Method Development for Large-Scale Prediction and Modeling of Protein-Protein Interactions

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Hadeer Elsayed
aus Gizeh, Ägypten

Tübingen
2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	31.10.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Oliver Kohlbacher
2. Berichterstatter:	Jun.-Prof. Dr. Andreas Dräger
3. Berichterstatter:	Prof. Dr. Christian von Mering

For Youssef Elsherif

"Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself "

Francis Crick

Abstract

Protein-protein interactions (PPIs), which mediate almost all biological processes, show a high level of complexity and heterogeneity. Despite seven decades of developing experimental and computational techniques, proteome-wide characterization of PPIs remains one of the grand challenges in biosciences. In this cumulative thesis, we address the questions of prediction and three-dimensional modeling of PPIs on the proteome level through four distinct projects.

The first project encompasses an extensive survey of experimental and computational methods that enable, at least in principle, high-throughput PPI characterization at the proteome level. We emphasized their domains of validity, limitations, and different yields over the same set of proteins. We underpinned that the efficient integration between these approaches would provide better insights into the cellular interactome due to their diverse yet complementary outcomes.

In the second project, we addressed the challenges of the evolutionary couplings approach, EVcomplex, for proteome-wide PPI prediction and structure elucidation using a fast alignment concatenation method and a probabilistic inference. This expanded the tractability of the approach five-fold over the *Escherichia coli (E. coli)* proteome and to the eukaryotic interactions. We then employed EVcomplex 2.0 to explore the PPI network of the *E. coli* cell envelope proteome, predicting and resolving hundreds of novel and known membrane PPIs that are challenging to characterize experimentally. We proposed *de novo* structural models of the Flagellar Hook-Filament Junction and Tol/Pal System and demonstrated a successful application on the eukaryotic human spliceosome complex.

The third project focused on exploring the interactome of the endolysosomal compartment using cross-linking mass spectrometry, bioinformatics analysis, and structural modeling. We developed a pipeline for structural evaluation and validation of the experimentally identified cross-links based on their known structures and models. Employing this pipeline, we analyzed the first cross-linking dataset of the endolysosomal compartment of HEK293 cells. We supported our results with validation of newly discovered interactions (e.g., ATP6V1D-FZD9,

GNB4-FLOT2) and structural models of interactions with known biological relevance (e.g., FLOT1-FLOT2 hetero-dimer, PPT1 tetramer).

Finally, we developed a novel framework, XLEC, integrating sparse data from evolutionary coupling (EC) and cross-linking mass spectrometry (XL-MS) for efficient interactome profiling and enhanced structural modeling of PPI. XLEC integrates information from both approaches in restraint-based modeling of complex structures and a subsequent machine learning-based model for interaction prediction. We applied XLEC to data from murine mitochondrial proteomes and showed improvement in its performance upon comparison with the individual approaches. Using XLEC, we generated hundreds of *de novo* PPI models revealing novel structural insights into the mammalian mitochondrial interactome.

The research presented in this cumulative thesis has made a significant stride in the exploration of proteome-wide PPIs. It successfully addressed and resolved challenges of their characterization, offering new insights into the complex landscape of protein interactions, and paving the way for further advancements in the field.

Zusammenfassung

Protein-Protein-Interaktionen (PPIs), die nahezu alle biologischen Prozesse dirigieren, zeigen ein hohes Maß an Komplexität und Heterogenität. Trotz sieben Jahrzehnten der Entwicklung experimenteller und rechnergestützter Techniken bleibt die proteomweite Charakterisierung von PPIs eine der großen Herausforderungen in den Biowissenschaften. In dieser kumulativen Doktorarbeit behandeln wir die Fragen der Vorhersage und dreidimensionalen Modellierung von PPIs auf Proteom-Ebene durch vier verschiedene Projekte, die in den folgenden Absätzen beschrieben werden.

Das erste Projekt umfasst eine umfangreiche Untersuchung experimenteller und rechnergestützter Methoden, die zumindest prinzipiell eine hochdurchsatzfähige Charakterisierung von PPIs auf Proteom-Ebene ermöglichen. Wir haben ihre Anwendungsbereiche, Grenzen und unterschiedlichen Ausbeuten bei derselben Menge von Proteinen hervorgehoben. Wir haben betont, dass eine effiziente Integration dieser Ansätze bessere Einblicke in das zelluläre Interaktom ermöglichen würde, durch die Vereinigung ihrer unterschiedlichen, aber komplementären Ergebnisse.

Im zweiten Projekt haben wir uns mit den Herausforderungen des Ansatzes vom Evolutionary Coupling tool EVcomplex für die proteomweite Vorhersage von PPIs und Strukturaufklärung befasst. Dabei haben wir eine schnelle Methode basierend auf der Konkatenation von Alignments und eine probabilistische Inferenz verwendet. Dies hat die Anwendbarkeit des Ansatzes für das *Escherichia coli (E. coli)*-Proteom und für die eukaryotischen Interaktionen um das Fünffache erweitert. Anschließend haben wir EVcomplex 2.0 verwendet, um das PPI-Netzwerk des *E. coli*-Zellhüllen-Proteoms zu erforschen. Dabei haben wir Hunderte von neuartigen und bekannten Membran-PPIs vorhergesagt und aufgelöst, die experimentell schwer zu charakterisieren sind. Wir haben *de novo* Strukturmodelle des Flagellar Hook-Filament Junction und des Tol/Pal-Systems vorgeschlagen und eine erfolgreiche Anwendung auf das eukaryotische menschliche Spleißosom-Komplex demonstriert.

Das dritte Projekt konzentrierte sich auf die Erforschung des Interaktoms des endolysosomalen Kompartiments unter Verwendung von Crosslinking-Massenspektrometrie, bioinformatischer Analyse und Strukturmodellierung. Wir haben eine Pipeline zur strukturellen Bewertung und Validierung der experimentell identifizierten Crosslinks entwickelt, basierend auf bekannten Strukturen und Modellen. Unter Verwendung dieser Pipeline haben wir den ersten Crosslinking-Datensatz des endolysosomalen Kompartiments von HEK293-Zellen analysiert. Wir haben unsere Ergebnisse durch die Validierung neu entdeckter Interaktionen (z. B. ATP6V1D-FZD9, GNB4-FLOT2) und Strukturmodelle von Interaktionen mit bekannter biologischer Relevanz (z. B. FLOT1-FLOT2-Heterodimer, PPT1-Tetramer) unterstützt.

Schließlich haben wir einen neuartigen Workflow namens, XLEC, entwickelt, der spärliche Daten aus Evolutionary Couplings (EC) und Crosslinking-Massenspektrometrie (XL-MS) zur effizienten Profilierung des Interaktoms und zur verbesserten Strukturmodellierung von PPIs integriert. XLEC kombiniert Informationen beider Ansätze in eine Modellierungsmethode komplexer Strukturen basierend auf Beschränkungen und ein darauf aufbauendes Modell maschinellen Lernens zur Vorhersage von Interaktionen. Wir haben XLEC auf Daten aus murinen mitochondrialen Proteomen angewendet und eine Verbesserung der Leistung im Vergleich zu den einzelnen Ansätzen gezeigt. Unter Verwendung von XLEC haben wir Hunderte von *de novo*-PPI-Modellen generiert, die neue strukturelle Einblicke in das mitochondriale Interaktom von Säugetieren ermöglichen.

Die in dieser kumulativen Doktorarbeit vorgestellte Forschung hat einen bedeutenden Fortschritt in der proteomweiten Erforschung von PPIs erzielt. Sie hat sich erfolgreich mit Herausforderungen in der Charakterisierung von PPIs befasst und sie gelöst, hat neue Erkenntnisse über die komplexe Landschaft der Proteininteraktionen geboten und den Weg für weitere Fortschritte auf diesem Gebiet geebnet.

Acknowledgments

I am deeply grateful to Prof. Dr. Oliver Kohlbacher for the opportunity to pursue my PhD in his lab. His decency, intelligence, and support for my independence and innovative freedom have created an ideal environment for my research. Oliver has empowered me with his guidance, provision of computational resources, and valuable collaborations which set me on the path to success from the very beginning.

I am immensely grateful to Prof. Dr. Andrei Lupas for hosting me in his lab during my PhD. His belief in my abilities and continuous support have been a constant source of motivation. Andrei's open-door policy, consistent availability, and invaluable mentorship have guided me throughout my research journey and will continue to shape my career in many significant ways.

I am truly grateful for my son, Youssef Elsherif, to whom I dedicate this piece of work. The pursuit of my PhD would have been much more challenging, without his understanding, patience, and unconditional love. I am forever indebted for his presence in my life which has been a constant reminder of what truly matters. I am committed to being the best mom I could be for him and to making him proud of the person I have become.

I would like to express my gratitude to my exceptional PhD partners, Dr. Eugen Netz, Dr. Anna Green, and Dr. Jasjot Singh, and my collaborators Dr. Debora Marks and Dr. Dominic Winter. Our collaboration has been nothing short of transformative, as we spent countless hours engaging in deep discussions, collaborating on various projects, and producing impactful research together.

I would like to extend my heartfelt appreciation to my close friends and acquaintances who have made my life during my PhD journey easier and filled it with joy. I am especially grateful to Dr. Nadia Samak, Dr. Amal Kasry, Dr. Layla Hiramatsu, Kyle Williams, Sibylle Patheiger, Zeena Mubarak, Dr. Joe Joiner, Dr. Anasuya Moitra, Danielle Rose, Dr. Ernesto Duran, Dr. Maria Contreras, Dr. Marco Podobnik, Dr. Yevgen Levdanckey, Dr. Alina Stein, Anupam Gautam, Amit Landge and Soubhik Sanyal for their friendship, support and the beautiful memories

we have shared throughout my journey.

Finally, I would like to thank members of Kohlbacher Lab and Lupas Lab in particular, Claudia Walter, Karin Lehmann, Dr. Timo Sachenberg, Dr. Vikram Alva, Dr. Birte Hernandez, Dr. Joana Pereira, Dr. Kyowon Jeong, Dr. Axel Walter, and Dr. Samual Wein for their support, stimulating discussions and feedback on my work. I would also like to thank all the program coordinators and the student representatives at the Max Planck Institute for Biology Tübingen.

General Remarks

- The research conducted for this thesis was carried out as part of The International Max Planck Research School (IMPRS) 'From Molecules to Organisms'.
- In accordance with the standard scientific protocol, I will use the personal pronoun ***we*** to indicate the reader and the writer, or my scientific collaborators and myself.
- We use the symbol '-' to indicate the interaction of molecules in a complex. Thus, the notation FlgL-FlgK would indicate a binary complex between the proteins FlgL and FlgK.
- Computational resources were provided by the Orchestra High Performance Compute Cluster at Harvard Medical School, which is funded by the NIH (NCRR 1S10RR028832-01) and the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

Contents

Abstract **v**

Zusammenfassung **vii**

Acknowledgments **ix**

General Remarks **xi**

1 Introduction **1**
1.1 Protein-protein interaction 1
1.2 Sequence Co-evolution 3
1.3 Cross-linking Mass Spectrometry 7

2 Objectives **11**

3 Results **13**
3.1 High-throughput PPI characterization techniques 14
3.2 Sequences co-evolution and the *Escherichia coli* cell envelope interactome 16
3.2.1 Introduction 17
3.2.2 Results 17
3.2.3 Conclusion 21
3.3 Cross-linking mass spectrometry and the human endolysosomal interactome . . 22
3.3.1 Introduction 23
3.3.2 Results 23
3.3.3 Conclusion 27
3.4 Cross-linking, evolutionary coupling, and the mouse mitochondrial interactome 28
3.4.1 Introduction 28
3.4.2 Results 28
3.4.3 Conclusion 35

4 Conclusion **37**

Bibliography **39**

5 List of Publications **47**

6 Contributions **49**

7 Abbreviations **51**

Appendix **52**

A Paper one: Exploring protein-protein interactions at the proteome level **53**
Article 53
Supplementary 68

B Paper Two: Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences **73**
Article 73
Supplementary 86

C Paper Three: Cross-linking of the endolysosomal system reveals potential flotillin structures and cargo **97**
Article 97
Supplementary 116

D Paper Four: XLEC: Large-scale prediction and modeling of protein-protein interactions by combining sequence co-evolution and cross-linking data **133**
Article 133

Chapter 1

Introduction

1.1 Protein-protein interaction

Proteins are fundamental molecules to the infrastructure, workforce, and sustainability of biological life. Almost all biological processes are mediated by proteins. These proteins do not operate autonomously but rather interact with other molecules such as proteins, lipids, metabolites, and nucleic acids. Protein-protein interaction (PPI) is one of the major types of interaction with over 80% of all proteins estimated to function in complexes with other proteins.[1] PPI can be defined as the direct physical association between two or more protein molecules, involving various residues, secondary structures, and surfaces. This association is mediated by a range of biophysical interactions including electrostatic forces, hydrogen bonding, and hydrophobic effects. The set of all biologically relevant PPIs in a cellular system is referred to as an interactome.

The characterization of a cellular interactome is crucial for our comprehension of the underlying biological processes and essential in facilitating therapeutic intervention under pathological conditions. Nevertheless, interactomes show a high degree of heterogeneity that has been challenging life scientists for decades to fully characterize. Our knowledge about PPIs including which proteins interact with one another and how they do so, is currently limited and fragmentary. There is hardly any three-dimensional information for 60-80% of the known PPIs in humans, bacteria, and yeast.[2] With the massively growing amount of sequence data, the gap in our knowledge about PPIs is widening.

To demonstrate the extent of the complexity of the cellular interactome, let us consider a cellular proteome of N proteins. Such a proteome can potentially form interactions between all $N(N+1)/2$ pairwise interactions. The biologically relevant interactome, albeit unknown in size, is like a needle in a haystack. The *E. coli* proteome (UniProt ID:UP000000625), for

instance, contains 4,448 proteins and thus has a search space of 9,894,576 protein pairs but only an estimated interactome of about 10,000 PPIs.[2] Besides the massive search space, protein interactions show high degrees of spatial, temporal, and conditional diversity.

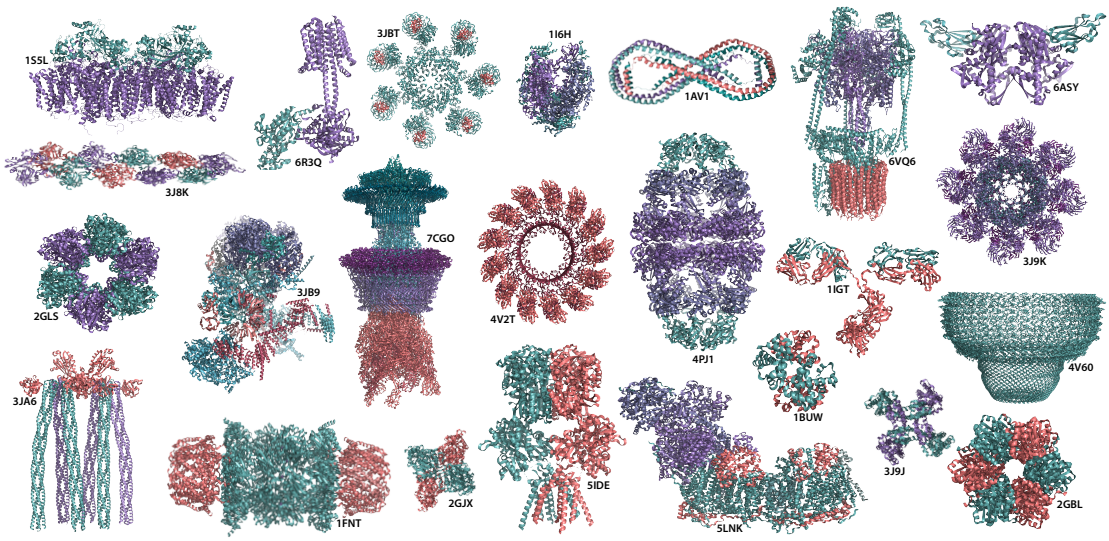


Figure 1.1: Molecular machinery of life.[3] Gallery of supermolecular complexes formed through protein-protein interactions.

Proteins may function individually or interact with one or more copies of identical proteins, different proteins, or both, either simultaneously, at different time points, or conditions. Based on the sequence similarity between the interacting chains, proteins can be involved in homo-interactions and/or hetero-interactions where homo-interactions occur between identical protein chains. It has been estimated that 45% of eukaryotic proteins and 60% of prokaryotic ones are involved in homo-interactions.[4] Based on the individual stability of the interacting proteins, they can form obligate and/or non-obligate interactions, where the subunits of non-obligate interactions are stable on their own, while those of obligate interactions are not. They fold and bind at the same time to form stable complexes. Hence, they are referred to as two-state folders.[5] Protein interactions can be permanent or transient. In permanent interactions, once proteins interact they stay permanently complexed, whilst in transient interactions proteins can associate and dissociate reversibly multiple times.

Proteins can also maintain intrinsically disordered regions that do not form a stable folded structure on their own, but rather can change conformations to adapt to different binding partners. Furthermore, proteins can interact with each other to form architectures of supramolecular machines or mediate a pathway via consecutive interactions (Fig. 1.1). Some proteins need to be expressed in their correct stoichiometry, undergo chemical modification (e.g., phosphory-

lation, glycosylation), conformational change, or get transported to their target compartment to be able to interact with their partners.[6]

Due to this level of complexity and versatility, discovering the complete maps of cellular interactomes remains a challenge to life sciences. Several computational and experimental approaches that have been developed to investigate PPIs show different domains of validity and limitations. A quick glance at these approaches shows that highly confident experimental techniques (e.g., X-ray crystallography, NMR, Cryo-EM) are low-throughput methods. High-throughput experimental approaches (e.g., Yeast Two-Hybrid, Affinity Purification Mass spectrometry, Phage Display, Protein Microarrays) tend to remove the proteins from their native environment, which is essential to the protein structure and functionality. They are often populated by false positives and yield no structure information.[7,8] Computational approaches can offer initial screens on cellular interaction based on genetic features of interacting proteins but still suffer from reliability issues and high false-positive rates.[8,9] Nevertheless, two approaches have attracted our attention due to their abilities to address these PPI characterization challenges, namely Evolutionary Coupling and Cross-linking Mass Spectrometry. In the following sections, we are highlighting these two approaches.

1.2 Sequence Co-evolution

Francis Crick once wrote, "Biologists must constantly keep in mind that what they see was not designed, but rather evolved."[10] This idea was established on the ground that the evolutionary trajectory of a protein through the sequence space is strongly constrained by its structure, function, and interaction.[11,12] Viewing proteins as evolved molecules has unraveled troves of information about them and has been guiding biological research in the last few decades. Sequence co-evolution is one of the main manifestations of such evolutionary pressure in which interacting residues exhibit patterns of correlated mutations during the evolution of a single protein or between interacting proteins.[13–15] For instance, if one protein developed a mutation that hinders an essential interaction from happening, the interacting protein is likely to develop a proximate and compensating mutation to maintain the interaction. Conversely, detecting signals of correlated mutations along the evolutionary history of proteins presents evidence of residue-residue interaction which is referred to as Evolutionary Coupling (EC).[16–18]

Since the inception of this idea, several algorithms have been developed to mine multiple sequence alignments, mimicking the evolutionary history of the proteins, for patterns of correlated mutations and utilize them for structure prediction of proteins and PPIs (Fig. 1.2). Statistical coupling analysis (SCA),[19] mutual information (MI),[20] the observed minus-expected-squared method (OMES),[21] direct coupling analysis (DCA),[22,23] and protein sparse

inverse covariance (PSICOV)[24] are examples of the major evolutionary coupling algorithms.

Pseudo Likelihood Maximization Direct Coupling Analysis (plmDCA)[25,26] has become one of the most efficient algorithms to predict true contacts between co-evolved residues using a global probability model. PlmDCA has been reported to outperform the early local covariance models (e.g., Statistical Coupling Analysis, Mutual Information) that were dominated by phylogenetic and entropic biases and indirect transitive correlations. Furthermore, plmDCA offers a cheap and rapid template-free approach that makes use of the sequence availability and offers structural and functional information about inter-residue contacts within a protein or across the interface of PPI. Both of the EVcomplex (`https://evcomplex.hms.harvard.edu`) and GREMLIN (`http://gremlin.bakerlab.org`) servers offer plmDCA calculations for PPI prediction and differ only in their alignment, ranking, and docking protocols. As EVcomplex was collaboratively developed by Marks Lab at Harvard Medical School and Kohlbacher Lab at the University of Tübingen, where I conducted my PhD research, our emphasis will be on EVcomplex.

The generic workflow of EVcomplex starts with building multiple sequence alignment (MSA) for a pair of query sequences individually and then concatenates them based on matching species. For the resulting concatenated MSA, EVcomplex computes the coupling between every residue pair in a maximum entropy model. To explain how EVcomplex extracts evolutionary couplings from an MSA of protein using plmDCA,[25,26] let us consider that $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_N)$ is the amino acid sequence of a protein of length N. Each σ_i represents the amino acid composition at position i and can assume values in $q = 1, 2, \ldots, 21$ for all 20 standard proteinogenic

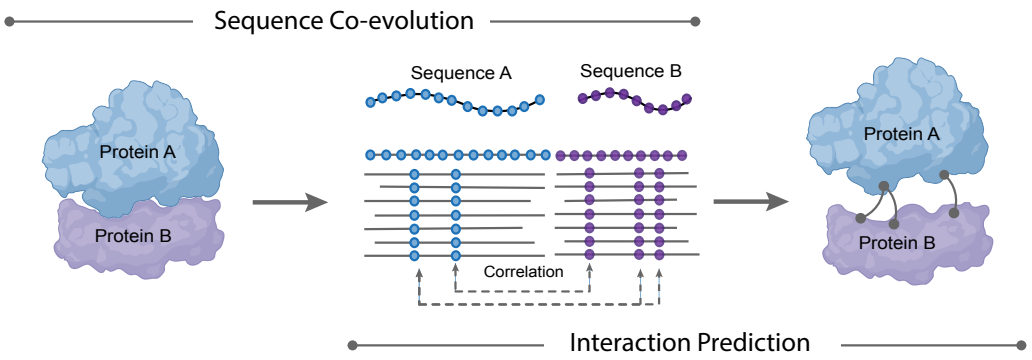


Figure 1.2: Conceptual illustration of sequence co-evolution in protein-protein interaction.

amino acids and the gap character.

Accordingly, an MSA with M sequences can be written as the integer array $\{\sigma^{(m)}\}_{m=1}^{M}$ where each row represents a sequence, and each column represents a position. The empirical frequencies for individual and pairwise positions can be computed as follows.

$$f_i(k) = \frac{1}{M}\sum_{m=1}^{M}\delta(\sigma_i^{(m)}, k) \tag{1.1}$$

$$f_{ij}(k,l) = \frac{1}{M}\sum_{m=1}^{M}\delta(\sigma_i^{(m)}, k)\delta(\sigma_j^{(m)}, l) \tag{1.2}$$

Where $f_i(k)$ is the fraction of sequences that have amino acid k on position i. Similarly, $f_{ij}(k,l)$ is the fraction of sequences that hold amino acid combination (k,l) in the position pair (i,j). $\delta(a,b)$ is the Kronecker Delta function taking values of

$$\delta(a,b) = \begin{cases} 1, & \text{if } a = b, \\ 0, & \text{if } a \neq b. \end{cases}$$

Then the connected correlations can be calculated as

$$c_{ij}(k,l) = f_{ij}(k,l) - f_i(k)f_j(l) \tag{1.3}$$

A generalized Potts model is the simplest probabilistic model $P(\sigma)$ which can reproduce the empirically observed $f_i(k)$ and $f_{ij}(k,l)$.

$$P(\sigma) = \frac{1}{Z}\exp\left(\sum_{i=1}^{N} h_i(\sigma_i) + \sum_{i\leq j\leq N} J_{ij}(\sigma_i, \sigma_j)\right) \tag{1.4}$$

This approach can maximize the probability of the input dataset $\{\sigma^{(m)}\}_{m=1}^{M}$ with respect to the model parameters $h_i(\sigma)$ and $J_{ij}(\sigma_i, \sigma_j)$. The vectors $h_i(\sigma)$ represent the biases of each position i towards specific amino acids, while the matrices $J_{ij}(\sigma_i, \sigma_j)$ describe the coupling or influence between the possible amino acid identities at positions i and j. In order to apply

this model, $h_i(\sigma)$ and $J_{ij}(\sigma_i, \sigma_j)$ must be inferred from the data. However, it is intractable to exactly compute a global partition function Z which is a normalization constant that sums over all possible sequences of length N. Pseudo-likelihood breaks down the exponential time complexity of the global partition function, which sums over all assignments to all nodes, by replacing it with local partition function, which sums over possible assignments to single nodes of the sequence. This approximation makes the couplings significantly easier to compute in general graphical models.

To distill the evolutionary couplings between every two residues in single numbers, we take the Frobenius norm (FN) of the transformed $J_{ij}(\sigma_i, \sigma_j)$ matrices which sums across all 2121 combinations of amino acids $k, l \in q = 1, 2, ..., 21$ in positions i and j. The resulting scores S_{ij}^{FN} are subsequently adjusted for confounding biases, including phylogeny or insufficient sampling, and normalized yielding the final EC score known as the Corrected Norm, S_{ij}^{CN}.

$$S_{ij}^{FN} = ||J_{ij}||_2 = \sqrt{\sum_{k}^{q} \sum_{l}^{q} J_{ij}^{'}(k,l)^2} \tag{1.5}$$

$$S_{ij}^{CN} = S_{ij}^{FN} - \frac{S_{i\cdot}^{FN} S_{\cdot j}^{FN}}{S_{\cdot\cdot}^{FN}} \tag{1.6}$$

Where $\cdot$ represents the row and column averages. This yields a distribution of scores corresponding to the evolutionary couplings between every residue pair within the sequence and usually consists of two main regions (Fig. 1.3.A): I) A normal-like distribution that is highly populated with residue pairs that are not coupled and tend to be distant within the three-dimensional structure. II) A positive right trail of the higher-scored outliers between residue pairs which are coevolving and usually capture three-dimensional proximity. Higher-scored residue outliers capture structural information more accurately than lower-scored outliers. Fig. 1.3.B illustrates an example contact map comparison between the predicted evolutionary coupling contacts and the known complex structures of the methionine-importing transmembrane transporter complex (MetN-MetI). A contact map is a two-dimensional graphical representation where each point represents an amino acid residue pair that is in close proximity, typically around 5 Å, on a three-dimensional structure of a protein or a protein complex.

Nevertheless, a few drawbacks of the EC approach limit its ability for PPI characterization, including: I) Protein interactions that lack sufficient homologous sequences (e.g., orphan

proteins) are untraceable by the approach.[27] II) It is a challenge to select the correct interacting homolog among same-species paralogs. Selecting them based on operon co-localization excludes about 80% of the interaction space in prokaryotes alone and omits eukaryotic interactions due to the lack of operonic structures.[28] III) Interactions with fast evolutionary rates (e.g., transient interactions) are hardly detectable by the approach.[29] IV) It is difficult to differentiate between intra- and inter-chain residue pairs for homo-complexes.

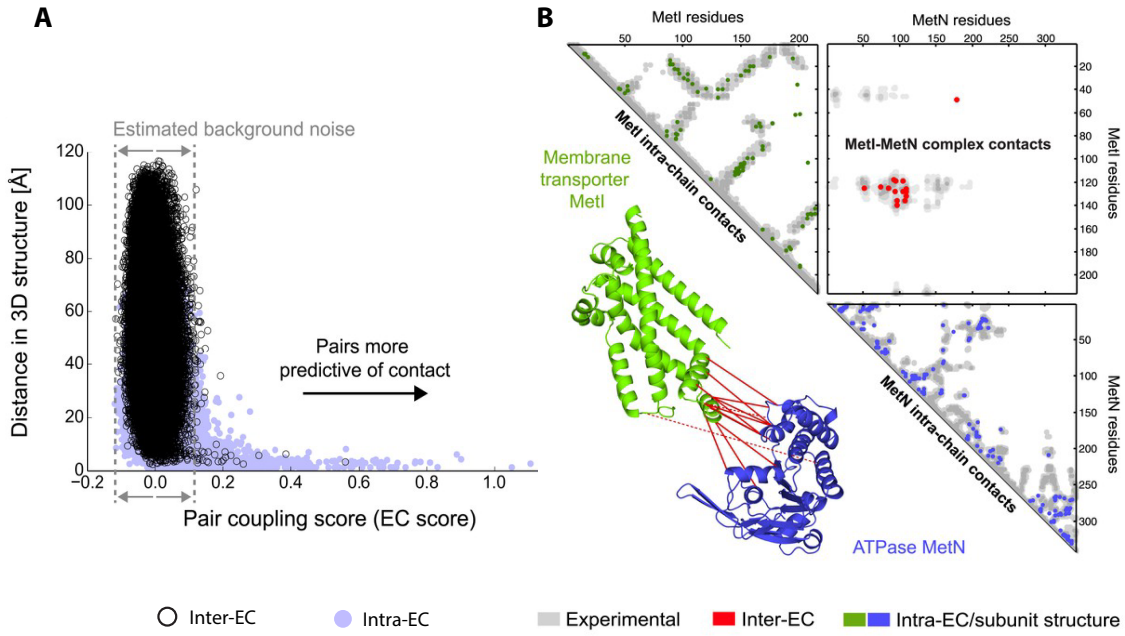


Figure 1.3: (A) Intra and inter EC distribution for the methionine-importing transmembrane transporter heterocomplex from *E. coli* with high coupling scores largely correspond to proximal pairs in 3D. (B) EVcomplex predictions and contact map comparison to the crystal structure (PDB ID: 3tui). Adapted from Hopf et al., 2014.[28]

1.3 Cross-linking Mass Spectrometry

Cross-linking (XL) is an established biochemical labelling technique.[30–32] The emergence of mass spectrometry (MS) technology empowered the approach towards efficient identification of the cross-linked proteins and precise assignment of the linked residues.[33] In cross-linking mass spectrometry (XL-MS), two or more proteins in a complex are covalently cross-linked *in vitro* or *in vivo* (Fig. 1.4). The used cross-linkers are selected based on their chemical reactivity, water solubility, cleavability, membrane permeability, length as well as other criteria. Cross-linked proteins are then subjected to enzymatic digestion resulting in a pool of peptides with a small fraction of cross-linked peptide pairs. Enrichment of these cross-linked peptides

is performed by size exclusion chromatography, ion exchange chromatography, or purification through affinity tags. Cross-linked peptides are then analyzed by mass spectrometry.[34]

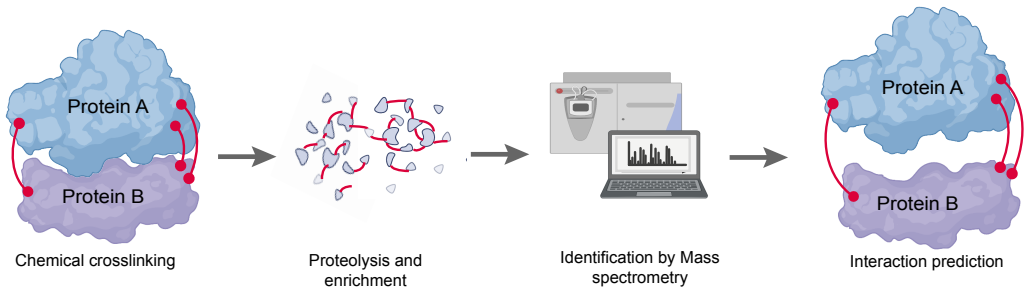


Figure 1.4: Conceptual illustration of Cross-linking Mass Spectrometry in protein-protein interaction.

XL-MS has gained a lot of interest due to its relative simplicity of implementation and reproducibility. For quite some time, XL-MS was successfully applied to study individual purified proteins and protein complexes. Nevertheless, It has recently shown great success in characterizing proteins and PPIs on the proteome level. XL-MS can prove the physical proximity of two interacting proteins even within larger complexes and is not limited to pairwise interactions. It can stabilize and capture information about transient or low-affinity interactions that are difficult to be detected by other approaches.

Additionally, XL-MS can provide structural information about the topological orientation of the interacting proteins reflecting information about cellular architectures and super-molecular machinery in native conditions. It provides an upper bound for the physical distance between the cross-linked residues based on the length of the used cross-linker. These distance restraints can be used for structural modeling of the interactions.

Despite these powerful abilities of XL-MS, it still suffers from some major drawbacks that limit its applicability. The identification of the cross-linked peptides is computationally expensive because of the quadratic expansion of the possible peptide pairing search space with the number of peptides in the database. In addition to the presence of false-positives, XL-MS suffers from a high false-negative rate by capturing mainly the highly abundant proteins and leaving out many present but unidentified cross-linked peptides. There is always a trade-off between sensitivity and specificity in the identification of cross-linked peptides.

Furthermore, XL-MS entails the risk of identifying non-specific interactions since being spatially close does not necessarily imply functionally relevant interactions. Finally, even though XL-MS yields a set of residue-residue distance restraints that can be used for modeling, a few sparse, long-distance ($\sim$ 35 Å) and mainly solvent-accessible restraints might not be sufficient to model the accurate topology of the interaction because it enables a large degree of freedom in the six dimensions of translations and rotation.

Chapter 2

Objectives

The principal objective of this thesis is to address the challenges of prediction and three-dimensional modelling of protein-protein interaction at the proteome level. This cumulative dissertation encompasses four research projects and their corresponding publications that approach these objectives in diverse yet complementary ways. These projects are summarized in the following paragraphs.

The first project aims to I) Survey the currently known high-throughput approaches for PPI characterization. II) Demonstrate their potential, at least in principle, to be scaled up to screen complete proteomes for PPI prediction. III) Highlight their domain of validity, limitation, and diverse outcomes over the same set of proteins. IV) Emphasize the potential integration of these approaches in improving our interactome resolution.

The second project focuses on addressing the challenges of the evolutionary coupling approach, EVcomplex, to enhance its ability to predict and structurally resolve PPIs. This involves I) Providing better concatenation criteria to select the most probable interacting homologs, especially in the presence of same-species paralogs, instead of relying on the co-existence within operonic structures or genome distance that significantly limits the applicability of the approach. II) Developing a more efficient inference method for the prediction of PPI and residue-residue contacts, moving away from the usage of global threshold that exhibits high true-negative rates and doesn't account for the different evolutionary rates of the proteins. Additionally, this project also aims to expand the tractability of the EVcomplex over prokaryotic proteomes and demonstrate its successful application in eukaryotic interactions. Furthermore, we aim to address scaling up EVcomplex to the proteome level screens and assessing its performance when applying to *E. coli* cell envelope proteome.

The third project aims to leverage cross-linking mass spectrometry (XL-MS) for the large-scale discovery of PPI under near-native conditions. Through a collaborative effort with Winter Lab at the Institute for Biochemistry and Molecular Biology (IBMB), we obtained the first cross-linking dataset of the endolysosomal compartment of HEK293 cells. We took on ourselves the responsibility of the sequence and structural bioinformatics analysis of this dataset. In the absence of similar tools, one of the main objectives of this project is to develop pipelines that accept the cross-linking data as input and perform two main functions: I) Structural evaluation and validation of the identified cross-linking data on their known structures and/or AlphaFold models, if existing. II) Three-dimensional modelling of the identified interactions either using docking or AlphaFold-Multimer.

The final project of this dissertation aims to develop an integrative framework, XLEC, that combines sparse data from cross-linking mass spectrometry (XL-MS) and evolutionary coupling (EC) for improved interaction prediction and modelling of PPI. To assess the efficiency of XLEC to tackle these questions, we strive to conduct a comparative analysis of the individual approaches and the machine learning approach, AlphaFold-Multimer. Additionally, this study intends to show a successful application of XLEC on a large-scale cross-linking dataset from mammalian mitochondria and provide novel models and structural insights into its interactome.

Chapter 3

Results

The work of this thesis is addressing the challenges of characterizing PPIs and structurally resolving their three-dimensional complexes on the large scale of proteomes. To accomplish this, we employed methods such as pipeline development, bioinformatics analysis, and structural modeling as well as investigating complexes of substantial biological importance. In this chapter, we outline the research of this cumulative thesis and discuss its corresponding results.

To begin, we provide a comprehensive review of the high-throughput experimental and computational approaches for the characterization of PPIs on the proteome level. Subsequently, we further develop the pipeline and the inference method of the evolutionary coupling approach, EVcomplex 2.0, for enhanced prediction and structure elucidation of PPIs. We showcase the successful application of EVcomplex 2.0 in exploring the protein interaction network of the *E. coli* cell envelope proteome.

Furthermore, we present and investigate the first LC-MS/MS experiment of the endolysosomal compartment of HEK293 cells identifying hundreds of novel interactions and their structural features under semi-native conditions. For this, we developed a pipeline for the structural evaluation and validation of these identified cross-links on known structures and/or AlphaFold models. Additionally, we established a pipeline for large-scale structural modeling of the identified interactions.

Lastly, we introduce an integrative framework, XLEC, that combines data from cross-linking experiments and evolutionary coupling for large-scale prediction and structure modeling of PPIs. We demonstrate its enhanced performance when compared to individual approaches of cross-linking and evolutionary coupling. We demonstrate the successful application of XLEC on the mouse mitochondrial interactome revealing novel structural insights into its PPIs.

3.1 High-throughput PPI characterization techniques

Parts of this section have also been published in the following review article:
Elhabashy H., Merino, F., Alva, V., Kohlbacher, O. and Lupas, A.N., 2022. Exploring protein-protein interactions at the proteome level. *Structure, 30*, p.462.

Given the essential role of PPIs in a wide range of biological processes and diseases, the ability to comprehensively characterize these PPIs is crucial for understanding how biological systems function and how to therapeutically intervene, shall the need appear. Various experimental and computational approaches have made it possible to systematically investigate PPIs at a large scale of proteomes. In this review article (Appendix A), we provide an overview of the currently known medium- to high-throughput approaches for characterizing PPIs at the proteome level. Fig. 3.1 shows conceptual illustrations of such computational and experimental techniques. We discussed their domain of validity, limitations, and unique outcome over the same set of proteins in a proteome. We also shed light on the benefits of higher interactome coverage and resolution upon their potential integration.

Despite the early discovery of biological interactions such as the antibody-antigen[35] and the trypsin-inhibitor,[36] the investigation of physical interactions between proteins played a minor role throughout the early 20th century.[37] It was lagging behind the topic that became the centre of attention in the research of biochemistry, the structure of proteins. With the accumulation of evidence about their essentiality in biological life in the 1940s, the physical associations between proteins started to gain more interest and grew alongside advances in technological and computational power.

The emergence of whole-genome sequencing and proteomics techniques in the 1990s allowed for systematic analysis of PPIs on the proteome-level. However, as Sydney Brenner once wrote at the beginning of the Human Genome Project: "Getting the sequence will be the easy part as only technical issues are involved. The hard part will be finding out what it means because this poses intellectual problems of how to understand the participation of the genes in the functions of living cells."[38]

Despite more than seven decades of technological and computational advances, discovering the complete maps of cellular interactomes remains a challenge to life scientists. This can be attributed to their high degrees of spatial, temporal, and conformational heterogeneity. We categorized the PPI characterization approaches into two groups, experimental and

computational (Appendix A). Nevertheless, this categorization might seem arbitrary due to the interdependency of the experimental approaches and computations.

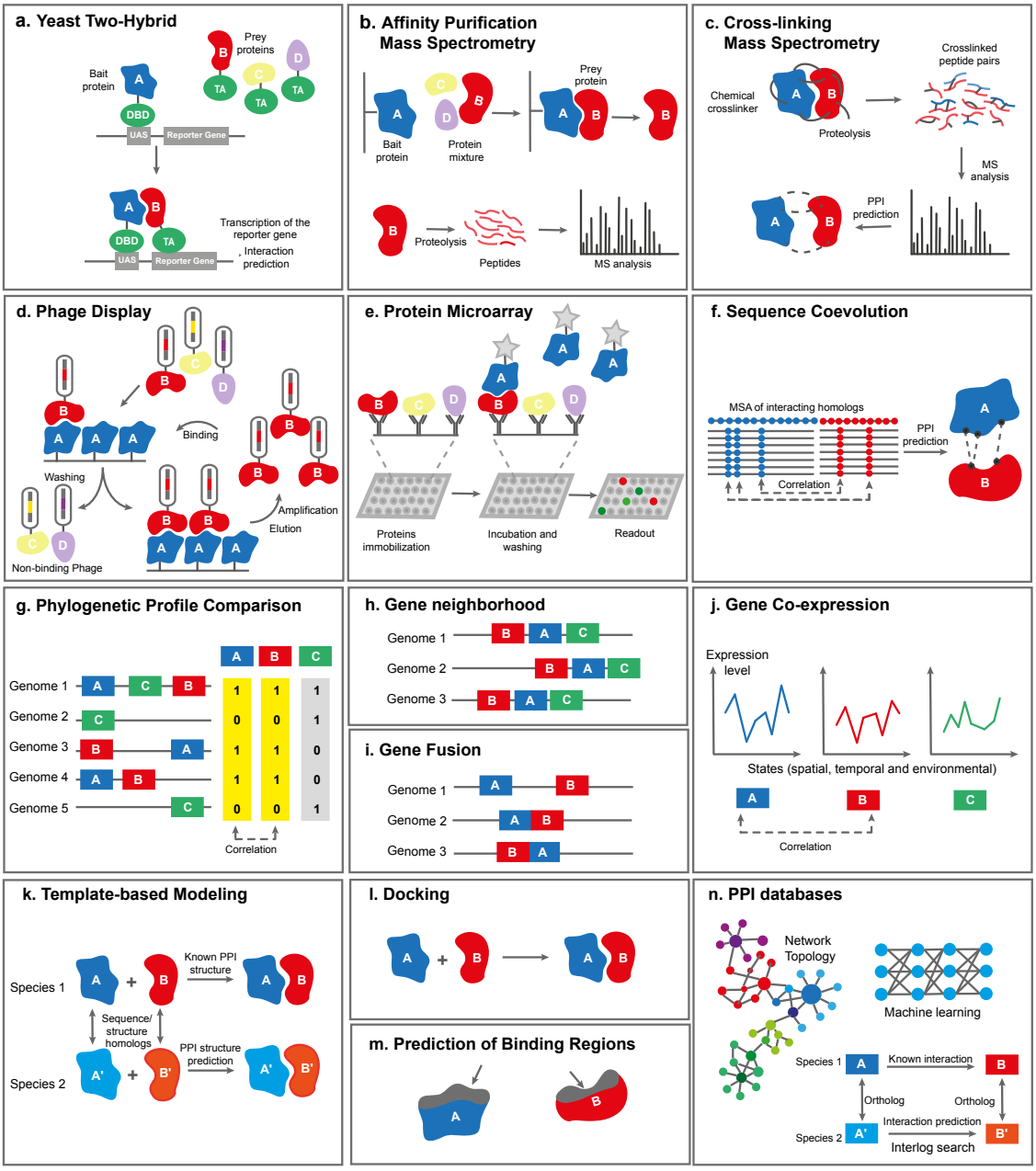


Figure 3.1: Conceptual illustration of the medium-to-high-throughput methods for characterization of protein-protein interactions.[3]

Computational methods that use concepts like gene proximity, gene fusion, gene co-expression, and phylogenetic profiling can provide an initial and fast screen for a proteome-wide protein

interaction network. However, they often confuse between functionally associated proteins or proteins with other sources of correlation and proteins with direct physical interactions.

Highly confident experimental approaches like co-crystallization and cryo-EM can provide the highest resolution of protein interaction, but they are the most labor and time-intensive methods with uncertain success rates. High-throughput experimental approaches (such as two-hybrid, affinity purification mass spectrometry, protein microarrays, and protein display) can efficiently screen large sets of protein pairs and reveal networks of probable physical interactions. Nevertheless, they are not done under native conditions, reflect mainly high-affinity interactions, and yield no structural information.

Methods such as cross-linking mass spectrometry and sequence coevolution can predict direct physical interaction from native and evolutionary origins, respectively. Besides they are labor efficient on the proteome level, they provide structural resolution of the protein complex on the residue level that can be utilized for complex modeling purposes. Machine learning approaches have recently made a breakthrough in the structure prediction of proteins and PPIs via integrative approaches that leverage databases of experimentally generated structures and the evolutionary trajectory of proteins in sophisticated algorithms yielding a high level of accuracy comparable with that of experimental methods.[39–42]

As we look into the future, the continued advancement of technology and computational power holds great potential for unraveling the complexities of PPIs. Integrating the strengths of different experimental and computational methods, will likely provide more comprehensive insights into the molecular organization of cellular systems.[43] Harnessing the power of machine learning, combined with experimental techniques, has the potential to revolutionize our understanding of PPIs, enabling us to map cellular interactomes with higher resolution and accuracy across entire proteomes.

3.2 Sequences co-evolution and the *Escherichia coli* cell envelope interactome

Parts of this section have also been published in the following article:
Green, A.G.*, **Elhabashy H.***, Brock, K.P., Maddamsetti, R., Kohlbacher, O. and Marks, D.S., 2021. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nature communications, 12*(1), p.1396. *Co-first authors.

3.2.1 Introduction

The characterization of PPIs on the global scale of proteomes and to the precise level of atomic resolution has been a long-standing goal in molecular biology. Numerous experimental and computational techniques have been developed to identify which proteins interact on the proteome level, but only a hand full of them can yield structural information about protein complexes. Coevolutionary methods have been proven successful in predicting and resolving the three-dimensional structures of PPIs by substantiating the evolutionary history of the proteins to infer the key interacting residues within a protein or a PPI. In this study (Appendix B), we addressed two of the main challenges of the evolutionary coupling approach, EVcomplex,[28,44] for large-scale prediction and structure elucidation of pairwise PPI without reliance on co-operonic positions, phylogeny, or docking. Fig. 3.2.A and 3.2.B illustrate the generic workflow of EVcomplex as described in section 1.2 and demonstrate its upscaling concept on the proteome level.

3.2.2 Results

The first challenge of EVcomplex in characterizing PPIs presented itself at the step of concatenating the multiple sequence alignments of query protein pairs. Pairing and concatenating the protein sequences are done by matching species, but often the presence of paralogs makes it difficult to select the correct interacting partners. The former usage of genome location or operonic structure to define the interacting partners within species limited the applicability of the approach to only genome-proximate interacting proteins and mostly excluded eukaryotic interactions.[28,45–47] To solve this, we replaced the reliance on genome distance by selecting a single sequence per species identified by the highest identity to the query sequence. This improvement expanded the applicability of EVcomplex by approximately five-fold over the *E. coli* proteome, from 10% to 53%, covering more interactions between proteins translated from genes that are distantly spread across the genome. In addition, it enabled the applicability of the approach to eukaryotic interactions which lack the operonic structures within their genomes.

The second challenge originated from the small fraction of the biologically relevant protein interactions within an enormous interaction search space leading to the prediction of false positives even with small false-positive rates. For example, *E.coli* has an estimated interactome of 10,000 PPIs and an interaction search space of more than nine million protein pairs.[2] The previous implementation of a high global threshold resulted in a high rate of true negatives of unpredicted yet coevolving interactions while could not eliminate high-scored false positives from predicted interactions. To increase the recall and minimize the number of false positive interactions, we developed a logistic regression model that uses a combination of structure-

agnostic features including EVcomplex score, the sum of intra-protein EC coupling constraints, the rank of inter-EC relative to intra-EC pairs, and sequence conservation (Fig. 3.2.C).

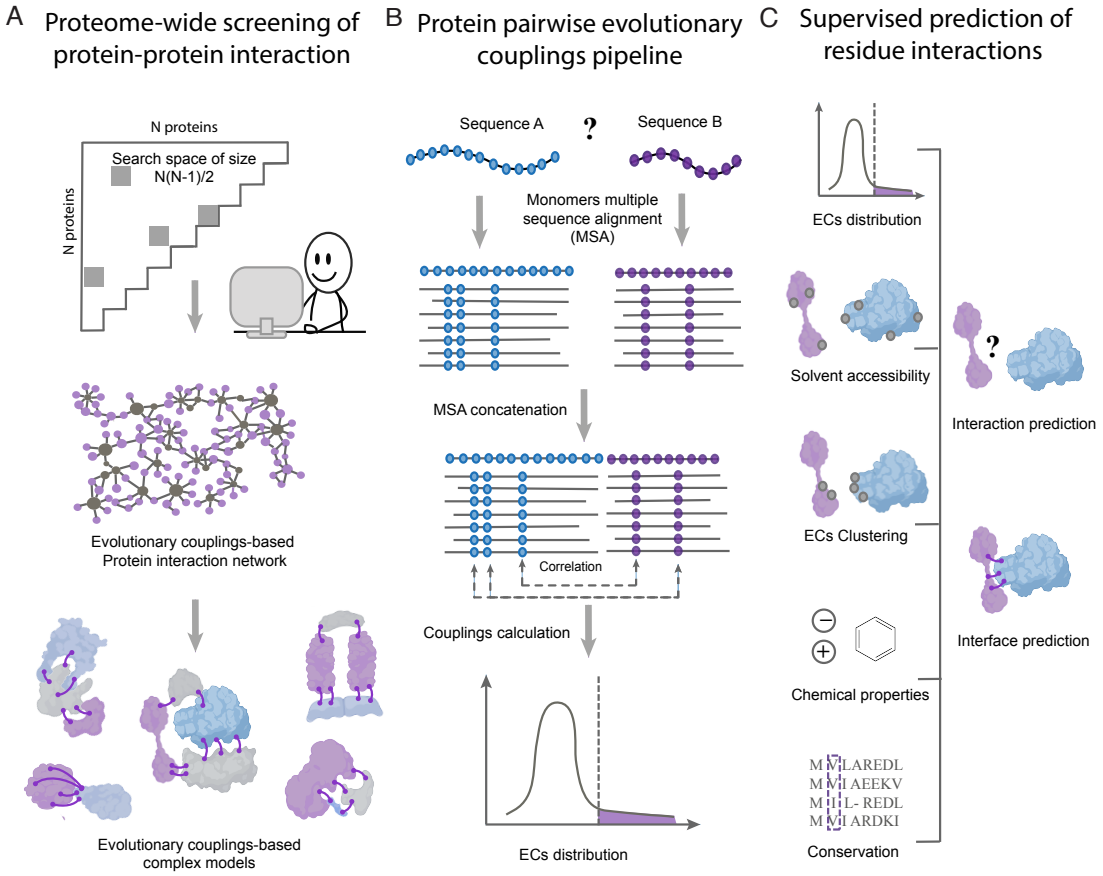


Figure 3.2: (A) Scaling up EVcomplex 2.0 to a proteome-wide screening of protein-protein interaction. (B) The generic pipeline of EVcomplex 2.0 for protein pair-wise evolutionary coupling calculation. (C) Supervised learning model, that uses features such as EVcomplex 2.0 output, solvent accessibility, clustering, chemical properties, and conservation, to estimate the probability of protein-protein interaction, and the probability of residue interaction.

