubiquitin-like modifier (SUMO) ligases

Reversible covalent modifications of proteins with ubiquitin and SUMO modulate protein-protein

interactions, cellular localization, and stability (*52*). SUMO E3 ligases facilitate SUMO transfer, and Siz1, Siz2, Mms21, and Zip3 are the known SUMO ligases in budding yeast (*52*). Our model of the Siz2 and Mms21 SUMO ligase complex (fig. S21A) suggests that both E3s could act jointly to modify DNA associated substrates perhaps through the DNA binding SAP domain of Siz2 (*53*) or involving the Mms21 (Nse2) containing Smc5–6 complex which modulates DNA recombination, replication and repair (*54*,*55*). The Smc5–6 complex contains another RING-finger E3 ligase-like subunit, Nse1 (*56*) that interacts with Nse3 and Nse4. Our model of the yeast Nse1–Nse3–Nse4 complex (fig. S21B) is similar to a structure determined for the *Xenopus laevis* complex, despite the sequences of the yeast and Xenopus proteins being too distant for similarity to be detectable by BLAST.

SUMO-targeted ubiquitin ligases (STUbLs) are ubiquitin ligases that recognize SUMO-modified proteins. A STUbL consisting of the Slx8 ubiquitin ligase and the associated protein Slx5 functions in proteasome-mediated turnover of several proteins associated with DNA replication, repair and chromosome structure (*57*-*59*). Our model of the Slx5-Slx8 complex (fig. S21C) provides insight into how these two proteins may collectively recognize their substrates. In addition, we generated a lower confidence but intriguing model of a previously undescribed complex between Slx8 and Cue3 (Coupling of ubiquitin conjugation to endoplasmic reticulum (ER) degradation protein 3) (fig. S21D), possibly linking ubiquitination of substrates to protein degradation in ER.

Complexes involved in chromosome segregation

The heterodecameric complex DASH/Dam1 (Dam1c) is composed of 10 proteins: Ask1, Dad1, Dad2, Dad3, Dad4, Dam1, Duo1, Hsk3, Spc19, and Spc34 which come together to form a “T” shape, and can further oligomerize into rings (*60*,*61*). During mitosis, these heterodecamers

strengthen the attachment between kinetochores and microtubules (*62*) by oligomerizing to form either partial or complete rings around microtubules and further contacting kinetochore components (*63*-*65*). Microtubules are required for in-vivo ring formation, but a structure of the Dam1c ring complex from *Chaetomium thermophilum* was determined in the absence of microtubules using monovalent salts (*66*). We generated structure models of nine binary complexes (Dad2-Ask1, Dad2-Hsk3, Dad2-Spc1, Dad4-Hsk3, Dam1-Duo1, Duo1-Dad1, Spc19-Dad1, Spc34-Duo1, and Spc34-Spc19) that encompass several members of Dam1c (fig. S22). These complexes are largely consistent with the Dam1c structure, suggesting that the findings from the thermophile structure can likely be extended to *S. cerevisiae*. We went beyond previous structural data by predicting the structure of a potential inter-decamer interaction between Spc19 and Dad1 involving a flexible loop of Spc19 and the N-terminal region of Dad1, which could be important for ring formation in vivo (*66*).

Complexes involved in molecule transport and membrane trafficking

The small membrane protein Ksh1 is conserved across eukaryotes, essential for growth, and plays an unknown role in protein secretion (*67*). We predicted structures of complexes between Ksh1 and two membrane proteins reported to form a complex: Yos1 and Yip1. This complex also includes Yif1 and interacts with Rab GTPases (*68*) (Fig. 3). These structures suggest Ksh1 is a fourth member of this enigmatic complex essential to the secretory pathway, and explains how Ksh1 can play a role in secretion despite its small size of 72 amino acids.

The vacuolar transporter chaperone (VTC) is a 5-subunit complex that synthesizes polyphosphate to regulate cellular phosphate levels (*69*). Structures are only known for some soluble portions of this complex, including the catalytic domain of the Vtc4 subunit (*70*). Our

model of the previously non-structurally characterized Vtc1–Vtc4 subcomplex suggests that the cytosolic active site is positioned by the complex to feed the polyphosphate product through a membrane pore into the lumen of the lysosome (Fig. 3).

The ESCRT-III complex is involved in a number of cellular membrane remodeling pathways, including receptor downregulation, membrane repair, and cell division (*71*,*72*). Our predicted interface between the Vps2 and Vps24 subunits of the ESCRT-III complex resembles the polymerization interface of a different ESCRT-III subunit Snf7 (*73*), providing insight into the roles of these previously uncharacterized ESCRT-III subunits, and highlighting the generality of this mode of interaction in ESCRT-III complexes. Notably, previously unpublished mutations (fig. S23) in Vps24 that prevent ESCRT function in multivesicular body sorting are located on the predicted interface between Vps2 and Vps24, supporting our model and the functional importance of the Vps2–Vps24 interaction. Vps55 and Vps68 are conserved membrane proteins that are important for endosomal cargo sorting; our predicted structure (Fig. 2) of their interaction provides clues about the mechanism of their function (*74*).

The GARP complex is a multisubunit tethering complex (MTC) that mediates docking and fusion of vesicles with the Golgi apparatus (*75*). Our approach generated models for binary complexes involving the four GARP subunits, and we further modeled the entire complex (fig. S24A). In this model, the four subunits assemble through a four-helix bundle. In each of the three larger subunits, Vps52, Vps53, and Vps54, C-terminal domains comprising “CATCHR” folds emanate from the bundle. This architecture resembles portions of the cryo-EM structure of the Exocyst complex, a distinct MTC that mediates fusion of vesicles at the plasma membrane (*76*), which possesses two separate four-helix bundles organizing its eight subunits. In our prediction, the “CATCHR” domains appear to be somewhat flexibly linked to the central four-helix bundle, and hence we overlaid the structure predictions for Vps52, Vps53, and Vps54, respectively, onto the

central four-helix bundle (Fig. 5C and fig. S24B). The resulting model has a striking resemblance to previously published 2D classes (fig. S24C) from a negative-stain EM analysis of the GARP complex (*77*). These predictions will facilitate structure-guided experiments to elucidate the mechanism of MTC function.

Golgi-resident protein, Grh1, forms a tethering complex with Uso1 and Bug1 that interacts with the COPII coat protein complex, Sec23/Sec24. The tether is thought to participate in COPII vesicle capture (*78*,*79*), but the mechanism remains unclear. The C-terminus of Grh1 contains a predicted intrinsically disordered region (IDR) with a net positively charged cluster and a triple-proline motif (fig. S25A,B). Our model of the Sec23–Grh1 complex contains an interface between the Sec23 gelsolin domain and the PPP motif of Grh1 (*80*), and an interface between the Grh1 IDR and Sec23 involving a disorder-to-helical transition (fig. S25C). A similar multivalent interface also drives interaction between Sec23 and the COPII coat scaffolding protein, Sec31 (*81*). Our model suggests that the combinatorial multivalent interaction between Grh1 and Sec23 may compete with the interaction between Sec31 and Sec23 to promote vesicle uncoating; consistent with this model, Grh1 is recruited to GST-Sec23, dependent on the IDR, and competes for Sec31 binding (fig. S25D).

SNARE proteins drive intracellular membrane fusion between transport vesicles and organelles (*82*). Our predicted complex structure between the SNARE Sed5 and the uncharacterized transmembrane protein Sft2 unexpectedly predicted an interaction between transmembrane domains of the two proteins (Fig. 3). SNARE localization is thought to occur through interactions of cytoplasmic domains with cytoplasmic sorting factors, but this prediction, together with genetic evidence (*83*), suggests SNARE localization or function may be subject to additional mechanisms via interactions with transmembrane protein regulators. Membrane fusion requires the formation of a 4-helix bundle (called the SNARE complex) between the vesicle SNARE and

the target membrane SNAREs (*84*,*85*). The bundle is formed by the SNARE motifs, which are 60-70 amino acids with heptad repeats and the ability to form coiled-coil structures. Models of binary complexes of SNARE-motif-containing proteins frequently differ from their classic conformation in the SNARE four-helical bundle (fig. S26A), probably because all four chains are required to form the stable complex (*86*). Indeed modeling the four SNARE proteins (Ufe1, Use1, Sec20, and Sec22) that are known to mediate the fusion between Golgi-derived retrograde transport vesicles with ER (*87*), together resulted in a complex that resembles a typical SNARE complex (*84*) (fig. S26B,C). This example highlights the potential pitfalls of modeling only binary complexes when the functional assembly involves more than two chains.

GPI transamidase complex

Glycosylphosphatidylinositol transamidase (GPI-T) is a pentameric enzyme complex of unknown structure (*88*-*90*) which catalyzes the attachment of GPI anchors to the C-terminus of specific substrate proteins, based on recognition of a C-terminal signal peptide (*91*). GPI-T catalyzes the removal of this signal sequence, replacing it with a new amide bond to an ethanolamine phosphate in the GPI anchor. The five subunits of *S. cerevisiae* GPI-T are Gpi8 (which contains the catalytic active site), Gpi16, Gaa1, Gpi17, and Gab1 (88,*92*,*93*). Our large-scale modeling approach generated models for the following binary complexes: Gpi8-Gpi17, Gab1-Gaa1, Gab1-Gpi17, and Gaa1-Gpi16. We subsequently modeled the full-length, pentameric GPI-T in one shot starting from the sequences of all components (Fig. 5E). Several features of this model are consistent with previous characterization of this enzyme. *S. cerevisiae* GPI-T can be purified as a core heterotrimer, containing only Gpi8, Gpi16, and Gaa1 (*92*); our GPI-T model confirms extensive interactions between the soluble domains of these three subunits. This model also recapitulates the disulfide bond between Gpi8 (Cys85) and Gpi16 (Cys202), previously characterized for human GPI-T (*94*) (the existence of this

disulfide bond in yeast GPI-T has been called into question (*90*)). Gaa1 is essential for binding of the GPI anchor to GPI-T (*95*) and the hydrophobic Gab1 is also predicted to participate in anchor recognition (*88*). Our model positions the transmembrane regions of Gaa1 and Gab1 against each other. The catalytic dyad in Gpi8 (Cys199 and His157) faces these transmembrane domains, and abuts against a highly conserved face of Gaa1, proposed to recognize the GPI anchor glycans (*96*,*97*). In our model, the positions of these subunits are consistent with binding of the GPI anchor to position the modifying amine in the Gpi8 active site for catalysis. Gpi16 is immediately adjacent to these interactions and is likely to also be involved in anchor recognition. In vivo, GPI-T is expected to be a dimer of pentamers, with dimerization occurring on one face of the caspase-like Gpi8 subunit (*92*,*97*,*98*). This decameric complex was too large for us to model computationally; however the pentameric complex we present here leaves open the dimerization face of Gpi8, consistent with probable dimerization. It also suggests that Gaa1 and Gpi17 would participate in dimerization of this enzyme. The functional role of Gpi17 has been elusive, but our model now suggests Gpi17 together with Gpi8 and Gpi16, forms a recognition channel for the C-terminal GPI-T signal peptide (fig. S27). In humans, mutations in GPI-T subunits are associated with neurodevelopmental disorders (*99*). Each subunit contributes to different cancer mechanisms, in some cases by perturbing GPI anchoring of specific receptors and in others by separating from GPI-T to alter disparate signal transduction pathways (*89*). Now, with a structural model in hand, these mechanisms can be examined at a molecular level.

Limitations of the current method

As with any new method, it is important when interpreting the results (our large set of predicted complex structures) to keep in mind the limitations of the approach. First, our study is not comprehensive, so conclusions should not be drawn about absences; in particular we

eliminated proteins that arose from recent duplication due to difficulty in identifying orthologs in other organisms, and thus only surveyed 2/3 of the entire yeast proteome. Second, the approach likely misses interactions restricted to a small set of organisms, or which vary rapidly during evolution, due to weaker co-evolutionary signals. Third, the approach likely works less well for transient interactions which generally involve smaller and weaker interfaces which may be under lower selective pressure, in particular those involving intrinsically disordered regions which are poorly represented in the PDB. The majority of known interactions identified by our approach are likely obligate assemblies and involve ordered structural elements. Fourth, interactions between single hydrophobic or amphipathic helices, such as single transmembrane helices or coiled coils, may be overpredicted (in initial studies of human complexes, interactions solely between single-pass transmembrane regions appear to be over represented). Fifth, and perhaps most importantly, for proteins that form high-order obligate protein complexes, binary complex models may be quite inaccurate, as illustrated by the SNARE example.

Conclusion

Our approach extends the range of large scale deep learning based structure modeling from monomeric proteins to protein assemblies. As highlighted by the above examples, following up on the many new complexes presented here should advance understanding of a wide range of eukaryotic cellular processes and provide new targets for therapeutic intervention. The methods can be extended directly to large scale mapping of interactions in the human proteome, but considerably more compute time will be required given the much larger total number of protein pairs, and models may be somewhat less accurate due to weaker co-evolutionary signal for the subset of human proteins unique to higher eukaryotes and for the many closely related paralogs arising from gene duplication. Investigating interactions of individual proteins or subsets of proteins, for example, deorphanization of orphan receptors, should be immediately accessible

using our approach provided there are sufficient sequence homologues. Training RF and AF on protein complexes should further improve performance of both methods (*100*), particularly for protein pairs with fewer homologues and/or weaker and more transient interactions, and reduce the dependence on ortholog identification. Together with the advances in monomeric structure prediction, our results herald a new era of structural biology in which computation plays a fundamental role in both interaction discovery and structure determination.

Methods

As described in detail in the Supplemental Methods, we developed a multistep bioinformatics and deep learning pipeline for identifying pairs of proteins likely to interact and modeling the three dimensional structures of the corresponding protein complexes. The steps of this pipeline are illustrated schematically in Fig. 1A. First, comprehensive orthologous groups of genes were generated and yeast genes were mapped to these groups; second, multiple sequence alignments of orthologous sequences were generated for each pair of yeast proteins; third, contact probability was computed for each protein pair using RoseTTAFold; and fourth, interaction probability was re-evaluated and complex structures were modeled using AlphaFold. The experimental data-guided PPI screening pipeline is very similar except that in the third stage, instead of using RoseTTAFold, we used experimental data primarily derived from large-scale screens to identify PPI candidates.

Acknowledgements

We thank Eric Horvitz, Nick V. Grishin, Hahnboem Park, and James H. Thomas for helpful discussions, Luki Goldschmidt and Aaron Guillory for computing resource management, and Lance Stewart for logistical support. Additionally, we are grateful to Martin Bard, Trisha N Davis,

David G Drubin, Maitreya J Dunham, Scott D Emr, Frederick Hughson, James Hurley, Kenji Murakami, Nobuhiro Nakamura, Eva Nogales, Randy Schekman, Shu-ou Shan, Soyeon Showman, Kaoru Sugasawa, and Sho Suzuki for their correspondence and biological expertise. We thank Stephen Burley, Brinda Vallat, and John Westbrook at the RCSB Protein Data Bank and Torsten Schwede, Gerardo Tauriello, Andrew Waterhouse, and Stefan Bienert at SWISS-MODEL for hosting our model structures at ModelArchive.

Funding

This work was supported by Microsoft (MB, DB, and Azure compute time and expertise), Amgen (DB and IH), Southwestern Medical Foundation (JP and QC), the Washington Research Foundation (MB and QC), Howard Hughes Medical Institute (DB, SB, SK, and generous compute time on Janelia), National Science Foundation (NSF) Cyberinfrastructure for Biological Research (CIBR, Award # DBI 1937533 to DB and IA), CPRIT training grant (RP210041 to JZ), UK Medical Research Council (MRC_UP_1201/10 to EAM.), HHMI Gilliam Fellowship (DJB), the Deutsche Forschungsgemeinschaft (KI-562/11-1 and KI-562/7-1 to CK), NIH/NIGMS (R21AI156595 to SO, R35GM136258 to JCF, R35NS097333 to JR, R35GM118026 and R01CA221858 to ECG), HHMI fellowship of the Damon Runyon Cancer Research Foundation (DRG2273-16 to SB and DRG2389-20 to KL), AIRC investigator and the European Research Council Consolidator (IG23710 and 682190 to DBr), the Defense Threat Reduction Agency (HDTRA1-21-1-0007 to DB). We also thank The National Energy Research Scientific Computing Center (NERSC) for providing computing time (project m3962 at NERSC).

Author contributions

QC and DB conceived the research; JP and QC prepared the sequence alignments used in the screen; MB implemented the RoseTTAFold pipeline; MB and SO repurposed AlphaFold for

complex modeling; JP, JZ, and QC designed the PPI screening procedure; IRH, MB, IA, and QC carried out the screen; IRH, AK, and QC analyzed and presented the results; IRH, AK, QC and DB coordinated the collaborative efforts; TJN, SB, SRB, VGS, XHL, KL, ZZ, DJB, UR, JK, ISF, BS, DB, JR, CK, ECG, SB, SK, EAM, JCF, and TLH provided biological insights on specific examples; QC and DB drafted the manuscript while all other authors contributed to the description of specific examples; all authors discussed the results and commented on the manuscript.

Competing interests

Authors declare that they have no competing interests.

Data and materials availability

Data and materials availability: Structures of highly confident pairs with accompanying pMSAs and metadata are available at ModelArchive: https://modelarchive.org/doi/10.5452/ma-bak-cepc. RoseTTAFold two-track version is available at https://github.com/RosettaCommons/RoseTTAFold or Zenodo (*101*). AlphaFold was fetched from https://github.com/deepmind/alphafold on July 16th, 2021 (v2.0.0). Code for a GPU implementation of DCA and the modifications to the AlphaFold predictions script are provided in Supplemental Methods.

References

1. T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, Y. Sakaki, A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 4569–4574 (2001).

2. S. R. Collins, P. Kemmeren, X.-C. Zhao, J. F. Greenblatt, F. Spencer, F. C. P. Holstege, J. S. Weissman, N. J. Krogan, Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. *Mol. Cell. Proteomics*. **6**, 439–450 (2007).

3. T. Reguly, A. Breitkreutz, L. Boucher, B.-J. Breitkreutz, G. C. Hon, C. L. Myers, A. Parsons, H. Friesen, R. Oughtred, A. Tong, C. Stark, Y. Ho, D. Botstein, B. Andrews, C. Boone, O. G. Troyanskya, T. Ideker, K. Dolinski, N. N. Batada, M. Tyers, Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. *J. Biol.* **5**, 11 (2006).

4. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, J. M. Rothberg, A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*. **403**, 623–627 (2000).

5. H. Yu, P. Braun, M. A. Yildirim, I. Lemmens, K. Venkatesan, J. Sahalie, T. Hirozane-Kishikawa, F. Gebreab, N. Li, N. Simonis, T. Hao, J.-F. Rual, A. Dricot, A. Vazquez, R. R. Murray, C. Simon, L. Tardivo, S. Tam, N. Svrzikapa, C. Fan, A.-S. de Smet, A. Motyl, M. E. Hudson, J. Park, X. Xin, M. E. Cusick, T. Moore, C. Boone, M. Snyder, F. P. Roth, A.-L. Barabási, J. Tavernier, D. E. Hill, M. Vidal, High-quality binary protein interaction map of the yeast interactome network. *Science*. **322**, 104–110 (2008).

6. O. Kuchaiev, M. Rasajski, D. J. Higham, N. Przulj, Geometric de-noising of protein-protein interaction networks. *PLoS Comput. Biol.* **5**, e1000454 (2009).

7. A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, M. Gerstein, Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* **18**, 529–536 (2002).

8. J. P. Mackay, M. Sunde, J. A. Lowry, M. Crossley, J. M. Matthews, Protein interactions: is seeing believing? *Trends Biochem. Sci.* **32**, 530–531 (2007).

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

10. A. G. Green, H. Elhabashy, K. P. Brock, R. Maddamsetti, O. Kohlbacher, D. S. Marks, Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat. Commun.* **12**, 1396 (2021).

11. S. Ovchinnikov, H. Kamisetty, D. Baker, Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife*. **3**, e02030 (2014).

12. T. A. Hopf, C. P. I. Schärfe, J. P. G. L. M. Rodrigues, A. G. Green, O. Kohlbacher, C. Sander, A. M. J. J. Bonvin, D. S. Marks, Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife*. **3** (2014), doi:10.7554/eLife.03430.

13. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. **373**, 871–876 (2021).

14. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583–589 (2021).

15. A. Meyer, M. Schartl, Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. *Curr. Opin. Cell Biol.* **11**, 699–704 (1999).

16. I. V. Grigoriev, R. Nikitin, S. Haridas, A. Kuo, R. Ohm, R. Otillar, R. Riley, A. Salamov, X. Zhao, F. Korzeniewski, T. Smirnova, H. Nordberg, I. Dubchak, I. Shabalov, MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* **42**, D699–704 (2014).

17. M. Spingola, L. Grate, D. Haussler, M. Ares Jr, Genome-wide bioinformatic and molecular analysis of introns in Saccharomyces cerevisiae. *RNA*. **5**, 221–234 (1999).

18. E. M. Zdobnov, D. Kuznetsov, F. Tegenfeldt, M. Manni, M. Berkeley, E. V. Kriventseva, OrthoDB in 2020: evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **49**, D389–D393 (2021).

19. A. Clum, M. Huntemann, B. Bushnell, B. Foster, B. Foster, S. Roux, P. P. Hajek, N. Varghese, S. Mukherjee, T. B. K. Reddy, C. Daum, Y. Yoshinaga, R. O'Malley, R. Seshadri, N. C. Kyrpides, E. A. Eloe-Fadrosh, I.-M. A. Chen, A. Copeland, N. N. Ivanova, DOE JGI Metagenome Workflow. *mSystems*. **6** (2021), doi:10.1128/mSystems.00804-20.

20. D. P. Wall, H. B. Fraser, A. E. Hirsh, Detecting putative orthologs. *Bioinformatics*. **19**, 1710–1711 (2003).

21. R. Oughtred, J. Rust, C. Chang, B.-J. Breitkreutz, C. Stark, A. Willems, L. Boucher, G. Leung, N. Kolas, F. Zhang, S. Dolma, J. Coulombe-Huntington, A. Chatr-Aryamontri, K. Dolinski, M. Tyers, The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **30**, 187–200 (2021).

22. H. Huang, B. M. Jedynak, J. S. Bader, Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput. Biol.* **3**, e214 (2007).

23. S. Keeney, C. N. Giroux, N. Kleckner, Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell*. **88**, 375–384 (1997).

24. B. de Massy, Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes. *Annu. Rev. Genet.* **47**, 563–599 (2013).

25. H. Murakami, S. Keeney, Regulating the formation of DNA double-strand breaks in meiosis. *Genes Dev.* **22** (2008), pp. 286–292.

26. C. Arora, K. Kee, S. Maleki, S. Keeney, Antiviral protein Ski8 is a direct partner of Spo11 in meiotic DNA break formation, independent of its cytoplasmic role in RNA metabolism. *Mol. Cell*. **13**, 549–559 (2004).

27. C. C. Bouuaert, S. E. Tischfield, S. Pu, E. P. Mimitou, E. Arias-Palomo, J. M. Berger, S. Keeney, Structural and functional characterization of the Spo11 core complex. *Nat. Struct. Mol. Biol.* **28**, 92–102 (2021).

28. F. Halbach, P. Reichelt, M. Rode, E. Conti, The yeast ski complex: crystal structure and RNA channeling to the exosome complex. *Cell*. **154**, 814–826 (2013).

29. S. Steiner, J. Kohli, K. Ludin, Functional interactions among members of the meiotic initiation

complex in fission yeast. *Curr. Genet.* **56**, 237–249 (2010).

30. S. Tessé, A. Storlazzi, N. Kleckner, S. Gargano, D. Zickler, Localization and roles of Ski8p protein in Sordaria meiosis and delineation of three mechanistically distinct steps of meiotic homolog juxtaposition. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12865–12870 (2003).

31. T. Robert, A. Nore, C. Brun, C. Maffre, B. Crimi, H.-M. Bourbon, B. de Massy, The TopoVIB-Like protein family is required for meiotic DNA double-strand break formation. *Science*. **351**, 943–949 (2016).

32. K. D. Corbett, P. Benedetti, J. M. Berger, Holoenzyme assembly and ATP-mediated conformational dynamics of topoisomerase VI. *Nat. Struct. Mol. Biol.* **14**, 611–619 (2007).

33. L. Salem, N. Walter, R. Malone, Suppressor analysis of the Saccharomyces cerevisiae gene REC104 reveals a genetic interaction with REC102. *Genetics*. **151**, 1261–1272 (1999).

34. M. R. Sullivan, K. A. Bernstein, RAD-ical New Insights into RAD51 Regulation. *Genes* . **9** (2018), doi:10.3390/genes9120629.

35. J. S. Filippo, P. Sung, H. Klein, Mechanism of Eukaryotic Homologous Recombination. *Annual Review of Biochemistry*. **77** (2008), pp. 229–257.

36. U. Roy, Y. Kwon, L. Marie, L. Symington, P. Sung, M. Lisby, E. C. Greene, The Rad51 paralog complex Rad55-Rad57 acts as a molecular chaperone during homologous recombination. *Molecular Cell*. **81** (2021), pp. 1043–1057.e8.

37. A. B. Conway, T. W. Lynch, Y. Zhang, G. S. Fortin, C. W. Fung, L. S. Symington, P. A. Rice, Crystal structure of a Rad51 filament. *Nat. Struct. Mol. Biol.* **11**, 791–796 (2004).

38. K. Sugasawa, J.-I. Akagi, R. Nishi, S. Iwai, F. Hanaoka, Two-step recognition of DNA damage for mammalian nucleotide excision repair: Directional binding of the XPC complex and DNA strand scanning. *Mol. Cell*. **36**, 642–653 (2009).

39. G. Kokic, A. Chernev, D. Tegunov, C. Dienemann, H. Urlaub, P. Cramer, Structural basis of TFIIH activation for nucleotide excision repair. *Nat. Commun.* **10**, 2885 (2019).

40. J. R. Thompson, Z. C. Ryan, J. L. Salisbury, R. Kumar, The Structure of the Human Centrin 2-Xeroderma Pigmentosum Group C Protein Complex. *Journal of Biological Chemistry*. **281** (2006), pp. 18746–18752.

41. T. van Eeuwen, Y. Shim, H. J. Kim, T. Zhao, S. Basu, B. A. Garcia, C. D. Kaplan, J.-H. Min, K. Murakami, Cryo-EM structure of TFIIH/Rad4–Rad23–Rad33 in damaged DNA opening in nucleotide excision repair. *Nat. Commun.* **12**, 1–17 (2021).

42. T. Riedl, F. Hanaoka, J.-M. Egly, The comings and goings of nucleotide excision repair factors on damaged DNA. *EMBO J.* **22**, 5293–5303 (2003).

43. S. Klinge, F. Voigts-Hoffmann, M. Leibundgut, N. Ban, Atomic structures of the eukaryotic ribosome. *Trends Biochem. Sci.* **37**, 189–198 (2012).

44. A. G. Hinnebusch, I. P. Ivanov, N. Sonenberg, Translational control by 5’-untranslated regions of eukaryotic mRNAs. *Science*. **352**, 1413–1416 (2016).

45. J. A. Saba, K. Liakath-Ali, R. Green, F. M. Watt, Translational control of stem cell function. *Nat. Rev. Mol. Cell Biol.* (2021), doi:10.1038/s41580-021-00386-2.

46. S. Klinge, J. L. Woolford Jr, Ribosome assembly coming into focus. *Nat. Rev. Mol. Cell Biol.* **20**, 116–131 (2019).

47. K. M. Mulvaney, C. Blomquist, N. Acharya, R. Li, M. J. Ranaghan, M. O'Keefe, D. J. Rodriguez, M. J. Young, D. Kesar, D. Pal, M. Stokes, A. J. Nelson, S. S. Jain, A. Yang, Z. Mullin-Bernstein, J. Columbus, F. K. Bozal, A. Skepner, D. Raymond, S. LaRussa, D. C. McKinney, Y. Freyzon, Y. Baidi, D. Porter, A. J. Aguirre, A. Ianari, B. McMillan, W. R. Sellers, Molecular basis for substrate recruitment to the PRMT5 methylosome. *Mol. Cell*. **81**, 3481–3495.e7 (2021).

48. Z. L. Watson, F. R. Ward, R. Méheust, O. Ad, A. Schepartz, J. F. Banfield, J. H. Cate, Structure of the bacterial ribosome at 2 Å resolution. *Elife*. **9** (2020), doi:10.7554/eLife.60482.

49. J. M. Małecki, M.-F. Odonohue, Y. Kim, M. E. Jakobsson, L. Gessa, R. Pinto, J. Wu, E. Davydova, A. Moen, J. V. Olsen, B. Thiede, P.-E. Gleizes, S. A. Leidel, P. Ø. Falnes, Human METTL18 is a histidine-specific methyltransferase that targets RPL3 and affects ribosome biogenesis and function. *Nucleic Acids Res.* **49**, 3185–3203 (2021).

50. F. Dragon, J. E. G. Gallagher, P. A. Compagnone-Post, B. M. Mitchell, K. A. Porwancher, K. A. Wehner, S. Wormsley, R. E. Settlage, J. Shabanowitz, Y. Osheim, A. L. Beyer, D. F. Hunt, S. J. Baserga, A large nucleolar U3 ribonucleoprotein required for 18S ribosomal RNA biogenesis. *Nature*. **417**, 967–970 (2002).

51. L. R. Kenner, A. A. Anand, H. C. Nguyen, A. G. Myasnikov, C. J. Klose, L. A. McGeever, J. C. Tsai, L. E. Miller-Vedam, P. Walter, A. Frost, eIF2B-catalyzed nucleotide exchange and phosphoregulation by the integrated stress response. *Science*. **364**, 491–495 (2019).

52. S. Jentsch, I. Psakhye, Control of nuclear activities by substrate-selective and protein-group SUMOylation. *Annu. Rev. Genet.* **47**, 167–186 (2013).

53. I. Psakhye, S. Jentsch, Protein group modification and synergy in the SUMO pathway as exemplified in DNA repair. *Cell*. **151**, 807–820 (2012).

54. D. Menolfi, A. Delamarre, A. Lengronne, P. Pasero, D. Branzei, Essential Roles of the Smc5/6 Complex in Replication through Natural Pausing Sites and Endogenous DNA Damage Tolerance. *Mol. Cell*. **60**, 835–846 (2015).

55. S. Agashe, C. R. Joseph, T. A. C. Reyes, D. Menolfi, M. Giannattasio, A. Waizenegger, B. Szakal, D. Branzei, Smc5/6 functions with Sgs1-Top3-Rmi1 to complete chromosome replication at natural pause sites. *Nat. Commun.* **12**, 2111 (2021).

56. G. De Piccoli, J. Torres-Rosell, L. Aragón, The unnamed complex: what do we know about Smc5-Smc6? *Chromosome Res.* **17**, 251–263 (2009).

57. I. Psakhye, F. Castellucci, D. Branzei, SUMO-Chain-Regulated Proteasomal Degradation Timing Exemplified in DNA Replication Initiation. *Mol. Cell*. **76**, 632–645.e6 (2019).

58. A. Waizenegger, M. Urulangodi, C. P. Lehmann, T. A. C. Reyes, I. Saugar, J. A. Tercero, B. Szakal, D. Branzei, Mus81-Mms4 endonuclease is an Esc2-STUbL-Cullin8 mitotic substrate impacting on genome integrity. *Nat. Commun.* **11**, 5746 (2020).

59. I. Psakhye, D. Branzei, SMC complexes are guarded by the SUMO protease Ulp2 against SUMO-chain-mediated turnover. *Cell Rep.* **36**, 109485 (2021).

60. J. J. L. Miranda, P. De Wulf, P. K. Sorger, S. C. Harrison, The yeast DASH complex forms closed rings on microtubules. *Nat. Struct. Mol. Biol.* **12**, 138–143 (2005).

61. S. Westermann, A. Avila-Sakar, H.-W. Wang, H. Niederstrasser, J. Wong, D. G. Drubin, E. Nogales, G. Barnes, Formation of a dynamic kinetochore- microtubule interface through assembly of the Dam1 ring complex. *Mol. Cell*. **17**, 277–290 (2005).

62. C. L. Asbury, D. R. Gestaut, A. F. Powers, A. D. Franck, T. N. Davis, The Dam1 kinetochore complex

harnesses microtubule dynamics to produce force and movement. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 9873–9878 (2006).

63. V. H. Ramey, A. Wong, J. Fang, S. Howes, G. Barnes, E. Nogales, Subunit organization in the Dam1 kinetochore complex and its ring around microtubules. *Mol. Biol. Cell*. **22**, 4335–4342 (2011).

64. J. O. Kim, A. Zelter, N. T. Umbreit, A. Bollozos, M. Riffle, R. Johnson, M. J. MacCoss, C. L. Asbury, T. N. Davis, The Ndc80 complex bridges two Dam1 complex rings. *Elife*. **6** (2017), doi:10.7554/eLife.21069.

65. C. T. Ng, L. Deng, C. Chen, H. H. Lim, J. Shi, U. Surana, L. Gan, Electron cryotomography analysis of Dam1C/DASH at the kinetochore-spindle interface in situ. *J. Cell Biol.* **218**, 455–473 (2019).

66. S. Jenni, S. C. Harrison, Structure of the DASH/Dam1 complex shows its role at the yeast kinetochore-microtubule interface. *Science*. **360**, 552–558 (2018).

67. F. Wendler, A. K. Gillingham, R. Sinka, C. Rosa-Ferreira, D. E. Gordon, X. Franch-Marro, A. A. Peden, J.-P. Vincent, S. Munro, A genome-wide RNA interference screen identifies two novel components of the metazoan secretory pathway. *EMBO J.* **29**, 304–314 (2010).

68. M. Heidtman, C. Z. Chen, R. N. Collins, C. Barlowe, Yos1p is a novel subunit of the Yip1p-Yif1p complex and is required for transport between the endoplasmic reticulum and the Golgi complex. *Mol. Biol. Cell*. **16**, 1673–1683 (2005).

69. Y. Desfougères, R. U. Gerasimaitė, H. J. Jessen, A. Mayer, Vtc5, a Novel Subunit of the Vacuolar Transporter Chaperone Complex, Regulates Polyphosphate Synthesis and Phosphate Homeostasis in Yeast. *J. Biol. Chem.* **291**, 22262–22275 (2016).

70. M. Hothorn, H. Neumann, E. D. Lenherr, M. Wehner, V. Rybin, P. O. Hassa, A. Uttenweiler, M. Reinhardt, A. Schmidt, J. Seiler, A. G. Ladurner, C. Herrmann, K. Scheffzek, A. Mayer, Catalytic core of a membrane-associated eukaryotic polyphosphate polymerase. *Science*. **324**, 513–516 (2009).

71. M. Vietri, M. Radulovic, H. Stenmark, The many functions of ESCRTs. *Nat. Rev. Mol. Cell Biol.* **21**, 25–42 (2020).

72. J. H. Hurley, ESCRTs are everywhere. *EMBO J.* **34**, 2398–2407 (2015).

73. S. Tang, W. M. Henne, P. P. Borbat, N. J. Buchkovich, J. H. Freed, Y. Mao, J. C. Fromme, S. D. Emr, Structural basis for activation, assembly and membrane binding of ESCRT-III Snf7 filaments. *Elife*. **4** (2015), doi:10.7554/eLife.12548.

74. C. Schluter, K. K. Y. Lam, J. Brumm, B. W. Wu, M. Saunders, T. H. Stevens, J. Bryan, E. Conibear, Global analysis of yeast endosomal transport identifies the vps55/68 sorting complex. *Mol. Biol. Cell*. **19**, 1282–1294 (2008).

75. S. Siniossoglou, H. R. Pelham, An effector of Ypt6p binds the SNARE Tlg1p and mediates selective fusion of vesicles with late Golgi membranes. *EMBO J.* **20**, 5991–5998 (2001).

76. K. Mei, Y. Li, S. Wang, G. Shao, J. Wang, Y. Ding, G. Luo, P. Yue, J.-J. Liu, X. Wang, M.-Q. Dong, H.-W. Wang, W. Guo, Cryo-EM structure of the exocyst complex. *Nat. Struct. Mol. Biol.* **25**, 139–146 (2018).

77. H.-T. Chou, D. Dukovski, M. G. Chambers, K. M. Reinisch, T. Walz, CATCHR, HOPS and CORVET tethering complexes share a similar architecture. *Nat. Struct. Mol. Biol.* **23**, 761–763 (2016).

78. R. Behnia, F. A. Barr, J. J. Flanagan, C. Barlowe, S. Munro, The yeast orthologue of GRASP65 forms a complex with a coiled-coil protein that contributes to ER to Golgi traffic. *J. Cell Biol.* **176**, 255–261 (2007).

79. M. Schuldiner, S. R. Collins, N. J. Thompson, V. Denic, A. Bhamidipati, T. Punna, J. Ihmels, B. Andrews, C. Boone, J. F. Greenblatt, J. S. Weissman, N. J. Krogan, Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*. **123**, 507–519 (2005).

80. W. Ma, J. Goldberg, TANGO1/cTAGE5 receptor as a polyvalent template for assembly of large COPII coats. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 10061–10066 (2016).

81. V. G. Stancheva, X.-H. Li, J. Hutchings, N. Gomez-Navarro, B. Santhanam, M. M. Babu, G. Zanetti, E. A. Miller, Combinatorial multivalent interactions drive cooperative assembly of the COPII coat. *J. Cell Biol.* **219** (2020), doi:10.1083/jcb.202007135.

82. T. C. Südhof, J. E. Rothman, Membrane fusion: grappling with SNARE and SM proteins. *Science*. **323**, 474–477 (2009).

83. S. Conchon, X. Cao, C. Barlowe, H. R. Pelham, Got1p and Sft2p: membrane proteins involved in traffic to the Golgi complex. *EMBO J.* **18**, 3934–3946 (1999).

84. R. B. Sutton, D. Fasshauer, R. Jahn, A. T. Brunger, Crystal structure of a SNARE complex involved in synaptic exocytosis at 2.4 A resolution. *Nature*. **395**, 347–353 (1998).

85. R. Jahn, R. H. Scheller, SNAREs--engines for membrane fusion. *Nat. Rev. Mol. Cell Biol.* **7**, 631–643 (2006).

86. J. Rizo, Mechanism of neurotransmitter release coming into focus. *Protein Sci.* **27**, 1364–1391 (2018).

87. L. Burri, O. Varlamov, C. A. Doege, K. Hofmann, T. Beilharz, J. E. Rothman, T. H. Söllner, T. Lithgow, A SNARE required for retrograde transport to the endoplasmic reticulum. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 9873–9877 (2003).

88. Y. Hong, K. Ohishi, J. Y. Kang, S. Tanaka, N. Inoue, J.-I. Nishimura, Y. Maeda, T. Kinoshita, Human PIG-U and yeast Cdc91p are the fifth subunit of GPI transamidase that attaches GPI-anchors to proteins. *Mol. Biol. Cell*. **14**, 1780–1789 (2003).

89. D. G. Gamage, T. L. Hendrickson, GPI transamidase and GPI anchored proteins: oncogenes and biomarkers for cancer. *Crit. Rev. Biochem. Mol. Biol.* **48**, 446–464 (2013).

90. L. Yi, G. Bozkurt, Q. Li, S. Lo, A. K. Menon, H. Wu, Disulfide Bond Formation and N-Glycosylation Modulate Protein-Protein Interactions in GPI-Transamidase (GPIT). *Sci. Rep.* **8**, 45912 (2017).

91. P. Moran, I. W. Caras, A nonfunctional sequence converted to a signal for glycophosphatidylinositol membrane anchor attachment. *J. Cell Biol.* **115**, 329–336 (1991).

92. P. Fraering, I. Imhof, U. Meyer, J. M. Strub, A. van Dorsselaer, C. Vionnet, A. Conzelmann, The GPI transamidase complex of Saccharomyces cerevisiae contains Gaa1p, Gpi8p, and Gpi16p. *Mol. Biol. Cell*. **12**, 3295–3306 (2001).

93. K. Ohishi, N. Inoue, T. Kinoshita, PIG-S and PIG-T, essential for GPI anchor attachment to proteins, form a complex with GAA1 and GPI8. *EMBO J.* **20**, 4088–4098 (2001).

94. K. Ohishi, K. Nagamune, Y. Maeda, T. Kinoshita, Two subunits of glycosylphosphatidylinositol transamidase, GPI8 and PIG-T, form a functionally important intermolecular disulfide bridge. *J. Biol. Chem.* **278**, 13959–13967 (2003).

95. S. Vainauskas, A. K. Menon, A conserved proline in the last transmembrane segment of Gaa1 is required for glycosylphosphatidylinositol (GPI) recognition by GPI transamidase. *J. Biol. Chem.* **279**, 6540–6545 (2004).

96. U. Meyer, M. Benghezal, I. Imhof, A. Conzelmann, Active site determination of Gpi8p, a caspase-related enzyme required for glycosylphosphatidylinositol anchor addition to proteins. *Biochemistry*. **39**, 3461–3471 (2000).

