Proteome-scale discovery of protein interactions with residue-level resolution using sequence coevolution

Anna G. Green[1*] and Hadeer Elhabashy[2,7*], Kelly P. Brock[1], Rohan Maddamsetti[1], Oliver Kohlbacher[2,3,4,5,7**], Debora S. Marks[1,6**]

[*] AGG and HE contributed equally to this manuscript. Please cite as Green and Elhabashy *et al.*, or Elhabashy and Green, *et al.*
** Joint corresponding authors (oliver.kohlbacher@uni-tuebingen.de, debbie@hms.harvard.edu)

[1] Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[2] Biomolecular Interactions, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany.
[3] Department of Computer Science, University of Tübingen, WSI/ZBIT, Sand 14, 72076 Tübingen, Germany
[4] Quantitative Biology Center, University of Tübingen, Auf der Morgenstelle 8, 72076 Tübingen, Germany
[5] Institute for Translational Bioinformatics, University Hospital Tübingen, Sand 14, 72076 Tübingen, Germany
[6] Broad Institute of Harvard and MIT
[7] Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14,72076, Tübingen

Abstract

The majority of protein interactions in most organisms are unknown, and experimental methods for determining protein interactions can yield divergent results. Here we use an orthogonal, purely computational method based on sequence coevolution to discover protein interactions at large scale. In the model organism *Escherichia coli*, 53% of protein pairs in the proteome are eligible for our method given currently available sequenced genomes. When assaying the entire cell envelope proteome, which is understudied due to experimental challenges, we found 620 likely interactions and their predicted structures, increasing the space of known interactions by 529. Our results show that genomic sequencing data can be used to predict and resolve protein interactions to atomic resolution at large scale. Predictions and code are freely available at https://marks.hms.harvard.edu/ecolicomplex and https://github.com/debbiemarkslab/EVcouplings

Introduction

A longstanding goal of molecular and structural biology is to determine which proteins interact with each other at residue resolution. Experimental methods have given the research community a rich characterization of many molecular interactions[1-3], but the vast majority of all interactions are undiscovered or not structurally solved. Estimates extrapolated from yeast two-hybrid studies[3] suggest that in the model organism *Escherichia coli*, about one in 1,000 of all possible protein pairs interact (10^4 / 10^7). Less than 40% of the protein interactions in *E. coli* have been experimentally observed[3] and at most 4% have been structurally characterized, even in related species. Since these experimental methods are labor- and cost-intensive, there has been

increased interest in computational approaches, alone or in combination with experiments. Homology modeling, a computational approach that exploits one known interaction to infer the structure of a similar interaction between homologous pairs, fills some of the knowledge gap[4], but the challenge remains for *de novo* prediction from sequence alone when homologous pairs of structures are unavailable.

Recently, approaches such as EVcouplings have successfully leveraged the vast corpus of natural sequences to determine 3D structures, by using probabilistic graphical models to infer coevolved positions. These coupled positions can be sufficient to fold single proteins[5-7] and RNAs[8], without the use of homologous structures, and to predict protein interactions *de novo* [9,10]. Ideally, this approach would extend to determine interactions between any pair of proteins, but previous work has been limited to predicting inter-protein contacts between co-operonic proteins, and therefore was calibrated on and usable only for select, small datasets[9-11]. In this work, we solve the challenges of concatenation and statistical scoring to provide a new method that integrates information from residue-level scores to predict protein interaction **(Fig. 1)**. Residue-level coevolution is then used as restraints for molecular docking to provide atomic-resolution models of protein interactions. As a case study, we highlight 620 interactions involving cell envelope proteins, which are some of the most challenging to determine experimentally, and demonstrate the applicability of our method to 53% of the *E. coli* proteome as of 2019.

Results

E. coli has many undiscovered protein interactions that are amenable to EVcouplings

47 Most protein interactions remain undiscovered or structurally unresolved, even for well-
48 studied model organisms such as *E. coli.* Of the estimated 10,000 protein interactions in *E. coli*,
49 approximately 3,946 have been observed experimentally[3], involving about 50% of *E. coli* proteins,
50 hence leaving a large fraction of interactions unknown (**Fig. 1**). The majority of the protein
51 monomers in the *E. coli* proteome (3,186 out of a total of 4,391) have high-quality monomer
52 alignments and are therefore amenable to EVcouplings analysis (Methods). For the 67% that have
53 an experimental monomer structure (either in *E. coli* or for a homologous sequence from another
54 organism), 78% have reasonable precision of the top ECs (60% for the top L/2 ECs, where L is
55 the protein length) (**Supplementary Table 1, Supplementary Fig. 1**). We therefore restrict our
56 computational predictions of interactions to the set of 3,186 (*i.e.,* 75% of proteins, covering 53%
57 of the total possible interaction space).
58
59 **Computing coevolution across proteins**
60 This work addresses two main challenges in determining large-scale interactions across a
61 whole genome. First, to avoid the use of operonic structure and to test *any* protein pair, we
62 constructed alignments of pairs of proteins from different organisms by identifying the nearest
63 orthologs using reciprocal best hits[12] (Methods). While the "reciprocal best hit" approach has many
64 assumptions, it errs on the conservative side – we are more likely to get false negatives than
65 positives – as inter-protein coupling scores are lower when pairings are incorrect. Second, the
66 challenge of identifying true interactions is compounded by their relative sparsity compared to all
67 possible pairs – an estimated 0.1% of all pairs of proteins have a direct interaction (**Fig. 1**).
68 Therefore, even small false positive rates will result in large numbers of predicted interactions that
69 are false.

To minimize the number of false positive interactions, we trained and calibrated on both positive and negative benchmarks: a 'gold standard' non-redundant set of 560 interactions with known structure, and a set of 2,500 non-redundant "non-interacting" protein pairs with no signal for interaction in yeast two-hybrid experiments (**Supplementary Table 2**, Methods). Evolutionary couplings (ECs) computed across these 560 pairs can recall correct interactions, at the expense of a large number of false positives from the set of 2,500 non-interacting protein pairs (**Fig. 2A**). The accuracy of the model increases when we include other structure-agnostic features in a logistic regression: EC score, sum of intra-protein EC coupling constraints, rank of inter-EC relative to intra-EC pairs, sequence conservation, and hydrophilicity (**Supplementary Table 3,** Methods). Specifically, if we consider a raw EC score threshold that limits our false positive rate to 0.1%, we recover only 12.5% of interacting complexes in the benchmark set. By contrast, the simple logistic regression score achieves a recall of 21% of true interacting complexes with a false positive rate of 0.1%. When the structure of at least one of the monomers is known, the logistic regression model achieves a recall of 26% after incorporation of accessible surface area and precision of intra-molecular evolutionary couplings as features (**Supplementary Fig. 2A**). Our model outperforms yeast two hybrid experiments, achieving a recall of 35% versus the 29% obtained experimentally at a false positive rate of 1%[3].