We calibrated the model on a non-redundant benchmark dataset of 561 positive protein pairs with known structures and 3,987 non-interacting protein pairs. The model achieved a recall of 20.7% at a false positive rate of 0.1% on the held-out test set (Fig.3.3.A & 3.3.B). The recall further increased to 22.6% at a false-positive rate of 0.1% in cases where the subunits of structures are known. Fig.3.3.C & 3.3.D are illustrating example performance of the model when predicting the interfaces between ABC transporter permease and ATP binding subunit (Y1470-Y1471) and between the DNA primase (PriS-PriL) in comparison to their known struc-

tures PDB ID: 2NQ2 chains C and A and PDB ID: 1ZT2 chain A and B, respectively. Our model shows comparable performance to Yeast Two-Hybrid experiments,[2], achieving a recall of 27.9% compared to the experimental result of 29% at a false positive rate of 1%. Furthermore, our model offers the additional advantage of providing residue-level resolution to the interface contacts.

As a case study, we employed EVcomplex 2.0 to screen for PPIs involving cell envelope and membrane proteins in *E. coli*, which are some of the most challenging to determine experimentally.[48] We predicted and resolved the interface of 504 interactions, 243 of which have previously been unknown, thus doubling the number of tentative interactions. We utilized the predicted couplings to generate docked models of complexes with particular biological interest, among them the Tol-Pal systems and the Flagellar Hook-Filament Junction in more depth.

The Tol-Pal system consists of five core proteins: TolA, TolQ, TolR, TolB, and Pal which are involved in maintaining the integrity of the outer membrane of Gram-negative bacteria.[49] Using EVcomplex 2.0, we predicted and resolved the interaction between TolR-TolB[50] and we also resolved the interface contacts for the known interactions TolA-TolB[51] and TolB-Pal.[49] We built structural models for all pairwise interactions and the five-subunit complex (Fig. 3.4). Our interaction model of TolB-Pal accurately replicates its known structure (PDB ID: 2HQS).[49] We also generated a molecular model between the cell division coordinator CpoB that interacts with the TolB subunit of the Tol-Pal system.[52]

We proposed a model for the flagellar hook-filament junction, which is involved in ensuring smooth connectivity between the hook and the filament of the bacterial flagellum.[53] The junction consists of two rings: one ring is made of a homocomplex of the FlgK protein and the other is made of a homocomplex of the FlgL protein. In 2017, an eleven-mer ring model of the FlgK ring from *Campylobacter jejuni* was reported.[54] However, it was not known how FlgL forms its ring and how the two rings fit together to form the junction. Using evolutionary coupling, we resolved the interface between FlgK and FlgL and built an interaction model between an eleven-mer ring of FlgL inside the previously reported eleven-mer ring of FlgK showing the first evolutionary-based model of the Flagellar Hook-Filament Junction (Fig. 3.5).

We also demonstrated the plausibility of the sequence co-evolution-based method, EVcomplex 2.0, on eukaryotic PPIs. We used our benchmark set of 1,675 PPIs to assess the ability of EVcomplex 2.0 for interface prediction in eukaryotic interaction. Out of the 1,675 protein pairs in our dataset, 977 were found predominantly in eukaryotes. Approximately 11% of the eukaryotic interactions have at least one correctly predicted interface contact, and 4% have three or more true positive interface contacts. We also achieved successful predictions for a

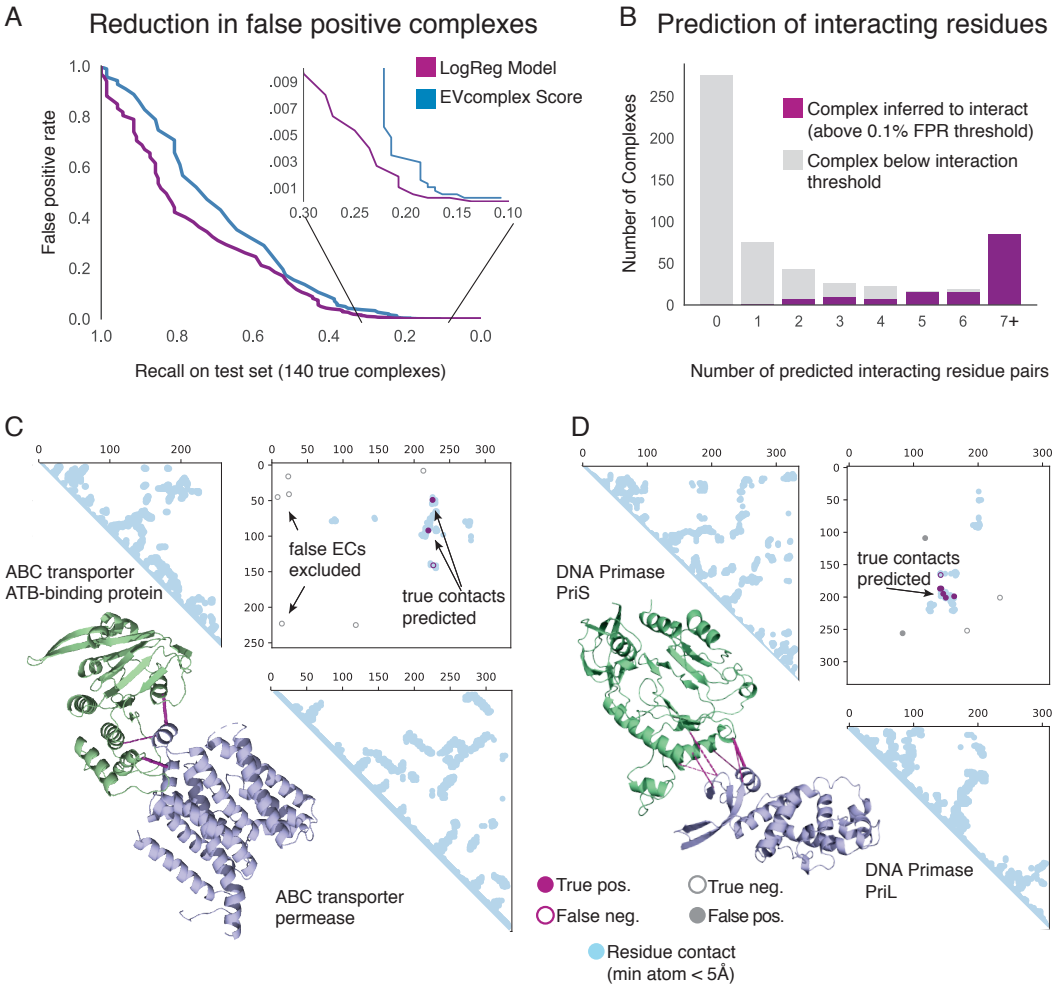


Figure 3.3: Our logistic regression model (purple) reduces the false positive rate compared to the previous EVcomplex score (blue) when predicting the interacting residues and interacting proteins. (A) Recall on the held-out fraction of the positive benchmark set (x-axis) versus the false positive rate on the held-out dataset of non-interacting complexes (y-axis). (B) The prediction of protein interaction is based on the prediction of interacting residues versus the number of predicted interacting residue pairs for complexes inferred to interact (purple) or not inferred to interact (gray) based on our stringent protein complex prediction threshold. (C) Example performance on known interaction between ABC transporter permease and ATP binding subunit (Y1470-HAEIN) with respect to its known structure PDB ID: 2NQ2 chains C and A. (D) Example performance on known interaction between DNA primase (PriS-PriL) with respect to its known structure PDB ID: 1ZT2 chain A and B.[44]

eukaryotic-exclusive complex, the human spliceosome. The spliceosome is an RNA-protein complex, which splices intronic sequences from immature mRNA in eukaryotes.[55] EVcomplex

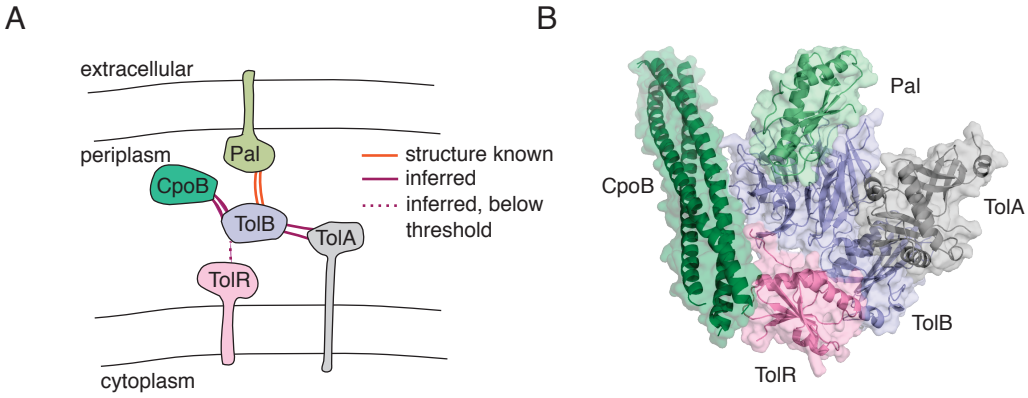


Figure 3.4: (A) Schematic of the proteins interaction network between the five core proteins of the Tol-Pal system, TolA, TolQ, TolR, TolB, and Pal, and the CpoB protein. (B) A complete model of the Tol-Pal complex inferred by restraints-based docking of known monomer structures based on the predicted interface contacts using EVcomplex 2.0.

2.0 correctly predicted the interface between Prp38-MFAP1 whose interactions contribute to the activation of the spliceosome,[56] LSm2-MFAP1, and Prp38-LSm5 which could plausibly interact at the time of spliceosome rearrangement during activation. Our results demonstrate the plausibility of using sequence co-evolution-based methods for eukaryotic protein interaction. For more details about the approach, our assessment, and results, see (Appendix B).

3.2.3 Conclusion

In conclusion, this study addressed key challenges for the evolutionary coupling approach, EVcomplex, for the prediction and three-dimensional modeling of PPIs on the proteome level. By employing a reciprocal highest identity approach, we resolved the issue of pairing the potential interacting homologs as well as expanded the applicability of EVcomplex 2.0 to eukaryotes and increased coverage in the *E. coli* proteome. Additionally, we developed a logistic regression model that incorporated structure-agnostic features to minimize false-positive interactions. We showcase the successful application of EVcomplex 2.0 on cell envelope and membrane proteins in *E. coli*, which are some of the most challenging to determine experimentally, doubling the number of tentative interactions. We proposed structural models for interaction with known biological relevance such as the Tol-Pal system, the flagellar hook-filament junction, and the eukaryotic-exclusive spliceosome complex.

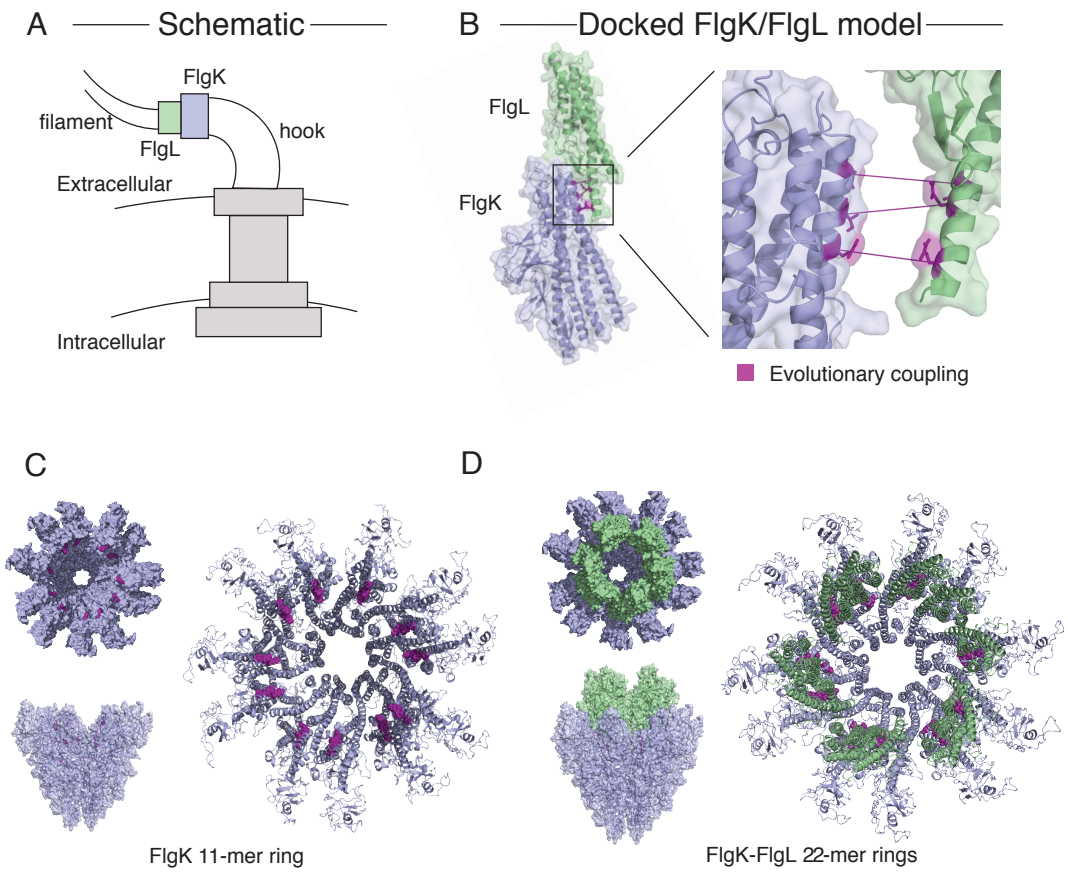


Figure 3.5: (A) Schematic of the flagellar hook-filament junction shows the two interacting rings. One ring is made of FlgL proteins (green) and the other is made of FlgK proteins (purple).[53] (B) A hetero-dimeric docked model of FlgL-FlgK using evolutionary coupling restraints. (C) A previously reported eleven-mers ring model of FlgK ring from *Campylobacter jejuni*.[54] Positions of the predicted interface contacts with the FlgL ring are highlighted in magenta. (D) A proposed model of the hetero-oligomers interaction between FlgL (green) and FlgK (purple) forming the flagellar hook-filament junction.

3.3 Cross-linking mass spectrometry and the human endolysosomal interactome

Parts of this section have also been published in the following article:
Singh, J., **Elhabashy H.**, Muthukottiappan, P., Stepath, M., Eisenacher, M., Kohlbacher, O., Gieselmann, V. and Winter, D., 2022. Cross-linking of the endolysosomal system reveals potential flotillin structures and cargo. *Nature Communications, 13*(1), p.6212.

3.3.1 Introduction

Lysosomes are cellular organelles that are involved in various cellular processes, mainly macromolecule degradation and metabolic signaling.[57,58] Protein interactions play an essential role in mediating these functionalities as well as other lysosome-related processes, including cell adhesion/migration and interaction with other cellular compartments.[59–64] Lysosomal dysfunction can lead to a range of neurodegenerative diseases and cancer.[65] Gene mutations in lysosomal proteins can induce about 70 rare and frequently devastating diseases, known as lysosomal storage disorders (LSDs).[66] Despite their crucial roles, the majority of lysosomal PPIs associated with the pathogenesis of these diseases are poorly characterized. Given the central role of lysosomes, we dedicated this project (Appendix C) to exploring the lysosomal proteome and interactome under near-native conditions.[67]

3.3.2 Results

Through a collaborative effort with the Winter Lab at the Institute for Biochemistry and Molecular Biology (IBMB), we investigated the first LC-MS/MS dataset of the endolysosomal compartment of HEK293 cells. Researchers from the Winter Lab created this dataset by applying the MS-cleavable cross-linker disuccinimidyl sulfoxide (DSSO) to lysosome-/early endosome-enriched fractions. To achieve this, they performed an enrichment condition for lysosomes using superparamagnetic iron oxide nanoparticles (SPIONs) overcoming the challenge of relatively low abundance of lysosomal proteins in whole-cell XL-LC-MS/MS studies. Lysosomes were cross-linked both in intact (IT) and disrupted (DR) states to address the limited membrane permeability of the used MS-cleavable cross-linker DSSO. They assessed lysosomal intactness, recovery, and enrichment over 384 plates of HEK293 cells across three biological replicates. After enzymatic digestion and peptide fractionation, the samples underwent LC-MS/MS analysis (Fig. 3.6). The resulting spectra were analyzed and cross-links were assigned using XlinkX software[68] with a false discovery rate (FDR) set at 5%. For more details about the methodology, see (Appendix C).

We assumed the responsibility for the sequence and structural analysis of this dataset. For this, we have developed a bioinformatics pipeline for structural evaluation and validation of the identified cross-links based on their known structures and AlphaFold models (Fig. 3.7).[39,69] The pipeline accepts the output of the XL-MS experiment as an input, typically in Excel or CSV format, containing information about the cross-linked proteins such as the UniProt IDs of the cross-linked proteins, the sequences of the cross-linked peptides, and the number of the cross-linked residues.

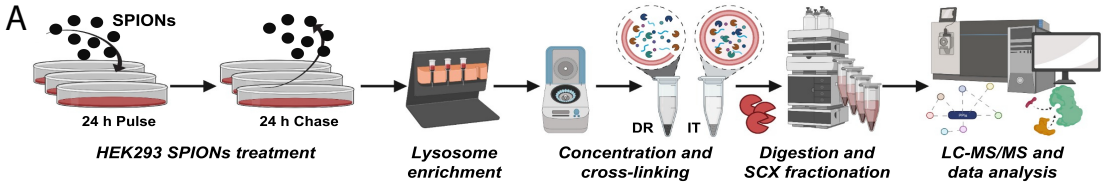


Figure 3.6: Experimental workflow for the XL-LC-MS/MS analysis of lysosome-enriched fractions.

To initiate the process, the pipeline parse the protein sequences via the application programming interface (API) from the UniProt web server (`https://www.uniprot.org`).[70] The protein sequences are subsequently queried using BLAST 2.9.0[71] against a locally downloaded Protein Data Bank (PDB).[72] The hit criteria included a minimum *e*-value threshold of 10^{-5}, restriction to the same species, and a minimum alignment coverage of 50%. Nevertheless, it is worth noting that users have the flexibility to customize the software versions and inclusion criteria according to their preferences. If a structural hit exists, the corresponding PDB file is retrieved from the Protein Data Bank web server (`https://www.rcsb.org`). In the absence of structural hits for intra cross-links, we retrieve predicted structure models from the AlphaFold Protein Structure Database (`https://alphafold.ebi.ac.uk`).[39,69] Similarly, in the absence of complex structural hits for inter cross-links, we model the complexes either with AlphaFold-Multimer[39,69] or with HADDOCK 2.2[73] restraints-based docking, whereas HADDOCK requires known monomer structures.

Following this, the pipeline extracts the targeted protein chains from the PDB files and proceeds with restraints mapping and evaluation. We used TopoLink[74] to evaluate the Euclidean and topological CA-CA distances between the cross-linked residues on their corresponding structures. The pipeline then processes the output of TopoLink (06/2019) and generates the final CSV output file that contains values of structural evaluation of the identified cross-links. Additionally, the pipeline automatically creates a visualization for the cross-linking restraints mapped on their corresponding structures using PyMol 2.9, a molecular visualization software. Notably, aside from the time required for the modeling procedure, our pipeline exhibits high efficiency and rapid processing capabilities, making it an ideal choice for large-scale analysis of cross-linking data and the modeling of identified interactions.

We identified 2,467 unique residue-residue cross-links including 524 cross-links covering 111 proteins assigned to the lysosomal compartment. Of the 524 lysosomal cross-linked peptides, 270 links stem from the same protein sequence (intra-link) and 254 links stem from different protein sequences (inter-link) (Fig. 3.8). Our pipeline structurally mapped 161

unique cross-links to the resolved structures of 34 lysosomal and lysosome-associated proteins. Of them, 87% achieved the acceptable distance range (< 35 Å) of the used cross-linker, confirming the validity of our dataset.

Based on the identified inter-links, we constructed a network of 254 cross-links involved lysosomal proteins. Comparison to known interactions revealed an overlap of 30%, while 70% of the cross-linked proteins demonstrated potential novel interactions. Only 26% of potential lysosomal PPIs were included in the STRING database.[75] Concerning lysosomal and lysosome-associated proteins, we identified the highest numbers of cross-links for the V-ATPase,

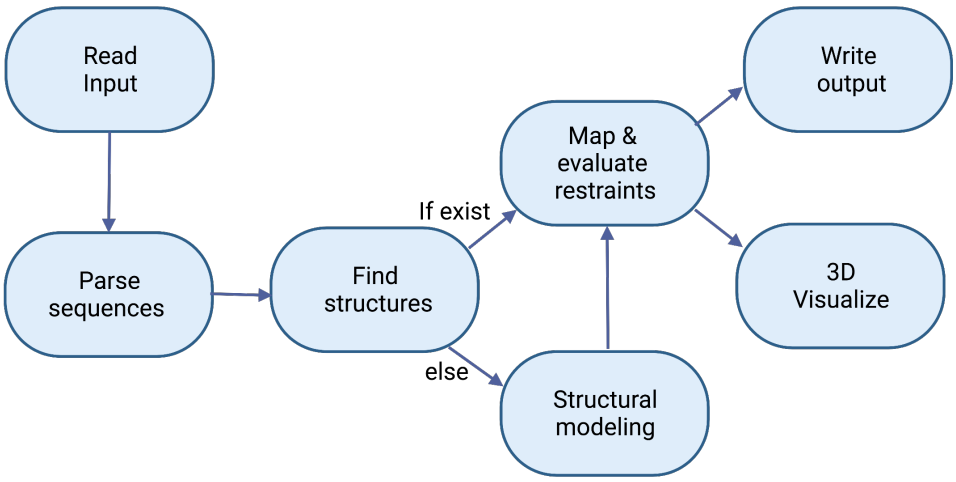


Figure 3.7: Bioinformatics workflow for the structural analysis and validation of cross-linking data.

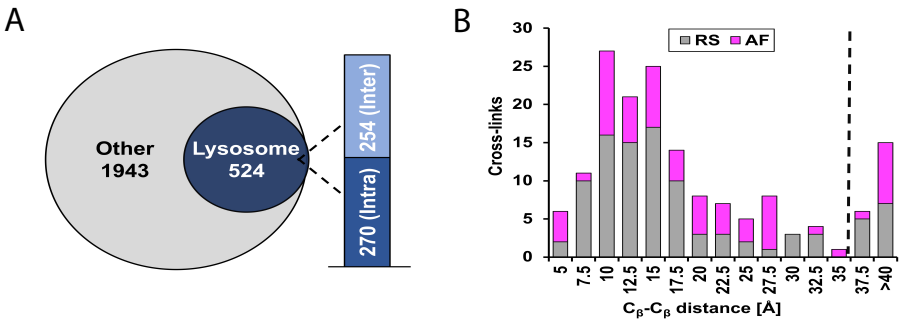


Figure 3.8: (A) Classification of unique cross-linked residue pairs. The pie-chart shows that 524 out of the identified 2,467 cross-links were assigned to the lysosomal compartment. Of these, 270 are intra-protein-links, and 254 are inter-proteins-links. (B) Distances of the mapped cross-links to experimental structures and AlphaFold models for lysosomal proteins, bin size: 2.5 Å.

the flotillins, mTORC1, and the syntaxins.

For the V-ATPase complex, we identified 29 cross-links covering 13 unique proteins of the 35-mer V-ATPase. We integrated 21 cross-links into a cryoEM structure of V-ATPase[61] (PDB ID: 6WM2) (Fig. 3.9.A), and the remaining eight cross-links, mapped on missing regions, were integrated onto full-length AlphaFold models upon aligning them with the assembly. About 90% of the mapped cross-links have fulfilled DSSO distance constraints. Three cross-links that exceed the distance limit correspond to the interaction between subunits A and E, which undergo conformational changes during the ATP hydrolysis.[61]

Most identified PPIs were associated with the soluble V1 part of subunit D (ATP6V1D) which can be attributed to the capability of V1 to dissociate from the lysosomal membrane-embedded V0 part exposing the D subunit to interactions.[76] The Winter Lab verified the identified interaction between ATP6V1D and frizzled 9 (FZD9), a member of the WNT signaling pathway, by co-immunoprecipitation. FZD9 was shown to be sorted to late endosomes/lysosomes after its internalization by endocytosis,[77] and the ATP6AP2 subunit of the V-ATPase was reported to interact with FZD841.

We identified 18 cross-links between peptides that originated from the PPT1 protein, a member of the palmitoyl protein thioesterase family. It has been reported that glycosylation-deficient PPT1 variants are devoid of enzymatic activity due to improper folding.[78,79] When mapping the cross-links on the PPT1 crystal structure (PDB ID: 3GRO) which forms a homo-dimeric complex, eleven intra-cross-links fulfilled the distance constraints on this structure and nine cross-links exceeded them indicating the possibility of an alternative oligomerization state (Fig. 3.9.C). This hypothesis aligns with a previous report on the maximal enzymatic activity of the PPT1 complex at a molecular weight that corresponds with a tetrameric assembly (>100 kDa).[79] We used the possible tetrameric-originated cross-links as distance restraints to build a tetrameric structural model of PPT1, using HADDOCK 2.2,[73,80] which has fulfilled 100% of the identified cross-links for PPT1 (Fig. 3.9.D).

Our dataset contains 29 unique cross-links for FLOT1 and FLOT2, which are highly conserved lipid raft-associated proteins found in nearly every vertebrate cell.[81] The Winter Lab verified the interaction between FLOT1 and FLOT2 using Western blotting and co-localization by immunostaining. We built a hetero-dimeric model of FLOT1- FLOT2 using ColabFold,[82] a variant of AlphaFold, which has shown excellent agreement with all identified cross-links (Fig. 3.9.B). We observed 18 substrates for FLOT1 and FLOT2 including the latrophilin family of adhesion G protein-coupled receptors GNB4. The Winter Lab validated the interaction between GNB4 with its cross-linked partner FLOT2 as well as FLOT1, indicating the binding of GNB4

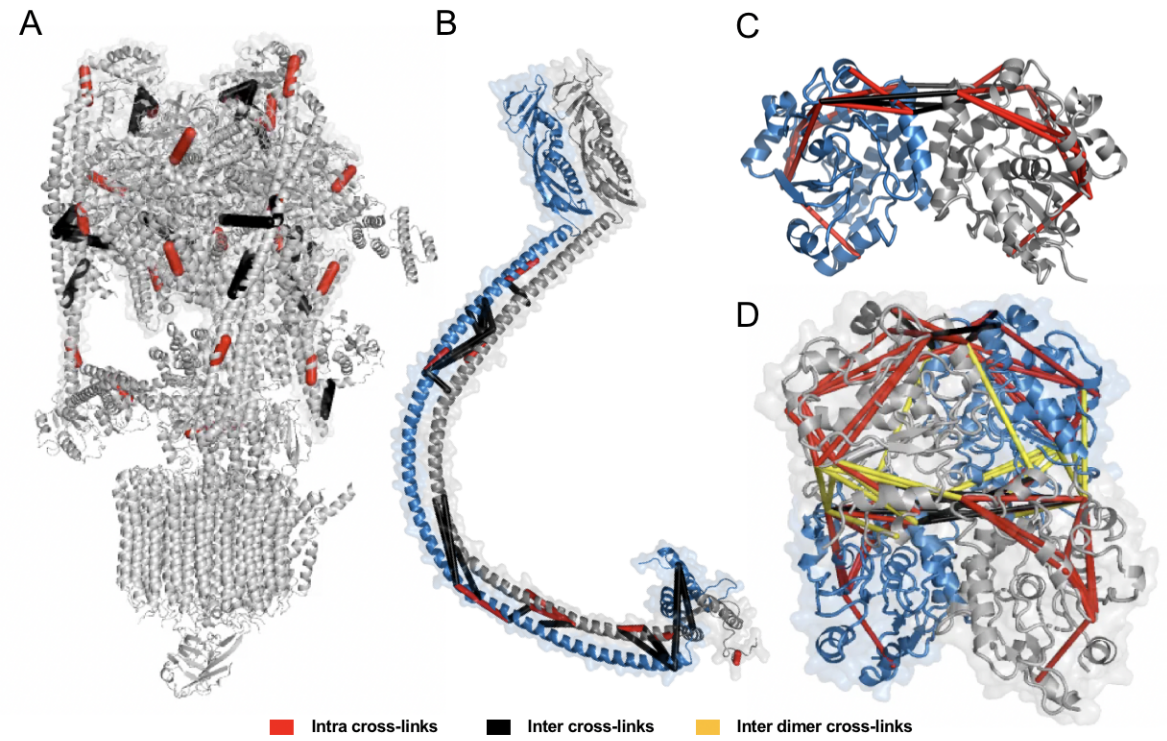


Figure 3.9: (A) Refined structure of the V-ATPase complex (PDB ID:6WM2). Identified cross-links were integrated into the V-ATPase cryo-EM structure. (B) Hetero-dimeric model of FLOT1-FLOT2 interaction (ColabFold). The model satisfies the distance constraints of all intra and inter cross-links. (C) Reported homo-dimeric human PPT1 structure (PDB ID: 3GRO). (D) Tetrameric PPT1 model representing the most favorable energetic state and fulfilling DSSO's distance constraints for all 18 cross-links

to FLOT1/FLOT2 heteromeric assembly.

3.3.3 Conclusion

In summary, this study presented the first LC-MS/MS dataset of the endolysosomal compartment of HEK293 cells. A robust bioinformatics pipeline was developed to evaluate and validate the identified cross-links using known structures and AlphaFold models. Our results revealed novel structural insights into known PPIs within the lysosomal compartments such as V-ATPase, PPT1, and FLOT1/FLOT2 and expanded our knowledge of the novel interactions. We supported results with experimental validation via co-immunoprecipitation and structural modeling for complexes with important biological relevance. This study highlights the potential of cross-linking proteomics and structural bioinformatics analysis for unraveling the complexity of the lysosomal compartment and identifying novel protein interactions.

3.4 Cross-linking, evolutionary coupling, and the mouse mitochondrial interactome

Parts of this section have been included in the manuscript:
Elhabashy H., Netz E., and Kohlbacher O., 2023. XLEC: Large-scale prediction and modeling of protein-protein interactions by combining sequence co-evolution and cross-linking data. *Manuscript under review*.

3.4.1 Introduction

The high level of heterogeneity of PPIs underscores the significance of adopting versatile and integrative approaches to address the challenges associated with their characterization and prediction. In this study (Appendix D), we developed an integrative approach called XLEC, **C**ross-**L**inking and **E**volutionary **C**oupling, that combines sparse data from cross-linking mass spectrometry experiments and evolutionary coupling calculations for enhanced prediction and modeling of PPIs (Fig. 3.10). The consensus outcome from both approaches boosts the confidence in discovering new interactions via evidence of physical proximity within native conditions and signals of sequence co-evolution. The identified distance restraints from both approaches unravel diverse yet complementary structural aspects of the protein complexes ranging from the topological arrangement of the interacting domains to the key interacting residues across the physical interface.

3.4.2 Results

The XLEC workflow employs a series of procedures to ensure increased accuracy in modeling and identification of PPIs (Fig. 3.11). Starting with a dataset of cross-linked protein pairs, XLEC utilizes EVcomplex 2.0 to investigate residue-residue co-evolution for every protein pair found to be cross-linked. XLEC then performs XL-based and EC-based docking separately using HADDOCK 2.2 default protocol. For each docking procedure, XLEC shortlists the top ten resulting models based on the HADDOCK scoring function and the top model from each pairwise-RMSD cluster within a threshold of 5 Å for further analysis.

To refine the used restraints and, subsequently, improve the resulting models, XLEC geometrically evaluates both the XL and EC restraints on the short-listed models. XL restraints that achieve less than 35 Å and EC restraints that achieve less than 15 Å in the model with the highest number of satisfied restraints are accepted for the XLEC-based docking. This procedure

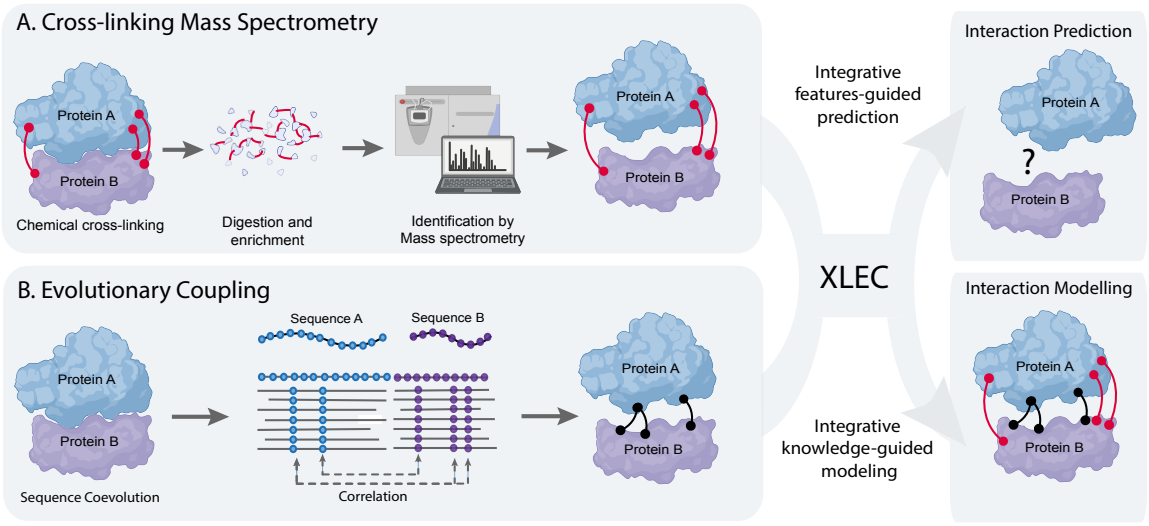


Figure 3.10: Conceptual illustration of XLEC approach for characterization of protein-protein interaction.

enables false positive elimination by excluding the restraints that are geometrically in consensus with the majority of the applied restraints, assuming true positives. It also enriches EC signals by rescuing lower-scored restraints in consistency with the top-scored restraints. Using a hybrid set of the updated XL and EC restraints, XLEC models all binary protein interactions using HADDOCK 2.2 restraints-based docking. Following the modeling procedure, XLEC employs a logistic regression classifier to differentiate between physically interacting proteins (class 1 = interface) and proximate proteins with no significant physical interface (class 0 = no interface).

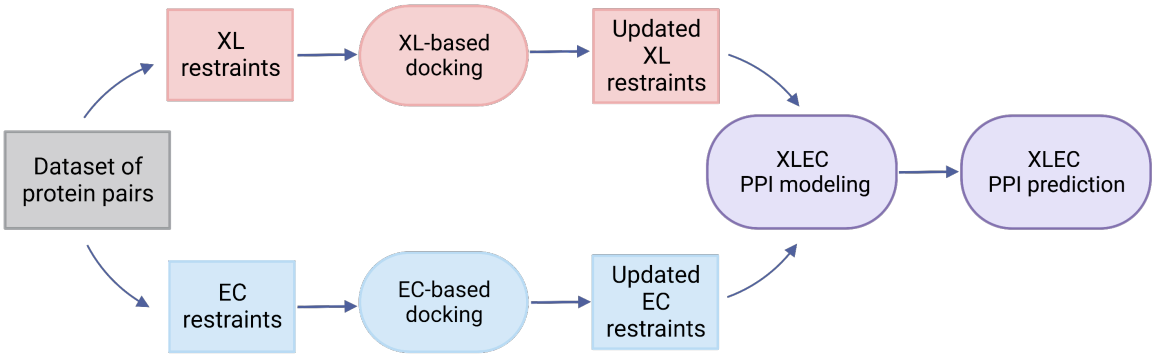


Figure 3.11: Workflow of the XLEC pipeline for modeling and prediction of PPIs.

We applied XLEC on a cross-linking dataset generated by combining two previously reported XL-MS experiments on the mouse mitochondrial interactome.[83,84] A non-redundant

combination of the inter-protein cross-links from the datasets consists of 2,616 cross-links forming an interaction network of 770 protein pairs across 324 unique proteins. Out of 324 protein sequences, only 238 have three-dimensional structures, comprising 41 PDB structures and 197 homology models. We created a benchmark dataset of 85 PPIs with complex structural resolution, including 67 PPIs with significant physical interfaces and 18 protein pairs without, and a prediction set of $\sim$ 400 protein pairs with no structural resolution for their complexes but rather their subunits. Both datasets share the criteria of having sufficient multiple sequence alignment for evolutionary coupling calculations and a minimum of one structurally mapped cross-link.

We modeled the 85 PPIs of our benchmark dataset using multiple HADDOCK 2.2[73,80] docking protocols (XL-based docking, EC-based docking, and XLEC-based docking) and the machine learning-based approach AlphaFold-Multimer V2.1,[41] both with and without the usage of templates. For more details about the modeling approaches, see (Appendix D). We assessed the performance of the used approaches based on two main criteria; the ability of the approach to model the interaction and its ability to predict the interaction, in that order.

For our assessment of model quality, we evaluated the overall geometric fit of the resulting models to the target structures using the CAPRI assessment protocol that uses metrics of Ligand Root Mean Square Diffusion (L-RMSD), Interface RMSD (I-RMSD) and the fraction of interface native contacts (F_nat).[85] Additionally, we used the DockQ score which is a linear combination of the three prior CAPRI metrics to evaluate model qualities.[86]

Table 3.1: Assessment of model quality. The performance of different interaction modeling approaches qualified as protein complex counts and percentages with respect to the interaction benchmark dataset in terms of CAPRI categories and DockQ scores.

Docking protocol	Used metric	High quality	Medium quality or better	Acceptable quality or better
XL-based docking	CAPRI	0 (0%)	0 (0%)	2 (3.0%)
	DockQ	0 (0%)	0 (0%)	3 (4.5%)
EC-based docking	CAPRI	0 (0%)	6 (9.0%)	12 (17.9%)
	DockQ	0 (0%)	4 (6.0%)	12 (17.9%)
XLEC-based docking	CAPRI	2 (3.0%)	8 (11.9%)	13 (19.4%)
	DockQ	1 (1.5%)	6 (9.0%)	10 (14.9%)

The XL-based docking produced models with CAPRI-acceptable quality for 3.0% of the benchmark interaction set and models with acceptable DockQ scores ($0 < DockQ < 0.23$) for 4.5%. No models achieved medium or high quality for this approach (Table 3.1). EC-based

docking has shown better performance, resulting in models with at least CAPRI and DockQ acceptable quality for 17.9% of the benchmark set. It generated models with CAPRI medium quality models for 9.0% and DockQ medium quality ($0.49 \leq DockQ < 0.8$) for 6% of the benchmark set and no high-quality models.

Upon employing XLEC to model our benchmark set, we generated models with acceptable quality or better for 19.4% and medium quality or better for 11.9% of the complexes, in terms of CAPRI categories. In terms of DockQ score, XLEC generated models with at least acceptable quality for 14.9% and at least medium quality for 9.0% over the same complexes set. Comparing to EC-based docking, it is evident that XLEC achieved a higher rate (CAPRI: 33% and DockQ 50%) of the medium to higher quality structures (Fig.3.12.A). Similar to XL- and EC-based dockings, it was challenging for XLEC to achieve a significant number of high-quality structures (CAPRI: 3.0% and DockQ: 1.5%).

Fig. 3.12.B, 3.12.C, and 3.12.D show exemplary models from all three described protocols of XL-, EC-, and XLEC-based dockings compared to their known PDB structures: I) Interaction model between subunits 2 and 7 of cytochrome b-c1 complex with respect to its PDB structure (ID: 4D6T, chains O and F). II) Interaction between the 75 kDa subunit of NADH-ubiquinone oxidoreductase and NADH dehydrogenase 1 alpha subcomplex subunit 2 of the mitochondrial NADH dehydrogenase complex with its PDB structure (ID: 6G2J, chains G and S). III) Interaction between subunits 7B and 4 of cytochrome c oxidase complex with PDB structure (ID: 2Y69, chains K and D). In these three complexes, even though the XL restraints are all consistent with the known structure, XL-based docking struggled to achieve an acceptable quality due to their sparsity. While the short-range interface contacts from EC achieved acceptable to medium-quality models. XLEC-based models showed increased model quality from medium to high accuracy.

Unsurprisingly, the machine learning-based approach of AlphaFold-Multimer significantly surpassed all the described restraints-based docking protocols, generating 56.7% of the complexes with acceptable quality or better, 49.3% with at least medium qualities, and 20.9% with highly accurate qualities as determined by CAPRI metrics. In terms of DockQ score, AlphaFold-Multimer generated 56.7% of the complexes with acceptable quality or better, 35.8% with at least medium qualities, and 14.9% with highly accurate quality.

While this excellent performance can be associated with the prior training and template usage by AlphaFold-Multimer, it can be tricky to extrapolate it to complexes, outside our benchmarking set, with no structure resolution in the databases. This claim was evident by the decrease in performance upon running AlphaFold-Multimer on the benchmark dataset with-

out the use of templates which still does not mitigate the fact that AlphaFold-Multimer was aware of the complex structures due to its training. Based on the CAPRI assessment, 49.3% of the complexes achieved acceptable quality or better, 43.3% achieved at least medium quality, and 11.7% achieved high accuracy. Based on DockQ categories, the results show 49.5% of benchmark complexes achieved acceptable qualities better, 20.9% achieved at least medium qualities, and 8.9% achieved highly accurate qualities.

In the absence of known complex structures, one can monitor AlphaFold-Multimer quality measures of pLDDT, pTM, ipTM, and PAE and decide on the reliability of the resulting model.[39]

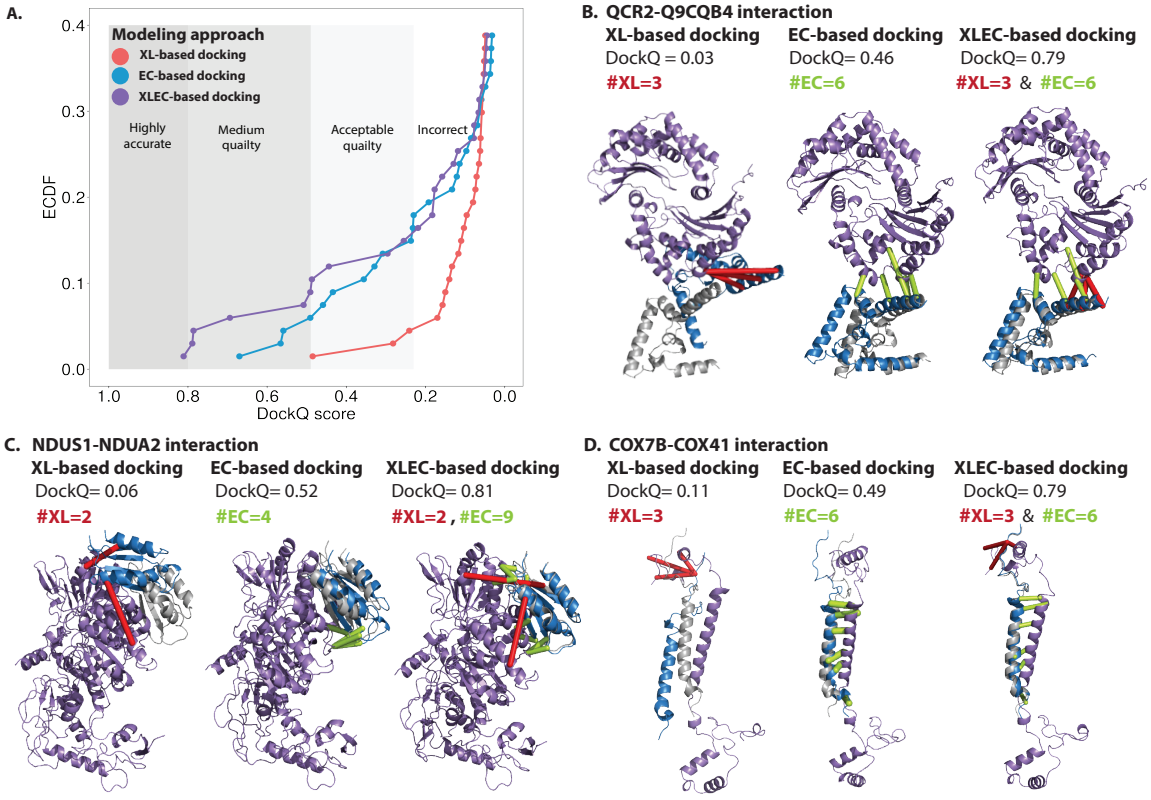


Figure 3.12: Benchmark set interaction models as predicted by XL-based docking, EC-based docking, and XLEC-based docking. (A) The panel shows the empirical cumulative distribution function (ECDF) versus DockQ score, illustrating the performance of the different modeling approaches on our benchmark set of complexes of known structures. (B) Docked models of QCR2-Q9CQB4 interaction as compared to PDB structure (ID:4D6T, chains O and F). (C) Docked models of NDUS1-NDUA2 interaction as compared to PDB structure (ID:6G2J, chains G and S). (D) Docked models of COX7B-COX41 interaction as compared to PDB structure (ID:2Y69, chains K and D). For every model, we show the predicted location of the ligand subunit (in blue) compared to its equivalent subunit (in gray) from the known complex structure after the backbone atoms of receptor subunits (in purple) were superimposed.

We employed AlphaFold-Multimer to model our prediction dataset of 417 PPIs with no known structural resolution for their complexes. For only 5% of the prediction dataset, AlphaFold-Multimer achieved a pTM based-confidence score of more than 0.5 in comparison to 40% in the benchmark dataset. Coupled with the monomer structures, XLEC-based docking can offer an alternative approach when AlphaFold-Multimer fails to yield a reliable model for complexes with no structural templates.

Using XLEC-based docking and AlphaFold-Multimer, we generated more than 400 *de novo* PPI models revealing novel insights into the mitochondrial interactome. Fig. 3.13.A and 3.13.B show models of XLEC in agreement with models from AlphaFold-Multimer for the interaction between DNA topoisomerase I (TOP1M) and electrogenic aspartate/glutamate antiporter (SLC25A13) and the interaction between the heat shock chaperone (TRAP1) and the stress-70 chaperone (GRP75). Fig. 3.13.C and 3.13.D show different structural models between XLEC and AlphaFold-Multimer for the interaction between alanine aminotransferase 2 (ALAT2) and Hexokinase-1 (HXK1) and the interaction between cytochrome b-c1 complex subunit 2 (QCR2) and aconitate hydratase (ACON).

For our assessment of interaction prediction, we trained logistic regression models to classify physically interacting proteins from non-interacting pairs on our benchmark dataset of 85 protein pairs. Of these, 67 protein pairs are physically interacting with 10 heavy atom interface contacts within 5 Å and 18 protein pairs have lesser or no interface contacts. This imbalance in the dataset presumably reflects the interacting to non-interacting proteins ratio within a cross-linking dataset. We used five-fold cross-validation to evaluate the performance of the classifiers.

Table 3.2: Features of the logistic regression classifiers

Classifiers	# XL	# XL < 30Å	EC_{max}	# EC < 10Å	MSA N_{eff}	MSA %coverage	EC Kurtosis & Skewness	AF pTM-score	HADDOCK score	# HADDOCK clusters
XL	✓	–	–	–	–	–	–	–	–	–
EC	–	–	✓	–	✓	✓	✓	–	–	–
XLEC	✓	✓	✓	✓	✓	✓	✓	–	✓	✓
XLEC & AF	✓	✓	✓	✓	✓	✓	✓	✓	–	–
AF	–	–	–	–	–	–	–	✓	–	–

We trained two baseline models on the sequence-based features of XL and EC separately. The XLEC-based predictor was trained on features from cross-linking and evolutionary coupling complemented with structural features including the numbers of justified XL and EC restraints when mapped on the generated models and metrics of intrinsic model quality. We introduced this novel approach that utilizes the structural features to guide the prediction of the interaction based on the premise that structural consistency between restraints of different

experimental and evolutionary evidence origins supports the interaction. Table 3.2 shows the features utilized by each model. We also trained one predictor on the pTM-based confidence scores of AlphaFold-Mutlimer and another on the structural features of AlphaFold-Multimer.