97. D. G. Gamage, Y. Varma, J. L. Meitzler, R. Morissette, T. J. Ness, T. L. Hendrickson, The soluble domains of Gpi8 and Gaa1, two subunits of glycosylphosphatidylinositol transamidase (GPI-T), assemble into a complex. *Arch. Biochem. Biophys.* **633**, 58–67 (2017).

98. J. L. Meitzler, J. J. Gray, T. L. Hendrickson, Truncation of the caspase-related subunit (Gpi8p) of Saccharomyces cerevisiae GPI transamidase: dimerization revealed. *Arch. Biochem. Biophys.* **462**, 83–93 (2007).

99. T. T. M. Nguyen, Y. Murakami, S. Mobilio, M. Niceta, G. Zampino, C. Philippe, S. Moutton, M. S. Zaki, K. N. James, D. Musaev, W. Mu, K. Baranano, J. R. Nance, J. A. Rosenfeld, N. Braverman, A. Ciolfi, F. Millan, R. E. Person, A.-L. Bruel, C. Thauvin-Robinet, A. Ververi, C. DeVile, A. Male, S. Efthymiou, R. Maroofian, H. Houlden, S. Maqbool, F. Rahman, N. V. Baratang, J. Rousseau, A. St-Denis, M. J. Elrick, I. Anselm, L. H. Rodan, M. Tartaglia, J. Gleeson, T. Kinoshita, P. M. Campeau, Bi-allelic Variants in the GPI Transamidase Subunit PIGK Cause a Neurodevelopmental Syndrome with Hypotonia, Cerebellar Atrophy, and Epilepsy. *Am. J. Hum. Genet.* **106**, 484–495 (2020).

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

101. M. Baek, L. Heo, R. Ndem, neilfleckSCRI, *RosettaCommons/RoseTTAFold: RoseTTAFold update: Including the simpler version for PPI screening* (2021; https://zenodo.org/record/5639837).

102. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

103. S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez, R. D. Finn, HMMER web server: 2018 update. *Nucleic Acids Res.* **46**, W200–W204 (2018).

104. F. Gabler, S. Nam, S. Till, M. Mirdita, M. Steinegger, J. Söding, A. N. Lupas, V. Alva, Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Current Protocols in Bioinformatics*. **72** (2020), , doi:10.1002/cpbi.108.

105. F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, M. Weigt, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U. S. A.* **108**, E1293–301 (2011).

106. D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, C. von Mering, The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2020).

107. L. Käll, A. Krogh, E. L. L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338**, 1027–1036 (2004).

108. C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535–9 (2006).

109. E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin, UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

110. K. Katoh, J. Rozewicki, K. D. Yamada, MAFFT online service: multiple sequence alignment,

interactive sequence choice and visualization. *Brief. Bioinform.* **20**, 1160–1166 (2019).

111. X. Robert, P. Gouet, Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Research*. **42** (2014), pp. W320–W324.

112. S. Banjade, S. Tang, S. D. Emr, Genetic and Biochemical Analyses of Yeast ESCRT. *Methods Mol. Biol.* **1998**, 105–116 (2019).

113. A. Y. Madrona, D. K. Wilson, Structure of Ski8p, a WD repeat protein involved in mRNA degradation and meiotic recombination (2004), , doi:10.2210/pdb1sq9/pdb.

114. M. D. Nichols, K. A. DeAngelis, J. L. Keck, J. M. Berger, Structure of the DNA topoisomerase vi a subunit (1999), , doi:10.2210/pdb1d3y/pdb.

115. K. D. Corbett, P. Benedetti, J. M. Berger, Crystal structure of the topoisomerase VI holoenzyme from Methanosarcina mazei (2007), , doi:10.2210/pdb2q2e/pdb.

116. F. Halbach, P. Reichelt, M. Rode, E. Conti, Crystal structure of the S. cerevisiae Ski2-3-8 complex (2013), , doi:10.2210/pdb4buj/pdb.

Supplemental Figures

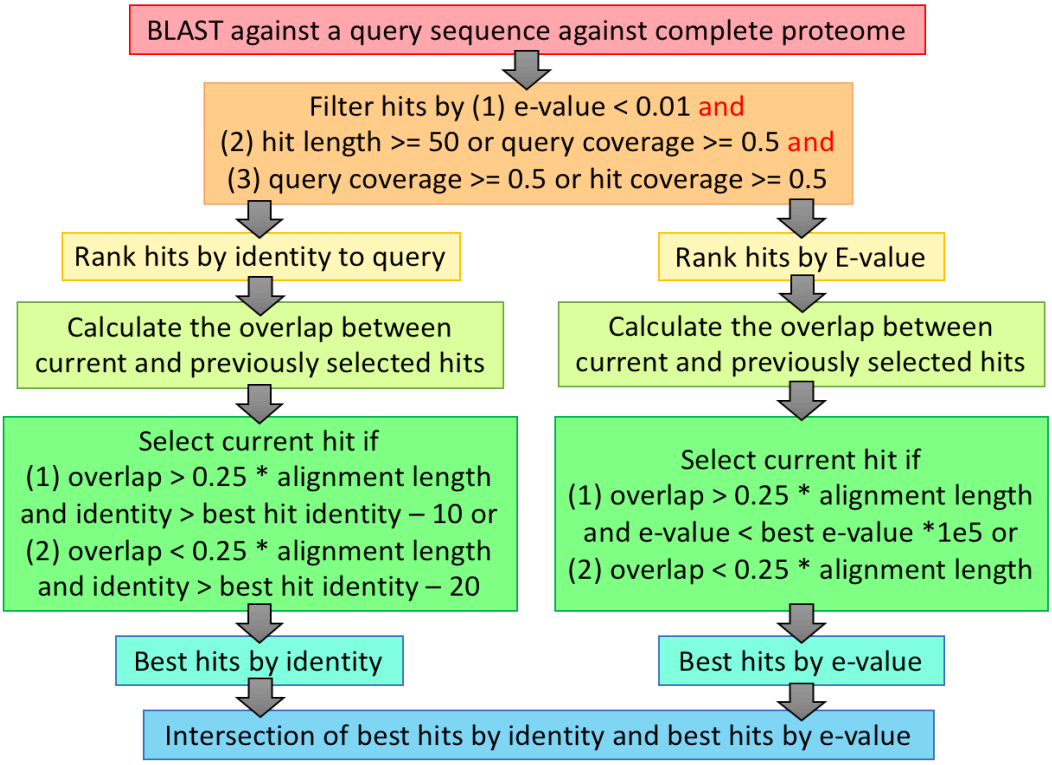


Figure S1. Procedure for identifying the best hit to a query protein in the complete proteome of a different species

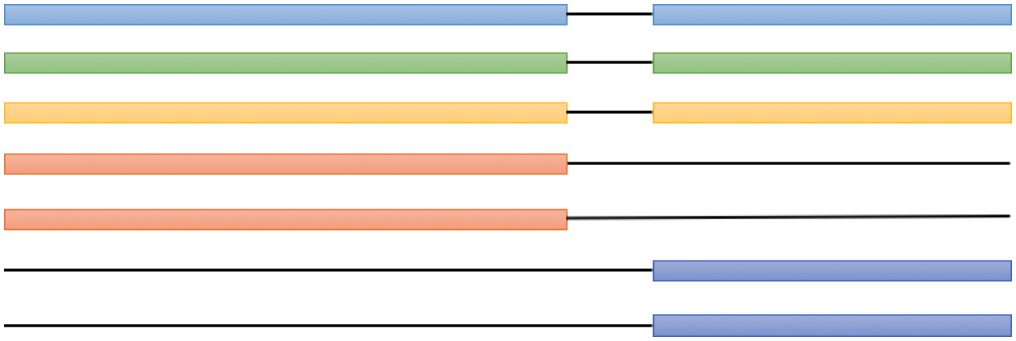


Figure S2. Diagram for paired multiple sequence alignments

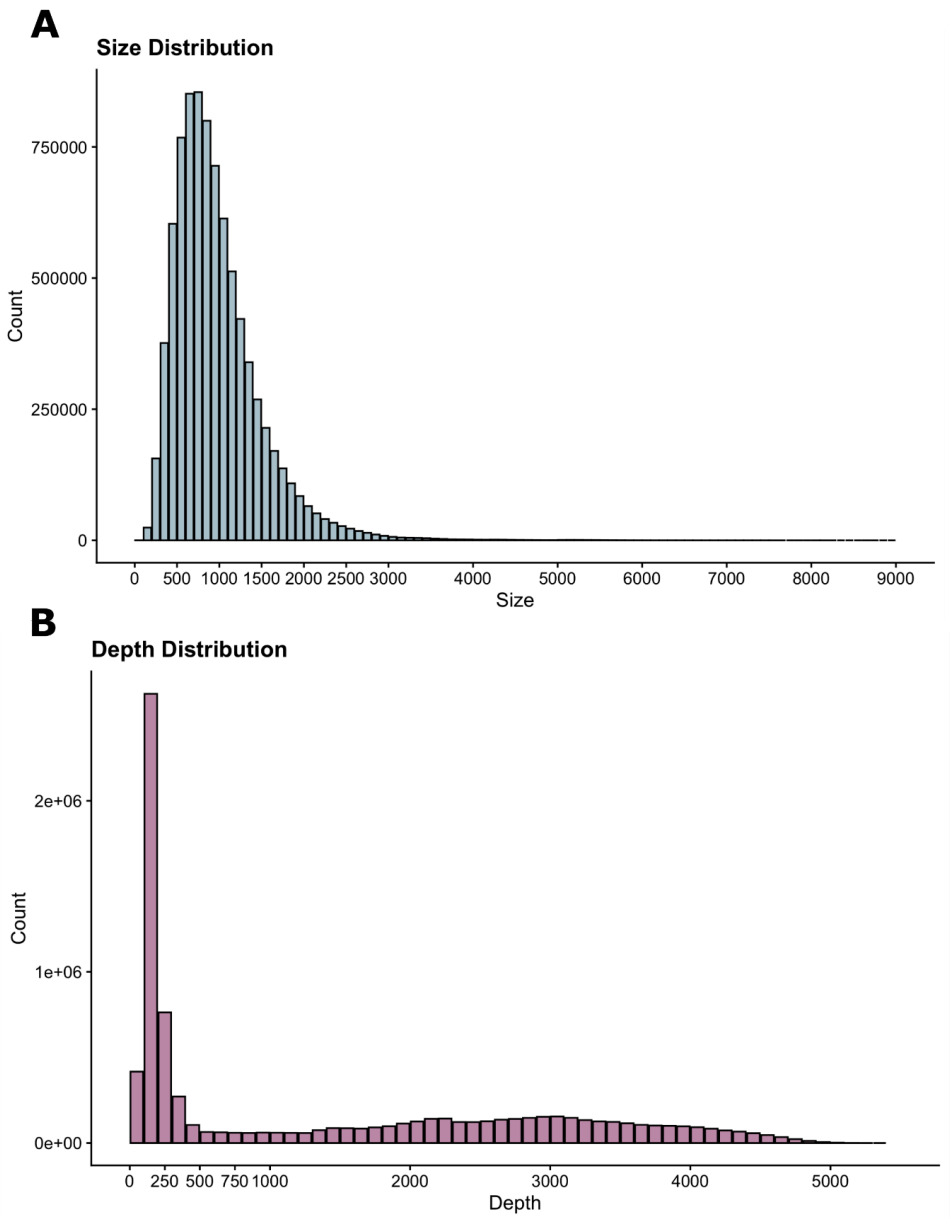


Figure S3. Length (A) and depth (B) distributions of pMSA

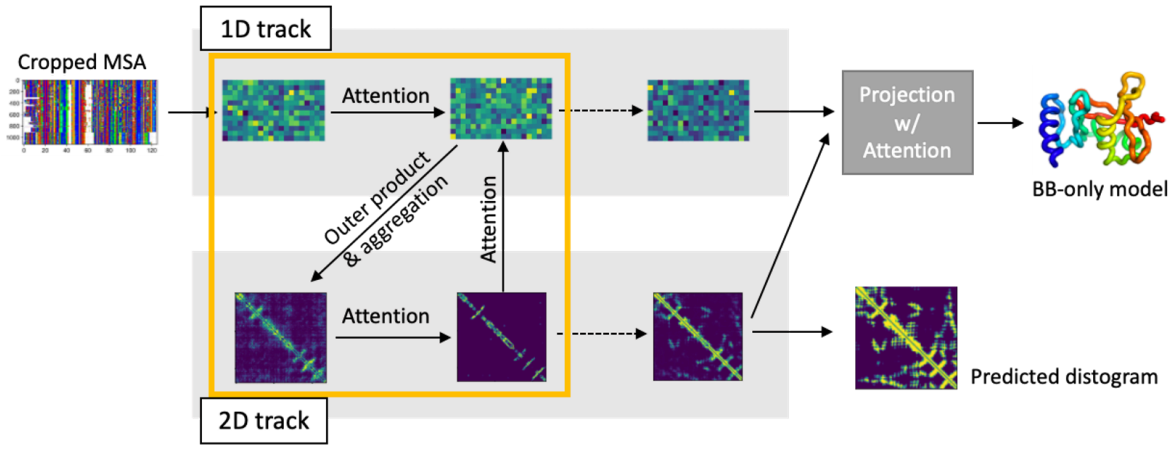


Figure S4. A lighter-weight RoseTTAFold two-track (RF2t) model

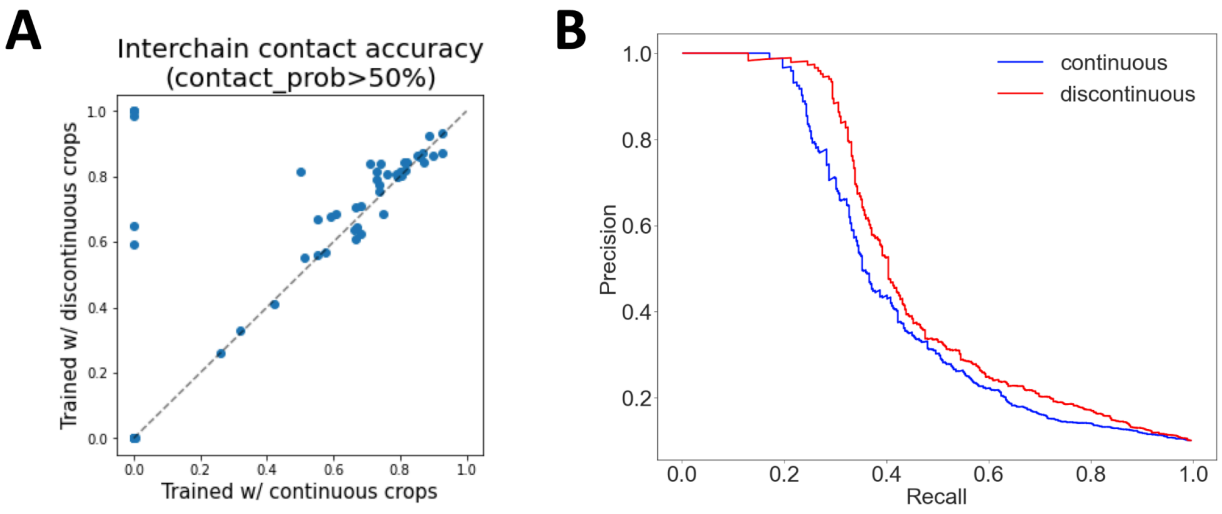


Figure S5. Performance comparison between the two-track RoseTTAFold model trained with discontinuous crops and continuous crops

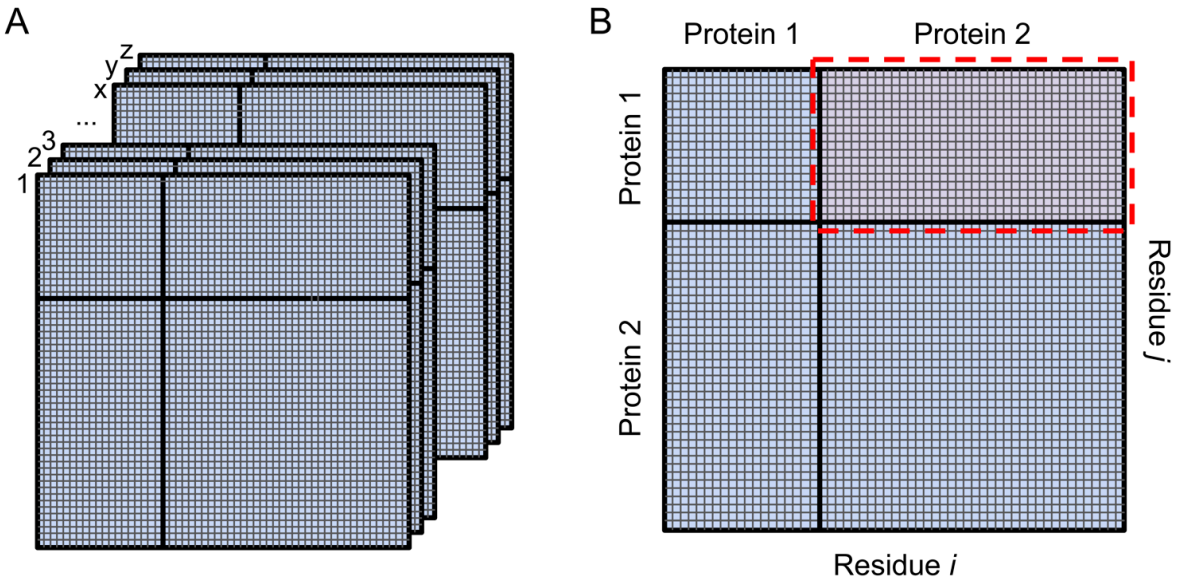


Figure S6. Distance matrix diagrams

(**A**) The 3D matrix of dimension len(prot1) + len(prot2) by len(prot1) + len(prot2) by number of distance bins (37 for RoseTTAFold binned 2-20Å / 0.5Å). (**B**) The summed 2D matrix (d < 12Å) from panel A. The submatrix inside the red dashed line contains the contact probability for residues between two proteins.

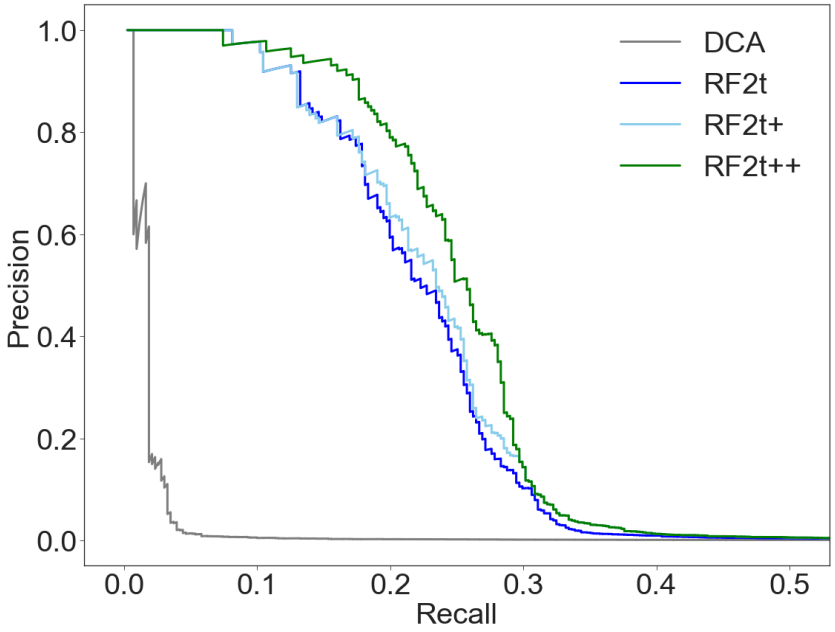


Figure S7. Precision-recall curve for different modifications of RoseTTAFold scores evaluated on the gold-standard set (table S1)

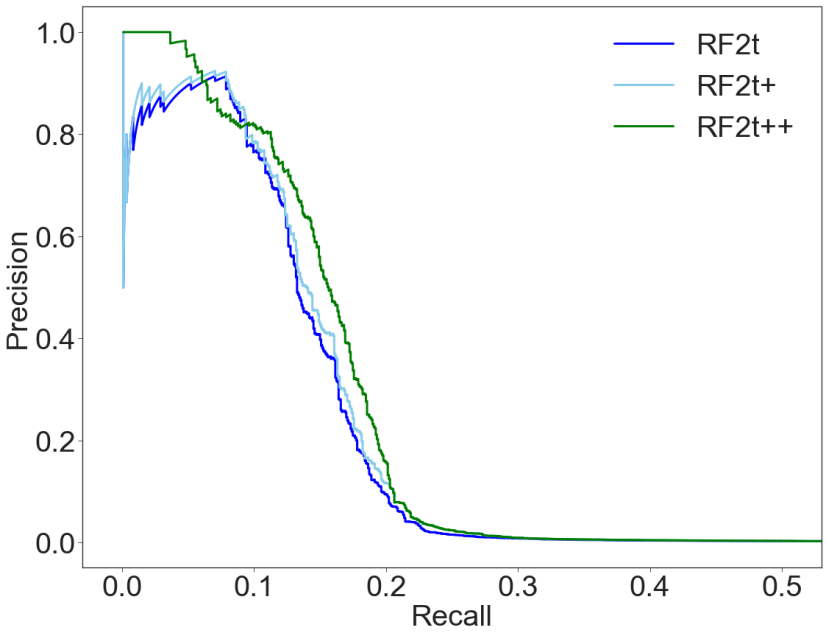


Figure S8. Precision-recall curve for different modifications of RoseTTAFold scores evaluated on the literature-curated set (table S1)

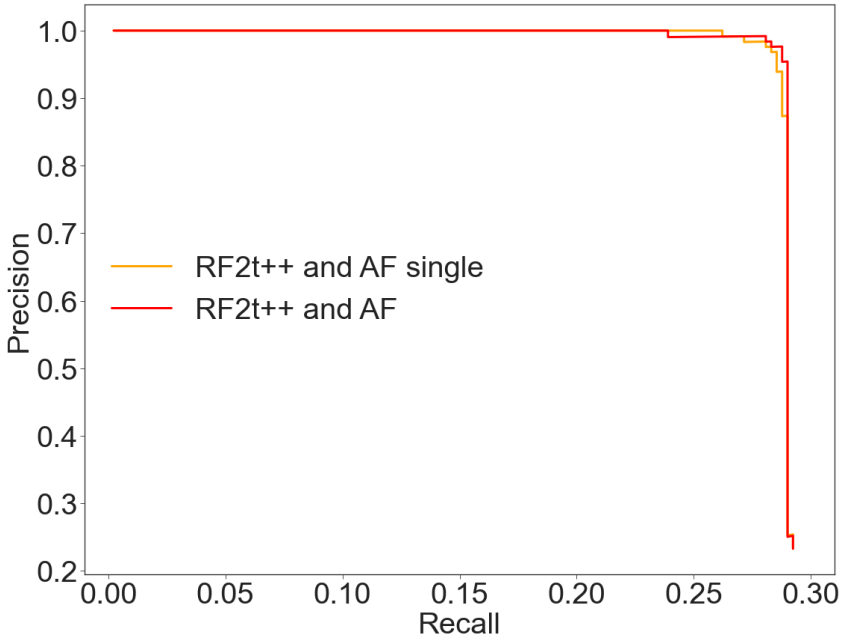


Figure S9. Precision-recall curve for different modifications of AlphaFold scores evaluated on the gold-standard set (table S1).

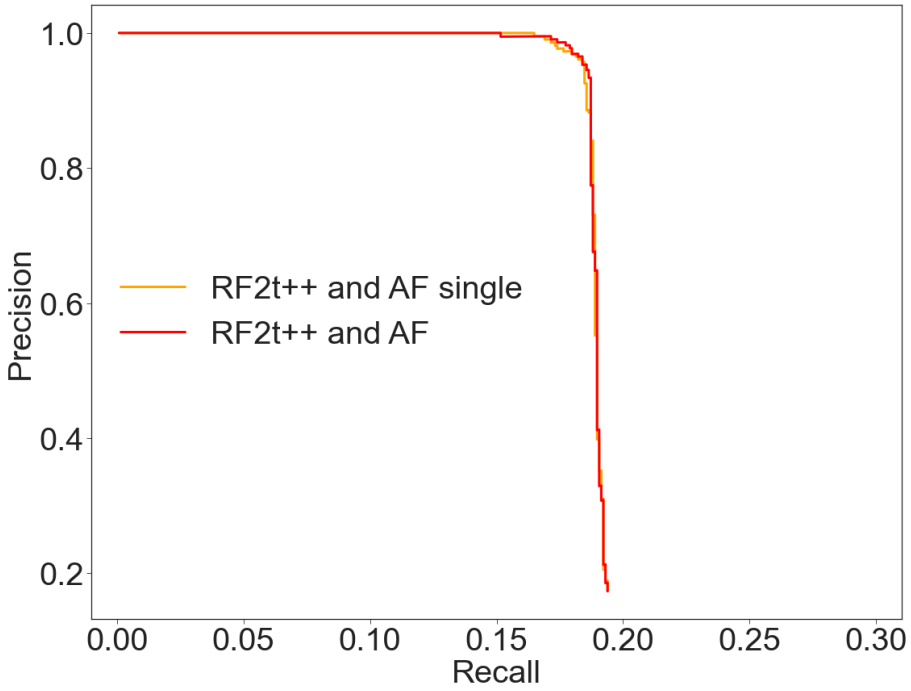


Figure S10. Precision-recall curve for different modifications of AlphaFold scores evaluated on the literature-curated set (table S1)

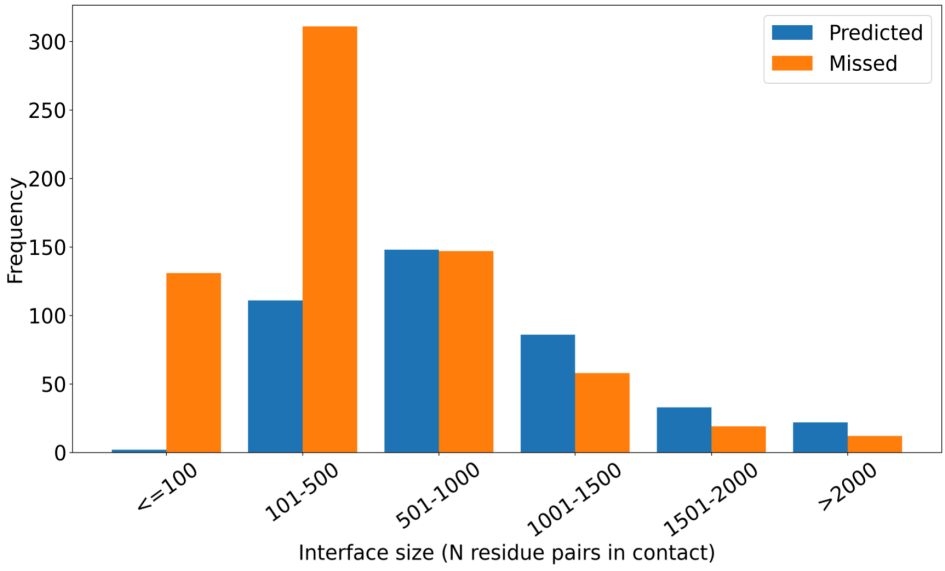


Figure S11. Interface size (in number of amino acids) of experimentally determined complexes predicted or missed in our screen

We mapped predicted PPIs onto known complexes deposited in the PDB. Blue bars indicate the number of complexes correctly predicted by our screen to interact while orange bars indicate complexes that were missed by our screen.

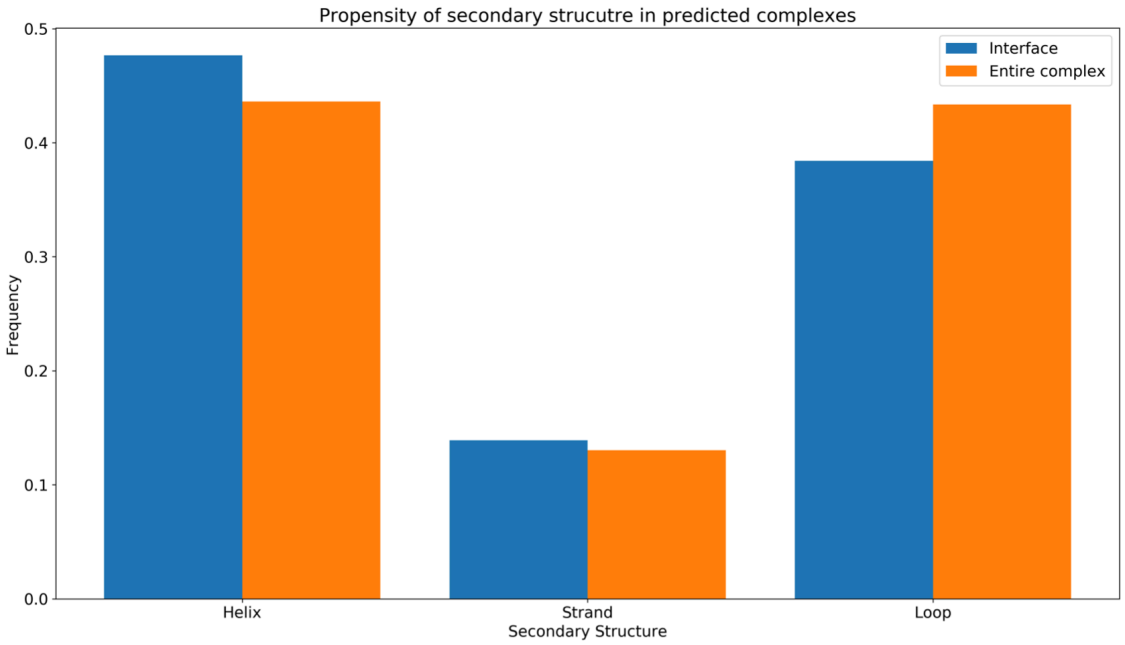


Figure S12. Secondary structure of our models of protein complexes

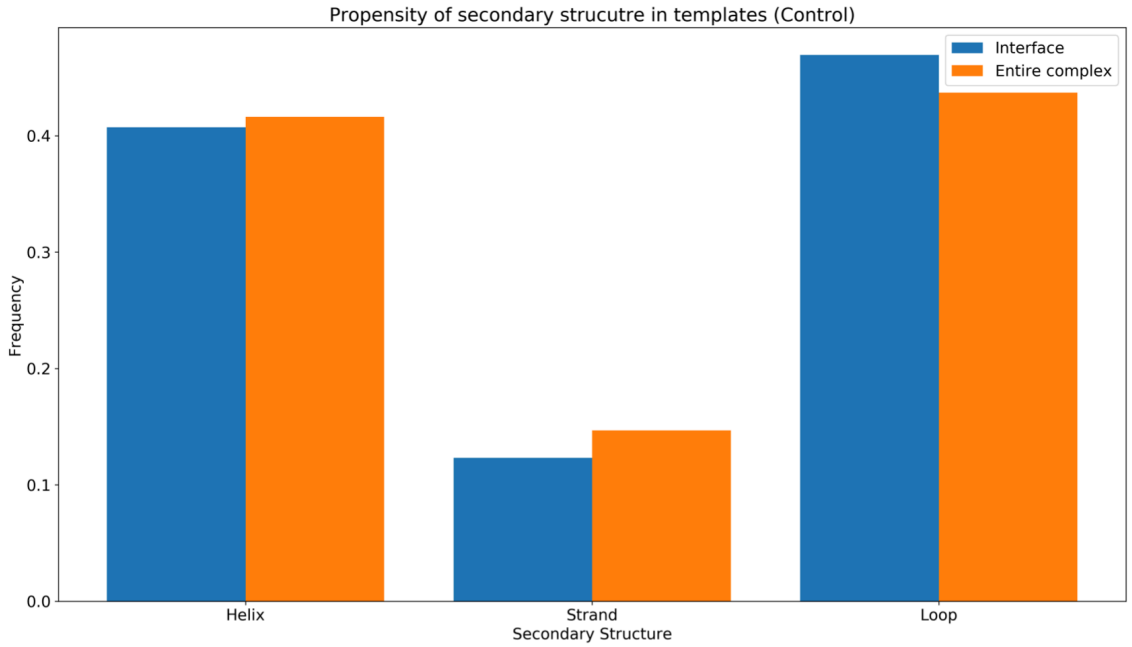


Figure S13. Secondary structure of experimentally determined yeast protein complexes in the PDB

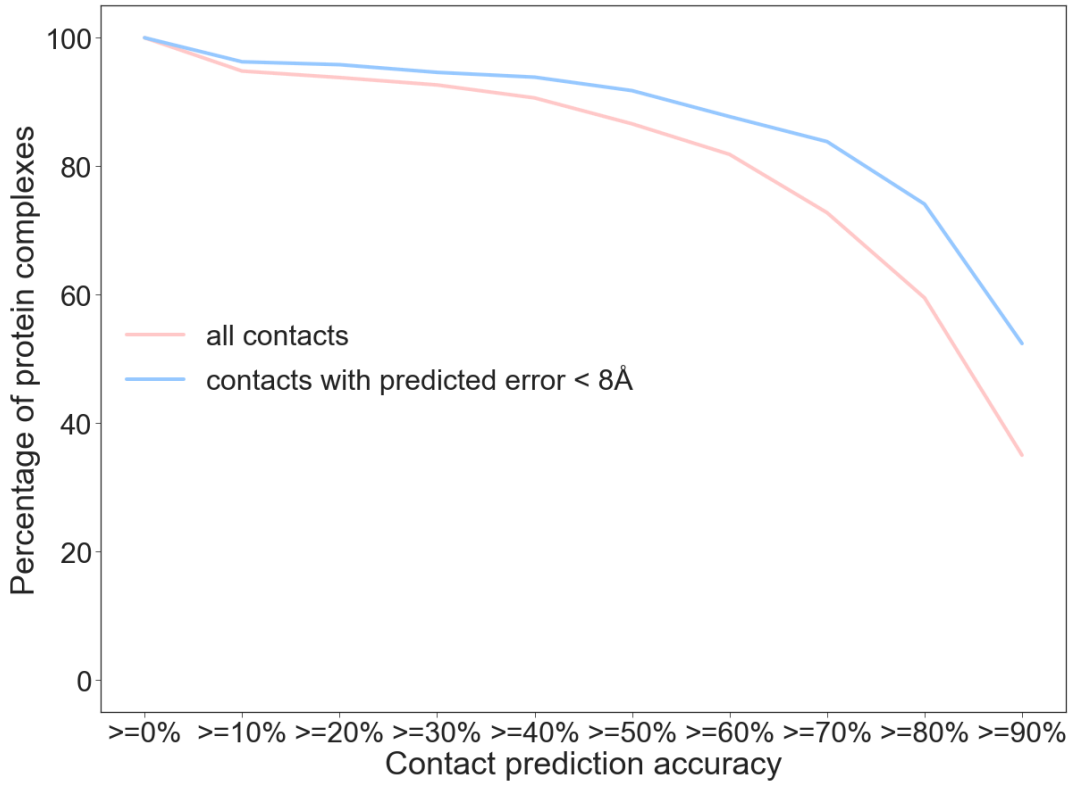


Figure S14. The percentage of protein complexes that are above different contact prediction accuracy cutoffs

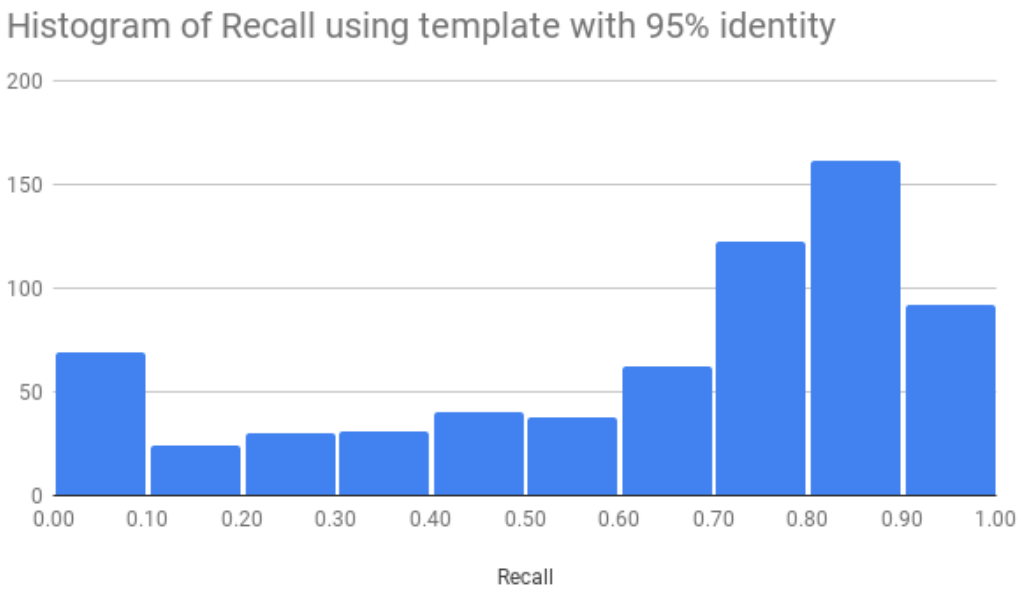


Figure S15. Distribution of percentage of contacts in experimental structures that are predicted correctly in our models

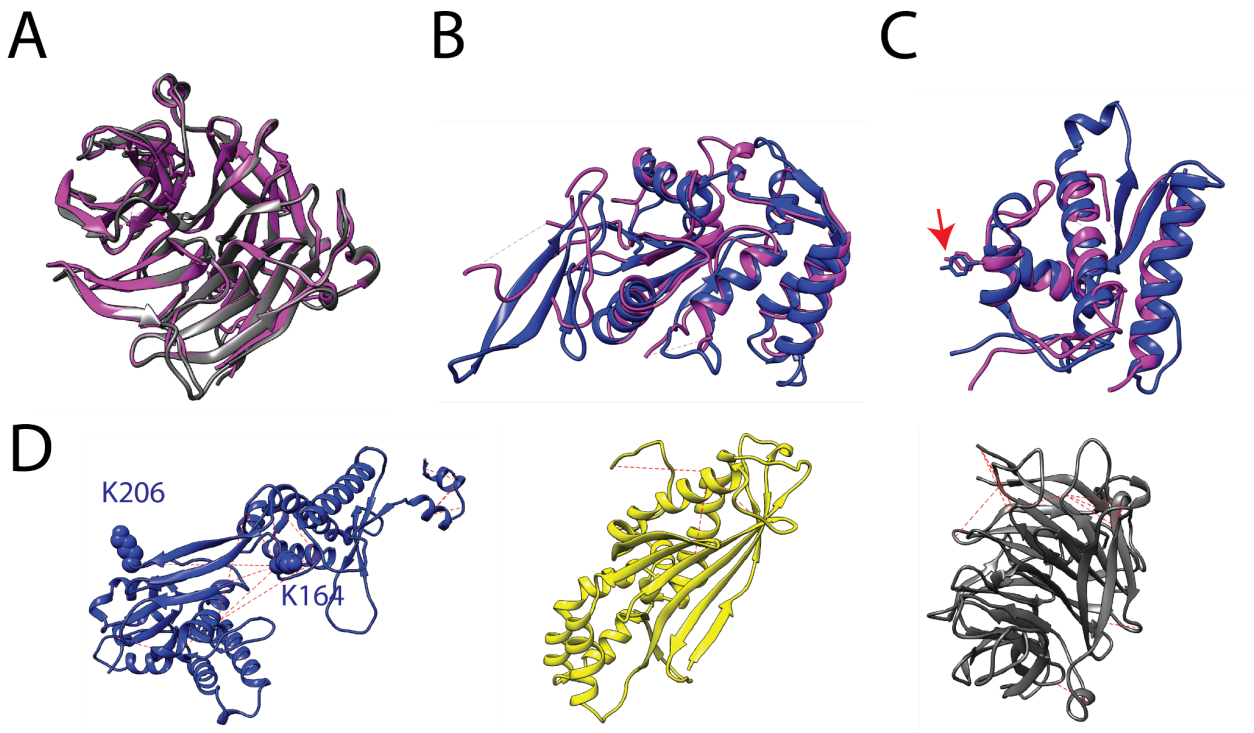


Figure S16. Ski8 complex monomer models

(**A**) Comparison of the AF model for Ski8 (gray) and a crystal structure (magenta, pdb: 1sq9 (*113*)). The root mean square deviation (RMSD) between the aligned structures is 1.1 Å. (**B**) Comparison of the predicted Spo11 Toprim domain structure (blue; residues 173 to 398) with the Toprim domain from *M. mazei* Top6A (magenta; residues 147 to 367 from pdb 2q2e (*115*)). The RMSD is 1.1 Å. (**C**) Comparison of the predicted WHD for Spo11 (blue; residues 24 to 172) with the WHD for *M. mazei* Top6A (magenta; residues 15 to 146). The RMSD is 1.2 Å. The catalytic tyrosines are positioned nearly identically for the two proteins (arrow). (**D**) Comparison of models to intramolecular crosslinking data for Spo11 (blue), Rec102 (yellow), and Ski8 (gray). Red dashed lines connect the α-carbons of lysine pairs that were observed to be crosslinked in a recent study of the Spo11-Ski8-Rec102-Rec104 complex (*27*). In the Spo11 model, 14 out of 15 cross-linked lysine pairs with high mass spectrometry counts (≥10) are within the crosslinker range limit of 27.4 Å. For the only exception (K164-K206, predicted distance of 30.7 Å), both lysines are in predicted loop regions which may be conformationally flexible. Similarly, all of the high-frequency crosslinked lysine pairs are within the crosslinking distance limit in the AF models for Rec102 (3 lysine pairs) and Ski8 (14 pairs).