The logistic regression model also outperforms raw ECs at detecting residue contacts when the interaction of the protein pairs is presumed known. Considering only the protein pairs in the positive benchmark set, at a score threshold that gives a precision of 80% of true inter-protein ECs, we recover 54% of the true ECs in our dataset using just raw EC scores, 65% of true ECs with the logistic regression model, and 78% with the logistic regression model when incorporating features of known monomer structures (**Fig. 2D, Supplementary Fig. 2B**).

93

94 **Prediction of membrane protein interactions**

95 Since the cell membrane contains many protein interactions essential for life, but is
96 notoriously difficult to study experimentally[13], we targeted the cell envelope proteome for detailed
97 analysis. We based our analysis on 1,583 proteins previously described as localized to the *E. coli*
98 cell envelope[1], constituting ~1.25 million possible interacting pairs. We assayed each compartment
99 of the *E. coli* cell envelope proteome with itself and with adjacent compartments (Methods) (**Fig.**
100 **3A**), for a total of 939,159 protein pairs. After monomer sequence alignment, concatenation, and
101 EC calculation, 191,999 proteins pairs comprised of 1,053 proteins (566 non-redundant protein
102 families) pass quality thresholds for analysis. The majority (771) of these proteins are inner
103 membrane proteins. 49% of these proteins have at least a partial structure known of itself or a
104 homologous protein.

105 We predict 529 not previously known protein interactions in the *E. coli* cell envelope as
106 well as the details of their interacting residues **(Fig. 3B, Supplementary Table 4, Supplementary**
107 **Fig. 3)**. Despite having chosen a stringent false positive rate threshold of 0.1%, we expect a
108 minority (30%) of these interactions may be false positives due to the large size of the tested space.
109 Indeed, in our negative assessment dataset of 12,333 protein pairs in the membrane unlikely to
110 interact, we detect interaction for 57 (0.5%). Of 171 independent families of protein interactions
111 with known interaction and known crystal structure that can be analyzed by our method, we
112 recover interaction signal for 29 of them, consistent with recall at this threshold from our
113 calibrations.

114 For protein interactions that have previously been observed experimentally, we report
115 significant residue-level coevolution for 450 pairs (at least 3 residue pairs above the 80% precision

116 threshold in calibration). These predictions can be used to disambiguate directly interacting
117 proteins in complexes identified via AP-MS, such as the Tol-Pal complex reported below. 91 of
118 these protein pairs would have been blindly identified as interacting by our stringent scoring
119 system.

120 The protein interactions in the *E. coli* membrane with the highest interaction score are
121 involved in essential cellular processes; for instance, the RodA and penicillin-binding protein cell
122 wall polymerases, which we had previously characterized in a detailed study[14], as well as
123 components of the electron transport chain, components of the ATP synthase, and components of
124 the bacterial flagellum.

125

126 **Atomic resolution models of membrane protein complexes**

127 The inter-protein residue pairs identified by our method can be used as restraints for
128 molecular docking, to resolve the 3D structure of protein complexes[9,10], (**Fig. 3C-E).** Here, we
129 provide detailed resolution and characterization of newly structurally resolved protein complexes
130 that score highly in our interaction model.

131 We provide structural models for high-scoring, previously structurally uncharacterized
132 interacting pairs. We also resolve the MlaC–MlaD proteins **(Fig. 3F)**, which are involved in inner
133 membrane phospholipid transport, and whose interaction was previously observed
134 experimentally[1]. We resolve the interface for the newly discovered interaction between MppA and
135 MltG **(Fig. 3G)**, which are involved in murein synthesis in the cell wall, and between the
136 peptidoglycan biosynthesis protein DacC and the peptidoglycan polymerase Pbp2 **(Fig. 3H)**.
137 While this docked model only satisfies one of five inter-molecular restraints, it is possible that
138 movement by the Pbp2 protein during synthesis may change the configuration of the interface.

139 Our method resolves the interaction interface between the flagellar proteins FlgL and FlgK,
140 which form the junction between the flagellar filament and flagellar hook[15]. Extrapolating from
141 our docked model of the monomers and a previously produced model of the FlgK ring from
142 *Campylobacter jejuni,* we infer the configuration of an 11-mer ring of FlgL inside an 11-mer ring
143 of FlgK[16] **(Fig. 4, Supplementary Fig. 4)**.

144 We constructed an atomic model of the Tol/Pal system (**Fig. 5**), a protein machine that
145 spans the inner and outer membranes of Gram-negative bacteria. These proteins may play a role
146 in membrane constriction during bacterial cell division[17,18]. For the known interacting interface
147 between TolB and Pal[19], we accurately recapitulate the known structure (PDB ID 2HQS). We
148 predict interactions and residue contacts for the known interaction between TolA and TolB[3,20] as
149 well as for the interactions of TolA, TolR, and TolQ in the membrane[21]. The protein CpoB interacts
150 with the Tol system and the cell division septum site[18], and here we provide a molecular model of
151 its interaction with TolB.

152

153 **Discussion**

154 In this work, we demonstrate a method for inferring protein interactions at residue-level
155 detail from genomic sequences alone. We provide a probabilistic scoring method, calibrated on
156 large, non-redundant datasets, which can be used to determine the probability of interaction of two
157 proteins. If the protein interaction is known, or inferred by the protein interaction score, the residue
158 interaction score can be used directly to determine residues in contact at the interface. These
159 methods can be freely accessed via our webserver or by downloading the source code [22]
160 (https://github.com/debbiemarkslab/EVcouplings).

161 Our method assumes that protein interactions are conserved across species, an assertion
162 which is certainly not true for all complexes and will lead to false negatives in the case of newly
163 evolved or poorly conserved interactions. In addition, any method to predict physical contacts in
164 protein interactions must be calibrated on known structures, and therefore may inherit biases from
165 the set of proteins that have been structurally characterized. In particular, crystal structures are
166 known to be biased against membrane proteins and transient interactions, which could lead to false
167 negative predictions.

168 With current sequence databases, we are able to apply this method to 53% of the protein
169 pairs in the *E. coli* proteome. By using our best reciprocal criterion, we limit ourselves to one
170 ortholog per genome, which limits the amount of sequence diversity that can be found for families
171 with multiple paralogs per genome. Growth in sequence databases and development of fast
172 concatenation techniques to handle paralogous sequences will increase the number of protein pairs
173 that can be analyzed.

174 Method development using coevolution to detect protein-protein interactions has thus far
175 been tested on a handful of well-studied protein-protein interactions[11,23-25], reaffirming our ability
176 to predict well-conserved, well-characterized protein interactions. While there are many possible
177 avenues for continuing to improve coevolution-based prediction of protein interactions, we suggest
178 that future development should begin with large, non-redundant biological datasets such as ours.
179 By identifying and diagnosing cases where existing methods fail on real datasets, new
180 developments will lead to greater opportunities to discover unknown protein interactions.