Combining evidence of interaction from both approaches of XL and EC, alongside the structural features of the models they can yield combined, resulted in slightly higher performance with an AUC of 0.73 compared to 0.64 (XL alone) and 0.68 (EC alone) (Fig. 3.14). AlphaFold-

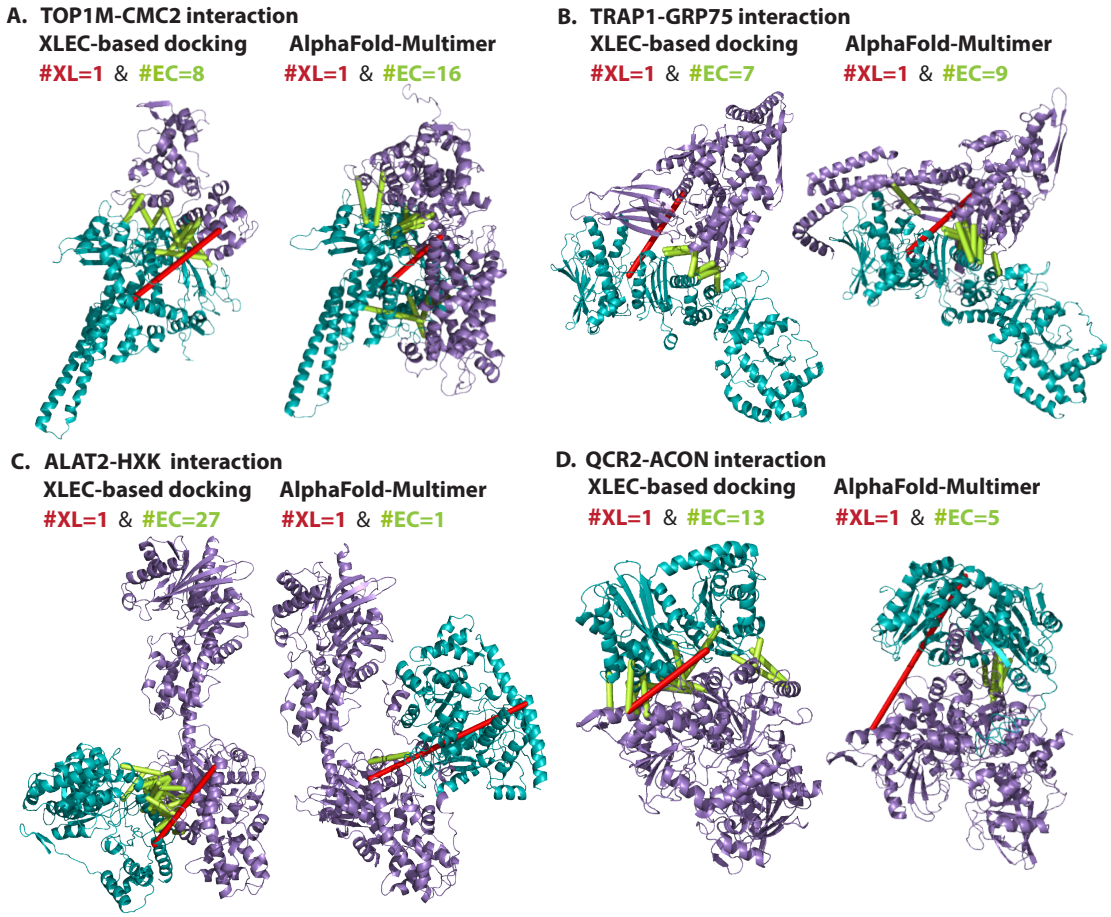


Figure 3.13: Interaction models from the prediction set as predicted by XLEC-based docking and AlphaFold-Multimer. For every model, we show the predicted location of the ligand subunit (in teal) and receptor subunits (in purple). Cross-linking restraints (XL) are shown in red and the evolutionary coupling restraints (EC) are shown in light green. (A) Interaction model of the DNA topoisomerase I (TOP1M) and electrogenic aspartate/glutamate antiporter (SLC25A13/CMC2). (B) Interaction model between the heat shock chaperone (TRAP1) and the stress-70 chaperone (GRP75). (C) Interaction model between the alanine aminotransferase 2 (ALAT2) and Hexokinase-1 (HXK1). (D) Interaction model between cytochrome b-c1 complex subunit 2 (QCR2) and aconitate hydratase (ACON).

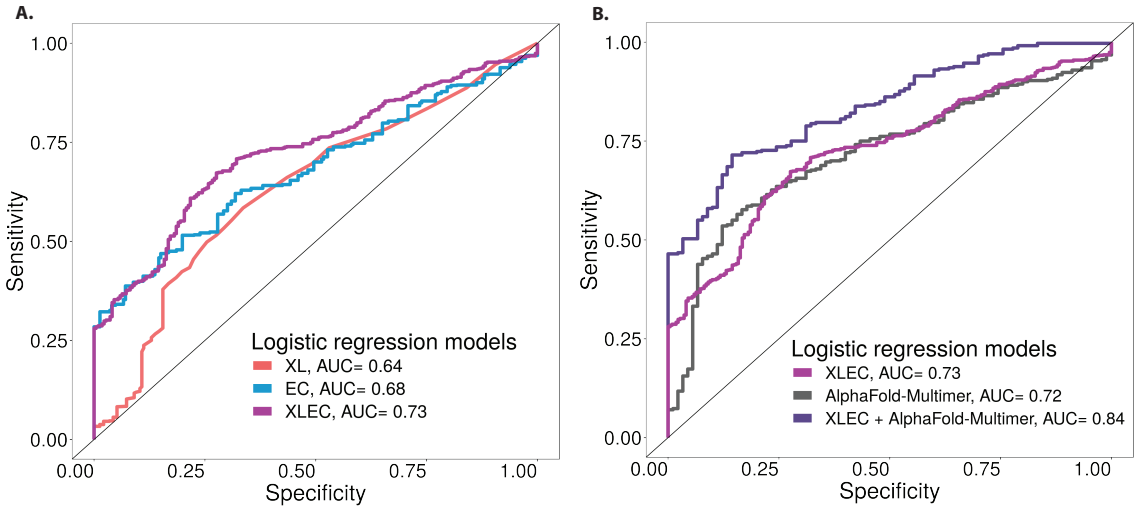


Figure 3.14: The performance of the trained interaction predictors. (A) The Receiver Operating Characteristic (ROC) curves for the developed logistic regression classifier of protein interaction and their associated Area Under the Curve (AUC) for XL-, EC- , and XLEC-based predictors. (B) shows the ROC curves with their corresponding AUC of XLEC, AlphaFold-Mulitmer, and XLEC & AlphaFold-Mulitmer based predictors.

Multimer-based classifier slightly underperforms the XLEC classifier achieving ROC-AUC of 0.72. The best predictor combined AlphaFold-Multimer as well as XLEC features achieving ROC-AUC of 0.84. The consensus between cross-linking restraints, evolutionary coupling features, and an accurate model that satisfies both of them, made a clear difference between interacting and non-interacting protein pairs. These results clearly show that integrating XLEC and ML-based methods harnesses a synergy for enhanced prediction of PPIs.

3.4.3 Conclusion

In this study, we developed an integrative approach called XLEC (Cross-Linking and Evolutionary Coupling) to enhance the large-scale prediction and modeling of PPIs. XLEC combines the diverse structural aspects provided by both approaches into structural models of the protein complexes. Additionally, we trained logistic regression models that use consensus sequence- and structure-based evidence from both approaches to classify interacting protein pairs. XLEC yielded the best predictor and modeling approach when compared to its constituting approaches individually. We applied XLEC to analyze a dataset of the mouse mitochondrial interactome, generating models for benchmark complexes and achieving high-quality results, showcasing its effectiveness in understanding PPIs and their structural characteristics. We showed that XLEC can steer PPI discovery towards higher confidence due to the consensus between native XL-MS and EC evidence. This performance further increases when using the

geometrical information from structural models generated by knowledge-based docking or AlphaFold-Multimer.

Chapter 4

Conclusion

Protein-protein interactions (PPIs) play essential roles in almost all biological processes and diseases. The molecular understanding of PPIs is fundamental to unravel the underlying processes that harness living organisms and therapeutically intervene, should the need arise. Despite significant advancements in experimental and computational techniques over the past seven decades, achieving a comprehensive understanding of the proteome-wide PPIs remains an elusive goal. The study of PPI requires answering some major questions, including; What proteins interact with each other? What is the structural basis of the interaction? What are the functional relevance and kinetics of the interaction? How to regulate the interaction?

Throughout the research of this thesis, we have made significant strides in characterizing PPIs on the large scale of proteomes. Our investigations have encompassed the development of novel methodologies, the refinement, and exploitation of existing techniques, as well as the integration of experimental and computational approaches. These comprehensive approaches have enabled us to expand our knowledge of PPIs for challenging interactomes such as those of *E.coli* cell envelope, mammalian mitochondria, and human lysosomes.

Specifically, we conducted a comprehensive review of the high-throughput techniques for PPI characterization. We highlighted their domains of validity and limitation over the same set of proteins and the promising outcome of their potential integration. We spotted the light on two approaches, Evolutionary Coupling (EC) and Cross-Linking Mass Spectrometry (XL-MS), due to their high-throughput yields that exceed the interaction discovery to its the structural resolution and benefits of their potential integration.

Through pipeline development, we expanded the applicability of the evolutionary coupling approach, EVcomplex 2.0, fivefold among prokaryotic proteins and made the method applicable to many eukaryotic proteins upon implementing a new concatenation procedure and

probabilistic inference. Using the developed approach, we screened the protein interactome of the entire *E. coli* cell envelope, identifying and resolving hundreds of novel and known PPIs. We built structural models for protein complexes with known biological importance including the Tol-Pal system and the Flagellar Hook-Filament Junction.[53] We also demonstrated the successful application of the approach on, the eukaryotic-exclusive complex, the human spliceosome.[55]

Additionally, we investigated the first protein interaction network of the endolysosomal compartment from the HEK293 cells identified by LC-MS/MS experiment. For this, we developed a pipeline for structural evaluation and validation of cross-linking data based on their known structures and models. We supported our results with validation of newly discovered interactions (e.g., ATP6V1D-FZD9, GNB4-FLOT2) and structural models of known biological relevance (e.g., FLOT1-FLOT2 heterodimer, PPT1 tetramer). This study enhanced our understanding of the endolysosomal interactome, shedding light on previously unexplored PPIs and their structural characteristics.

Furthermore, we developed an integrative framework called XLEC that synergistically combines large-scale data from Evolutionary Coupling (EC) and Cross-Linking Mass Spectrometry (XL-MS) for enhanced prediction and structural modeling of PPIs. Our assessment showed the increased performance of XLEC in identifying and characterizing protein complexes compared to the individual approaches. We highlighted the unique advantages of integrating XLEC with machine learning approaches such as AlphaFold-Multimer. Using XLEC, we generated novel structural insights into the mouse mitochondrial interactome with hundreds of structurally resolved and modeled PPIs.

Overall, this research emphasizes the importance of a multi-faceted and integrative approach to tackle the challenges associated with PPI characterization and prediction. The heterogeneous nature of PPIs necessitates continued efforts to refine and advance existing methodologies and develop new techniques that will allow us to unlock the full potential of proteome-wide PPIs characterization. While significant progress has been made, there is still much to explore and discover in this field: How to robustly characterize an interactome on a systems level to completeness? What strategies to employ to define biological relevancy and identify regulatory mechanisms underlying PPI networks? How to capture the dynamic nature of PPIs under native conditions?

In recent years, machine learning has emerged as a powerful tool that holds great promise for accelerating and enhancing the study of PPIs. By leveraging vast amounts of experimental data and sophisticated algorithms, machine learning approaches can comprehend complex corrections and provide novel insights into PPIs. However, challenges remain in the appli-

cation of machine learning to PPI research, such as ensuring the reliability and accuracy of predictions and estimating data bias of the used algorithms. Additionally, the interpretability of machine learning approaches for PPI prediction and modeling and what they tell us about protein folding and interaction are areas of great interest and ongoing investigation.

As we look to the future, it is clear that the study of PPIs will continue to be a rich area of research, offering endless opportunities for exploration and discovery. Through collaborative efforts between experimental and computational researchers and combining the strengths of their approaches, we can unlock new insights, accelerate discoveries, and ultimately deepen our understanding of the intricate world of PPIs.

Bibliography

[1] Berggård T., Linse S., and James P. (2007). Methods for the detection and analysis of protein—protein interactions. *Proteomics*, 7(16):2833–2842. 1

[2] Rajagopala S. V., et al. (2014). The binary protein—protein interaction landscape of *Escherichia coli*. *Nature Biotechnology*, 32(3):285–290. 1, 2, 17, 19

[3] Elhabashy H., Merino F., Alva V., Kohlbacher O., and Lupas A. N. (2022). Exploring protein—protein interactions at the proteome level. *Structure*. 2, 15

[4] Gaber A. and Pavšič M. (2021). Modeling and structure determination of homo—oligomeric proteins: An overview of challenges and current approaches. *International Journal of Molecular Sciences*, 22(16):9081. 2

[5] Keskin, Ozlem N. T. and Gursoy A. (2016). Predicting protein—protein interactions from the molecular to the proteome level. *Chemical Reviews*, 116(8):4884–4909. 2

[6] Shoemaker B. A. and Panchenko A. R. (2007). Deciphering protein—protein interactions. Part I. Experimental Techniques and Databases. *PLOS Computational Biology*, 3(3). 3

[7] Mi Zhou Q. L. and Wang R. (2016). Current experimental methods for characterizing protein—protein interactions. *ChemMedChem*, 11(8):738–756. 3

[8] Nealon J. O., Philomina L. S., and McGuffin L. J. (2017). Predictive and experimental approaches for elucidating protein—protein interactions and quaternary structures. *International Journal of Molecular Sciences*, 18(12):2623. 3

[9] Salwinski L. and Eisenberg D. (2003). Computational methods of analysis of protein—protein interactions. *Current Opinion in Structural Biology*, 13(3):377–382. 3

[10] Crick F. *What Mad Pursuit: A Personal View of Scientific Discovery*. Penguin books. Penguin (1990). 3

[11] Ingles-Prieto A., et al. (2013). Conservation of protein structure over four billion years. *Structure*, 21(9):1690–1697. 3

[12] Göbel U., Sander C., Schneider R., and Valencia A. (1994). Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18(4):309–317. 3

[13] Yanofsky C., Horn V., and Thorpe D. (1964). Protein structure relationships revealed by mutational analysis. *Science*, 146(3651):1593–1594. 3

[14] Benner S. A. and Gerloff D. (1991). 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. *Advances in Enzyme Regulation*, 31:121–181.

[15] Taylor W. R. and Hatrick K. (1994). Compensating changes in protein multiple sequence alignments. *Protein Engineering, Design and Selection*, 7(3):341–348. 3

[16] Fitch W. M. and Markowitz E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics*, 4(5):579–593. 3

[17] Chiu D. K. and Kolodziejczak T. (1991). Inferring consensus structure from nucleic acid sequences. *Bioinformatics*, 7(3):347–352.

[18] Korber B. T., Farber R. M., Wolpert D. H., and Lapedes A. S. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proceedings of the National Academy of Sciences*, 90(15):7176–7180. 3

[19] Lockless S. W. and Ranganathan R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299. 3

[20] Dunn S. D., Wahl L. M., and Gloor G. B. (2008). Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340. 3

[21] Fodor A. A. and Aldrich R. W. (2004). On evolutionary conservation of thermodynamic coupling in proteins. *Journal of Biological Chemistry*, 279(18):19046–19050. 3

[22] Weigt M., White R. A., Szurmant H., Hoch J. A., and Hwa T. (2009). Identification of direct residue contacts in protein—protein interaction by message passing. *Proceedings of the National Academy of Sciences*, 106(1):67–72. 3

[23] Morcos F., et al. (2011). Direct—coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301. 3

[24] Jones D. T., Buchan D. W., Cozzetto D., and Pontil M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28(2):184–190. 4

[25] Balakrishnan S., Kamisetty H., Carbonell J. G., Lee S.-I., and Langmead C. J. (2011). Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078. 4

[26] Ekeberg M., Lövkvist C., Lan Y., Weigt M., and Aurell E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Physical Review E*, 87(1):012707. 4

[27] Mao W., Kaya C., Dutta A., Horovitz A., and Bahar I. (2015). Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution. *Bioinformatics*, 31(12):1929–1937. 7

[28] Hopf T. A., et al. (2014). Sequence co—evolution gives 3D contacts and structures of protein complexes. *eLife*, 3:e03430. 7, 17

[29] Mintseris J. and Weng Z. (2005). Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences*, 102(31):10930–10935. 7

[30] Clegg C. and Hayes D. (1974). Identification of neighbouring proteins in the ribosomes of *Escherichia coli*: A topographical study with the cross—linking reagent dimethyl suberimidate. *European Journal of Biochemistry*, 42(1):21–28. 7

[31] Sun T.-T., Bollen A., Kahan L., and Traut R. R. (1974). Topography of ribosomal proteins of the *Escherichia coli* 30S subunit as studied with reversible cross—linking reagent methyl 4—mercaptobutyrimidate. *Biochemistry*, 13(11):2334–2340.

[32] Chang F. and Flaks J. G. (1972). The specific cross—linking of two proteins from the *Escherichia coli* 30S ribosomal subunit. *Journal of Molecular Biology*, 68(1):177–180. 7

[33] Rappsilber J. (2011). The beginning of a beautiful friendship: cross—linking/mass spectrometry and modelling of proteins and multi—protein complexes. *Journal of Structural Biology*, 173(3):530–540. 7

[34] Leitner A. (2016). Cross—linking and other structural proteomics techniques: how chemistry is enabling mass spectrometry applications in structural biology. *Chemical Science*, 7(8):4792–4803. 8

[35] Ehrlich P. (1891). Experimentelle untersuchungen über immunität. i. ueber ricin. *DMW-Deutsche Medizinische Wochenschrift*, 17(32):976–979. 14

[36] Hedin S. (1906). Trypsin and antitrypsin. *Biochemical Journal*, 1(10):474. 14

[37] Braun P. and Gingras A.-C. (2012). History of protein—protein interactions: From egg—white to complex networks. *Proteomics*, 12(10):1478–1498. 14

[38] Brenner S. *Loose ends*, volume 5. Elsevier (1995). 14

[39] Jumper J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589. 16, 23, 24, 32

[40] Baek M., et al. (2021). Accurate prediction of protein structures and interactions using a three—track neural network. *Science*, 373(6557):871–876.

[41] Evans R., et al. (2022). Protein complex prediction with AlphaFold—Multimer. *BioRxiv*, pages 2021–10. 30

[42] Chowdhury R., et al. (2022). Single—sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, pages 1–7. 16

[43] Robinson C. V., Sali A., and Baumeister W. (2007). The molecular sociology of the cell. *Nature*, 450(7172):973–982. 16

[44] Green A. G., et al. (2021). Large—scale discovery of protein interactions at residue resolution using co—evolution calculated from genomic sequences. *Nature Communications*, 12(1):1–12. 17, 20

[45] Ovchinnikov S., Kamisetty H., and Baker D. (2014). Robust and accurate prediction of residue—residue interactions across protein interfaces using evolutionary information. *eLife*, 3:e02030. 17

[46] Feinauer C., Szurmant H., Weigt M., and Pagnani A. (2016). Inter—protein sequence co—evolution predicts known physical interactions in bacterial ribosomes and the Trp operon. *PLOS One*, 11(2):e0149166.

[47] Zeng H., et al. (2018). ComplexContact: a web server for inter—protein contact prediction using deep learning. *Nucleic Acids Research*, 46(W1):W432–W437. 17

[48] Díaz-Mejía J. J., Babu M., and Emili A. (2008). Computational and experimental approaches to chart the *Escherichia coli* cell—envelope—associated proteome and interactome. *FEMS Microbiology Reviews*, 33(1):66–97. 19

[49] Bonsor D. A., et al. (2009). Allosteric β—propeller signalling in TolB and its manipulation by translocating colicins. *The EMBO Journal*, 28(18):2846–2857. 19

[50] Walburger A., Lazdunski C., and Corda Y. (2002). 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(3):695–708. 19

[51] Lloubès R., et al. (2001). The Tol—Pal proteins of the *Escherichia coli* cell envelope: an energized system required for outer membrane integrity? *Research in Microbiology*, 152(6):523–529. 19

[52] Gray A. N., et al. (2015). Coordination of peptidoglycan synthesis and outer membrane constriction during *Escherichia coli* cell division. *eLife*, 4:e07118. 19

[53] Homma M., DeRosier D. J., and Macnab R. M. (1990). Flagellar hook and hook—associated proteins of *Salmonella typhimurium* and their relationship to other axial components of the flagellum. *Journal Of Molecular Biology*, 213(4):819–832. 19, 22, 38

[54] Bulieris P. V., Shaikh N. H., Freddolino P. L., and Samatey F. A. (2017). Structure of FlgK reveals the divergence of the bacterial hook—filament junction of *Campylobacter*. *Scientific Reports*, 7(1):1–12. 19, 22

[55] Will C. L. and Lührmann R. (2011). Spliceosome structure and function. *Cold Spring Harbor perspectives in biology*, 3(7):a003707. 20, 38

[56] Ulrich A. K., Seeger M., Schütze T., Bartlick N., and Wahl M. C. (2016). Scaffolding in the Spliceosome via single α—helices. *Structure*, 24(11):1972–1983. 21

[57] Ballabio A. and Bonifacino J. S. (2020). Lysosomes as dynamic regulators of cell and organismal homeostasis. *Nature Reviews Molecular Cell Biology*, 21(2):101–118. 23

[58] Shin H. R. and Zoncu R. (2020). The lysosome at the intersection of cellular growth and destruction. *Developmental Cell*, 54(2):226–238. 23

[59] Liu G. Y. and Sabatini D. M. (2020). mTOR at the nexus of nutrition, growth, ageing and disease. *Nature Reviews Molecular Cell Biology*, 21(4):183–203. 23

[60] Cabukusta B. and Neefjes J. (2018). Mechanisms of lysosomal positioning and movement. *Traffic*, 19(10):761–769.

[61] Wang L., Wu D., Robinson C. V., Wu H., and Fu T.-M. (2020). Structures of a complete human V—ATPase reveal mechanisms of its assembly. *Molecular Cell*, 80(3):501–511. 26

[62] Garg S., et al. (2011). Lysosomal trafficking, antigen presentation, and microbial killing are controlled by the Arf—like GTPase Arl8b. *Immunity*, 35(2):182–193.

[63] Balderhaar H. J. k. and Ungermann C. (2013). CORVET and HOPS tethering complexes—coordinators of endosome and lysosome fusion. *Journal of Cell Science*, 126(6):1307–1316.

[64] Skowyra M. L., Schlesinger P. H., Naismith T. V., and Hanson P. I. (2018). Triggered recruitment of ESCRT machinery promotes endolysosomal repair. *Science*, 360(6384):eaar5078. 23

[65] Fraldi A., Klein A. D., Medina D. L., and Settembre C. (2016). Brain disorders due to lysosomal dysfunction. *Annual Review of Neuroscience*, 39(1):277–295. 23

[66] Platt F. M., dÁzzo A., Davidson B. L., Neufeld E. F., and Tifft C. J. (2018). Lysosomal storage diseases. *Nature Reviews Disease Primers*, 4(1):1–25. 23

[67] Singh J., et al. (2022). Cross-linking of the endolysosomal system reveals potential flotillin structures and cargo. *Nature Communications*, 13(1):6212. 23

[68] Klykov O., et al. (2018). Efficient and robust proteome-wide approaches for cross-linking mass spectrometry. *Nature protocols*, 13(12):2964–2990. 23

[69] Varadi M., et al. (2022). AlphaFold protein structure database: massively expanding the structural coverage of protein—sequence space with high—accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444. 23, 24

[70] Consortium U. (2015). UniProt: a hub for protein information. *Nucleic acids research*, 43(D1):D204–D212. 24

[71] Altschul S. F., Gish W., Miller W., Myers E. W., and Lipman D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410. 24

[72] Helen M. Berman, John Westbrook Z. F. G. G. T. N. B. H. W. I. N. S. and Bourne P. E. (2000). The protein data bank. *Nucleic Acids Res*, 28:235–242. 24

[73] Dominguez C., Boelens R., and Bonvin A. M. (2003). HADDOCK: a protein—protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737. 24, 26, 30

[74] Ferrari A. J., et al. (2019). TopoLink: evaluation of structural models using chemical crosslinking distance constraints. *Bioinformatics*, 35(17):3169–3170. 24

[75] Szklarczyk D., et al. (2021). The STRING database in 2021: customizable protein—protein networks, and functional characterization of user—uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612. 25

[76] Maxson M. E. and Grinstein S. (2014). The vacuolar—type H+—ATPase at a glance—more than a proton pump. *Journal of Cell Science*, 127(23):4987–4993. 26

[77] Grainger S., et al. (2019). EGFR is required for Wnt9a—Fzd9b signalling specificity in haematopoietic stem cells. *Nature Cell Biology*, 21(6):721–730. 26

[78] Bellizzi III J. J., et al. (2000). The crystal structure of palmitoyl protein thioesterase 1 and the molecular basis of infantile neuronal ceroid lipofuscinosis. *Proceedings of the National Academy of Sciences*, 97(9):4573–4578. 26

[79] Lyly A., et al. (2007). Glycosylation, transport, and complex formation of palmitoyl protein thioesterase 1 (PPT1)—distinct characteristics in neurons. *BMC Cell Biology*, 8(1):1–14. 26

[80] Van Zundert G., et al. (2016). The HADDOCK2. 2 web server: user—friendly integrative modeling of biomolecular complexes. *Journal of Molecular Biology*, 428(4):720–725. 26, 30

[81] Rivera-Milla E., Stuermer C., and Málaga-Trillo E. (2006). Ancient origin of reggie (flotillin), reggie—like, and other lipid—raft proteins: convergent evolution of the SPFH domain. *Cellular and Molecular Life Sciences CMLS*, 63(3):343–357. 26

[82] Mirdita M., et al. (2022). ColabFold: making protein folding accessible to all. *Nature Methods*, pages 1–4. 26

[83] Schägger H. and Pfeiffer K. (2000). Supercomplexes in the respiratory chains of yeast and mammalian mitochondria. *The EMBO Journal*, 19(8):1777–1783. 29

[84] Liu F., Lössl P., Rabbitts B. M., Balaban R. S., and Heck A. J. (2018). The interactome of intact mitochondria by cross—linking mass spectrometry provides evidence for coexisting respiratory supercomplexes. *Molecular & Cellular Proteomics*, 17(2):216–232. 29

[85] Lensink M. F., Méndez R., and Wodak S. J. (2007). Docking and scoring protein complexes: CAPRI 3rd edition. *Proteins: Structure, Function, and Bioinformatics*, 69(4):704–718. 30

[86] Basu S. and Wallner B. (2016). DockQ: a quality measure for protein—protein docking models. *PLOS One*, 11(8):e0161879. 30

Chapter 5

List of Publications

Elhabashy H., Netz E., and Kohlbacher O., 2023. XLEC: Large-scale prediction and modeling of protein-protein interactions by combining sequence co-evolution and cross-linking data. *Manuscript under review.*

Singh, J., **Elhabashy H.**, Muthukottiappan, P., Stepath, M., Eisenacher, M., Kohlbacher, O., Gieselmann, V. and Winter, D., 2022. Cross-linking of the endolysosomal system reveals potential flotillin structures and cargo. *Nature Communications, 13*(1), p.6212.

Elhabashy H., H., Merino, F., Alva, V., Kohlbacher, O. and Lupas, A.N., 2022. Exploring protein—protein interactions at the proteome level. *Structure, 30*, p.462.

Green, A.G.*, **Elhabashy H.***, Brock, K.P., Maddamsetti, R., Kohlbacher, O. and Marks, D.S., 2021. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nature communications, 12*(1), p.1396. *Co-first authors

Biggers, L., **Elhabashy H.**, Ackad, E. and Yousef, M.S., 2020. Molecular dynamics simulations of an engineered T4 lysozyme exclude helix to sheet transition, and provide insights into long distance, intra—protein switchable motion. *Protein Science, 29*(2), pp.542-554.

Chapter 6

Contributions

All ideas, approaches and results presented in this work were developed and discussed with my supervisors Prof. Dr. Oliver Kohlbacher (O.K.) and Prof. Dr. Andrei Lupas (A.L.). The following co-workers also contributed to the different projects:

- Dr. Felipe Merino (F.M.)
- Dr. Vikram Alva (V.A.)
- Dr. Anna Green (A.G.)
- Dr. Kelly Brock (K.B.)
- Dr. Rohan Maddamsetti (R.M.)
- Prof. Dr. Debora Marks (D.M.)
- Dr. Jasjot Singh (J.S.)
- Pathma Muthukottiappan (P.M.)
- Markus Stepath (M.S.)
- Martin Eisenacher (M.E.)
- Prof. Dr. Dominic Winter (D.W.)
- Prof. Dr. Volkmar Gieselmann (V.G.)
- Eugen Netz (E.N.)

Paper one: Exploring protein-protein interactions at the proteome level

H.E., V.A., and F.M. developed the concept and outline for the article, with advice from A.L. and O.K. All of the authors wrote the paper and discussed and edited the final version.

Paper Two: Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences

A.G. and H.E. contributed to the conception and design of the research; acquisition, analysis, and interpretation of data; manuscript drafting and revision. A.G. contributed to the creation of software and developed the scoring model. H.E. performed the structural analysis and modeling of the complexes. K.B. obtained and analyzed human proteome alignment data and assisted with the manuscript. R.M. obtained and analyzed *E. coli* proteome alignment data and assisted with study design. D.M. and O.K. contributed to the conception and design of the research, data interpretation, manuscript writing, and supervision of the study. All authors discussed the results and contributed to the final manuscript.

Paper Three: Cross-linking of the endolysosomal system reveals potential flotillin structures and cargo

Conceptualization: J.S. and D.W.; methodology: J.S., V.G., D.W.; bioinformatic analysis and structural modeling: J.S., H.E., M.S., and O.K.; investigation: J.S. and P.M.; writing-original draft: J.S. and D.W.; writing-review & editing: J.S., H.E., P.M., M.E., O.K., V.G., and D.W.; visualization: J.S., H.E., and P.M.; funding acquisition: J.S. and D.W.; resources: V.G. and D.W.; supervision: D.W. All authors contributed to the final manuscript.

Paper Four: XLEC: Large-scale prediction and modeling of protein-protein interactions by combining sequence co-evolution and cross-linking data

H.E. and O.K. developed the concept and outline for the study. H.E. wrote the code, performed the calculations, did the analysis and structural modeling. H.E., E.N. and O.K. wrote and revised the manuscript. O.K. supervised the study. All authors read, discussed and edited the final version of this manuscript.

Chapter 7

Abbreviations

3D	*Three-dimensional*
APMS	*Affinity Purification Mass Spectrometry*
BLAST	*Basic Local Alignment Search Tool*
BN-PAGE	*Blue Native-Polyacrylamide Gel Electrophoresis*
CNS	*Crystallography and NMR system*
Cryo-EM	*Cryo-Electron Microscopy*
DSSO	*Disuccinimidyl Sulfoxide*
DR	*Disrupted*
DCA	*Direct Coupling Analysis*
EC	*Evolutionary Coupling*
E. coli	Escherichia coli
F_nat	*Fraction of Interface Native Contacts*
GAL4	*Yeast Transcription Factor*
GAL4-BD	*Yeast Transcription Factor Binding Domain*
GAL4-AD	*Yeast Transcription Factor Activation Domain*
HADDOCK	*High Ambiguity Driven Protein-Protein Docking*
HUPO	*Human Proteome Organization*
Co-IP	*Co-Immunoprecipitation*
IT	*Intact*
IMPRS	*International Max Planck Research School*
iBAQ	*Intensity-based Absolute Quantification*
I-RMSD	*Interface Root Mean Square Deviation*
L-RMSD	*Ligand Root Mean Square Deviation*
LSDs	*Lysosomal Storage Disorders*
MSA	*Multiple Sequence Alignment*
MS	*Mass Spectrometry*

MVP	*Major Vault Protein*
MI	*Mutual Information*
NMR	*Nuclear Magnetic Resonance.*
OMES	*Observed Minus Expected Squared*
PIN	*Protein Interaction Network*
PPI	*Protein-Protein Interaction*
PFAM	*Protein Families Database*
PDB	*Protein Data Bank*
PLL	*Poly-L-Lysine*
plmDCA	*Pseudo Likelihood Maximization Direct Coupling Analysis*
PBS	*Phosphate Buffered Saline*
PNS	*Post-Nuclear Supernatant*
PSI	*Proteomics Standard Initiative*
PSICOV	*Precise Structural Contact Prediction using Sparse Inverse Covariance*
RMSD	*Root Mean Square Deviation*
STRING	*Search Tool for the Retrieval of Interacting Genes/Proteins*
SPR	*Surface Plasmon Resonance*
SPIONs	*Super-Paramagnetic Iron Oxide Nanoparticles*
SCX	*Strong Cation-Exchange*
SCA	*Statistical Coupling Analysis*
UniProt	*Universal Protein Resource database*
VS	*Virtual Screening*
WGCNA	*Weighted Correlation Network Analysis*
XL	*Cross-linking*
XLEC	*Cross-linking and Evolutionary Coupling*
XL-MS	*Cross-linking Mass Spectrometry*
Y2H	*Yeast Two-Hybrid*

Appendix A

Paper one: Exploring protein-protein interactions at the proteome level

CellPress

Structure

Review

Exploring protein-protein interactions at the proteome level

Hadeer Elhabashy,[1,2,3,*] Felipe Merino,[1] Vikram Alva,[1] Oliver Kohlbacher,[2,3,4,5] and Andrei N. Lupas[1,*]
[1]Department of Protein Evolution, Max Planck Institute for Biology, Tübingen, Germany
[2]Department of Computer Science, University of Tübingen, WSI/ZBIT, Sand 14, 72076 Tübingen, Germany
[3]Institute for Bioinformatics and Medical Informatics, University of Tübingen, 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
*Correspondence: hadeer.elhabashy@tuebingen.mpg.de (H.E.), andrei.lupas@tuebingen.mpg.de (A.N.L.)
https://doi.org/10.1016/j.str.2022.02.004

SUMMARY

Proteins are central to all of the processes of life. For their activity, they almost invariably need to interact with other macromolecules, be they nucleic acids, membranes, glycans, or other proteins. The interaction between proteins is indeed the most common mode of macromolecular interaction underpinning living systems. To understand these systems at a molecular level, it is therefore essential to identify and characterize their constituent protein-protein interactions. Despite an unprecedented growth in our knowledge of complete proteomes across all domains of life, both at the sequence level and increasingly at the structure level, the inherently low accuracy and molecular resolution of many techniques have made the characterization of protein-protein interactions one of the grand challenges of molecular biology. In this review, we survey both computational and experimental techniques for the medium- to high-throughput characterization of protein-protein interactions and discuss the potential of integrative approaches, given recent advances in sequence analysis and structure prediction.

INTRODUCTION

"Almost all aspects of life are engineered at the molecular level, and without understanding molecules we can only have a very sketchy understanding of life itself," as Francis Crick noted (Crick, 1988). As the main mediators of function in living systems, proteins are the key molecules we need to understand. Their activity is dependent on their interactions with other molecules, frequently with other proteins. These may be other copies of the same protein (homo-oligomers) or of different proteins (hetero-oligomers), or sometimes both, particularly in complex molecular machines. The network of all protein-protein interactions (PPIs) is referred to as the interactome. The interactome is much smaller than the sum of all possible interactions because proteins are gregarious and could associate in many ways, only a subset of which are biologically meaningful. At the same time, the interactome is much more complex than the mere number of interactions. While some proteins form stable, permanent assemblies, many proteins, for example, in signal transduction, only interact transiently and reversibly, exhibiting an extensive range of affinities and lifespans. PPIs are often also dependent on the presence of specific modifications that are introduced during or after translation (e.g., phosphorylation, glycosylation, methylation, lipidation, ubiquitination), allowing the cell to enable or suppress particular interactions in a regulated and, many times, reversible way. A further parameter is conditional structuring on a scaffold or around a specific ligand. In many cases, the scaffold is a membrane or a nucleic acid, which induces a productive conformation in the interacting proteins. Most commonly, however, the scaffold is another structured protein. Particularly in multicellular eukaryotes, many proteins contain intrinsically disordered regions, which may become at least locally structured upon encountering their partner. Correspondingly, many folded protein domains exist to mediate such interactions. Some are globular and bind specific sequence motifs on partner proteins, whereas others are repetitive arrays and generally use more distributed sequence features for recognition. Other key parameters modulating PPIs are the availability of the interacting proteins in the correct stoichiometry for assembly, their presence in specific cellular compartments for interaction, and their coordinated expression. Because of this versatility, the same interactome shows considerable spatial, temporal, and dynamic variations, making its accurate description challenging.

Despite the fundamental importance of PPIs, the field was slow in developing because of the need to establish an understanding of proteins and the complex nature of their interactions. The first PPIs were described in the late 19th century, with observations of trypsin and its inhibitor and of antibodies against plant toxins, but the recognition that both partners in these interactions were proteins was only achieved several decades later (Braun and Gingras, 2012; Northrop, 1922; Ehrlich, 1891). The emergence of molecular biology after 1950 provided steady progress in the techniques needed to establish PPIs, culminating in the use of DNA libraries for the direct detection of PPIs by phage display (Smith, 1985) and yeast two-hybrid (Y2H) screening (Van Criekinge and Beyaert, 1999). The rise of

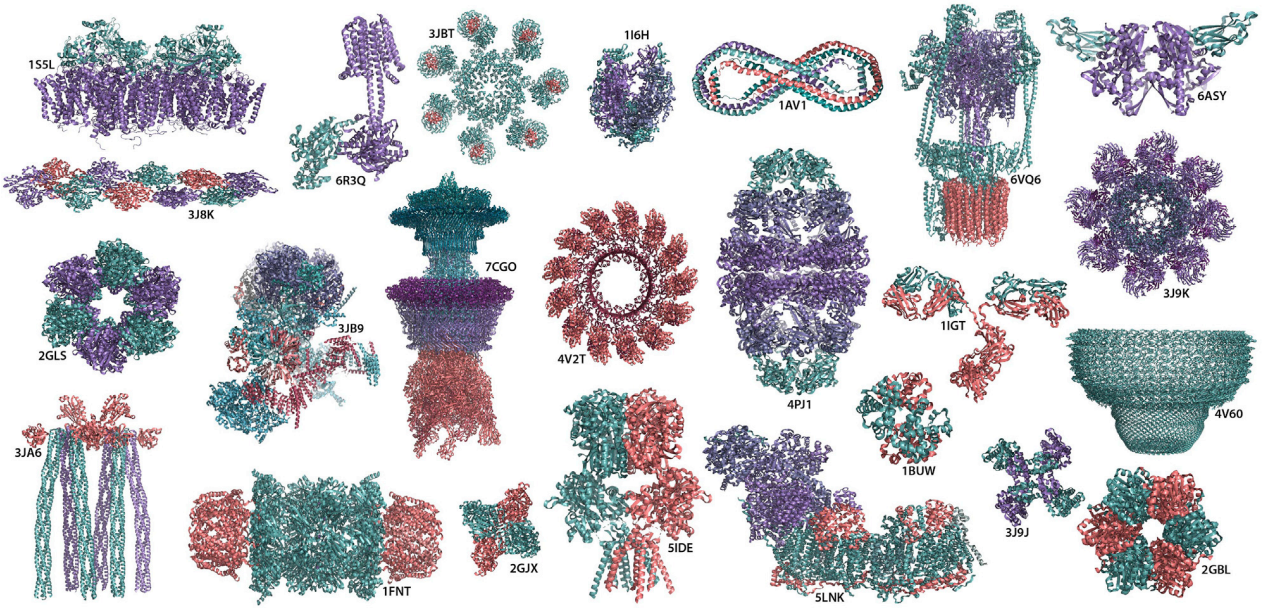


Figure 1. Molecular machinery of life

Gallery of supramolecular complexes formed through protein-protein interactions. PDB IDs of shown structures from the left to right are as follows 1S5L: photosynthetic oxygen evolving center; 3J8K: tilted state of actin T2; 2GLS: glutamine synthetase; 3JA6: novel kinase conformational switch in bacterial chemotaxis signaling; 6R3Q: membrane adenylyl cyclase bound to an activated stimulatory G protein; 3JB9: yeast spliceosome; 1FNT: 20s proteasome from yeast in complex with the proteasome activator pa26 from *Trypanosoma brucei*; 3JBT: the Apaf-1 apoptosome; 7CGO: the flagellar motor-hook complex from *Salmonella*; 2GJX: human β-hexosaminidase A; 1I6H: RNA polymerase II elongation complex; 4V2T: membrane-embedded pleurotolysin pore with 13-fold symmetry; 5IDE: GluA2/3 AMPA receptor heterotetramer (model I); 1AV1: human apolipoprotein A-I; 4PJ1: human mitochondrial chaperonin symmetrical 'football' complex; 5LNK: ovine respiratory complex I; 6VQ6: mammalian V-ATPase from rat brain; 1IGT: immunoglobulin; 1BUW: s-nitroso-nitrosyl human hemoglobin A; 6ASY: human endoplasmic reticulum chaperone BiP; 3J9K: dark apoptosome in complex with the Dronc CARD domain; 4V60: rat liver half-vault; 3J9J: capsaicin receptor TRPV1; and 2GBL: circadian clock protein KaiC

genomics in the 1990s finally enabled the system-wide analysis of PPI networks. Correspondingly, in the last 3 decades, the continuous development of experimental and computational methods for characterizing PPIs at increasing throughput levels has occurred. As sequencing technology became better and less expensive, a growing number of complete genomes from different organisms became available and comparative genomics revealed interesting organizational patterns: interacting proteins were often found in proximity within genomes, and their patterns of occurrence across genomes turned out to be correlated (Keskin et al., 2016).

While the large-scale prediction or validation of whether two proteins interact has become routine, predicting how they interact on the structural level proved to be significantly more challenging due to the low-throughput nature of structure determination techniques (X-ray crystallography, NMR, cryoelectron microscopy [cryo-EM]) (Figure 1). Consequently, even in the reasonably well-investigated model organism *Escherichia coli*, only approximately 9% of probable PPIs have been characterized at a structural level to date (Green et al., 2021), and this number is yet lower for the human interactome (Luck et al., 2020). In recent years, cross-linking mass spectrometry (XL-MS) and sequence coevolution analyses have alleviated this issue by providing not only the identity of the interacting proteins but also residue-level information about the interaction. Particularly with the advent of machine learning approaches in sequence analysis, coevolutionary correlations enabled recent breakthroughs in protein structure prediction, allowing the modeling of proteins at an accuracy that rivals experimentally determined structures (Jumper et al., 2021; Pereira et al., 2021). The study of PPIs is poised for a revolution; combining the predicted structural models for a given proteome with experimental data will accelerate the structural characterization of whole interactomes dramatically.

In this article, we discuss developments in the experimental and computational exploration of PPIs at a systems level. Our emphasis is on methods that allow, at least in principle, for the characterization of interactions at the proteome level, particularly those that can uncover information about the structural characteristics of protein complexes. For this, we have divided methods into two groups, experimental and computational, even though all of the methods rely on experimental data and use (sometimes extensive) computation. The distinction may therefore seem arbitrary, but it is in fact based on whether the experimental data were gathered specifically to elucidate PPIs ("experimental methods"), or whether they were determined for other purposes (e.g., genomic and expression data) and are now mined computationally for the inference of PPIs ("computational methods").

EXPERIMENTAL METHODS FOR PPI ELUCIDATION

High-throughput methods for interaction screening

The methods described in this section are generally used for large-scale surveys and reveal the presence of an interaction

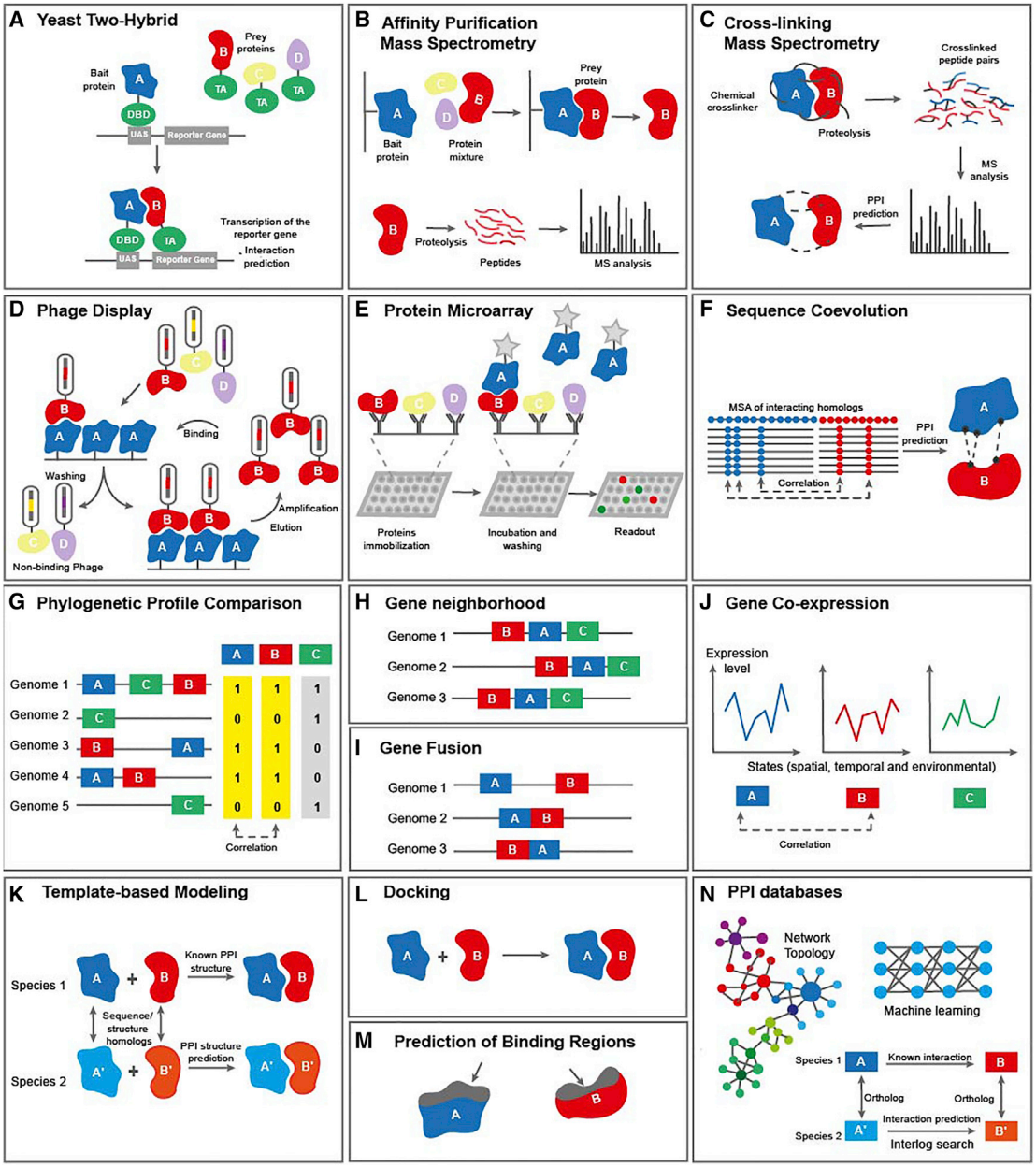


Figure 2. High-throughput PPIs methods
Conceptual illustration of the high-throughput methods for PPI prediction and structural characterization described in the text.

without revealing its molecular details. They provide catalogs of potential pairwise interactions as a starting point for the validation and structural characterization of biologically interesting connections. The main challenge facing these methods is the substantial proportion of false positives.

Two-hybrid methods

The two-hybrid technique exploits the modular nature of the yeast transcription factor GAL4 to induce the expression of a reporter gene when two proteins of interest interact (Fields and Song, 1989). GAL4 consists of two distinct domains: a DNA-binding domain (GAL4-BD) and a transcription activation domain (GAL4-AD). For the prototypical Y2H screening, cells lacking a functional GAL4 gene are transformed with two compatible expression vectors, each coding for a fusion protein comprising either GAL4-BD (bait) or GAL4-AD (prey). Successful interaction between bait and prey brings both domains of GAL4 into proximity and induces the transcription of the reporter gene (Van Criekinge and Beyaert, 1999) (Figure 2A). Y2H is one of the most common high-throughput techniques for PPI detection because it is comparatively inexpensive and easy to automate. Its biggest drawbacks are that (1) it requires nuclear localization of bait and prey and therefore is not suitable for membrane

proteins, extracellular proteins, or proteins localized in other subcellular compartments; (2) using fusion proteins always carries the risk of altering the function of the original protein, or hiding interaction sites, leading to missed interactions; and (3) PPIs modulated by post-translational modifications may be incorrectly detected when those modifications are not possible in yeast cells. To alleviate these issues, several novel alternative Y2H systems have been developed, including mammalian Y2H (Luo et al., 1997), bacterial Y2H (Karimova et al., 1998), split-luciferase (Ozawa et al., 2001), split-ubiquitin (Thaminy et al., 2004), insect Y2H (Mon et al., 2009), and next-generation sequencing Y2H systems (Suter et al., 2015).