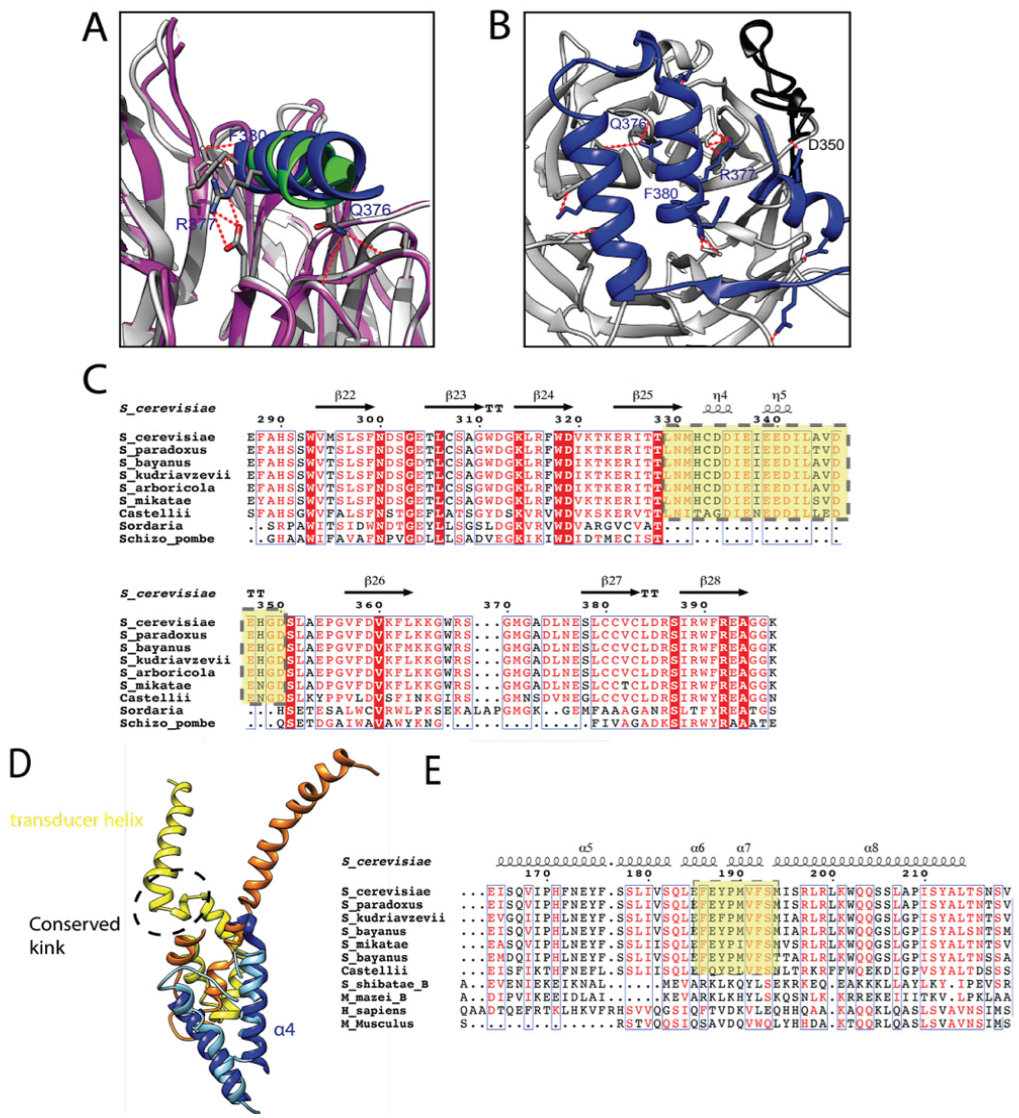


Figure S17. Interfaces of Spo11 with Ski8 and Rec102

(**A**) The AF model for Ski8 (gray) interaction with an alpha helix containing the conserved Spo11 QREIF$_{380}$ motif (blue) is compared with a model (magenta and green) previously generated based on a complex of Ski8 with a similar peptide sequence in Ski3 (*27*)(*116*)). Red dashed lines indicate predicted hydrogen bonds in the RF model. (**B**) Spo11 residues 323-381 (blue) form extensive contacts with Ski8. The black segment is a part of Ski8 that appears to extend the interaction surface. Within this segment, Ski8 D350 is predicted to form a hydrogen bond with Spo11 Q323 and a salt bridge with Spo11 R320. Predicted hydrogen bonds are indicated by red dashed lines. (**C**) Sequence alignment of fungal Ski8 orthologs. The black region in panel B corresponds to an insertion (residues 329-350, highlighted in yellow) that is found in *Saccharomyces* species but not in *S. macrospora* or *S. pombe*, in which Ski8 homologs also interact with Spo11. (**D**) Comparison of the Spo11–Rec102 and Top6A–Top6B interfaces. In the RF model, Spo11 residues 42 to 121 are shown in blue (a β-sheet in this region omitted for clarity) and Rec102 residues 164 to 229 are shown in yellow. From the crystal structure of the *M. mazei* Top6A–Top6B complex, Top6A residues 15 to 68 are shown in cyan and Top6B residues 440 to 508 are shown in orange. The helix labeled “α4” is the first helix in the WHD of Spo11. (**E**) Multiple sequence alignment of transducer helices from yeast Rec102, mammalian Top6BL, and archaeal Top6B proteins. The black box highlights the sequence corresponding to the predicted kink in the transducer helix of the yeast proteins, not apparent in Top6B or Top6BL.

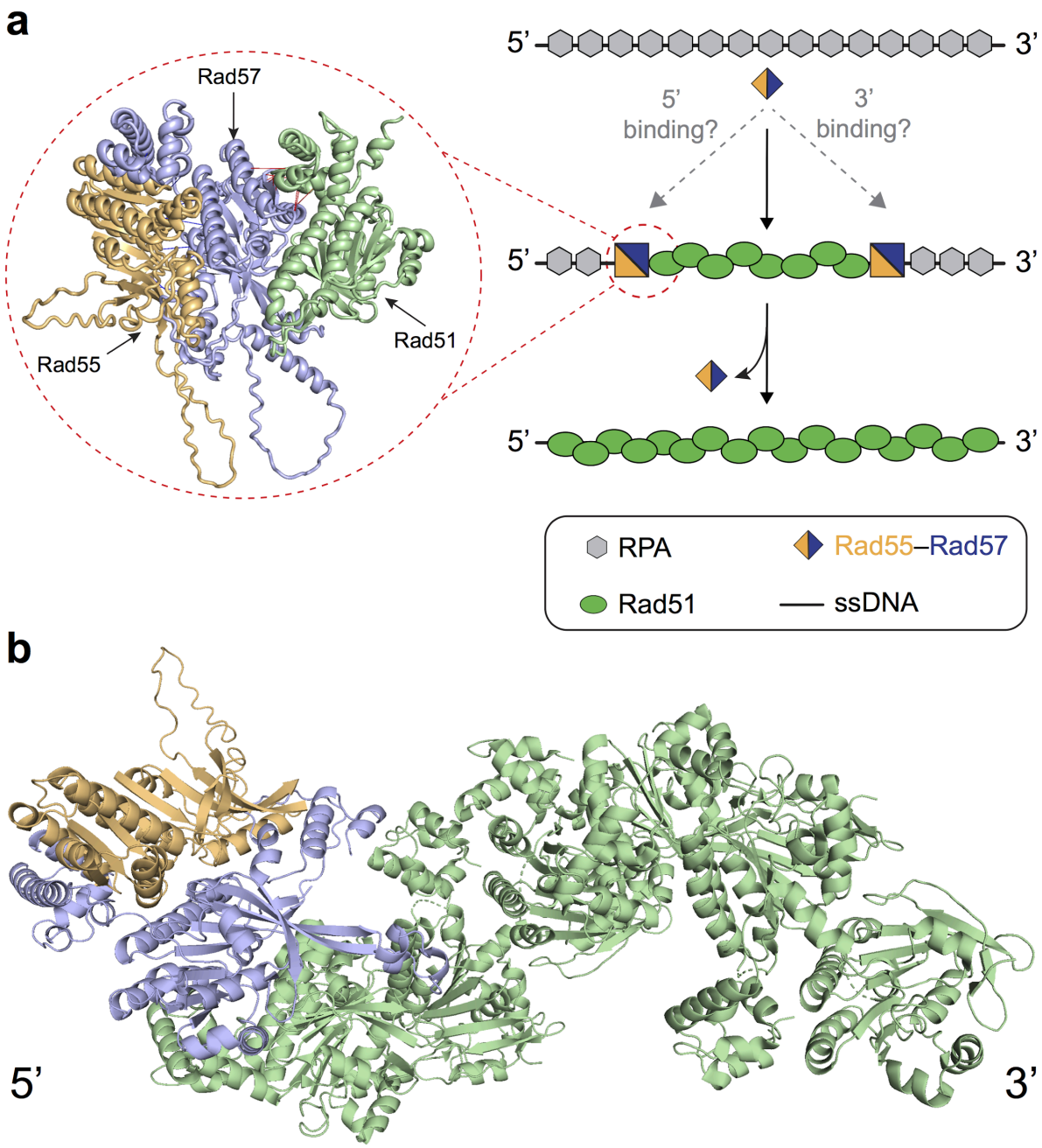


Figure S18. Model for Rad55–Rad57 binding to the Rad51–ssDNA filament

(**A**) Schematic for Rad55–Rad57 mediated Rad51 filament assembly during homologous recombination. Rad51 recombinase must displace replication protein A (RPA) bound to single–stranded DNA (ssDNA) to form Rad51–ssDNA filaments that carry out DNA recombination. Rad55–Rad57 acts as a chaperone, facilitating a faster and more extensive displacement of RPA by Rad51. The specific polarity of interaction between Rad55–Rad57 and the Rad51–ssDNA filament is unknown, but is expected to influence whether filament growth is stimulated in the 5'→3' or 3'→5' direction. Our model suggests that Rad55–Rad57 may be transiently binding at the 5' end of a Rad51–ssDNA filament through an interaction between Rad57 and Rad51 (inset). (**B**) Putative model for Rad55–Rad57 binding at the 5' end of a Rad51 filament.

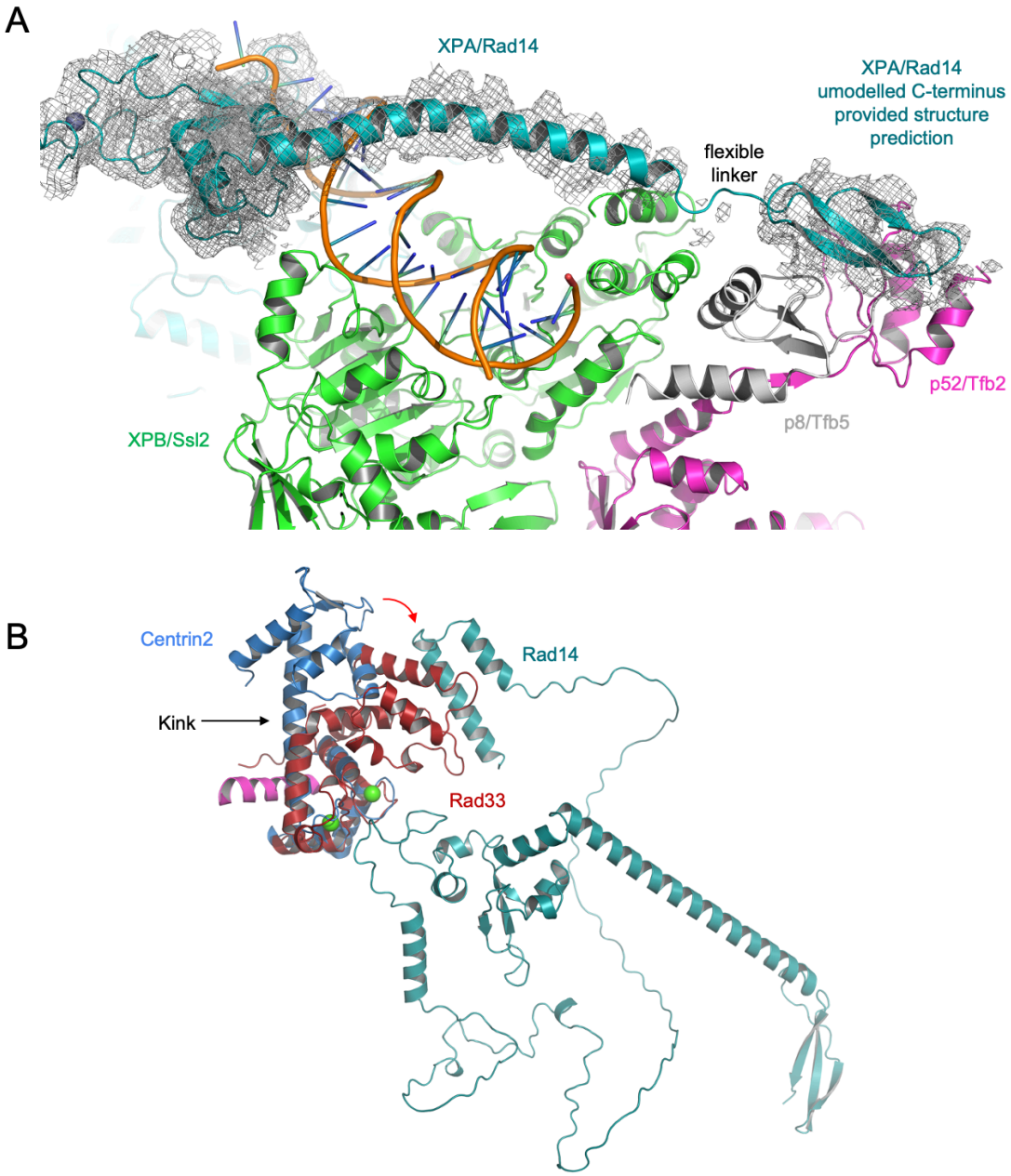


Figure S19. Structural analysis of the Rad33-Rad14 complex model

(**A**) C-terminus of Rad14 within the EM density of 6ro4. (**B**) Superposition of Centrin2 (blue, 2ggm) and Rad33 (red). Rad14 is shown in teal. All models are displayed as cartoons. The structures show a high degree of similarity. The model of Rad33-Rad14 shows a distinct kink in the middle of an extended helix that connects the N- and C- terminus of Centrin2 (Rad33).

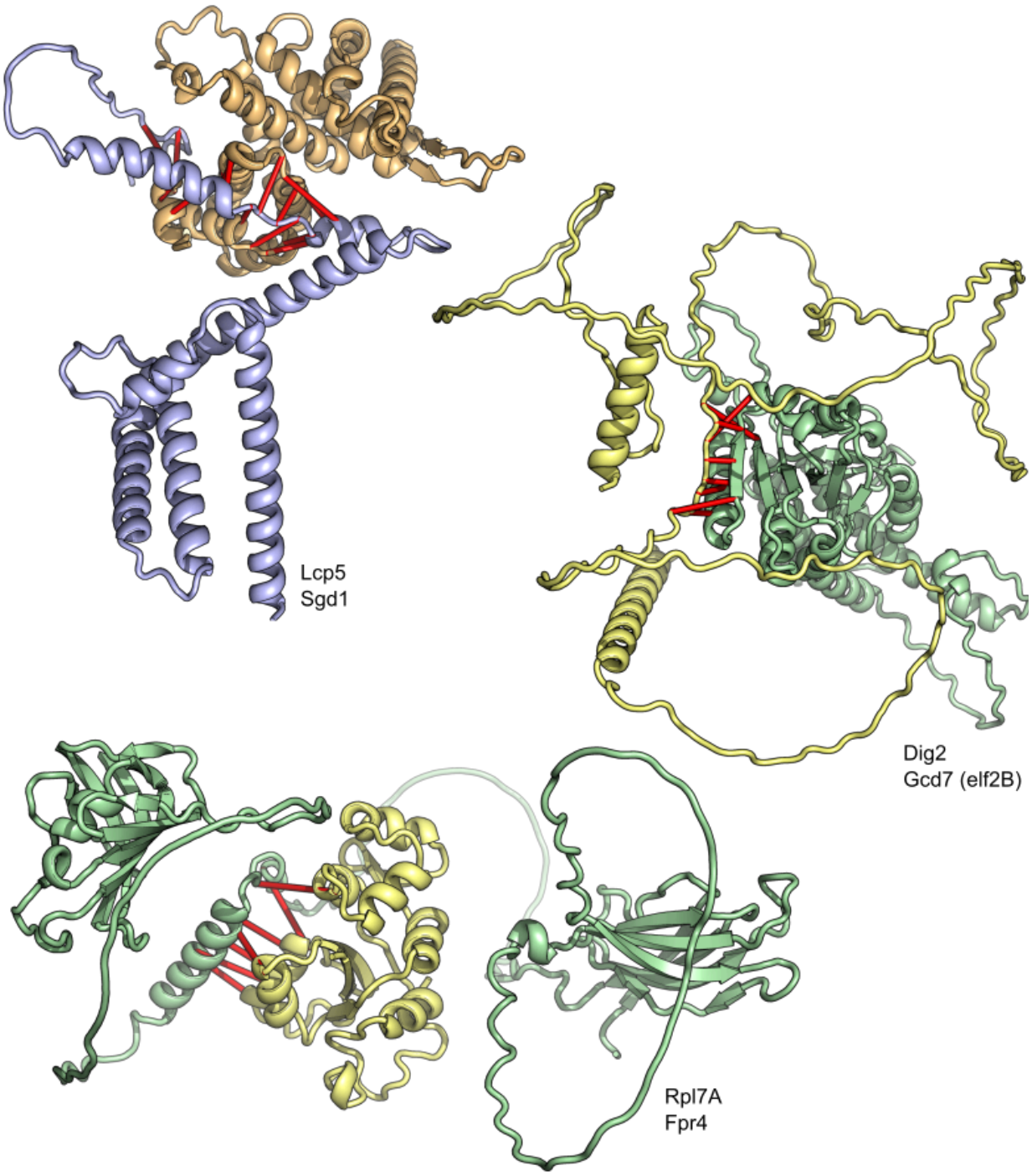


Figure S20. Structures of complexes involved in translation and ribosome regulation

Top predicted residue-residue contacts are indicated with bars. Pair color indicates the method of identification from Fig. 1; experiment-guided pairs are yellow and green and "*de novo*" pairs are blue and light-orange.

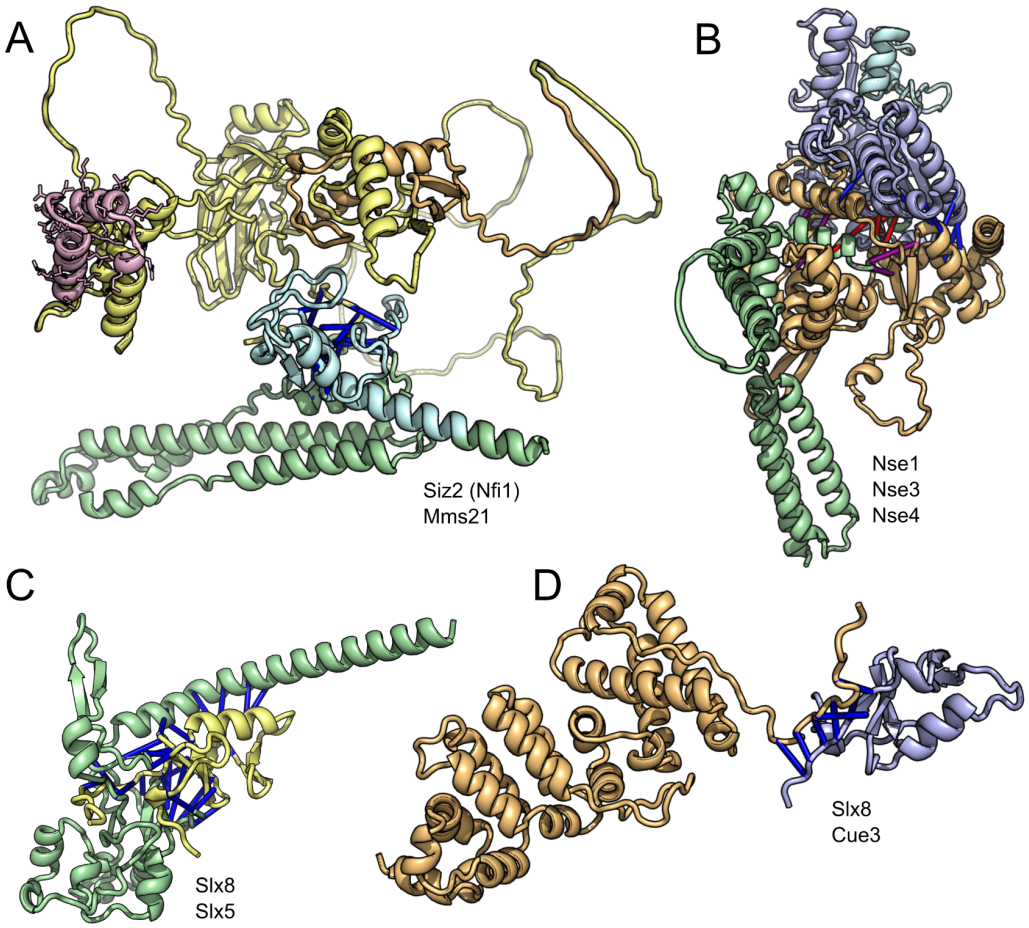


Figure S21. Complexes involving SUMO and ubiquitin ligases

Top predicted residue-residue contacts are indicated with bars. Pair color indicates the method of identification from Fig. 1; experiment- guided pairs are yellow and green (panel A) and "*de novo*" pairs are blue and light-orange (panel C). Highly disordered regions removed for visual clarity. (**A**) Sap domain of Siz2 is highlighted in light-pink with sticks, Siz2 Zn finger is in light-orange and Mms21 Zn finger is in light-teal. (**B**) Nse1 Zn finger is in light-teal. (**C**,**D**) colored as in Fig. 2.

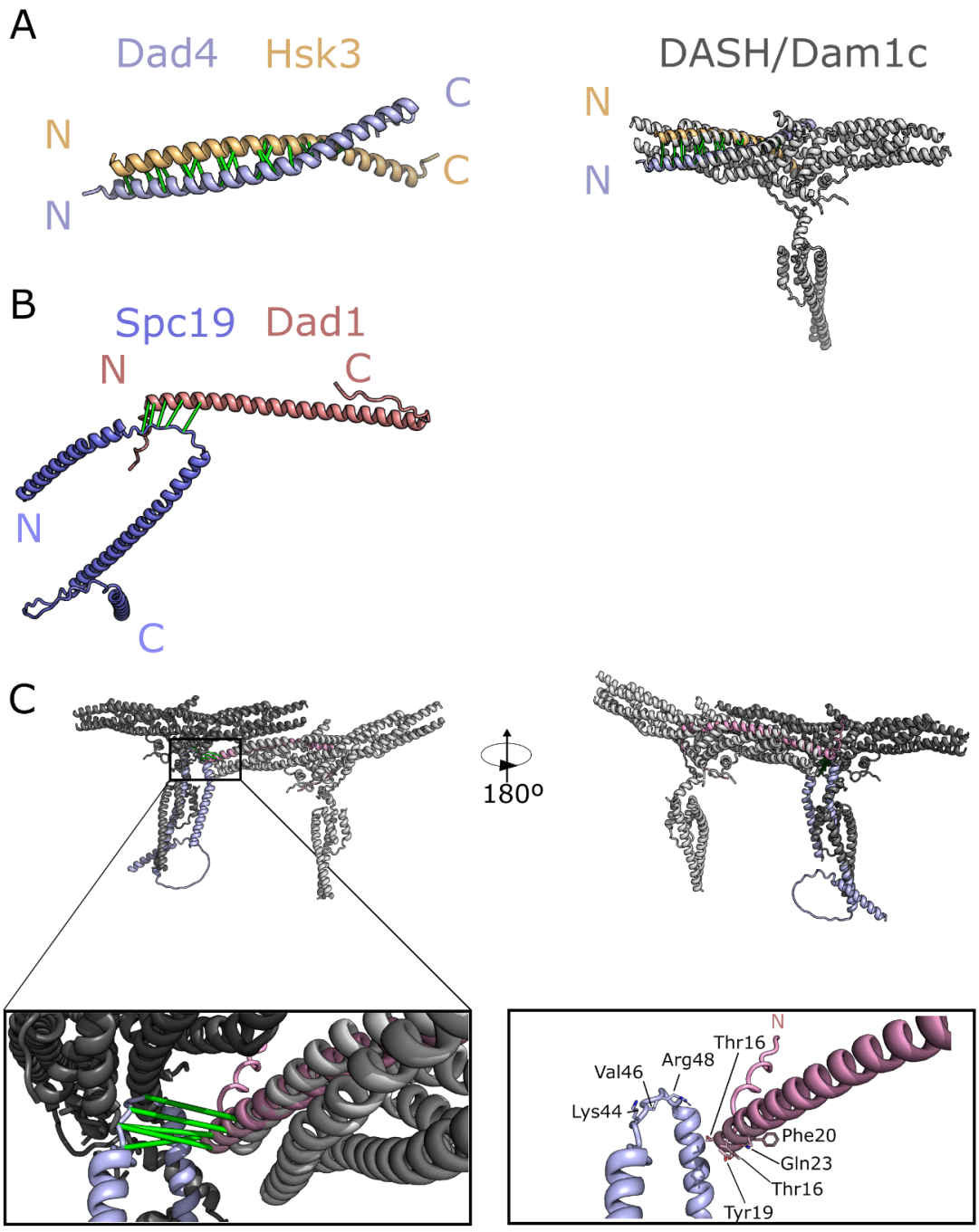
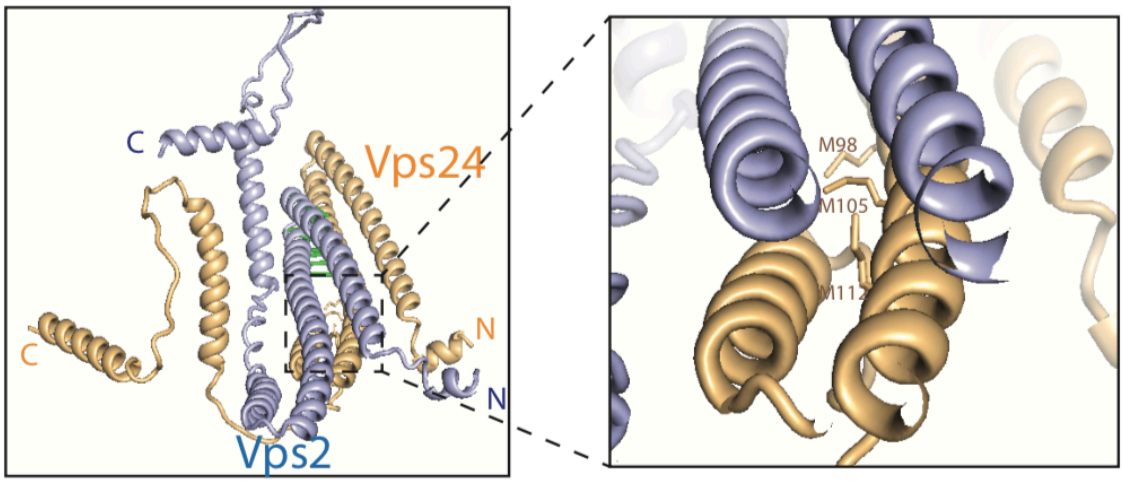


Figure S22. Predicted inter and intra-decamer interactions in the DASH/Dam1 complex

(A) Left: Predicted dimer between Dad4 (blue) and Hsk3 (gold) proteins, with predicted contacts shown in green. Right: The Dad4-Hsk3 dimer aligned with the structure of the DASH/Dam1 decamer complex from *C. thermophilus* (grey;PDB:6CFZ; (*66*)). (**B**) Left: Predicted dimer between Spc19 (blue) and Dad1 (pink) with predicted contacts shown in green. (**C**) Left and right: The Spc19 Dad1 dimer aligned with the structure of the DASH/Dam1 decamer complex from *C. thermophilus* (grey; PDB:6CFZ; (*66*)) in the context of ring formation. The image on the right has been rotated 180˚ about the Y axis. Bottom: A zoomed view of the Dad1-Spc19 interactions with (left) and without (right) the *C. thermophilus* structure visible.



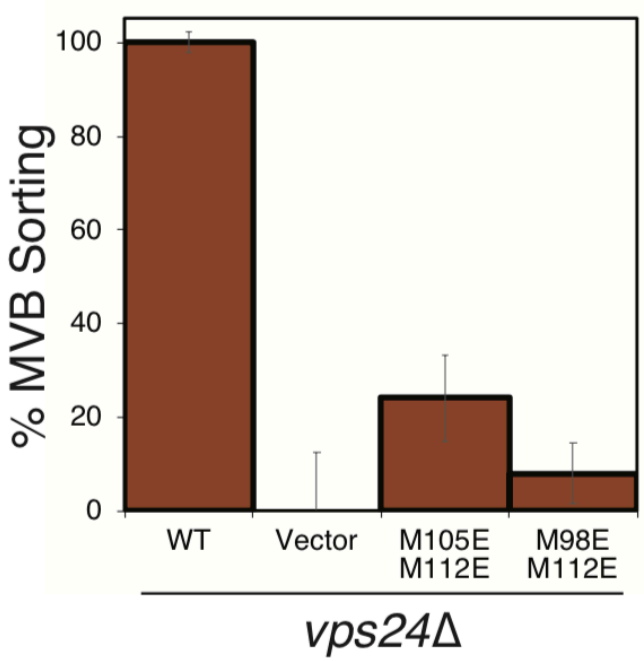


Figure S23. Predicted Vps2-Vps24 complex structure is consistent with unpublished mutagenesis data

Predicted ESCRT-III interface mutations inhibit cargo sorting functions. (**A**) Predicted structure of the yeast ESCRT-III complex Vps24-Vps2. Figure on the right represents a zoomed-in image of the box with dotted-lines. Residues M98, M105 and M112 in the helix-3 region of Vps24 are highlighted in "sticks" representation. (**B**) Data represent flow-cytometry cargo sorting assay in *S. cerevisiae*. The methionine transporter Mup1 tagged with pHluorin (Mup1-pHluorin) was used as cargo. Upon sorting to the vacuole, the fluorescence of pHluorin is quenched, which is quantified as 100% for WT and 0% for the empty vector. Error bars represent standard deviation of three independent experiments.

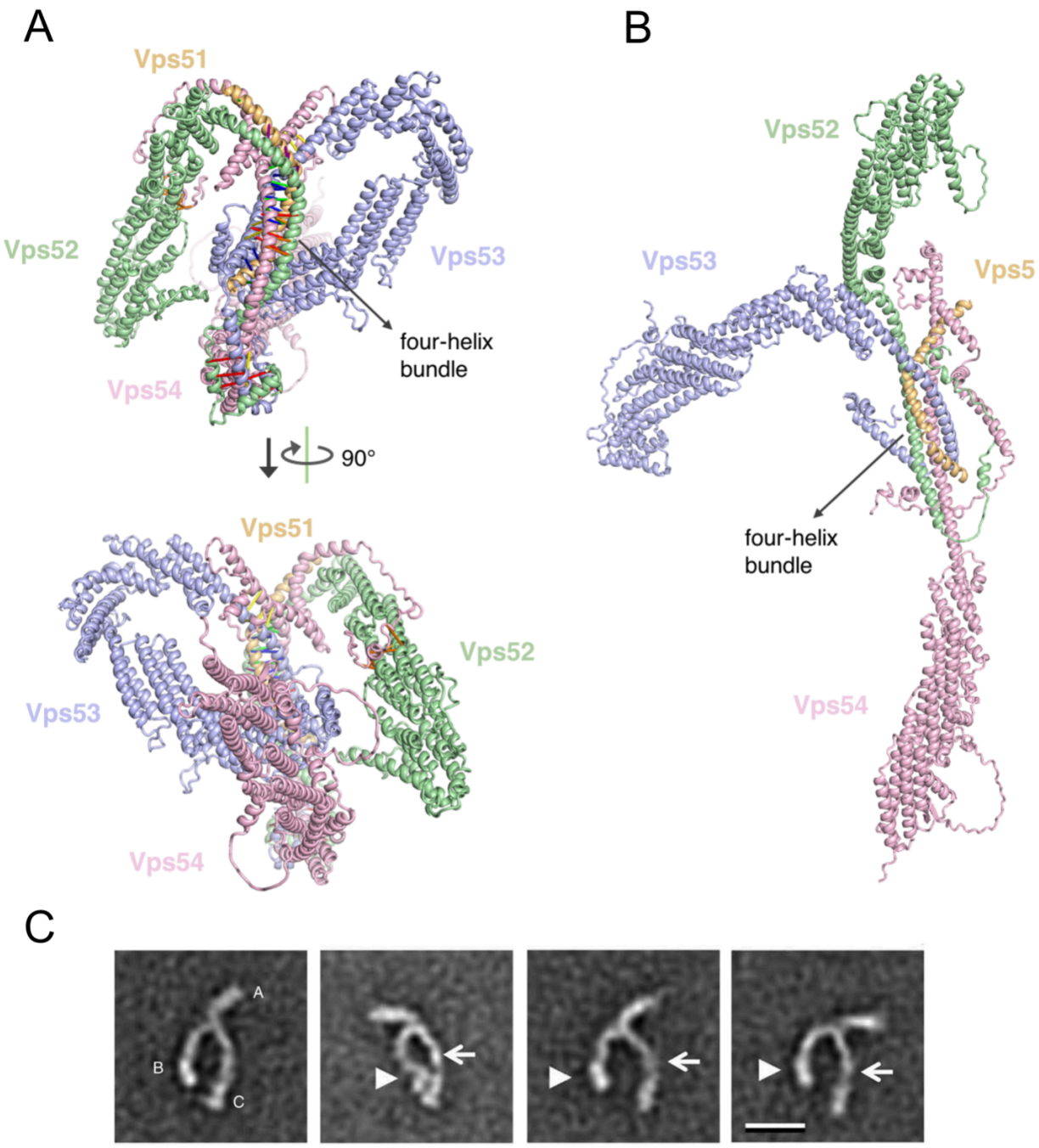


Figure S24. Predicted GARP complex structure is consistent with negative stain data

(**A**) Predicted model of the GARP complex is shown in two orientations. Predicted residue-residue contacts are indicated with bars. (**B**) Model constructed by superimposing individual AF2 predictions for Vps52, Vps53, and Vps54 onto the central four-helix bundle. The resulting model shows a different overall architecture. (**C**) 2D class averages of GARP complex with minor adaptations (77). Reprinted by permission from Springer Nature Customer Service Center GmbH: Springer Nature, *Nature Structural and Molecular Biology*, CATCHR, HOPS and CORVET tethering complexes share a similar architecture, H-T Chou, D. Dukovski, M.G. Chambers, K.M. Reinisch, and T. Walz, 2016.

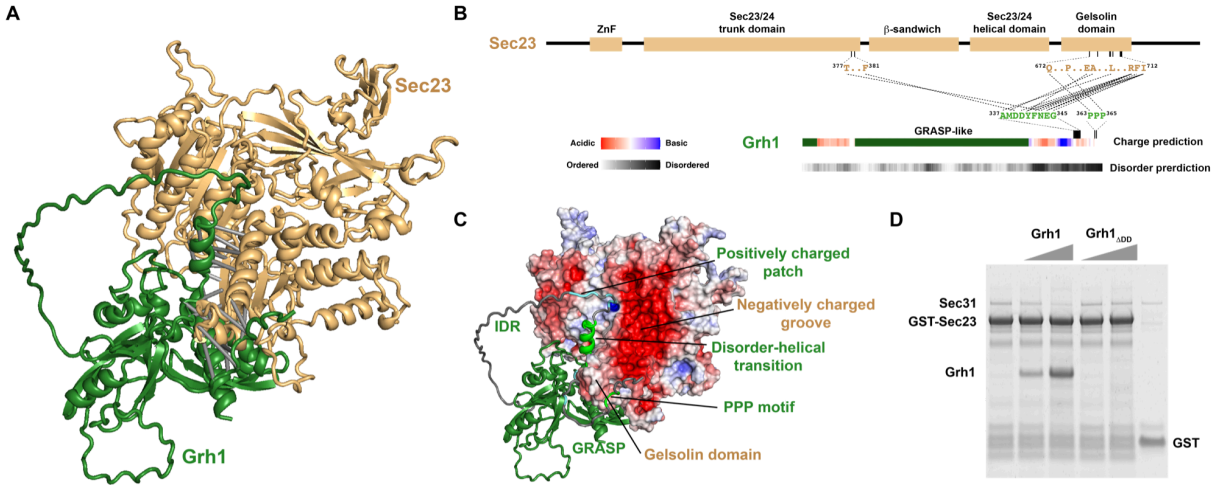


Figure S25. Predicted multivalent complex between Grh1-Sec23

(**A**) Predicted model of the Sec23-Grh1 complex showing predicted residue-residue contacts as grey bars. (**B**) Domain structure and predicted interactions of Sec23 (top) and Grh1 (bottom). Grh1 contains a C-terminal intrinsically disordered region (IDR) that hosts a positively charged cluster and two sites predicted to interact with the gelsolin domain in Sec23 (dashed lines). (**C**) Predicted structure of the Sec23-Grh1 complex with Sec23 rendered by surface charge. Two predicted interaction interfaces include a helical motif involving disorder-to-helical transition (light green) and a polyproline PPP motif (dark green). The flanking positively charged patch in the IDR may sample the negatively charged groove in Sec23 to anchor the multivalent interface. (**D**) GST pulldown assay illustrating the importance of interaction between C-terminal IDR of Grh1 and Sec23. Full length Grh1 was able to compete with outer coat protein Sec31 upon recruitment by Sec23; deletion of the C-terminal IDR which harbours the predicted multivalent interface completely abolished the recruitment of Grh1 to Sec23.

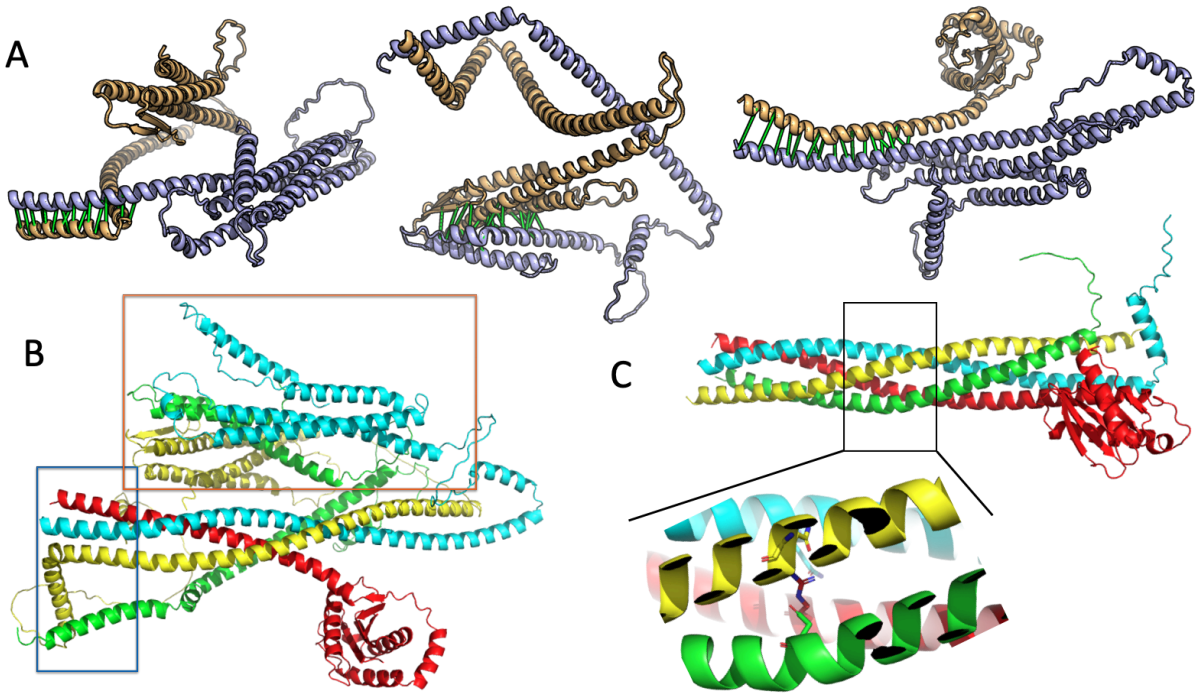


Figure S26. Modeling of a SNARE complex consisting of Use1, Ufe1, Sec20, and Sec22

The complex is involved in the fusion between Golgi-derived retrograde transport vesicles and the endoplasmic reticulum. (**A**) Binary complexes of UFE1-SEC20 (left), USE1-SEC20 (middle), UFE1-SEC22 (right). (**B**) A UFE1-USE1-SEC20-SEC22 complex modeled using full-length proteins with AF. UFE1: cyan, USE1: green, SEC20: yellow, SEC22: red. UFE1, SEC20, and SEC22 form a three-helical bundle through part of their SNARE motifs, while the SNARE motif of USE1 is however not included in this helical bundle. The transmembrane helices from these proteins are roughly located within the blue box. These transmembrane helices and additional likely flexible helices within the red box were excluded to build a SNARE model in (**C**)**.** (**C**) A UFE1-USE1-SEC20-SEC22 complex (upper) modeled by AF using only the regions that tend to form a helical bundle in panel B. We inspected the residues mediating the interaction between the SNARE motifs. In general, the interactions involved layers of four hydrophobic residues (one residue from each helix), but in the middle of the four-helix bundle there is a polar layer formed by Arg157 of SEC22, Gln289 of UFE1, Gln237 of SEC20, Asp183 of USE1 (lower). This arrangement resembles that observed for the neuronal SNARE complex, where there is a polar layer formed by one Arg and three Gln residues in the middle, and hydrophobic layers elsewhere (*84*).