181

182
183 **Works Cited**

184 1 Babu, M. et al. Global landscape of cell envelope protein complexes in Escherichia coli.
185 Nature biotechnology **36**, 103 (2018).
186 2 Hu, P. et al. Global functional atlas of Escherichia coli encompassing previously
187 uncharacterized proteins. PLoS Biol **7**, e96, doi:10.1371/journal.pbio.1000096 (2009).
188 3 Rajagopala, S. V. et al. The binary protein-protein interaction landscape of Escherichia
189 coli. Nat Biotechnol **32**, 285-290, doi:10.1038/nbt.2831 (2014).
190 4 Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. & Schwede, T. Modeling protein
191 quaternary structure of homo- and hetero-oligomers beyond binary interactions by
192 homology. Sci Rep **7**, 10480, doi:10.1038/s41598-017-09654-8 (2017).
193 5 Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation.
194 PLoS One **6**, e28766, doi:10.1371/journal.pone.0028766 (2011).
195 6 Hopf, T. A. et al. Three-dimensional structures of membrane proteins from genomic
196 sequencing. Cell **149**, 1607-1621, doi:10.1016/j.cell.2012.04.012 (2012).
197 7 Ovchinnikov, S. et al. Large-scale determination of previously unsolved protein
198 structures using evolutionary information. Elife **4**, e09248, doi:10.7554/eLife.09248
199 (2015).
200 8 Weinreb, C. et al. 3D RNA and Functional Interactions from Evolutionary Couplings.
201 Cell **165**, 963-975, doi:10.1016/j.cell.2016.03.030 (2016).
202 9 Hopf, T. A. et al. Sequence co-evolution gives 3D contacts and structures of protein
203 complexes. Elife **3**, doi:10.7554/eLife.03430 (2014).
204 10 Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-
205 residue interactions across protein interfaces using evolutionary information. Elife **3**,
206 e02030, doi:10.7554/eLife.02030 (2014).
207 11 Feinauer, C., Szurmant, H., Weigt, M. & Pagnani, A. Inter-Protein Sequence Co-
208 Evolution Predicts Known Physical Interactions in Bacterial Ribosomes and the Trp
209 Operon. PLoS One **11**, e0149166, doi:10.1371/journal.pone.0149166 (2016).
210 12 Zeng, H. et al. ComplexContact: a web server for inter-protein contact prediction using
211 deep learning. Nucleic Acids Res **46**, W432-W437, doi:10.1093/nar/gky420 (2018).
212 13 Díaz-Mejía, J. J., Babu, M. & Emili, A. Computational and experimental approaches to
213 chart the Escherichia coli cell-envelope-associated proteome and interactome. FEMS
214 Microbiology Reviews **33**, 66-97, doi:10.1111/j.1574-6976.2008.00141.x (2008).
215 14 Sjodt, M. et al. Structure of the peptidoglycan polymerase RodA resolved by
216 evolutionary coupling analysis. Nature **556**, 118-121, doi:10.1038/nature25985 (2018).
217 15 Homma, M., DeRosier, D. J. & Macnab, R. M. Flagellar hook and hook-associated
218 proteins of Salmonella typhimurium and their relationship to other axial components of
219 the flagellum. Journal of Molecular Biology **213**, 819-832,
220 doi:https://doi.org/10.1016/S0022-2836(05)80266-9 (1990).
221 16 Bulieris, P. V., Shaikh, N. H., Freddolino, P. L. & Samatey, F. A. Structure of FlgK
222 reveals the divergence of the bacterial Hook-Filament Junction of Campylobacter.
223 Scientific Reports **7**, 15743, doi:10.1038/s41598-017-15837-0 (2017).
224 17 Gerding, M. A., Ogata, Y., Pecora, N. D., Niki, H. & De Boer, P. A. The trans-envelope
225 Tol–Pal complex is part of the cell division machinery and required for proper outer-
226 membrane invagination during cell constriction in E. coli. Molecular microbiology **63**,
227 1008-1025 (2007).

18 Gray, A. N. et al. Coordination of peptidoglycan synthesis and outer membrane constriction during Escherichia coli cell division. elife **4**, e07118 (2015).
19 Bonsor, D. A. et al. Allosteric β-propeller signalling in TolB and its manipulation by translocating colicins. The EMBO journal **28**, 2846-2857 (2009).
20 Walburger, A., Lazdunski, C. & Corda, Y. The Tol/Pal system function requires an interaction between the C-terminal domain of TolA and the N-terminal domain of TolB. Molecular microbiology **44**, 695-708 (2002).
21 Lloubès, R. et al. The Tol-Pal proteins of the Escherichia coli cell envelope: an energized system required for outer membrane integrity? Research in microbiology **152**, 523-529 (2001).
22 Hopf, T. A. et al. The EVcouplings Python framework for coevolutionary sequence analysis. biorxiv, doi:10.1101/326918 (2018).
23 Bitbol, A. F., Dwyer, R. S., Colwell, L. J. & Wingreen, N. S. Inferring interaction partners from protein sequences. Proc Natl Acad Sci U S A **113**, 12180-12185, doi:10.1073/pnas.1606762113 (2016).
24 Gueudre, T., Baldassi, C., Zamparo, M., Weigt, M. & Pagnani, A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. Proc Natl Acad Sci U S A **113**, 12186-12191, doi:10.1073/pnas.1607570113 (2016).
25 Bitbol, A. F. Inferring interaction partners from protein sequences using mutual information. PLoS Comput Biol **14**, e1006401, doi:10.1371/journal.pcbi.1006401 (2018).
26 Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. J Mol Biol **372**, 774-797, doi:10.1016/j.jmb.2007.05.022 (2007).
27 UniProt Consortium, T. UniProt: the universal protein knowledgebase. Nucleic Acids Res **46**, 2699, doi:10.1093/nar/gky092 (2018).
28 Velankar, S. et al. SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. Nucleic Acids Res **41**, D483-489, doi:10.1093/nar/gks1258 (2013).
29 El-Gebali, S. et al. The Pfam protein families database in 2019. Nucleic Acids Res **47**, D427-D432, doi:10.1093/nar/gky995 (2019).
30 Berman, H. M. et al. in International Tables for Crystallography Volume F: Crystallography of biological macromolecules 675-684 (Springer, 2006).
31 Eddy, S. R. Accelerated Profile HMM Searches. PLOS Computational Biology **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).
32 Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I. & Langmead, C. J. Learning generative models for protein fold families. Proteins **79**, 1061-1078, doi:10.1002/prot.22934 (2011).
33 Ekeberg, M., Lovkvist, C., Lan, Y., Weigt, M. & Aurell, E. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys **87**, 012707, doi:10.1103/PhysRevE.87.012707 (2013).
34 Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**, 2577-2637, doi:doi:10.1002/bip.360221211 (1983).
35 The nature of the accessible and buried surfaces in proteins. Journal of Molecular Biology **105**, 1 - 12, doi:https://doi.org/10.1016/0022-2836(76)90191-1 (1976).
36 Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. **12**, 2825-2830 (2011).