Display methods

Phage display is one of the oldest high-throughput methods used to screen PPIs (Smith, 1985; Sidhu et al., 2003). In this approach, a gene of interest is fused to the coat protein gene of the phage, leading to the phage's displaying the fused protein on its surface. The *E. coli* filamentous bacteriophages M13, f1, and fd are widely used for phage display. Libraries of phages displaying different prey proteins are then tested for binding against a particular bait protein. Phages binding the bait are captured and amplified through consecutive passages through a bacterial host and identified by sequencing the DNA of the phages (Figure 2D) (Zhou et al., 2016). Phage display allows large-scale screening and can handle large libraries containing as many as 10^{10} different variants (Willats, 2002). One distinct advantage of phage display is that the combinatorial search for interacting partners can be combined with multiple rounds of amplification, decreasing the noise in the technique and allowing the identification of very high-affinity binders. By enabling the detection of high-affinity antibody interactions, phage display has played an essential role in biomedical research and contributed to the fields of immunology, cell biology, drug discovery, and pharmacology (Nagano and Tsutsumi, 2021). However, there are some technical drawbacks to the approach: (1) it can be difficult and time-consuming to generate high-quality libraries for large-scale PPI screening (Caberoy et al., 2010); (2) it can have unintended effects on the folding of the displayed protein; and (3) using fusion proteins may also hinder the correct folding and binding between the interacting domains. Some of these issues can be addressed using other genotype-phenotype conjugation systems, such as yeast display (Boder and Wittrup, 1997), bacterial display (Francisco et al., 1993), ribosome display (Hanes and Plückthun, 1997), and mRNA display (Roberts and Szostak, 1997; Nemoto et al., 1997).

Affinity purification (AP) coupled to MS

Display and Y2H methods cannot be used to characterize interactions that depend on additional components. A classical alternative to explore PPIs made by specific target proteins is to use AP-MS (for a review, see Dunham et al., 2012). In AP-MS, cells, tissue, or even subcellular compartments containing the protein of interest are lysed under conditions meant to preserve even weak interactions. The target protein is then specifically captured, along with some of the interacting proteins, and analyzed with MS to identify all of the co-purified proteins (Figure 2B). Contrary to other experimental techniques, every protein in a complex would theoretically be identified, even if another protein mediates their interaction with the bait. The capture of the protein complex can be achieved in many ways, most commonly by co-immunoprecipitation or purification-specific tags (e.g., FLAG; Hopp et al., 1988). Further specificity can be achieved using multiple tags and sequential APs (e.g., Puig et al., 2001). Regardless of whether antibodies or additional tags are used, there are a few drawbacks to this approach: (1) PPIs could mask the epitope being recognized during purification, essentially rendering that particular interaction silent; (2) proteins from different cellular compartments could theoretically mix during purification, raising the possibility of non-physiological interactions; and (3) weak interactions, or those requiring specific ions or small molecules, could be lost during the purification.

Proximity-based labeling

In proximity-based labeling, originally known as BioID, a protein of interest is fused to a mutant version of the biotin-ligase BirA* (Roux et al., 2012). BirA* leaks activated biotin (biotinoyl-5′-AMP) from its active site, which readily attaches to proteins within an ~10-nm distance (Kim et al., 2014). Biotin-labeled proteins are then purified using streptavidin and identified by MS. A large drawback is that cells need to be grown for several hours with biotin to obtain proper labeling. Thus, BioID is not suitable for temporal studies but rather provides an integrated view of all of the possible PPIs of a particular protein. One way by which this issue has been addressed involves replacing BirA* with the enzyme ascorbate peroxidase (APEX) (Rhee et al., 2013). In the presence of hydrogen peroxide, APEX catalyzes the oxidation of biotin-phenol into biotin-phenoxyl. Biotinilation is thus triggered by a quick exposure to H_2O_2, reducing the labeling time to ~1 min. As with BioID, this results in the biotinylation of proteins within an ~20-nm reach (Martell et al., 2012). The lack of a purification step means that even weak and transient interactions can be detected. Moreover, the technique is also applicable to membrane proteins. The versatility of this technique has quickly placed it as a good option for PPI characterization. However, proximity-based labeling is not dependent on physical PPIs, but simply on their vicinity and is thus prone to false positives. Improvements to proximity-based labeling have therefore strongly centered on generating smaller and more efficient enzymes to reduce the impact of the required protein fusion or improving labeling speed and efficiency (Lam et al., 2015; Kim et al., 2016; Ramanathan et al., 2018; Branon et al., 2018).

Protein microarrays

Protein microarrays were originally proposed by Chang (1983) as an array of antibodies meant to profile cell-surface antigens in parallel (Figure 2E). However, any soluble protein could be attached to these chips. The development of tools to print these arrays at high densities opened the possibility of high-throughput proteome-wide screening (MacBeath and Schreiber, 2000; Lee et al., 2003; Michaud et al., 2003; Zhu et al., 2001). The detection of interactions on protein arrays is typically achieved by fluorescent markers added to the analyte before the assay. Label-free PPI detection is also possible (for a review on the topic, see Ray et al., 2010), but often at the expense of throughput. Surface plasmon resonance (SPR) is one of the few methods that allows for binding detection in large arrays with the added advantage that binding kinetics can also be determined from a single experiment, and it has been successfully applied to rank antibody binding according to their affinity (Wassaf et al., 2006) and serum profiling (Lausted et al., 2008). The need for protein purification

for array production and the long-term stability of the proteins limit the applicability of the method to large proteomes. An interesting solution to both of these problems came with the development of *in situ*-generated arrays, in which coding DNA for each protein is added to the arrays and transformed into proteins by *in vitro* translation (He and Taussig, 2001; for a review, see He et al., 2008b). This tremendously simplifies array production since no purification step is required, and only a suitable *in vitro* translation system is necessary. Further developments in which the template DNA is also immobilized allows for the direct translation of an array of DNA into a protein array (Ramachandran et al., 2004), even multiple times (He et al., 2008a). The use of these types of assays has allowed, for example, the profiling of several plant transcription factors against 12,000 different proteins (Yazaki et al., 2016).

Experimental elucidation of structural details of protein complexes

Besides the mere information of whether two proteins interact, it is often important to understand how they interact—in other words, obtain information on the structural details of the interaction. In contrast to the methods from the previous section, these methods tend to be used less in high-throughput settings. They are usually focused on the characterization of interactions within cellular subsystems or even individual complexes.

XL-MS

XL-MS combines the chemical cross-linking of proteins with the rapid identification of proteins through MS. While cross-linking paired with gel electrophoretic approaches were used to reveal proximity and interaction between ribosomal proteins already in the early 1970s (Chang and Flaks, 1972; Clegg and Hayes, 1974; Sun et al., 1974), its combination with MS-based proteomics (Aebersold and Mann, 2003) has gained general applicability for PPI detection in recent years (Back et al., 2003; Trakselis et al., 2005; Sinz, 2006, 2007; Mouradov et al., 2008; Tang and Bruce, 2009; Singh et al., 2010a, 2010b). XL-MS can identify both the existence of a PPI and the structural details by revealing the localization of the linked residues (Rappsilber, 2011). In XL-MS experiments, two or more proteins in a complex are covalently cross-linked *in vitro* or *in vivo*. Cross-linked proteins are then digested enzymatically, resulting in a pool of peptides with a small fraction of cross-linked peptide pairs. After the enrichment of these cross-linked peptides, the identity of the peptides is determined by (tandem) MS (Leitner, 2016) (Figure 2C). XL-MS has gained much interest because it is fast, cost-effective, and relatively simple. XL-MS has been applied successfully to study individual purified proteins and protein complexes. Since the native cellular environment plays a crucial role in protein conformations and interactions that are not observable with purified proteins and complexes, there have been recent advances in the high-throughput *in vivo* investigation of proteome-wide interactions using XL-MS (Rinner et al., 2008; Liu et al., 2015, 2018). For instance, recent studies profiling the mitochondrial interactome achieved ~50% proteome coverage and identified thousands of cross-linked peptides with structural information and hundreds of *de novo* PPIs (Schweppe et al., 2017; Liu et al., 2018).

XL-MS combines the scalability of high-throughput methods with the ability to reveal structural details of the interactions in an *in vivo* context. The spatial proximity information (based on the length of the cross-linker) provides essential information for structural modeling of the interactions. Nevertheless, XL-MS techniques have some drawbacks: (1) the identification of cross-linked peptide pairs is one of the main challenges for high-throughput XL-MS since the possible peptide pairing search space increases quadratically with the number of peptides in the database; (2) they suffer from a high false negative rate caused by the sensitivity of MS and are thus biased to the most abundant cross-linked peptides; and (3) the existence of a cross-link does not necessarily imply functionally relevant interactions, but it can be caused by non-specific interactions as well.

Cryo-EM for structural proteomics

Recent advances in electron detectors and image processing algorithms have revolutionized the use of single-particle analysis and cryo-EM for structure determination, allowing the routine determination of macromolecular structures at near-atomic resolution (Kühlbrandt, 2014). Without the requirement for crystallization, this technique allows the structural characterization of hard-to-obtain proteins or macromolecular complexes in near-native conditions. Since sample heterogeneity can be, at least in theory, sorted computationally, the idea of using cryo-EM to determine the structures of complex mixtures has always been present in the community. Although a straightforward idea, a real-world application is challenging. Combining classification with *ab initio* reconstructions is a difficult task, especially for very complex samples. Moreover, guessing the sequence of a protein based merely on its density is a problem that has only recently been addressed, albeit with very promising results. Su and colleagues identified and obtained several near-atomic resolution structures from a single sample of partly purified proteins, using an iterative approach of classification and refinement called "build and retrieve" (Su et al., 2021). Likewise, Ho and colleagues used a similar approach to obtain multiple structures from enriched fractions of cell lysates (Ho et al., 2020). To identify the proteins, they developed an algorithm called CryoID, which compares the sequences guessed from the maps to a database. CryoID successfully identified proteins for resolutions better than 4 Å, and could identify several chains in a protein complex, paving the way for cryo-EM-based structural proteomics.

Recently, cryoelectron tomography has experienced an algorithmic revolution of its own; modern processing methods allow the determination of near-atomic resolution structures directly from particles, or sub-tomograms in this case, inside intact cells (Tegunov et al., 2021; Himes and Zhang, 2018). Observing protein *in situ* has the advantage that complexes are kept intact, allowing for the structural characterization of proteins in their native environment, thus providing an unprecedented level of detail in the inner working of cells and their pathologies (see, for instance, Mahamid et al., 2016; Guo et al., 2018; Allegretti et al., 2020). Could sub-tomogram averaging be used to identify complexes *de novo* in their native environments? Certainly, for well-understood samples such as sarcomeres, this could be possible, as highlighted by the recent identification of the binding site of nebulin in the thin filament (Wang et al., 2021). Currently, near-atomic resolution can be reached with complexes present in very large amounts and it is likely impossible to reach them for low-abundance complexes. There is, however, a possible way

forward. With the advent of high-accuracy protein structure predictions (Jumper et al., 2021), it has become possible to compare experimental maps to atomic model databases instead of sequences, facilitating protein identification at lower resolutions.

COMPUTATIONAL TOOLS FOR PPI PREDICTION

The data generated using experimental methods for PPI elucidation have been integrated and curated in a number of databases. These databases are essential resources for large-scale analyses of interactions, but also a convenient way to identify known and potential interaction partners for individual proteins. However, there are also purely computational methods available, which permit the prediction of highly probable interactions between two proteins as well as insights into structural details of such a complex.

Protein-protein interaction databases

The wealth of PPI data generated throughout the years have largely been made available in public PPI databases (Nakajima et al., 2021). In these databases, interactions from various sources, datasets, as well as experimentally or computationally substantiated PPIs are stored and integrated (Figure 2N). However, different databases focus on different aspects—for instance, on specific techniques, organisms, or degrees of curation or annotation (Table S1). It was estimated that error rates in the data curation and false positive rates in high throughput can reach up to 45% (Huang et al., 2007). Some PPI databases (e.g., STRING) use information from biological network topology and text mining to improve data quality (Blaschke et al., 2001; Przulj et al., 2004; Papanikolaou et al., 2015). To render these databases (and related tools) interoperable, several data exchange formats have been proposed by the Proteomics Standards Initiative (PSI) of the Human Proteome Organization (HUPO), in particular, the XML-based format PSI-MI XML (miXML) and the tabular format PSI-MI TAB (miTab) (Deutsch et al., 2017). PPI databases are very useful for an initial screen and prioritization of potential PPIs for further validation.

Prediction of PPIs from genome data

Physically interacting proteins are often co-located within genomes, and their genetic proximity can sometimes lead to gene fusion events. Furthermore, in prokaryotes, interacting genes tend to cluster in conserved operons (Mushegian and Koonin, 1996). In eukaryotes, the co-localization of functionally associated genes is less pronounced, but still relevant (Dandekar et al., 1998; Lee and Sonnhammer, 2003; Al-Shahrour et al., 2010; Dávila López et al., 2010). Early on, methods using phylogenetic comparisons of gene co-occurrence could be used to predict association, either functional or physical, between non-homologous proteins (Pellegrini et al., 1999). Over the last 2 decades, multiple statistical approaches have been developed for phylogenetic profile comparison, including naive (Pellegrini et al., 1999; Huynen et al., 2000; Glazko and Mushegian, 2004; Brilli et al., 2008), phylogeny-aware (Vert, 2002; Barker and Pagel, 2005; Barker et al., 2007; Kensche et al., 2008; Csurös, 2010), and consensus methods (Tremblay et al., 2021). Phylogenetic profile comparison has led to numerous discoveries in the functional linkage of uncharacteristic proteins to conserved pathways (Kensche et al., 2008; Jothi et al., 2007). Phylogenetic analysis of gene fusion events can also be used to infer direct PPIs (Chia and Kolatkar, 2004). Fusion events have been reported for almost all of the domains of life and is evidenced through the large number of multi-domain proteins (Apic et al., 2001). Structural analysis of high-throughput interaction data has shown that gene fusion events can optimize and simplify the topology of protein complex assemblies (Marsh et al., 2013). As these approaches have low prediction accuracy, they are primarily used as additional lines of evidence or for initial and large-scale screening preceding more advanced interaction prediction methods.

Co-expression analysis

Many PPIs require the expression of the interaction partners in the correct stoichiometry (Shoemaker and Panchenko, 2007). The correlated expression of genes can thus serve as an indication of a PPI (Figure 2J). Furthermore, the co-expression of interacting proteins may co-evolve across species, a fact that can be integrated into predictions as well (Fraser et al., 2004). Several methods have been developed to disentangle co-expression signals between interacting proteins (Kemmeren et al., 2002; Steffen et al., 2002; Zhou et al., 2002; Kim et al., 2003; Stuart et al., 2003; Schäfer and Strimmer, 2005; Chuang et al., 2008; Zhang and Horvath, 2005) (Table S1). Since its introduction in 2005, the weighted gene co-expression network analysis (WGCNA) has become one of the most effective and widely used methods for gene co-expression network analysis (Zhang and Horvath, 2005; Langfelder and Horvath, 2008; Zhao et al., 2010). WGCNA has been successfully used to unravel important functional correlations in cancer research (Giulietti et al., 2017; Li et al., 2019; Liu et al., 2021), proteomics, and transcriptomics (Zheng et al., 2021; Bordini et al., 2021; Liu et al., 2021; He et al., 2021). High-throughput experiments have made it possible to measure the expression profiles of genes on the genome level, which can be easily accessible via open libraries or web servers, but they can also be noisy and populated by false positives.

Prediction of structural details of protein complexes

In this section, we address methods that provide insights into interactions at atomic resolution. While interfaces obtained by template-based modeling and docking potentially carry substantial uncertainties, the determination of complex structures by cryo-EM (or co-crystallization, which we do not address here) provide a gold standard for PPI characterization.

Sequence coevolution

The evolutionary history of proteins through sequence space is strongly constrained by their structure, function, and interaction. In 1994, researchers reported the detection of evolutionary patterns of correlated mutations that contain information about protein structure (Göbel et al., 1994). Three years later, the same observation was confirmed for PPIs (Pazos et al., 1997). If a PPI is essential for the survival of an organism, then mutations in one interaction partner require a compensating mutation in the binding partner to ensure survival. Sequence coevolution analysis reconstructs the evolutionary history of proteins from extensive multiple sequence alignments and extracts correlated

mutations from this information. These correlations, the so-called evolutionary couplings, can be used to identify interacting residues between pairs of proteins (Figure 2F). Various algorithms have been developed, including statistical coupling analysis (SCA; Lockless and Ranganathan, 1999)), mutual information (MI; Gloor et al., 2005), the observed minus-expected-squared method (OMES; Fodor and Aldrich, 2004), protein sparse inverse covariance (PSICOV; Jones et al., 2012), and direct coupling analysis (DCA) with mean-field and pseudo-likelihood maximization (Weigt et al., 2009; Morcos et al., 2011; Balakrishnan et al., 2011; Ekeberg et al., 2013) (Table S1). Sequence coevolution calculations offer a purely sequence-based approach that can predict and resolve PPIs to the key interacting residues at the interface. These signals can be used effectively to improve the results of protein-protein docking. The approach can be automated and scaled up as a proteome-wide interactome profiling technique. The approach was recently used to screen fully or partially the *E. coli* interactome (Cong et al., 2019; Green et al., 2021). However, evolutionary coupling approaches require sufficient depth and diversity of the underlying multiple sequence alignments, limiting their applicability in some cases (Mao et al., 2015). The presence of paralogs can render the identification of the interacting sequence pairs difficult. Transient interactions have shown a faster evolutionary rate than obligate interactions, making them hardly tractable by the approach (Mintseris and Weng, 2005). The approach also struggles with predicting homo-complexes when it comes to differentiating between intra- and inter-chain coevolved residues for identical sequences. Some attempts to overcome this issue make use of available structures or NMR data of the subunits involved in the interaction (Yan and Huang, 2021; Sala et al., 2020).

Docking

Protein-protein docking (PPD) denotes the computational prediction of the three-dimensional (3D) structure of a complex, given the 3D structure of the interaction partners (Vakser, 2014; Huang, 2014; Pagadala et al., 2017) (Figure 2L). PPD can be traced back to the early days of molecular modeling, in the 1970s, when global rigid body search algorithms were used to predict complex structures (Wodak and Janin, 1978; Greer and Bush, 1978). In the 1990s, algorithmic advancements, combined with improvements in energy functions to score different docking poses, enabled large-scale docking experiments. Fast Fourier transform-based docking rendered the global search for suitable docking poses computationally more tractable (Katchalski-Katzir et al., 1992). The global search algorithms use structural and biophysical features of the two protein surfaces, including shape, charges, or hydrophobicity, to identify poses with high complementarity between the interacting proteins. These poses are then usually refined using more sophisticated energy functions while relaxing the rigidity of the proteins. This relaxation permits local rearrangements of clashing side chains, or even larger structural rearrangements. Although the template-free docking, the so-called *ab initio* docking, can yield good models of protein complexes, it is usually computationally costly since it requires extensive sampling and can often become trapped in local minima. The use of prior experimental and computational knowledge can dramatically reduce the search space and thereby improve the accuracy of both these approaches (Li and Huang, 2021). Methods that predict the binding interfaces of individual proteins can be of great use in this respect (Table S1). These methods are trained on sequence- or structure-based interface properties, such as conservation, hydrophobicity, electrostatic potential, and solvent accessibility, to distinguish binding sites from the rest of the protein surface (Keskin et al., 2016) (Figure 2M). Although knowledge-based docking has shown much success for proteome-wide predictions (Kundrotas et al., 2012), proteins that undergo conformational change upon interaction remain an important challenge. To tackle them, investigators have used an ensemble of the unbound and bound conformations of proteins to predict and analyze PPIs (Kuzu et al., 2013; Halakou et al., 2017).

Template-based prediction

This approach models PPIs using homologous complexes of known structure as templates (Figure 2K). It is based on the expectation that PPIs are conserved across evolutionary time and, indeed, PPI interfaces appear to be better conserved than the rest of the protein surface (Tsai et al., 1996). In a majority of cases, complexes with sequence identities of more than 40% maintain similar interaction modes (Aloy et al., 2003; Tsai et al., 1996). Efforts have been made to expand the approach to detect templates of unclear homology by showing that similar binding modes can occur in the absence of detectable sequence similarity (Mukherjee and Zhang, 2011; Sinha et al., 2010; Tuncbag et al., 2011; Zhang et al., 2012; Kundrotas et al., 2012; Guerler et al., 2013). For instance, template PPIs with low sequence identities may still yield acceptable models if the subunit interfaces of the template complex are well resolved (Rodrigues et al., 2013) and our current knowledge of protein-protein interfaces are sufficient to cover most PPIs (Kundrotas et al., 2012; Gao and Skolnick, 2010). Such methods based on local similarity can generate high-quality PPI models based on low sequence identities and are more popular for large-scale predictions (Muratcioglu et al., 2015). Template-based PPI modeling is a special case of homology modeling, in which the models for the constituent proteins are co-optimized in the complex (Szilagyi and Zhang, 2014). As such, it faces all of the challenges of homology modeling, such as the detection and correct alignment of distant homologs, and additionally, a dearth of methods that can efficiently refine the modeled full-length complexes. Nevertheless, the current exponential increase in experimentally solved structures of protein complexes suggests that most PPIs will either have a solved structure or a high-quality structural template in the foreseeable future (Kundrotas et al., 2012). Template-based approaches are fast enough to be scaled up to the proteome level, especially with efficient optimization of the refinement steps (Table S1). For instance, the PRISM software suite was used to generate more than 60,000 putative interactions among 6,170 target proteins based on structures in the PDB (Tuncbag et al., 2011; Baspinar et al., 2014).

Machine learning-based prediction

In 2020, we witnessed a revolution in the field of molecular biology when DeepMind presented AlphaFold version 2.0 (Jumper et al., 2021), a machine learning approach that incorporates homology, structural information from the evolutionary history, a sequence-structure relationship, and force-field refinement to raise the bar of protein structure prediction accuracy to a median

of 2.1 Å (AlQuraishi, 2021). This advance was made possible through decades of efforts in experimental structure biology and genome sequencing, which built the datasets needed to teach machine learning networks how to fold a protein artificially. Hundreds of thousands of proteins have been modeled using AlphaFold and made publicly available in the AlphaFold Protein Structure Database (https://alphafold.ebi.ac.uk/). Academic research groups followed suit and built another machine learning approach, RoseTTAFold (Baek et al., 2021), the performance of which is beginning to close the gap between the leading academic groups and DeepMind. The rise of such powerful methods has finally established computational modeling next to experimental approaches as a reliable source of structural information. In 2021, just months after the introduction of AlphaFold, these methods were applied to the prediction of PPIs. AlphaFold-Multimer showed increased accuracy in predicting multimeric interfaces while maintaining high single-chain accuracy for a large dataset of 4,433 protein complexes (Evans et al., 2021). A combination of RoseTTAFold and AlphaFold was used to screen 8.3 million pairs of *Saccharomyces cerevisiae* proteins and model approximately 712 known and *de novo* protein assemblies spanning almost all of the key processes in eukaryotic cells (Humphreys et al., 2021). ColabFold, offering both AlphaFold and RoseTTAFold to the non-expert user, used a fast multiple sequence alignment generation by MMseqs2 to predict homo- and hetero-complexes (Mirdita et al., 2021). One study showed an improved performance of PPI docking by combining AlphaFold and ClusPro (Ghani et al., 2021). Despite the unprecedented success of such approaches and the extent to which they have empowered research in various branches of the life sciences, they do occasionally produce erroneous models and therefore still need to be followed up by in-depth analysis and validation, until we establish reliable metrics for model quality.

Outlook

In this article, we have surveyed approaches taken to identify and characterize PPIs at a systems level. Each approach has its strengths and limitations, which are often complementary between methods. Thus, the computational analysis of genomic data, for example, for the co-localization of genes on the chromosome, their co-expression, or their co-occurrence in different genomes, provides a first and fast overview of potential interactions, which can be collected to obtain genome-wide representations of protein networks. At the experimental level, such overviews can be obtained through methods such as phage display, Y2H screening, and AP, which provide large sets of probable physical interactions, covering many proteins, but reflecting mainly high-affinity interactions and lacking any structural information on the complexes formed. These methods are comparatively fast, but prone to produce false positives at a rate that may exceed one-third of the identified interactions (Huang et al., 2007). Thus, their main use is to substantiate potential interactions and prioritize them for more detailed analysis with methods that have a lower throughput, but provide a much more reliable and detailed description of the interaction. Among these, co-crystallization and cryo-EM provide the highest resolution of the complex formed through the interaction, but they are also the most labor intensive and have uncertain success rates. Methods such as XL-MS and sequence coevolution are less labor intensive and have more plannable success rates, but they provide medium-resolution data on the distance of residues at or near the interface and must be coupled with knowledge of the individual protein structures to assemble complexes.

The diversity of the approaches taken by the different methods underpins their different outcomes over the same set of proteins. For example, applying four different methods to the proteome of *E. coli* substantially expanded the cumulative list of probable PPIs in that organism (Babu et al., 2012; Weisbrod et al., 2013; Rajagopala et al., 2014; Cong et al., 2019). This complementarity has prompted proposals to combine methods to obtain much better insight into the architecture of cellular complexes (Robinson et al., 2007) but has thus far been exploited mainly for the study of individual complexes, albeit sometimes of a very large, multisubunit nature (Alber et al., 2007; Rout and Andrej, 2019). We see that considerable further benefit is achievable by integrating the individual approaches with a view of obtaining higher-resolution representation of interactions over entire proteomes, and such integration will have to be achieved at the computational level. This is because (1) each experimental method requires dedicated infrastructure and highly trained staff, making it difficult to establish more than one method at a high level within one research group, (2) the diverse nature of the obtained data requires statistical approaches, such as Bayesian inference, for their integration, (3) the datasets are of considerable complexity, especially when gathered over large sets of proteins, so that extraction of complex correlations in the data exceeds the biological range of the human brain and requires machine learning approaches. Computational approaches are already essential for the interpretation of the raw data in all of the experimental methods described here. All of the methods are enhanced by simple reference to the computational interactomes predicted primarily from genomic data, such as in the STRING database, but these do not function in an integrative way. What we specifically refer to here is to the need for computational methods specifically aimed at integrating disparate data into a coherent, high-resolution view of whole protein networks.

We think that machine learning will be at the base of the new computational methods we propose here, since it has already demonstrated its power to extract complex correlations from large datasets. Most recently, it was used to model proteins with a native structure at a level of accuracy that is comparable with that of experimental methods (Jumper et al., 2021; Pereira et al., 2021). The machine-learning program AlphaFold version 2.0 (https://github.com/deepmind/alphafold), is being used to build a dedicated database of structure models for all of the proteins in the UniRef90 sequence database. The AlphaFold Protein Structure Database provides structural coordinates that can be used for the interpretation of interactome data in the same way in which experimental coordinates are already used now, but on a much larger scale. Recently, AlphaFold has been shown to also model many homo- and hetero-oligomers at atomic accuracy (Evans et al., 2021), suggesting that the efficient inference of interactomes will become possible by computational means, subject at least initially to verification by experimental methods. The extent of completeness of such inferred interactomes is difficult to evaluate at this point. The datasets on which we can train

networks may not be sufficient at present to allow the proteome-wide computational prediction of PPIs at an acceptable level of accuracy. This may be the case, particularly for interactions that are transient, low-affinity, or dependent on the conditional structuring of one partner, which may require substantial input and validation from experimental methods. We are therefore led to wonder whether computational approaches can take over the exploration of protein-protein interactions in the foreseeable future or whether experimental methods will continue to be needed for years to come.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.str.2022.02.004.

ACKNOWLEDGMENTS

The authors would like to thank the members of the Department of Protein Evolution at the Max Planck Institute for Developmental Biology and Applied Bioinformatics group at Eberhard Karls University of Tübingen for discussion and feedback. This work was supported by institutional funds from the Max Planck Society. H.E. is funded by the International Max Planck Research School (IMPRS) "From Molecules to Organisms" and Eberhard Karls University of Tübingen.

AUTHOR CONTRIBUTIONS

H.E., V.A., and F.M. developed the concept and outline for the article, with advice from A.N.L. and O.K. All of the authors wrote the paper and discussed and edited the final version.

SUPPORTING CITATIONS

The following references appear in the supplemental information: Ohue et al., 2014; Oughtred et al., 2021; Ovchinnikov et al., 2014; Patil et al., 2011; Peri et al., 2004; Porollo and Meller, 2007; Reille et al., 2018; Ripoche et al., 2017; Rolland et al., 2014; Jiménez-Garcis et al., 2018; Saha et al., 2018; Salwinski et al., 2004; Szklarczyk et al., 2021; Torchala et al., 2013; Tovchigrechko and Vakser, 2006; Tran et al., 2018; Vries et al., 2006; Yan et al., 2020; Zhang et al., 2011; Zhang et al., 2013; Chen et al., 2017; Christoffer et al., 2021; Croft et al., 2011; Giurgiu et al., 2019; Hermjakob et al., 2004; Hopf et al., 2014; Kamburov et al., 2009; Kamisetty et al., 2013; Kawabata, 2016; Kozakov et al., 2017; Licata et al., 2012; Lyskov and Gray, 2008; Meyer et al., 2013; Mosca et al., 2013; Obayashi et al., 2018; Obayashi et al., 2019; Ofran and Rost, 2007; Szklarczyk et al., 2021

REFERENCES

Aebersold, R., and Mann, M. (2003). Mass spectrometry-based proteomics. Nature *422*, 198–207.

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., et al. (2007). The molecular architecture of the nuclear pore complex. Nature *450*, 695–701.

Allegretti, M., Zimmerli, C.E., Rantos, V., Wilfling, F., Ronchi, P., Fung, H.K.H., Lee, C.W., Hagen, W., Turoňová, B., Karius, K., et al. (2020). In-cell architecture of the nuclear pore and snapshots of its turnover. Nature *586*, 796–800.

Aloy, P., Ceulemans, H., Stark, A., and Russell, R.B. (2003). The relationship between sequence and interaction divergence in proteins. J. Mol. Biol. *332*, 989–998.

AlQuraishi, M. (2021). Machine learning in protein structure prediction. Curr. Opin. Chem. Biol. *65*, 1–8.

Al-Shahrour, F., Minguez, P., Marqués-Bonet, T., Gazave, E., Navarro, A., and Dopazo, J. (2010). Selection upon genome architecture: conservation of functional neighborhoods with changing genes. PLoS Comput. Biol. *6*, e1000953.

Apic, G., Gough, J., and Teichmann, S.A. (2001). An insight into domain combinations. Bioinformatics *17*, S83–S89.

Babu, M., Kagan, O., Guo, H., Greenblatt, J., and Emili, A. (2012). Identification of protein complexes in Escherichia coli using sequential peptide affinity purification in combination with tandem mass spectrometry. J. Vis. Exp. *69*. https://doi.org/10.3791/4057.

Back, J.W., de Jong, L., Muijsers, A.O., and de Koster, C.G. (2003). Chemical cross-linking and mass spectrometry for protein structural modeling. J. Mol. Biol. *331*, 303–313.

Baek, M., Frank, D.M., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876.

Balakrishnan, S., Kamisetty, H., Jaime, G.C., Lee, S.I., and langmead, C.J. (2011). Learning generative models for protein fold families. Proteins *79*, 1061–1078.

Barker, D., and Pagel, M. (2005). Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput. Biol. *1*, e3.

Barker, D., Meade, A., and Pagel, M. (2007). Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. Bioinformatics *23*, 14–20.

Baspinar, A., Cukuroglu, E., Nussinov, R., Keskin, O., and Gursoy, A. (2014). PRISM: a web server and repository for prediction of protein-protein interactions and modeling their 3D complexes. Nucleic Acids Res. *42*, W285–W289.

Blaschke, C., Hoffmann, R., Oliveros, J.C., and Valencia, A. (2001). Extracting information automatically from biological literature. Comp. Funct. Genomics *2*, 310–313.

Boder, E.T., and Wittrup, K.D. (1997). Yeast surface display for screening combinatorial polypeptide libraries. Nat. Biotechnol. *15*, 553–557.

Bordini, M., Zappaterra, M., Soglia, F., Petracci, M., and Davoli, R. (2021). Weighted gene Co-expression network analysis identifies molecular pathways and hub genes involved in broiler white striping and wooden breast myopathies. Scientific Rep. *11*, 1776.

Branon, T.C., Bosch, J.A., Sanchez, A.D., Udeshi, N.D., Svinkina, T., Carr, S.A., Feldman, J.L., Perrimon, N., and Ting, A.Y. (2018). Efficient proximity labeling in living cells and organisms with TurboID. Nat. Biotechnol. *36*, 880–887.

Braun, P., and Gingras, A.C. (2012). History of protein-protein interactions: from egg-white to complex networks. Proteomics *12*, 1478–1498.

Brilli, M., Mengoni, A., Fondi, M., Bazzicalupo, M., Liò, P., and Fani, R. (2008). Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network. BMC Bioinformatics *9*, 551.

Caberoy, N.B., Zhou, Y., Jiang, X., Alvarado, G., and Li, W. (2010). Efficient identification of tubby-binding proteins by an improved system of T7 phage display. J Mol Recognit *23*, 74–83.

Chang, F.N., and Flaks, J.G. (1972). The specific cross-linking of two proteins from the Escherichia coli 30 S ribosomal subunit. J. Mol. Biol. *68*, 177–180.

Chang, T.W. (1983). Binding of cells to matrixes of distinct antibodies coated on solid surface. J. Immunological Methods *65*, 217–223.

Chen, J.Y., Pandey, R., and Nguyen, T.M. (2017). HAPPI-2: a comprehensive and high-quality map of human annotated and predicted protein interactions. BMC Genomics *18*, 182.

Chia, J.M., and Kolatkar, P.R. (2004). Implications for domain fusion protein-protein interactions based on structural information. BMC Bioinformatics *5*, 161.

Christoffer, C., Chen, S., Bharadwaj, V., Aderinwale, T., Kumar, V., Hormati, M., and Kihara, D. (2021). LZerD webserver for pairwise and multiple protein-protein docking. Nucleic Acids Res. *49*, W359–W365.

Chuang, C.L., Jen, C.H., Chen, C.M., and Shieh, G.S. (2008). A pattern recognition approach to infer time-lagged genetic interactions. Bioinformatics *24*, 1183–1190.

Clegg, C., and Hayes, D. (1974). Identification of neighbouring proteins in the ribosomes of Escherichia coli. A topographical study with the cross-linking reagent dimethyl suberimidate. Eur J Biochem *42*, 21–28.

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

Crick, F.H. (1988). What Mad Pursuit: A Personal View of Scientific Discovery (Basic Books).

Croft, D., O'Kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2011). Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res. *39*, D691–D697.

Csurös, M. (2010). Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics *26*, 1910–1912.

Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. *23*, 324–328.

Dávila López, M., Martínez Guerra, J.J., and Samuelsson, T. (2010). Analysis of gene order conservation in eukaryotes identifies transcriptionally and functionally linked genes. PLoS One *5*, e10654.

Deutsch, E.W., Orchard, S., Binz, P.A., Bittremieux, W., Eisenacher, M., Hermjakob, H., Kawano, S., Lam, H., Mayer, G., Menschaert, G., et al. (2017). Proteomics standards initiative: fifteen years of progress and future work. J. Proteome Res. *16*, 4288–4298.

Dunham, W.H., Mullin, M., and Gingras, A.C. (2012). Affinity-purification coupled to mass spectrometry: basic principles and strategies. Proteomics *12*, 1576–1590.

Ehrlich, P. (1891). Experimentelle Untersuchungen über Immunität. II. Ueber Abrin. Deutsche Medizinische Wochenschrift *17*, 1218–1219.

Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. Phys Rev E Stat Nonlin Soft Matter Phys *87*, 012707.

Evans, R., O'Neill, M., Alexander, P., Antropova, N., Senior, A., Green, T., Augustin, Ž., Russ, B., Sam, B., Jason, Y., et al. (2021). Protein Complex Prediction with AlphaFold-Multimer, Preprint at. bioRxiv. https://doi.org/10.1101/2021.10.04.463034.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340*, 245–246.

Fodor, A.A., and Aldrich, R.W. (2004). On evolutionary conservation of thermodynamic coupling in proteins. J. Biol. Chem. *279*, 19046.

Francisco, J.A., Campbell, R., Iverson, B.L., and Georgiou, G. (1993). Production and fluorescence-activated cell sorting of Escherichia coli expressing a functional antibody fragment on the external surface. Proc. Natl. Acad. Sci. U S A *90*, 10444–10448.

Fraser, H.B., Hirsh, A.E., Wall, D.P., and Eisen, M.B. (2004). Coevolution of gene expression among interacting proteins. Proc. Natl. Acad. Sci. U S A *101*, 9033–9038.

Gao, M., and Skolnick, J. (2010). Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. Proc. Natl. Acad. Sci. U S A *107*, 22517–22522.

Ghani, U., Desta, I., Jindal, A., Khan, O., Jones, G., Kotelnikov, S., Padhorny, D., Vajda, S., and Kozakov, D. (2021). Improved Docking of Protein Models by a Combination of Alphafold2 and ClusPro, Preprint at. bioRxiv. https://doi.org/10.1101/2021.09.07.459290.

Giulietti, M., Occhipinti, G., Principato, G., and Piva, F. (2017). Identification of candidate miRNA biomarkers for pancreatic ductal adenocarcinoma by weighted gene Co-expression network analysis. Cell Oncol (Dordr) *40*, 181–192.

Giurgiu, M., Reinhard, J., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Ruepp, A. (2019). CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. *47*, D559–D563.

Glazko, G.V., and Mushegian, A.R. (2004). Detection of evolutionarily stable fragments of cellular pathways by hierarchical clustering of phyletic patterns. Genome Biol. *5*, 1–13.

Gloor, G.B., Martin, L.C., Wahl, L.M., and Dunn, S.D. (2005). Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry *44*, 7156–7165.

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

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

Greer, J., and Bush, B.L. (1978). Macromolecular shape and surface maps by solvent exclusion. Proc. Natl. Acad. Sci. U S A *75*, 303–307.

Guerler, A., Govindarajoo, B., and Zhang, Y. (2013). Mapping monomeric threading to protein-protein structure prediction. J. Chem. Inf. Model. *53*, 717–725.

Guo, Q., Lehmer, C., Martínez-Sánchez, A., Rudack, T., Beck, F., Hartmann, H., Pérez-Berlanga, M., Frottin, F., Hipp, M.S., Hartl, F.U., et al. (2018). In situ structure of neuronal C9orf72 poly-GA aggregates reveals proteasome recruitment. Cell *172*, 696–705.e12.

Halakou, F., Kilic, E.S., Cukuroglu, E., Keskin, O., and Gursoy, A. (2017). Enriching traditional protein-protein interaction networks with alternative conformations of proteins. Sci Rep *7*, 7180.

Hanes, J., and Plückthun, A. (1997). In vitro selection and evolution of functional proteins by using ribosome display. Proc. Natl. Acad. Sci. U S A *94*, 4937–4942.

He, M., and Taussig, M.J. (2001). Single step generation of protein arrays from DNA by cell-free expression and in situ immobilisation (PISA method). Nucleic Acids Res. *29*, E73-3.

He, M., Stoevesandt, O., Palmer, E.A., Khan, F., Ericsson, O., and Taussig, M.J. (2008). Printing protein arrays from DNA arrays. Nat. Methods *5*, 175–177.

He, M., Stoevesandt, O., and Taussig, M.J. (2008). In situ synthesis of protein arrays. Curr. Opin. Biotechnol. *19*, 4–9.

He, Y., Wang, Z., Ge, H., Liu, Y., and Chen, H. (2021). Weighted gene Co-expression network analysis identifies genes related to anthocyanin biosynthesis and functional verification of hub gene SmWRKY44. Plant Sci. Int. J. Exp. Plant Biol. *309*, 110935.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., et al. (2004). IntAct: an open source molecular interaction database. Nucleic Acids Res. *32*, D452–D455.

Himes, B.A., and Zhang, P. (2018). emClarity: software for high-resolution cryo-electron tomography and subtomogram averaging. Nat. Methods *15*, 955–961.

Ho, C.M., Li, X., Lai, M., Terwilliger, T.C., Beck, J.R., Wohlschlegel, J., Goldberg, D.E., Fitzpatrick, A.W.P., and Zhou, Z.H. (2020). Bottom-up structural proteomics: cryoEM of protein complexes enriched from the cellular milieu. Nat. Methods *17*, 79–85.

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

Hopp, T.P., Prickett, K.S., Price, V.L., Libby, R.T., March, C.J., Cerretti, D.P., Urdal, D.L., and Conlon, P.J. (1988). A short polypeptide marker sequence useful for recombinant protein identification and purification. Nat. Biotechnol. *6*, 1204–1210. https://doi.org/10.1038/nbt1088-1204.

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

Huang, S.Y. (2014). Search strategies and evaluation in protein-protein docking: principles, advances and challenges. Drug Discov. Today *19*, 1081–1096.

Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T.J., Banjade, S., Bagde, S.R., Stancheva, V.G., et al. (2021). Structures of core eukaryotic protein complexes, Preprint at. bioRxiv. https://doi.org/10.1101/2021.09.30.462231.

Huynen, M., Snel, B., Lathe, W., 3rd, and Bork, P. (2000). Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. Genome Res. *10*, 1204–1210.

Jones, D.T., Buchan, D.W., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics *28*, 184–190.

Jothi, R., Przytycka, T.M., and Aravind, L. (2007). Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. BMC Bioinformatics *8*, 173.

Jiménez-García, B., Jorge, R.T., Romero-Durana, M., Vidal, M., Jiménez-González, D., and Fernández-Recio, J. (2018). LightDock: a new multi-scale Approach to protein-protein docking. Bioinformatics *34*, 49–55.

Jumper, J., Evans, R., Alexander, P., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature. https://doi.org/10.1038/s41586-021-03819-2.

Kamburov, A., Wierling, C., Lehrach, H., and Herwig, R. (2009). ConsensusPathDB–a database for integrating human functional interaction networks. Nucleic Acids Res. *37*, D623–D628.

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). 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.

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

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A.A., Aflalo, C., and Vakser, I.A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc. Natl. Acad. Sci. U S A *89*, 2195–2199.

Kawabata, T. (2016). HOMCOS: an updated server to search and model complex 3D structures. J. Struct. Funct. Genomics *17*, 83–99.

Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders, R., Brazma, A., Holstege, F.C., and Holstege, P. (2002). Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol. Cell *9*, 1133–1143.

Kensche, P.R., van Noort, V., Dutilh, B.E., and Huynen, M.A. (2008). Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface *5*, 151–170.

Keskin, O., Tuncbag, N., and Gursoy, A. (2016). Predicting protein-protein interactions from the molecular to the proteome level. Chem. Rev. *116*, 4884–4909.

Kim, M.D., Han, K.C., Kang, H.A., Rhee, S.K., and Seo, J.H. (2003). Coexpression of BiP increased antithrombotic hirudin production in recombinant Saccharomyces cerevisiae. J. Biotechnol. *101*, 81–87.

Kim, D.I., Birendra, K.C., Zhu, W., Motamedchaboki, K., Doye, V., and Roux, K.J. (2014). Probing nuclear pore complex architecture with proximity-dependent biotinylation. Proc. Natl. Acad. Sci. U S A *111*, E2453–E2461.

Kim, D.I., Jensen, S.C., Noble, K.A., Kc, B., Roux, K.H., Motamedchaboki, K., and Roux, K.J. (2016). An improved smaller biotin ligase for BioID proximity labeling. Mol. Biol. Cell *27*, 1188–1196.

Kozakov, D., Hall, D.R., Xia, B., Porter, K.A., Padhorny, D., Yueh, C., and Beglov, D. (2017). The ClusPro web server for protein-protein docking. Nat. Protoc. *12*, 255–278.

Kühlbrandt, W. (2014). Biochemistry. The resolution revolution. Science *343*, 1443–1444.

Kundrotas, P.J., Zhu, Z., Janin, J., and Vakser, I.A. (2012). Templates are available to model nearly all complexes of structurally characterized proteins. Proc. Natl. Acad. Sci. U S A *109*, 9438–9441.

Kuzu, G., Gursoy, A., Nussinov, R., and Keskin, O. (2013). Exploiting conformational ensembles in modeling protein-protein interactions on the proteome scale. J. Proteome Res. *12*, 2641–2653.

Lam, S.S., Martell, J.D., Kamer, K.J., Deerinck, T.J., Ellisman, M.H., Mootha, V.K., and Ting, A.Y. (2015). Directed evolution of APEX2 for electron microscopy and proximity labeling. Nat. Methods *12*, 51–54.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics *9*, 559.

Lausted, C., Hu, Z., and Hood, L. (2008). Quantitative serum proteomics from surface plasmon resonance imaging. Mol Cell Proteomics *7*, 2464–2474.

Lee, J.M., and Sonnhammer, E.L. (2003). Genomic gene clustering analysis of pathways in eukaryotes. Genome Res. *13*, 875–882.

Lee, Y., Lee, E.K., Cho, Y.W., Matsui, T., Kang, I.C., Kim, T.S., and Han, M.H. (2003). ProteoChip: a highly sensitive protein microarray prepared by a novel method of protein immobilization for application of protein-protein interaction studies. Proteomics *3*, 2289–2304.

Leitner, A. (2016). Cross-linking and other structural proteomics techniques: how chemistry is enabling mass spectrometry applications in structural biology. Chem. Sci. *7*, 4792–4803.

Li, H., and Huang, S.Y. (2021). Protein–protein docking with interface residue restraints. Chin. Phys. B *30*, 018703.

Li, B., Pu, K., and Wu, X. (2019). Identifying novel biomarkers in hepatocellular carcinoma by weighted gene Co-expression network analysis. J. Cell Biochem. https://doi.org/10.1002/jcb.28420.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., Castagnoli, L., and Cesareni, G. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. *40*, D857–D861.

Liu, F., Rijkers, D.T., Post, H., and Heck, A.J. (2015). Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. Nat. Methods *12*, 1179–1184.

Liu, F., Lössl, P., Rabbitts, B.M., Balaban, R.S., and Heck, A.J.R. (2018). The interactome of intact mitochondria by cross-linking mass spectrometry provides evidence for coexisting respiratory supercomplexes. Mol Cell Proteomics *17*, 216–232.

Liu, B., Zhan, Y., Chen, X., Hu, X., Wu, B., and Pan, S. (2021). Weighted gene Co-expression network analysis can sort cancer-associated fibroblast-specific markers promoting bladder cancer progression. J. Cell Physiol. *236*, 1321–1331.

Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science *286*, 295–299.

Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408.

Luo, Y., Batalao, A., Zhou, H., and Zhu, L. (1997). Mammalian two-hybrid system: a complementary approach to the yeast two-hybrid system. BioTechniques *22*, 350–352.

Lyskov, S., and Gray, J.J. (2008). The RosettaDock server for local protein-protein docking. Nucleic Acids Res. *36*, W233–W238.

MacBeath, G., and Schreiber, S.L. (2000). Printing proteins as microarrays for high-throughput function determination. Science *289*, 1760–1763.

Mahamid, J., Pfeffer, S., Schaffer, M., Villa, E., Danev, R., Cuellar, L.K., Förster, F., Hyman, A.A., Plitzko, J.M., and Baumeister, W. (2016). Visualizing the molecular sociology at the HeLa cell nuclear periphery. Science *351*, 969–972.

Mao, W., Kaya, C., Dutta, A., Horovitz, A., and Bahar, I. (2015). Comparative study of the effectiveness and limitations of current methods for detecting sequence coevolution. Bioinformatics *31*, 1929–1937.

Marsh, J.A., Hernández, H., Hall, Z., Ahnert, S.E., Perica, T., Robinson, C.V., and Teichmann, S.A. (2013). Protein complexes are under evolutionary selection to assemble via ordered pathways. Cell *153*, 461–470.

Martell, J.D., Deerinck, T.J., Sancak, Y., Poulos, P.T.L., Mootha, V.K., Sosinsky, G.E., Ellisman, M.H., Mootha, V.K., Sosinsky, G.E., Ellisman, M.H., et al.

(2012). Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. Nat. Biotechnol. *30*, 1143–1149.

Meyer, M.J., Das, J., Wang, X., and Yu, H. (2013). INstruct: a database of high-quality 3D structurally resolved protein interactome networks. Bioinformatics *29*, 1577–1579.

Michaud, G.A., Salcius, M., Zhou, F., Bangham, R., Bonin, J., Guo, H., Snyder, M., Predki, P.F., and Schweitzer, B.I. (2003). Analyzing antibody specificity with whole proteome microarrays. Nat. Biotechnol. *21*, 1509–1512.

Mintseris, J., and Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. Proc. Natl. Acad. Sci. U S A *102*, 10930–10935.

Mirdita, M., Ovchinnikov, S., and Martin, S. (2021). ColabFold - making protein folding accessible to all, Preprint at. bioRxiv. https://doi.org/10.1101/2021.08.15.456425.

Mon, H., Sugahara, R., Hong, S.M., Lee, J.M., Kamachi, Y., Kawaguchi, Y., and Kusakabe, T. (2009). Analysis of protein interactions with two-hybrid system in cultured insect cells. Anal. Biochem. *392*, 180–182.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Debora, S.M., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. U S A *108*, E1293–E1301.

Mosca, R., Céol, A., and Aloy, P. (2013). Interactome3D: adding structural details to protein networks. Nat. Methods *10*, 47–53.

Mouradov, D., King, G., Ross, I.L., Forwood, J.K., Hume, D.A., Sinz, A., Martin, J.L., Kobe, B., and Huber, T. (2008). Protein structure determination using a combination of cross-linking, mass spectrometry, and molecular modeling. Methods Mol. Biol. *426*, 459–474.

Mukherjee, S., and Zhang, Y. (2011). Protein-protein complex structure predictions by multimeric threading and template recombination. Structure *19*, 955–966.