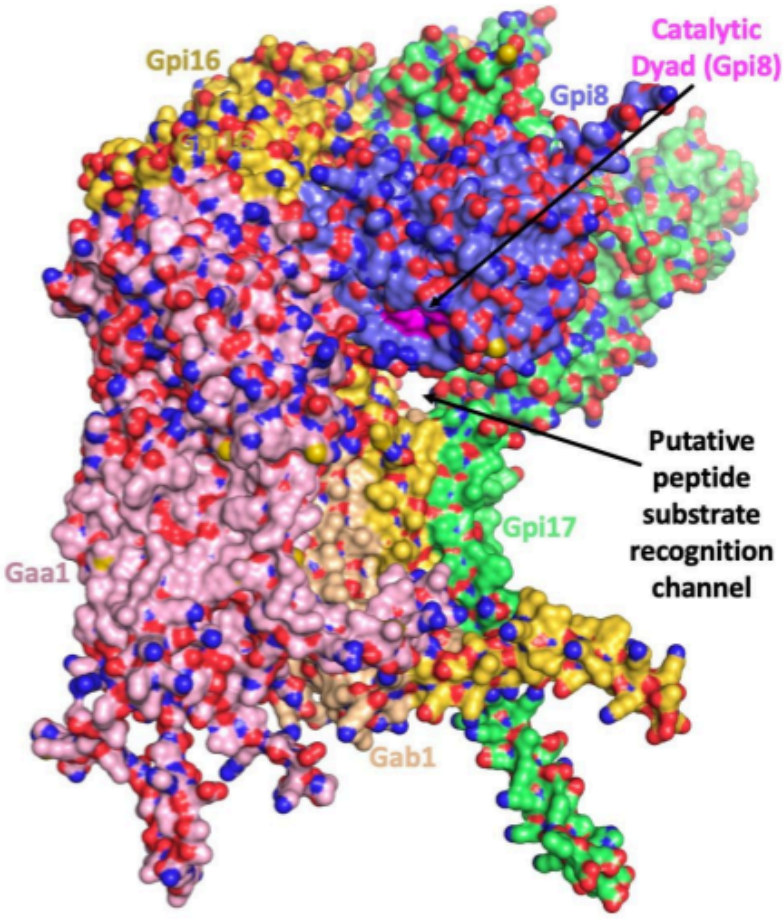


Figure S27. C-terminal GPI-T signal sequence recognition tunnel suggested by complex model of GPI transamidase

GPI-T recognizes a C-terminal signal sequence composed of a pattern of small, hydrophilic, and then hydrophobic residues. The model of GPI-T reveals a putative channel between the Gpi8, Gpi16, and Gpi17 subunits. The position of this tunnel is shown in a cartoon model (Fig. 5E) and in a surface model (shown above) of GPI-T. The side chains of Ala168 in Gpi8 and Phe500 in Gaa1 were hidden for better visualization of this channel. The N-terminal signal peptides were removed from Gpi8 (residues 1-22) and Gpi16 (residues 1-19). For Gpi8, only residues 23-306 are included in the model. Each subunit is color coded as indicated, with the catalytic dyad in Gpi8 (Cys199 and His157) highlighted in magenta.

Supplemental Tables

Table S1. Datasets we obtained from Yeast Interactome Database

Name	Description
Y2H set	The union of CCSB-YI1, Ito-core, and Uetz-screen
APMS set	Co-complex membership associations, Combined-AP/MS
Literature-curated set	Literature-curated interactions, LC-multiple
Gold-standard set	Positive Reference Set (PRS) and Binary Gold standard set (Binary-GS)

Table S2: Extended annotations for modeled PPIs in Fig. 2.

Protein 1	Protein 2	Group	Protein 1 Annotation	Protein 2 Annotation
Nse1	Nse3	DNA Repair	Non-structural maintenance of chromosomes element 1	Non-structural maintenance of chromosome element 3
Rad33	Rad14	DNA Repair	DNA repair protein RAD33	DNA repair protein RAD14
Rmi1	Top3	DNA Repair	RecQ-mediated genome instability protein 1	DNA topoisomerase 3
Rev7	Rev1	DNA Repair	DNA polymerase zeta processivity subunit	DNA repair protein REV1
Rad57	Rad55	DNA Repair	DNA repair protein RAD57	DNA repair protein RAD55
Ddc1	Rad17	Mitosis, Meiosis, or DNA Damage	DNA damage checkpoint protein 1	DNA damage checkpoint control protein RAD17
Ddc1	Mec3	Mitosis, Meiosis, or DNA Damage	DNA damage checkpoint protein 1	DNA damage checkpoint control protein MEC3
Mec3	Rad17	Mitosis, Meiosis, or DNA Damage	DNA damage checkpoint control protein MEC3	DNA damage checkpoint control protein RAD17
Spo11	Rec102	Mitosis, Meiosis, or DNA Damage	Meiosis-specific protein SPO11	Meiotic recombination protein REC102
Ski8	Spo11	Mitosis, Meiosis, or DNA Damage	Antiviral protein SKI8	Meiosis-specific protein SPO11
Bub2	Tem1	Mitosis, Meiosis, or DNA Damage	Mitotic check point protein BUB2	Protein TEM1
Bfa1	Bub2	Mitosis, Meiosis, or DNA Damage	Mitotic check point protein BFA1	Mitotic check point protein BUB2
Taf3	Taf10	Transcription	Transcription initiation factor TFIID subunit 3	Transcription initiation factor TFIID subunit 10
Ssu72	Pta1	Transcription	RNA polymerase II subunit A C-terminal domain phosphatase SSU72	Pre-tRNA-processing protein PTA1
Npa3	Gpn3	Transcription	GPN-loop GTPase 1	GPN-loop GTPase 3
Ino2	Ino4	Transcription	Protein INO2	Protein INO4
Kti12	Elp2	Transcription	Killer toxin insensitivity protein 12	Elongator complex protein 2
Sas10	Noc4	Translation	Something about silencing protein 10	Nucleolar complex protein 4
Lcp5	Bfr2	Translation	U3 small nucleolar ribonucleoprotein protein LCP5	Protein BFR2 (Brefeldin A resistance protein 2)
Pcc1	Kae1	Translation	EKC/KEOPS complex subunit PCC1	tRNA N6-adenosine threonylcarbamoyltransferase
Ncs2	Ncs6	Translation	Cytoplasmic tRNA 2-thiolation protein 2	Cytoplasmic tRNA 2-thiolation protein 1
Rpl12B	Rmt2	Translation	60S ribosomal protein L12-B	Protein arginine N-methyltransferase 2

Table S3: Extended annotations for modeled PPIs from Fig. 3.

Protein 1	Protein 2	Group	Protein 1 Annotation	Protein 2 Annotation
Yif1	Yip1	Protein & Ion Transport	Protein transport protein YIF1	Protein transport protein YIP1
Ksh1	Yos1	Protein & Ion Transport	Protein Kish	Protein transport protein YOS1
Tca17	Trs130	Protein & Ion Transport	TRAPP-associated protein TCA17	Trafficking protein particle complex II-specific subunit 130
Vtc1	Vtc4	Protein & Ion Transport	Vacuolar transport chaperone 1	Vacuolar transporter chaperone 4
Vps68	Vps55	Protein & Ion Transport	Vacuolar protein sorting-associated protein 68	Vacuolar protein sorting-associated protein 55
Sed5	Sft2	Protein & Ion Transport	Integral membrane protein SED5	Protein transport protein SFT2
Yip1	Ksh1	Protein & Ion Transport	Protein transport protein YIP1	Protein kish
Cbp3	Cbp6	Mitochondria	Protein CBP3, mitochondrial	Cytochrome B pre-mRNA-processing protein 6
Cox11	Cox19	Mitochondria	Cytochrome c oxidase assembly protein COX11, mitochondrial	Cytochrome c oxidase assembly protein COX19
Tim17	Mgr2	Mitochondria	Mitochondrial import inner membrane translocase subunit TIM17	Protein MGR2 (Mitochondrial genome-required protein 2)
Tim44	Tim23	Mitochondria	Mitochondrial import inner membrane translocase subunit TIM44	Mitochondrial import inner membrane translocase subunit TIM23
Tim23	Tim17	Mitochondria	Mitochondrial import inner membrane translocase subunit TIM23	Mitochondrial import inner membrane translocase subunit TIM17
Spc3	Spc1	Protein Translocation	Signal peptidase complex subunit SPC3	Signal peptidase complex subunit SPC1
Sec11	Spc3	Protein Translocation	Signal peptidase complex catalytic subunit SEC11	Signal peptidase complex subunit SPC3
Spc2	Sec11	Protein Translocation	Signal peptidase complex subunit SPC2	Signal peptidase complex catalytic subunit SEC11
Srp68	Srp72	Protein Translocation	Signal recognition particle subunit SRP68	Signal recognition particle subunit SRP72
Srp21	Srp14	Protein Translocation	Signal recognition particle subunit SRP21	Signal recognition particle subunit SRP14
Pac2	Tub1	Tubulin	Protein PAC2	Tubulin alpha-1 chain
Rbl2	Tub2	Tubulin	Tubulin-specific chaperone A	Tubulin beta chain
Ftr1	Fet5	Iron Transport	Plasma membrane iron permease	Iron transport multicopper oxidase FET5

Table S4: Extended annotations for modeled PPIs in Fig. 4.

Protein 1	Protein 2	Group	Protein 1 Annotation	Protein 2 Annotation
Mdh2	Mdh1	Enzyme	Malate dehydrogenase	Malate dehydrogenase
Rip1	Mzm1	Enzyme	Cytochrome b-c1 complex subunit Rieske, mitochondrial	Mitochondrial zinc maintenance protein 1, mitochondrial
Sdh2	Sdh8	Enzyme	Succinate dehydrogenase [ubiquinone] iron-sulfur subunit, mitochondrial	Succinate dehydrogenase assembly factor 4, mitochondrial
Gpi2	Gpi19	Enzyme (GPI)	Phosphatidylinositol N-acetylglucosaminyltransferase GPI2 subunit	Phosphatidylinositol N-acetylglucosaminyltransferase subunit GPI19
Gpi14	Pbn1	Enzyme (GPI)	GPI mannosyltransferase 1	Protein PBN1 (Protease B non-derepressible protein 1)
Gpi1	Arv1	Enzyme (GPI)	Phosphatidylinositol N-acetylglucosaminyltransferase subunit GPI1	Protein ARV1
Pga1	Gpi18	Enzyme (GPI)	GPI mannosyltransferase 2 subunit PGA1	GPI mannosyltransferase 2
Rip1	Cmc4	Uncharacterized	Cytochrome b-c1 complex subunit Rieske, mitochondrial	Cx9C motif-containing protein 4, mitochondrial
Rpp0	Ylr287C	Uncharacterized	60S acidic ribosomal protein P0	Uncharacterized protein YLR287C
Dus4	Ypl108W	Uncharacterized	tRNA-dihydrouridine(20a/20b) synthase [NAD(P)+]	Uncharacterized protein YPL108W
Ynr021W	Ypr063C	Uncharacterized	UPF0674 endoplasmic reticulum membrane protein YNR021W	Uncharacterized protein YPR063C
Nyv1	Ygr016W	Uncharacterized	Vacuolar v-SNARE NYV1	Uncharacterized membrane protein YGR016W
Rvs167	Ypl077C	Uncharacterized	Reduced viability upon starvation protein 167	Uncharacterized protein YPL07C

Table S5: Extended annotations for modeled PPIs in Fig. 5

Protein 1	Protein 2	Protein 3	Protein 1 Annotation	Protein 2 Annotation	Protein 3 Annotation
Nhp10	Ies3	Ies5	Non-histone protein 10	Ino eighty subunit 3	Ino eighty subunit 5
Cgi121	Bud32	Kae1	EKC/KEOPS complex subunit CGI121	EKC/KEOPS complex subunit BUD32	tRNA N6-adenosine threonylcarbamoyltransferase
Yra2	Tho1	Sub2	RNA annealing protein YRA2	THO complex subunit 1	ATP-dependent RNA helicase SUB2
Aim7	Arc15	Arc40	Protein AIM7	Actin-related protein 2/3 complex subunit 5	Actin-related protein 2/3 complex subunit 1
Sen2	Sen15	Sen34	tRNA-splicing endonuclease subunit SEN2	tRNA-splicing endonuclease subunit SEN15	tRNA-splicing endonuclease subunit SEN34
Sen15	Sen54	Sen34	tRNA-splicing endonuclease subunit SEN15	tRNA-splicing endonuclease subunit SEN54	tRNA-splicing endonuclease subunit SEN34
Rpc19	Rpb10	Rpb3	DNA-directed RNA polymerases I and III subunit RPAC2	DNA-directed RNA polymerases I, II, and III subunit RPABC5	DNA-directed RNA polymerase II subunit RPB3

Panel	Protein	Protein Annotation
B	Rad57	DNA repair protein RAD57
B	Rad55	DNA repair protein RAD55
B	Rad51	DNA repair protein RAD51
C	Vps53	Vacuolar protein sorting-associated protein 53
C	Vps51	Vacuolar protein sorting-associated protein 51
C	Vps52	Vacuolar protein sorting-associated protein 52
C	Vps54	Vacuolar protein sorting-associated protein 54
D	Rad33	DNA repair protein Rad33
D	Rad14	DNA repair protein Rad14
E	Gpi8	GPI transamidase component GPI8
E	Gab1	GPI transamidase component GAB1
E	Gpi17	GPI transamidase component GPI17
E	Gaa1	GPI transamidase component GAA1
E	Gpi16	GPI transamidase component GPI16

Chapter 3: PROTEIN INTERACTIONS IN HUMAN PATHOGENS REVEALED THROUGH DEEP LEARNING

A version of this chapter has been accepted at *Nature Microbiology* and may be found on bioRxiv prior to print:

Ian R. Humphreys[†], Jing Zhang[†], Minkyung Baek[†,#], Yaxi Wang[†], Aditya Krishnakumar, Jimin Pei, Ivan Anishchenko, Catherine A. Tower, Blake A. Jackson, Thulasi Warrier, Deborah T. Hung, S. Brook Peterson, Joseph D. Mougous, Qian Cong[#], and David Baker[#]. "Essential and virulence related protein interactions of pathogens revealed through deep learning". *bioRxiv*, 2024.

[†]These authors contributed equally to this work.

Abstract

Identification of bacterial protein–protein interactions and predicting the structures of the complexes could aid in the understanding of pathogenicity mechanisms and developing treatments for infectious diseases. Here, we developed RoseTTAFold2-Lite, a rapid deep learning model that leverages residue-residue coevolution and protein structure prediction to systematically identify and structurally characterize protein-protein interactions at the proteome-wide scale. Using this pipeline, we searched through 78 million pairs of proteins across 19 human bacterial pathogens and identified 1923 confidently predicted complexes involving essential genes and 256 involving virulence factors. Many of these complexes were not previously known; we experimentally tested 12 such predictions, and half of them were validated. The predicted interactions span core metabolic and virulence pathways ranging from post-transcriptional modification to acid neutralization to outer membrane machinery and should contribute to our understanding of the biology of these important pathogens and the design of drugs to combat them.

Introduction

Understanding the biology of pathogenic bacteria is important for human health and therapeutics. Protein-protein interactions (PPIs) are central to biological processes, but many interactions remain unknown, especially for non-model organisms. High-throughput experiments such as the two-hybrid screen and affinity purification coupled with mass spectrometry (AP/MS) have been used to identify PPIs in a variety of organisms (*1*–*3*). However, such methods can fail to reveal transient interactions and be plagued by non-specific interactions in non-physiological conditions, which result in discrepancies between experiments along with high false-positive

and false-negative rates (*4*, *5*). Interacting proteins often co-evolve, and hence amino acid coevolution can be exploited to assess the likelihood that two proteins interact with each other. Coevolutionary information between proteins extracted from paired multiple sequence alignments (pMSAs) of orthologous proteins (*6*–*8*) has been used to systematically identify PPIs in prokaryotes at an accuracy that rivals experimental screens (*7*). Supplementing coevolution with deep learning (DL) based structure prediction methods has further increased the accuracy of PPI prediction, enabling large-scale prediction of PPIs in yeast (*9*) and humans (*10*, *11*).

We set out to systematically identify and structurally characterize PPIs in pathogenic bacteria. We selected 19 bacterial pathogens (table S1) that span 6 phyla and are the leading causes of pathogen-associated deaths in humans (12). These organisms are associated with infections in skin (*Staphylococcus aureus*), gastrointestinal tract (*Clostridioides difficile*, *Helicobacter pylori*, *Listeria monocytogenes*, and *Salmonella typhimurium*), respiratory system (*Legionella pneumophila, Mycobacterium tuberculosis, Pseudomonas aeruginosa, and Streptococcus pneumoniae*), urinary and genital tracts (*Chlamydia trachomatis* and *Mycoplasma genitalium*), and the plague (*Yersinia pestis).* For most of these organisms, large-scale experimental screens have identified essential genes (EGs) virulence factors (VFs); these results are summarized in the Database of Essential Genes (DEG) (*12*) and Virulence Factor DataBase (VFDB) (*13*). We focused on EGs and VFs because the former provides targets for drug development to inhibit essential cellular functions and treat infectious diseases, while the latter may explain molecular mechanisms of pathogenicity. Comparative analysis showed significant overlap in the sets of EGs between different pathogens, but each pathogen still harbors ~100 unique EGs (table S2). In contrast, VFs differ considerably between species, suggesting a diversity of virulence mechanisms which we attempt to capture with our set of phylogenetically diverse species (table S2).

Results

Computational pipeline for proteome-wide PPI identification

To screen through hundreds of millions of protein pairs for PPIs, we first sought to increase the computational efficiency of PPI identification without compromising accuracy. We previously developed a 2-track RoseTTAFold (RF 2-track) network that is a simplified version of RoseTTAFold (*14*). Although RF 2-track was not trained to model protein complexes or distinguish interacting from non-interacting proteins, residue-residue distograms produced by this network enable the detection of PPIs on a proteome-wide scale at an accuracy that far exceeds statistical analysis of coevolution between proteins (*9*). Similarly, we and others have used AlphaFold (AF) (*15*) to evaluate interactions identified in lower accuracy large-scale screens (*9*–*11*, *16*, *17*); the computational cost of AF prohibits its application on a protome-wide scale. AF-multimer (*18*) was trained to model 3D structures of known protein complexes, and consequently, it tends to predict PPIs between non-interacting pairs, displaying a worse performance in distinguishing true PPIs from random pairs than AF (**Fig. 1B** top).

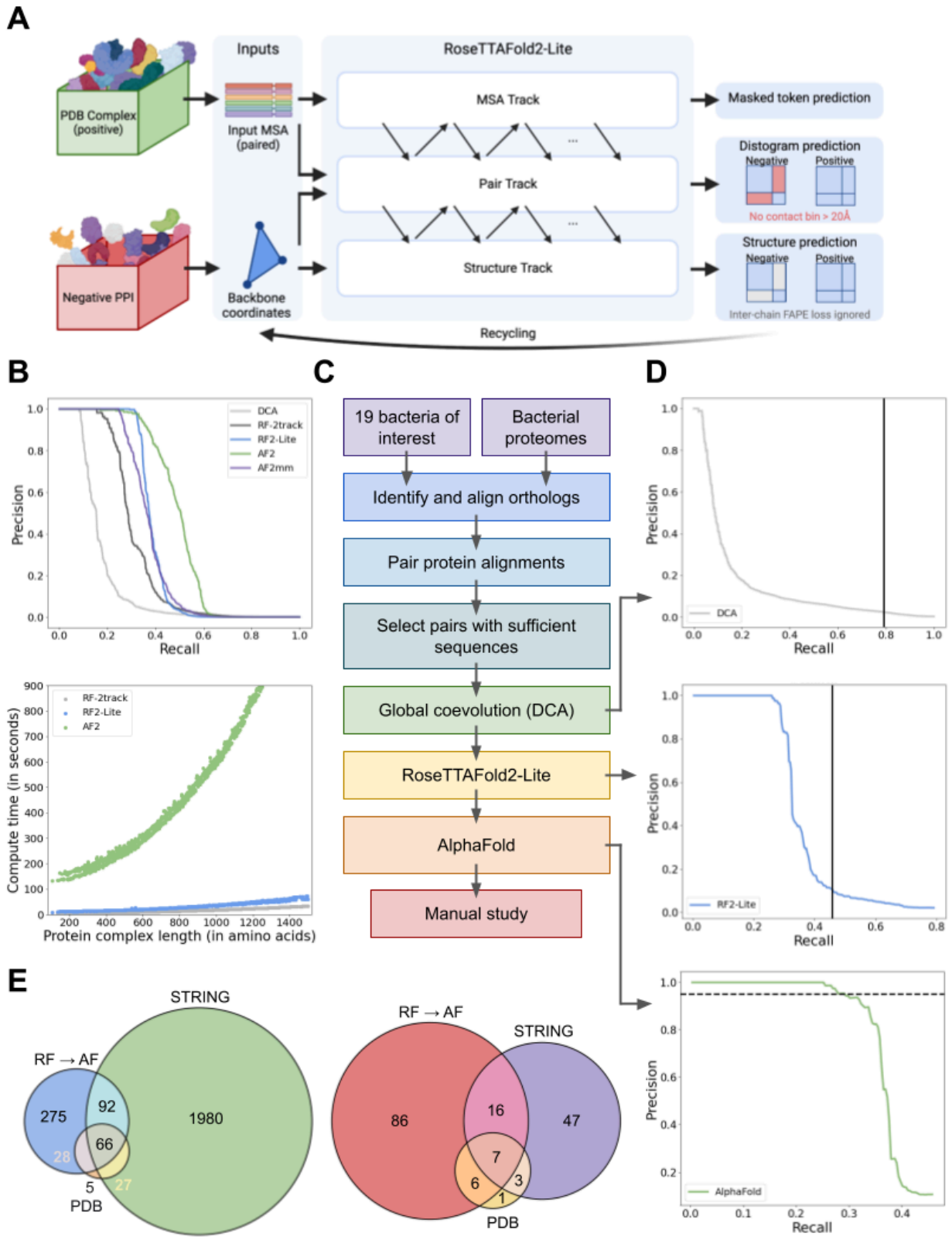


Figure 1: Protein-protein interaction identification by coevolution and deep learning methods
(A) Overview of the RF2-Lite network architecture. (**B**) Benchmark performance of PPI prediction

methods. Top: precision and recall curves of DCA (grey), RF 2-track (black), RF2-Lite (blue), AF (green), and AF-multimer (purple) in distinguishing true PPIs from random protein pairs. For different methods, we used the pMSAs generated by our bioinformatic pipeline (Supplemental Methods). We applied each method on a benchmark set of 1000 randomly selected positive control pairs and 10,000 negative control pairs (Supplemental Methods). The precision and recall curve for this benchmark is in fig. S6A. Real signal-to-noise ratio for the PPI screen is on the order of 1:1000 (*1*); to reflect the impact of a much larger set of non-interacting pairs, we upsampled the negative control set to 1,000,000 by randomly sampling 100 “pseudo” interacting probabilities from the Gaussian distribution around each real interacting probability we obtained for the negative controls with a standard deviation of 0.1. Bottom: runtime comparison of different methods. **(C)** Schematic overview of our PPI screen pipeline. **(D)** Precision and recall curves at different stages in the pipeline. Top: DCA on PPI prediction; solid black vertical line represents the recall cutoff in this stage. Middle: RF2-Lite screen procedure on the 'pilot set'; solid black vertical line indicates the recall cutoff at this stage. Bottom: AF screen procedure on the 'pilot set'; dashed horizontal line shows the precision cutoff, i.e., 0.95. **(E)** Summary of predicted PPIs for the “pilot set” that focuses on EGs and VFs. Left: interactions between interacting EGs in the 'pilot set' based on different evidence: blue, green, and orange circles represent our predicted pairs, functional interactions according to STRING (total score ≥ 900 and experimental score ≥ 400), and interacting pairs according to PDB (BLAST hit to complex in PDB e-value ≤ 0.00001, sequence identity ≥ 50% and coverage ≥ 50%), respectively. Right: PPIs involving VFs in the 'pilot set' supported by difference evidence: red, purple, and yellow circles represent our predictions, pairs according to STRING, and pairs according to PDB.

We hypothesized that a dedicated lighter-weight network trained on both interacting and non-interacting protein pairs that balances accuracy with speed could assist proteome-wide PPI screens. We revised the original RF network by introducing architectural improvements to increase accuracy while reducing the number of layers to enable the rapid computation necessary for large-scale screens (**Fig. 1A**; Supplemental Methods). We trained this network using a combination of (1) monomeric protein structures from Protein Data Bank (PDB), (2) AF models of UniRef50 sequences, (3) pairwise protein complex structures extracted from PDB, and (4) random non-interacting protein pairs. The four types of training data were mixed at a ratio of 1:3:2:2 (table S3). The model was trained using the masked language model (MLM) loss, distogram (dist) prediction loss, frame aligned point error (FAPE) loss, accuracy estimation loss, bond geometry loss, and van der Waals (vdW) energy loss. For the negative interaction examples, we ignored the inter-chain region for FAPE calculation and required the network to predict the distogram to be in the “non-interacting bin” for the inter-chain region. We designate the resulting network as RoseTTAFold2-Lite (RF2-Lite) as it resembles the RoseTTAFold2 architecture but has many fewer parameters (*19*). RF2-Lite has improved performance in distinguishing true PPIs over the previous RF 2-track: at the same precision, the recall for true

PPIs by RF2-Lite is in between RF 2-track and AF (**Fig. 1B** top; fig. S6-S7). Despite this increase in accuracy, RF2-Lite's speed is still comparable to RF 2-track, and it requires about 20-fold less compute time than AF (**Fig. 1B** bottom).

We combined direct coupling analysis (DCA) (*20*), RF2-Lite, and AF (**Fig. 1C**; Supplemental Methods) to identify and model interacting proteins and applied this pipeline to the 19 human pathogens listed in table S2. To monitor the performance of our pipeline, we assembled a set of positive controls and a ~700-fold larger negative set based on information from the STRING database (Supplemental Methods).

We constructed a database of 44,871 representative bacterial proteomes/genomes (one per species) obtained from NCBI and used the reciprocal best hit criteria (*21*) to identify an orthologue for every protein in each proteome (fig. S1). We aligned these orthologous sequences (*22*, *23*), and for each protein pair in each of the 19 pathogens (fig. S2), we concatenated their MSAs by connecting sequences of the same species to generate pMSAs (fig. S3). We removed proteins whose monomeric structure could not be confidently modeled by AF (average pLDDT < 50 in AFDB) and filtered the pMSAs based on their depth and quality (figs. S4,S5): of the total 140.2 million protein pairs, we selected 77.9 million (56%) with higher monomer structure and MSA quality.

We assessed the residue-residue coevolution for the selected pairs using DCA, and found that the 7.7 million (10%) high-scoring protein pairs by DCA contained 79% of the positive controls (**Fig. 1D** top). Among these 7.7 million pairs, we initially focused on a "pilot set" of 0.14 million pairs involving at least one VF (according to VFDB) and 0.83 million pairs of EGs (according to DEG). We removed redundancy in this set by clustering proteins from the 19 species into orthologous groups using OrthoMCL (*24*). If the orthologs of a protein pair were present in multiple species, we selected only one pair with the highest DCA score, resulting in a total of

457,310 representative PPI candidates.

We used RF2-Lite to identify confident PPIs from the “pilot set” and observed that we could achieve a recall of 28% at a precision of 95% when an RF2-Lite contact probability cutoff of 0.74 was used (**Fig. 1D** middle). We investigated whether using a loose RF2-Lite cutoff (contact probability 0.05) to select candidate PPIs (46,609, around 10% selected) for AF could improve recall. The RF2-Lite → AF pipeline only improved the recall to 29% at 95% precision (**Fig. 1D** bottom) at the cost of using 3 times more computer resources than simply relying on RF2-Lite to detect PPIs (table S4). Thus, the contribution of AF in distinguishing true PPIs from random pairs is limited, but it remains essential for obtaining high-quality 3D structures for the predicted protein complexes.

The successive use of DCA (selecting top 10%), RF2-Lite (cutoff: 0.05), and AF (cutoff: 0.9) collectively reduced the total number of random pairs by nearly 10,000 fold, resulting in 562 highly confident predictions from the “pilot set”. The identified binary protein complexes include 461 protein complexes involving EGs (**Fig. 1E** left) and 115 involving VFs (**Fig. 1E** right). Further investigation of these interactions may be useful for understanding the mechanisms of pathogenicity and developing disease prevention and treatment strategies. The vast majority (19%) of predicted protein complexes from the “pilot set” did not have experimental 3D structures in PDB (BLAST e-value ≤ 0.00001, identity ≥ 50%, and coverage ≥ 50% for both proteins), and half do not have confident experimental support according to STRING (*25*) (table S5).

To gain more structural and functional insights into these pathogens, we applied the RF2-Lite → AF pipeline to an additional 3.82 million pairs involving essential proteins and biological processes of therapeutic interest, such as the outer membrane machinery (table S6). This search resulted in an additional 3,051 predicted PPIs. To facilitate downstream studies, we

deposited all confident models to ModelArchives (modelarchive.org/doi/10.5452/ma-bak-evip) and provided additional metadata in the supplemental data file S1. Inspection of the predicted PPIs revealed a small number of proteins (in particular ferredoxin and rubredoxin) with predicted interactions between many random proteins, likely constituting small false positive hubs. We removed 405 PPIs involving such potential false positive hubs before deposition to ModelArchive.

It is difficult to cover even a small fraction of the biological insights that can be revealed from these 3D structures of protein complexes in one paper. In the following sections, we first describe experimental validation for a subset of predictions and then highlight examples that illustrate some of the biological insights revealed by the identification of putative PPIs and computational modeling of protein complexes.

Experimental validation

To corroborate our benchmarking analyses, which suggest that our predicted interactions should be quite accurate, we selected two sets of predicted interactions for experimental characterization. We biased these selections towards PPIs with no prior experimental evidence or strong functional associations because validating such interactions could provide new biological insights. The first set (table S7) was selected based on statistical methods (GREMLIN) for PPI detection, prior to the development and application of the DL methods. This set was used to probe the accuracy of statistical (DCA and GREMLIN (*26*)) versus DL methods for PPI detection. The second set (table S8) was selected from our final set of predicted 3,613 PPIs, with a goal of evaluating the accuracy of our current entire pipeline.

We selected the first dataset using the following criteria: (1) at least 20 kb apart (with a minimum

of 20 intervening genes), (2) not having homologous complexes in the PDB, (3) not predicted to have the same molecular function, (4) not annotated as part of the same biological pathway, and (5) not strongly supported by STRING (combined score < 800). All eleven pairs show strong coevolution according to DCA and GREMLIN, but five pairs were not predicted to interact by RF2-Lite or AF (fig. S11). A bacterial two-hybrid (B2H) system (*27*) coupled with a quantitative β-Galactosidase assay (*28*) was employed to measure interactions for these eleven pairs (fig. S12).

Despite the strong support by DCA and GREMLIN, the five pairs not predicted to interact by RF2-Lite or AF did not show evidence of interaction using the B2H assay (fig. S11). Among the six pairs supported by RF2-Lite or AF, significant reporter activation indicative of interaction was detected for two: one is between iron-sulfur cluster binding protein lpg2881 (Uniprot: Q5ZRK0) and uncharacterized protein lpg0371 (Uniprot: Q5ZYK1) from *L. pneumophila*; another is between ribosomal silencing factor RsfS (PA4005; Uniprot: Q9HX22) and PhoH-like protein domain-containing protein YbeZ (PA3981; Uniprot: Q9HX38) from *P. aeruginosa* (**Fig. 2A**). For one additional pair, nucleoid-associated protein lmo2703 (Uniprot: Q8Y3X6) and signal recognition particle protein Ffh (Uniprot: Q8Y695) from *L. monocytogenes*, we were unable to assess the interaction experimentally due to false-positive reporter activation when only one protein was expressed (fig. S12). The remaining three pairs failed to generate a positive reporter signal; however, false negative results from B2H assays do not necessarily rule out the existence of a genuine interaction due to possible failures in protein expression and folding of the fusion proteins, and lack of sensitivity of the screen to weak and transient interactions.

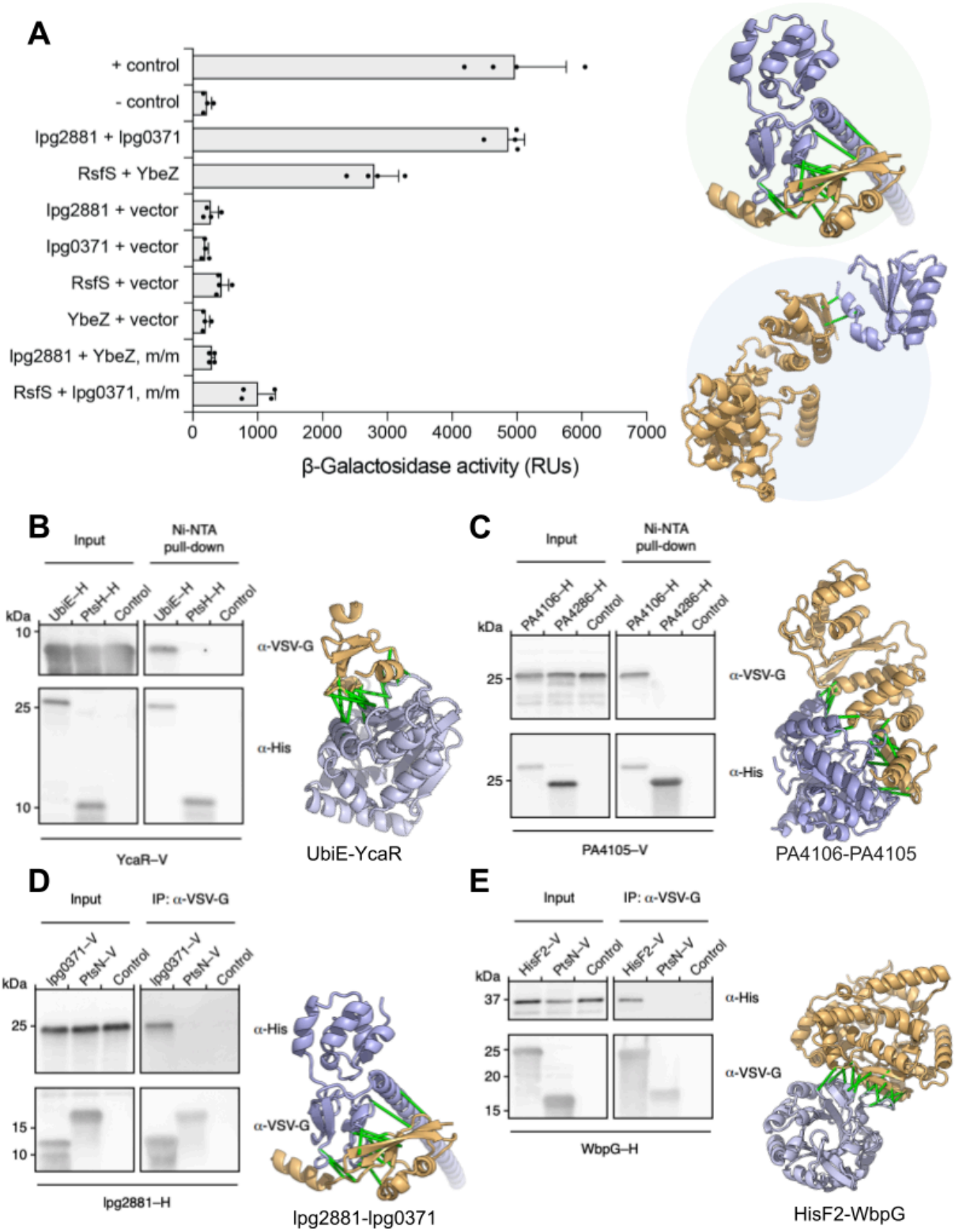


Figure 2: Experimental validation of selected protein-protein interactions
(A) Interactions assessed by B2H that measures β−galactosidase activity resulting from activation of the *lacZ* reporter gene due to the interaction between two tested proteins that are fused to two domains of a transcription activator. *E. coli* expressing T25-zip and T18-zip fusion proteins was used as a positive control (+ control), and *E. coli* harboring empty T25 and T18 plasmids was used as a negative control (-

control). m/m = mix-and-match control. RUs (relative units) = Luminescence / OD_{600} per hour. Error bars indicate ±s.d. ($n = 2$ biological replicates each with 2 technical replicates). Computed models of experimentally validated PPIs (lpg2881 + lpg0371, and RsfS + YbeZ) are shown on the right: top, iron-sulfur cluster binding protein lpg2881 (Q5ZRK0) and uncharacterized protein lpg0371 (Q5ZYK1) from *L. pneumophila*; bottom, ribosomal silencing factor RsfS (Q9HX22) and PhoH-like protein domain-containing protein YbeZ (Q9HX38) from *P. aeruginosa.* **(B-E)** Interactions validated by co-immunoprecipitation (Co-IP)/pull-down. Predicted interacting partners in each PPI pair are heterologously expressed and tagged (–H, hexahistidine; –V, VSV-G epitope). A random bait protein was included as a negative control for each experiment. Control lanes correspond to samples with prey proteins and beads added without any bait proteins. Each positive interaction is supported by two independent Co-IP/pull-down experiments. **(B)** Ubiquinone biosynthesis C-methyltransferase UbiE (P0A887) and protein of unknown function YcaR (P0AAZ7) from *E. coli*. **(C)** Uncharacterized protein PA4106 (Q9HWS2) and a putative transcriptional factor PA4105 (Q9HWS3) from *P. aeruginosa*. **(D)** lpg2881 and lpg0371 from *L. pneumophila*, a pair that is tested positive by B2H as well. **(E)** Putative imidazole glycerol phosphate synthase subunit hisF2 (P72139) and LPS biosynthesis protein WbpG (Q9HZ78) from *P. aeruginosa*. In all the panels, connecting green bars are between representative residue-residue contacts at the interfaces predicted from the summed AlphaFold probability for distance bins below 12Å.