37 Krivov, G. G., Shapovalov, M. V. & Dunbrack Jr., R. L. Improved prediction of protein side-chain conformations with SCWRL4. Proteins: Structure, Function, and Bioinformatics **77**, 778-795, doi:10.1002/prot.22488 (2009).

38 Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. Journal of the American Chemical Society **125**, 1731-1737, doi:10.1021/ja026939x (2003).

39 Brunger, A. T. Version 1.2 of the Crystallography and NMR system. Nature Protocols **2**, 2728, doi:10.1038/nprot.2007.406 https://www.nature.com/articles/nprot.2007.406#supplementary-information (2007).

Figure 1: Scaling EVcouplings methods to full bacterial genomes
(A) The search problem for binary protein-protein interactions in *Escherichia coli* involves finding all of the estimated 10^4 true interactions out of 10^7 possible pairs. Only approximately 300 of these true direct interactions have structural resolution in any organism. (B) Evolutionary couplings learned directly from protein sequences can resolve interfaces. Sequence alignments of both monomeric proteins are created and concatenated by the best reciprocal hit procedure before inference of evolutionary couplings. Raw evolutionary coupling scores can be combined with features of their distribution, biochemical properties, and sequence entropy to improve inference. (C) A benchmark dataset of all non-redundant protein interactions with known interface structure.

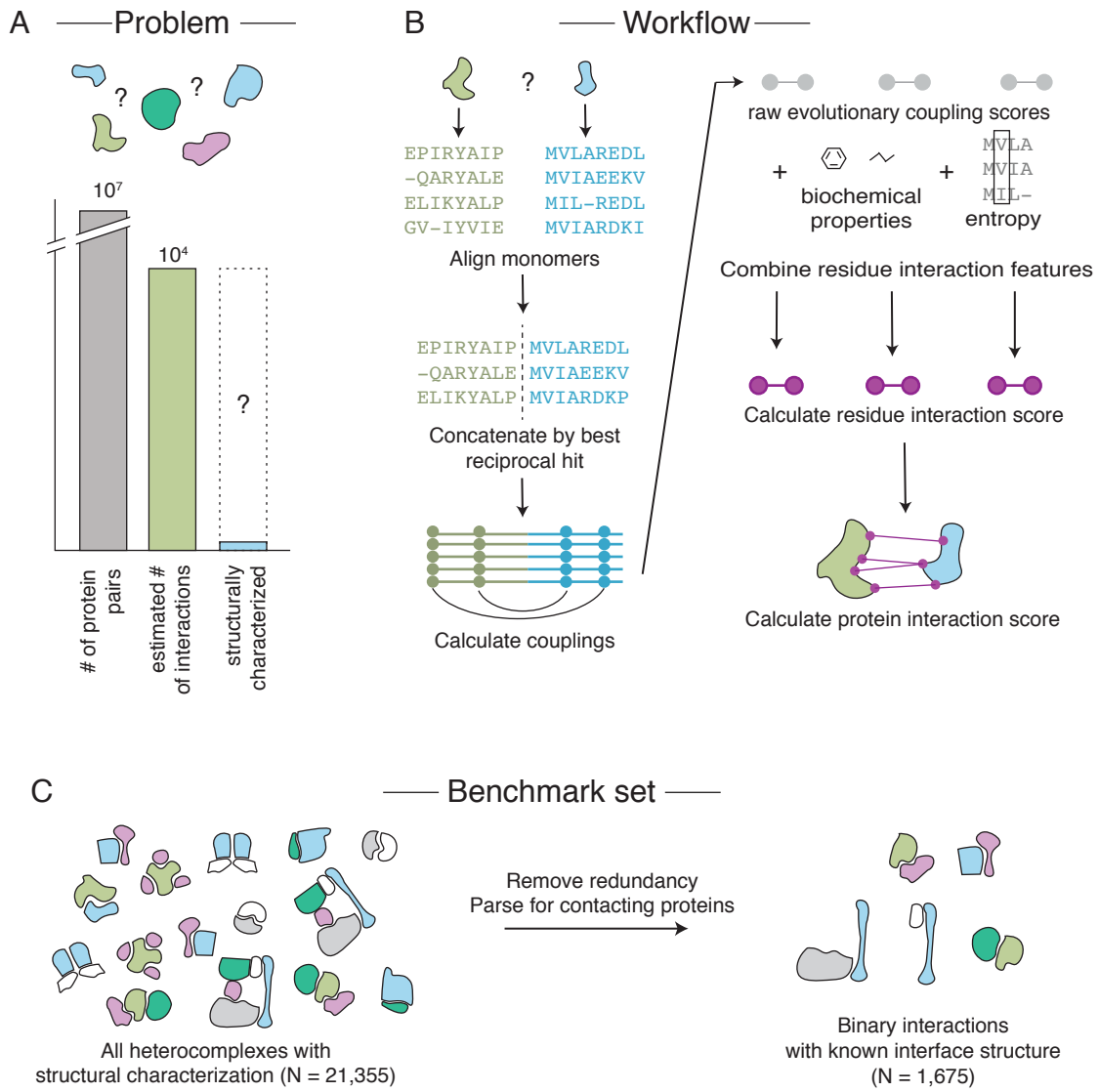


Figure 2: Model predicts interacting residues and interacting proteins with high precision
(A) Our logistic regression model (purple) reduces the false positive rate compared to raw ECs (grey). The x-axis shows the recall on the positive benchmark set of 560 true complexes, and the y-axis shows the false positive rate on the held-out dataset of non-interacting complexes at the score threshold that gives each recall value. (B) Once the interaction of the proteins is known, our model has higher recall of interacting residues than raw ECs. For protein complexes with inferred interaction at a 40% recall cutoff (N=224), the precision versus recall of the top 10 ECs is plotted. ECs are considered true if inter-residue minimum atom > 8Å, and false if > 12Å. (C) Example performance on known interaction between and ABC transporter permease and ATP binding subunit (UniProt IDs Y1470_HAEIN and Y1471_HAEIN, PDB ID: 2NQ2 chains C and A. (D) Example performance on known interaction between and DNA primase PriS and PriL (UniProt IDs PRIS_SULSO and PRIL_SULSO, PDB ID: 1ZT2 chain A and B)