Muratcioglu, S., Guven-Maiorov, E., Keskin, Ö., and Gursoy, A. (2015). Advances in template-based protein docking by utilizing interfaces towards completing structural interactome. Curr. Opin. Struct. Biol. *35*, 87–92.

Mushegian, A.R., and Koonin, E.V. (1996). Gene order is not conserved in bacterial evolution. Trends Genet *12*, 289–290.

Nagano, K., and Tsutsumi, Y. (2021). Phage display technology as a powerful platform for antibody drug discovery. Viruses *13*. https://doi.org/10.3390/v13020178.

Nakajima, N., Akutsu, T., and Nakato, R. (2021). Databases for protein-protein interactions. Methods Mol. Biol. *2361*, 229–248.

Nemoto, N., Miyamoto-Sato, E., Husimi, Y., and Yanagawa, H. (1997). In vitro virus: bonding of mRNA bearing puromycin at the 3'-terminal end to the C-terminal end of its encoded protein on the ribosome in vitro. FEBS Lett. *414*, 405–408.

Northrop, J.H. (1922). The inactivation OF trypsin : II. The equilibrium between trypsin and the inhibiting substance formed BY its action ON proteins. J. Gen. Physiol. *4*, 245–260.

Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y., and Kinoshita, K. (2018). ATTED-II in 2018: a plant coexpression database based on investigation of the statistical property of the mutual rank index. Plant Cell Physiol. *59*, 440.

Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S., and Kinoshita, K. (2019). COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. Nucleic Acids Res. *47*, D55–D62.

Ofran, Y., and Rost, B. (2007). ISIS: interaction sites identified from sequence. Bioinformatics *23*, e13–e16.

Ohue, M., Matsuzaki, Y., Uchikoga, N., Ishida, T., and Akiyama, Y. (2014). MEGADOCK: an all-to-all protein-protein interaction prediction system using tertiary structure data. Protein Pept. Lett. *21*, 766–778.

Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Dolma, S., Coulombe-Huntington, J., et al. (2021). The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. A Publ. Protein Soc. *30*, 187–200.

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

Ozawa, T., Kaihara, A., Sato, M., Tachihara, K., and Umezawa, Y. (2001). Split luciferase as an optical probe for detecting protein-protein interactions in mammalian cells based on protein splicing. Anal. Chem. *73*, 2516–2521.

Pagadala, N.S., Syed, K., and Tuszynski, J. (2017). Software for molecular docking: a review. Biophys Rev *9*, 91–102.

Papanikolaou, N., Pavlopoulos, G.A., Theodosiou, T., and Iliopoulos, I. (2015). Protein-protein interaction predictions using text mining methods. Methods *74*, 47–53.

Patil, A., Nakai, K., and Nakamura, H. (2011). HitPredict: a database of quality assessed protein-protein interactions in nine species. Nucleic Acids Res. *39*, D744–D749.

Pazos, F., Helmer-Citterich, M., Ausiello, G., and Valencia, A. (1997). Correlated mutations contain information about protein-protein interaction. J. Mol. Biol. *271*, 511–523.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U S A *96*, 4285–4288.

Pereira, J., Simpkin, A.J., Hartman, M.D., Rigden, D.J., Keegan, R.M., and Lupas, A.N. (2021). High-accuracy protein structure prediction in CASP14. Proteins. https://doi.org/10.1002/prot.26171.

Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., Surendranath, V., Muthusamy, B., Gandhi, T.K., Chandrika, K.N., Deshpande, N., Suresh, S., et al. (2004). Human protein reference database as a discovery resource for proteomics. Nucleic Acids Res. *32*, D497–D501.

Porollo, A., and Meller, J. (2007). Prediction-based fingerprints of protein-protein interactions. Proteins *66*, 630–645.

Przulj, N., Wigle, D.A., and Jurisica, I. (2004). Functional topology in a network of protein interactions. Bioinformatics *20*, 340–348.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M., and Séraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods (San Diego, Calif.) 24*, 218–229. https://doi.org/10.1006/meth.2001.1183.

Rajagopala, S.V., Sikorski, P., Kumar, A., Mosca, R., Vlasblom, J., Arnold, R., Franca-Koh, J., Pakala, S.B., Phanse, S., Ceol, A., et al. (2014). The Binary Protein-Protein Interaction Landscape of Escherichia Coli. Nat. Biotechnol. *32*, 285–290.

Ramachandran, N., Hainsworth, E., Bhullar, B., Eisenstein, S., Rosen, B., Lau, A.Y., Walter, J.C., and LaBaer, J. (2004). Self-assembling protein microarrays. Science *305*, 86–90.

Ramanathan, M., Majzoub, K., Rao, D.S., Neela, P.H., Zarnegar, B.J., Mondal, S., Roth, J.G., Gai, H., Kovalski, J.R., Siprashvili, Z., et al. (2018). RNA-protein interaction detection in living cells. Nat. Methods *15*, 207–212.

Rappsilber, J. (2011). The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. J. Struct. Biol. *173*, 530–540.

Ray, S., Mehta, G., and Srivastava, S. (2010). Label-free detection techniques for protein microarrays: prospects, merits and challenges. Proteomics *10*, 731–748.

Reille, S., Garnier, M., Robert, X., Gouet, P., Martin, J., and Launay, G. (2018). Identification and visualization of protein binding regions with the ArDock server. Nucleic Acids Res. *46*, W417–W422.

Rhee, H.W., Zou, P., Udeshi, N.D., Martell, J.D., Mootha, V.K., Carr, S.A., and Ting, A.Y. (2013). Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. Science *339*, 1328–1331.

Rinner, O., Seebacher, J., Walzthoeni, T., Mueller, L.N., Beck, M., Schmidt, A., Mueller, M., and Aebersold, R. (2008). Identification of cross-linked peptides from large sequence databases. Nat. Methods *5*, 315–318.

Ripoche, H., Laine, E., Ceres, N., and Carbone, A. (2017). JET2 viewer: a database of predicted multiple, possibly overlapping, protein-protein interaction sites for PDB structures. Nucleic Acids Res. *45*, D236–D242.

Roberts, R.W., and Szostak, J.W. (1997). RNA-peptide fusions for the in vitro selection of peptides and proteins. Proc Natl Acad Sci U S A *94*, 12297–12302.

Robinson, C.V., Sali, A., and Baumeister, W. (2007). The molecular sociology of the cell. Nature *450*, 973–982.

Rodrigues, J.P., Melquiond, A.S., Karaca, E., Trellet, M., van Dijk, M., van Zundert, G.C., Schmitz, C., de Vries, S.J., Bordogna, A., Bonati, L., Kastritis, P.L., and Bonvin, A.M. (2013). Defining the limits of homology modeling in information-driven protein docking. Proteins *81*, 2119–2128.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212–1226.

Rout, M.P., and Sali, A. (2019). Principles for integrative structural biology studies. Cell *177*, 1384–1403.

Roux, K.J., Kim, D.I., Raida, M., and Burke, B. (2012). A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. J. Cell Biol. *196*, 801–810.

Saha, C.K., Sanches Pires, R., Brolin, H., Delannoy, M., and Catherine Atkinson, G. (2018). Predicting functional associations using flanking genes (FlaGs), Preprint at. bioRxiv. https://doi.org/10.1101/362095.

Sala, D., Cerofolini, L., Fragai, M., Giachetti, A., Luchinat, C., and Rosato, A. (2020). A protocol to automatically calculate homo-oligomeric protein structures through the integration of evolutionary constraints and NMR ambiguous contacts. Comput. Struct. Biotechnol. J. *18*, 114–124.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The database of interacting proteins: 2004 update. Nucleic Acids Res. *32*, D449–D451.

Schäfer, J., and Strimmer, K. (2005). An empirical bayes approach to inferring large-scale gene association networks. Bioinformatics *21*, 754–764.

Schweppe, D.K., Chavez, J.D., Lee, C.F., Caudal, A., Kruse, S.E., Stuppard, R., Marcinek, D.J., Shadel, G.S., Tian, R., and Bruce, J.E. (2017). Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. Proc Natl Acad Sci U S A *114*, 1732–1737.

Shoemaker, B.A., and Panchenko, A.R. (2007). Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Comput. Biol. *3*, e42.

Sidhu, S.S., Fairbrother, W.J., and Deshayes, K. (2003). Exploring protein-protein interactions with phage display. Chembiochem *4*, 14–25.

Singh, P., Panchaud, A., and Goodlett, D.R. (2010a). Chemical cross-linking and mass spectrometry as a low-resolution protein structure determination technique. Anal. Chem. *82*, 2636–2642.

Singh, R., Park, D., Xu, J., Hosur, R., and Berger, B. (2010b). Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. Nucleic Acids Res. *38*, W508–W515.

Sinha, R., Kundrotas, P.J., and Vakser, I.A. (2010). Docking by structural similarity at protein-protein interfaces. Proteins *78*, 3235–3241.

Sinz, A. (2006). Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein-protein interactions. Mass Spectrom. Rev. *25*, 663–682.

Sinz, A. (2007). Investigation of protein-ligand interactions by mass spectrometry. ChemMedChem *2*, 425–431.

Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science *228*, 1315–1317.

Steffen, M., Petti, A., Aach, J., D'haeseleer, P., and Church, G. (2002). Automated modelling of signal transduction networks. BMC Bioinformatics *3*, 34.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science *302*, 249–255.

Su, C.C., Lyu, M., Morgan, C.E., Bolla, Bolla, J.R., Robinson, C.V., and Yu, E.W. (2021). A 'Build and Retrieve' methodology to simultaneously solve cryo-EM structures of membrane proteins. Nat. Methods *18*, 69–75.

Sun, T.T., Bollen, A., Kahan, L., and Traut, R.R. (1974). Topography of ribosomal proteins of the Escherichia coli 30S subunit as studied with the reversible cross-linking reagent methyl 4-mercaptobutyrimidate. Biochemistry *13*, 2334–2340.

Suter, B., Zhang, X., Pesce, C.G., Mendelsohn, A.R., Dinesh-Kumar, S.P., and Mao, J.H. (2015). Next-generation sequencing for binary protein-protein interactions. Front Genet *6*, 346. https://doi.org/10.3389/fgene.2015.00346.

Szilagyi, A., and Zhang, Y. (2014). Template-based structure modeling of protein-protein interactions. Curr. Opin. Struct. Biol. *24*, 10–23.

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

Tang, X., and Bruce, J.E. (2009). Chemical cross-linking for protein-protein interaction studies. Methods Mol. Biol. *492*, 283–293.

Tegunov, D., Xue, L., Dienemann, C., Cramer, P., and Mahamid, J. (2021). Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 Å in cells. Nat. Methods *18*, 186–193.

Thaminy, S., Miller, J., and Stagljar, I. (2004). The split-ubiquitin membrane-based yeast two-hybrid system. Methods Mol. Biol. *261*, 297–312.

Torchala, M., Moal, I.H., Chaleil, R.A., Fernandez-Recio, J., and Bates, P.A. (2013). SwarmDock: a server for flexible protein-protein docking. Bioinformatics *29*, 807–809.

Tovchigrechko, A., and Vakser, I.A. (2006). GRAMM-X public web server for protein-protein docking. Nucleic Acids Res. *34*, W310–W314.

Trakselis, M.A., Alley, S.C., and Ishmael, F.T. (2005). Identification and mapping of protein-protein interactions by a combination of cross-linking, cleavage, and proteomics. Bioconjug. Chem. *16*, 741–750.

Tran, N.V., Greshake Tzovaras, B., and Ebersberger, I. (2018). PhyloProfile: dynamic visualization and exploration of multi-layered phylogenetic profiles. Bioinformatics *34*, 3041–3043.

Tremblay, B.J.-M., Lobb, B., and Andrew, C.D. (2021). PhyloCorrelate: inferring bacterial gene-gene functional associations through large-scale phylogenetic profiling. Bioinformatics. https://doi.org/10.1093/bioinformatics/btaa1105.

Tsai, C.J., Lin, S.L., Wolfson, H.J., and Nussinov, R. (1996). A dataset of protein-protein interfaces generated with a sequence-order-independent comparison technique. J. Mol. Biol. *260*, 604–620.

Tuncbag, N., Gursoy, A., Nussinov, R., and Keskin, O. (2011). Predicting protein-protein interactions on a proteome scale by matching evolutionary and structural similarities at interfaces using PRISM. Nat. Protoc. *6*, 1341–1354.

Vakser, I.A. (2014). Protein-protein docking: from interaction to interactome. Biophys J *107*, 1785–1793.

Van Criekinge, W., and Beyaert, R. (1999). Yeast two-hybrid: state of the art. Biol. Procedures Online *2*, 1–38.

Vert, J.P. (2002). A tree kernel to analyse phylogenetic profiles. Bioinformatics *18 Suppl 1*, S276–S284.

de Vries, S.J., van Dijk, A.D., and Bonvin, A.M. (2006). WHISCY: what information does surface conservation yield? Application to data-driven docking. Proteins *63*, 479–489.

Wang, Z., Grange, M., Pospich, S., Wagner, T., Kho, A.L., Gautel, M., and Raunser, S. (2021). Structure of the thin filament in native skeletal muscles reveals its interaction with nebulin and two distinct conformations of myosin, Preprint at. bioRxiv. https://doi.org/10.1101/2021.10.06.463400.

Wassaf, D., Kuang, G., Kopacz, K., Wu, Q.L., Nguyen, Q., Toews, M., Cosic, J., Jacques, J., Wiltshire, S., Lambert, J., et al. (2006). High-throughput affinity ranking of antibodies using surface plasmon resonance microarrays. Anal. Biochem. *351*, 241–253.

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

Weisbrod, C.R., Chavez, J.D., Eng, J.K., Yang, L., Zheng, C., and Bruce, J.E. (2013). In vivo protein interaction network identified with a novel real-time cross-linked peptide identification strategy. J. Proteome Res. *12*, 1569–1579.

Willats, W.G. (2002). Phage display: practicalities and prospects. Plant Mol. Biol. *50*, 837–854.

Wodak, S.J., and Janin, J. (1978). Computer analysis of protein-protein interaction. J. Mol. Biol. *124*, 323–342.

Yan, Y., and Huang, S.Y. (2021). Accurate prediction of inter-protein residue-residue contacts for homo-oligomeric protein complexes. Brief. Bioinformatics. https://doi.org/10.1093/bib/bbab038.

Yan, Y., Tao, H., He, J., and Huang, S.Y. (2020). The HDOCK server for integrated protein-protein docking. Nat. Protoc. *15*, 1829–1852.

Yazaki, J., Galli, M., Kim, A.Y., Nito, K., Aleman, F., Chang, K.N., Carvunis, A.R., Quan, R., Nguyen, H., Song, L., et al. (2016). Mapping transcription factor interactome networks using HaloTag protein arrays. Proc Natl Acad Sci U S A *113*, E4238–E4247.

Zhang, B., and Horvath, S. (2005). A general framework for weighted gene Co-expression network analysis. Stat. Appl. Genet Mol Biol. *4*, Article17.

Zhang, Q.C., Deng, L., Fisher, M., Guan, J., Honig, B., and Petrey, D. (2011). PredUs: a web server for predicting protein interfaces using structural neighbors. Nucleic Acids Res. *39*, W283–W287.

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., et al. (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature *490*, 556–560.

Zhang, Q.C., Petrey, D., Garzón, J.I., Deng, L., and Honig, B. (2013). PrePPI: a structure-informed database of protein-protein interactions. Nucleic Acids Res. *41*, D828–D833.

Zhao, W., Langfelder, P., Fuller, T., Dong, J., Li, A., and Hovarth, S. (2010). Weighted gene coexpression network analysis: state of the art. J. Biopharm. Stat. *20*, 281–300.

Zheng, P.-F., Chen, L.-Z., Guan, Y.-Z., and Liu, P. (2021). Weighted gene Co-expression network analysis identifies specific modules and hub genes related to coronary artery disease. Sci. Rep. *11*, 6711.

Zhou, X., Kao, M.C., and Wong, W.H. (2002). Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci U S A *99*, 12783–12788.

Zhou, M., Li, Q., and Wang, R. (2016). Current experimental methods for characterizing protein-protein interactions. ChemMedChem *11*, 738–756.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., et al. (2001). Global analysis of protein activities using proteome chips. Science *293*, 2101–2105.

Supplementary Material

Exploring Protein-Protein Interactions at the Proteome Level

Hadeer Elhabashy [a,b,c,*], Felipe Merino [a], Vikram Alva [a], Oliver Kohlbacher [b,c,d,e], Andrei N Lupas [a,*]

a Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany
b Department of Computer Science, University of Tübingen, WSI/ZBIT, Sand 14, 72076, Tübingen, Germany
c Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14,72076, Tübingen, Germany
d Quantitative Biology Center, University of Tübingen, Auf der Morgenstelle 8, 72076 Tübingen, Germany
e Institute for Translational Bioinformatics, University Hospital Tübingen, Sand 14, 72076 Tübingen, Germany

* Corresponding authors

Email addresses: hadeer.elhabashy@tuebingen.mpg.de (Hadeer Elhabashy), vikram.alva@tuebingen.mpg.de (Vikram Alva), felipe.merino@tuebingen.mpg.de (Felipe Merino), oliver.kohlbacher@uni-tuebingen.de (Oliver Kohlbacher), andrei.lupas@tuebingen.mpg.de (Andrei N Lupas)

Table 1: Computational tools for characterization of PPIs

Approach	Tools	Reference	Web Resource	Description
Genomic-based	**EFI-GNT**	(Gerlt 2017; Zallot et al. 2018)	https://efi.igb.illinois.edu/efi-gnt	Genome neighborhood-based prediction of functional associations
	FlaGs	(Saha et al. 2018)	http://webflags.se	Genome neighborhood-based prediction of functional association in microbial and viral genomes
	PhyloCorrelate	(Tremblay, Lobb, and Doxey 2021)	https://phylocorrelate.uwaterloo.ca	Pairwise phylogenetic profile comparison across bacterial genomes
	PhyloProfile	(Tran, Greshake Tzovaras, and Ebersberger 2018)	https://github.com/BIONF/PhyloProfile	Multi-layered phylogenetic profile comparison
Gene Co-expression	**ATTED**	(Obayashi et al. 2018)	https://atted.jp	Plant Coexpression Database
	COXPRESdb	(Obayashi et al. 2019)	https://coxpresdb.jp	Coexpression database for 11 animal species
	WGCNA	(Langfelder and Horvath 2008)	https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA	Weighted gene co-expression network analysis
Sequence coevolution	**EVcomplex**	(Hopf et al. 2014; Green et al. 2021)	https://evcouplings.org/complex	PPI prediction and structure elucidation
	GREMLIN	(Ovchinnikov, Kamisetty, and Baker 2014; Kamisetty, Ovchinnikov, and Baker 2013)	http://gremlin.bakerlab.org/cplx_submit.php	PPI prediction and structure elucidation
Docking	**ClusPro**	(Kozakov et al. 2017; Ghani et al. 2021)	https://cluspro.org	FFT-based docking protein-protein docking, recently enhanced with AlphaFold v2.0 combination
	DOT2	(Roberts et al. 2013)	https://www.sdsc.edu/CCMS/DOT	Protein-protein docking
	GRAMM-X	(Tovchigrechko and Vakser 2006)	http://vakser.compbio.ku.edu/resources/gramm/grammx	FFT-based docking protein-protein docking
	HADDOCK	(Dominguez et al. 2003; de Vries et al. 2010)	https://wenmr.science.uu.nl/haddock2.4	Knowledge driven protein-protein docking
	HDOCK	(Yan et al. 2020)	http://hdock.phys.hust.edu.cn	Hybrid template-based and *Ab initio* protein-protein docking
	InterEvDock3	(Quignot et al. 2021)	https://bioserv.rpbs.univ-paris-diderot.fr/services/InterEvDock3	Hybrid template and evolutionary-based protein-protein docking

	LightDock	(Jiménez-García et al. 2018)	https://lightdock.org	Knowledge driven protein-protein docking
	LZerD	(Christoffer et al. 2021)	https://lzerd.kiharalab.org	Sequences to complexes protein-protein docking
	MEGADOCK 4.0	(Ohue et al. 2014)	https://www.bi.cs.titech.ac.jp/megadock	FFT-based protein-protein docking
	RosettaDOCK	(Lyskov and Gray 2008)	https://rosie.graylab.jhu.edu/docking2	Local protein-protein docking
	SwarmDOCK	(Torchala et al. 2013)	https://bmm.crick.ac.uk/~svc-bmm-swarmdock	Flexible protein-protein docking
Prediction of binding regions	**ArDock**	(Reille et al. 2018)	https://ardock.ibcp.fr	Prediction and the visualization of binding site on protein surfaces
	JET2 Viewer	(Ripoche et al. 2017)	http://www.jet2viewer.upmc.fr/jet2_viewer/report/index.html	Pre-computed interfaces of PDB chains
	PredUs	(Q. C. Zhang et al. 2011)	http://honig.c2b2.columbia.edu/predus	Prediction of Protein Interfaces Using Structural Alignment
	PROFISIS	(Ofran and Rost 2007)	https://rostlab.org/owiki/index.php/PROFisis	ML-based prediction of protein interfaces from sequences
	SPPIDER	(Porollo and Meller 2007)	http://sppider.cchmc.org	ML-based prediction of protein interfaces from structures
	WHISCY	(de Vries, van Dijk, and Bonvin 2006)	https://wenmr.science.uu.nl/whiscy	Conservation and structure-based interface prediction
Template-based prediction	**COTH**	(Mukherjee and Zhang 2011)	https://zhanglab.dcmb.med.umich.edu/COTH	Multiple-chain protein threading algorithm
	HOMCOS	(Kawabata 2016)	https://homcos.pdbj.org	Complex threading and template-based docking
	INstruct	(Meyer et al. 2013)	http://instruct.yulab.org	Database of protein interaction network with structural resolution
	Interactome3D	(Mosca, Céol, and Aloy 2013)	https://interactome3d.irbbarcelona.org	Structural annotation of PPI networks
	PrePPI	(Q. C. Zhang et al. 2013)	https://honiglab.c2b2.columbia.edu/PrePPI	Database of known and predicted PPIs
	PRISM	(Tuncbag et al. 2011; Baspinar et al. 2014)	http://cosbi.ku.edu.tr/prism	Structural similarity and conservation-based PPIs prediction
	Struct2Net	(R. Singh et al. 2010)	http://cb.csail.mit.edu/cb/struct2net/webserver	Structure-based threading approach

Protein-protein interaction databases	**BioGRID**	(Oughtred et al. 2021)	https://thebiogrid.org	Biomedical database for PPIs
	CCSB	(Rolland et al. 2014)	http://interactome.dfci.harvard.edu/index.php?page=home	Human interactome map by Y2H screening
	ConsensusPathDB	(Kamburov et al. 2009)	http://cpdb.molgen.mpg.de	Database for human functional interactome
	CORUM	(Giurgiu et al. 2019)	https://mips.helmholtz-muenchen.de/corum	Database of manually annotated mammalian PPIs
	DIP	(Salwinski et al. 2004)	https://dip.doe-mbi.ucla.edu/dip	Database for experimentally determined PPIs
	HAPPI-2	(Chen, Pandey, and Nguyen 2017)	http://discovery.informatics.uab.edu/HAPPI	Human PPIs
	HitPredict	(Patil, Nakai, and Nakamura 2011)	http://www.hitpredict.org	Database for experimentally PPIs
	HPRD	(Peri et al. 2004)	http://www.hprd.org	Human protein reference database
	IntAct	(Hermjakob et al. 2004)	https://www.ebi.ac.uk/intact	Database for PPIs from literature curation or direct user submissions
	MINT	(Licata et al. 2012)	https://mint.bio.uniroma2.it	Database for experimentally verified PPIs from the scientific literature
	Reactome	(Croft et al. 2011)	https://reactome.org	Database of reactions, pathways, and biological processes
	STRING	(Szklarczyk et al. 2021)	https://string-db.org	Database for known and predicted physical interactions and functional associations
Machine learning-based modeling of protein complexes	**ColabFold**	(Mirdita, Ovchinnikov, and Steinegger 2021)	https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/beta/AlphaFold2_advanced.ipynb	AlphaFold v2.0-based PPI structure prediction

Appendix B

Paper Two: Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-21636-z OPEN

Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences

Anna G. Green [1,8], Hadeer Elhabashy[2,3,4,8], Kelly P. Brock[1], Rohan Maddamsetti [1], Oliver Kohlbacher [2,3,4,5,6]✉ & Debora S. Marks [3,7]✉

Increasing numbers of protein interactions have been identified in high-throughput experiments, but only a small proportion have solved structures. Recently, sequence coevolution-based approaches have led to a breakthrough in predicting monomer protein structures and protein interaction interfaces. Here, we address the challenges of large-scale interaction prediction at residue resolution with a fast alignment concatenation method and a probabilistic score for the interaction of residues. Importantly, this method (EVcomplex2) is able to assess the likelihood of a protein interaction, as we show here applied to large-scale experimental datasets where the pairwise interactions are unknown. We predict 504 interactions de novo in the *E. coli* membrane proteome, including 243 that are newly discovered. While EVcomplex2 does not require available structures, coevolving residue pairs can be used to produce structural models of protein interactions, as done here for membrane complexes including the Flagellar Hook-Filament Junction and the Tol/Pal complex.

[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] Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany. [4] Department of Computer Science, University of Tübingen, WSI/ZBIT, Sand 14, 72076 Tübingen, Germany. [5] Quantitative Biology Center, University of Tübingen, Auf der Morgenstelle 8, 72076 Tübingen, Germany. [6] Institute for Translational Bioinformatics, University Hospital Tübingen, Sand 14, 72076 Tübingen, Germany. [7] Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA. [8] These authors contributed equally: Anna G. Green and Hadeer Elhabashy. ✉email: oliver.kohlbacher@uni-tuebingen.de; debbie@hms.harvard.edu

A longstanding goal of molecular biology is to determine the three-dimensional structure of protein interactions at atomic resolution. However, despite the 'resolution revolution' in cryo-electron microscopy[1], atomic resolution of complexes is still labor intensive, lagging far behind the number of known protein interactions in any one organism[2–4]. For instance, even in the well-studied proteome of *Escherichia coli*, less than half of probable protein interactions have been identified[4] and less than 9% have been structurally characterized, even accounting for similarity to 3D structures of complexes in related species. Hence there is continued interest in computational approaches that can accelerate the discovery of protein interactions in 3D at atomic resolution.

There have been many experimental[5–7] and computational methods[8–10] to identify which proteins interact within an organism to scale, but the only computational methods able to determine both interactions and their precise, residue-resolution interfaces are based on coevolution. Coevolutionary methods such as EVcouplings[11,12] and others[13] have been successful in determining 3D structures by leveraging the vast corpus of natural sequences using probabilistic graphical models to infer candidate pairs of interacting residues. These coupled positions can be sufficient to fold single proteins[11–13] and RNAs[14], without the use of homologous structures, and even resolve protein interaction interfaces de novo[15,16]. However, in the previous work that resolved interacting residues between proteins, coevolution was typically benchmarked on fewer than 100 examples and limited by using co-operonic proteins or phylogeny to identify sequence pairs. Therefore, de novo prediction was limited to select, small datasets or individual runs on a webserver[15–18]. While this work was in preparation, new work was published that used proteome-scale coevolution to provide a dataset of 804 candidate interactions in *E. coli*[19]. Though these predictions are valuable, the methodology is not available for queries of new protein pairs or new large-scale predictions (see Results for detailed comparison).

Here we address the challenges of concatenation and statistical scoring to provide a new method that integrates information from residue-level scores to determine whether two proteins interact and the residues involved in their interaction (Fig. 1) without reliance on co-operonic positions, phylogeny, or docking. We estimate that the scope is 53% of all possible protein pairs for a bacterial genome such as *E. coli*, compared to previous methods that rely on genomic position (~10%). As proof of principle, we present results on a new benchmark set, which allows us to assess the specificity and sensitivity of the method. As a case study, we predict 504 interactions involving cell envelope and membrane proteins, which are some of the most challenging to determine experimentally. We provide docked models of these complexes, some of which are of particular biological interest and are discussed in more depth (the Tol/Pal systems and the Flagellar Hook-Filament Junction). We also provide predicted protein interactions at residue resolution for an AP/MS dataset where the precise details of interactions are unknown, as well as residue-level predictions for a Yeast Two-Hybrid dataset. To demonstrate the potential for eukaryotic complexes, we also show successful predictions for eukaryotic-exclusive complexes including the human spliceosome.

Results

Most protein interactions in *E. coli* do not have structural resolution. Most protein interactions within an organism remain structurally unresolved, even for well-studied model organisms such as *E. coli*. Of the estimated 10,000 protein interactions in *E. coli*, approximately 3,946 have been observed[4], involving about 50% of all *E. coli* proteins. Of these protein-protein interactions, 9% have an experimentally determined structure or can be inferred trivially from solved homologous interactions[19], hence leaving a large fraction of interactions unknown or without structural resolution (Fig. 1).

The majority of the protein monomers in the *E. coli* proteome (3,189 out of a total of 4,391) have high-quality monomer alignments and are therefore amenable to EVcomplex2 (Methods). We verify that these alignments are of high quality by testing the precision of the top ECs for those monomers with an experimental structure (either in *E. coli* or another organism), finding that 78% have reasonable precision of the top ECs (60% for the top L ECs, where L is the protein sequence length) (Supplementary Data 1, Supplementary Fig. 1). We therefore restrict our computational predictions of co-evolution to the set of 3,189 monomers with high-quality alignments (i.e., 75% of proteins, covering 53% of possible interaction space).

Probabilistic score allows flexibility for computing residue resolution interactions. This work addresses two main challenges for determining residue resolution of interactions. The first challenge is to avoid the use of genome location or operonic structure and hence be positioned to investigate any protein pair for interacting residues. To solve this, we constructed alignments of pairs of proteins from different organisms by identifying a single protein per species with reciprocal highest identity to the query sequence (Methods). This increases the number of protein interactions that can be tested in *E. coli* by approximately fivefold across the proteome when compared to reliance on genome distance and opens the door to genome-scale interactions in eukaryotic organisms (Supplementary Note 1). This scope and improved coverage will continue to increase as sequencing becomes even more commonplace, as this is a major remaining limitation.

The second challenge is calibrating a method that can both predict residue contacts for known interactions and also use predicted residue contacts to detect protein interactions with high confidence. True interactions are extremely sparse, representing a small fraction of a very large number of possible combination of proteins; for instance, it has been estimated that only about 10,000 of the nine million protein pairs in *E. coli* will interact[4] (Fig. 1). Therefore, even small false positive rates will result in large numbers of predicted interactions that are false. To minimize the number of false positive interactions, we calibrate on both positive and negative benchmarks: a 'gold standard' non-redundant set of 561 interactions with known structure, and a set of 3,987 non-redundant "non-interacting" protein pairs with no signal for interaction in prior high throughput experiments[2,4] (Supplementary Data 2, Methods).

On these benchmarks, our previous state-of-art method, the EVcomplex score threshold[15], has a 13% false positive rate, resulting in large numbers of incorrect pairs predicted to interact. These and other methods[15,16,18] may still be useful if the interaction is already presumed known, but the false positive rate is prohibitively high for unbiased proteome-scale screens. We therefore built a semi-supervised learning method combining the unsupervised EVcomplex score with structure-agnostic features of coevolving residues in a logistic regression: the EVcomplex score, sum of intra-protein EC coupling constraints, the rank of inter-EC relative to intra-EC pairs, and sequence conservation (Supplementary Data 3, Methods). This model achieves a recall of 20.7% at a false positive rate of 0.1% on the held-out test set and outperforms the previous EVcomplex score across all false positive rate thresholds. For cases where the three-dimensional structure of both monomers, is known, the recall can be further increased to 22.6% at a false-positive rate of 0.1% (Supplementary

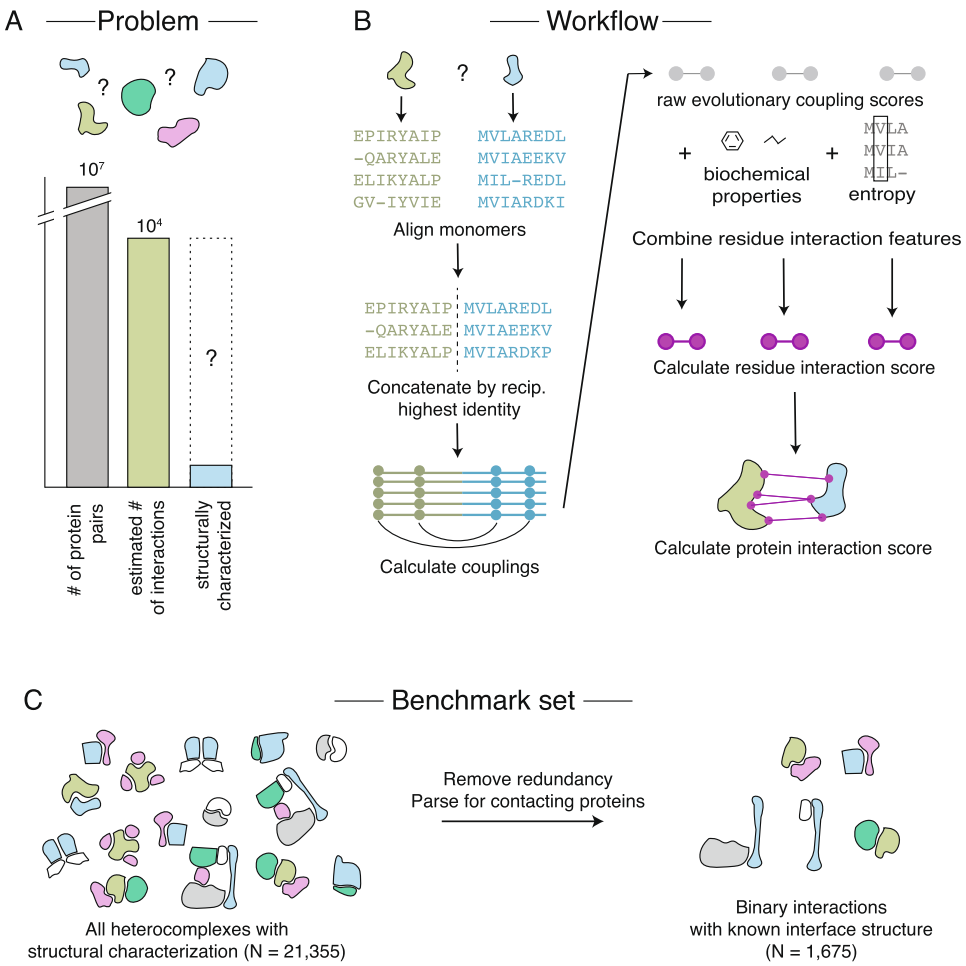


Fig. 1 Scaling EVcouplings methods to full bacterial genomes. A The search problem for binary protein-protein interactions in *Escherichia coli* involves finding all of the estimated 10^4 true interactions out of 10^7 possible pairs. Only approximately 9% of these true direct interactions have a crystal structure solved in *E. coli* or a homologous structure in another organism. **B** Evolutionary couplings learned directly from protein sequences can resolve interfaces. Sequence alignments of both monomeric proteins are created and concatenated by the reciprocal highest identity 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.

Fig. 3A). We provide full tables of the performance of these scores across all false positive rates, so that users can select thresholds appropriate for their application (Supplementary Data 4). Our model performs comparably to Yeast Two-Hybrid experiments, achieving a recall of 27.9% versus the 29% obtained experimentally at a false positive rate of 1%[4], with the added feature of providing residue-level interaction for all protein pairs found.

The logistic regression model also outperforms the previous scoring method 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 59.9% of the true ECs in our held-out test set using our previous scoring method, 65.6% of true ECs with the logistic regression model, and 69.3% with the logistic regression model when incorporating features of known monomer structures (Fig. 2, Supplementary Fig. 3). Both our previous EVcomplex score and our new models outperform other previously used scoring methods (raw EC score[19], Z-score[20], and raw EVcomplex score[15]) at both predicting interacting proteins and identifying interface residue contacts (Supplementary Fig. 3).

Recently, another computational method using co-evolution to search for proteome scale interactions at residue resolution was published[19]. Based on their reported false-positive rates on their benchmark set, our method is less sensitive for determining the interaction of proteins (our method: 25% recall at FPR 0.34% (Supplementary Data 5), their method: 36% recall at FPR 0.34%). However, this was not a head-to-head comparison and their increased recall may be due to the fact that they did not filter their positive benchmark set for redundancy. In a head-to-head comparison, our method performed better for predicting inter-protein contacts given a known interaction (58.1% vs 75.3% precision across all residues for their reported interacting complexes with known structure, Methods, Supplementary Data 5). Our increased precision comes at some expense in terms of recall, in some cases because accurate inter-protein ECs fall below the chosen scoring threshold, and in other cases possibly because numerous paralogs rendered our concatenation method inaccurate (4% and 18% of the dataset, respectively) (Supplementary Note 1, Supplementary Fig. 4). These comparisons should be considered provisional as their pipeline[19] is not yet available, so we were not able to test their method on our benchmark dataset, nor were we able to adjust their scoring threshold. An additional advantage of our model is that we provide distinct scoring methods for when interactions are known and unknown. We anticipate this will be important for

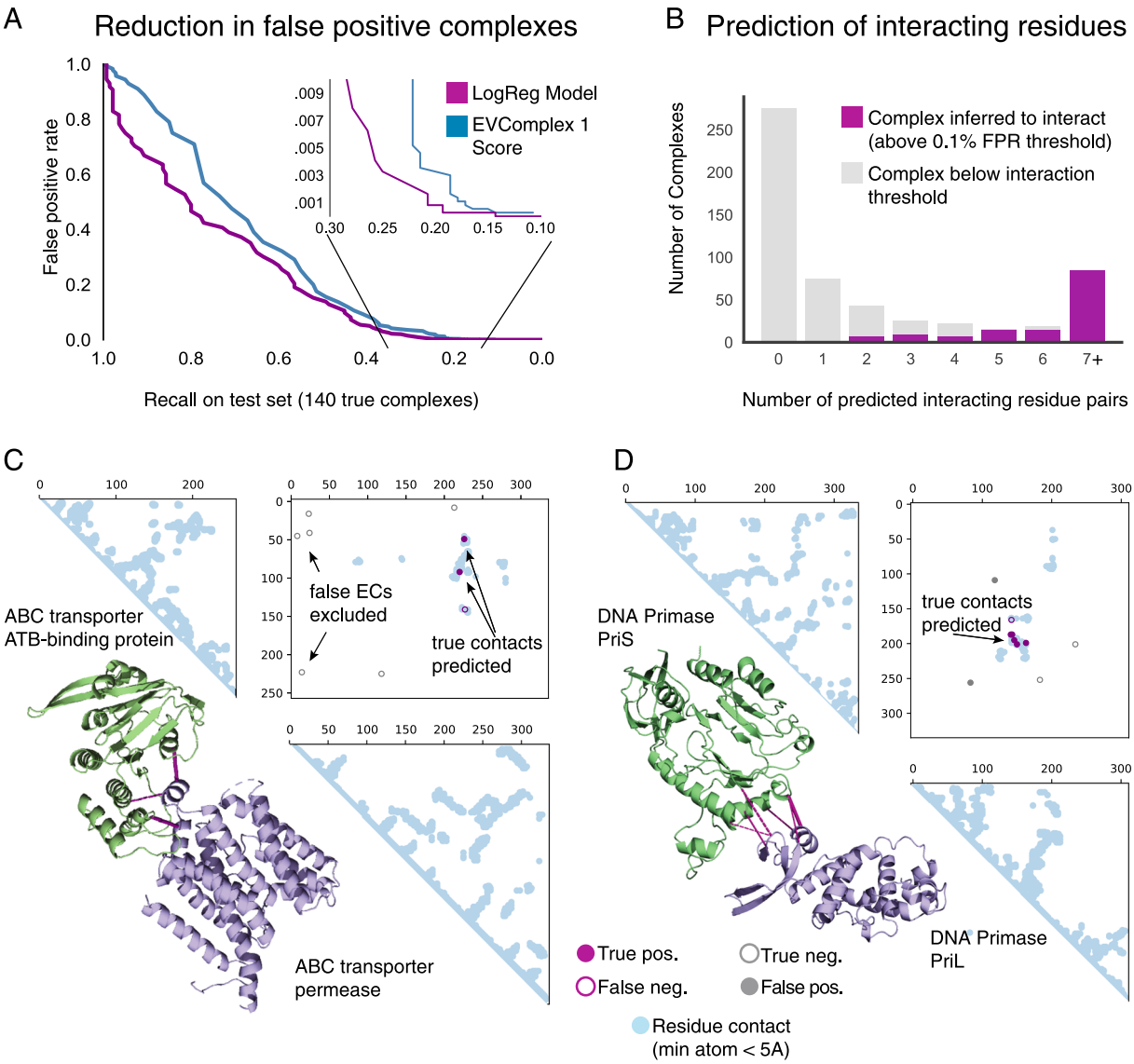


Fig. 2 Model predicts interacting residues and interacting proteins with high precision. A Recall on the held-out fraction of the positive benchmark set (x-axis) and false positive rate on the held-out dataset of non-interacting complexes (y-axis) at a score threshold that gives the corresponding recall. Our logistic regression model (purple) reduces the false positive rate compared to the previous EV Complex score (blue). **B** Prediction of protein interaction is based on the prediction of interacting residues. Number of predicted interacting residue pairs for complexes inferred to interact (purple) or not inferred to interact (gray) based on our stringent protein complex prediction threshold. **C** Example performance on known interaction between ABC transporter permease and ATP binding subunit (UniProt IDs: Y1470_HAEIN and Y1471_HAEIN, PDB ID: 2NQ2 chains C and A [https://www.rcsb.org/structure/2NQ2]. **D** Example performance on known interaction between DNA primase PriS and PriL (UniProt IDs: PRIS_SULSO and PRIL_SULSO, PDB ID: 1ZT2 chain A and B [https://www.rcsb.org/structure/1ZT2]).

users who have candidate interactions, e.g., from mass spectrometry experiments.

New predicted membrane protein interactions with strong co-evolution. Since the cell membrane contains many protein interactions essential for life, but is notoriously difficult to study experimentally[21], we targeted the cell envelope proteome for detailed analysis. We based our analysis on 1583 proteins previously described as localized to the *E. coli* cell envelope[2]. We assayed each compartment of the *E. coli* cell envelope proteome with itself and with adjacent compartments (Methods) (Fig. 3A), for a total of 939,159 protein pairs. After monomer sequence alignment, concatenation, and EC calculation, 198,534 protein pairs comprised of 1053 proteins (566 non-redundant protein families) pass quality thresholds for analysis. The majority (771) of these proteins are inner membrane proteins. 49% of these proteins have at least a partial structure known of themselves or a homologous protein.

We predict 504 interactions in the *E. coli* cell envelope as well as the details of their interacting residues, including 243 interacting pairs not previously observed experimentally (Fig. 3B, Supplementary Data 7, Supplementary Fig. 3). Despite having chosen a stringent false positive rate threshold of 0.1%, we expect that up to 39.1% of these interactions may be false positives due to the large size of the tested space. Of 305 known membrane protein interactions with known structures that can be analyzed by our method, we recover an interaction signal for 76 of them (Supplementary Fig. 5), similar to the recall expected based on our calibrations (24.9% here vs 20.7% in calibration). The protein

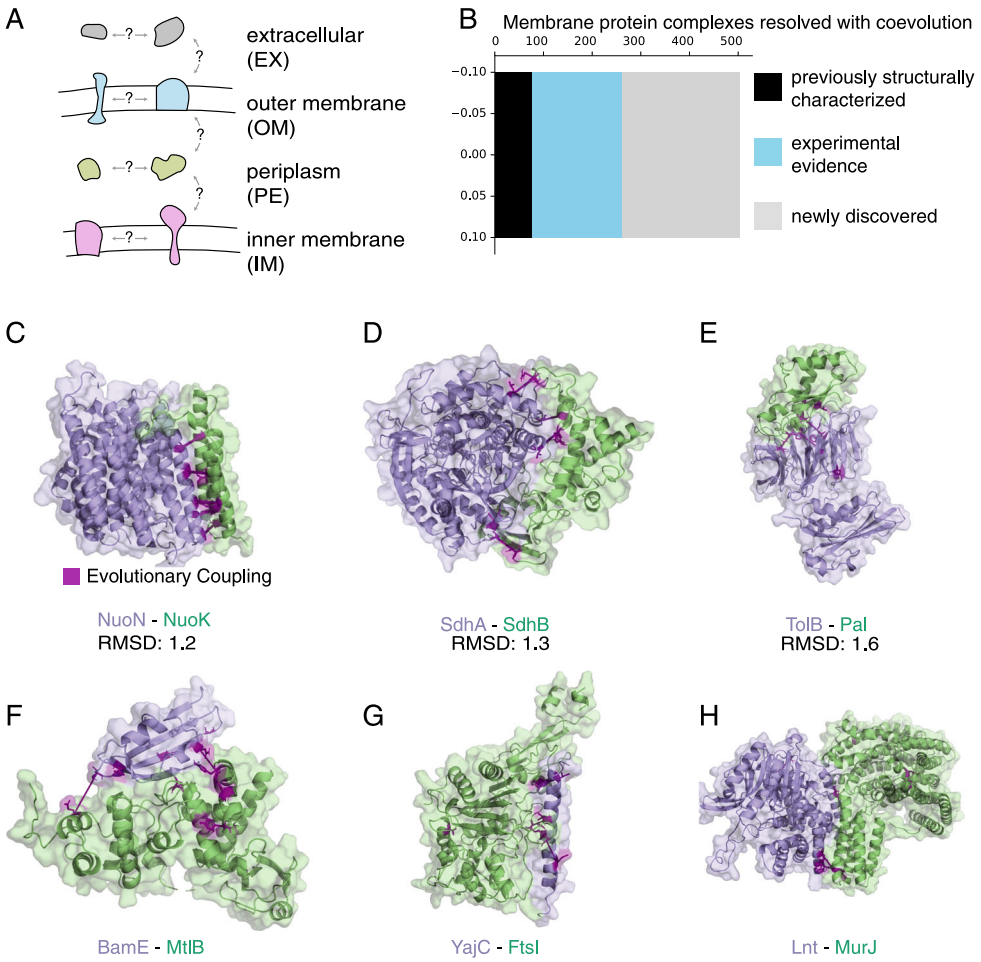


Fig. 3 Discovery of hundreds of new interactions in the *E. coli* membrane proteome. A We searched a high-value subset of the 10^7 possible interactions in the *E. coli* proteome by searching membrane compartments with themselves and with adjacent membrane compartments. **B** We found 504 high-scoring protein interactions in the cell envelope, including 75 with structural characterization and 186 with previous experimental evidence (and no structural characterization). **C**–**E** 3D configurations of previously structurally characterized interactions are accurately predicted by molecular docking with inferred restraints. RMSDs calculated by comparison to known structures with PDB IDs 3RKO [https://www.rcsb.org/structure/3RKO], 2WU2 [https://www.rcsb.org/structure/2WU2], and 2HQS [https://www.rcsb.org/structure/2HQS], respectively. **F**–**H** Example of three docked models of newly resolved protein complexes: BamE/MltB, YajC/FtsI, and Lnt/MurJ. Evolutionarily coupled residues used as restraints in docking are shown in magenta and connected with solid lines.

interactions in the *E. coli* membrane with the highest interaction score are involved in essential cellular processes; for instance, the RodA and penicillin-binding protein cell wall polymerases, which we previously characterized in a detailed study[22], as well as components of the electron transport chain, components of the ATP synthase, and components of the bacterial flagellum.

Although AP/MS has recently been used to identify 12,807 co-purifying membrane proteins, it is not always known which proteins interact with which other proteins within their respective complexes[2]. We therefore computed the probability score of the pairs within each complex to determine interacting pairs, finding significant residue-level coevolution for 872 pairs (three or more ECs above the 80% precision residue interaction threshold), 262 of which would have been blindly identified using our stringent protein interaction scoring threshold (Supplementary Data 7). These predictions can be used to disambiguate directly interacting proteins in complexes identified via AP/MS, such as the Tol-Pal complex reported below.

Yeast Two-Hybrid experiments have provided a rich dataset of candidate interactions in *E. coli*[4]. EVcomplex2 can both increase confidence in the inferred interactions and also provide residue resolution for the predicted complexes. For 21% ($N = 559$) of complexes found to interact by Yeast Two-Hybrid, we find at least one EC above the 80% precision residue interaction threshold (and 198 complexes have three or more ECs) (Supplementary Data 7). This does not necessarily mean that complexes with no residue coevolution signal do not interact, as lack of signal may be due to chosen threshold in our benchmarking or to lack of evolutionary conservation of complexes.

Docking of newly resolved membrane-associated interactions. The inter-protein residue pairs identified by our method can be used as restraints for molecular docking, to resolve the 3D structure of protein complexes[15,16] (Fig. 3C–E). In addition, docking lends additional confidence to ECs below the residue interaction scoring threshold: ECs below the threshold end up satisfied in the final model for 58% of the docked benchmark complexes (Supplement, Supplementary Fig. 6, Supplementary Data 8, Supplementary Data 10).