For both PPIs validated by our B2H assays, there are no published data directly supporting functional or physical interactions between the two proteins. However, in both cases, existing evidence indirectly suggests that the interactions could be biologically significant. The pair of proteins from *L. pneumophila* (lpg2881-lpg0371; Q5ZRK0-Q5ZYK1) are homologous to proteins of the Rnf electron transport complex (RnfB with 53% sequence identity and RnfH with 36% sequence identity, respectively). The function of these proteins in *L. pneumophila* is unclear because this species appears to lack the other components of the complex, and one of the proteins, lpg0371, also shares homology with the antitoxin component of the RatAB toxin-antitoxin module. However, in species that encode the complete Rnf complex, RnfB and RnfH directly interact (*29*). The interacting pair from *P. aeruginosa* consists of the ribosomal silencing factor RsfS and the PhoH-like protein domain-containing protein YbeZ. Under nutrient depletion or during stationary phase growth, RsfS binds to ribosomal protein L14, ultimately preventing the association of the 30S and 50S ribosomal subunits and repressing translation (*30*). This facilitates adaptation to low nutrient conditions and promotes survival during the stationary phase. The function of YbeZ is less well-characterized, but it interacts with the RNase YbeY, and both proteins are required for processing and maturation of the 16S rRNA (*31*). Our

finding that YbeZ and RsfS interact suggests that the regulation of ribosome assembly and ribosome subunit processing may be linked in *P. aeruginosa.*

The second validation set consists of six protein pairs (table S8) lacking homologous protein complexes in the PDB, with little support in STRING (only one pair STRING > 600), and distant in the genome (separated by > 100 genes) in half the cases. We focused on proteins consisting primarily of globular domains (percentage of residues from non-globular domains < 20%), as such proteins are more amenable to heterologous expression-based assays. Using co-immunoprecipitation (Co-IP) assays, we detected an interaction between four of the six pairs. These include a pair we had previously validated by B2H, Q5ZRK0-Q5ZYK1, a distally encoded pair from *E. coli* and two proximally encoded pairs from *P. aeruginosa*. *E coli* UbiE catalyzes a carbon-methyl transfer reaction in the biosynthesis of ubiquinone (coenzyme Q) and menaquinone (vitamin K2) (*32*), while YcaR is a small protein detected as differentially expressed in multiple proteomics studies but to which no function has been assigned (*33*, *34*). *P. aeruginosa* PA4105-PA4106 (Q9HWS3-Q9HWS2) are uncharacterized proteins with no clear homologs of known functions based on primary sequence comparisons, but a FoldSeek search (*35*) revealed structural similarity between these proteins and TglI and TglH from *P. syringae pv*. *maculicola* (*P. syringae*) which form a complex that catalyzes the removal of cysteine β-methylene (β-CH2) from TglA-Cys, a step in the biosynthesis of the natural product 3-thiaglutamate (3-thiaGlu) (*36*, *37*). *P. aeruginosa* Q9HZ78-P72139 are an amidotransferase essential in B-band LPS biosynthesis (WbpG, Q9HZ78), and a predicted imidazole glycerol phosphate synthase subunit (HisF2, P72139). It was previously proposed that HisF2, together with HisH2, delivers ammonia to WbpG (*38*), a hypothesis our interaction finding supports. The PtsH-PtsN (Q9HVV2-Q9HVV4) pair with the highest support by STRING (score = 959) failed to generate a positive Co-IP signal (fig. S13); PtsH is a histidine-phosphorylatable phosphocarrier protein encoded adjacent to PtsN, a nitrogen regulatory protein with a phosphotransferase

component, and the interaction between these proteins may be transient, and thus difficult to detect by co-IP.

These experimental data support the *in silico* benchmark in suggesting that the DL methods have greater accuracy than statistical methods in PPI discovery, identifying additional components for well-known biological pathways and accelerate the characterization of proteins of unknown function. In the following sections, we provide an overview of the much larger set of interactions predicted by the DL methods but not yet experimentally validated; to illustrate the insights that can be gained from these data, we provide biological context for selected interaction pairs and higher-order assemblies.

Binary interactions

From the total set of 3613 predicted binary PPIs, 1686 (47%) have homologous complexes in PDB (BLAST e-value ≤ 0.00001 for both proteins), 1862 (52%) are supported by strong functional association according to STRING (total score ≥ 900), and 1284 (36%) are supported by both PDB and STRING; the remaining 1349 (37%, 3613 − (1686 + 1862 − 1284)) to our knowledge, are unknown PPIs. Although such previously unsupported PPIs might contain a higher fraction of false predictions, the high precision on our benchmark sets suggests a significant fraction of the new predictions are likely correct. We identify 166 putative interactions that involve uncharacterized proteins (all Pfam domains are uncharacterized; Supplemental Methods), the majority of these pairs (149) include an interaction partner of known functional domains, and 131 (117 with known partners) not well described previously (STRING combined score < 900 and BLAST e-value to PDB chains > 0.00001).

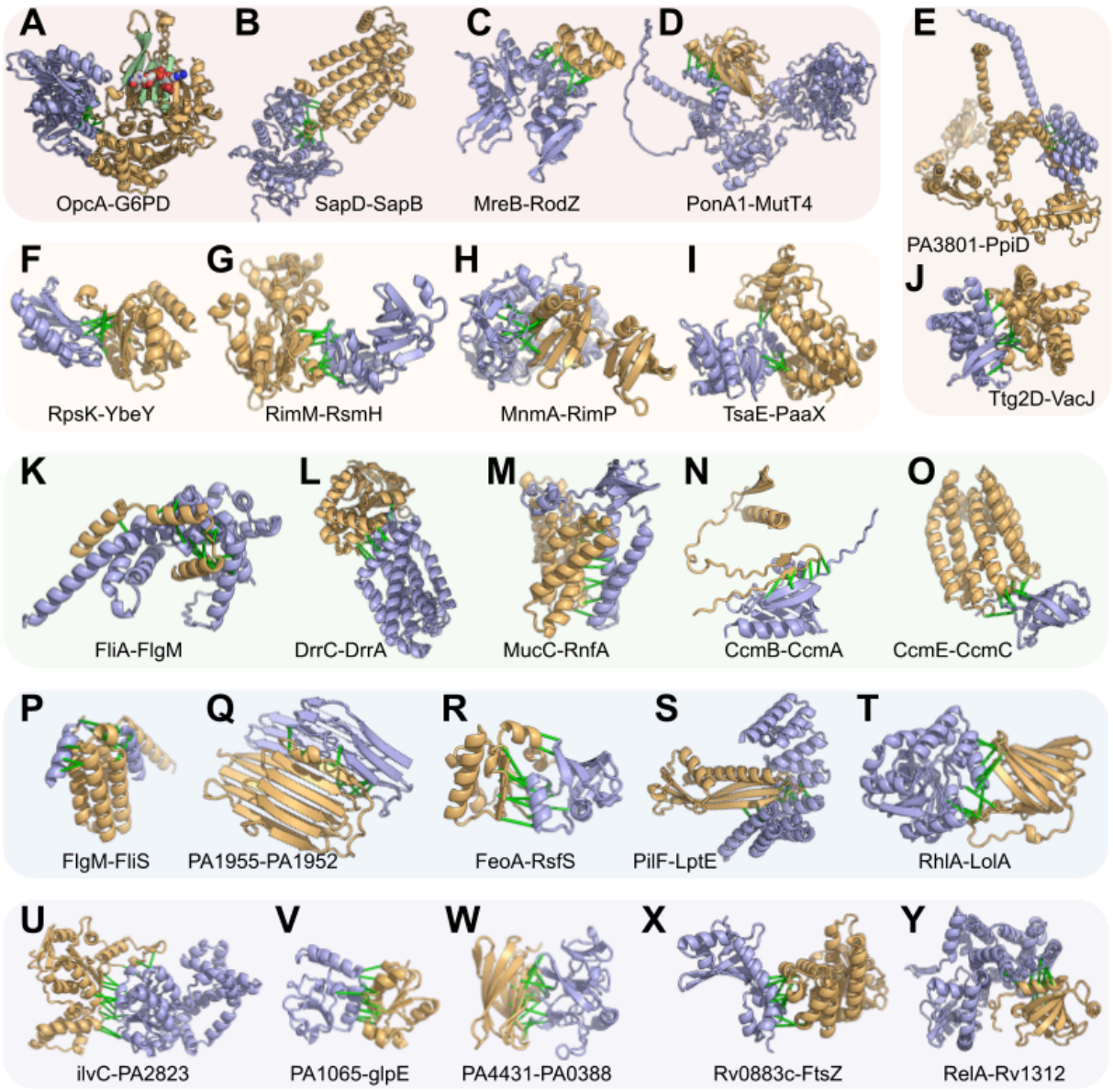


Figure 3: Computed models of binary protein complexes
(A-J) Interactions involving essential genes. **(A)** Interaction with an enzyme where the enzymatic site is highlighted in light green with an NAD moiety. **(B-D)** additional interactions involving essential genes. **(E,J)** Interactions involving transport pathways. **(F-I)** Transcription and translation. **(K-T)** Interactions involving virulence factors. **(U-Y)** Interactions with uncharacterized proteins. In all models, the first protein is in blue, and the second is in gold. Green bars are between representative residue-residue contacts at the interfaces predicted from the summed AlphaFold probability for distance bins below 12Å. Additional information (organisms and UniProt annotations) is in table S9.

1923 of the predicted PPIs include one or more EGs. Examples of predicted interactions among EGs without homologous complexes in the PDB are highlighted in **Fig. 3A-J** and table S9. In some cases, the predicted PPIs support previous findings from the literature. For example, we

predict an interaction between glucose-6-phosphate 1-dehydrogenase 2 (G6PD2) and OxPP (oxidative pentose pathway) cycle protein OpcA (**Fig. 3A**). G6PD2 is an isozyme of G6PD, a member of the pentose phosphate pathway, catalyzing the oxidation of G6P to 6-phosphogluconolactone while converting NADP+ to NADPH and protecting cells from oxidative stress (*39*). OpcA has been implicated as an allosteric activator of G6PD (*40*), but, to our knowledge, the binding site remains unknown. Our predicted interface places OpcA away from the active site of G6PD, consistent with allosteric modulation of activity (fig. S19). We predict an interaction between 30S ribosomal protein S11 (rpsK), a surface-exposed ribosomal protein that forms part of the mRNA binding cleft which recognizes the Shine-Dalgarno sequence (*41*, *42*), and YbeY, a highly conserved endoribonuclease which has been linked to numerous processes such as 16S rRNA maturation, 70S control, and regulation of mRNA (*43*) (**Fig. 3F**). In some bacteria, YbeY plays a key role in virulence and cell stress (*44*). Our predicted structure of S11-YbeY with an interface mediated by S11 β-strands agrees with previous work that identified S11-YbeY interaction by bacteria 2-hybrid, coimmunoprecipitation, and mutational analyses (*45*). The 3D model of the S11-YbeY complex may lend further insights into how YbeY coordinates cleavage of the rRNA precursor during 16S maturation.

256 of the predicted PPIs contain VFs (according to VFDB and Uniprot Keywords) that participate in pathogen colonization, nutrient acquisition, and evasion of host immunity (*46*). Secreted VFs rarely interact with endogenous proteins of a pathogen; consistent with this, we did not detect many PPIs involving VFs, and those we did identify mostly involve structural components of flagella (considered virulence factors in many bacteria (40)) and bacterial secretion systems (**Fig. 3K-T**). We also identified other interactions related to flagella function, for example, between the anti-sigma factor FlgM, a negative regulator of flagellin synthesis, and flagellar secretion chaperone (FliS) (**Fig. 3P**), an interaction supported by a previous experimental study (*47*) but without 3D structure information. Our 3D models, in agreement with

previous observations (*47*), revealed that FlgM can compete with flagellin (FliC, major structural component of the flagella) for the same interface on FliS; FlgM uses its C-terminal helices to interact with FliS, which could prevent its interaction with the flagellar sigma factor FliA. The FliS-FlgM interaction might provide a negative feedback mechanism to control the expression of flagellin: when intracellular flagellin is abundant, it outcompetes FlgM in binding the anti-sigma factor FlgM, and the release of FlgM antagonizes the activity of sigma factor FliA, turning off the expression of late-stage flagellar genes, including flagellin (FliC).

We identify 149 putative interactions (**Fig. 3U-Y**) between uncharacterized proteins (according to Pfam domains) and functionally annotated binding partners such as ketol-acid reductoisomerase IlvC, thiosulfate sulfurtransferase GlpE, ubiquinol-cytochrome c reductase, cell division protein FtsZ, and bifunctional guanosine pentaphosphate [(p)ppGpp] synthase/hydrolase RelA. These predicted interaction partners provide contextual hypotheses about the function of these uncharacterized proteins, 72 of which are essential to pathogen survival, to guide further experimental studies aimed at elucidating their functions.

Multicomponent protein complexes

In many cases, the predicted binary interactions form larger sets, suggesting the formation of higher-order assemblies. For example, in our set of 3613 predicted interactions, we found 206 trimeric protein complexes where each component is predicted to directly interact with the other two. 1545 (40%) of the predicted binary interactions involve proteins that have multiple interacting partners, which allows us to build higher-order protein complexes by concatenating the MSAs of multiple proteins and modeling them together through AF.

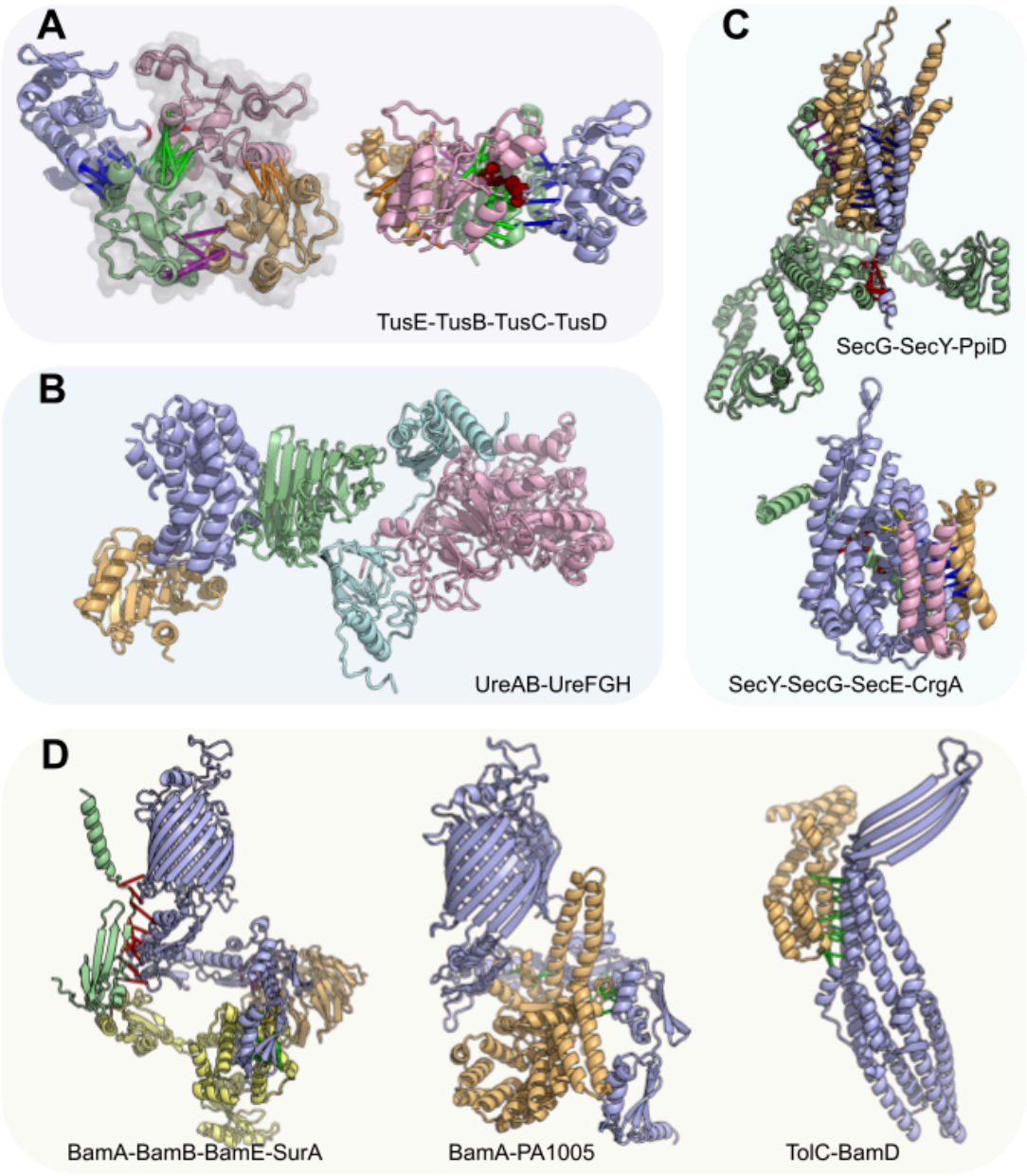


Figure 4: Computed models for multi-component protein complexes
(A) *H. pylori* tRNA 2-thiouridine synthesizing protein complex. Left: a model of the TusE(blue)-TusB(gold)-TusC(green)-TusD(pink) complex overlaid with the TusBCD PDB structure (2D1P, shown in semi-transparent grey). Right: an alternative view of this complex. **(B)** The UreAB-UreFGH complex (colored in cyan, pink, blue, gold, green, respectively) in *H. pylori* assembled through multiple subcomplexes: UreFGH, UreAB, and UreAH. **(C)** Accessory components of the Sec translocon. Top: *P. aeruginosa* SecG(blue)-SecY(gold)-PpiD(green) complex. Bottom: *M. tuberculosis* SecY(blue)-SecG(gold)-SecE(green)-CrgA(pink) complex. **(D)** Accessory components of the *P. aeruginosa and S. typhimurium* outer membrane B-barrel assembly machinery. Left: interaction between SurA (yellow) and Bam proteins (BamA: blue, BamB: gold, BamE: blue). Middle: BamA (blue) and PA1005 (gold), a putative BepA orthologue. Right: interaction between TolC (blue) and BamD (gold). In all schematics, green, red, yellow, and magenta bars connect representative residue-residue contacts at the interfaces predicted from the summed AlphaFold probability for distance bins below 12Å.

tRNA modification and sulfur transfers in the 2-thio modification complex of E. coli

Transfer RNAs (tRNA) play critical roles in protein synthesis and are often decorated with post-transcriptional modifications that contribute to efficient protein synthesis (*48*). Wobble positions are hotspots of such modifications. In glutamate, glutamine, and lysine tRNAs, the wobble uridine is modified to 5-methylaminomethyl-2-thiouridine (mnm^5s^2U) by tRNA 2-thiouridine synthesizing proteins (Tus) (*49*); which include: TusA, TusB, TusC, TusD, TusE, and tRNA-specific 2-thiouridylase (MnmA). Cysteine desulfurase (IscS) is essential for 2-thio modification in *E. coli (49)*. IscS transfers sulfur from cysteine to TusA, which is transferred to TusD of the TusBCD complex via TusE and subsequently to MnmA, which incorporates the sulfur into the tRNA (*49*), (*50*). The structure of the IscS-TusA dimer and sulfur transfer mediating heterohexameric complex, TusBCD, has been co-crystallized (*50*, *51*), but structural details for other components of this system are poorly understood. We predicted the structures of the TusE-MnmA and TusE-TusC complexes (fig. S20) and assembled a model of the full TusBCDE heterotetramer which contextualizes the interaction of TusE with TusBCD (**Fig. 4A**; fig. S21). Our model places TusE close to TusC and TusD, with a confidently predicted TusC-TusE interface (fig. S21E) and is consistent with the hypothesis that Cys108 of TusE accepts sulfur from Cys78 of TusD (*49*, *50*), but also suggests that TusC serves as a scaffold to bring TusD and TusE to close proximity. We also predict the structure of the TusE-MnmA interaction and find that TusE cannot interact with TusBCD and MnmA simultaneously due to overlap in the interfaces with MnmA and TusD (fig. S20A-F).

A two-step nickel transfer in H. pylori urease complex

Urease hydrolyses urea into ammonia and is broadly conserved in bacteria and eukaryotes. In

H. pylori, urease neutralizes gastric acid and facilitates gut colonization (*52*), thus proteins in the urease complex are considered VFs. While most bacterial ureases have three chains (UreA, UreB, and UreC), *H. Pylori* urease has two due to the fusion of UreA and UreB orthologues (*53*). The UreAB(C) system has four accessory proteins: UreE, UreF, UreG, and UreH (*54*). We predict a UreA-UreH interaction and use it to assemble a model of a UreAB-UreFGH pentamer (**Fig. 4B**; fig. S22E). The UreAB(C) and UreFGH substructures have been determined experimentally (*53*), (*55*), and our predictions are consistent with these (fig. S22A-D). During urease maturation, UreFGH receives nickel from UreE, but how this occurs remains poorly understood. Two hypotheses are (a) that UreE transfers nickel to UreFGH complex (*56*) or (b) that upon binding GTP, UreG dissociates from UreFGH, receives nickel from UreE, and subsequently interacts with the inactive UreFH to activate the complex (*55*). Superimposing our UreE-UreG model onto the UreFGH complex shows that UreE clashes with UreF, indicating that UreE cannot directly interact with the UreFGH complex. Therefore, our observation supports the latter hypothesis wherein UreG likely receives nickel separately from UreFH (fig. S23) (*55*).

Interactors of the Sec translocon

The Sec translocon machinery transports proteins across the plasma membrane. The Sec translocon channel is a heterotrimeric complex composed of SecYEG, which operates in tandem with SecA, a RecA-like ATPase that moves peptides through the SecY channel in a process similar to Sec61 translocon in eukaryotes (*57*). We predict interactions between the Sec translocon and peptidyl-prolyl cis/trans isomerase D (ppiD) (**Fig. 4C** top; fig. S24), which has been identified as the most prominent interactor of SecYEG by AP/MS (*58*) and co-immunoprecipitation (*59*). In our model of the SecYG-ppiD trimer, ppiD primarily interacts with SecY through the transmembrane helices while coming close to SecG via a small loop. We

also predict interactions between Sec and CrgA, a transmembrane protein and a component of the divisome (**Fig. 4C** bottom). We find that the CrgA-SecY interface occurs near the lateral gate of SecY (*60*) (fig. S25A), potentially occluding Sec translocation. We hypothesize that during bacterial division, CrgA binds Sec to regulate and recruit translocation machinery near the cell division site, this latter hypothesis is further supported by the predicted interaction between CrgA and SecE (fig. S25) and a less confident prediction of CrgA-SecG interaction that fell slightly below our cutoff.

Outer membrane β-barrel assembly machinery of P. aeruginosa and V. cholera

In Gram-negative bacteria, the β-barrel assembly machinery (BAM) is essential for the folding and insertion of outer membrane β-barrel proteins (*61*, *62*). BAM consists of an outer membrane-spanning β-barrel, BamA, that interacts with four periplasmic lipoproteins, BamB, BamC, BamD, and BamE, to form a five-component complex (computed interactions and structures agree with known experimental data (fig. S26)) (*61–65*). This complex has recently garnered increased attention as a potential therapeutic target, especially since the discovery of Darobactin, a novel antimicrobial compound that binds along the lateral gate of BamA to inhibit outer membrane protein (OMP) biogenesis (*66*, *67*).

The function of BAM is assisted by several other proteins, including the chaperone survival factor A (SurA) and periplasmic chaperone 17-kilodalton protein (Skp). SurA plays an important role in facilitating the recruitment of unfolded OMPs from the periplasm to the BAM complex (*68*). Both our BAM-SurA model and a recently published study using an orthogonal approach to ours (*69*) place SurA in the same position to simultaneously interact with BamA, BamB, and BamE (**Fig. 4D** left). Additionally, we predict an interaction between Skp and SurA (fig. S27), which in addition to their roles in maintaining the solubility of unfolded OMP proteins, may act in

tandem to disassemble oligomeric OMPs that have aggregated (*70*).

We also predict an interaction between BamA and PA1005 (Uniprot: Q9I4W8) (**Fig. 4D** middle), a possible orthologue of β-barrel assembly-enhancing protease (BepA) (fig. S28). *E. coli* BepA is a periplasmic zinc-metallopeptidase with an important role in OM homeostasis and is involved in the degradation of BamA in the absence of SurA (*71*). BepA has been shown to interact with BAM (*72*), and further cross-linking experiments suggest that BepA C-terminal tetratricopeptide repeat (TPR) domain is inserted into the periplasmic region of BamA, below the β-barrel (*71*). Our computed model agrees with the proposed broad interface between BamA and BepA, provides structural details into the BamA-BepA interaction, and also suggests that when BepA is in complex with BamA, BAM is unable to assemble into its active form due to steric clashes between BepA and periplasmic Bam lipoproteins.

TolC is an outer membrane protein that homo-trimerizes to form a large outer membrane export tunnel that interacts with inner membrane translocases (*73*, *74*). The catalytic β-barrel domain of BamA binds substrates along the β-barrel seam during OMP folding, and in this process, the N-terminal of the β-barrel likely swings outward (*75*, *76*). The interaction between BamA and TolC has been recognized as an essential step in the assembly of TolC which occurs in a SurA-independent manner (*77*, *78*). We predict an interaction between BamD and TolC (**Fig. 4D** right), which, when superimposed onto the BAM complex (fig. S29), depicts how the β-sheets of TolC interact with the N-terminal strand of the BamA β-barrel seam. Our computed model shows how TolC could be folded by the BAM complex and suggests that BamD may potentially replace SurA to stabilize or recruit TolC to BAM.

Discussion

RF2-Lite is a new DL network for PPI prediction that is optimized to balance the accuracy and speed necessary for large-scale applications. We integrated RF2-Lite into a pipeline for proteome-wide PPI detection and modeling. We applied this pipeline to an array of human bacterial pathogens, resulting in several thousand predicted PPIs and their 3D structure models. Over one thousand of our predictions are previously unknown, and both our benchmark and experimental validation suggest that a significant fraction of these new PPIs are likely correct and should provide novel biological insights. The 3D structure models of protein complexes generated in our study provide mechanistic details for numerous essential cellular pathways and virulence factors.

Our results demonstrate the potential of computational methods in elucidating the 3D interactome and gaining functional insights for any organism. However, there is still considerable room for improvement in reducing the false positive and false negative rates. As a consequence of the false negatives, our predictions are not comprehensive: the absence of interactions should not be overinterpreted. Although we sought to be conservative and predict only highly confident PPIs, false positives unavoidably exist in our datasets. If each protein on average interacts with only 1 partner, 80% of the predictions in our final dataset are expected to be correct based on our benchmark. Some predicted PPIs, if true, appear to be transient based on the function of proteins, and hence could be difficult to detect with experimental methods like Co-IP (without cross-linking). Based on our limited experimental validation, one should expect that ⅔ of our predicted interactions would give a positive signal in Co-IP experiments. By directly training not only on the PDB but also on larger sets of protein pairs where direct interactions are confidently known to occur (and not occur), as was done by Motmaen et al. 2023 for peptide-MHC complexes (*79*), it should be possible to increase prediction accuracy across a broad spectrum of interaction modalities.

Methods

We have built upon our previously developed multi-step bioinformatics and DL pipelines for identifying pairs of interacting proteins within the proteome of an organism (*7*, *9*) to improve the scalability and accuracy of predictions. The architecture of the new RF2-Lite, which was trained both on monomeric proteins and protein complexes, is outlined in **Fig. 1A,** and the major steps of the bioinformatics pipeline are listed in **Fig. 1C**. Based on our positive and negative controls, PPIs identified by our pipeline have a predicted precision of 95% based on the assumption that each protein directly interact with 5 other proteins. However, if the signal-to-noise ratio is much lower, e.g., the average number of direct interacting partners for each protein is 1, the estimated precision falls to 80%. A detailed description of our methodology is provided in the Supplemental Methods.

Acknowledgments

We thank N.V. Grishin, E. Horvitz, and H. Park for helpful discussions, L. Goldschmidt and A. Guillory for computing resource management, and L. Stewart and L. Stuart for logistical support. Additionally, we are grateful to T.G. Bernhardt, Y.O. Elshenawi, X. Liu, G.V. Mukamolova, K.M. Ottemann, M.L. Reniere, and N.R. Salama for their correspondence and biological expertise.

We acknowledge funding from Bill and Melinda Gates Foundation #OPP1156262 (I.R.H.) and Washington Research Foundation and Translational Research Fund (to M.B.). This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00210147 to M.B.), and J.Z. was supported by CPRIT training grant RP210041. The Defense Threat Reduction Agency grant HDTRA1-21-1-0007

(A.K. and I.A.), Audacious Project at the Institute for Protein Design (A.K.), Spark Therapeutics (I.A.), and National Institute of Allergy and Infectious Diseases (NIAID) Federal Contracts HHSN272201700059C & 75N93022C00036 (I.A.). NIH R01AI145954 (to J.D.M), Defense Advanced Research Projects Agency Biological Technologies Office Program: Harnessing Enzymatic Activity for Lifesaving Remedies (HEALR) under cooperative agreement No. HR0011-21-2-0012 (to J.D.M), and I-2095-20220331 (to Q.C.) from the Welch Foundation. J.D.M and D.B. are Howard Hughes Medical Institute investigators and Q.C. is a Southwestern Medical Foundation endowed scholar.

Author contributions

QC and DB conceived the research and contributed equally; IRH and JZ prepared the sequence alignments used in the screen; MB designed and trained RoseTTAFold2-Light; IRH, JZ, JP, IA, and QC designed the PPI screening procedure; IRH and JZ carried out the screen; IRH, JZ, MB, AK, and QC analyzed and presented computational results; YW, CAT, and BAJ conducted laboratory experiments; YW analyzed and presented experimental results; IRH, YW, TW, DTH, SBP, and JDM provided biological insights on specific examples; IRH, YW, SBP, JM, QC, and DB drafted the manuscript; all authors discussed the results and commented on the manuscript.

Competing Interests

The authors declare they have no competing interests.

Data and materials availability

Structures of highly confident pairs with accompanying metadata are available at the ModelArchive: modelarchive.org/doi/10.5452/ma-bak-evip. Other high-order protein complexes are shared at: https://conglab.swmed.edu/pathogens/. RF2-Lite is available at https://github.com/SNU-CSSB/RF2-Lite. AlphaFold was obtained from https://github.com/deepmind/alphafold on 16 July 2021 (v2.0.0).

References

1. S. V. Rajagopala, P. Sikorski, A. Kumar, R. Mosca, J. Vlasblom, R. Arnold, J. Franca-Koh, S. B. Pakala, S. Phanse, A. Ceol, R. Häuser, G. Siszler, S. Wuchty, A. Emili, M. Babu, P. Aloy, R. Pieper, P. Uetz, The binary protein-protein interaction landscape of Escherichia coli. *Nat. Biotechnol.* **32**, 285–290 (2014).

2. P. Uetz, L. Giot, G. Cagney, T. A. Mansfield, R. S. Judson, J. R. Knight, D. Lockshon, V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin, D. Conover, T. Kalbfleisch, G. Vijayadamodar, M. Yang, M. Johnston, S. Fields, J. M. Rothberg, A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*. **403**, 623–627 (2000).

3. G. Butland, J. M. Peregrín-Alvarez, J. Li, W. Yang, X. Yang, V. Canadien, A. Starostine, D. Richards, B. Beattie, N. Krogan, M. Davey, J. Parkinson, J. Greenblatt, A. Emili, Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*. **433**, 531–537 (2005).

4. A. M. Edwards, B. Kus, R. Jansen, D. Greenbaum, J. Greenblatt, M. Gerstein, Bridging structural biology and genomics: assessing protein interaction data with known complexes. *Trends Genet.* **18**, 529–536 (2002).

5. J. P. Mackay, M. Sunde, J. A. Lowry, M. Crossley, J. M. Matthews, Protein interactions: is seeing believing? *Trends Biochem. Sci.* **32**, 530–531 (2007).

6. S. Ovchinnikov, H. Kamisetty, D. Baker, Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife*. **3**, e02030 (2014).

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

8. A. G. Green, H. Elhabashy, K. P. Brock, R. Maddamsetti, O. Kohlbacher, D. S. Marks, Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nat. Commun.* **12**, 1396 (2021).

9. I. R. Humphreys, J. Pei, M. Baek, A. Krishnakumar, I. Anishchenko, S. Ovchinnikov, J. Zhang, T. J. Ness, S. Banjade, S. R. Bagde, V. G. Stancheva, X.-H. Li, K. Liu, Z. Zheng, D. J. Barrero, U. Roy, J. Kuper, I. S. Fernández, B. Szakal, D. Branzei, J. Rizo, C. Kisker, E. C. Greene, S. Biggins, S. Keeney, E. A. Miller, J. C. Fromme, T. L. Hendrickson, Q. Cong, D. Baker, Computed structures of core eukaryotic protein complexes. *Science*. **374**, eabm4805 (2021).

10. J. Pei, J. Zhang, Q. Cong, Human mitochondrial protein complexes revealed by large-scale coevolution analysis and deep learning-based structure modeling, , doi:10.1101/2021.09.14.460228.

11. J. Zhang, J. Pei, J. Durham, T. Bos, Q. Cong, Computed cancer interactome explains the effects of somatic mutations in cancers. *Protein Sci.* **31**, e4479 (2022).

12. R. Zhang, H.-Y. Ou, C.-T. Zhang, DEG: a database of essential genes. *Nucleic Acids Res.* **32**, D271–2 (2004).

13. L. Chen, J. Yang, J. Yu, Z. Yao, L. Sun, Y. Shen, Q. Jin, VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* **33**, D325–8 (2005).

14. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. DeGiovanni, J. H. Pereira, A. V. Rodrigues, A. A. van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, R. J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network. *Science*. **373**, 871–876 (2021).

15. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature*. **596**, 583–589 (2021).

16. M. Gao, D. Nakajima An, J. M. Parks, J. Skolnick, AF2Complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* **13**, 1744 (2022).

17. D. F. Burke, P. Bryant, I. Barrio-Hernandez, D. Memon, G. Pozzati, A. Shenoy, W. Zhu, A. S. Dunham, P. Albanese, A. Keller, R. A. Scheltema, J. E. Bruce, A. Leitner, P. Kundrotas, P. Beltrao, A. Elofsson, Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* (2023), doi:10.1038/s41594-022-00910-8.

18. R. Evans, M. O'Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, O. Ronneberger, S. Bodenstein, M. Zielinski, A. Bridgland, A. Potapenko, A. Cowie, K. Tunyasuvunakool, R. Jain, E. Clancy, P. Kohli, J. Jumper, D. Hassabis, Protein complex prediction with AlphaFold-Multimer, , doi:10.1101/2021.10.04.463034.

19. M. Baek, I. Anishchenko, I. R. Humphreys, Q. Cong, D. Baker, F. DiMaio, Efficient and accurate prediction of protein structure using RoseTTAFold2. *bioRxiv* (2023), p. 2023.05.24.542179.

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

21. D. P. Wall, H. B. Fraser, A. E. Hirsh, Detecting putative orthologs. *Bioinformatics*. **19** (2003), pp. 1710–1711.

22. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

23. F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, D. G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

24. F. Chen, A. J. Mackey, C. J. Stoeckert Jr, D. S. Roos, OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.* **34**, D363–8 (2006).

25. D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, C. von Mering, STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–52 (2015).

26. H. Kamisetty, S. Ovchinnikov, D. Baker, Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 15674–15679 (2013).

27. G. Karimova, J. Pidoux, A. Ullmann, D. Ladant, A bacterial two-hybrid system based on a reconstituted signal transduction pathway. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 5752–5756 (1998).

28. G. Karimova, E. Gauliard, M. Davi, S. P. Ouellette, D. Ladant, Protein-Protein Interaction: Bacterial Two-Hybrid. *Methods Mol. Biol.* **1615**, 159–176 (2017).

29. L. Zhang, O. Einsle, Architecture of the NADH:ferredoxin oxidoreductase RNF that drives Biological Nitrogen Fixation. *bioRxiv* (2022), p. 2022.07.08.499327.

30. R. Häuser, M. Pech, J. Kijek, H. Yamamoto, B. Titz, F. Naeve, A. Tovchigrechko, K. Yamamoto, W. Szaflarski, N. Takeuchi, T. Stellberger, M. E. Diefenbacher, K. H. Nierhaus, P. Uetz, RsfA (YbeB)

proteins are conserved ribosomal silencing factors. *PLoS Genet.* **8**, e1002815 (2012).

31. Y. Xia, Y. Weng, C. Xu, D. Wang, X. Pan, Z. Tian, B. Xia, H. Li, R. Chen, C. Liu, Y. Jin, F. Bai, Z. Cheng, O. P. Kuipers, W. Wu, Endoribonuclease YbeY Is Essential for RNA Processing and Virulence in Pseudomonas aeruginosa. *MBio*. **11** (2020), doi:10.1128/mBio.00659-20.

32. P. T. Lee, A. Y. Hsu, H. T. Ha, C. F. Clarke, A C-methyltransferase involved in both ubiquinone and menaquinone biosynthesis: isolation and identification of the Escherichia coli ubiE gene. *J. Bacteriol.* **179**, 1748–1754 (1997).

33. M. Božik, P. Cejnar, M. Šašková, P. Nový, P. Maršík, P. Klouček, Stress response of Escherichia coli to essential oil components - insights on low-molecular-weight proteins from MALDI-TOF. *Sci. Rep.* **8**, 13042 (2018).

34. M. Sultonova, B. Blackmore, R. Du, O. Philips, J. A. Paulo, J. P. Murphy, Integrated changes in thermal stability and proteome abundance during altered nutrient states in Escherichia coli and human cells. *Proteomics*. **22**, e2100254 (2022).

35. M. van Kempen, S. S. Kim, C. Tumescheit, M. Mirdita, J. Lee, C. L. M. Gilchrist, J. Söding, M. Steinegger, Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* (2023), doi:10.1038/s41587-023-01773-0.

36. C. P. Ting, M. A. Funk, S. L. Halaby, Z. Zhang, T. Gonen, W. A. van der Donk, Use of a scaffold peptide in the biosynthesis of amino acid-derived natural products. *Science*. **365**, 280–284 (2019).

37. Y. Zheng, X. Xu, X. Fu, X. Zhou, C. Dou, Y. Yu, W. Yan, J. Yang, M. Xiao, W. A. van der Donk, X. Zhu, W. Cheng, Structures of the holoenzyme TglHI required for 3-thiaglutamate biosynthesis. *Structure*. **31**, 1220–1232.e5 (2023).

38. L. Feng, S. N. Senchenkova, J. Tao, A. S. Shashkov, B. Liu, S. D. Shevelev, P. R. Reeves, J. Xu, Y. A. Knirel, L. Wang, Structural and genetic characterization of enterohemorrhagic Escherichia coli O145 O antigen and development of an O145 serogroup-specific PCR assay. *J. Bacteriol.* **187**, 758–764 (2005).

39. J. M. Sandoval, F. A. Arenas, C. C. Vásquez, Glucose-6-phosphate dehydrogenase protects Escherichia coli from tellurite-mediated oxidative stress. *PLoS One*. **6**, e25573 (2011).

40. K. D. Hagen, J. C. Meeks, The unique cyanobacterial protein OpcA is an allosteric effector of glucose-6-phosphate dehydrogenase in Nostoc punctiforme ATCC 29133. *J. Biol. Chem.* **276**, 11477–11486 (2001).

41. T. Kaminishi, D. N. Wilson, C. Takemoto, J. M. Harms, M. Kawazoe, F. Schluenzen, K. Hanawa-Suetsugu, M. Shirouzu, P. Fucini, S. Yokoyama, A snapshot of the 30S ribosomal subunit capturing mRNA via the Shine-Dalgarno interaction. *Structure*. **15**, 289–297 (2007).

42. G. Z. Yusupova, M. M. Yusupov, J. H. Cate, H. F. Noller, The path of messenger RNA through the ribosome. *Cell*. **106**, 233–241 (2001).

43. A. I. Jacob, C. Köhrer, B. W. Davies, U. L. RajBhandary, G. C. Walker, Conserved bacterial RNase YbeY plays key roles in 70S ribosome quality control and 16S rRNA maturation. *Mol. Cell*. **49**, 427–438 (2013).

44. M. Vercruysse, C. Köhrer, B. W. Davies, M. F. F. Arnold, J. J. Mekalanos, U. L. RajBhandary, G. C. Walker, The highly conserved bacterial RNase YbeY is essential in Vibrio cholerae, playing a critical role in virulence, stress regulation, and RNA processing. *PLoS Pathog.* **10**, e1004175 (2014).

45. M. Vercruysse, C. Köhrer, Y. Shen, S. Proulx, A. Ghosal, B. W. Davies, U. L. RajBhandary, G. C. Walker, Identification of YbeY-Protein Interactions Involved in 16S rRNA Maturation and Stress

Regulation in Escherichia coli. *MBio*. **7** (2016), doi:10.1128/mBio.01785-16.

46. B. B. Finlay, S. Falkow, Common themes in microbial pathogenicity revisited. *Microbiol. Mol. Biol. Rev.* **61**, 136–169 (1997).

47. C. Barembruch, R. Hengge, Cellular levels and activity of the flagellar sigma factor FliA of Escherichia coli are controlled by FlgM-modulated proteolysis. *Mol. Microbiol.* **65**, 76–89 (2007).

48. T. Suzuki, The expanding world of tRNA modifications and their disease relevance. *Nat. Rev. Mol. Cell Biol.* **22**, 375–392 (2021).

49. Y. Ikeuchi, N. Shigi, J.-I. Kato, A. Nishimura, T. Suzuki, Mechanistic insights into sulfur relay by multiple sulfur mediators involved in thiouridine biosynthesis at tRNA wobble positions. *Mol. Cell*. **21**, 97–108 (2006).

50. T. Numata, S. Fukai, Y. Ikeuchi, T. Suzuki, O. Nureki, Structural basis for sulfur relay to RNA mediated by heterohexameric TusBCD complex. *Structure*. **14**, 357–366 (2006).

51. R. Shi, A. Proteau, M. Villarroya, I. Moukadiri, L. Zhang, J.-F. Trempe, A. Matte, M. E. Armengod, M. Cygler, Structural basis for Fe-S cluster assembly and tRNA thiolation mediated by IscS protein-protein interactions. *PLoS Biol.* **8**, e1000354 (2010).