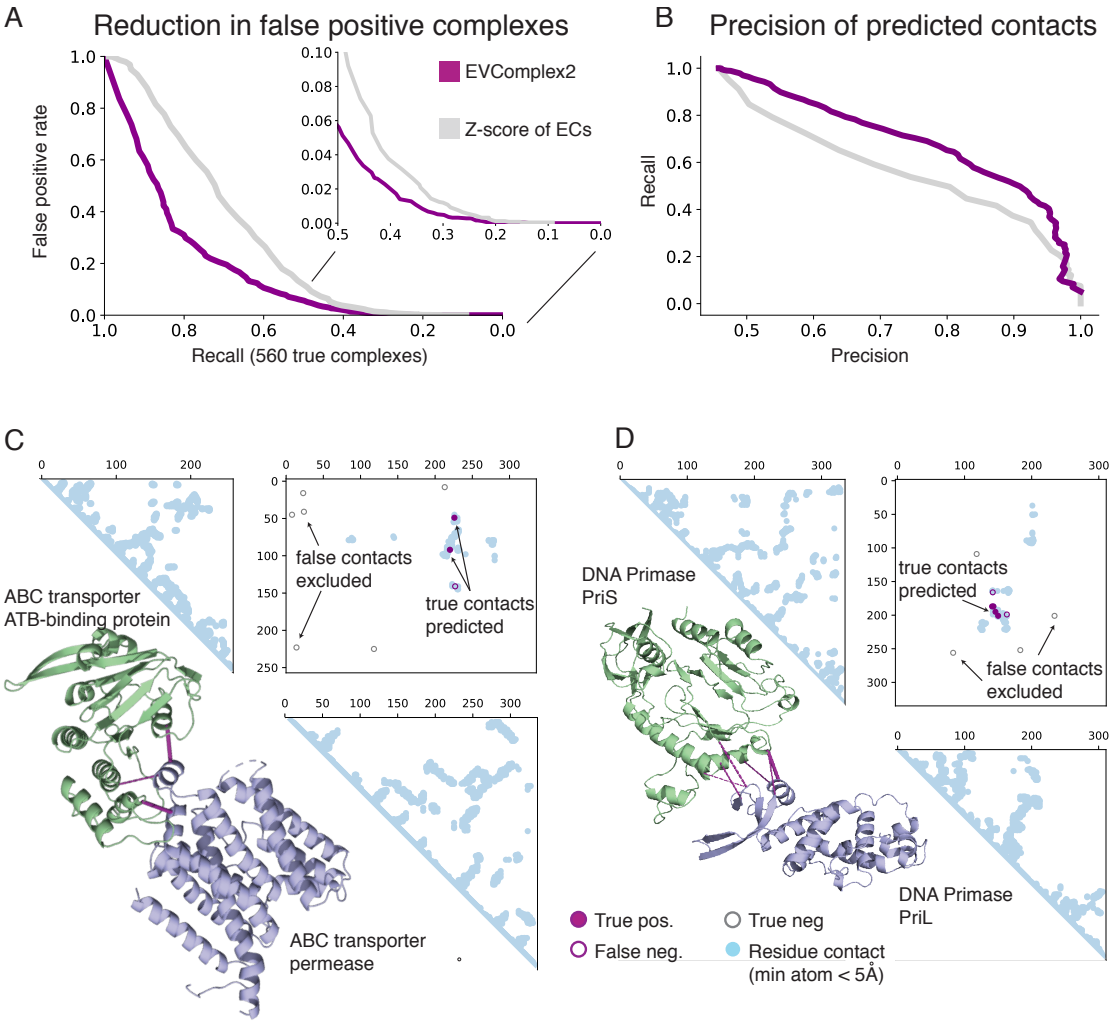


Figure 3: Discovery of hundreds of new interactions in the *E. coli* membrane proteome
(A) We searched a high-value subset of the 10^7 possible interactions in the *E. coli* proteome by searching membrane compartments with themselves and with adjacent membrane compartments. (B) We found 620 high-scoring protein interactions in the cell envelope, including 29 with structural characterization and 92 with previous experimental evidence. (C-E) 3D configurations of previously structurally characterized interactions accurately predicted by molecular docking with inferred restraints. Known structures are shown in grey. RMSDs calculated by comparison to known structures with PDB IDs 3RKO, 2WU2, and 2HQS, respectively. (F-H) Example of three docked models of newly resolved protein complexes: MlaC/MlaD, MppA/MltG, and DacC/Pbp2.