Of all inferred membrane protein interactions, we docked 59 pairs that had structures of both monomers available, including 36 with no previous structure of their interaction (Supplementary

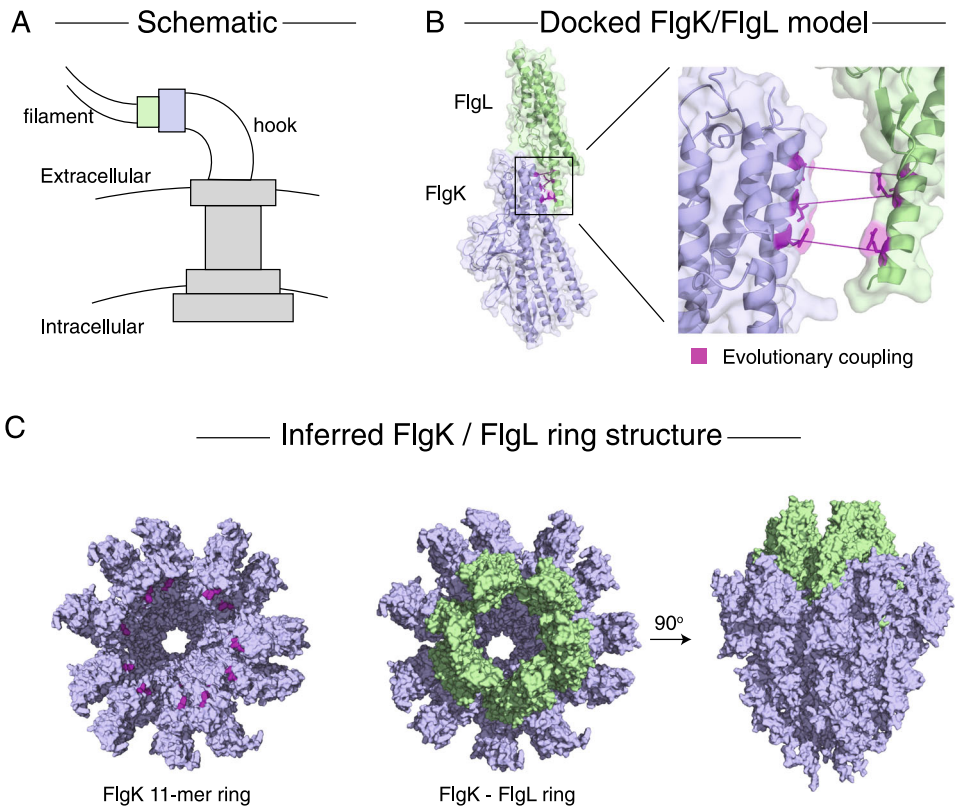


Fig. 4 Model of the flagellar hook-filament junction. A Schematic of the orientation of the bacterial flagellum. The proteins FlgL (green) and FlgK (blue) form two rings which create the junction between the hook and filament of the flagellum[24]. **B** Docked model of FlgL and FlgK using evolutionary couplings. PDB structures of homologous proteins from *Salmonella typhimurium* were used in docking (PDB IDs 2D4Y [https://www.rcsb.org/structure/2D4Y], and 2D4X [https://www.rcsb.org/structure/2D4X], respectively). Predicted interface residues are highlighted in purple. **C** A previously inferred model of the FlgK ring from *Campylobacter jejuni*[24] 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.

Data 9, Supplementary Data 11). This includes the proteins BamE and MltB (Fig. 3F), which had previously been determined to be in the same complex in an AP/MS experiment but the details of their interaction were unknown[2]. We determine an interaction and resolve the interface for the peptidoglycan biosynthesis protein FtsI and the subunit of the Sec translocation complex YajC (Fig. 3G), as well as for Lnt and MurJ (Fig. 3H), involved in lipoprotein maturation and peptidoglycan biosynthesis, respectively.

Our method resolves the interaction interface between flagellar proteins FlgL and FlgK, which form the junction between the flagellar filament and flagellar hook[23]. Extrapolating from our docked model of the monomers and a previous model of the FlgK ring from *Campylobacter jejuni*[24], we infer the configuration of an 11-mer ring of FlgL inside an 11-mer ring of FlgK[24] (Fig. 4, Supplementary Fig. 7).

We constructed an atomic model of the Tol/Pal system (Fig. 5), a molecular machine spanning the inner and outer membranes of Gram-negative bacteria. These proteins may play a role in membrane constriction during bacterial cell division[25]. For the known interacting interface between TolB and Pal[26], we accurately recapitulate the known structure (PDB ID: 2HQS [https://www.rcsb.org/structure/2HQS]). We predict interaction and residue contacts for the known interaction between TolA and TolB[4,27] as well as residue contacts and a likely interaction of TolR and TolB in the membrane[28]. For CpoB, which interacts with the Tol system and the cell division septum site[29], we provide a molecular model of its interaction with TolB.

Current scope of EVcomplex for interactions exclusive to Eukaryotes. The prediction of protein interactions using coevolution is not restricted to bacterial genomes. However, predicting protein interactions using coevolutionary methods in eukaryotes is difficult due to the smaller number of sequenced genomes and higher numbers of paralogous genes compared to bacteria. We used our positive benchmark set, which contains structures from all domains of life, to assess the prospects of EVcomplex2 for predicting protein interactions at residue resolution in eukaryotes (Methods). Of the 1,675 protein pairs in our dataset, 977 are found predominantly in eukaryotes (more than 90% of concatenated sequences are eukaryotic in origin) (Fig. 6). As expected, a higher fraction of the eukaryotic protein complexes were excluded from downstream analyses due to low sequence diversity compared to bacterial protein complexes (78% versus 53% excluded, respectively). Approximately 11% of the 977 eukaryotic protein interactions have at least one interacting residue pair correctly predicted by EVcomplex2, and 4% have three or more correctly predicted (Fig. 6).

An important eukaryotic complex is the spliceosome, an RNA–protein complex which splices intronic sequences from immature mRNA in eukaryotes[30]. EVcomplex2 correctly predicts the interface between Prp38 and MFAP1, whose interactions contribute to the activation of the spliceosome[31] (Fig. 6). Excluding alignments with low sequence diversity, the next two highest-ranked interactions in the spliceosome are between LSm2 and MFAP1 as well as between Prp38 and LSm5, which could plausibly come into contact during rearrangement of the

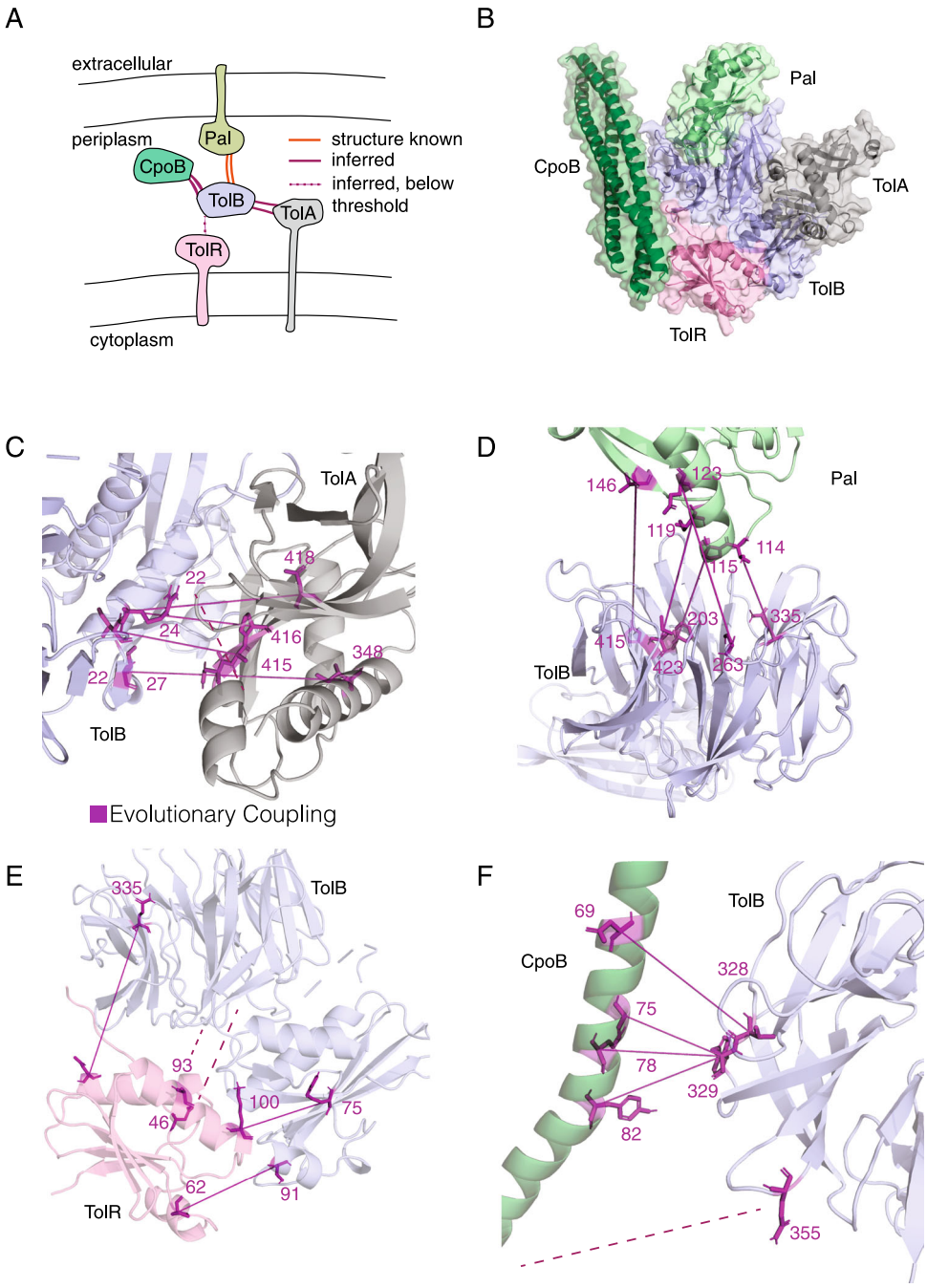


Fig. 5 Atomic resolution model of the Tol-Pal complex. A Schematic of the proteins involved in the Tol-Pal complex[2] (TolABR, CpoB, and Pal). Interactions with previously solved interfaces are shown in orange and interactions inferred by our method are shown in purple. **B** Complete model of the Tol-Pal 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 **C**–**F** 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 are missing from the solved structure. Structures used are 1TOL_A (TolA) [https://www.rcsb.org/structure/1TOL], 2HQS_A (TolB) [https://www.rcsb.org/structure/2HQS], 5BY4_A (TolR) [https://www.rcsb.org/structure/5BY4], 2HQS_H (Pal) [https://www.rcsb.org/structure/2HQS], and 2XDJ_A (CpoB) [https://www.rcsb.org/structure/2XDJ].

spliceosome during activation. These results demonstrate the possibility of using sequence coevolution-based methods for eukaryotes.

We assessed the current prospects for predicting whether proteins interact and their residue resolution on the 74,449 proteins in the human proteome, by aligning their constituent PFAM domains[32] (Methods). We find that 42% of proteins in the human proteome can be aligned with medium sequence diversity in at least one domain, and 20% of proteins can be aligned with the high diversity cutoff used for *E. coli* in all of their domains. These represent approximate upper and lower bounds on the number of proteins that would be amenable to coupling-based analysis with any concatenation method. Of the 3×10^9 possible interactions in the human proteome, this means that between 4% and 17% of protein interactions are currently accessible to EVcomplex2.

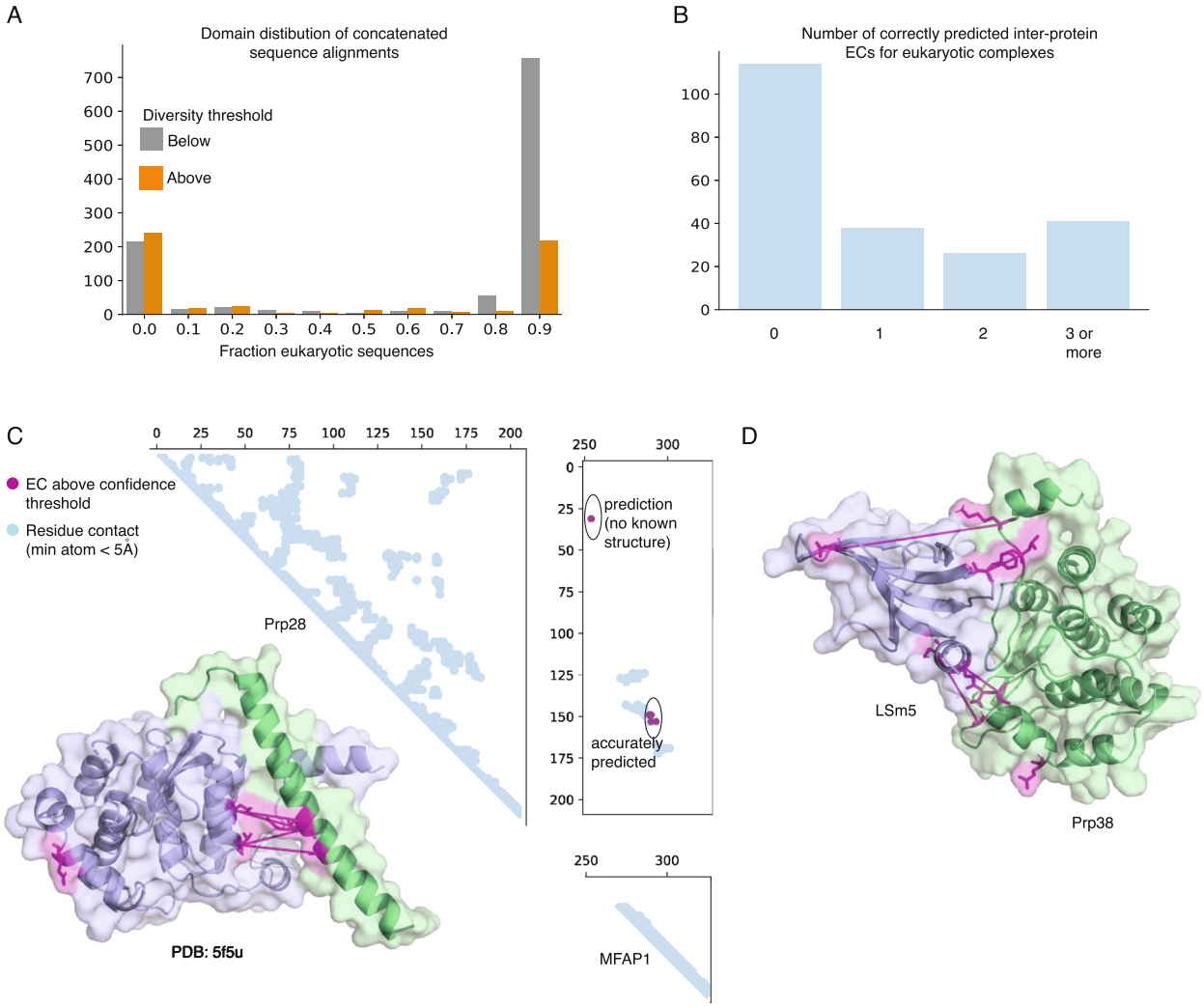


Fig. 6 Predicted protein interactions in eukaryotic proteomes. A Distribution of eukaryotic sequences in concatenated sequence alignments. Shown in orange are alignments that passed our sequence diversity threshold, and in gray are those that did not. **B** Number of correctly predicted inter-protein ECs for eukaryotic-exclusive complexes above the sequence diversity threshold. Eukaryotic-exclusive complexes are defined as complexes whose concatenated sequence alignment is at least 90% eukaryotic sequences. Inter protein ECs are defined as correct if their minimum atom distance is <8 Å. **C** The human spliceosome proteins Prp38 and MFAP1 have a known interface correctly predicted by our method, and have a protein interaction detected. The inter-protein ECs above the 80% precision threshold used throughout the paper are shown in purple. Known protein structure (PDB ID: 5F5U, [https://www.rcsb.org/structure/5F5U]) is used to visualize the subunits. **D** Schematic of predicted interaction between Lsm5 and Prp38 with top 5 inter-protein ECs.

Discussion

In this work, we demonstrate a method for inferring protein interactions at residue-level detail from genomic sequences alone. We provide a probabilistic scoring method, calibrated on large, non-redundant datasets, which can be used to determine the probability of interaction of two proteins. If the protein interaction is known or inferred by the protein interaction score, the residue interaction score can be used directly to determine residues in contact at the interface. These methods can be freely accessed via our website (https://marks.hms.harvard.edu/ecolicomplex/) or by downloading the source code (https://github.com/debbiemarkslab/EVcouplings)[33].

The highest reciprocal identity concatenation method introduced here applies to more complexes than concatenation methods relying on genomic distance to correctly pair interacting paralogs[15], and perform comparably to a recent method that relies on reciprocal best hits for entire bacterial genomes[19]. Our method only requires the sequence annotation already present in the UniProt database and therefore has the advantage of not requiring curation and all-vs-all sequence searches of full genomes[19], phylogeny construction[18], or extensive iterations[34,35], rendering it very fast with no dependencies. We show that cases where our concatenation method is outperformed tend to have larger numbers of paralogs, suggesting that other more computationally intensive techniques could be selected based on prior knowledge about the protein pair.

Our method assumes that interactions between orthologous proteins are conserved across species, an assertion which is certainly not true for all complexes and will lead to false negatives in the case of newly evolved or poorly conserved interactions. In addition, since our method has been calibrated on experimentally resolved structures, it may inherit their biases. In particular, crystal structures are known to be biased against membrane proteins, alternative conformal states, and transient interactions, which could lead to false-negative predictions. This bias is likely to be more pronounced in eukaryotes, which tend to have more multi-domain and intrinsically disordered proteins than prokaryotes[36].

With current sequence databases, 53% of the protein pairs in the *E. coli* proteome are eligible for EVcomplex2. By using our

highest reciprocal identity criterion, we restrict ourselves to one ortholog per genome, which limits the amount of sequence diversity that can be found for families with multiple paralogs per genome. This limitation is most consequential in eukaryotes, where there are fewer sequenced genomes to draw from. Previous studies showed that eukaryotic protein complexes could be successfully predicted if they had prokaryotic homologs[20,37], and our study extends this finding to show that some eukaryotic-exclusive complexes now have sufficient sequence diversity for prediction. Growth in sequence databases and the development of fast concatenation techniques to handle paralogous sequences will increase the number of protein pairs that can be analyzed. Two recent papers have shown that sequence diversity generated in the lab can be used to determine protein structures[38,39], opening an interesting future avenue for using experiments to determine the structure of protein complexes.

Method development using coevolution to resolve protein-protein interactions has thus far been tested on a handful of well-studied protein-protein interactions[15–17,34,35,40], reaffirming our ability to predict well-conserved, well-characterized protein interactions. While there are many possible avenues for continuing to improve coevolution-based prediction of protein interactions, we suggest that future development should begin with large, non-redundant biological datasets such as ours. By identifying and diagnosing cases where existing methods fail on real datasets, new developments will lead to greater opportunities to discover unknown protein interactions.

Methods

Creation of positive benchmark set. We built a dataset of non-redundant heterodimeric protein-protein interactions with resolved crystal structures. A list of all heteromeric structures in the PDBePISA was downloaded[41] (download date: 2/19/2018) and passed through the RCSB Protein Data Bank with the following query: "Sequence Length is between 30 and 1200 and Resolution is between 1.0 and 5.0 and Representative Structures at 30% Sequence Identity." IDs were mapped to UniProt[42] identifiers using the SIFTS database[43] (download date 2/20/2018). Only complexes with at least two chains that map to different UniProt identifiers were kept. Single-chain and fusion proteins were removed because they may have non-specific interface contacts not found in nature.

For each protein, we extracted the PFAM domains[32] annotated in that protein, which we call the PFAM set for that protein. We then consider a protein-protein interaction as unique if and only if the interacting proteins constitute a pair of PFAM sets not yet seen in our database, yielding 4154 non-redundant pairs of interactions. We then consider only contacting pairs of proteins, defined as a protein pair with at least 20 pairs of residues with a minimum atom distance less than five angstroms. Because each unique interaction can be represented by many crystal structures, representative crystals were chosen based on the number of other interfaces present in that crystal – i.e., complexes with more subunits were prioritized. We removed any cases where the two proteins map to the same chain, where the two proteins share any PFAM domain, or where either of the proteins was less than 30 amino acids in length, resulting in a final list of 1,675 protein complexes with known interfaces (Supplementary Data 2).

Creation of negative benchmark set. A total of 2,500 negative examples were selected randomly from all pairs of complexes with no signal for interaction in a large-scale Yeast Two-Hybrid experiment in *E. coli*[4], as well as all pairs of proteins whose proteins were found in different APMS benchmark complexes[1]. We verified that no pair of proteins had the same PFAM set as any other pair of proteins in the negative set and that no pair had the same PFAM set as any pair of proteins in the positive benchmark set. We further removed protein pairs with any links in the STRING database[44], protein pairs previously inferred to interact (including protein-ribosome pairs)[19], and proteins from membrane protein export apparatuses (as annotated in Babu et al., 2017): "BamABCDE Outer Membrane Protein Assembly Complex", "TatABCE protein export complex", "SecD-SecF-YajC-YidC Secretion Complex/Sec Translocation Complex", "Sec Translocation Complex/SecYEG translocase", "Sec Translocation Complex", "GspC-O secreton complex." The final dataset included 3987 protein pairs (Supplementary Data 2).

Monomer sequence alignment. For alignments of the positive benchmark set, all UniProt[42] identifiers corresponding to the sequence of interest were extracted from PDB[45]. For alignments of the *E. coli* genome, all identifiers from the reviewed *Escherichia coli* strain K12 reference proteome (UP000000625) were downloaded from UniProt. For both the positive benchmark set and *E. coli* genome set, the entire length of every protein was used for sequence alignment. Alignments were constructed using jackhmmer version 3.1b2[46] with five iterations against the UniProt database (downloaded February 2018), as implemented in the EVcouplings software package. Alignments were considered to have sufficient coverage if at least 80% of the columns were less than 50% gaps. A range of bit-scores were tested (0.1, 0.2, 0.3, 0.5, and 0.8 times protein length L). For each alignment, the N_{eff} was calculated by downweighing each sequence by the number of other sequences with more than 80% identity. The alignment with the largest N_{eff} that had 80% coverage of columns was selected for each protein. Proteins from the *E. coli* genome were deemed eligible for EVcomplex analysis if their monomer sequence alignment had at least 80% column coverage and $N_{eff}/L >= 2.5$, after characterizing the precision of monomer protein couplings across a range of sequence alignment diversities (Supplementary Fig. 1).

Concatenation. Monomer protein sequences were concatenated for EVcomplex2 analysis. We first annotate the species of origin for each sequence by extracting the species annotation from the UniProt[42] database. A single representative for each species was chosen based on the highest identity to the query sequence. After excluding closely related paralogs in the query genome with more than 90% sequence identity, we confirm that the candidate hit has the highest identity to the query sequence than to other paralogs from the query species found during our sequence alignment protocol. If for a single species, a reciprocal highest identity hit for both monomers can be identified, those hits are concatenated and added to the alignment.

Concatenated alignment quality control. To avoid analyzing protein pairs where the two proteins are homologous, we implemented four different quality control metrics. First, we remove all pairs of proteins where the first protein contains a PFAM domain that is found in the second protein. Then, we remove all pairs where our structure comparison protocol found overlapping hits to the same chain of the same PDB structure. Next, we remove all pairs that display high-scoring coevolution along a diagonal between the two proteins (i.e, between position i in protein A and the corresponding position i in protein B). We consider these contacts to be artifactual due to their very high scores relative to known interactions. Finally, we do not consider proteins from *E. coli* where there is a non-traditional amino acid annotated in the UniProt sequence, because these can indicate pseudogenes. To ensure high-diversity, low gap alignments for inference, we removed from consideration concatenated sequence alignments with N_{eff}/L less than 0.2 or coverage of columns less than 80%.

EC calculation. For calculation of evolutionary couplings (ECs), hits composed of more than 50% gaps were filtered from the alignment, and sequences with homologs more than 80% identical were downweighed to compute N_{eff}, the effective number of sequences[12]. ECs were calculated using pseudo-likelihood maximization[14,47,48]. The λ_J term was scaled by the number of amino acids minus one times the number of sites in the model minus one[49]. Pre- and post-processing was performed using the EVcouplings Python package[33].

Raw EC scores, sometimes called the Corrected Norm (CN) score, were calculated by performing the average product correction (APC) on the Frobenius Norm scores output by plmc. Diagonal ECs, where positions i and j are within 5 amino acids, were not included for further analysis. For analysis of raw ECs, the Z-score of each inter-residue EC was calculated using the distribution of all inter-residue ECs. The EVcomplex score was calculated by dividing each inter-protein EC by the absolute value of the minimum inter-protein EC score (for the raw EVcomplex score), and then normalizing for the N_{eff}/L of the concatenated alignment (for the EVcomplex score)[15]. The EVcomplex2 score is calculated by inputting the couplings scores and additional features into the model described under "Detecting inter-protein residue contacts", for predicting residue interactions when the protein interaction is presumed known, or by inputting the coupling scores and additional features into the model described under "Detecting protein interactions" for predicting protein interactions.

Comparison of ECs to experimental structures. To identify experimental structures for comparison of evolutionary couplings, the monomer sequence alignments were searched against the PDB database using *hmmsearch* version 3.1b2[46]. For *E. coli* monomer alignments, the top 20 were selected. For the validation dataset, all hits above a bitscore threshold of 0.2 L were selected. Once structures were selected for comparison, the minimum of the nearest atom inter-residue distances across all structures were used for comparison to ECs[11].

Solvent accessible surface area calculation. Relative and absolute solvent accessibilities were calculated using DSSP (CMBI version by M.L. Hekkelman/2010-10-21)[50] with a probe size of 1.4 Å and standard van der Waals radii[51]. If multiple structural hits were found for each protein, we take the mean relative solvent accessibility across all structures, to avoid biases caused by incomplete structures.

Conservation calculation. The conservation C for each column i in the concatenated sequence alignments were calculated as one minus the normalized entropy for the column i,

$$C = 1 - H/\log_2 20 \quad (1)$$

$$H = -\sum_{k=1}^{20} f_k \log_2 f_k \quad (2)$$

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

Detecting inter-protein residue contacts. We trained a logistic regression classifier to predict the probability that two residues i and j are in contact (distance < 8 Å), given their EC score and additional residue-level features. This model for residue interaction was trained on the positive benchmark dataset, which includes 560 protein complexes for which the concatenated sequence alignment has sufficient coverage and diversity. Because most of the inter-protein ECs are noise and do not correspond to real contacts[15,16], we take the top ten inter-protein ECs as input to the model. Residue pairs where one or both residues were missing in all structures were removed from the dataset because their distance is unknown. This resulted in 5,344 residue pairs with a defined minimum atom distance, 26% of which were within 8 Å.

140 (25%) of the 561 complexes had all of their ECs held out and used for testing. Models were trained using the Logistic Regression classifier in Scikit-learn v. 0.23.5[52], using the LIBLINEAR solver and L1 regularization. Regularization weight of 0.05 was chosen to maximize a tradeoff between recall and precision using ten-fold cross-validation on the training dataset.

The following features were evaluated for their ability to predict residue contacts in a logistic regression model with L1 regularization: the EVcomplex score[15]; the Z-score of the corrected norm (CN) score of all inter-protein ECs; the conservation of the more conserved for the two residues; the relative rank of the inter-protein ECs; the maximum of the intra-protein EC enrichment[11] for the two residues; the percent of residues in both positions that were hydrophobic (amino acids AVILMC), hydrophilic (amino acids STNQ), bulky (amino acids WFY), or charged (amino acids RHKDE) (4 separate features); and the combined hydrophilicity score at both positions. The features chosen for the final model based on non-zero weights in the L1 regression were the EVcomplex score[12], the conservation of the more conserved for the two residues, the relative rank of the first inter-protein EC, and the maximum of the intra-protein EC enrichment for the two residues[11] (Supplementary Data 3). The resulting model is referred to as the "structure-free" score.

We also trained a model that takes advantage of data from known monomer structures by incorporating two additional features: the minimum of the accessible surface area of both residues, and the precision of the ECs ranked above the current inter-protein EC. These two features were found to be significant in combination with the features found above. This model is referred to as the "structure-aware" score.

Detecting protein interactions. To build a model for determining whether two proteins interact, we use the score of the residue interaction model for the top ten inter-protein residue pairs as inputs to a logistic regression, as well as the relative rank of the highest-scoring inter protein EC (Supplementary Data 3). Our training dataset was composed of positive and negative examples. Positive examples were 420 protein pairs (75%) from the positive benchmark set, filtered so that all examples had at least one of the top ten inter-protein ECs within 8 Å, for a total of 322 complexes. Negative examples were 322 protein complexes randomly chosen from the negative benchmark set of Y2H negative complexes. Our test dataset was composed of the same 140 positive benchmark protein complexes withheld during the training of the residue interaction model, as well as all examples in the negative benchmark set not included during training.

Models were trained using the Logistic Regression classifier in Scikitlearn v. 0.23.5[52], using the LIBLINEAR solver and L1 regularization (weight = 0.1) to select features and regularization weight, chosen to maximize a tradeoff between recall and precision using ten-fold cross-validation. The final model was trained with L2 regularization using the LIBLINEAR solver in Scikitlearn v. 0.23.5.

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

Comparison to Cong et al., *Science*, 2019. Recall and false positive rate for prediction of interacting proteins was extracted from Cong et al., *Science*, 2019 Supplementary Table S3[19]. All protein pairs with a crystal structure listed were extracted from their Supplementary Table S8 ("Protein–protein interactions in *E. coli* identified in the co-evolution screen"). 331 of 339 protein pairs were successfully run using our pipeline, with the remaining 8 excluded due to technical issues or pathologies of the protein pair (Supplementary Data 4). Both our and their inter-protein ECs were compared to the same PDB structures as listed in their study, using a minimum atom distance cutoff of 8 Å to define interaction.

Our pipeline was considered to have predicted an inter-protein EC if the structure-free logistic regression score exceeded the 80% precision cutoff. To investigate the causes of discrepancies between prediction methods, we calculated the median number of paralogs per species for each of the constituent monomer alignments.

Cell envelope proteome. Localizations of *E. coli* proteins were extracted from a previous study[2]. Cell envelope proteins were separated into four categories: inner membrane-bound (integral or lipoprotein), periplasmic, outer membrane-bound (integral or lipoprotein), and extracellular. Proteins annotated as "membrane related" but without a specific location were excluded. Each of the four compartments was tested against itself and against proteins in the physically adjacent compartment(s): inner membrane versus periplasm, periplasm versus outer membrane, and outer membrane versus extracellular.

Docking. Side chains of all monomers were perturbed and optimized using SCWRL4[53] to reduce the effect of crystallization in a complex that may bias the docking procedure. Restraints-based docking was done using HADDOCK (High Ambiguity Driven biomolecular DOCKing) v2.2[54] and CNS[55]. This allows restricting the possible interaction search space using a more sophisticated treatment of conformational flexibility. The top five inter-ECs scores in both structure-aware and structure-free regression models were selected. Unambiguous distance restraints were applied on C-beta (except for glycine, where C-alpha was used) with an effective distance of 5 Å with an upper and lower bound of 2 Å. Default HADDOCK protocol was done starting with a rigid-body energy minimization (500 models), followed by semi-flexible refinement in torsion angle space (100 models), and ending with further refinement in explicit water (100 models). Upon binding, small conformational changes can be accounted for by HADDOCK through explicit flexibility during the molecular dynamics refinement. The resulting models were scored using the default HADDOCK scoring function, and the top ten best-scoring models, as well as the top models from RMSD clusters, are reported. To evaluate docking of our benchmark set where the structure of the interaction is known, we calculated Ligand-RMSD (L-RMSD) as well as Ligand centroid rotation angle and displacement in comparison to the reference structures.

Annotating domain of life for concatenated alignments. For all concatenated sequence alignments that comprise the positive benchmark set, the UniProt[42] accession numbers were mapped against the NCBI taxonomy database[56] We extracted the domain of life for each sequence in the database and calculated the percent of sequences in each alignment that were eukaryotic in origin.

Generating PFAM domain alignments. For estimating the number of possible human interaction pairs suitable for our analysis, we used a set of EVcouplings runs computed for all ~16 K PFAM domains in 2017 (PFAM release 30.0). A representative sequence from each domain family was chosen, and alignments were built for each sequence using jackhmmer[46] for five iterations at multiple length-normalized bit scores between 0.1 and 0.5 using the 2017 Uniref100 release[42]. Pseudo likelihood maximization was then used to compute evolutionary couplings scores for each possible pair of positions in the query.

Estimation of human protein pairs amenable to EVcomplex. We downloaded the set of 74,449 unique translation products annotated by UniProt human reference proteome UP000005640 (dated 12/2/2019). Each entry was then cross-referenced against the PFAM 32.0 release to identify component domains. Using the set of runs described above, we then identified how many domains in each protein had an alignment with high coverage (with over 80% of positions containing fewer than 30% gaps in the alignment) and a sufficiently diverse set of sequences (effective number of sequences per residue > = 1 or 5). These calculations were then used to report the number of human proteins amenable to EVcomplex reported in the text.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The benchmark and prediction datasets generated are available in the Supplementary Material, and full results are available online at https://marks.hms.harvard.edu/ecolicomplex/. The sequences and sequence annotation data analyzed during this study are available from UniProt (download date: Apr 1, 2017). The sequence location data analyzed during this study are available from ENA (download date: Feb 2017). The sequence taxonomy data analyzed during this study are available from the NCBI taxonomy database (download date: May 2020). The 3D-Structural information data analyzed during this study are available from the Protein Data Bank and SIFTS (download date Feb 1, 2018).

Code availability

EVcomplex2 is available as part of the EVcouplings Python Package, which is available on github under an MIT license: https://github.com/debbiemarkslab/EVcouplings and citable as https://doi.org/10.1093/bioinformatics/bty862.

Received: 22 January 2020; Accepted: 27 January 2021;
Published online: 02 March 2021

References

1. Kühlbrandt, W. The resolution revolution. *Science* **343**, 1443–1444 (2014).
2. Babu, M. et al. Global landscape of cell envelope protein complexes in *Escherichia coli*. *Nat. Biotechnol.* **36**, 103 (2018).
3. Hu, P. et al. Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. *PLoS Biol.* **7**, e96 (2009).
4. Rajagopala, S. V. et al. The binary protein-protein interaction landscape of *Escherichia coli*. *Nat. Biotechnol.* **32**, 285–290 (2014).
5. Rual, J.-F. et al. Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437**, 1173–1178 (2005).
6. Ho, Y. et al. Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183 (2002).
7. Butland, G. et al. Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433**, 531–537 (2005).
8. Li, Y., Calvo, S. E., Gutman, R., Liu, J. S. & Mootha, V. K. Expansion of biological pathways based on evolutionary inference. *Cell* **158**, 213–225 (2014).
9. Marcotte, C. & Marcotte, E. Predicting functional linkages from gene fusion. *Appl. Bioinf.* **1**, 1–8 (2002).
10. Pazos Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* **14**, 609–614 (2001).
11. Hopf, T. A. et al. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149**, 1607–1621 (2012).
12. Marks, D. S. et al. Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6**, e28766 (2011).
13. Ovchinnikov, S. et al. Large-scale determination of previously unsolved protein structures using evolutionary information. *Elife* **4**, e09248 (2015).
14. Weinreb, C. et al. 3D RNA and functional interactions from evolutionary couplings. *Cell* **165**, 963–975 (2016).
15. Hopf, T. A. et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, e03430 (2014).
16. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030 (2014).
17. Feinauer, C., Szurmant, H., Weigt, M. & Pagnani, A. Inter-protein sequence co-evolution predicts known physical interactions in bacterial ribosomes and the Trp operon. *PLoS ONE* **11**, e0149166 (2016).
18. Zeng, H. et al. ComplexContact: a web server for inter-protein contact prediction using deep learning. *Nucleic Acids Res.* **46**, W432–W437 (2018).
19. Cong, Q., Anishchenko, I., Ovchinnikov, S. & Baker, D. Protein interaction networks revealed by proteome coevolution. *Science* **365**, 185–189 (2019).
20. Rodriguez-Rivas, J., Marsili, S., Juan, D. & Valencia, A. Conservation of coevolving protein interfaces bridges prokaryote-eukaryote homologies in the twilight zone. *Proc. Natl Acad. Sci. USA* **113**, 15018–15023 (2016).
21. Díaz-Mejía, J. J., Babu, M. & Emili, A. Computational and experimental approaches to chart the *Escherichia coli* cell-envelope-associated proteome and interactome. *FEMS Microbiol. Rev.* **33**, 66–97 (2008).
22. Sjodt, M. et al. Structure of the peptidoglycan polymerase RodA resolved by evolutionary coupling analysis. *Nature* **556**, 118–121 (2018).
23. Homma, M., DeRosier, D. J. & Macnab, R. M. Flagellar hook and hook-associated proteins of *Salmonella typhimurium* and their relationship to other axial components of the flagellum. *J. Mol. Biol.* **213**, 819–832 (1990).
24. Bulieris, P. V., Shaikh, N. H., Freddolino, P. L. & Samatey, F. A. Structure of FlgK reveals the divergence of the bacterial Hook-Filament Junction of Campylobacter. *Sci. Rep.* **7**, 15743 (2017).
25. Gerding, M. A., Ogata, Y., Pecora, N. D., Niki, H. & De Boer, P. A. The trans-envelope Tol–Pal complex is part of the cell division machinery and required for proper outer-membrane invagination during cell constriction in *E. coli*. *Mol. Microbiol.* **63**, 1008–1025 (2007).
26. Bonsor, D. A. et al. Allosteric β-propeller signalling in TolB and its manipulation by translocating colicins. *EMBO J.* **28**, 2846–2857 (2009).
27. 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. *Mol. Microbiol.* **44**, 695–708 (2002).
28. Lloubès, R. et al. The Tol-Pal proteins of the *Escherichia coli* cell envelope: an energized system required for outer membrane integrity? *Res. Microbiol.* **152**, 523–529 (2001).
29. Gray, A. N. et al. Coordination of peptidoglycan synthesis and outer membrane constriction during *Escherichia coli* cell division. *elife* **4**, e07118 (2015).
30. Will, C. L. & Luhrmann, R. Spliceosome structure and function. *Cold Spring Harb. Perspect. Biol.* **3** (2011).
31. Ulrich, A. K. C., Seeger, M., Schutze, T., Bartlick, N. & Wahl, M. C. Scaffolding in the Spliceosome via Single alpha Helices. *Structure* **24**, 1972–1983 (2016).
32. El-Gebali, S. et al. The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427–D432 (2019).
33. Hopf, T. A. et al. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* (2018).
34. 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. USA* **113**, 12186–12191 (2016).
35. Bitbol, A. F., Dwyer, R. S., Colwell, L. J. & Wingreen, N. S. Inferring interaction partners from protein sequences. *Proc. Natl Acad. Sci. USA* **113**, 12180–12185 (2016).
36. Xue, B., Dunker, A. K. & Uversky, V. N. Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J. Biomol. Struct. Dyn.* **30**, 137–149 (2012).
37. Malinverni, D., Jost Lopez, A., De Los Rios, P., Hummer, G. & Barducci, A. Modeling Hsp70/Hsp40 interaction by multi-scale molecular simulations and coevolutionary sequence analysis. *eLife* **6**, e23471 (2017).
38. Rollins, N. J. et al. Inferring protein 3D structure from deep mutation scans. *Nat. Genet.* **51**, 1170–1176 (2019).
39. Schmiedel, J. M. & Lehner, B. Determining protein structures using deep mutagenesis. *Nat. Genet.* **51**, 1177–1186 (2019).
40. Bitbol, A. F. Inferring interaction partners from protein sequences using mutual information. *PLoS Comput. Biol.* **14**, e1006401 (2018).
41. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
42. UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **46**, 2699 (2018).
43. Velankar, S. et al. SIFTS: structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.* **41**, D483–D489 (2013).
44. Szklarczyk, D. et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43**, D447–D452 (2015).
45. Berman, H. M. et al. The protein data bank, 1999. in *International Tables for Crystallography Volume F: Crystallography of biological macromolecules* 675–684 (Springer, 2006).
46. Eddy, S. R. Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011).
47. Balakrishnan, S., Kamisetty, H., Carbonell, J. G., Lee, S. I. & Langmead, C. J. Learning generative models for protein fold families. *Proteins* **79**, 1061–1078 (2011).
48. 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 (2013).
49. Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proc. Natl Acad. Sci. USA* **110**, 15674 (2013).
50. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
51. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* **105**, 1–12 (1976).
52. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
53. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins: Struct., Funct., Bioinforma.* **77**, 778–795 (2009).
54. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein −protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737 (2003).
55. Brunger, A. T. Version 1.2 of the crystallography and NMR system. *Nat. Protoc.* **2**, 2728 (2007).
56. Federhen, S. The NCBI Taxonomy database. *Nucleic acids Res.* **40**, D136–D143 (2012).

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). A.G.G. was supported by National Science Foundation Graduate Research Fellowship DGE1144152 and National Institutes of Health NLM Training Grant T15LM007092. H.E. is funded by

the International Max Planck Research School (IMPRS) "From Molecules to Organisms." DM was supported by a Dean's Innovation Grant from Harvard Medical School and CZI Chan Zuckerberg Initiative 2018-191853.

Author contributions

A.G. and H.E. contributed to the conception and design of the research; acquisition, analysis, and interpretation of data; manuscript drafting and revision. A.G. contributed to the creation of software and developed the scoring model. H.E. performed the structural analysis and modeling of the complexes. K.B. obtain and analyzed human proteome alignment data and assisted with the manuscript. R.M. obtained and analyzed *E. coli* proteome alignment data and assisted with study design. D.M. and O.K. contributed to the conception and design of the research, data interpretation, manuscript writing, and supervision of the study. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21636-z.

Correspondence and requests for materials should be addressed to O.K. or D.S.M.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2021

Supplementary Material for:

Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences

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

[*] 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] Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14,72076, Tübingen
[4] Department of Computer Science, University of Tübingen, WSI/ZBIT, Sand 14, 72076 Tübingen, Germany
[5] Quantitative Biology Center, University of Tübingen, Auf der Morgenstelle 8, 72076 Tübingen, Germany
[6] Institute for Translational Bioinformatics, University Hospital Tübingen, Sand 14, 72076 Tübingen, Germany
[7] Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA

Supplementary Note 1

Comparison to operon-based sequence concatenation

In initial studies on detecting protein interaction at residue resolution from genome sequences, sequences were concatenated based on distance in their genome of origin, thus limiting the application to proteins found within the same operon in bacteria[1,2]. Our highest reciprocal identity concatenation method was designed to allow many more complexes to be accessible to coevolution-based methods than with our previous operon-based concatenation methods. To compare these two methods, we ran our previously published operon-based concatenation protocol[2] on our true positive benchmark dataset of 561 proteins that achieved sufficient sequence diversity in their concatenated alignment. This previous protocol pairs sequences if and only if they are (1) reciprocally the closest to one another on the genome versus any other paralog and (2) within 10,000 nucleotides. Genome location annotation was extracted

from the ENA database[3] (Download Date: Feb 2017), and protein sequences that could not be mapped to the ENA database were discarded in concatenation.

Of the 561 protein interactions in our positive benchmark dataset, 358 could not be concatenated to sufficient sequence diversity (N_{eff} / L >= 0.2) using operon distance because they are exclusively eukaryotic (**Methods)** or are too infrequently found in the same operon in bacterial genomes. For the remaining 203 complexes, we compared the accuracy of the top 10 ECs found using the two concatenation methods. 26% of protein pairs had the same accuracy with both methods, 39% had higher accuracy with highest reciprocal identity concatenation, and 35% had higher accuracy with the operon distance concatenation.

We investigated the reasons for the differences in performance by comparing the median number of paralogs found per species in the monomer sequence alignments. The complexes where performance is better with the operon-based concatenation method tend to have more paralogs in the monomer sequence alignments, (median of median per-species paralogs = 3) than those where performance is better with the highest reciprocal identity method (median of median per-species paralogs = 2). This indicates that while in general the highest reciprocal identity concatenation renders more complexes eligible and produces comparable or higher accuracy, bacterial complexes that are known to have large numbers of paralogs may be better predicted with an operon-based method.

Comparison of previously published scoring methods for ECs

Many scoring methods for inter-protein evolutionary couplings have been used in the literature, including the raw CN score (the result of applying the Average Product Correction to the Frobenius Norm score)[4], the Z-score of the raw CN scores,[5] and the EVcomplex score[2], a transformation of the CN score using the diversity of the concatenated sequence alignment [2]. Both the Z-score and the EVcomplex score are a monotonic transformation of the CN score. We assessed the performance of these scores at discovering true complexes and true residue contacts on our positive benchmark set of 561 complexes that achieved sufficient sequence diversity for prediction. Because these scoring methods are unsupervised, we used the entire benchmark set for assessment. We find that the corrected EVcomplex score performs best both at recovering true residue contacts, and recovering true complexes in our dataset (**Supplementary Fig. 2)**.

Comparison to Cong et al, *Science,* 2019

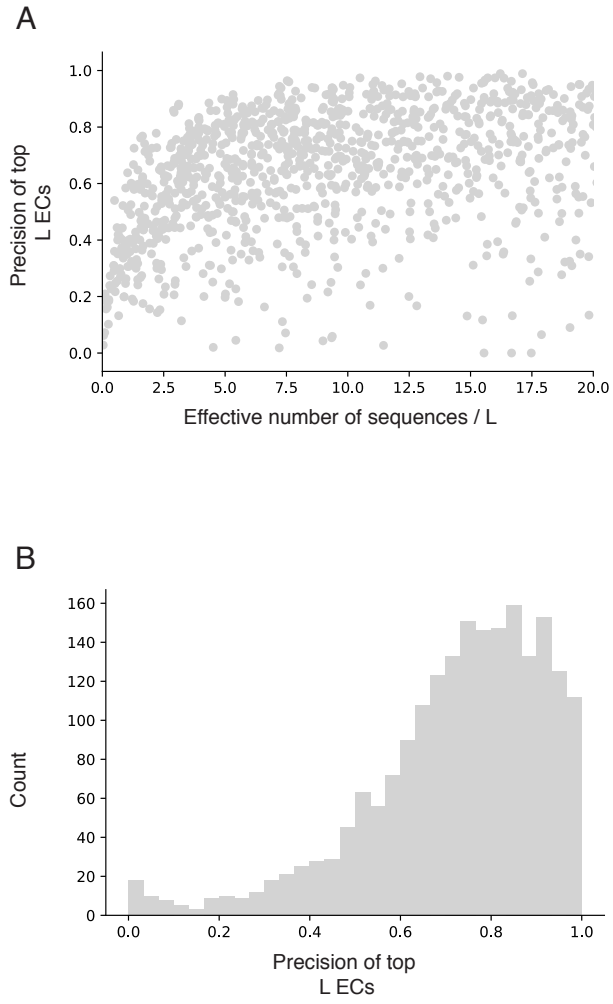
We performed a head-to-head comparison of interacting residue prediction by running the positive benchmark set from Cong et al., *Science*, 2019[4] using our pipeline. On their positive benchmark set of 339 complexes, our method reports higher-precision inter-protein ECs (58.1% vs 74.6%, across 220 complexes where we report significant ECs), at the expense of tending to report fewer ECs per complex (**Supplementary Fig. 4A, Supplementary Table 4)**. For the 109 remaining complexes where we report no significant ECs, 25 complexes were reported with a low precision (< 30%) prediction in the prior study as well, 21 complexes had high precision inter-protein ECs from our pipeline that fell below our defined confidence threshold, and 63

complexes were predicted with high precision by Cong *et al*. but not by our method. Those 63 pairs of proteins constitute only 25 non-redundant pairs (based on PFAM domain composition), and tend to have higher numbers of paralogs than complexes which performed well using our method **(Supplementary Fig. 4A)**.

Molecular docking on the positive benchmark set

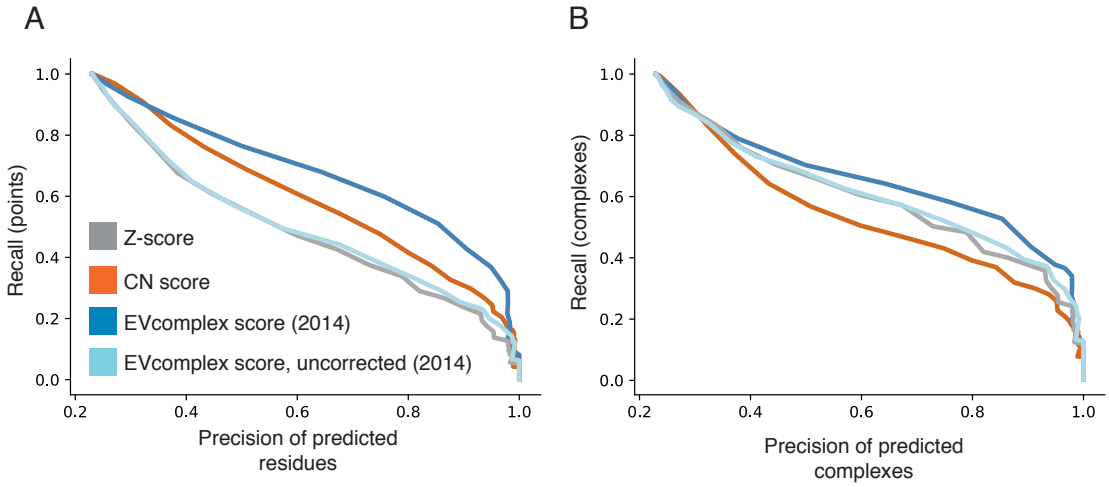
To assess whether molecular docking is capable of further increasing the accuracy of the predicted inter protein ECs, we performed molecular docking on a subset of our positive benchmark set, for a total of 173 complexes. Our docking procedure tends to produce accurate models, with 54% achieving a centroid ligand-RMSD < 10Å among those where at least three true restraints were used for docking, and 45% among all models **(Supplementary Figure 6)**. The more correct inter-protein ECs are used for docking, the more likely the models are to have low RMSD. Among these 173 protein complexes, we find that docking "rescues" correct ECs for 100 complexes – inter protein ECs that were below the confidence threshold end up in contact in the highest scoring docked model. This indicates that docking is capable of further boosting the signal for accurate inter-protein ECs, even if the top-scoring docked model is not itself accurate.