52. K. A. Eaton, C. L. Brooks, D. R. Morgan, S. Krakowka, Essential role of urease in pathogenesis of gastritis induced by Helicobacter pylori in gnotobiotic piglets. *Infect. Immun.* **59**, 2470–2475 (1991).

53. N. C. Ha, S. T. Oh, J. Y. Sung, K. A. Cha, M. H. Lee, B. H. Oh, Supramolecular assembly and acid resistance of Helicobacter pylori urease. *Nat. Struct. Biol.* **8**, 505–509 (2001).

54. E. L. Carter, N. Flugga, J. L. Boer, S. B. Mulrooney, R. P. Hausinger, Interplay of metal ions and urease. *Metallomics*. **1**, 207–221 (2009).

55. Y. H. Fong, H. C. Wong, M. H. Yuen, P. H. Lau, Y. W. Chen, K.-B. Wong, Structure of UreG/UreF/UreH complex reveals how urease accessory proteins facilitate maturation of Helicobacter pylori urease. *PLoS Biol.* **11**, e1001678 (2013).

56. M. A. Farrugia, L. Macomber, R. P. Hausinger, Biosynthesis of the urease metallocenter. *J. Biol. Chem.* **288**, 13178–13185 (2013).

57. T. A. Rapoport, Protein translocation across the eukaryotic endoplasmic reticulum and bacterial plasma membranes. *Nature*. **450**, 663–669 (2007).

58. B. Jauss, N.-A. Petriman, F. Drepper, L. Franz, I. Sachelaru, T. Welte, R. Steinberg, B. Warscheid, H.-G. Koch, Noncompetitive binding of PpiD and YidC to the SecYEG translocon expands the global view on the SecYEG interactome in. *J. Biol. Chem.* **294**, 19167–19183 (2019).

59. H. Götzke, I. Palombo, C. Muheim, E. Perrody, P. Genevaux, R. Kudva, M. Müller, D. O. Daley, YfgM is an ancillary subunit of the SecYEG translocon in Escherichia coli. *J. Biol. Chem.* **289**, 19089–19097 (2014).

60. B. Van den Berg, W. M. Clemons Jr, I. Collinson, Y. Modis, E. Hartmann, S. C. Harrison, T. A. Rapoport, X-ray structure of a protein-conducting channel. *Nature*. **427**, 36–44 (2004).

61. R. Voulhoux, M. P. Bos, J. Geurtsen, M. Mols, J. Tommassen, Role of a highly conserved bacterial protein in outer membrane protein assembly. *Science*. **299**, 262–265 (2003).

62. T. Wu, J. Malinverni, N. Ruiz, S. Kim, T. J. Silhavy, D. Kahne, Identification of a multicomponent complex required for outer membrane biogenesis in Escherichia coli. *Cell*. **121**, 235–245 (2005).

63. N. Noinaj, A. J. Kuszak, J. C. Gumbart, P. Lukacik, H. Chang, N. C. Easley, T. Lithgow, S. K. Buchanan, Structural insight into the biogenesis of β-barrel membrane proteins. *Nature*. **501**, 385–390 (2013).

64. L. Han, J. Zheng, Y. Wang, X. Yang, Y. Liu, C. Sun, B. Cao, H. Zhou, D. Ni, J. Lou, Y. Zhao, Y. Huang, Structure of the BAM complex and its implications for biogenesis of outer-membrane proteins. *Nat. Struct. Mol. Biol.* **23**, 192–196 (2016).

65. Y. Gu, H. Li, H. Dong, Y. Zeng, Z. Zhang, N. G. Paterson, P. J. Stansfeld, Z. Wang, Y. Zhang, W. Wang, C. Dong, Structural basis of outer membrane protein insertion by the BAM complex. *Nature*. **531**, 64–69 (2016).

66. Y. Imai, K. J. Meyer, A. Iinishi, Q. Favre-Godal, R. Green, S. Manuse, M. Caboni, M. Mori, S. Niles, M. Ghiglieri, C. Honrao, X. Ma, J. J. Guo, A. Makriyannis, L. Linares-Otoya, N. Böhringer, Z. G. Wuisan, H. Kaur, R. Wu, A. Mateus, A. Typas, M. M. Savitski, J. L. Espinoza, A. O'Rourke, K. E. Nelson, S. Hiller, N. Noinaj, T. F. Schäberle, A. D'Onofrio, K. Lewis, A new antibiotic selectively kills Gram-negative pathogens. *Nature*. **576**, 459–464 (2019).

67. H. Kaur, R. P. Jakob, J. K. Marzinek, R. Green, Y. Imai, J. R. Bolla, E. Agustoni, C. V. Robinson, P. J. Bond, K. Lewis, T. Maier, S. Hiller, The antibiotic darobactin mimics a β-strand to inhibit outer membrane insertase. *Nature*. **593**, 125–129 (2021).

68. J. G. Sklar, T. Wu, D. Kahne, T. J. Silhavy, Defining the roles of the periplasmic chaperones SurA, Skp, and DegP in Escherichia coli. *Genes Dev.* **21**, 2473–2484 (2007).

69. B. Schiffrin, J. M. Machin, T. K. Karamanos, A. Zhuravleva, D. J. Brockwell, S. E. Radford, A. N. Calabrese, Dynamic interplay between the periplasmic chaperone SurA and the BAM complex in outer membrane protein folding. *Communications Biology*. **5**, 1–15 (2022).

70. N. Chamachi, A. Hartmann, M. Q. Ma, A. Svirina, G. Krainer, M. Schlierf, Chaperones Skp and SurA dynamically expand unfolded OmpX and synergistically disassemble oligomeric aggregates. *Proc. Natl. Acad. Sci. U. S. A.* **119** (2022), doi:10.1073/pnas.2118919119.

71. Y. Daimon, C. Iwama-Masui, Y. Tanaka, T. Shiota, T. Suzuki, R. Miyazaki, H. Sakurada, T. Lithgow, N. Dohmae, H. Mori, T. Tsukazaki, S.-I. Narita, Y. Akiyama, The TPR domain of BepA is required for productive interaction with substrate proteins and the β-barrel assembly machinery complex. *Mol. Microbiol.* **106**, 760–776 (2017).

72. S.-I. Narita, C. Masui, T. Suzuki, N. Dohmae, Y. Akiyama, Protease homolog BepA (YfgC) promotes assembly and degradation of β-barrel membrane proteins in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **110**, E3612–21 (2013).

73. T. Thanabalu, E. Koronakis, C. Hughes, V. Koronakis, Substrate-induced assembly of a contiguous channel for protein export from E.coli: reversible bridging of an inner-membrane translocase to an outer membrane exit pore. *EMBO J.* **17**, 6487–6496 (1998).

74. V. Koronakis, A. Sharff, E. Koronakis, B. Luisi, C. Hughes, Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature*. **405**, 914–919 (2000).

75. D. Tomasek, S. Rawson, J. Lee, J. S. Wzorek, S. C. Harrison, Z. Li, D. Kahne, Structure of a nascent membrane protein as it folds on the BAM complex. *Nature*. **583**, 473–478 (2020).

76. M. T. Doyle, J. R. Jimah, T. Dowdy, S. I. Ohlemacher, M. Larion, J. E. Hinshaw, H. D. Bernstein, Cryo-EM structures reveal multiple stages of bacterial outer membrane protein folding. *Cell*. **185**, 1143–1156.e13 (2022).

77. J. Werner, R. Misra, YaeT (Omp85) affects the assembly of lipid-dependent and lipid-independent outer membrane proteins of Escherichia coli. *Mol. Microbiol.* **57**, 1450–1459 (2005).

78. D. Bennion, E. S. Charlson, E. Coon, R. Misra, Dissection of β-barrel outer membrane protein assembly pathways through characterizing BamA POTRA 1 mutants of Escherichia coli. *Mol. Microbiol.* **77**, 1153–1171 (2010).

79. A. Motmaen, J. Dauparas, M. Baek, M. H. Abedi, D. Baker, P. Bradley, Peptide-binding specificity prediction using fine-tuned protein structure prediction networks. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2216697120 (2023).

Supplemental Figures

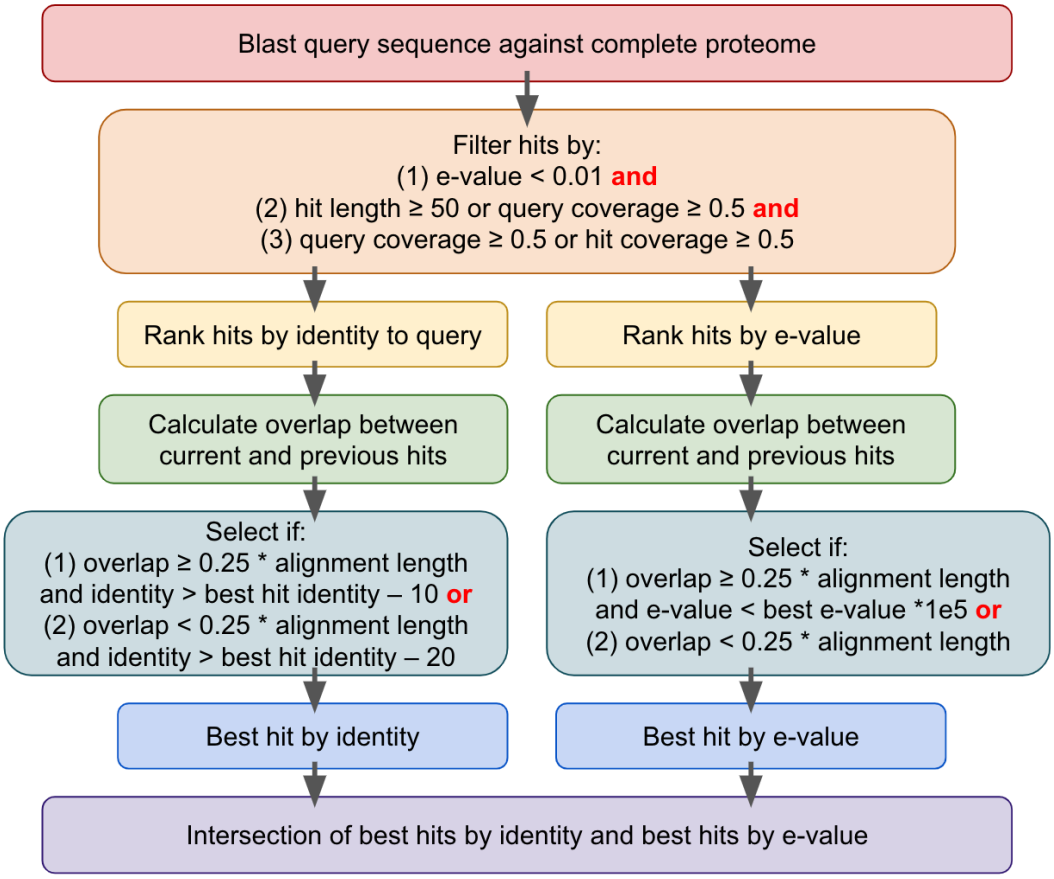


Figure S1: Flowchart of reciprocal best hits filtering criterion for complete proteomes

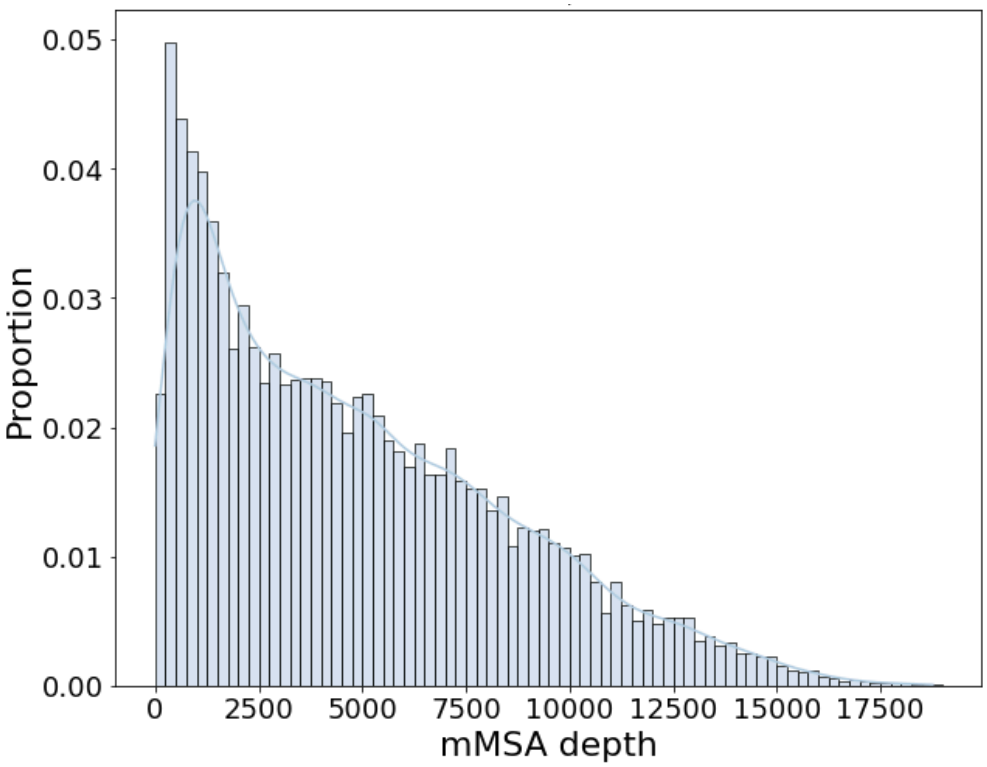


Figure S2: Representative distribution of monomeric MSA depth

15,000 proteins across the 19 proteomes were randomly selected to get a representative distribution of monomeric MSA (mMSA) depth. Alignments were filtered with hhfilter at 90% sequence identity and 75% coverage to remove redundancy.

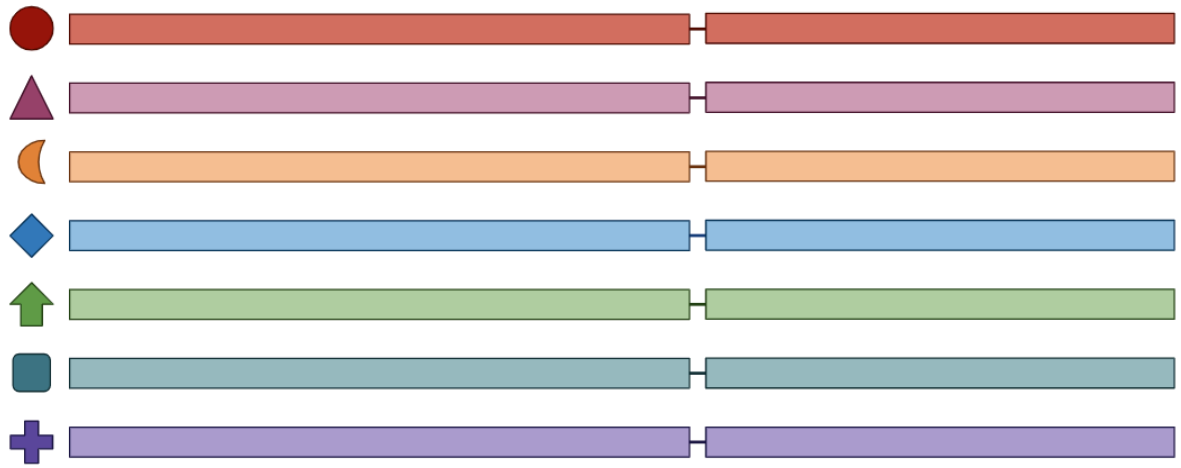


Figure S3: Paired multiple sequence alignment schematic

Proteins from the same proteome are colored in the same color. The dash the separation of two proteins and residue indexing gap that is implemented in RoseTTAFold and AlphaFold. However, there is no explicit gap implemented in the pMSAs between two proteins.

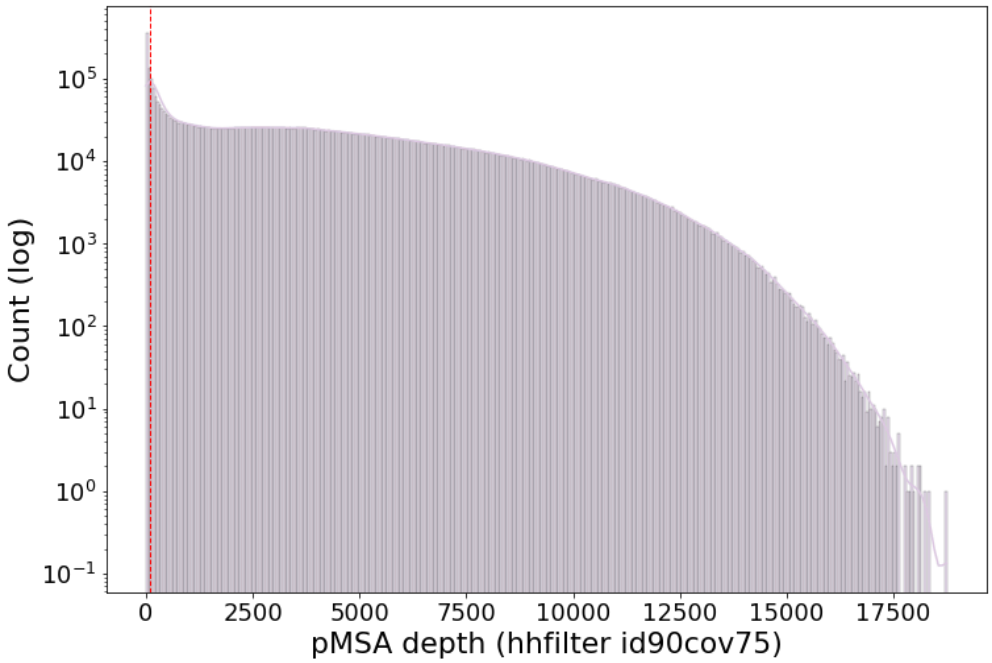


Figure S4: Distribution of pMSA depth

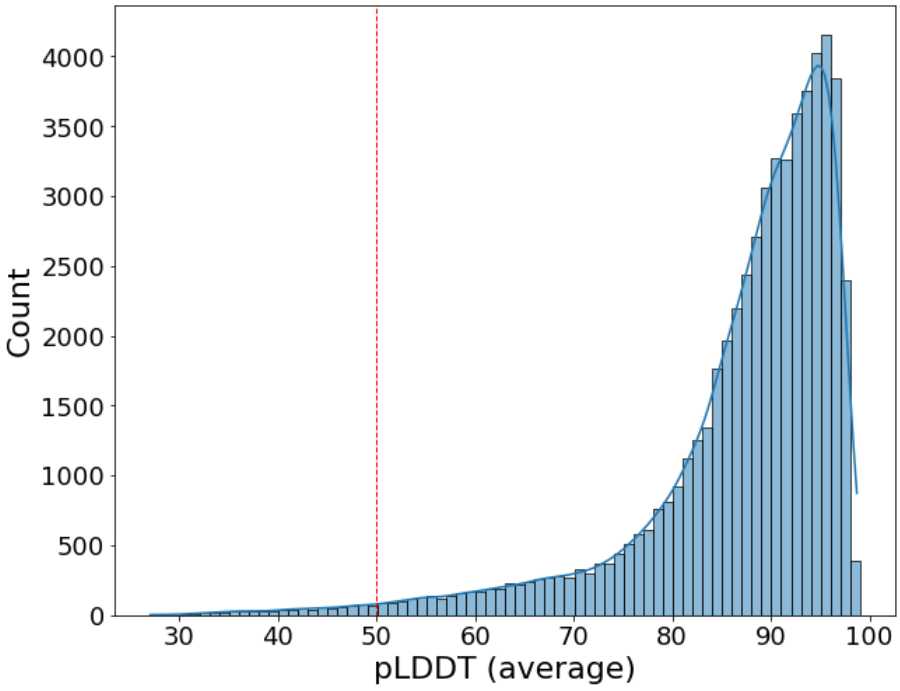


Figure S5: Distribution of AlphaFold monomer model average pLDDT

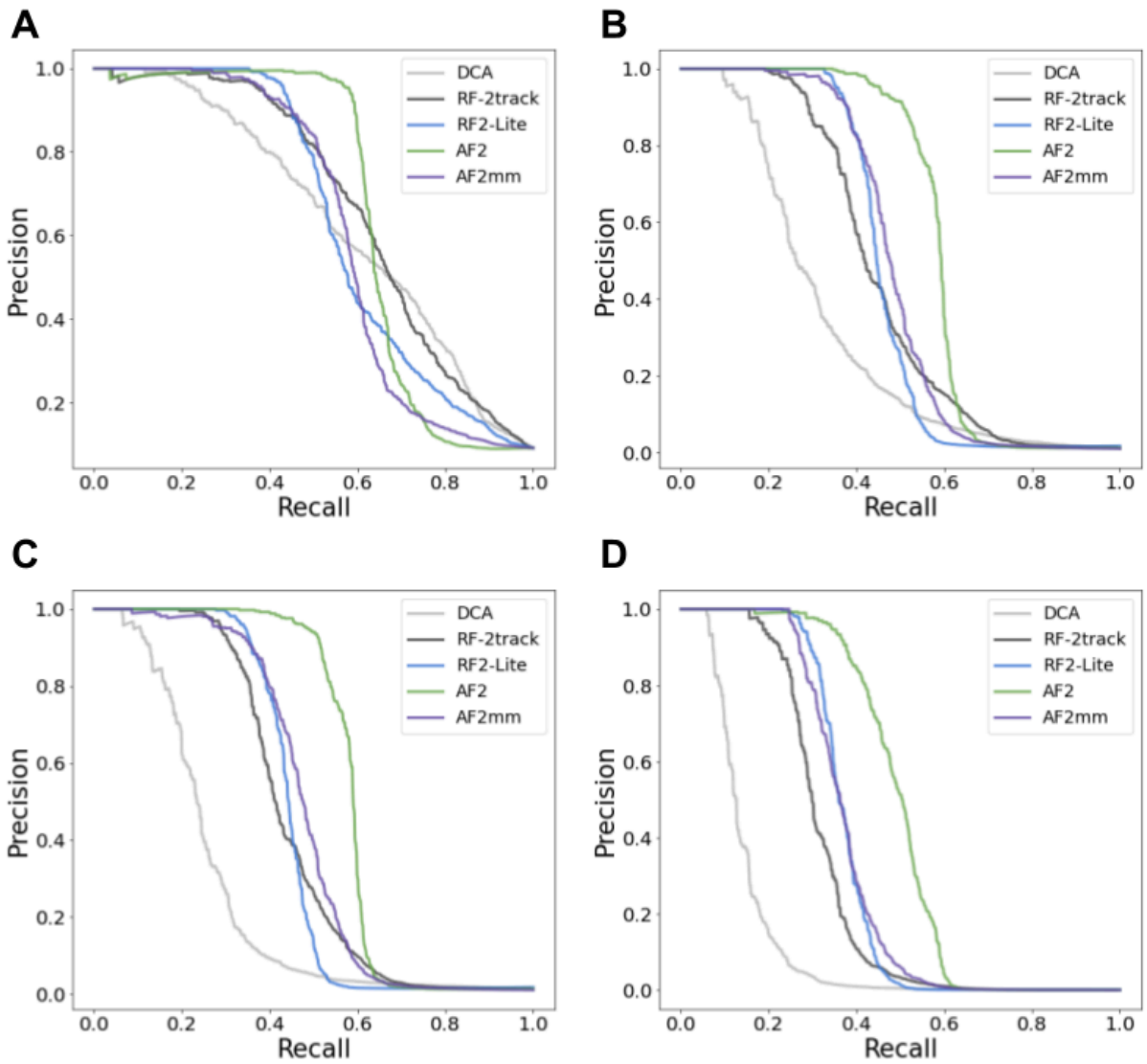


Figure S6: PPI screening methodology performance

Precision vs recall of different PPI screening tools. (**A**) Baseline 1000 positive:10,000 negative pair dataset. (**B**) 1000 positives with 100,000 negatives that were generated by randomly sampling 10 data points with 0.1 standard deviation around each of the 10,000 baseline negative examples. (**C**) 1000 positives with 100,000 negatives that were generated by randomly sampling 10 data points with 0.05 standard deviation around each of the 10,000 baseline negative examples. (**D**) 1000 positives with 1,000,000 negatives that were generated by randomly sampling 100 data points with 0.05 standard deviation around each of the 10,000 baseline negative examples.

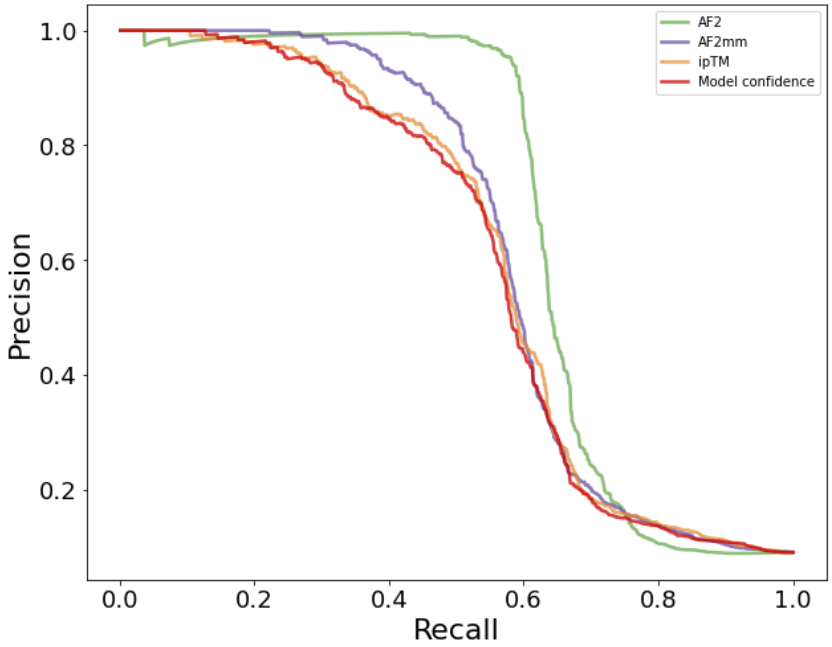


Figure S7: AlphaFold-multimer distance vs ipTM for PPI identification

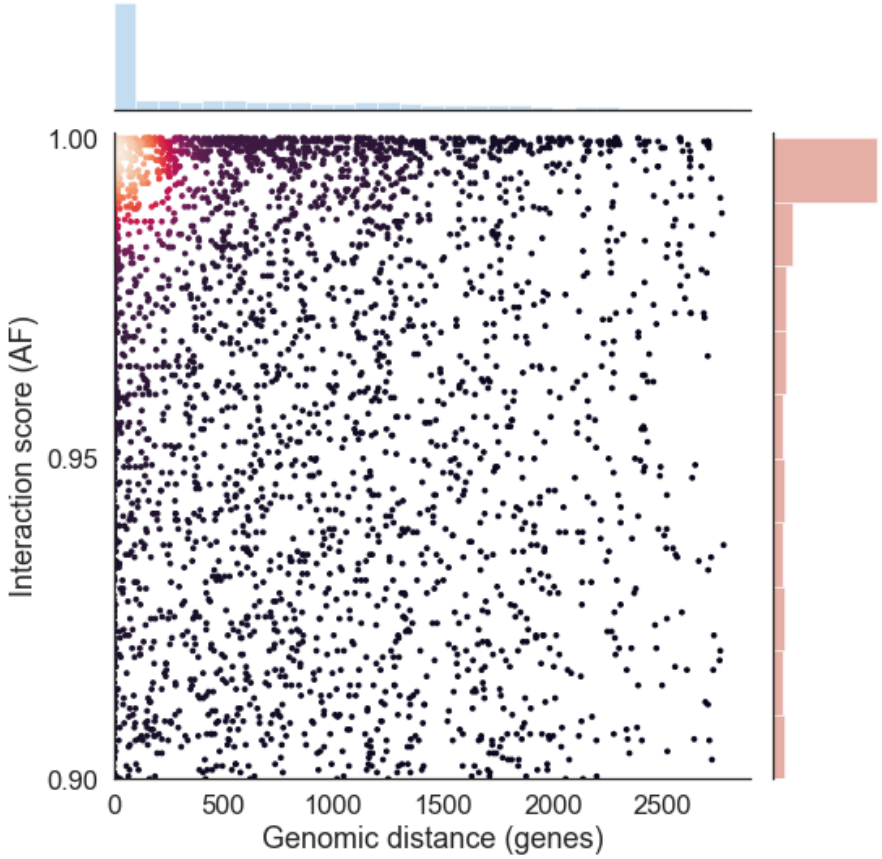


Figure S8: Predicted interactions by genomic distance

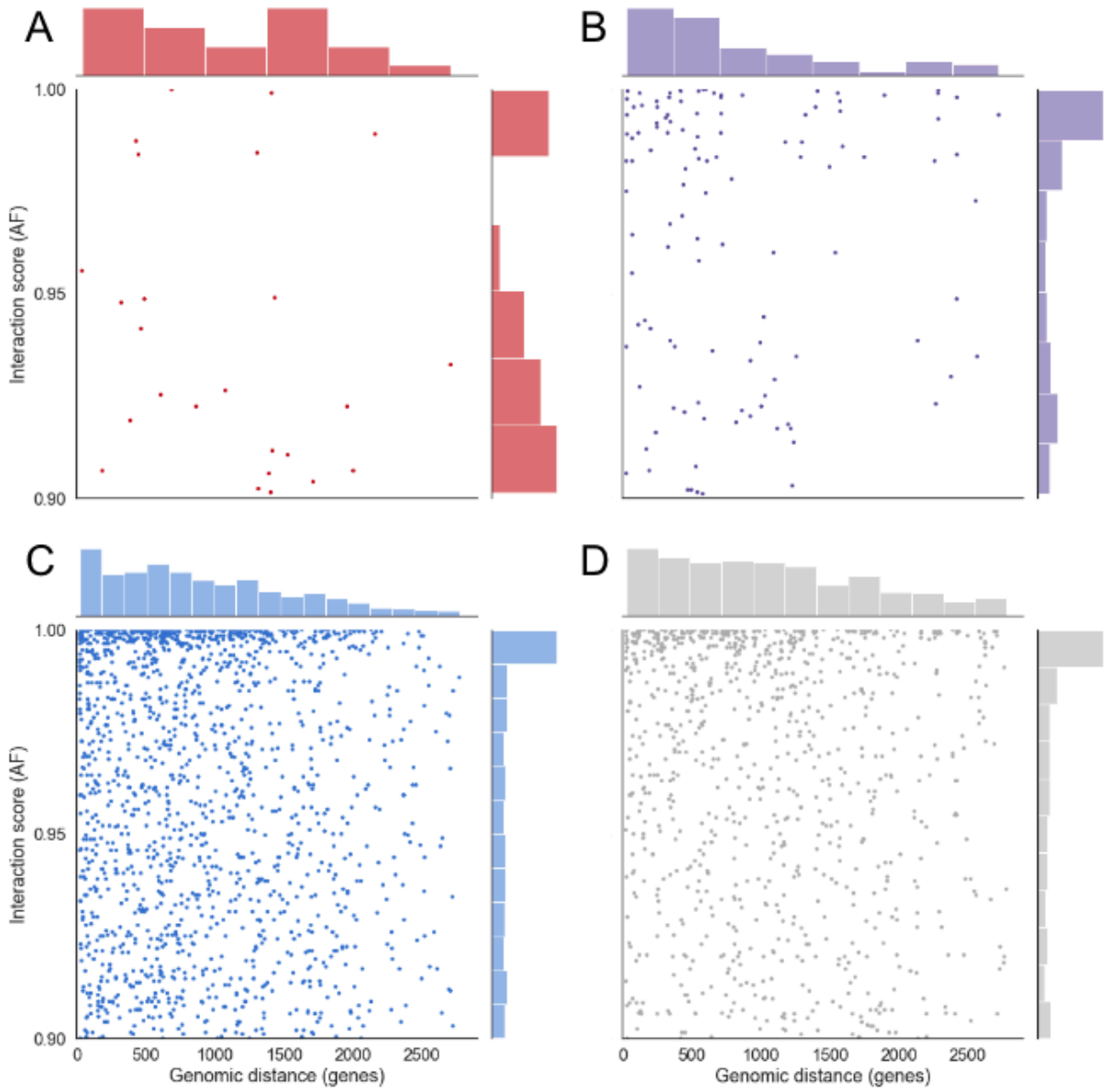


Figure S9: Predicted unique interactions by genomic distance

Each primary plot contains the PPI interaction score from the AF screen by the number of genes between the two proteins. Each point represents a pair of proteins. On the side axes of each primary plot, are histograms of these data to better depict the density. (**A**) Interaction between an EG and VF, (**B**) interactions between one VF and non-essential gene or between two VFs, (**C**) interactions between one EG and non-essential gene or between two EGs, and (**D**) interactions between two non-essential, non-virulence factors.

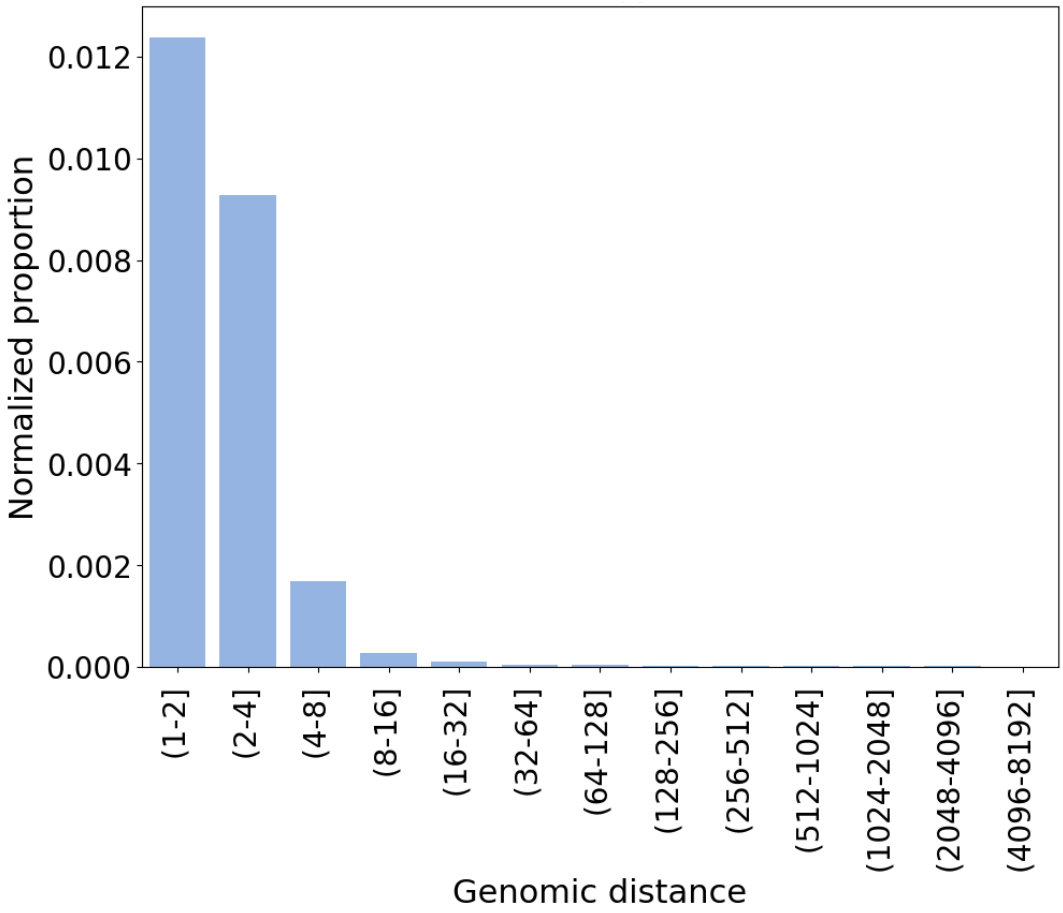
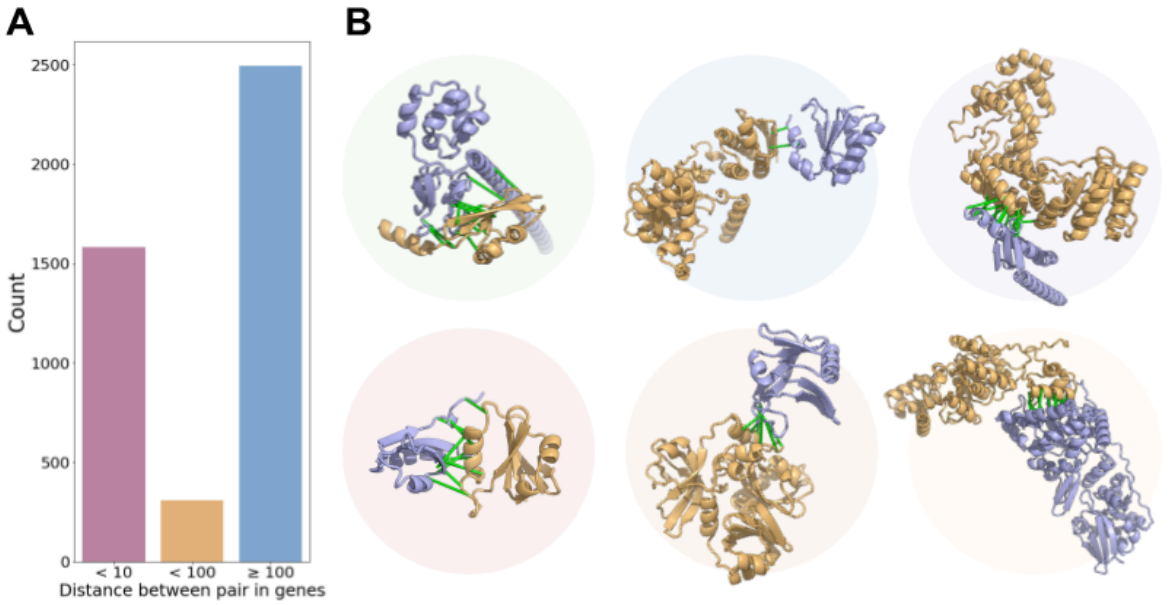


Figure S10: Normalized fraction of predicted interactions by genomic distance

Fraction of predicted interactions as a function of genomic distance between protein pairs. Genomic distance is calculated by the minimal number of genes between the coding genes of the two proteins. The proportion of interactions is the number of interactions confidently predicted in our study (AF score > 0.90) normalized to the number of pairs in the 19 pathogen proteomes that fall into each genomic distance bin.



C

	Distance	STRING	DCA	GREMLIN	RF2-Lite	AF	B2H
Q5ZRK0_Q5ZYK1	475	573	0.23	1	1.0	1.0	0.98
Q9HX22_Q9HX38	23	736	0.77	1	0.77	0.61	0.56
Q8Y3X6_Q8Y695	899	445	0.17	0.81	0.81	0.88	0.51
P60415_Q928N1	347	522	0.09	0.87	0.92	0.99	0.06
Q9HXJ0_Q9I5H0	2508	445	0.55	0.98	0.79	0.81	0.03
A0A0H2W8E5_A0A384LHF7	419	780	0.64	0.99	0.64	1.0	0.03
Q9HU12_Q9HX75	1227		0.14	0.83	0.01	0.27	0.03
P26401_Q8ZL53	1557		0.12	0.74	0.005	0.07	0.11
Q9KM04_Q9KUE5			0.13	0.62	0.001	0.03	0.03
Q5ZRT6_Q5ZTZ8	758	677	0.95	1	0.003	0.06	0.04
P0A4L3_Q92A74	700	295	0.38	0.74	0.04	0.03	0.07

Figure S11: Experimentally validated pairs, models, and metadata

Experimentally validated predicted interacting protein pairs encoded by distant genes. (**A**) Distance between predicted pairs of genes ≥ 0.8 AF score (~80% precision). (**B**) Predicted dimeric protein models with AF; left to right: Q5ZRK0-Q5ZYK1Q (green circle), Q9HX22-Q9HX38 (blue circle), Q8Y3X6-Q8Y695 (purple circle), P60415-Q928N1 (red circle), Q9HXJ0-Q9I5H0 (orange circle), A0A0H2W8E5-A0A384LHF7 (yellow circle); green bars between unique predicted interface residues ≤ 12Å. (**C**) Metadata table for pairs containing genetic distance, combined STRING scores, DCA and GREMLIN normalized to precision/recall curve, RF2-Lite and AF scores, and bacterial-two hybrid normalized to positive control (fig. S11). Annotations in table S7.