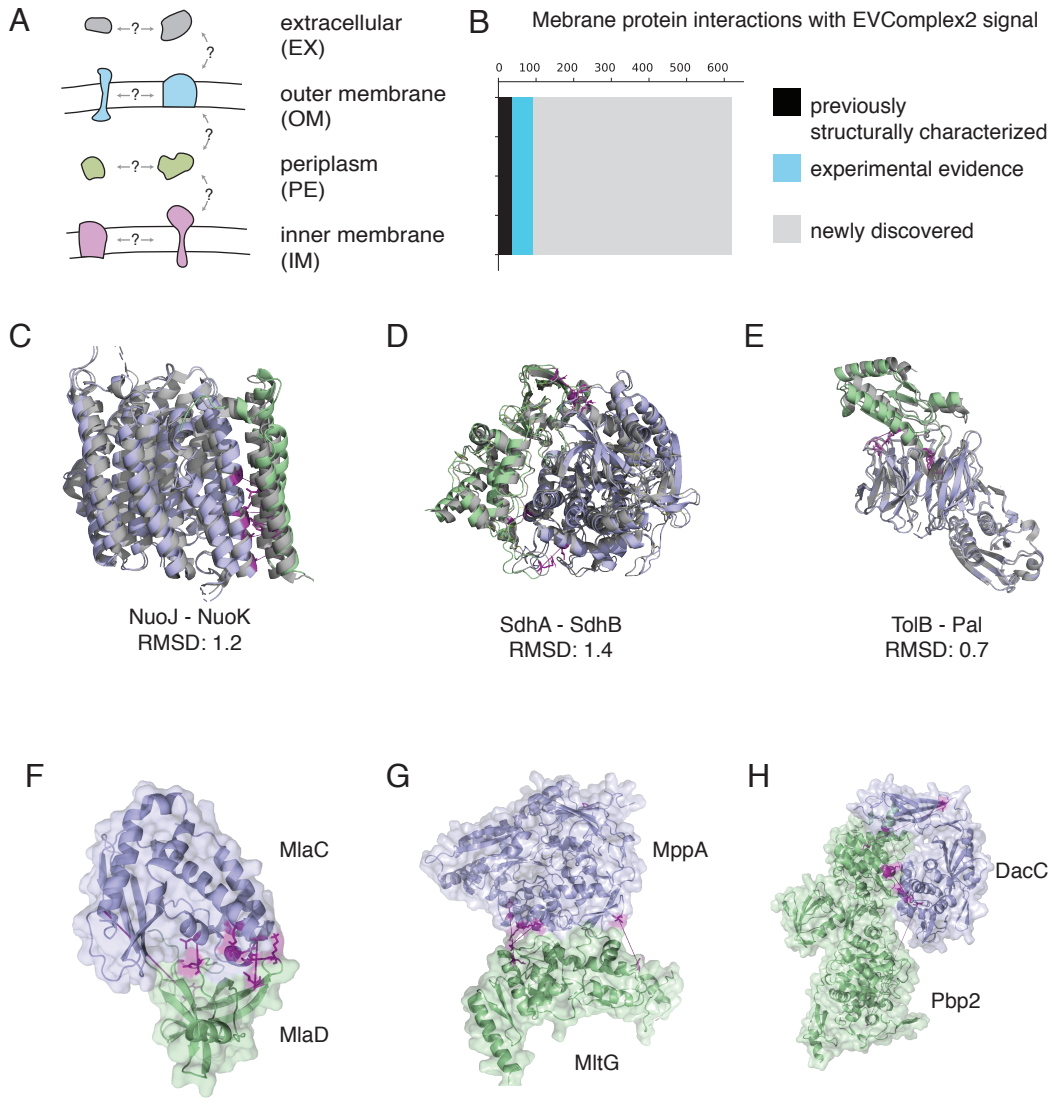


Figure 4: Model of the flagellar hook-filament junction
(A) Schematic of the orientation of the bacterial flagellum. The proteins FlgL (green) and FlgK (blue) form two rings which create the junction between the hook and filament of the flagellum[16]. (B) Docked model of FlgL and FlgK using evolutionary couplings. PDB structures of homologous proteins from *Salmonella typhimurium* were used in docking (PDB IDs 2D4Y and 2D4X, respectively). Predicted interface residues are highlighted in purple. (C) A previously inferred model of the FlgK ring from *Campylobacter jejuni*[16] was used to infer the structure of the entire hook-filament junction. Evolutionary coupled residues (purple) show the interface for FlgL ring insertion into the FlgK ring. By aligning our docked model to the *C. jejuni* ring, we show that an 11-mer ring of FlgL fits inside the FlgK structure.

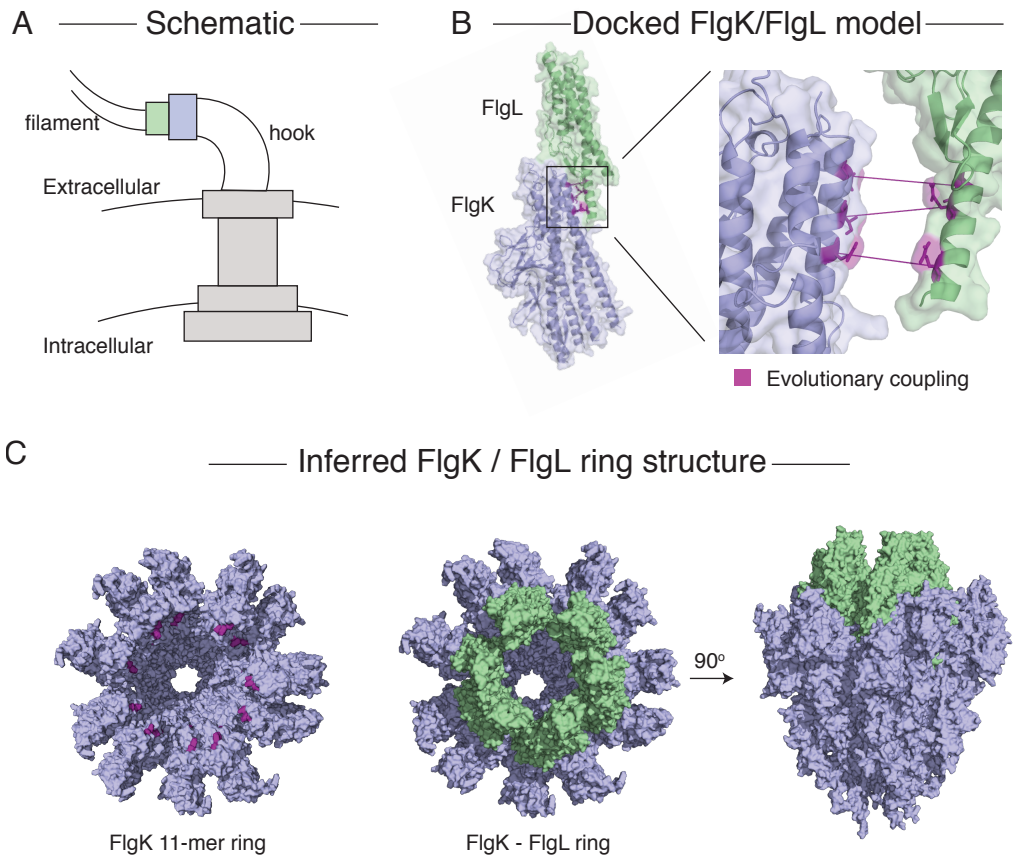
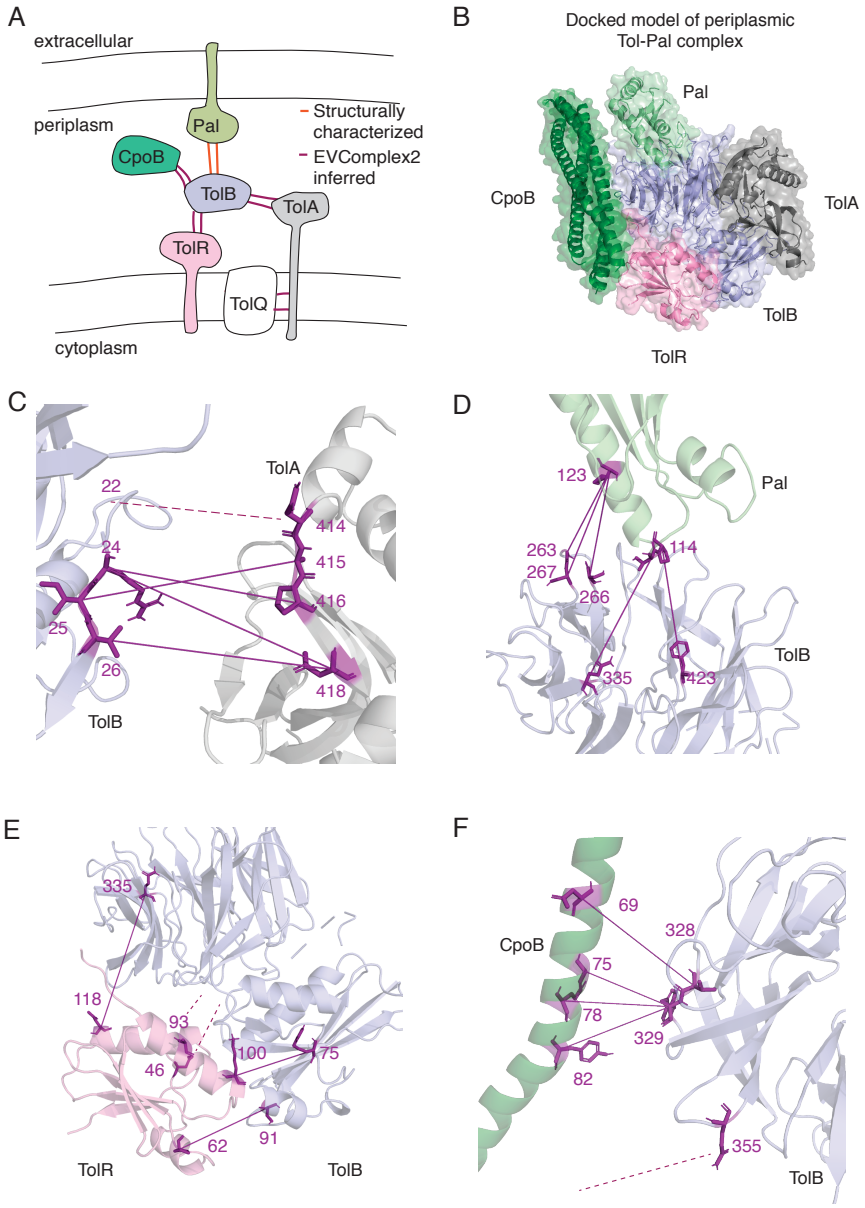