Supplementary Figure 1: High-quality sequence alignment of protein monomers from *E. coli*. (A) The precision of the top ECs increases as the effective number of sequences (N_{eff}) divided by sequence length (L) increases, with a plateau at approximately $N_{eff}/L = 2.5$. Each point corresponds to a monomeric protein in *E. coli* for which there is a crystal structure or structure of known homolog. The x-axis was truncated at 20 to show detail in the low range. (B) A histogram of the precision of the top L ECs for proteins whose sequence alignments had $N_{eff}/L >= 2.5$

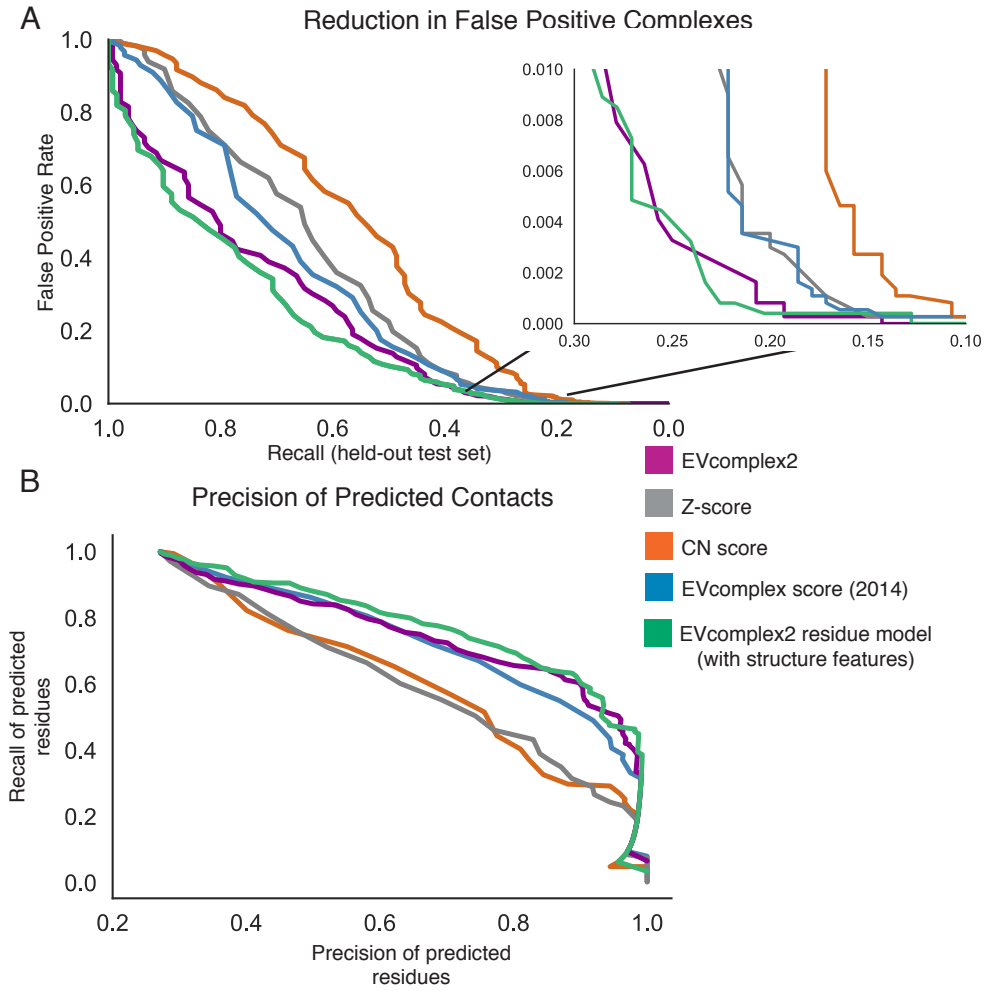


Supplementary Figure 2: Performance of previously published EC scoring methods on the benchmark dataset.

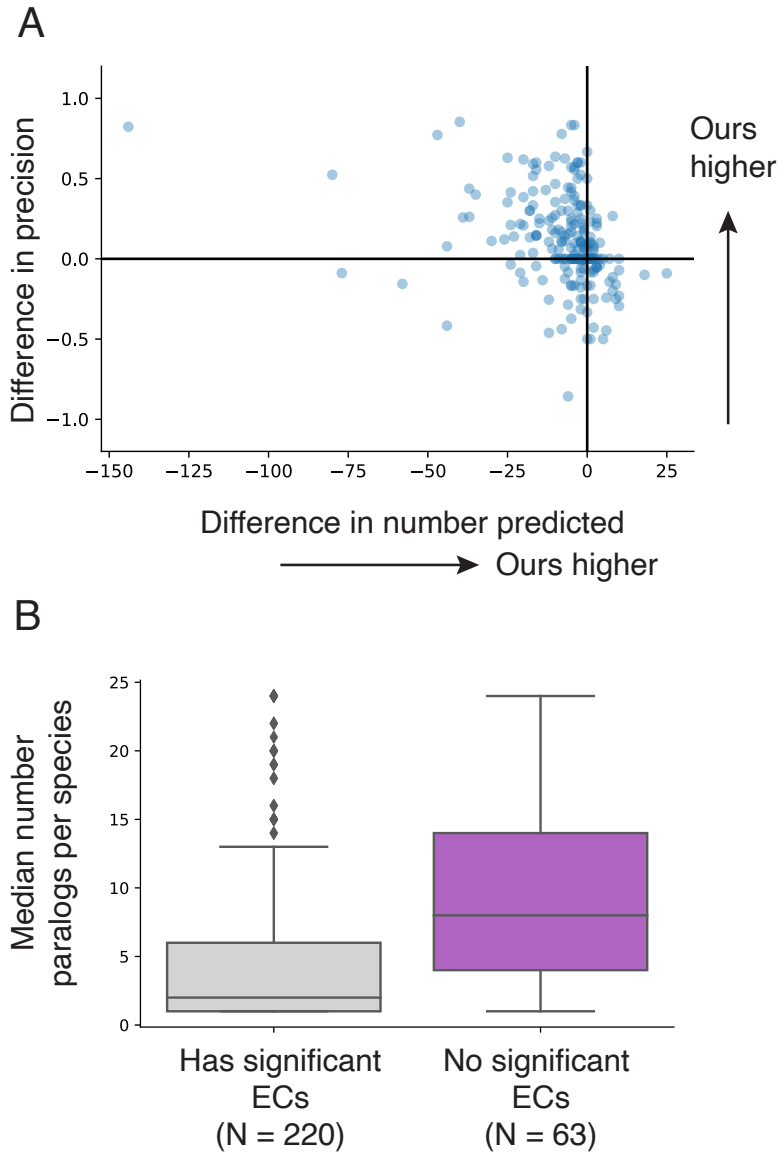
Performance of the corrected norm (CN) score, Z-score of the corrected norm score, and EVcomplex score (both corrected and uncorrected) on the entire benchmark dataset, considering the top 20 ECs for each complex. (A) Precision-recall curve for recovery of true residue contacts. Contacts are considered true if they are within 8Å (B) Precision-recall curve for recovery of true complexes. Complexes are considered 'true' if at least one of the top twenty ECs represents a contacting residue pair. Complexes are considered recovered if at least one EC is above the score threshold.



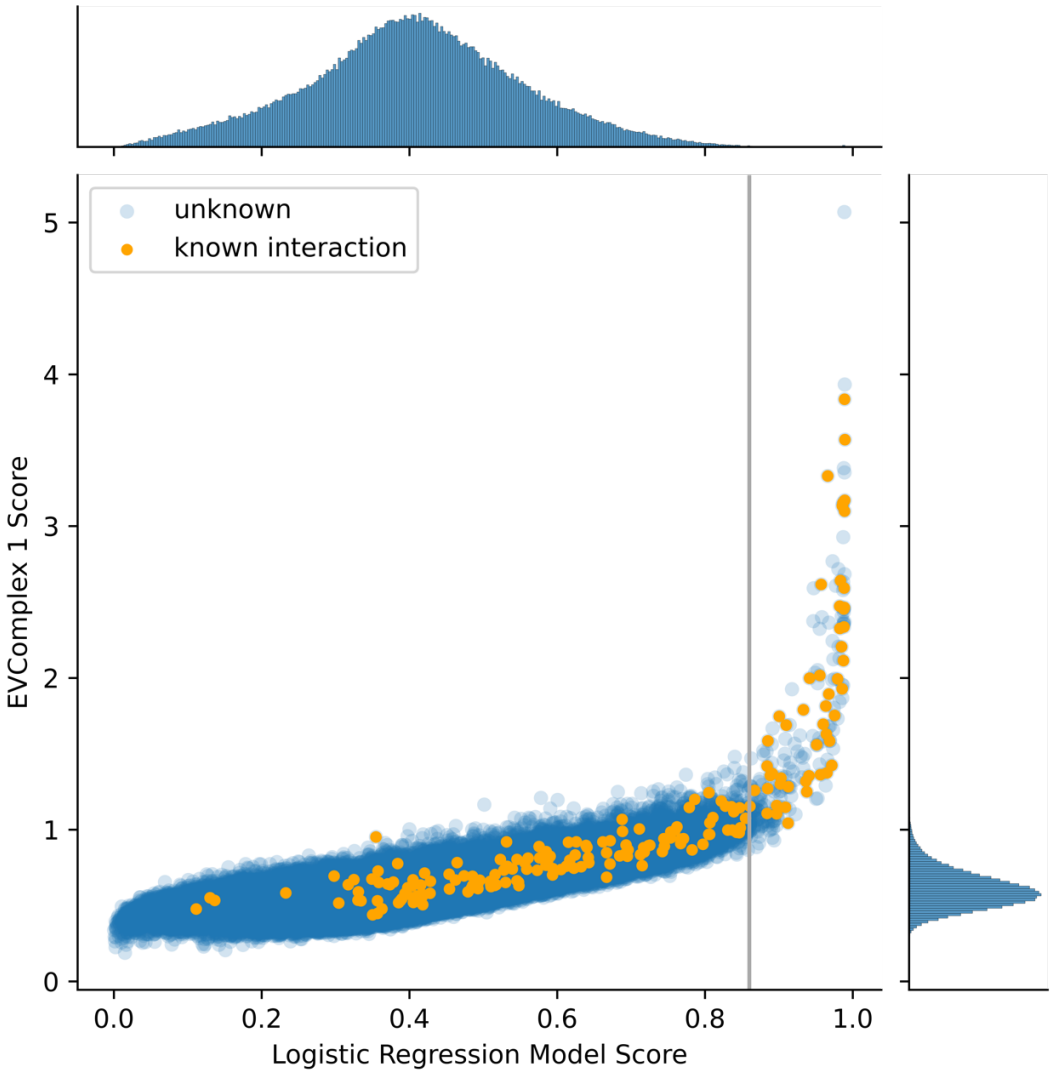
Supplementary Figure 3: Features of monomer protein structure further increase precision and recall on the held-out test set (A) Our logistic regression model (purple) reduces the false positive rate compared to unsupervised assessments of EC scores (grey, blue, and orange). A logistic regression model that also incorporates features of monomeric protein structure (green) further outperforms the structure-agnostic logistic regression (purple). The x-axis shows the recall on the held-out positive benchmark set, 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, incorporation of features of monomeric structures increases model recall for a given precision (purple), and precision is further increased when features based on monomer structures are introduced (green). The precision versus recall of the top 10 ECs for protein complexes with known interfaces is plotted. ECs are considered true if inter-residue minimum atom distance is less than 8Å.



Supplementary Figure 4: Comparison to Cong, et al. Science 2019. (A) Comparison of the precision and recall of our methods to those reported in Cong, *et al.* Science 2019, for 215 complexes from the Cong positive benchmark set for which we report significant inter-protein ECs under our structure-free residue scoring model. Our model tends to report higher precision ECs, at the expense of number of significant residues reported. (B) For those complexes reported with reasonable precision (>= 30%) in Cong *et al.* where our method predicts no significant ECs, we examine the median number of paralogs per species in the sequence alignment, and compare to the 215 with significant ECs using our method. Box plot center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers.

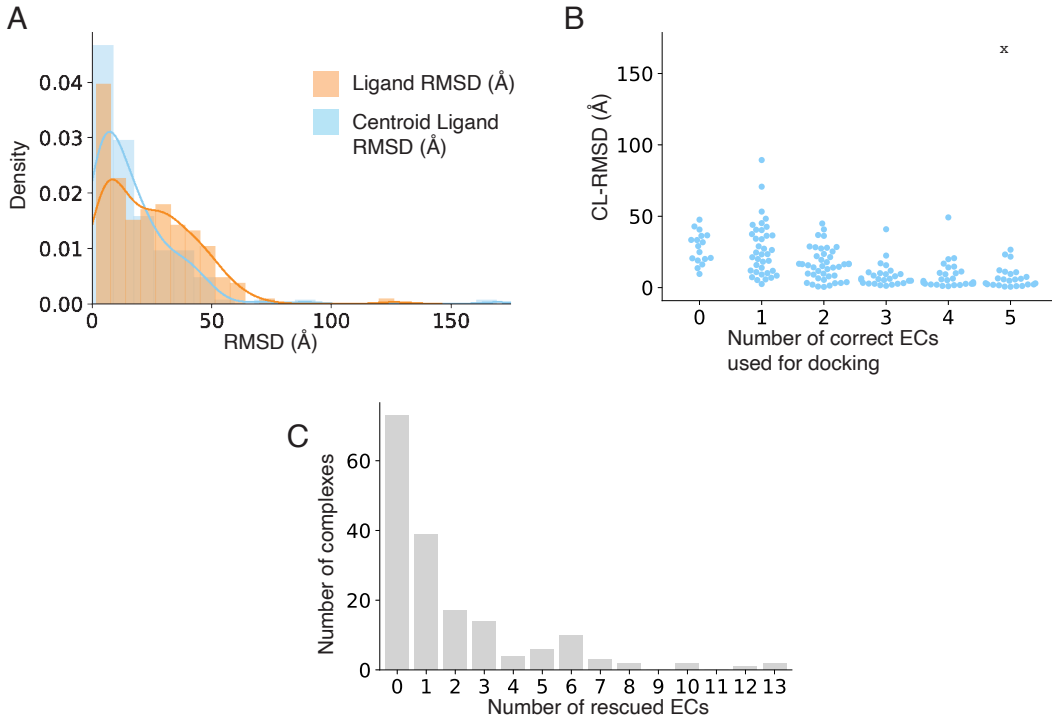


Supplementary Figure 5: Protein interaction scores for cell envelope proteome. Comparison of our new scoring method (Logistic Regression Model without structure-derived features), i.e. "EVcomplex2", to previous scoring method[2]. Orange dots indicate proteins found to interact in APMS experiments[7] with a solved crystal structure of their interaction, blue dots indicate all other proteins. Cutoff for inferring interaction used in this paper is shown as a grey line.



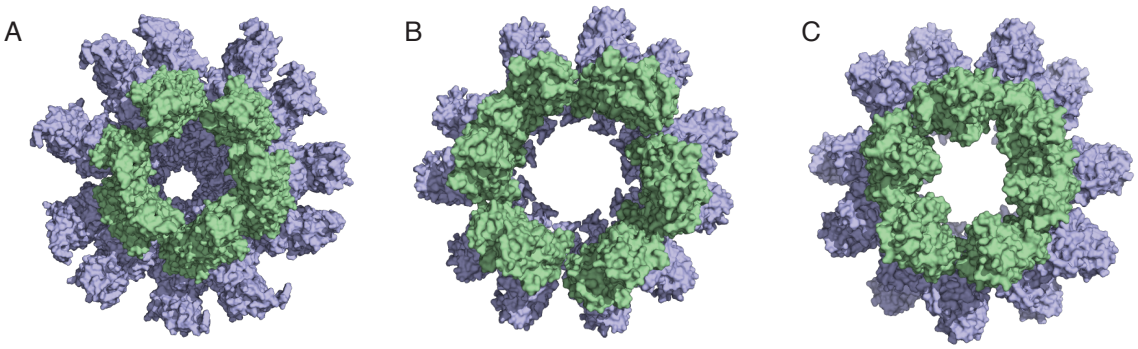
Supplementary Figure 6: Docking performance in the positive benchmark set

(A) RMSD values (Å) for docked models in the positive benchmark set (N=173). The RMSD of the top model, blindly selected based on the HADDOCK scoring function, for each protein complex is shown (B) Effect of the number of residues used for docking on the Centroid Ligand RMSD of the top model. Predictably, CL-RMSD is lower for models generated with more correct ECs (min. atom distance < 8Å). Of note is an outlier (represented here with an 'x'), RpoE/ChrR (PDB ID 2Q1Z [https://www.rcsb.org/structure/2Q1Z][8]), which is unsolvable with rigid-body docking because ChrR penetrates and loops around RpoE when the two are in complex – a major conformational change from the monomer state. (C) Number of true EC's below the structure-aware model residue interaction threshold that are "rescued" (*i.e.* have < 8Å min. atom distance) in the top docked model.



Supplementary Figure 7: Inferred ring structure of flagellar hook-filament junction from three species

Ring structure of FlgK (blue) and FlgL (green) from three different species (A) *Campylobacter jejuni,* (B) *Salmonella typhimurium,* (C) *Escherichia coli*. The *S. typhimurium* ring model was constructed by docking monomer structures 2D4Y [https://www.rcsb.org/structure/2D4Y] and 2D4X [https://www.rcsb.org/structure/2D4X] using evolutionary couplings, and then arranging eleven copies of the lowest energy model in a ring based on the coordinates of the *C. jejuni* FlgK ring model[9]. The *E. coli* ring model was created by making a homology model[10] of the *E. coli* FlgL and FlgK proteins against the monomer structures 2D4X and 2D4Y, performing docking using evolutionary couplings, and then arranging eleven copies of the lowest energy model in a ring based on the coordinates of the *C. jejuni* FlgK ring model. The *C. jejuni* ring was created using the structure 5XBJ [https://www.rcsb.org/structure/5XBJ] of FlgK and a homology model of the *C. jejuni* FlgL created with 2D4X as a template, performing docking using evolutionary couplings, and then arranging eleven copies of the lowest energy model in a ring based on the coordinates of the *C. jejuni* FlgK ring model.



Supplementary References

1 Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *Elife* **3**, e02030, doi:10.7554/eLife.02030 (2014).

2 Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3**, doi:10.7554/eLife.03430 (2014).

3 Pakseresht, N. *et al.* Assembly information services in the European Nucleotide Archive. *Nucleic Acids Research* **42**, D38-D43, doi:10.1093/nar/gkt1082 (2013).

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

5 Rodriguez-Rivas, J., Marsili, S., Juan, D. & Valencia, A. Conservation of coevolving protein interfaces bridges prokaryote-eukaryote homologies in the twilight zone. *Proc Natl Acad Sci U S A* **113**, 15018-15023, doi:10.1073/pnas.1611861114 (2016).

6 UniProt Consortium, T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **46**, 2699, doi:10.1093/nar/gky092 (2018).

7 Babu, M. *et al.* Global landscape of cell envelope protein complexes in Escherichia coli. *Nature biotechnology* **36**, 103 (2018).

8 Campbell, E. A. *et al.* A conserved structural module regulates transcriptional responses to diverse stress signals in bacteria. *Mol Cell* **27**, 793-805, doi:10.1016/j.molcel.2007.07.009 (2007).

9 Bulieris, P. V., Shaikh, N. H., Freddolino, P. L. & Samatey, F. A. Structure of FlgK reveals the divergence of the bacterial Hook-Filament Junction of Campylobacter. *Scientific Reports* **7**, 15743, doi:10.1038/s41598-017-15837-0 (2017).

10 Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. & Schwede, T. Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. *Sci Rep* **7**, 10480, doi:10.1038/s41598-017-09654-8 (2017).

Appendix C

Paper Three: Cross-linking of the endolysosomal system reveals potential flotillin structures and cargo

nature communications

Article

https://doi.org/10.1038/s41467-022-33951-0

Cross-linking of the endolysosomal system reveals potential flotillin structures and cargo

Received: 11 February 2022

Accepted: 6 October 2022

Published online: 20 October 2022

Check for updates

Jasjot Singh[1], Hadeer Elhabashy[2,3,4], Pathma Muthukottiappan[1], Markus Stepath[5,6], Martin Eisenacher[5,6], Oliver Kohlbacher[3,4,7], Volkmar Gieselmann[1] & Dominic Winter[1]

Lysosomes are well-established as the main cellular organelles for the degradation of macromolecules and emerging as regulatory centers of metabolism. They are of crucial importance for cellular homeostasis, which is exemplified by a plethora of disorders related to alterations in lysosomal function. In this context, protein complexes play a decisive role, regulating not only metabolic lysosomal processes but also lysosome biogenesis, transport, and interaction with other organelles. Using cross-linking mass spectrometry, we analyze lysosomes and early endosomes. Based on the identification of 5376 cross-links, we investigate protein-protein interactions and structures of lysosome- and endosome-related proteins. In particular, we present evidence for a tetrameric assembly of the lysosomal hydrolase PPT1 and a heterodimeric structure of FLOT1/FLOT2 at lysosomes and early endosomes. For FLOT1-/FLOT2-positive early endosomes, we identify >300 putative cargo proteins and confirm eleven substrates for flotillin-dependent endocytosis, including the latrophilin family of adhesion G protein-coupled receptors.

Lysosomes, the central lytic organelles of mammalian cells, are of crucial importance for cellular homeostasis. This is underscored by the detrimental consequences resulting from impairment of lysosomal function: mutations in genes encoding lysosomal proteins are causative for a group of ~70 rare and frequently devastating diseases, so-called lysosomal storage disorders (LSDs). Moreover, lysosomal dysfunction has been demonstrated in a number of more common conditions, including neurodegenerative diseases and cancer[1,2].

In addition to the long-known role of lysosomes in the degradation of intra- and extracellular substrates, more recent findings place them at the center of metabolic signaling. The major player in this context is the mammalian target of rapamycin complex 1 (mTORC1), whose activity is regulated at the lysosomal surface. This regulation is mediated by several protein complexes located in/at the lysosomal membrane, which integrate the activity of major signaling pathways, as well as the concentration of various metabolites[3]. Furthermore, protein complexes were shown to play a role in other lysosome-related processes, such as their transport, direct interaction with other cellular compartments, gene regulation, immunity, cell adhesion/migration, and plasma membrane repair[4].

For most of these functions, protein–protein interactions (PPIs) at the lysosomal membrane play a decisive role. Nutrient sensing and activation of mTORC1 are regulated by the interaction of at least 30 individual proteins[5], and lysosomal motility is controlled by the

[1]Institute for Biochemistry and Molecular Biology, Medical Faculty, University of Bonn, 53115 Bonn, Germany. [2]Department of Protein Evolution, Max-Planck-Institute for Developmental Biology, 72076 Tübingen, Germany. [3]Institute for Bioinformatics and Medical Informatics, University of Tübingen, 72076 Tübingen, Germany. [4]Department of Computer Science, University of Tübingen, 72076 Tübingen, Germany. [5]Medical Proteome-Center, Medical Faculty, Ruhr-University Bochum, 48801 Bochum, Germany. [6]Medical Proteome Analysis, Center for Protein Diagnostics, Ruhr-University Bochum, 48801 Bochum, Germany. [7]Institute for Translational Bioinformatics, University Hospital Tübingen, 72076 Tübingen, Germany.
e-mail: dominic.winter@uni-bonn.de

reversible association to microtubules through dynein and kinesin by several adaptor/scaffold complexes, such as BLOC1-related complex (BORC)[6]. The core feature of lysosomes, their acidic pH, is maintained by the 1.25 MDa vacuolar-type ATPase (V-ATPase) complex, which consists of 35 subunits (17 unique proteins), and catalyzes the transport of protons across the lysosomal membrane[7]. Delivery of certain lysosomal proteins is achieved by members of the homotypic fusion and protein sorting (HOPS)[8], the class C core vacuole/endosome tethering (CORVET)[9], as well as adaptor protein (AP) complexes, and the endosomal sorting complex required for transport (ESCRT) mediates repair of lysosomal membranes[10].

Also, for interactions of lysosomes with other organelles, protein complexes play an essential role. This includes fusion events with cargo delivery vesicles such as endosomes, phagosomes, and autophagosomes[4], exocytosis at the plasma membrane[11], or direct interactions with the endoplasmic reticulum[12], the Golgi apparatus[13] peroxisomes[14], RNA granules[15], and mitochondria[16]. The latter facilitates, for example, the exchange of small molecules and was shown to regulate events such as mitochondrial fusion and fission[4].

The majority of protein complexes that facilitate these processes are poorly characterized, and novel members/interactors are continuously being identified. Given the central role of lysosomes in metabolic regulation, and the high number of cellular structures they interact with, it is highly likely that several functionally important interactors of lysosomal proteins are still unknown.

Although structural data are available for a number of lysosomal luminal proteins and complexes in/at its membrane, three-dimensional information is still lacking for a significant fraction of the lysosomal proteome. The majority of existing structural data originates from crystallography experiments, heavily relying on affinity-purified proteins, or fragments thereof, from pro- or eukaryotic overexpression systems and their crystallization in vitro. The applicability of these structures to the in vivo situation remains therefore, in some instances, questionable[17].

A promising avenue to identify interactions of lysosomal proteins, and to obtain insights into their structure under physiological conditions, is chemical cross-linking in combination with mass spectrometry-based proteomics (XL-LC-MS/MS)[18]. In cross-linking experiments, a chemical linker forms covalent bonds between certain amino acids such as lysine. In subsequent MS analyses, these bonds are identified, providing direct proof for the interaction of proteins within a certain distance constraint, defined by the type of cross-linker[19]. This allows for the identification of PPIs, and hence localization, with high confidence. Compared to other commonly used approaches, such as immunoprecipitation (IP), proximity labeling[20], or lysosome enrichment[21], such data provide superior spatial evidence. Furthermore, the distance constraints of the cross-linker can serve as a basis for the molecular modeling of proteins and their complexes. This allows supplementing well-established techniques such as nuclear magnetic resonance (NMR), X-ray crystallography, or cryo-electron microscopy (cryo-EM), compensating for missing/incomplete data, and validating predicted protein structures[22].

So far, XL-LC-MS/MS experiments have been performed for samples of varying complexity, ranging from individual proteins[23] and multi-subunit complexes[24,25] to whole organelles[26,27] and cell/tissue lysates[28,29]. A dedicated analysis of lysosomal proteins by cross-linking has not been performed to date, which is certainly related to the fact that lysosomal proteins are of low abundance (estimated at 0.2% of cellular protein mass)[30].

In this work, we present a XL-LC-MS/MS dataset of the endolysosomal compartment of HEK293 cells, applying the MS-cleavable cross-linker disuccinimidyl sulfoxide (DSSO) to lysosome-/early endosome-enriched fractions. We present an interaction map of lysosomal proteins, of which we verify selected PPIs by co-IP and validate/extend existing protein structures. Based on the cross-linking data and integrative modeling, we further propose a tetrameric assembly of PPT1 and a heterodimeric structure for FLOT1/FLOT2. Finally, by affinity purification and MS analysis of FLOT1/FLOT2-positive early endosomes, we investigate the putative cargo of these vesicles and confirm selected candidates by co-IP and immunostaining.

Results

Cross-linking mass spectrometry analysis of lysosome-enriched fractions

In mammalian cells, the majority of lysosomal proteins are of relatively low abundance, and whole-cell XL-LC-MS/MS studies typically cover only a fraction of the lysosomal proteome (Supplementary Fig. 1a). A way to overcome this limitation is lysosome enrichment, which was shown by us to increase signal intensities for certain lysosomal proteins up to 100-fold relative to whole-cell lysates[31]. Accordingly, we enriched lysosomes by superparamagnetic iron oxide nanoparticles (SPIONs, Fig. 1a),[32,33] and established cross-linking conditions for lysosome-enriched fractions utilizing the MS-cleavable cross-linker DSSO[34]. Due to the limited membrane permeability of DSSO[35], we cross-linked lysosomes both in an intact (IT) and disrupted (DR) state, and determined optimal reaction conditions by silver staining and western blotting (Supplementary Fig. 1b, c). Subsequently, we enriched lysosomes from 384 plates of HEK293 cells across six biological replicates, and assessed lysosomal intactness, recovery, and enrichment (Fig. 1b, c). Using a non-cross-linked fraction of each sample, we acquired an LC-MS/MS reference dataset. In total, we identified 4181 proteins, of which 474 were assigned the term "lysosome" based on GO terms and UniProt classifiers in >3 runs, indicating an excellent performance of lysosome enrichment[36] (Supplementary Data 1 and 2). To assess the quantitative distribution of lysosomal proteins in our samples, we utilized these data to estimate absolute protein abundances by intensity-based absolute quantification (iBAQ)[37]. This revealed a threefold overrepresentation of lysosomal protein abundance relative to the whole dataset (Fig. 1d and Supplementary Data 2).

We cross-linked lysosome-enriched fractions in both the IT and DR state (three replicates each), followed by their proteolytic digestion, strong cation-exchange (SCX) peptide fractionation, and analysis by LC-MS/MS (Fig. 1a). Analysis of the XL-LC-MS/MS dataset with XlinkX[38] resulted in the assignment of 6580 cross-link spectral matches, originating from 4294 cross-linked peptides at a false discovery rate (FDR) of 5 % (Supplementary Fig. 1d–f and Supplementary Data 3). Out of the 2467 unique residue-to-residue cross-links, 524 identifications (270 intralinks between different residues of the same protein and 254 inter-links between two different proteins) originated from 111 proteins assigned to the lysosomal compartment, while the remaining cross-links were identified for proteins which are currently not connected to lysosomes, presenting potentially novel interaction partners (Fig. 1e, Supplementary Fig. 1g, h and Supplementary Data 3). Interestingly, only 25% of cross-links were found both in the IT and the DR state, with the latter contributing a larger fraction to the dataset, further demonstrating the limited membrane permeability of DSSO (Fig. 1f). A similar distribution was observed for the cross-links identified for non-lysosomal proteins contained in the dataset (Supplementary Fig. 1d), and we observed a better reproducibility in cross-link identification for the DR state (Supplementary Fig. 2f). Strikingly, while cross-link spectral matches (CSMs) of cytosolic proteins were almost equally distributed between both conditions, 91% of the dataset's CSMs assigned to lysosomal luminal proteins were identified in samples cross-linked in the DR state (Fig. 1g).

As lysosomal proteins are expressed at a dynamic abundance range encompassing three orders of magnitude[36], we further correlated CSMs, peptide spectral matches (PSMs), and iBAQ values. Even though higher abundant proteins tended to yield more CSMs, we did

https://doi.org/10.1038/s41467-022-33951-0

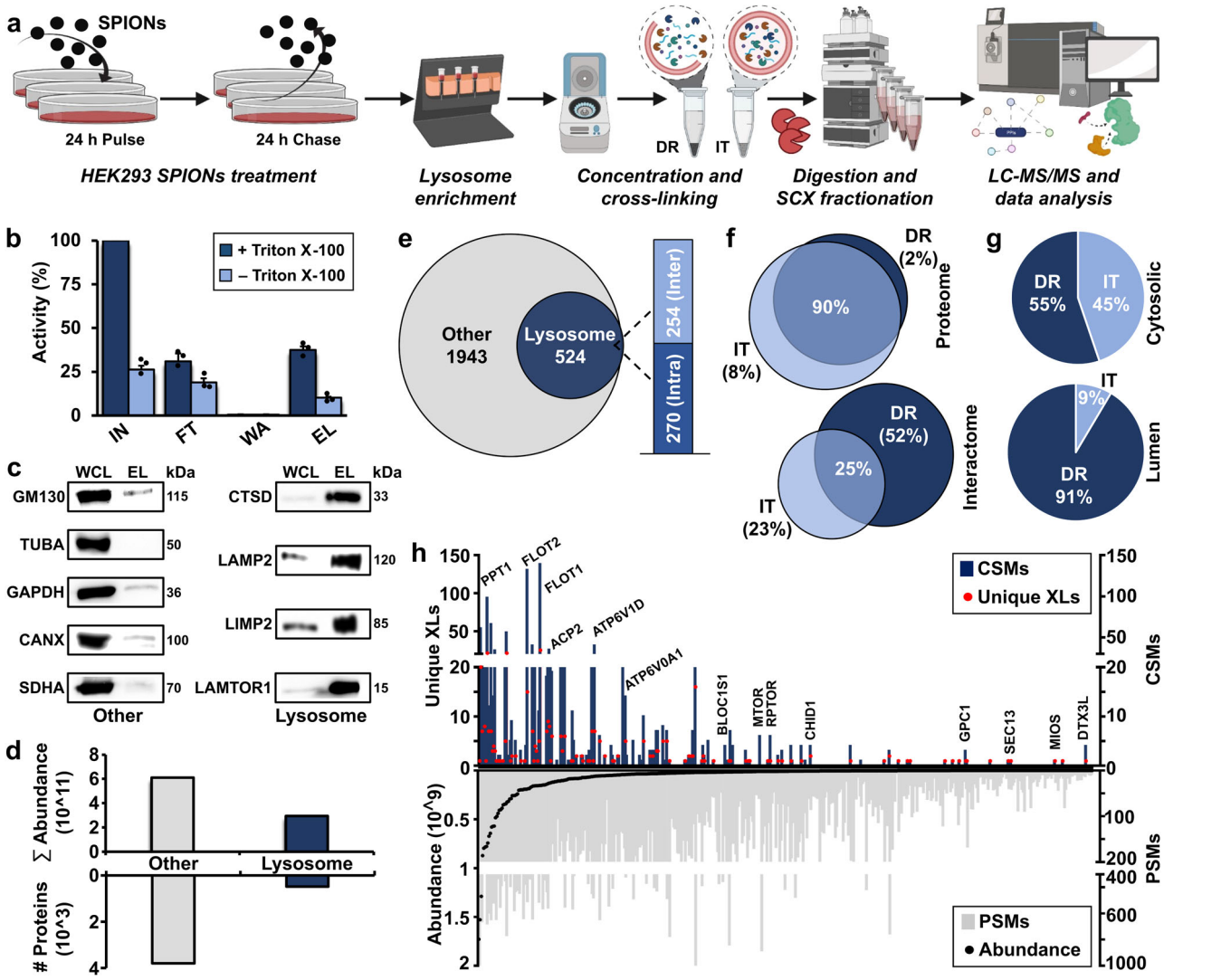


Fig. 1 | Cross-linking mass spectrometry analysis of lysosome-enriched fractions. a Experimental workflow for the XL-LC-MS/MS analysis of lysosome-enriched fractions. Created with BioRender.com. **b** Normalized β-hexosaminidase activities for individual fractions from lysosome enrichment by SPIONs. Data are presented as mean values + SD ($n = 3$, biologically independent samples over three independent experiments). **c** Western blot analysis of lysosome-enriched fractions for contamination by other organelles ($n = 2$). Lysosome: lysosomal proteins (CTSD, LAMP2, LIMP2, and LAMTOR1). Other: Golgi apparatus (GM130), cytoskeleton (TUBA), cytosol (GAPDH), endoplasmic reticulum (CANX), and mitochondria (SDHA). **d** Summed iBAQ abundances for proteins identified in lysosome-enriched fractions in ≥3 replicates. **e** Classification of unique cross-linked residue pairs. **f** Proteins detected in non-cross-linked lysosome-enriched fractions (proteome), and unique lysosomal cross-linked residue pairs (interactome) for DR and IT samples. **g** Localization of CSMs for 68 lysosomal proteins cross-linked in the DR and IT state. Cytosolic: proteins located at the cytosolic face of the lysosomal membrane; Lumen: lysosomal luminal proteins. **h** Correlation of cross-link identification and protein abundance for lysosomal proteins. CSMs and PSMs represent summed values of the analysis ($n = 6$, biologically independent samples over six independent experiments (3× DR and 3× IT)). SPIONs superparamagnetic iron oxide nanoparticles, DR disrupted, IT intact, SCX strong cation-exchange, IN input, FT flow through, W wash, EL eluate, WCL whole-cell lysate, iBAQ intensity-based absolute quantification, XLs cross-links, CSMs cross-link spectral matches, PSMs peptide spectral matches. Source data are provided as a Source Data file.

not identify a strong correlation between cross-link identification and protein abundance, showing that our dataset also covered proteins of low expression levels (Fig. 1h, Supplementary Fig. 1i and Supplementary Data 2 and 3). When we compared the average iBAQ abundance for proteins involved in intra- and inter-links, we observed a tendency towards identifying more intralinks in higher abundant proteins, which was less pronounced for lysosomal proteins (Supplementary Fig. 1j, k).

Finally, we investigated the distribution of CSMs across a shortlist of lysosomal proteins/complexes[21]. Most categories showed an equal distribution between DR and IT samples, with the exception of proteins involved in lysosomal substrate degradation (93%), heat-shock proteins (70%), and annexins (75%), for which more cross-links were annotated in the DR sample (Supplementary Fig. 1l).

Investigation of lysosomal protein–protein interactions

Based on all inter-links in the dataset, we constructed a network of 1008 proteins engaged in 1023 interactions, of which 254 involved lysosomal proteins (Fig. 2a, Supplementary Fig. 2a–c and Supplementary Data 3). Comparison to known interactions revealed an overlap of ~30%, confirming the validity of our dataset. While 34% of interactions of non-lysosomal proteins were included in STRING, only 26% of potential lysosomal PPIs have been reported previously (Fig. 2b). We classified lysosomal PPIs based on the interacting subcellular compartment, revealing an overrepresentation of nuclear and cytoplasmic/cytoskeletal proteins (Fig. 2c). With respect to lysosomal and lysosome-associated proteins, we identified the highest numbers of PPIs for the V-ATPase, the flotillins, mTORC1, and the syntaxins (Fig. 2d, e and Supplementary Fig. 2d, e).

https://doi.org/10.1038/s41467-022-33951-0

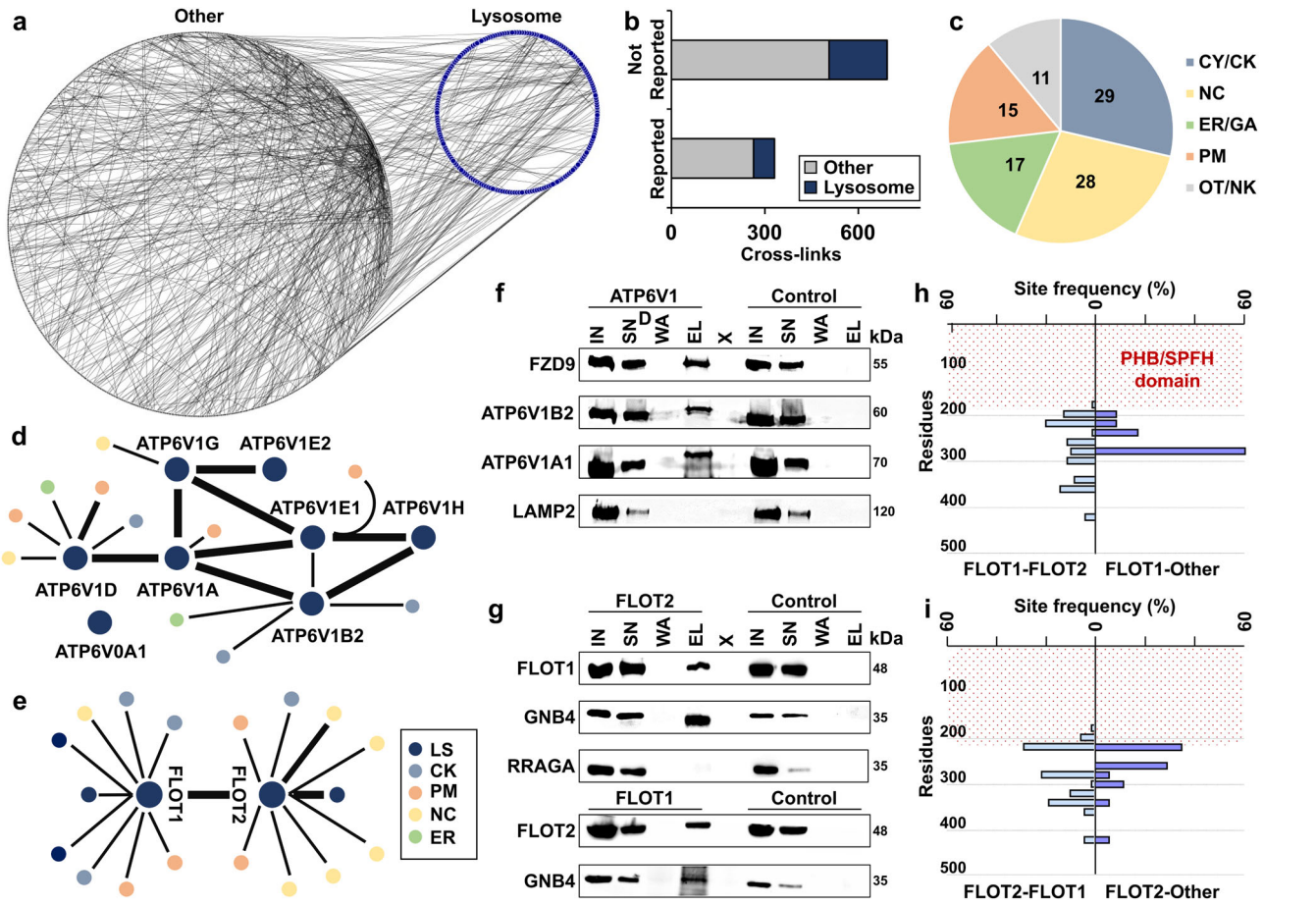


Fig. 2 | The human lysosomal interactome. a PPIs identified in the XL-LC-MS/MS dataset. Lysosomal proteins (blue dots), non-lysosomal proteins (gray dots), and PPIs (gray lines) are indicated. **b** Matching of PPIs to the STRING database. **c** Numbers of proteins from distinct subcellular localizations interacting with lysosomal proteins. **d**, **e** Interaction networks of the V-ATPase (**d**) and the flotillin (**e**) complex. **f** Co-IP of ATP6V1D and FZD9 ($n = 2$). ATP6V1B2 and ATP6V1A1 are known members of the V-ATPase complex; LAMP2 is a lysosomal membrane protein. Control: empty beads. **g** Co-IPs of FLOT1, FLOT2, and GNB4 ($n = 2$). RRAGA is a lysosomal membrane-associated protein. Control: empty beads. **h**, **i** Site frequency distribution for identified FLOT1 (**h**) and FLOT2 (**i**) cross-links. Site frequency represents the percentage of cross-links detected in bins of 20 residues each. The region indicated by red dots represents the PHB domain. CY cytoplasm, CK cytoskeleton, NC nucleus, ER endoplasmic reticulum, GA Golgi apparatus, PM plasma membrane, LS lysosome, OT others, NK not known, IN input, SN supernatant, W wash, EL eluate, X empty lane. Source data are provided as a Source Data file.

For the V-ATPase complex, we detected most PPIs for the D subunit of its soluble V1 part (ATP6V1D). This may be related to the capability of V1 to dissociate from the lysosomal membrane-embedded V0 part, which was shown to be involved in the regulation of V-ATPase activity[39], exposing the D subunit to interactions. The cross-link of ATP6V1D with frizzled 9 (FZD9), a member of the WNT signaling pathway, sparked our interest. FZD9 was shown to be sorted to late endosomes/lysosomes after its internalization by endocytosis[40], and the ATP6AP2 subunit of the V-ATPase was reported to interact with FZD8[41]. We, therefore, investigated the validity of this PPI by co-IP, confirming both the observed interaction of ATP6V1D with FZD9, as well as interaction with other subunits of the V-ATPase complex (Fig. 2f).

We further investigated the interaction of FLOT1, FLOT2, and GNB4, for which we identified an inter-link with FLOT2. We were able to co-IP FLOT1 and FLOT2, which are known to form heterooligomers[42], as well as GNB4 with its direct interactor FLOT2 as well as with FLOT1, indicating binding of GNB4 to FLOT1/FLOT2 heteromeric assemblies (Fig. 2g).

As we observed 18 different PPIs for FLOT1 and FLOT2 in total, we further investigated their distribution across both proteins. While FLOT1/FLOT2 inter-links were detected across most of the regions predicted to form a helical structure (amino acids 193–365 and 213–362 for FLOT1 and FLOT2, respectively)[43], the interaction with other proteins occurred almost exclusively in confined sections of <100 amino acids (Fig. 2h, i). The fact that we found FLOT1/FLOT2 inter-links across the whole region of the proteins shows that no sequence-dependent bias towards the cross-link reaction or detectability exists. Therefore, the presence of the majority of PPIs with proteins other than FLOT1/FLOT2 at a distinct section of the proteins suggests the presence of FLOT1/FLOT2 interaction hotspots.

Structural integration of cross-linker distance constraints suggests a tetrameric assembly of PPT1 in vivo

Cross-links between different amino acids provide distance information (the length of DSSO links is ~35 Å) that is helpful to validate (or infer) protein structures[44]. Initially, we used TopoLink[45] to match 161 unique cross-links to the resolved structures of 34 lysosomal and lysosome-associated proteins, confirming the validity of our dataset. The remaining 64 cross-links assigned to lysosomal proteins could not be integrated, as the respective regions have not been resolved yet. We matched these cross-links either to homology models based on available PDB structures from other organisms using SWISS-MODEL[46], or to the AlphaFold Protein Structure Database[47] (Fig. 3a, Supplementary Fig. 3 and Supplementary Data 4).

https://doi.org/10.1038/s41467-022-33951-0

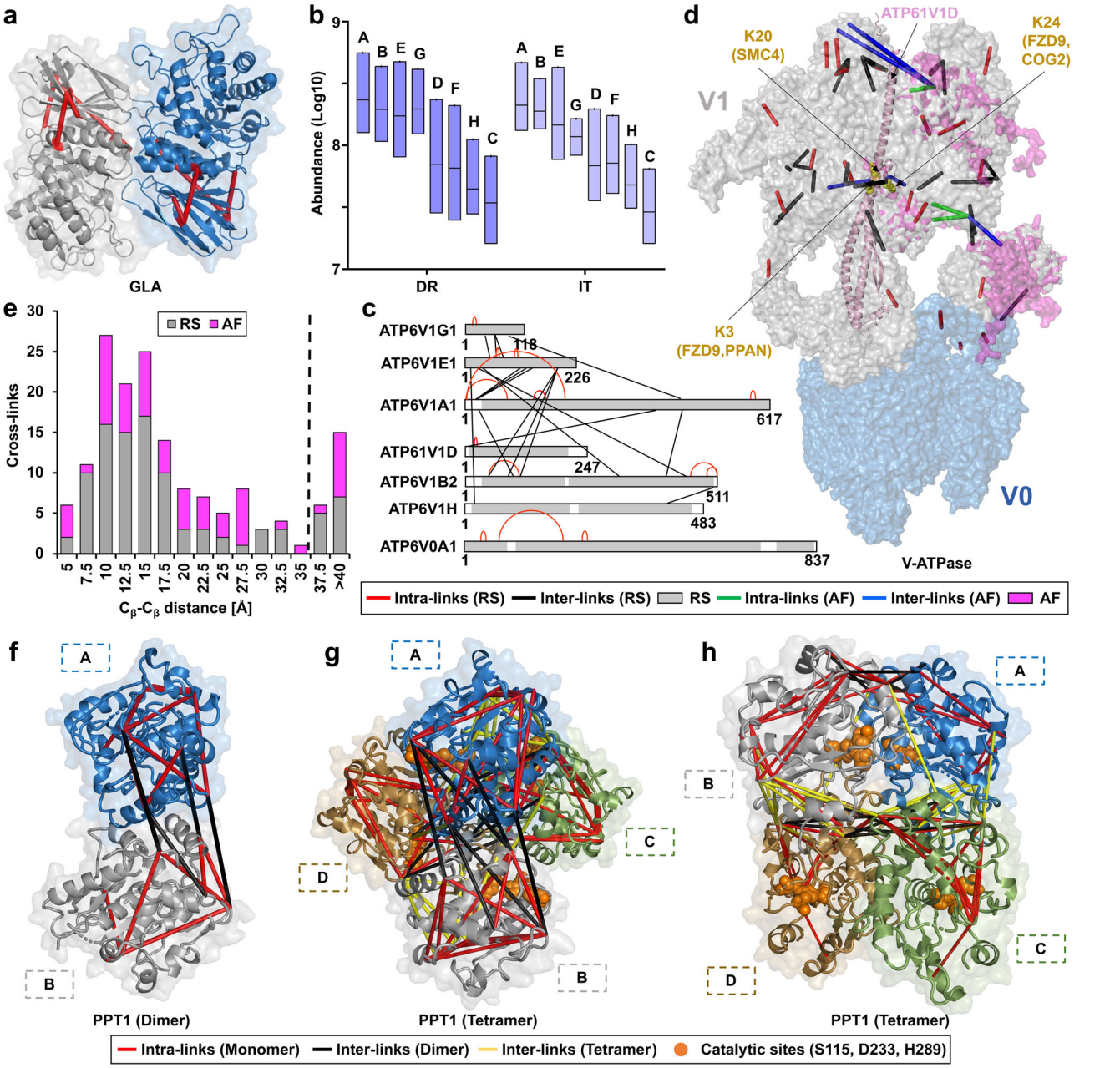


Fig. 3 | Cross-link-based structural refinement of the V-ATPase and PPT1 complex. a Matching of cross-links to the crystal structure of lysosomal alpha-galactosidase A (GLA). **b** Average iBAQ abundances for individual V-ATPase subunits based on the analysis of lysosome-enriched fractions in a non-cross-linked state. Each box depicts the interquartile range (IQR, the range between the 25th and 75th percentile, median = line). **c** Cross-links identified for individual V-ATPase subunits. Structurally resolved regions are colored gray, and unresolved regions are white. **d** Refined structure of the V-ATPase complex. Identified cross-links were integrated into the V-ATPase cryo-EM structure[7]. Missing regions were supplemented with predicted structures from AlphaFold2 based on the identified cross-links. **e** Distances for mapped cross-links to crystal structures and AlphaFold2 models for lysosomal proteins, bin size: 2.5 Å. **f** Reported homodimeric human PPT1 structure (PDB identifier 3GRO). **g**, **h** Tetrameric PPT1 model representing the most favorable energetic state and fulfilling DSSO's distance constraints for all 18 cross-links. A, B, C, and D indicate individual subunits. DR disrupted, IT intact, RS resolved structure, AF AlphaFold2. Source data are provided as a Source Data file.