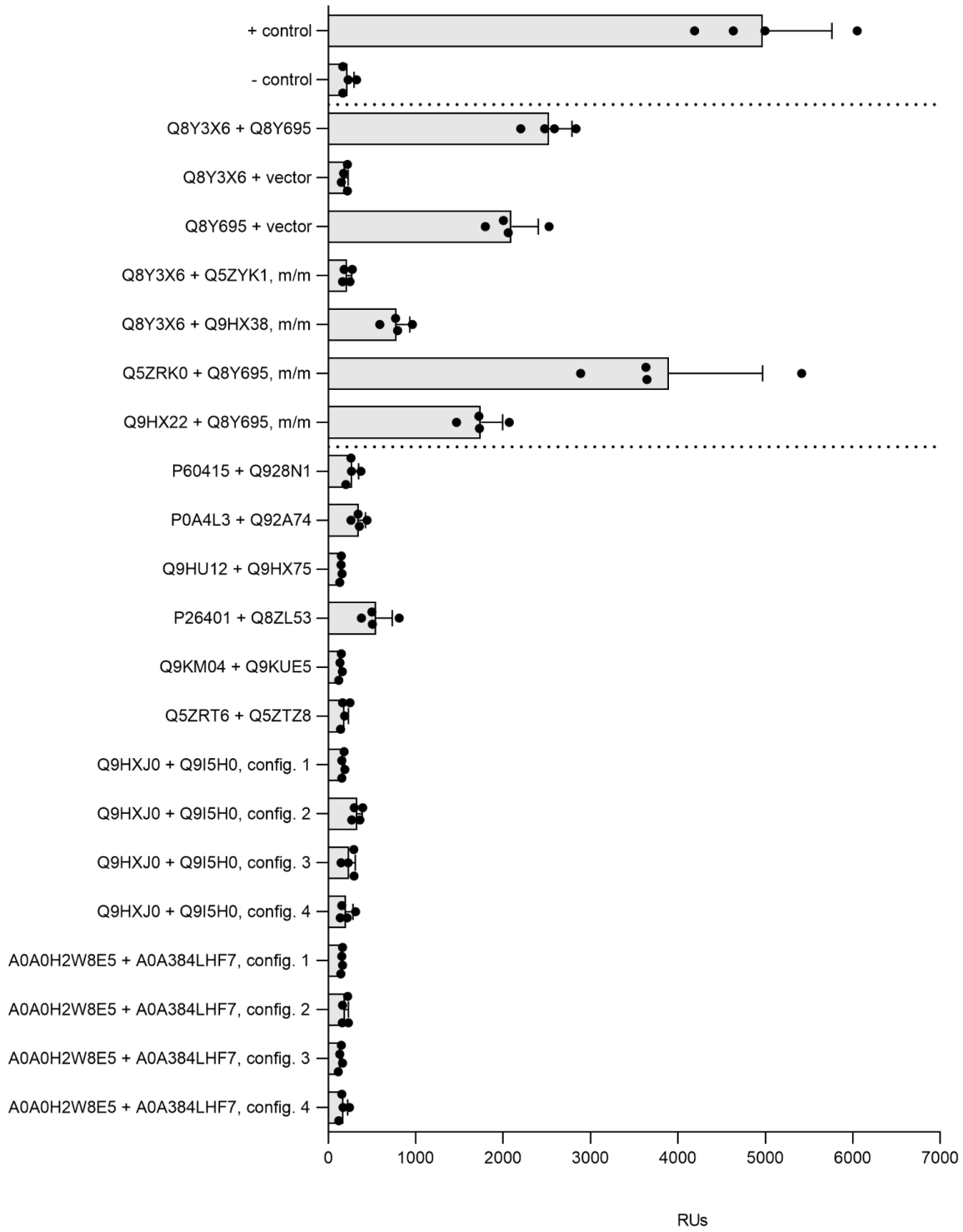


Figure S12: β-galactosidase activity of validated pairs by bacterial-two hybrid

Bacterial two-hybrid assays support a lack of interaction between coevolved proteins not predicted to interact by deep learning PPI methods. Assay and controls are as described in Fig. 2A. Config.1 ~ Config.4, different combinations of N-terminal or C-terminal fusions of T18 or T25 fragments with proteins of interest. Error bars indicate ±s.d. (n = 2 biological replicates each with n = 2 technical replicates). Annotations for proteins in table S6.

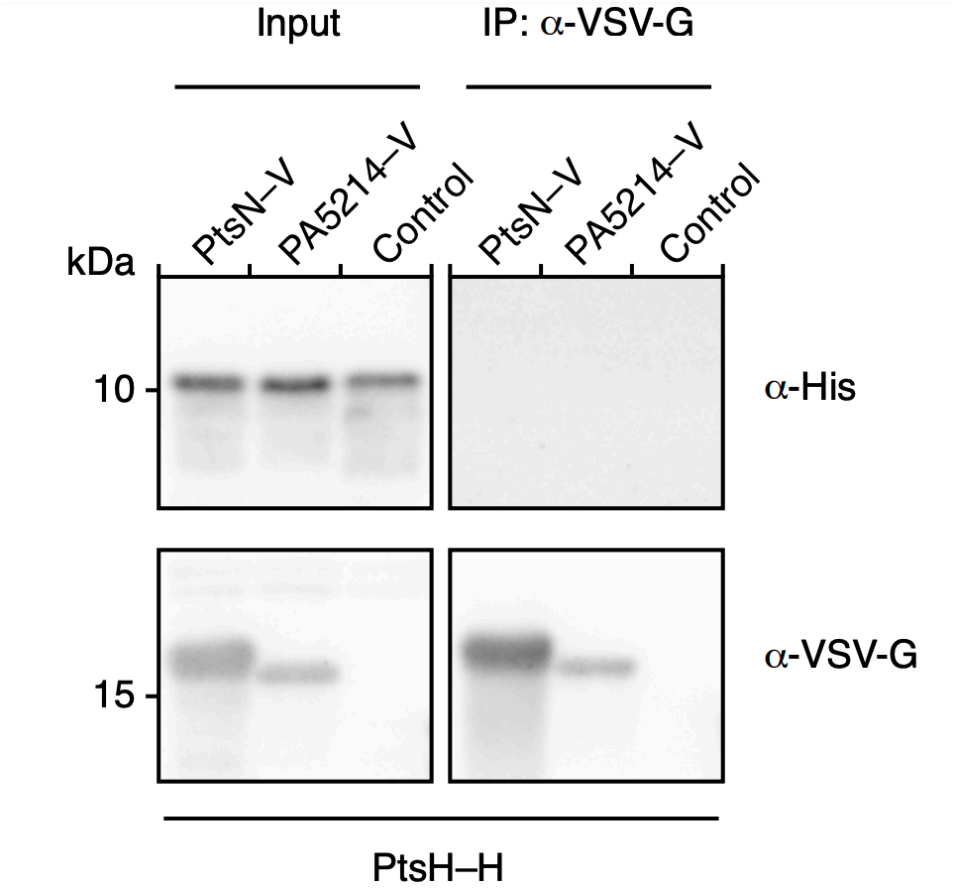


Figure S13: PtsH-PtsN negative Co-IP pulldown

α-VSV-G blot for PtsH-PtsN pulldown showing no signal by Co-IP despite strong string score and high PPI score suggesting potential challenges in experimentally validating some putative interaction pairs. N=2.

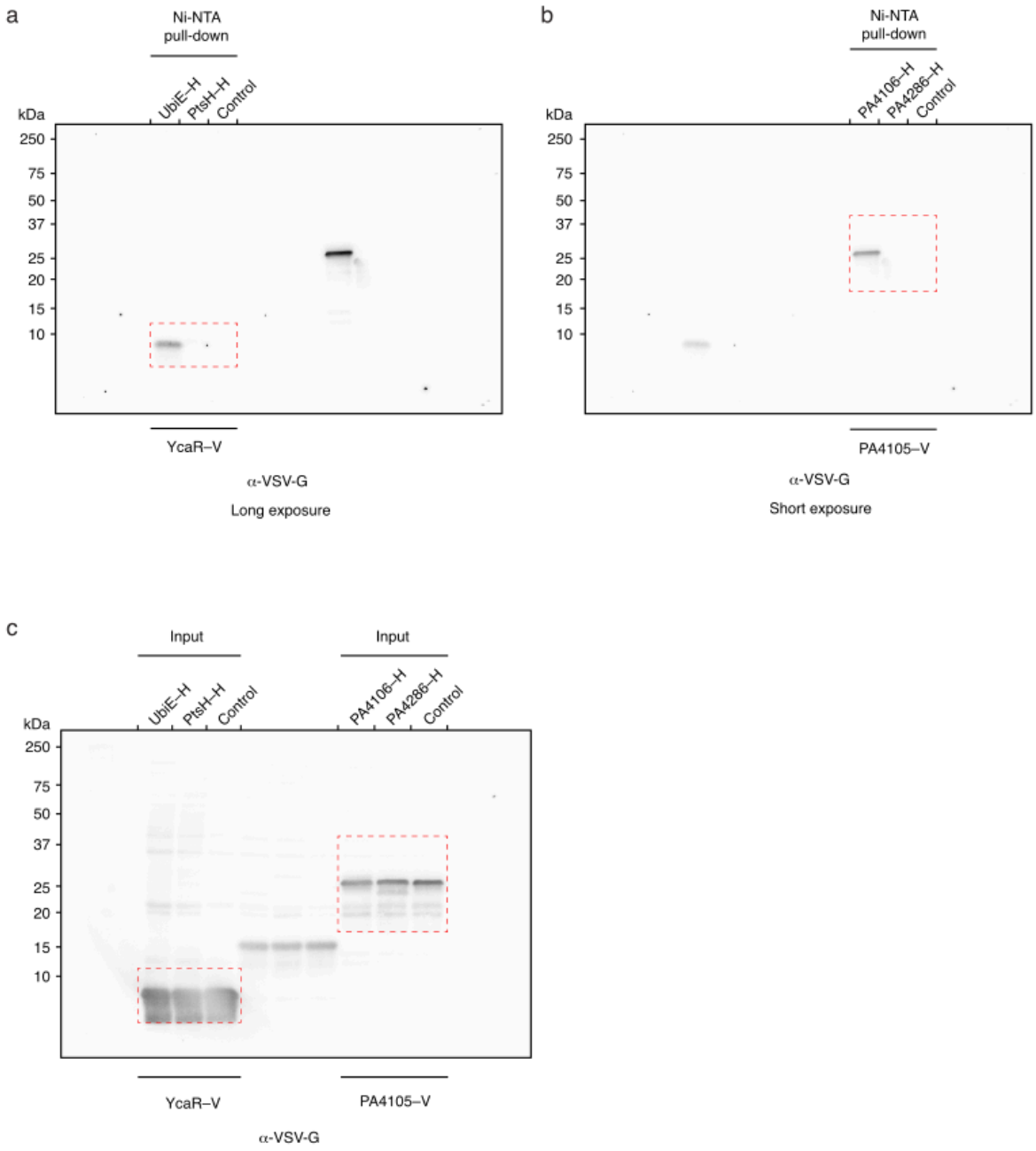


Figure S14: Uncropped western blot images for Figure 2: I

α-VSV-G blots for Fig 2a,b. The cropped region of each gel included in the main text figure is denoted by red dash boxes.

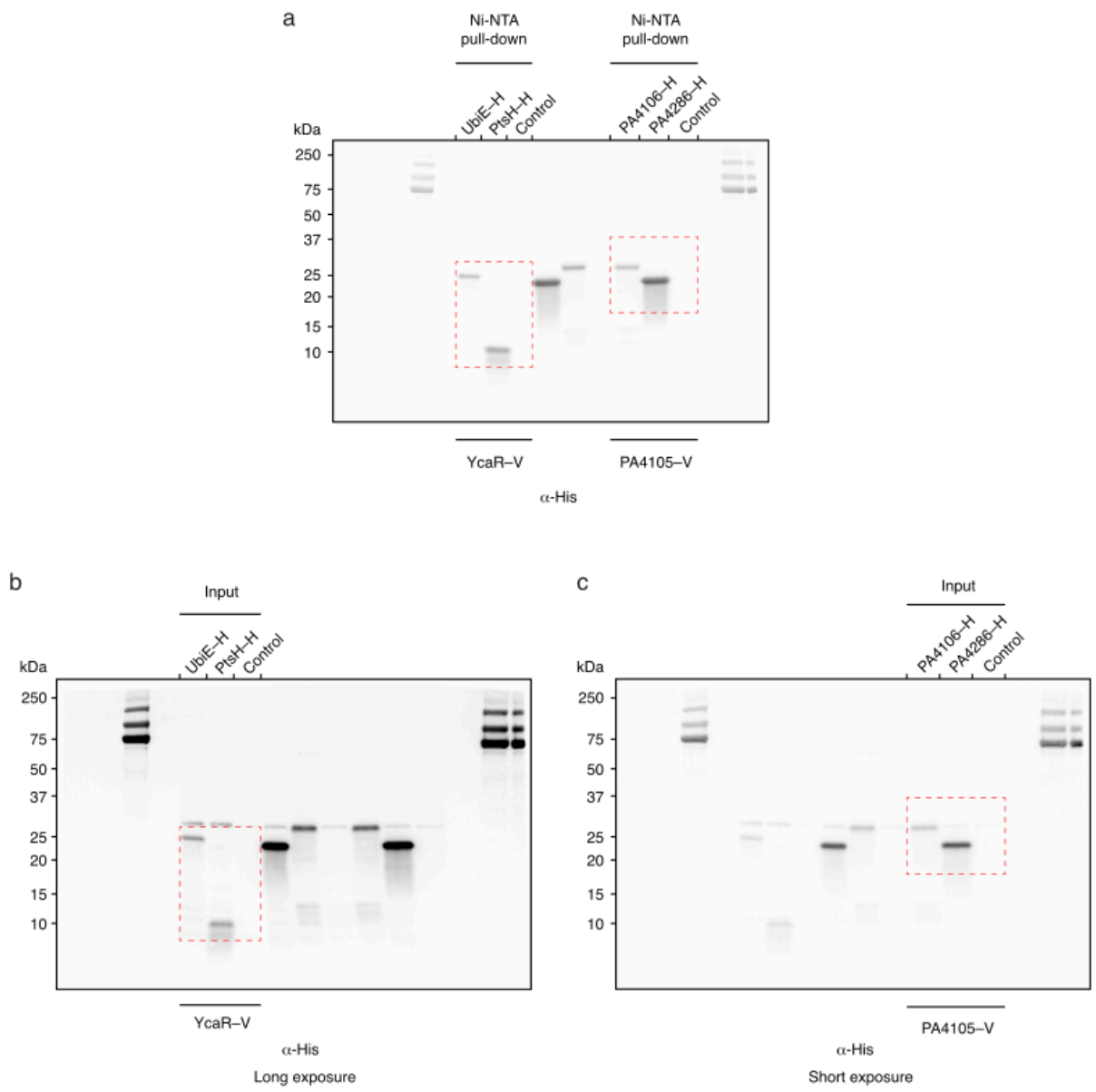


Figure S15: Uncropped western blot images for Figure 2: II

α-His-G blots for Fig 2a,b. The cropped region of each gel included in the main text figure is denoted by red dash boxes.

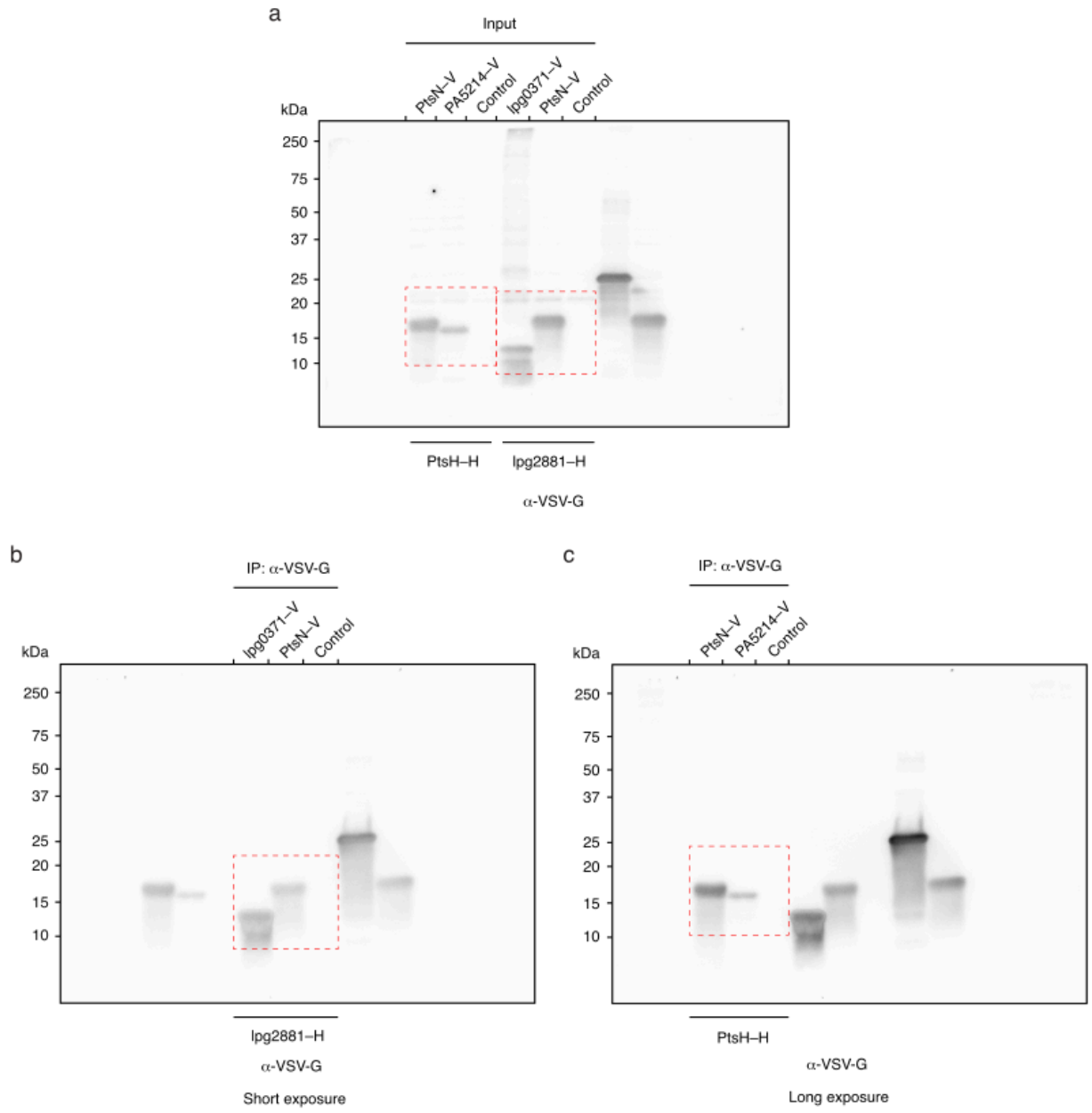


Figure S16: Uncropped western blot images for Figure 2: III

α-VSV-G blots for Fig 2d and supplemental figure S13. The cropped region of each gel included in the main text figure is denoted by red dash boxes.

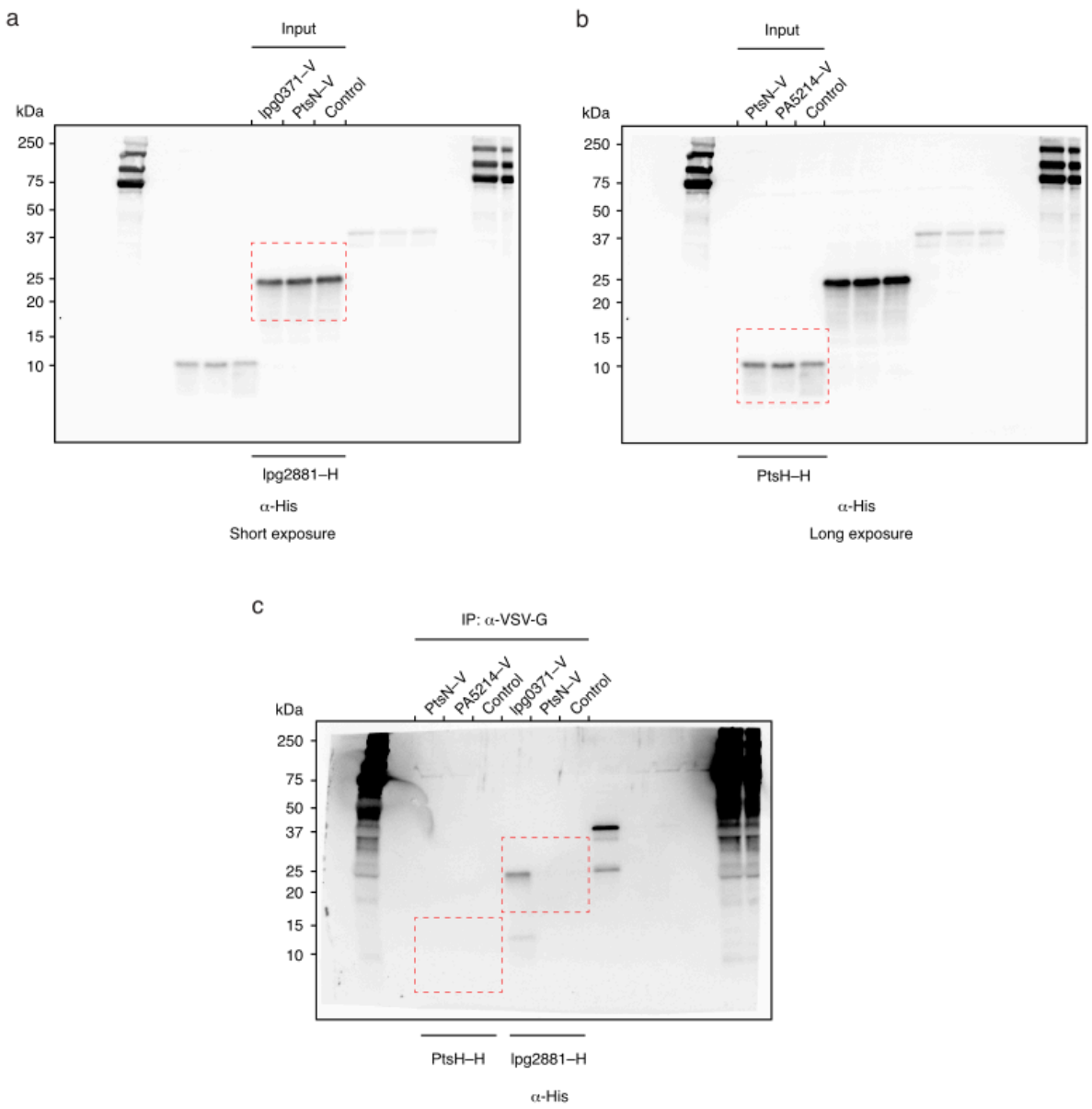


Figure S17: Uncropped western blot images for Figure 2: IV

α-His-G blots for Fig 2d and supplemental figure S13. The cropped region of each gel included in the main text figure is denoted by red dash boxes.

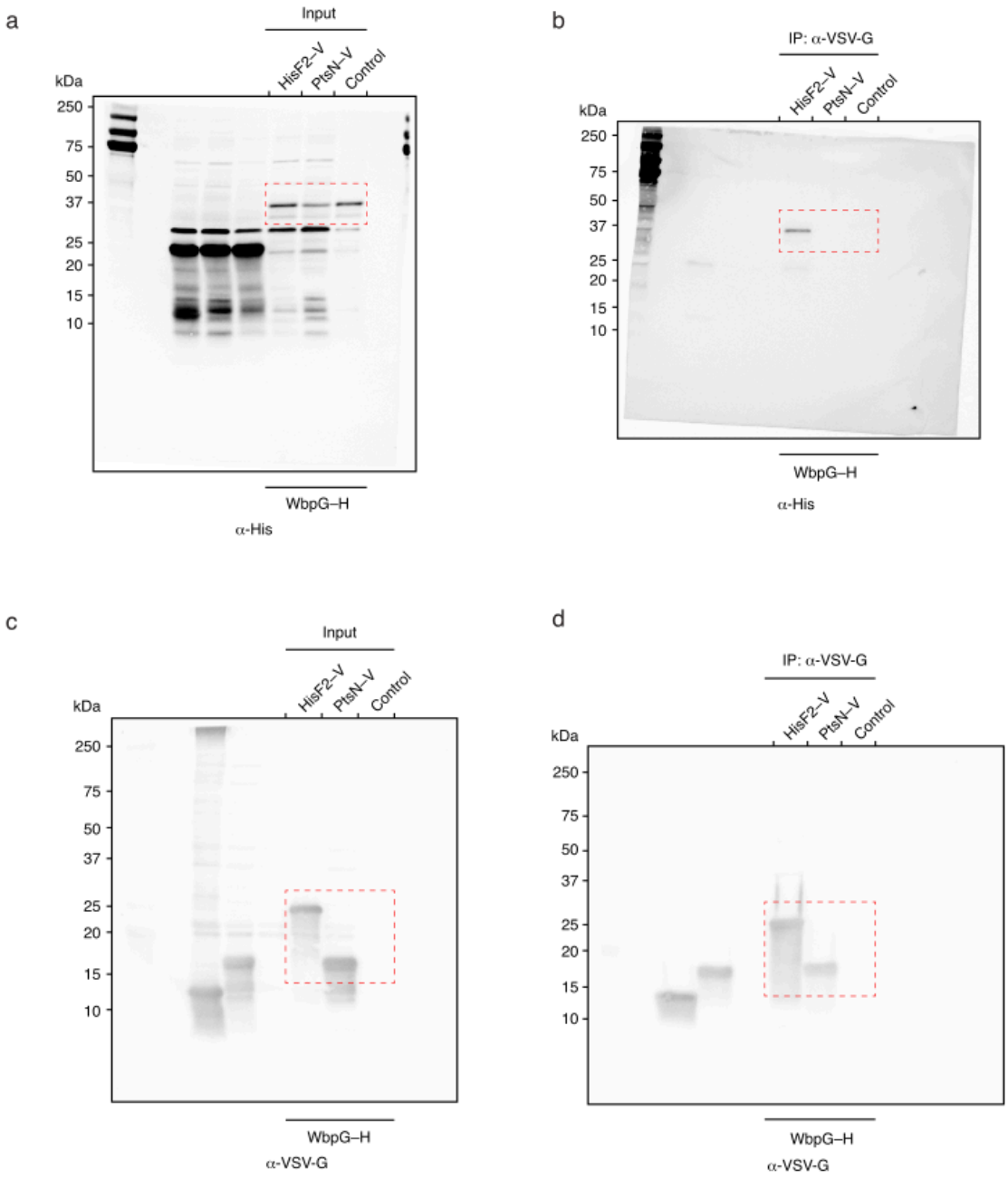


Figure S18: Uncropped western blot images for Figure 2: V

α-His and α-VSV-G blots for Fig 2e. The cropped region of each gel included in the main text figure is denoted by red dash boxes.

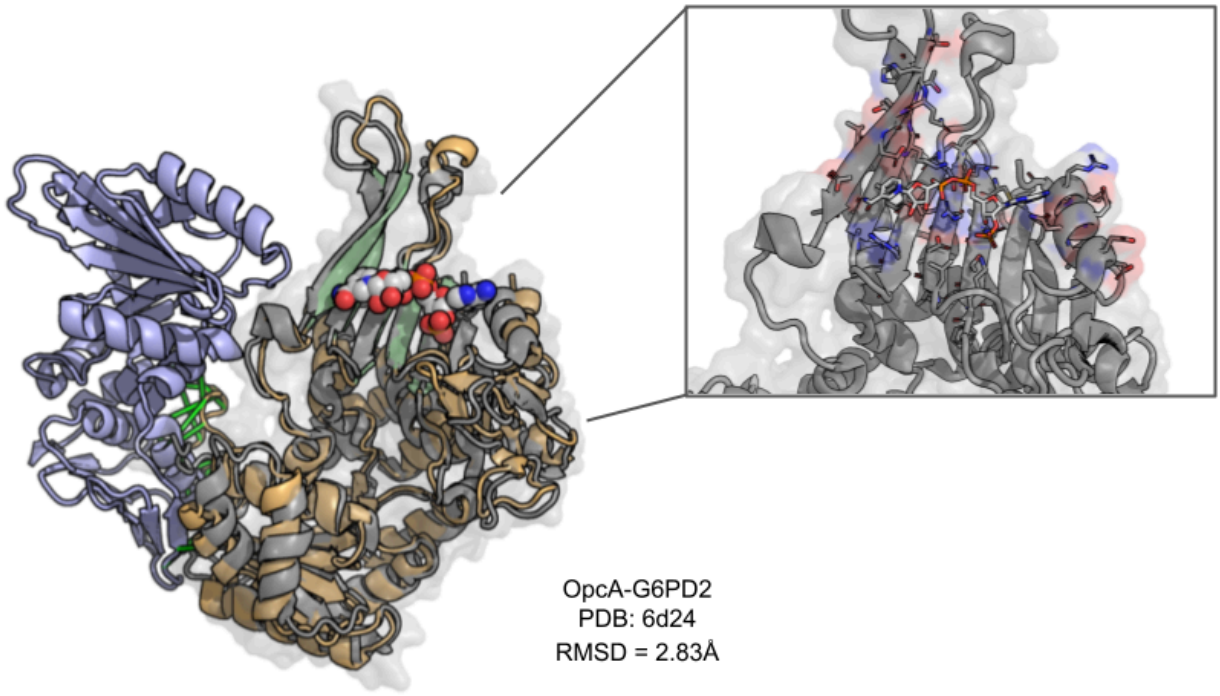


Figure S19: Glucose-6-phosphate 1-dehydrogenase and OPXX cycle protein

Predicted interaction of *M. tuberculosis* glucose-6-phosphate 1-dehydrogenase 2 (G6PD2) rendered in gold and OXPP cycle protein OpcA rendered in blue. The active site of the predicted structure is highlighted in green based on PDB annotation. The grey structure is an overlaid structure of G6PD (PDB: 6D24), focus box displays the PDB residues and NAD moiety.

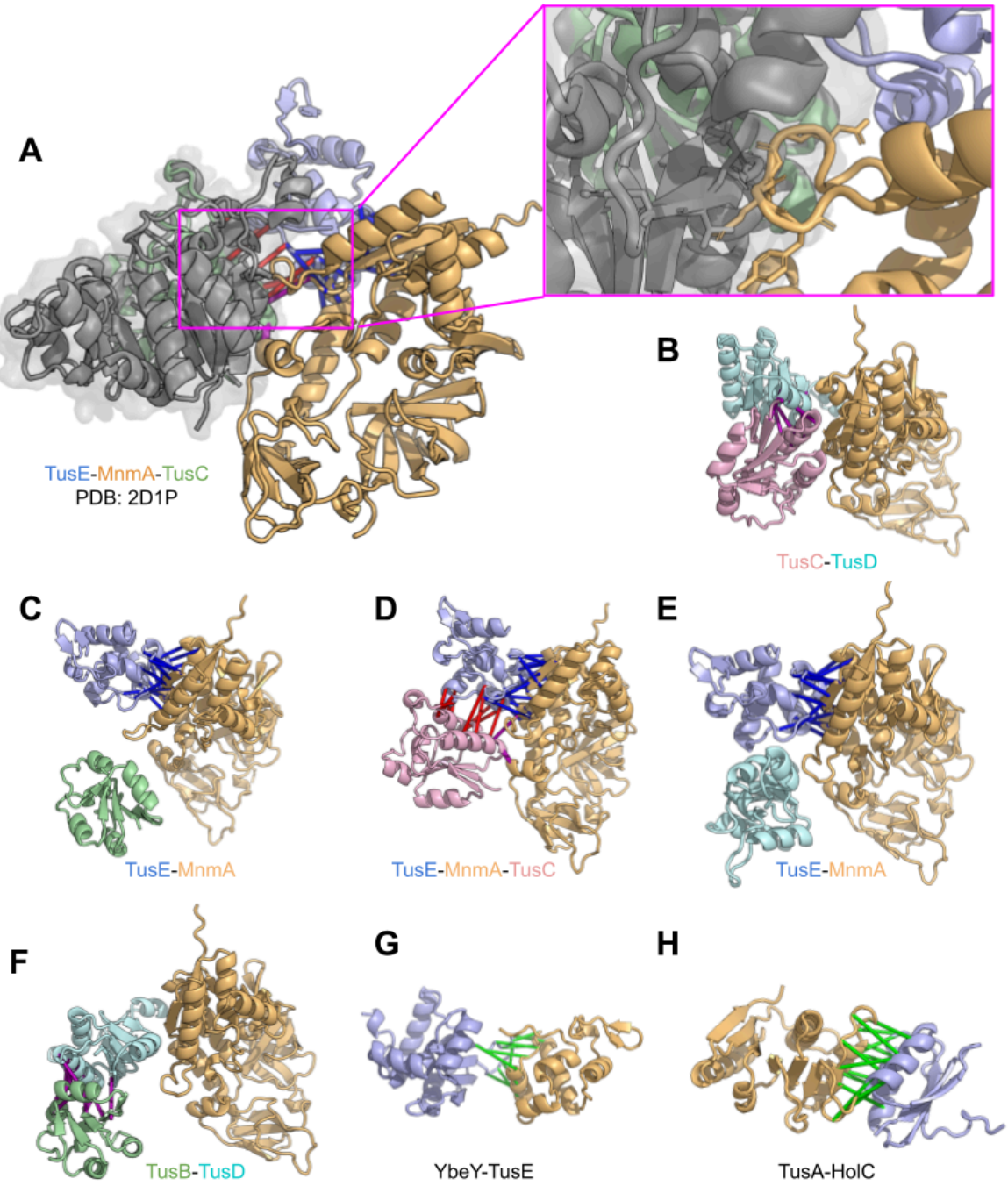


Figure S20: tRNA 2-thiouridine synthesizing complex (Tus) and MnmA trimers

Predictions of interactions with Tus in *E. coli*. (A) The modeled trimer of TusE-MnmA-TusC with TusBCD overlay (PDB: 2D1P) aligned to TusC showing that while MnmA may interact with TusEC, there are steric clashes with TusBCD; although it should be noted that there may be some flexibility to accommodate this interface which we are unable to capture though overlaid structures. (B-F) Attempted trimeric combinatorial interaction modeling of TusB, TusC, TusD, and TusE with MnmA; below each model the components which were predicted to interact are named. (G) Predicted interaction between YbeY-TusE. (H) Predicted interaction between TusA-HolC.

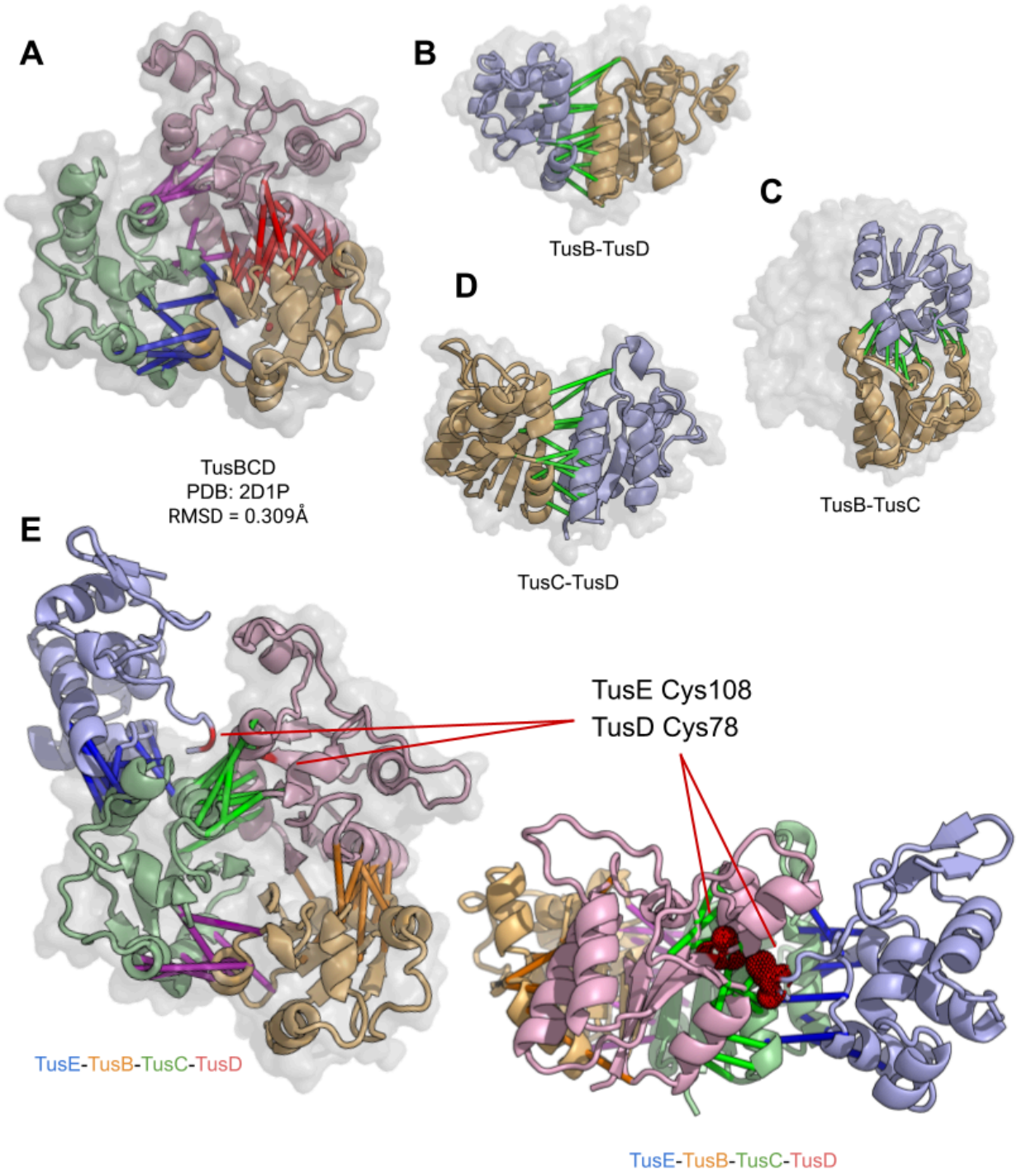


Figure S21: tRNA 2-thiouridine synthesizing complex (Tus)

Predictions of interactions with Tus in *E. coli*. (A-E) TusB, TusC, TusD interactions with experimentally determined TusBCD structure overlay in grey (PDB: 2D1P). (E) contains two highlighted cysteine residues which were identified as functionally relevant.

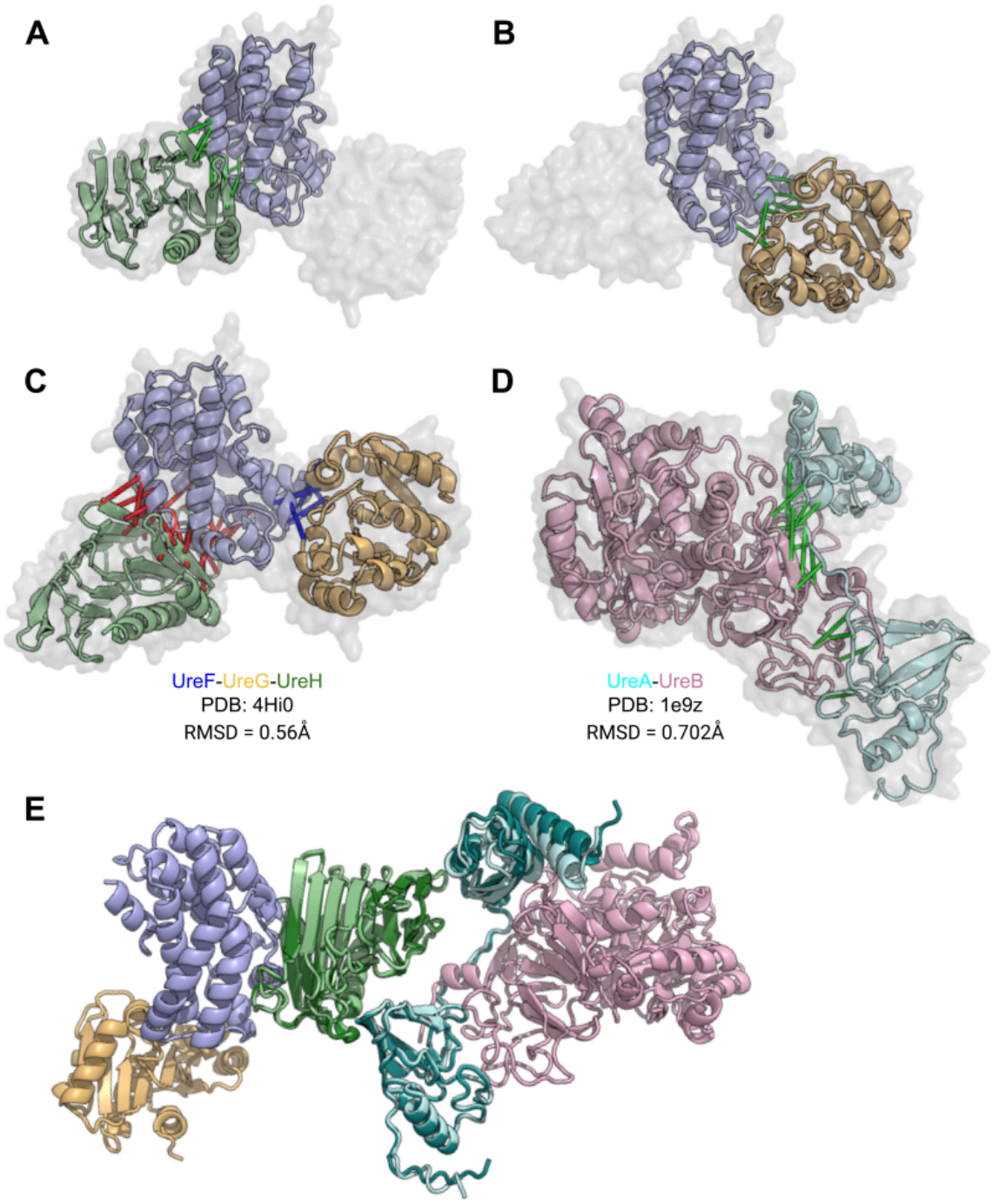


Figure S22: Urease oligomeric assembly generation

H. pylori UreAB and UreFGH complex validation through experimental structure overlays (PDB: 4Hi0, 1e9z) (*20*, *21*). (E) Pentameric UreAB-UreFGH complex assembled through multiple subcomplexes: UreFGH, UreAB, and UreAH aligned to UreAH depicted in dark teal and dark green of panel E. We note that there appear to be some steric clashes at the interface between UreA-UreH when aligned to PDB subcomplexes; however, this is not found in our predicted dimeric models. We would also like to note that we do also predict UreB-UreH, however, the interaction score from AlphaFold was 0.7, which is below our 95% precision threshold.