Figure 5: Atomic resolution model of the Tol-Pal complex produced by EVcouplings
(A) Schematic of the proteins involved in the Tol-Pal complex[1] (TolABRQ, CpoB, and Pal). Interactions with previously solved interfaces are shown in orange and interactions inferred by our method are shown in purple. (B-E) Residue resolution of TolB-Pal, TolB-TolA, TolR-TolB, and CpoB-TolB interfaces. The top 5 inferred interface contacts are shown in purple. Dashed lines indicate inferred contacts where one or more residues is missing from the solved structure. Structures used are 1TOL_A (TolA), 2HQS_A (TolB), 5BY4_A (TolR), 2HQS_H (Pal), and 2XDJ_A (CpoB). (F) Complete model of the Tol-Pal complex inferred by aligning results of docked pairwise models. Note that CpoB is inferred to be a trimer *in vivo* but was docked as a monomer for modeling purposes.



Acknowledgements

We thank members of the Marks, Sander, and Kohlbacher labs for discussion and feedback. We thank Dr. Thomas Hopf for input on EVcouplings methodology and Dr. Anastasiya Yakhnina for input about the Tol-Pal complex. Computational resources and support were provided by the Orchestra High Performance Compute Cluster at Harvard Medical School, which is funded by the NIH (NCRR 1S10RR028832-01). AGG was supported by National Science Foundation Graduate Research Fellowship DGE1144152. HE is funded by the International Max Planck Research School (IMPRS) "From Molecules to Organisms."

Note: while preparing this manuscript, Cong et al. published complimentary work, DOI: 10.1126/science.aaw6718

298 **Methods**

299 **Creation of positive benchmark set**

300 A list of all heteromeric structures in the PDBePISA was downloaded[26] (download date:
301 2/19/2018) and passed through the RCSB Protein Data Bank with the following query: "Sequence
302 Length is between 30 and 1200 and Resolution is between 1.0 and 3.0 and Representative
303 Structures at 30% Sequence Identity." IDs were mapped to UniProt [27] identifiers using the SIFTS
304 database[28] (download date 2/20/2018). Only complexes with at least two chains that map to
305 different UniProt identifiers were kept. Single-chain and fusion proteins were removed because
306 they may have non-specific interface contacts not found in nature.

307 For each protein, we extracted the PFAM domains[29] annotated in that protein, which we
308 call the PFAM set for that protein. We then consider a protein-protein interaction as unique if and
309 only if the interacting proteins constitute a pair of PFAM sets not yet seen in our database, yielding
310 4,154 non-redundant pairs of interactions. We then consider only contacting pairs, defined as a
311 protein pair with at least 20 pairs of residues with minimum atom distance < 5Å. Because each
312 unique interaction can be represented by many crystal structures, representative crystals were
313 chosen based on the number of other interfaces present in that crystal – *i.e.*, complexes with more
314 subunits were prioritized. We removed any cases where the two proteins map to the same chain,
315 where the two proteins share any PFAM domain, or where either of the proteins was less than 30
316 amino acids in length, resulting in a final list of 1,675 protein complexes with known interfaces
317 (**Supplementary Table 2)**.

318

319

320

321 **Creation of negative benchmark set**

322 2,500 negative examples were selected randomly from all pairs of complexes with no
323 signal for interaction in a large-scale Y2H experiment in *E. coli*[3]. We verified that no pair of
324 proteins had the same PFAM set as any other pair of proteins in the negative set, and that no pair
325 had the same PFAM set as any pair of proteins in the positive benchmark set **(Supplementary**
326 **Table 2)**. For a negative assessment dataset to evaluate the performance of our predictor on the *E.*
327 *coli* cell envelope, all pairs of proteins where the monomers are found in different AP-MS
328 benchmark complexes were assessed[1].

329

330 **Monomer Sequence Alignment**

331 For alignments of the positive benchmark set, all UniProt identifiers corresponding to the
332 sequence of interest were extracted from PDB[30]. For alignments of the *E. coli* genome, all
333 identifiers from the reviewed *Escherichia coli* (strain K12) reference proteome were downloaded
334 from UniProt[27]. The entire length of every protein was used for sequence alignment. Alignments
335 were constructed using jackhmmer[31] with five iterations against the UniProt database (downloaded
336 February 2018), as implemented in the EVcouplings software package. Alignments were
337 considered to have sufficient coverage if at least 80% of the columns were less than 50% gaps. A
338 range of bitscores were tested (0.1, 0.2, 0.3, and 0.5, and 0.8 times protein length), and the
339 alignment with the largest N_{eff} that had 80% coverage of columns was selected for each protein.
340 Proteins from the *E. coli* genome were deemed eligible for EVcomplex analysis if their monomer
341 sequence alignment had at least 80% column coverage and $N_{eff}/L >= 2.5$, after characterizing the
342 precision of monomer protein couplings across a range of sequence alignment diversities
343 (**Supplementary Fig. 1)**.

344 **Concatenation**

345 Monomer protein sequences were concatenated by extracting the species annotation from
346 the UniProt database. A single representative for each species was chosen based on best reciprocal
347 identity to the query sequence, after excluding closely related paralogs in the query genome with
348 more than 90% sequence identity. If, for a single species, a best reciprocal hit for both monomers
349 can be identified, those hits are concatenated and added to the alignment.