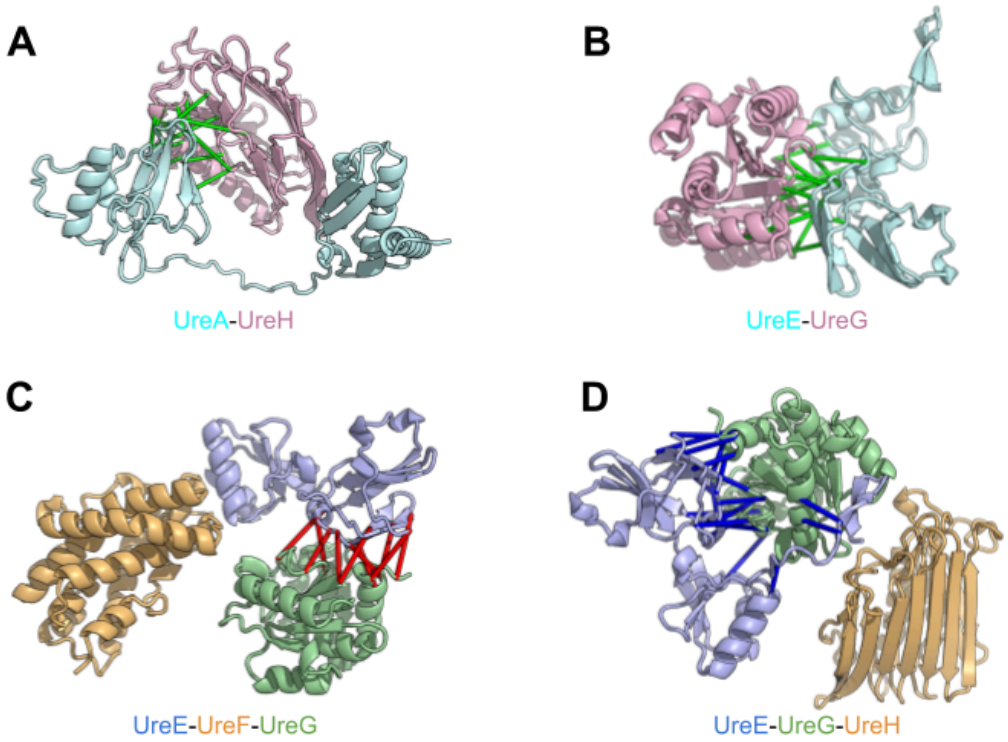


Figure S23: Urease trimeric interactions

Additional *H. pylori* urease predictions. (A) Dimeric interaction of UreA-UreH, which the previously described UreAB-UreFGH complex is aligned to. (B) UreE-UreG, dimeric interaction. (C,D) UreE-UreG interaction is modeled with third protein UreF or UreH, respectively, which are not predicted to interact with UreG in the presence of UreE.

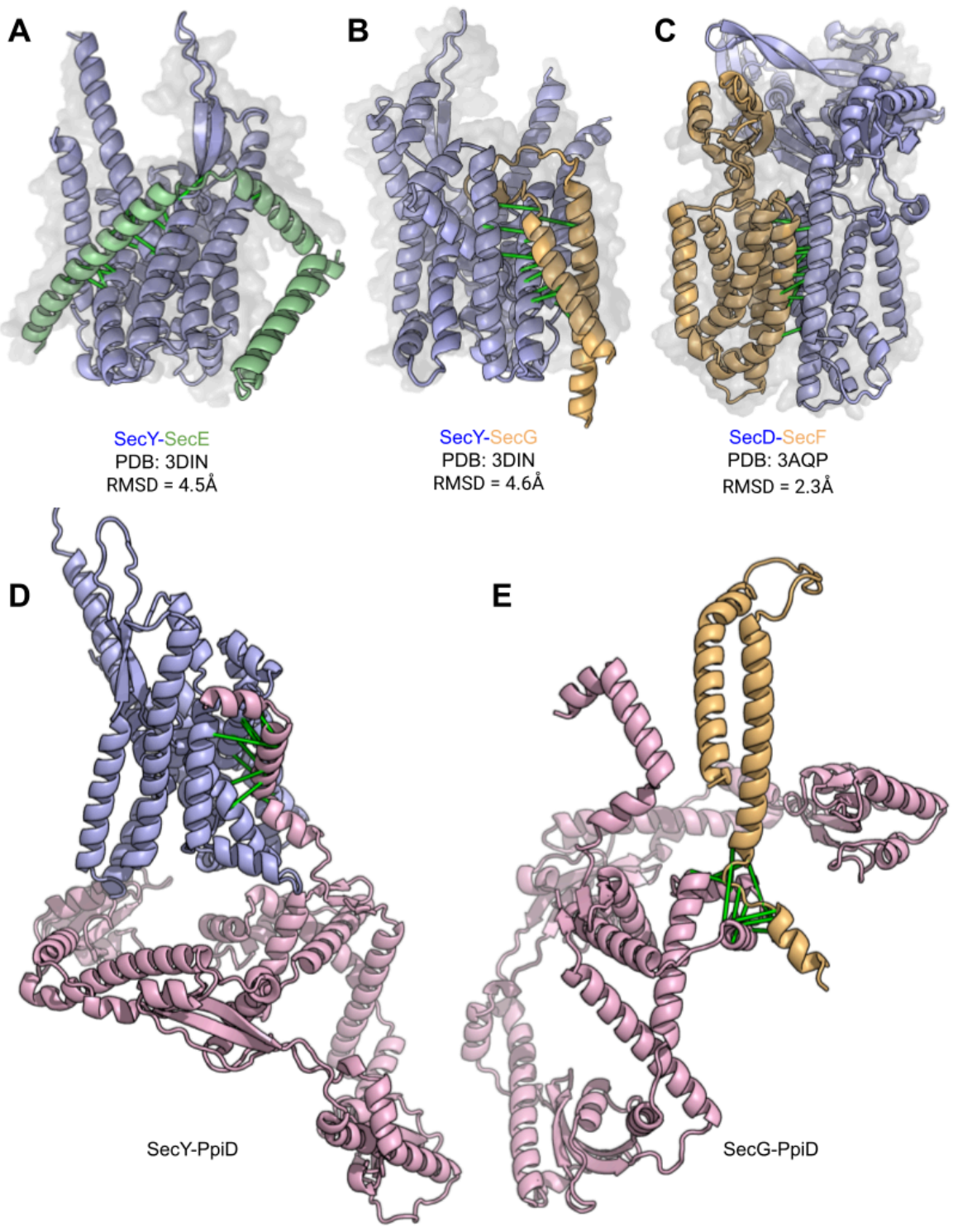


Figure S24: Sec translocon orthologous PDB validation and PpiD

Interactions with components of the Sec translocon in *P. aeruginosa*. (A,B) SecY-SecE and SecY-SecG dimers with overlay (PDB: 3DIN). (C) SecD-SecF dimeric interaction with overlay (PDB: 3AQP). (D,E) Predicted dimeric interactions with SecY/SecG and PpiD.

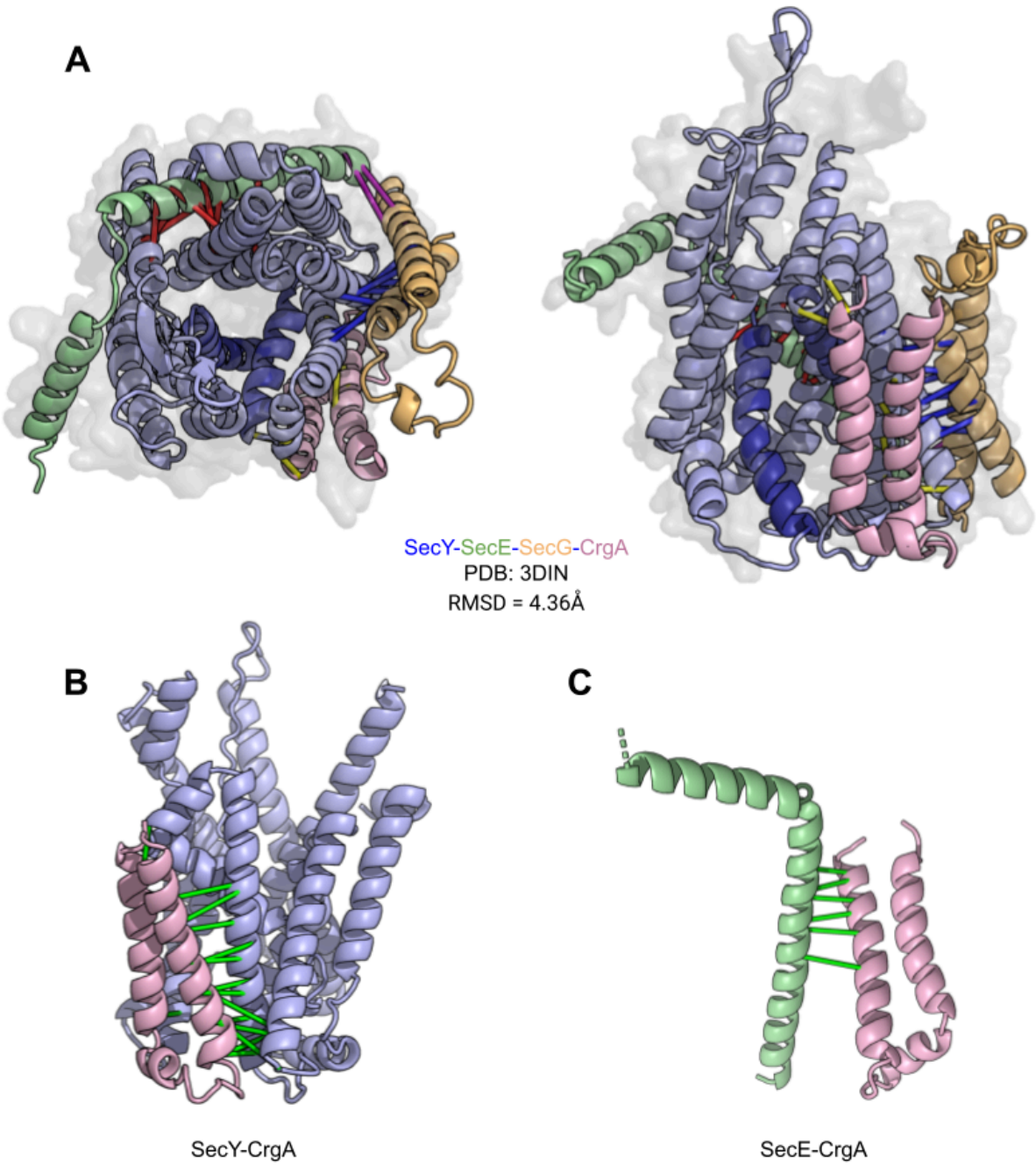


Figure S25: Sec translocon interactions with CrgA

Sec and CrgA interactions in *M. tuberculosis*. (A) One-shot prediction of SecYEG-CrgA with SecYEG structure overlay (PDB: 3DIN). Left top-down; right rotated 90 degrees. Transmembrane helices of two and seven colored in dark blue correspond to the lateral gate of SecY. (B,C) Dimeric interaction predictions between CrgA and Sec proteins SecY and SecE.

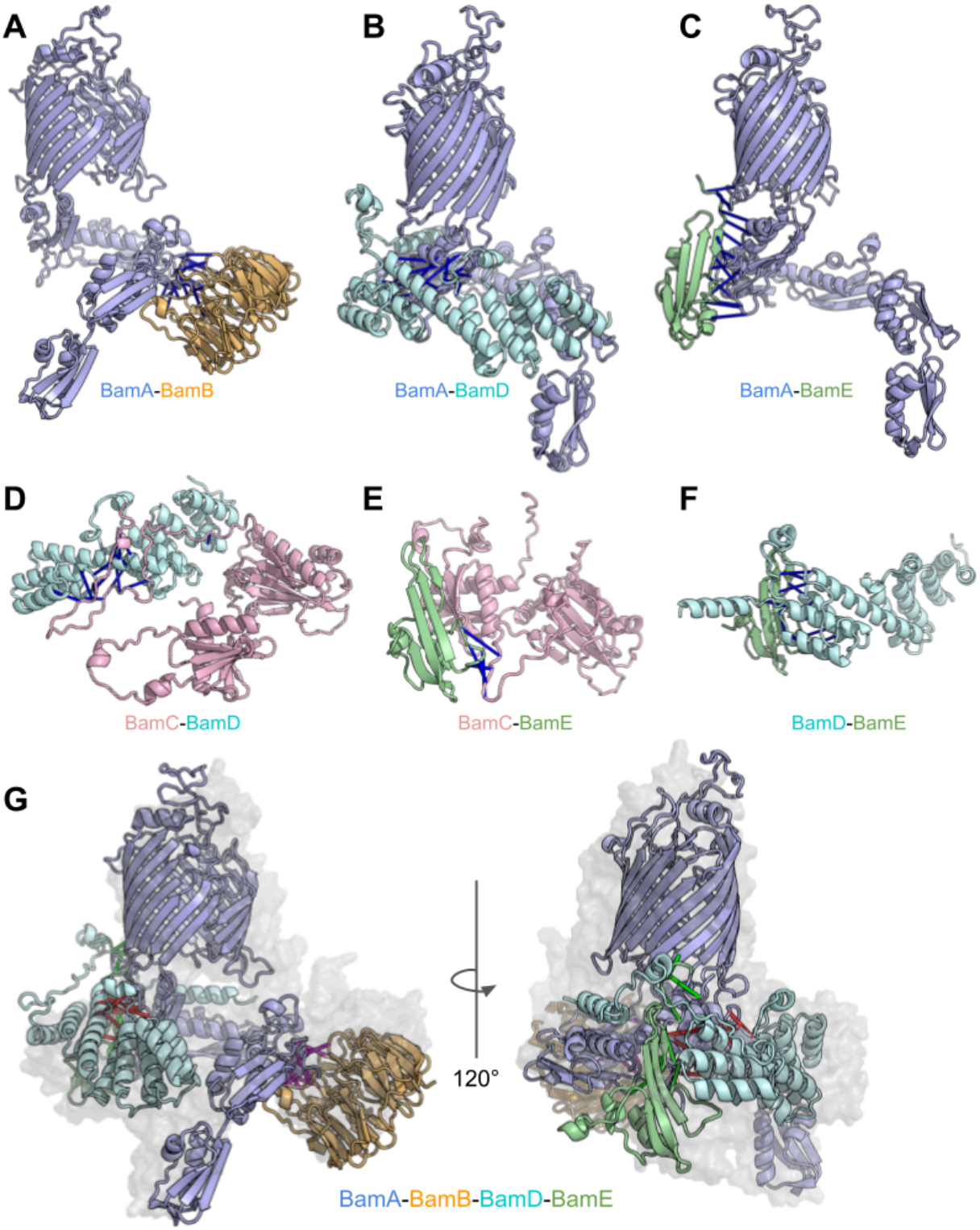


Figure S26: BAM complex orthologous PDB validation

Predicted interactions within the *P. aeruginosa* BamABDE complex. (A-F) predicted dimeric interactions of BamABDE. (G) Predicted one-shot tetrameric BamABDE complex with structure overlay (PDB: 5D0O).

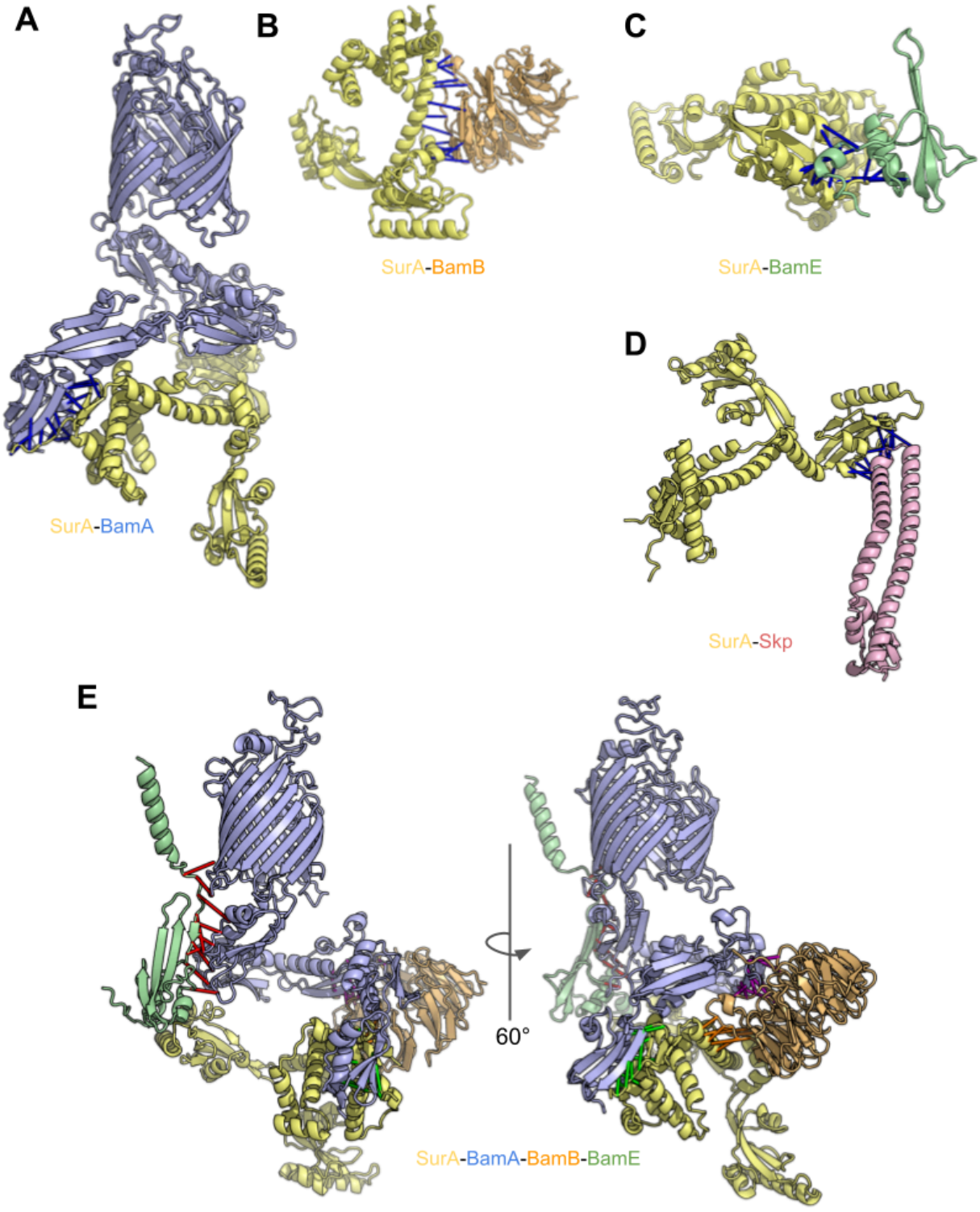


Figure S27: BAM complex and SurA interactions

Predicted interactions with *V. cholera* and *P. aeruginosa* Bam and SurA. (A-C) predicted dimeric interaction with SurA. (D) Predicted dimeric interaction between SurA-Skp. (E) Predicted one-shot tetrameric BamABE-SurA complex.

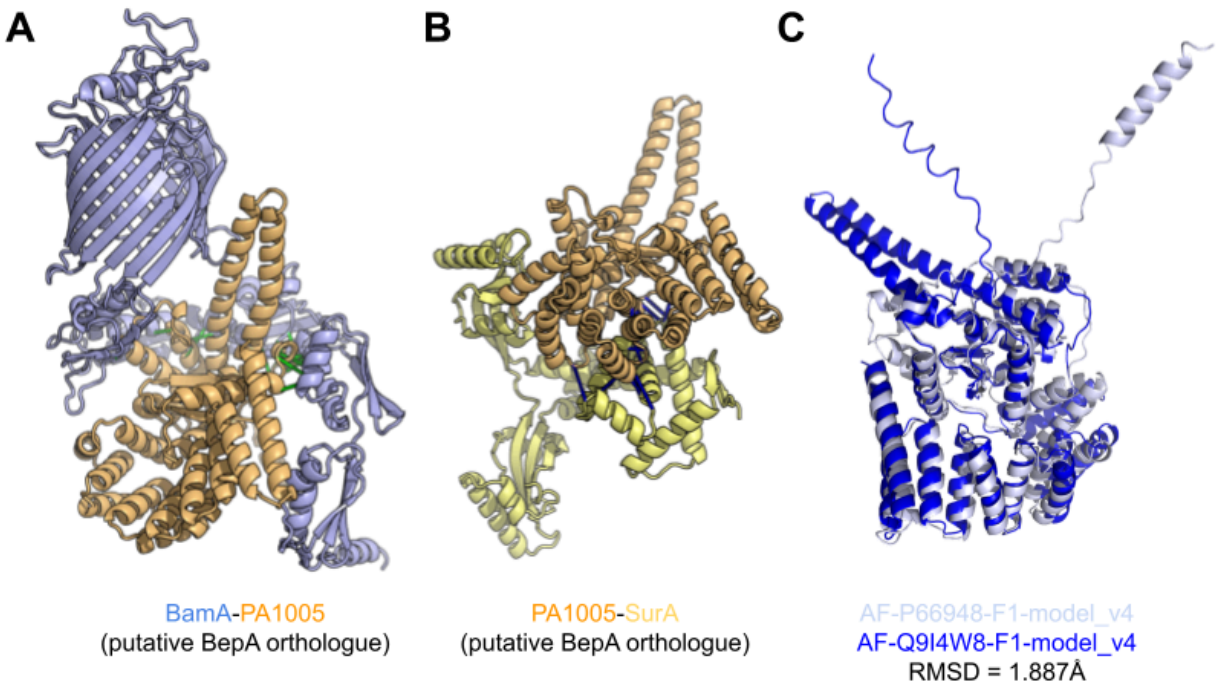


Figure S28: BepA putative orthologue identification and Bam/SurA interaction

Two interactions and monomeric structure of PA1005, a *P. aeruginosa* putative orthologue of BepA. (A) Predicted interaction between BamA-PA1005. (B) Predicted interaction between PA1005-SurA. (C) AlphaFold database V4 model of *P. aeruginosa* PA1005 (accession: Q9I4W8) in dark blue aligned to *E.coli* BepA (accession: AF-P66948) in light blue.

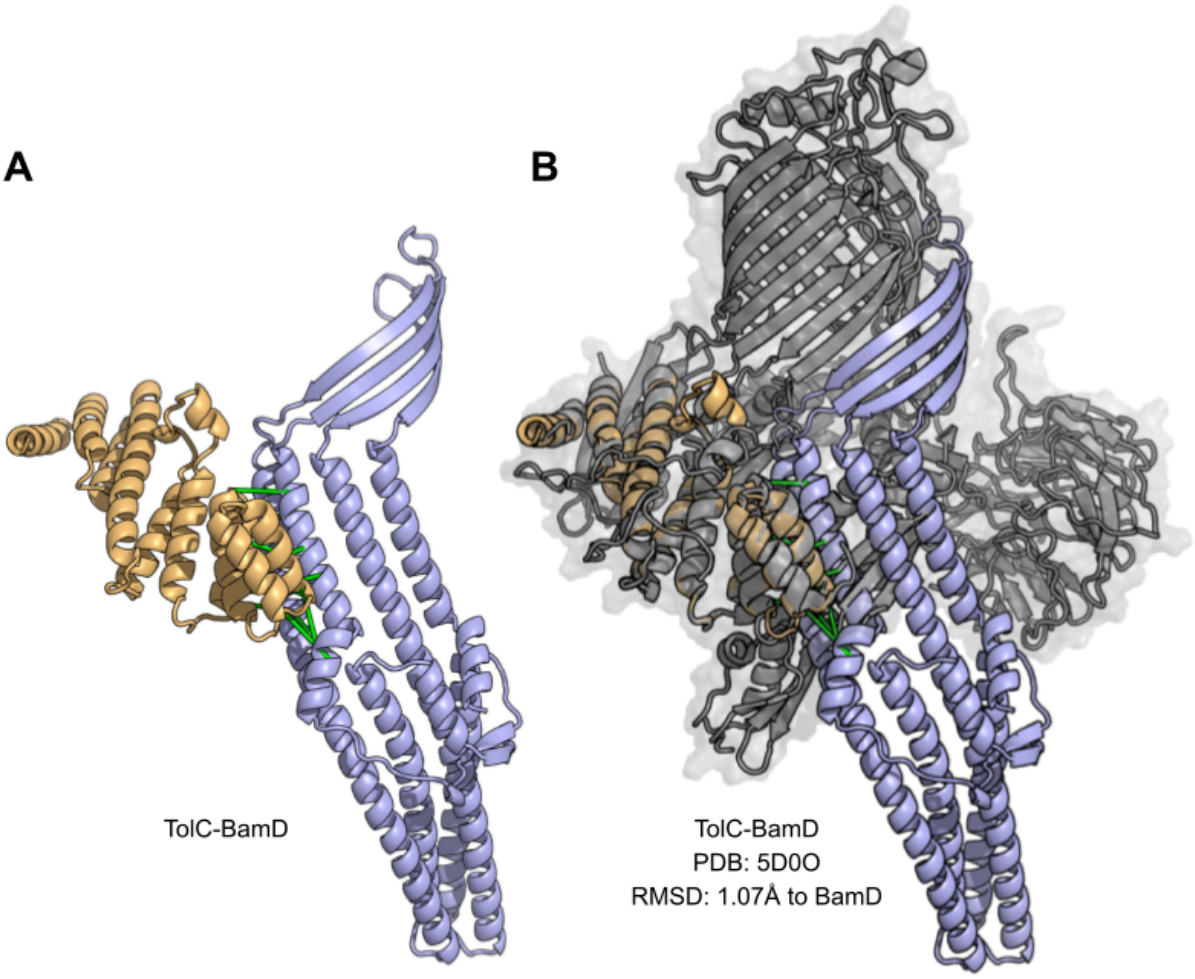


Figure S29: Folding of TolC by BAM complex

TolC interaction with BAM in *S. typhimurium*. (A) Predicted dimeric interaction of TolC-BamD. (B) Dimeric prediction of TolC-BamD aligned to BamD in a structure of BamABCDE complex (PDB: 5D0O); showing how the β-sheet of BamD aligns with the seam of the BamA β-barrel that folds nascent polypeptides and insertion into the outer membrane.

Supplemental Tables

Table S1: Proteome strains and Uniprot accessions

Pathogen (strain)	Uniprot proteome		
Aca (PHEA-2)	UP000007477	**Lpn** (ATCC 33152)	UP000000609
Bfr (YCH46)	UP000002197	**Mge** (G37)	UP000000807
Bhe (ATCC 49882)	UP000000421	**Mtu** (H37Rv)	UP000001584
Cdi (630)	UP000001978	**Nme** (MC58)	UP000000425
Ctr (A/HAR-13)	UP000002532	**Pae** (PA01)	UP000002438
Eco (K12)	UP000000625	**Sau** (PS 47)	UP000008816
Ftu (SCHU S4)	UP000001174	**Spn** (2070335)	UP000002642
Hpy (ATCC 700392)	UP000000429	**Sty** (LT2)	UP000001014
Lmo (EGD-e)	UP000000817	**Vch** (ATCC 39315)	UP000000584
		Ype (CO-92)	UP000000815

Table S2: Statistics for pathogens in our study

Abbr.	Organism	Total[1]	Essential genes (unique)[2]	Virulence factors (unique)[3]
Aca	*Acinetobacter calcoaceticus*	3598	566 (31)	34 (6)
Bfr	*Bacteroides fragilis*	4597	463 (129)	0
Bhe	*Bartonella henselae*	1466	3 (0)	45 (22)
Cdi	*Clostridioides difficile*	3762	1 (0)	13 (10)
Ctr	*Chlamydia trachomatis*	917	0	113 (95)
Eco	*Escherichia coli*	4262	1155 (98)	51 (15)
Ftu	*Francisella tularensis*	1528	495 (46)	89 (34)
Hpy	*Helicobacter pylori*	1553	324 (118)	113 (47)
Lmo	*Listeria monocytogenes*	2844	7 (0)	35 (14)
Lpn	*Legionella pneumophila*	2930	0	463 (324)
Mge	*Mycoplasma genitalium*	483	381 (101)	0
Mtu	*Mycobacterium tuberculosis*	3993	1138 (383)	97 (55)
Nme	*Neisseria meningitidis*	2001	534 (43)	56 (18)
Pae	*Pseudomonas aeruginosa*	5564	768 (105)	289 (88)
Sau	*Staphylococcus aureus*	2889	546 (32)	63 (36)
Spn	*Streptococcus pneumoniae*	2823	379 (83)	32 (21)
Sty	*Salmonella typhimurium*	4533	842 (50)	139 (69)
Vch	*Vibrio cholerae*	3783	838 (321)	141 (43)
Ype	*Yersinia pestis*	3909	134 (0)	120 (22)
Total:		57,435	8574 (1540)	1893 (919)

[1]Total number of proteins in the proteome. [2]Number of proteins mapped to essential genes in DEG. [3]Number of proteins mapped to virulence factors in VFDB. In parentheses are the number of unique EG or VF in each organism that lacks orthologues from other pathogens in our study.

Table S3: RoseTTAFold2-Lite training setup

	Initial training	Fine-tuning
Crop size	256	384
Batch size	32	32
Loss function	$3.0*Loss_{MLM} + 1.0*Loss_{dist} + 10.0*Loss_{FAPE} + 0.1*Loss_{accuracy}$	$3.0*Loss_{MLM} + 1.0*Loss_{dist} + 10.0*Loss_{FAPE} + 0.1*Loss_{accuracy} + 0.1*Loss_{bond} + 0.1*Loss_{vdW}$
Learning rate & scheduling	0.001 Linear warm-up for first 1000 optimization steps, then decay learning rate by 0.95 after every 15000 optimization steps	0.0005 No warm-up. Decay learning rate by 0.95 after every 15000 optimization steps
Examples per epoch	22400	22400
Number of epochs	200	100

Table S4: Recall of filtering pipeline

Method(s) of filtering			Recall
DCA			4.1%
DCA	RF2-Lite		28%
DCA	RF2-Lite	AF2	29%

Recall of filtering pipeline methods based on 95% precision of pilot-set benchmark based on DCA → RF2-Lite → AF2 successive filtering.

Table S5: Predicted interactions in STRING

Set	STRING (total)	STRING (exp)	New interactions	Existing interactions
Pilot	900	400	381	181
Pilot	700	400	376	186
Pilot	400	400	375	187
Pilot	900		287	275
Pilot	700		259	303
Pilot	400		234	328
All	900	400	2411	1202
All	700	400	2314	1299
All	400	400	2297	1316
All	900		1751	1862
All	700		1382	2231
All	400		1251	2362

Number of new/known pilot-set and all (extended-set + pilot-set) protein pairs predicted in this screen based on STRING scores at various cutoffs.

Table S6: RF2-Lite pilot-set and extended-set pairs by pathogen

Abbr.	Organism	Pilot-set	Extended-set	Total
Aca	*Acinetobacter calcoaceticus*	33,094	100,173	133,267
Bfr	*Bacteroides fragilis*	13,999	55,899	69,898
Bhe	*Bartonella henselae*	344	614	958
Cdi	*Clostridioides difficile*	823	687	1,510
Ctr	*Chlamydia trachomatis*	1,604	2,734	4,338
Eco	*Escherichia coli*	58,417	367,586	426,003
Ftu	*Francisella tularensis*	20,327	89,404	109,73
Hpy	*Helicobacter pylori*	9,570	55,725	65,295
Lmo	*Listeria monocytogenes*	717	296,807	297,524
Lpn	*Legionella pneumophila*	7,363	50,810	58,173
Mge	*Mycoplasma genitalium*	5,002	24,996	29,998
Mtu	*Mycobacterium tuberculosis*	104,177	1,096,732	1,200,909
Nme	*Neisseria meningitidis*	20,242	96,404	116,646
Pae	*Pseudomonas aeruginosa*	83,052	1,235,506	1,318,558
Sau	*Staphylococcus aureus*	25,257	57,957	83,21
Spn	*Streptococcus pneumoniae*	18,387	45,283	63,670
Sty	*Salmonella typhimurium*	31,144	127,260	158,40
Vch	*Vibrio cholerae*	19,698	98,932	118,630
Ype	*Yersinia pestis*	4,092	20,707	24,799
Total:		457,310	3,824,215	4,281,525

Pilot-set and extended-set protein pairs screened in this study. All pilot-set were selected from the top scoring pairs by DCA that were annotated as virulence factors or essential genes and run through RF2-Lite.

Table S7: Metadata of experimentally validated interactions by B2H

Org.	Uniprot	Locus	Gene	Annotations
Lpn	Q5ZRK0	lpg2881	-	Iron-sulfur cluster binding protein
	Q5ZYK1	lpg0371	-	UPF0125 protein lpg0371
Pae	Q9HX22	PA4005	RsfS	Ribosomal silencing factor RsfS
	Q9HX38	PA3981	YbeZ	PhoH-like protein domain-containing protein
Lmo	Q8Y3X6	lmo2703	-	Nucleoid-associated protein lmo2703
	Q8Y695	-	Ffh	Signal recognition particle protein
Lmo	P60415	lmo2054	-	UPF0298 protein lmo2054
	Q928N1	lmo2402	-	Lmo2402 protein
Pae	Q9HXJ0	PA3812	IscA	Iron-binding protein IscA
	Q9I5H0	PA0759	YgfZ	Folate-binding protein YgfZ
Ype	A0A0H2W8E5	YPO3902	-	Putative magnesium chelatase family protein
	A0A384LHF7	YPO0243	-	DprA winged helix domain-containing protein
Pae	Q9HU12	PA5177	-	Probable hydrolase
	Q9HX75	PA3941	-	Uncharacterized protein
Sty	P26401	STM2087	RfbV	Abequosyltransferase RfbV
	Q8ZL53	STM3707	YibD	Putative glycosyltransferase
Vch	Q9KM04	VC_A0585	-	Glutathione S-transferase, putative
	Q9KUE5	VC_0576	-	Stringent starvation protein A
Lpn	Q5ZRT6	lpg2792	TpiA	Triosephosphate isomerase
	Q5ZTZ8	lpg2010	Gmk	Guanylate kinase
Lmo	P0A4L3	lmo1233	TrxA	Thioredoxin
	Q92A74	-	Hup	Hup protein

Experimentally validated interaction pairs by B2H displayed in Fig. 2 and figs. S11-12 with organism abbreviations (see above), uniprot ID, gene locus, name, and annotations.

Table S8: Metadata of experimentally validated interactions by Co-IP

Org.	Uniprot	Locus Gene	Score	Contacts	Globularity	Genetic distance	STRING	Annotations
Eco	P0A887 P0AAZ7	ubiE ycaR	0.99	79	98% 92%	1408	287	Ubiquinone/menaquinone biosynthesis C-methyltransferase IspA family inner membrane protein; Involved in cell division
Pae	Q9HWS2 Q9HWS3	PA4106 PA4105	1.0	203	98% 98%	2	na	UPF0276 protein PA4106 Putative DNA-binding domain-containing protein
Pae	P72139 Q9HZ78	hisF2 wbpG	0.99	99	98% 100%	2	na	Putative imidazole glycerol phosphate synthase subunit hisF2 LPS biosynthesis protein WbpG
Lpn	Q5ZRK0 Q5ZYK1	lpg2881 lpg0371	1.0	114	86% 83%	475	573	Iron-sulfur cluster binding protein UPF0125 protein lpg0371
Pae	Q9HVV2 Q9HVV4	ptsH ptsN	0.96	54	100% 100%	3	959	Phosphocarrier protein HPr Nitrogen regulatory protein
Lpn	Q5ZW63 Q5ZX88	FlgJ Ttg2D	0.94	54	87% 90%	375	na	Muramidase, peptidoglycan hydrolase FlgJ Signal peptide protein, toluene tolerance protein Ttg2D

Experimentally validated interaction pairs by Co-IP displayed in Fig. 2 with organism abbreviations (see above), uniprot ID, locus/gene, AFscores, number of contacts at interface, globularity, genetic distance, STRING combined score, and uniprot annotations. The four pairs above the green line were found to interact by Co-IP; the two pairs below the red line were not detected to interact by Co-IP.

Table S9: Uniprot annotations of interactions in Figure 3

Fig. 3	Org.	Pair	Protein 1	Protein 2
A	Mtu	OpcA-G6PD2 (zwf2)	OXPP cycle protein OpcA	Glucose-6-phosphate 1-dehydrogenase 2 (G6PD 2)
B	Eco	SapD-SapB	Putrescine export system ATP-binding protein SapD	Putrescine export system permease protein SapB
C	Eco	MreB-RodZ	Cell shape-determining protein MreB	Cytoskeleton protein RodZ
D	Mtu	PonA-MutT4	Penicillin-binding protein 1A (PBP-1A)	Putative mutator protein MutT4
E	Pae	PA3801-PpiD	Ancillary SecYEG translocon subunit	Peptidyl-prolyl cis-trans isomerase D
F	Pae	RpsK-YbeY	30S ribosomal protein S11	Endoribonuclease YbeY
G	Bfr	RimH-RsmH	Ribosome maturation factor RimM	Ribosomal RNA small subunit methyltransferase H
H	Ftu	MnmA-RimP	tRNA-specific 2-thiouridylase MnmA	Ribosome maturation factor RimP
I	Eco	TsaE-PaaX	tRNA threonylcarbamoyl- adenosine biosynthesis protein TsaE	Transcriptional repressor PaaX
J	Aca	Ttg2D-VacJ	Putative toluene tolerance protein	VacJ family lipoprotein
K	Pae	FliA-FlgM	RNA polymerase sigma factor FliA (RNA polymerase sigma factor for flagellar operon)	Negative regulator of flagellin synthesis (Anti-sigma-28 factor)
L	Mtu	DrrC-DrrA	Probable doxorubicin resistance ABC transporter permease protein DrrC	Doxorubicin resistance ATP-binding protein DrrA
M	Pae	MucC-RnfA	Positive regulator for alginate biosynthesis MucC	Ion-translocating oxidoreductase complex subunit A
N	Spn	AMCSP13_001135 AMCSP13_000945	Fibronectin/fibrinogen binding domain protein	DNA gyrase subunit B domain protein
O	Lpn	CcmE-CcmC	Cytochrome c-type biogenesis protein CcmE	Heme exporter protein C
P	Pae	FlgM-FliS	Negative regulator of flagellin synthesis (Anti-sigma-28 factor)	Flagellar secretion chaperone FliSB
Q	Pae	Pa1955-Pa1952	Adhesin	Fimbrial biogenesis outer membrane usher protein
R	Lpn	FeoA-RsfS	Ferrous iron transporter A	Ribosomal silencing factor RsfS
S	Lpn	lpg1546-LptE	Fimbrial biogenesis and twitching motility protein PilF	LPS-assembly lipoprotein LptE
T	Pae	RhlA-LolA	3-(3-hydroxydecanoyloxy) decanoate synthase	Outer-membrane lipoprotein carrier protein
U	Pae	ilvC-PA2823	Ketol-acid reductoisomerase	DUF815 domain-containing protein
V	Pae	PA1065-glpE	DUF488 domain-containing protein	Thiosulfate sulfurtransferase GlpE
W	Pae	PA4431-PA0388	Ubiquinol-cytochrome c reductase iron-sulfur subunit	DUF4426 domain-containing protein
X	Mtu	Rv0883c-FtsZ	Uncharacterized protein	Cell division protein FtsZ
Y	Mtu	RelA-Rv1312	Bifunctional (p)ppGpp synthase/hydrolase RelA	Uncharacterized protein

Predicted interaction pairs displayed in Fig. 3 with uniprot mapping of gene name and protein annotations.