350

351 **Concatenated alignment quality control**

352 To avoid analyzing protein pairs where the two proteins are homologous, we implemented
353 three different quality control metrics. First, we remove all pairs of proteins where the first protein
354 contains a PFAM domain that is found in the second protein. Then, we remove all pairs where our
355 structure comparison protocol found overlapping hits to the same chain of the same PDB structure.
356 Finally, we remove all pairs that display high-scoring coevolution along a diagonal between the
357 two proteins (*i.e,* between position *i* in protein A and the corresponding position *i* in protein B).
358 We consider these contacts to be artifactual due to their very high scores relative to known
359 interactions. In addition, we do not consider proteins from *E. coli* where there is a non-traditional
360 amino acid annotated in the UniProt sequence, because these can indicate pseudogenes.

361 To ensure high-diversity, low gap alignments for inference, we removed from
362 consideration concatenated sequence alignments with N_{eff}/L less than 0.2 or coverage of columns
363 lesser than 80%.

364

365

366

367 **EC calculation**

368 For EVCouplings calculation, hits composed of more than 50% gaps were filtered from the
369 alignment, and sequences with homologs more than 80% identical were downweighed to compute
370 N_{eff}, the effective number of sequences[5]. ECs were calculated using pseudo-likelihood
371 maximization[32,33], as implemented by Weinreb *et al.* (2016)[8]. The λ_J term was scaled by the number
372 of amino acids minus one times the number of sites in the model minus one. Pre- and post-
373 processing was performed using the EVCouplings Python package[22].

374

375 **Comparison of ECs to experimental structures**

376 To identify experimental structures for comparison of evolutionary couplings, the
377 monomer sequence alignment were searched against the PDB database using *hmmsearch*[31]. For *E.*
378 *coli* monomer alignments, the top 20 were selected. For the validation dataset, all hits above a
379 bitscore threshold of 0.2 *L* were selected. Once structures were selected for comparison, the
380 minimum of the minimum atom inter-residue distances across all models were used for comparison
381 to ECs[6].

382

383 **Solvent accessible surface area calculation**

384 Relative and absolute solvent accessibilities were calculated using DSSP[34] with a probe
385 size of 1.4 Å and van der Waals radii from literature[35]. If multiple structural hits were found for
386 each protein, we take the mean relative solvent accessibility across all structures, to avoid biases
387 caused by incomplete structures.

388

389

390 **Conservation calculation**

391 The conservation C for each column i in the concatenated sequence alignments was
392 calculated as one minus the normalized entropy for the column i,

393 $$C = 1 - H/log_2 20$$

394 $$H = -\sum_{k=1}^{20} f_k log_2 f_k$$

395 Where f_k is the frequency of the k-th amino acid in column i.

396

397 **Detecting inter-protein residue contacts**

398 For analysis of raw ECs, the Z-score of each inter-residue EC was calculated using the
399 distribution of all inter-residue ECs.

400 The model for residue interaction was trained on the positive benchmark dataset. Our
401 dataset for includes 560 protein complexes for which the concatenated sequence alignment has
402 sufficient coverage and diversity. Because most of the inter-protein ECs are noise and do not
403 correspond to real contacts[9,10], we take the top 10 inter-protein ECs as input to the model. This
404 resulted in 5,334 residue pairs with a defined minimum atom distance, 26% of which were within
405 8Å.

406 We trained a logistic regression classifier to predict the probability that two residues i and
407 j are in contact (distance < 8Å), given their EC score and additional residue-level features. 25% of
408 the data was held out for testing. Models were trained using the Logistic Regression classifier in
409 Scikit-learn[36], using the liblinear solver and L1 regularization. Regularization weight of 0.05 was
410 chosen to maximize a tradeoff between recall and precision using 10-fold cross validation.

411 The features with non-zero weights were the Z-score of the APC-corrected FN score[5], the
412 conservation of the more conserved for the two residues, the relative rank of the first inter-protein
413 EC, the maximum of the intra-protein EC enrichment for the two residues[6], and the combined
414 hydrophilicity score of the amino acids at those two positions.

415 For a model to be used when monomer structures are known, we repeated the model
416 selection procedure with two additional features: the minimum of the accessible surface area of
417 both residues, and the minimum of the precision of the intra-protein ECs ranked above the current
418 inter-protein EC. These two features were found to be significant in combination with the features
419 found above.

420

421 **Detecting protein interactions**

422 To build a model for determining whether two proteins interact, we combine the score of
423 the residue interaction model for the top ten inter-protein residue pairs using a logistic regression
424 framework. The training dataset was composed of positive and negative examples. Positive
425 examples were protein pairs from the positive benchmark set for which at least three of the top 10
426 inter-protein ECs have distances within 8Å, for a total of 188 complexes. An equal number of
427 negative examples were chosen from the negative benchmark set. 25% of this combined dataset
428 was held out for testing.

429 We trained two models, using the residue interaction score with and without structure-
430 aware features. We also include the relative rank of the highest-scoring inter-protein EC as a
431 feature. Models were trained using the Logistic Regression classifier in Scikitlearn[36], using the
432 liblinear solver and L1 regularization (weight = 0.1), chosen to maximize a tradeoff between recall
433 and precision using 10-fold cross validation.

434

435 **Determination of false positive rates for interaction detection**

436 The protein pairs from the negative benchmark set that were excluded from the training
437 and test set were used to assess the false positive rates expected when the model is applied to large
438 datasets.

439

440 **Cell envelope proteome**

441 Localizations of *E. coli* proteins were extracted from a previous study[1]. Cell envelope proteins
442 were separated into four categories: inner membrane bound (integral or lipoprotein), periplasmic,
443 outer membrane bound (integral or lipoprotein), and extracellular. Proteins annotated as
444 "membrane related" but without a specific location were excluded. Each of the four compartments
445 was tested against itself and against proteins in the physically adjacent compartment(s): inner
446 membrane versus periplasm, periplasm versus outer membrane, and outer membrane versus
447 extracellular.

448

449 **Docking**

450 Side chains of all monomers were perturbed and optimized by using SCWRL4[37] to revoke any
451 effect of crystallization in a complex that may bias the docking procedure. Restraints-based
452 docking was done using HADDOCK (High Ambiguity Driven biomolecular DOCKing) v2.2[38]
453 and CNS[39]. This allows restricting the possible interaction search space using a more sophisticated
454 treatment of conformational flexibility. Top five inter-ECs scores in both of structure-aware and
455 structure-free regression models were selected. Unambiguous distance restraints were applied on
456 C-beta (except for glycine, where C-alpha was used) with an effective distance of 5 Å with upper

457 and lower bound of 2 Å. Default HADDOCK protocol was done starting with a rigid-body energy
458 minimization (500 models), followed by semi-flexible refinement in torsion angle space (100
459 models), and ending with further refinement in explicit water (100 models). The explicit flexibility
460 during the molecular dynamics refinement HADDOCK can account for small conformational
461 changes occurring upon binding. The resulting models were scored using the default HADDOCK
462 scoring function, and the top ten best-scoring models, as well as the top models from RMSD
463 clusters, are reported.
464
465