The complex for which we identified the highest number of cross-links was the V-ATPase, a 35-mer assembly of 17 unique proteins[7], of which 13 were covered in our dataset. While the soluble V1 part (eight different proteins) yielded 26 cross-links, only three were identified for the membrane-embedded V0 section (seven different proteins). After confirmation of the V1 part's correct stoichiometry through the individual subunit's iBAQ values in the non-cross-linked dataset ($A_3B_3E_3G_3D_1F_1C_1H_1$) (Fig. 3b and Supplementary Data 2), we mapped the identified cross-links to a recently published structure determined by cryo-EM[7]. Out of the 29 unique cross-links identified, 21 could readily be integrated into the published structure, while 8 originated from regions that were so far structurally not resolved (Fig. 3c). For these sections, we integrated the predicted full-length protein models by AlphaFold2, aligning them with the published V-ATPase subunit structures in the complex, based on the identified cross-links (Fig. 3d). In this combined model, we cover ~95% of the V-ATPase sequence, and 90% of cross-links fulfill DSSO's distance constraints. The remaining three cross-links were inter-links of the A and E subunit of the V1 part, which are known to undergo conformational changes during ATP hydrolysis[7].

Regarding mTORC1, we identified most cross-links for proteins related to Ragulator, a lysosomal membrane-associated complex that is crucial for mTORC1 activity[3]. For the N-terminal region of LAMTOR1, we identified three cross-links of the same lysine residue with different amino acids of LAMTOR3. Two of them violated DSSO's distance constraints relative to the crystal structure obtained in the presence of the Ragulator-associated RAG GTPases[48], indicating cross-linking of an alternative state, possibly representing Ragulator in the absence of RAG GTPases (Supplementary Fig. 3j). We further identified three intralinks from RRAGA that originated from the same lysine residue at its C-terminus, cross-linked to three different amino acids (Supplementary Fig. 3k). This could be related to high flexibility of this region, which is in accordance with the fact that it could not be covered in a previous crystallization study[49].

Of the 161 lysosomal cross-links identified, 21 exceeded DSSO's distance constraint (Fig. 3e and Supplementary Data 4). Surprisingly, nine of them originated from PPT1, a member of the palmitoyl protein thioesterase family. The current PPT1 crystal structure (PDB identifier 3GRO) resembles a homodimeric assembly of two identical subunits (Fig. 3f). Strikingly, while all eleven PPT1 intralinks fulfilled the distance constraints of this structure, the inter-links exceeded them, indicating the possibility of an alternative oligomerization state. This possibility is in line with a previous study detecting the maximal enzymatic activity of PPT1 at a complex size of >100 kDa[50]. We, therefore, used HADDOCK2.2 to perform restraint-based docking for the monomeric subunits of PPT1, extracted from 3GRO. This resulted in the prediction of a tetrameric PPT1 model, which fulfills the distance constraints for all 18 PPT1 cross-links (Fig. 3g, h).

Proposal of a heterodimeric FLOT1/FLOT2 model featuring extended alpha-helical domains

We identified the highest number of cross-links for the two members of the flotillin family, FLOT1 and FLOT2, which were also over-represented in the proteomic dataset of the lysosome-enriched fraction (Fig. 1h and Supplementary Data 2). FLOT1 and FLOT2 are lipid raft-associated proteins which are present in nearly every type of vertebrate cell, and are highly conserved among organisms[43]. We confirmed their co-enrichment with lysosomes by western blotting (Fig. 4a) and colocalization by immunostaining (Fig. 5a and Supplementary Fig. 5a), which is in agreement with previous EM studies detecting FLOT1 at the lysosomal surface[51,52]. It is well-established, that FLOT1 and FLOT2 form heterodimers with a 1:1 stoichiometry[53], but only partial structural information is available from NMR analyses of the N-terminal region of mouse Flot2 (PDB identifier 1WIN). Other experimental data on FLOT structures are not available, as purification of the full-length proteins is problematic[54].

Our dataset contains 29 unique cross-links for FLOT1 and FLOT2, including 11 intra- and 22 inter-links (Supplementary Data 3). In the first step, we predicted individual secondary structures for both FLOT1 and FLOT2 using PSIPRED (Fig. 4b)[55]. For the protein N-termini, this resulted in a cluster of beta-sheets, which is in accordance with their sequence homology to the stomatin/PHB/flotillin/HflK/C (SPFH) domains and the Flot2 NMR structure. The middle section features an extended α-helical region that is interrupted once in the case of FLOT2, while the C-terminus forms one beta-sheet and several short helices for both proteins. We further calculated coiled-coil probabilities for the helical regions using PCOILS[56], with window sizes of 14, 21, and 28 amino acids (Fig. 4b). Dependent on the region of both proteins, windows of 14 and 28 amino acids delivered the best results, with slightly different patterns for FLOT1 and FLOT2. Matching our data to full-length structural models from the AlphaFold Protein Structure Database showed excellent agreement, and all identified cross-links confirmed the predicted structures.

Subsequently, we built heterodimeric models of FLOT1, FLOT2, and their interaction using ColabFold[57], a variant of AlphaFold2. For additional post-modeling validation with experimental data, we mapped the cross-linking restraints on the five resulting models. All generated models satisfied all experimentally identified restraints based on the cross-links, increasing our confidence in these models. The resulting models feature closely aligned highly similar structures for both flotillins. In particular, they feature a globular N-terminal region consisting of SPFH domains (residues 1–162 of FLOT1 and FLOT2), a central linear α-helical region (residues 163–341 of FLOT1 and 163–350 of FLOT2), and a C-terminal α-helical coiled-coil structure (residues 342–427 of FLOT1 and 351–426 of FLOT2). The SPFH domains of both FLOT1 and FLOT2 present with antiparallel β-sheets, with six repeats each, and four partially exposed α-helices, forming an ellipsoidal-like globular domain (Fig. 4c and Supplementary Fig. 4). Based on the fact that the major helices were not interrupted, and that the C-terminus was in its most compact state, we selected model four, which was also supported by all FLOT1/FLOT2 cross-links detected (Fig. 4c and Supplementary Fig. 4c).

We highlighted several structural features in our model which are known for both flotillins (Fig. 4c). The *S*-palmitoylation and *N*-myristoylation sites, which are crucial for the membrane-association of both flotillins[58], are located at the SPFH domain's membrane interfaces. The tyrosine phosphorylation sites, which were shown to be crucial for flotillin-mediated endocytosis and FLOT1/FLOT2 interaction[59,60], are surface-exposed and in proximity to a basic motif (HQR) on the respective other flotillin. Moreover, the PDZ3 domains of both flotillins strongly co-localize, forming a combined feature, and the EA-rich motifs, which were predicted to mediate flotillin oligomerization[43,61], are distributed along the length of the central α-helical region[62]. Interestingly, the putative interaction hotspots (Fig. 2h, i) locate around the major bend observed in this structure, containing one tyrosine residue (Y238/Y241) in its center. These residues are located in highly conserved sequence motifs (A-X-A-X-L-A-pY-X-L-Q with X: [D/Q or E/Q]), possibly presenting a regulatory switch for FLOT1/FLOT2 PPIs by tyrosine phosphorylation. To further confirm the validity of our model and the presence of the predicted interaction hotspot, we generated deletion mutants lacking this entire region (amino acid residues 200–300) for both FLOT1 and FLOT2. We co-transfected both constructs and immunoprecipitated FLOT1/FLOT2 followed by the detection of GNB4 (Fig. 2g). Deletion of the interaction hotspot prevented binding of GNB4, indicating that this region is crucial for its interaction with FLOT1/FLOT2 (Fig. 4d).

Flotillins assemble in similar higher-order structures at lysosomes and endosomes

While FLOT1 was detected previously at the cytosolic face of lysosomes by EM[51,52], assemblies of both FLOT1 and FLOT2 were only shown for the plasma membrane or early endosomes, where they play a role in clathrin-independent endocytosis[63,64]. In line with these findings, we detected FLOT1 and FLOT2 to localize at the plasma membrane and to partially co-localize with lysosomes of HeLa and HEK293 cells (Fig. 5a and Supplementary Fig. 5a). In these analyses, we also observed numerous FLOT1/FLOT2 puncta, which did not co-localize with lysosomes, presenting putative FLOT1/FLOT2-positive endosomes.

In order to elucidate if the structural assembly of FLOT1 and FLOT2 differs between early endosomes and lysosomes, we performed cross-linking of early endosome-enriched fractions. To increase the number of FLOT1-/FLOT2-positive early endosomes, we co-transfected cells with FLAG-tagged FLOT1 and FLOT2, as their overexpression was shown to increase the number of FLOT-positive endosomes[42,53]. After confirmation of correct FLOT distribution utilizing FLAG-, GFP-, and mCherry-tagged versions of FLOT1 and FLOT2 (Supplementary Figs. 5b and 6a), we established SPIONs pulse-chase conditions for the enrichment of early endosomes (Fig. 5b). Analysis of the early endosome-enriched fraction by western blot and LC-MS/MS verified

https://doi.org/10.1038/s41467-022-33951-0

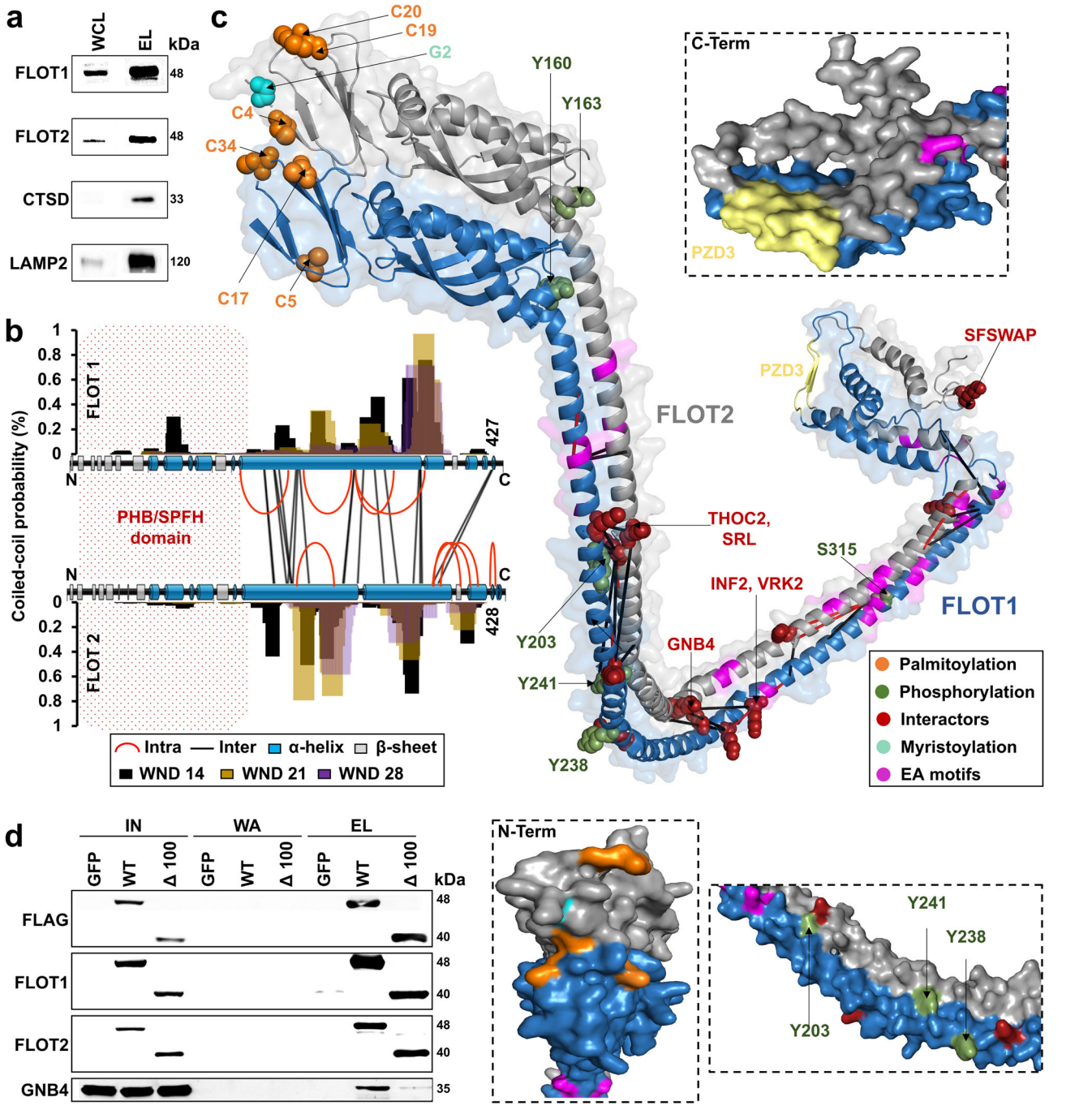


Fig. 4 | Proposed model for a heterodimeric FLOT1-FLOT2 complex. a Western blot analysis of lysosome-enriched fractions ($n = 3$). CTSD is a lysosomal luminal and LAMP2 a lysosomal transmembrane protein. **b** Identified cross-links, predicted secondary structures (PSIPRED), and coiled-coil probabilities (PCOILS) for FLOT1 and FLOT2. **c** Heterodimeric model of FLOT1-FLOT2 interaction (ColabFold). The model satisfies the distance constraints of all cross-links. **d** FLAG-IP from HEK293 cells transfected with either full-length FLAG-FLOT1 and FLAG-FLOT2 or versions lacking amino acid residues 200–300 (Δ100) ($n = 2$). WCL whole-cell lysate, EL eluate, PHB prohibitin homology domain, SPFH stomatin/PHB/flotillin/HflK/C domain, WND window, E glutamic acid, A alanine, PZD3 postsynaptic density protein-95/discs large/zonula occludens-1, IN input, WA wash, WT wild-type, Δ 100 mutant with a deletion of 100 amino acids (residues 200–300 of FLOT1/FLOT2). Source data are provided as a Source Data file.

the presence of marker proteins such as EEA1, the clathrin chains CLTA, CLTB, and CLTC, as well as the RAB-GTPases RAB5, RAB11, and RAB14. Furthermore, we detected FLOT1 and FLOT2, while markers for other organelles were depleted (Fig. 5c and Supplementary Data 5). We then performed enrichment of early endosomes from FLOT1/-FLOT2-FLAG overexpressing HEK293 cells, established their cross-linking, performed MS sample preparation followed by SCX fractionation, and investigated them by LC-MS/MS (Fig. 5d and Supplementary Fig. 6b–f). Importantly, western blot analysis of FLOT1/FLOT2 aggregation in response to different amounts of DSSO showed that the concentration we used for cross-linking of lysosome-enriched fractions does not result in flotillin over-cross-linking (Supplementary Fig. 5c, d), further supporting the validity of these data. In total, we identified 1081 cross-links from 414 unique proteins (Fig. 5e, Supplementary Fig. 6f and Supplementary Data 6). This early endosome dataset contains 15 unique cross-links for FLOT1 and FLOT2, which all matched to our

https://doi.org/10.1038/s41467-022-33951-0

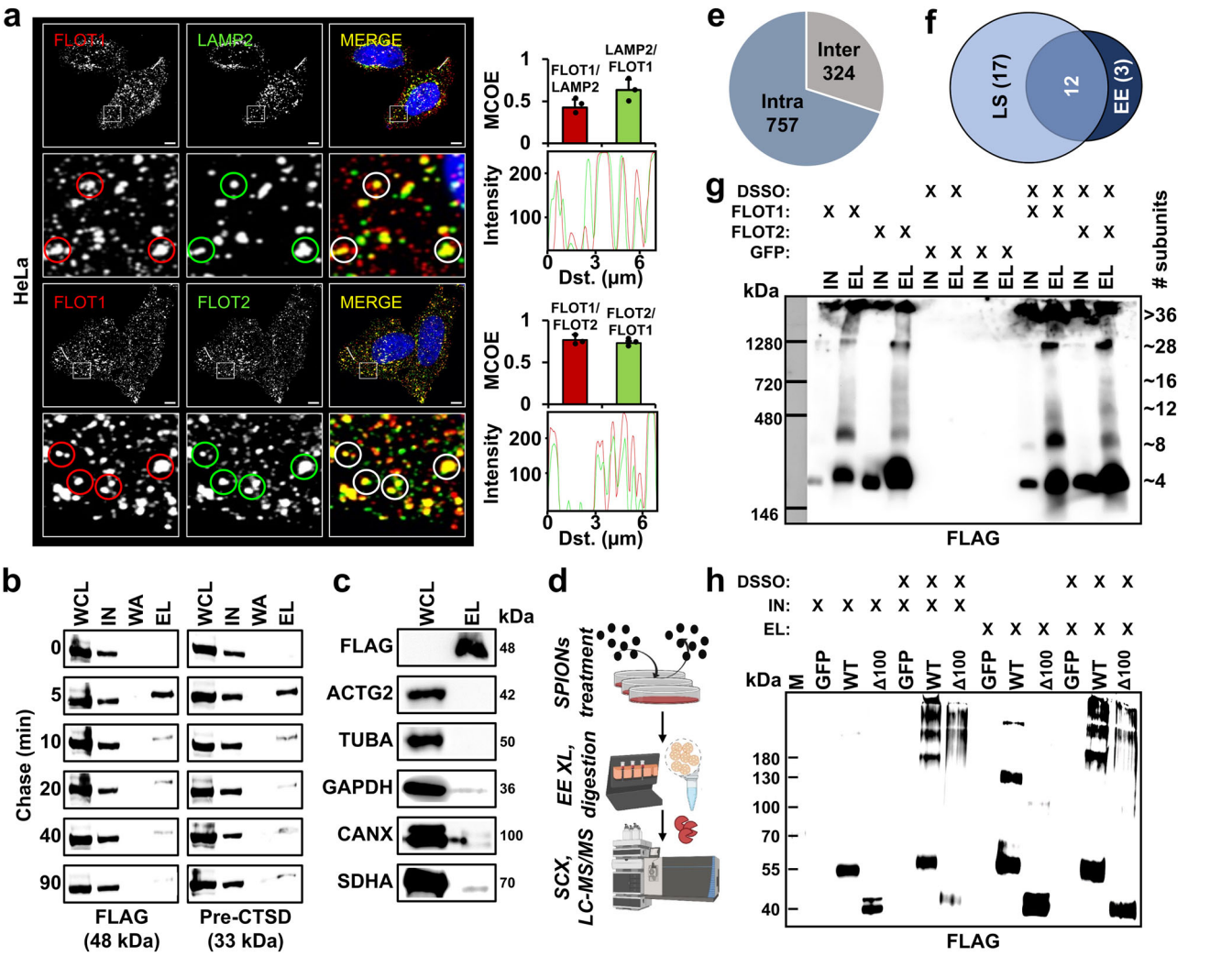


Fig. 5 | Investigation of higher-order flotillin assemblies in lysosome- and early endosome-enriched fractions. a HeLa cells were stained with antibodies for FLOT1, FLOT2, and LAMP2 followed by microscopy imaging ($n = 3$). Lower panels display a zoom-in of the regions indicated in the full picture. Mander's coefficients show the average degree of overlap of one protein population relative to another. Data are presented as mean values + SD ($n = 3$, cells examined over one experiment). Profile plots indicate the degree of colocalization for individual vesicles. **b** Analysis of FLOT1/FLOT2 in early endosomes ($n = 1$). Following pulsed SPIONs treatment, endosomes were enriched at the indicated chase times. Pre-CTSD serves as a marker for early endosomes. **c** Western blot analysis of early endosome-enriched fractions with marker proteins for different subcellular compartments ($n = 1$): Cytosol (ACTG2 and GAPDH), cytoskeleton (TUBA), endoplasmic reticulum (CANX), and mitochondria (SDHA). **d** Experimental workflow for early endosome enrichment and XL-LC-MS/MS analysis. Created with BioRender.com. **e** Overview of the cross-linking dataset obtained from early endosome-enriched fractions. **f** Overlap of unique FLOT1/FLOT2 cross-links for XL-LC-MS/MS analyses of early endosome- and lysosome-enriched fractions. **g** Western blot analysis of BN-PAGE-separated FLAG-IP eluates with/without cross-linking by DSSO ($n = 5$). IN/EL refers to individual FLAG-IP fractions. **h** Western blot analysis of SDS-PAGE separated FLAG-IP ELs with/without DSSO cross-linking for WT and FLOT1/FLOT2 residue 200–300 deletion mutants (Δ100) ($n = 3$). Scale bar = 5 µM; WCL whole cell lysate, IN input, WA wash, EL eluate, MCOE Mander's coefficient, Dst. distance, LS lysosomes, EE early endosomes, XL cross-linking, SPIONs superparamagnetic iron oxide nanoparticles, M marker, WT wild type. Source data are provided as a Source Data file.

predicted FLOT1-FLOT2 heterodimer within DSSO's distance constraints, and from which 80% overlapped with the cross-linking dataset from lysosome-enriched fractions (Fig. 5f). These data indicate that early endosome- and lysosome-localized flotillins assemble in a similar way.

It has been shown previously, that flotillins form higher-order assemblies[61,63]. We, therefore, performed blue native (BN)-PAGE experiments to investigate the size distribution of FLOT1/FLOT2 structures in a native and a cross-linked state (Fig. 5g). These analyses revealed that a significant amount of both FLOT1 and FLOT2 migrates at a range corresponding to a tetrameric assembly, while smaller fractions migrated at sizes consistent with higher-order structures exceeding 1 MDa. The cross-linking samples, which were generated with the same reaction conditions as the early endosome and lysosome experiments, presented with the same complex sizes, indicating that the cross-link data represent the native state.

Based on the combined 32 unique cross-links from the lysosome- and early endosome cross-linking datasets, we further attempted to model the tetrameric structure using either restraint-based docking or AlphaFold-Multimer. This did, however, not result in a convincing structural model. We further investigated in how far the deletion of the putative interaction hotspot influences the assembly of higher-order structures by co-transfection of both FLOT1/FLOT2 deletion constructs followed by IP, cross-linking and SDS-PAGE. Western blot analyses of these samples indicate that the absence of the interaction hotspot also influences the formation of higher-order assemblies migrating at sizes >180 kDa (Fig. 5h).

https://doi.org/10.1038/s41467-022-33951-0

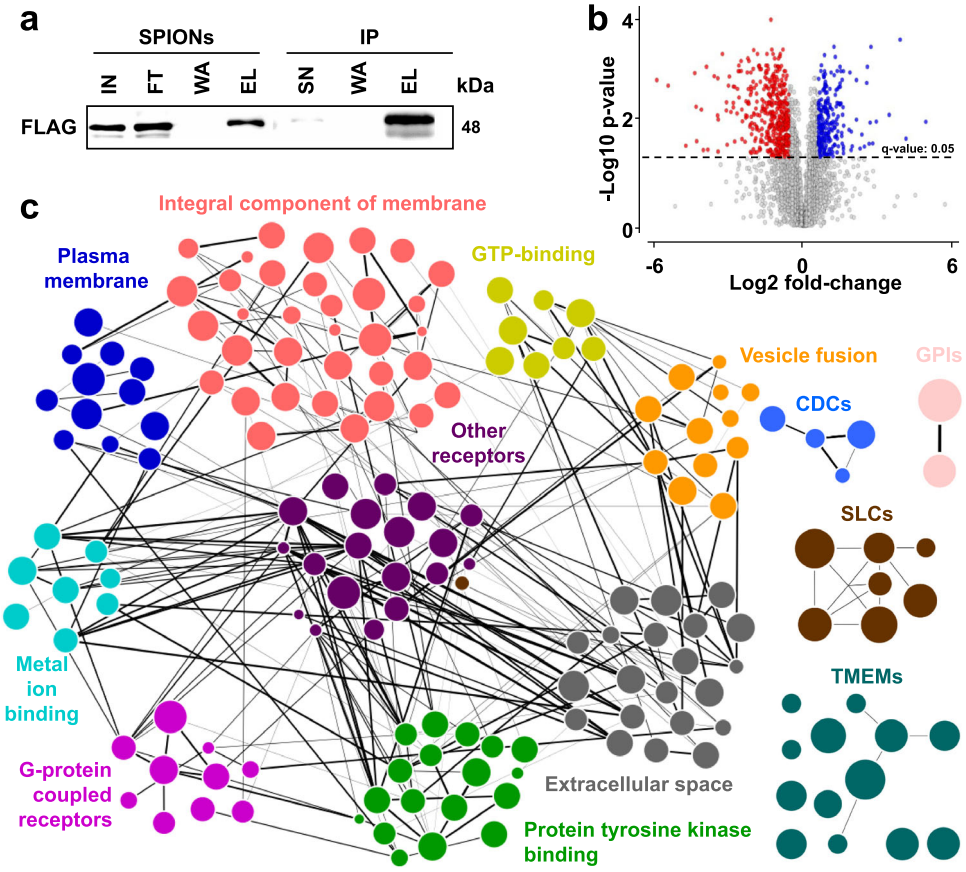


Fig. 6 | Identification of potential FLOT1-/FLOT2-positive early endosome cargo. a Western blot analysis of FLAG-FLOT1/FLOT2 in early endosome-enriched fractions (SPIONs) and intact endosome-IP fractions ($n = 1$). **b** Data-independent acquisition (DIA)-based protein abundance fold-change ratios of SPIONs/SPIONs +IP fractions ($n = 3$, biologically independent samples over three independent experiments). Significant differentially regulated proteins are indicated (cut-offs: q-value: <0.05, fold change: >1.5). **c** STRING-based PPI analysis of selected protein categories overrepresented in the SPIONs+IP fraction. Node size corresponds to the protein abundance and line thickness to the PPI confidence score. SPIONs superparamagnetic iron oxide nanoparticles, IP immunoprecipitation, IN input, SN supernatant, FT flow through, WA wash, EL eluate, PPI protein–protein interaction. Source data are provided as a Source Data file.

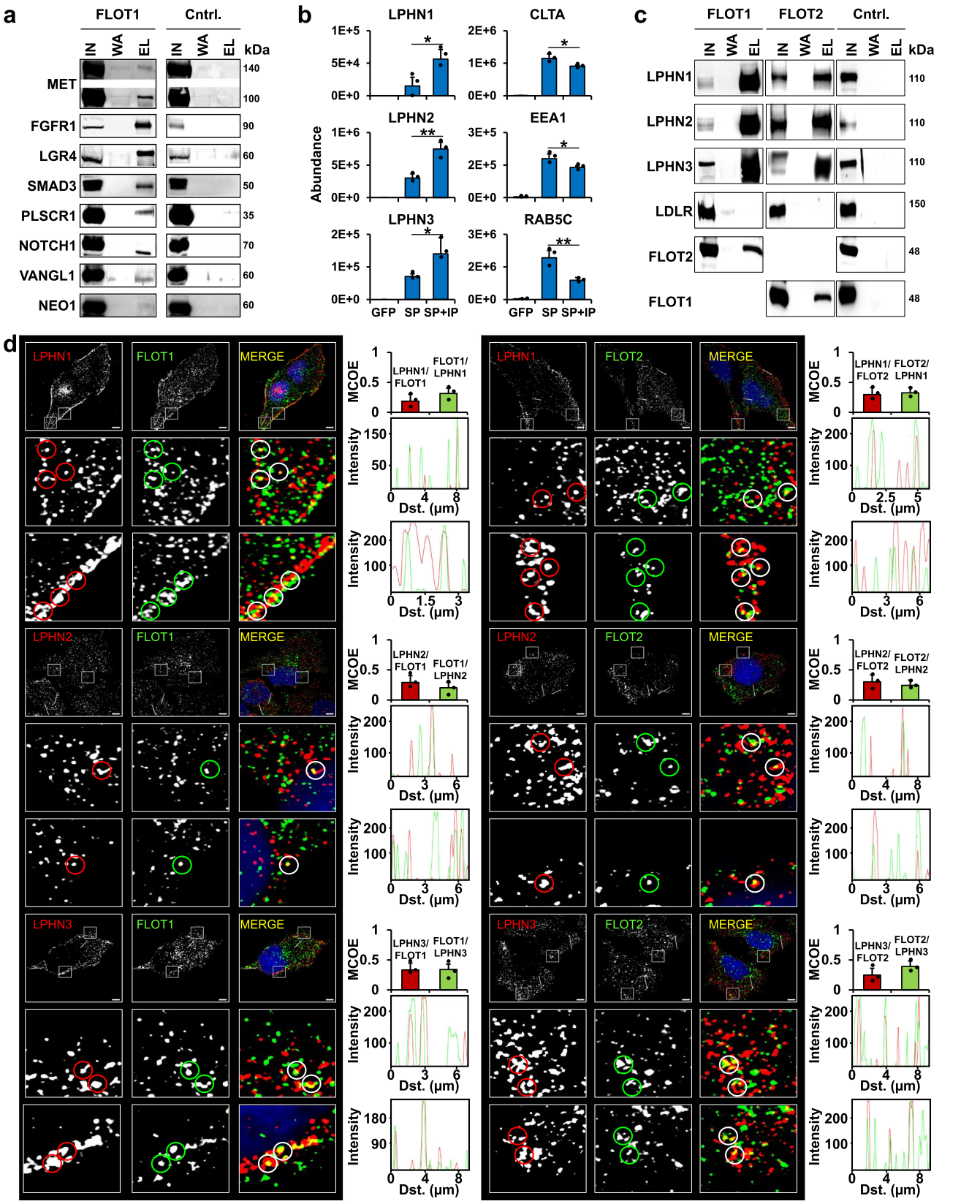
Analysis of flotillin-endosome cargo reveals an over-representation of membrane proteins and receptors

Flotillins have been proposed as defining structural components of an endocytic pathway independent of clathrin and caveolin[53,65], and were shown to co-localize with early endosomes[64]. In agreement with these findings, we detected their co-enrichment with endosomes (Fig. 5b, c). To investigate the putative cargo of FLOT1-/FLOT2-positive early endosomes, we combined SPIONs enrichment of early endosomes from HEK293 cells overexpressing FLAG-tagged FLOT1/FLOT2 and immunoprecipitation of FLAG-positive intact vesicles from these samples (Fig. 6a). Subsequently, we analyzed the resulting fractions by data-independent acquisition (DIA) LC-MS/MS and performed label-free quantification (Supplementary Fig. 7a–d). Based on 5089 proteins quantified across all conditions, we were able to define protein populations that were enriched/depleted in FLOT1-/FLOT2-positive endosomes relative to the total cellular pool of early endosomes (Fig. 6b and Supplementary Data 7). Importantly, the FLOT1-/FLOT2-depleted population of early endosomes contained the clathrin chains CLTA and CLTC, EEA1, and the endosome-related GTPases RAB5 and RAB11, confirming a separation of clathrin- and flotillin-containing early endosomes. We confirmed these findings by co-immunostaining of EEA1 with endogenous FLOT1/FLOT2 in HeLa and HEK293 cells as well as in cells and transfected with FLOT1-GFP, showing that FLOT1-/FLOT2-positive vesicles present a distinct population from EEA1-positive endosomes (Supplementary Figs. 7e, f and 8).

The enriched population, which presents potential cargo of FLOT1-/FLOT2-positive early endosomes, consists of 328 proteins (Supplementary Data 7). Of these, 189 were also identified in the lysosome-enriched fractions (Supplementary Data 2), possibly indicating their lysosomal destination. GO-enrichment analysis revealed a significant overrepresentation of several categories related to organelle and membrane proteins (Supplementary Fig. 7g), which is in agreement with a potential role of flotillins in the endocytosis and vesicular transport of plasma membrane proteins, demonstrated previously e.g., for NPC1L1[66,67]. We subsequently performed STRING analyses to sub-classify potential cargo proteins (Fig. 6c). This further revealed enrichment of different receptors types and members of the transmembrane protein (TMEM) family and the solute carrier (SLC) family of transporters (Supplementary Data 7).

For a subset of putative FLOT-positive endosome cargo proteins, we performed FLOT1 co-IPs to investigate a potential interaction. In total, we were able to IP eleven proteins with FLOT1 (Fig. 7a, c), encompassing eight receptors, two membrane proteins (PLSCR1 and VANGL1), and one protein interacting with a receptor (SMAD3). Among these, one group of receptors, the latrophilins, sparked our interest, as we detected all three members of this G protein-coupled receptor subfamily (LPHN1, LPHN2, and LPHN3) to be significantly enriched in FLOT1-/FLOT2-positive endosomes (Fig. 7b). We were able to co-IP all latrophilins with both FLOT1 and FLOT2, implicating their interaction

https://doi.org/10.1038/s41467-022-33951-0



with heterooligomeric FLOT structures or both monomers (Fig. 7c). We furthermore confirmed their colocalization to FLOT1-/FLOT2-positive endosomes by immunofluorescence in HeLa and HEK293 cells (Fig. 7d and Supplementary Fig. 9), providing additional evidence that the latrophilins present a potential cargo for flotillin-mediated endocytosis.

Discussion

In this study, we present a XL-LC-MS/MS analysis of lysosome-enriched fractions. In comparison to whole-cell analyses, we detected significantly higher numbers of cross-links for proteins reported previously to be related to the lysosome. This was especially the case for bona fide lysosomal proteins, which are localized in/at the organelle[68].

Fig. 7 | Confirmation of possible FLOT1-/FLOT2-positive early endosome cargo proteins. a Co-IP of FLOT1 with possible cargo proteins detected by DIA-LC-MS/MS analysis of FLOT1-/FLOT2-positive early endosomes ($n = 1$). **b** Average DIA protein abundance for LPHN1, LPHN2, LPHN3, CLTA, EEA1, and RAB5C. Data are presented as mean values + SD ($n = 3$, biologically independent samples over three independent experiments). Significance based on Student's unpaired two-sided t test; range: $^{*}P < 0.05$, $^{**}P < 0.01$; exact P values: LPHN1 (0.022), LPHN2 (0.004), LPHN3 (0.044), CLTA (0.047), EEA1 (0.035), RAB5C (0.008). **c** Co-IP of FLOT1 and FLOT2 with LPHN1, LPHN2, and LPHN3. LDLR serves as marker for plasma membrane and clathrin-mediated endocytosis. **d** Immunostaining of HeLa cells for FLOT1 and FLOT2 in combination with LPHN1, LPHN2, and LPHN3 ($n = 3$). Lower panels show a zoom-in for regions indicated in full-size images. Mander's coefficients were determined to assess signal overlap between individual populations. Data are presented as mean values + SD ($n = 3$, cells examined over one experiment). Profile plots indicate the degree of colocalization for individual vesicles. Scale bars = 5 μM; Cntrl. control, IN input, WA wash, EL eluate. Source data are provided as a Source Data file.

When considering this shortlist, we identified >100-fold more cross-links compared to previously published whole-cell studies[28,29]. A probable key factor in this context was our starting material, consisting of SPIONs-enriched lysosomes, as lysosomal proteins are typically of low abundance relative to the total proteome (estimated 0.2% of cellular protein mass[30]). We also identified many cross-links from proteins that are seemingly unconnected to lysosomes. These proteins present possible substrates which were degraded in the lysosomal lumen during the time-point of enrichment, members of protein complexes associated with the lysosomal membrane, or unspecifically enriched proteins binding to the magnetic columns, which could be due to the large amount of cells utilized as input for lysosome enrichment (64 plates per replicate).

With respect to individual lysosomal proteins, we did not observe a strong correlation of their abundance and the number of annotated cross-links. We identified, for example, no cross-links for the lysosome-associated membrane glycoproteins 1/2 (LAMP1 and LAMP2), which were estimated to contribute 50% of lysosomal protein mass[69], and were among the most abundant lysosomal proteins in our dataset. Also, for the lysosomal luminal hydrolase CTSD, which was detected with the third-highest iBAQ value, only two unique cross-links were detected, while we found the same number for the low-abundant protein CHID1 (~275-fold less abundant). This could be due to several reasons: for the cross-linking of intact lysosomes, the lack of membrane permeability of reactive DSSO prevents the analysis of lysosomal luminal proteins or domains (in case of membrane proteins), and LAMP1/2 only contain a short cytosolic domain (~10 amino acids). While analysis of disrupted lysosomes should allow for coverage of luminal proteins/domains, as was, e.g., the case for PPT1, modification of amino acids by glycosylation could be a reason for the lack of identified cross-links for certain proteins, as the lysosomal luminal regions of LAMP1/2, as well as CTSD, are highly glycosylated. This could be due to interference of glycosylation with the cross-linking reaction, proteolytic digestion, or its influence on peptide fragmentation. Furthermore, in order to allow for the identification of post-translationally modified cross-linked peptides by database searching, the molecular weight of each modification needs to be defined. This further prevents the identification of glycosylated peptides if no special emphasis is put on their analysis.

Based on the identified cross-links, we generated a protein interaction network containing 70% potentially novel PPIs. Compared to other approaches, cross-linking allows for identifying interacting residues between individual proteins in situ. An approach able to generate data under similar near-native conditions is proximity biotinylation, utilizing, e.g., BirA* or APEX2, which have been used in two different studies to investigate the lysosomal (surface) proteome and interactome in vivo[15,70]. In contrast to cross-linking, proximity biotinylation only allows to determine the presence of a protein within a defined radius relative to the respective fusion construct. It cannot identify, however, if two proteins are directly interacting, or which residues/domains are involved. This is exemplified by a recent study utilizing BirA* fused to eight different proteins, including LAMP1, LAMP2, and LAMP3, for investigation of the lysosomal proteome[70]. Comparison of the interaction partners, which were identified for the individual LAMPs, revealed that only 8–18% were unique, while the others were identified for at least two of them. From these data, it can be inferred that the latter proteins are in close proximity to the lysosomal membrane, but it cannot be identified with which of the LAMP proteins they directly interact. This is not only the case for these three examples but for all lysosomal proteins which may be utilized as bait. If a specific cross-link is found for a given lysosomal protein, on the other hand, the respective amino acids have to be within a distance of 35 Å, implying direct interaction. Therefore, compared to proximity biotinylation approaches, the interactome presented in this study provides a level of detail that is unprecedented for the analysis of lysosomal PPIs. Furthermore, cross-linking allows for the unbiased analysis of PPIs while proximity biotinylation requires the expression of fusion proteins, possibly inducing artefacts due to changes in abundance, confirmation, or distribution relative to the endogenous population. Proximity biotinylation requires, however, less input material, and allows for the identification of PPIs irrespective of the presence of lysine residues at a suitable distance, facilitating the identification of a larger spectrum of interaction partners with higher sensitivity.

Another common approach for the investigation of protein interactions is co-IP, which we used to confirm two PPIs of members of the largest interaction networks, namely the FLOT1/FLOT2 complex and the V-ATPase complex. We selected the interaction of ATP6V1D with FZD9, a G-protein-coupled multi-pass transmembrane receptor for WNT2[71], as another member of this family, FZD8, was previously identified to interact with ATP6AP2[41]. Intriguingly, it was shown in this study that correct V-ATPase function, and accordingly acidification of the lysosome, is necessary for WNT signaling, and that interaction of FZD8 with the V-ATPase complex plays a decisive role[72]. The direct interaction of ATP6V1D and FZD9 identified in this study substantiates this functional connection of WNT signaling and lysosomal acidification, with a possible role related to the latter. V-ATPase can only acidify lysosomes when the V0 part, which is integrated in the lysosomal membrane, pairs with the cytosolic V1 part. The independently assembled V1 part[7] can reversibly dissociate from V0, which appears to be a process that can be regulated through different types of stimuli[39]. Similarly, it is conceivable that frizzled proteins could control V-ATPase V0/V1 assembly, and thus possibly regulate lysosomal acidification, and hence WNT signaling.

The second major lysosomal complex covered in our dataset was mTORC1. Among others, we detected cross-links for the interaction of the same lysine in the N-terminal region of LAMTOR1 with different residues of LAMTOR3. Both proteins are members of the Ragulator complex, whose interaction with the RAG GTPases regulates mTORC1 activity[73]. It was shown previously, that the N-terminal region of LAMTOR1 could not be crystallized without the RAG GTPases, implicating that it could exist in an unordered state under these circumstances[48]. Matching of the three LAMTOR1/3 cross-links to the Ragulator crystal structure determined in the presence of RAG GTPases fulfilled only for one of them DSSO's distant constraints. A possible explanation for the other two cross-links is that they originated from a state where Ragulator was not interacting with the RAG GTPases, indicating an alternative structure of the LAMTOR1 N-terminal region in situ.

When matching of cross-links to a structure determined from overexpression or in vitro experiments results in violation of DSSOs distance constraints, it can indicate that this structure does not present the native form of a protein or complex. This applied to the inter-links matched to the dimeric model of PPT1 (PDB identifier 3GRO), which were responsible for 43% of overlength cross-links in our dataset. On the other hand, all intralinks matched the published structure. This prompted us to further investigate possible routes of PPT1 assembly. Our proposed tetrameric model satisfied all distance constraints, which is in line with previous findings detecting the majority of enzymatic activity in size exclusion chromatography fractions correlating with the molecular weight of a tetrameric assembly[50]. The observed discrepancy to the crystallography-based structure could be related to the expression system utilized (*Spodoptera frugiperda*), as it lacks the capability to glycosylate PPT1 properly. In line with this possibility, it was shown previously, that glycosylation-deficient PPT1 variants are devoid of enzymatic activity, which was attributed to improper folding[50,74].

The flotillins, which yielded the highest number of cross-links, were reported in several studies to interact with a large variety of proteins at different subcellular locations and to be involved in a plethora of processes[75]. FLOT1 has been demonstrated to play a role in clathrin/caveolin-independent endocytosis[65], and the cargo of such FLOT1-positive endosomes has been shown to be delivered to lysosomes[63,76]. Furthermore, FLOT1 was detected at the lysosomes' cytosolic face[51,52]. In accordance with these findings, we detected and verified the interaction of FLOT1/FLOT2 with guanine-nucleotide-binding protein subunit beta-4 (GNB4), a beta subunit of heterotrimeric G-proteins[77]. Variants of GNB4 were shown to cause Charcot–Marie–Tooth disease[78], and downregulation of GNB4 levels has been found in Gaucher Disease, a lysosomal storage disorder[79]. Association of GNB4 with lysosomes was implicated in a previous study[80], and a yeast two-hybrid screen revealed its direct interaction with LAMP2[81], which co-localizes with FLOT1 and the lysosomal surface[51]. Taken together, these data provide further evidence that GNB4 is located at lysosomes, and imply that flotillins could play a role in its localization and/or function.

With respect to the structure of flotillins, only rudimentary experimental information exists based on NMR analyses of the mouse Flot2 N-terminal region (PDB identifier 1WIN). For the remaining protein, until recently, the predicted structure featured a 180° turn in the central helix of both flotillins resulting in direct interaction of the proteins' N- and C-terminal domains[43]. The FLOT1 and FLOT2 structures predicted by AlphaFold2 disagree with this model, featuring an extended helix with a bend in its middle for both of the proteins. This is supported by the cross-links detected in our dataset. We did not observe long-distance intralinks, which would be indicative of a major fold resulting in the proximity of distant regions of the proteins, and the inter-links of both proteins behave in unison, confirming intermolecular interactions along an extended region of the heterodimer.

We detected interaction with GNB4, among others, in the putative interaction hotspot, which localized in both flotillins to the major bend in the extended alpha-helical structure. Intriguingly, both the FLOT1 and the FLOT2 tyrosine residue located in this structure were reported in PhosphoSitePlus[82] to be phosphorylated in >30 studies. Therefore, they could present a potential regulatory element for PPIs, e.g., to discriminate interactions regulated by tyrosine kinase- or G-protein-coupled receptor signaling, as it was shown previously that flotillins are involved in both types of signaling pathways[83]. Deletion of this interaction hotspot in both FLOT1 and FLOT2 did not only abrogate interaction with GNB4, but also interfered with the formation of higher-order assemblies, providing further evidence for a prominent role in the formation of PPIs for this region of the flotillins.

Except for a few proteins[66], no substrates for flotillin-mediated endocytosis are known to date. Our analysis of its putative cargo identified >300 proteins, including a large number of membrane proteins, receptors, and associated proteins. We were able to confirm interaction of FLOT1 with eleven putative cargo proteins by co-IP, lending further credibility to our dataset. This included membrane-associated as well as transmembrane proteins. Among those, the latrophilins stood out, as all three members of this family of adhesion G protein-coupled receptors[84] were found to be significantly enriched. Furthermore, we were able to co-IP them with both flotillins and observed their colocalization with FLOT1/FLOT2-positive vesicular structures, raising the possibility that they may be targets for flotillin-mediated endocytosis. The fact that only a sub-population of latrophilins co-localized with flotillin-positive vesicles implies that they are either rapidly delivered for lysosomal degradation after their endocytosis, or that they are also endocytosed by other mechanisms, depending, e.g., on clathrin or caveolin.

Methods

Cell culture and enrichment of lysosomes

HEK293 (CRL-1573) and HeLa (CCL-2) cells were obtained from ATCC. Tissue culture plates (10 cm) were coated with 0.5 mg/mL poly-L-lysine (PLL) in 1× phosphate-buffered saline (PBS) for 20 min at 37 °C. On each plate, 6×10^6 HEK293 cells were seeded in full medium supplemented with 10% (v/v) superparamagnetic iron oxide nanoparticles (SPIONs) solution (DexoMAG40 from Liquid Research Limited, core size: 8 nm, hydrodynamic size: 50 nm, Fe content: 10 mg/ml, coating: dextran 40 kDa) and incubated for 24 h. Subsequently, cells were washed three times with pre-warmed PBS, fresh full medium was added, and cells were incubated for 24 h. Prior to harvesting, cells were washed twice with ice-cold PBS, scraped off the plates in 2 mL each of ice-cold isolation buffer (250 mM sucrose, 10 mM HEPES-NaOH pH 7.4, 1 mM $CaCl_2$, 1 mM $MgCl_2$, 1.5 mM MgAc, 1× cOmplete EDTA-free protease-inhibitor cocktail), and pooled. Cell suspensions from four plates each were dounced with 25 strokes in a 15 mL dounce homogenizer, and nuclei, as well as intact cells, were pelleted by centrifugation for 10 min at 600×*g*, 4 °C. The supernatant (post-nuclear supernatant, PNS) was transferred to a new tube, the pellet resuspended in 3 mL of isolation buffer, and dounced and centrifuged again. The supernatant from this step was combined with the first one, and the pooled PNS was used for lysosome enrichment using LS columns in combination with a QuadroMACS magnet (both from Miltenyi Biotec). Columns were equilibrated with 1 mL 0.5% (w/v) bovine serum albumin (BSA) in PBS, the combined PNS of two cell culture plates was applied to one column, and the flow through was collected. After three washing steps with 1 mL isolation buffer each, columns were removed from the magnet, and lysosomes were eluted twice in 1 mL of isolation buffer using a plunger. Individual eluate fractions were centrifuged for 30 min at 20,000×*g*, 4 °C, the supernatants were discarded, the pellets were resuspended in isolation buffer, and for each biological replicate (individual experiments from different passage numbers), the pellets from 64 plates were pooled. Protein concentrations were determined using the DC protein assay. The efficiency of lysosome enrichment and lysosomal integrity was assessed using the β-hexosaminidase assay[85]. Fractions obtained from lysosome enrichment (25 µL each) were combined with 8 µL of 10% Triton X-100 or 8 µL of PBS, followed by the addition of 50 µL reaction solution (100 mM sodium citrate pH 4.6, 0.2% (w/v) BSA, 10 mM para-nitrophenyl-N-acetyl-2-B-D-glucosaminide) in a 96-well plate format. Subsequently, the plate was incubated for 15 min at 37 °C, and 200 µL of stop solution (0.4 M glycine-HCl, pH 10.4) was added to the sample. Absorbance was measured at 405 nm on a microplate reader.

Transfection of cells and enrichment of FLOT1-/FLOT2-positive early endosomes

In total, 64 tissue culture plates were coated with PLL (0.5 mg/mL), HEK293 cells were seeded at a density of 3.5×10^6 cells/plate, and cultivated as described above. After 24 h, cells were transfected with 6 μg of plasmid (1:1 mixture of FLOT1-FLAG and FLOT2-FLAG or GFP) using TurboFect. After 4 h, the cell culture medium was replaced, and cells were incubated for 48 h. In the first step, early endosome enrichment was performed in a similar way as lysosome enrichment using adjusted pulse-chase conditions. After addition of SPIONs solution to the cells (10% (v/v) final concentration), cells were incubated for 5 min at 37 °C, washed with pre-warmed PBS, and fresh full medium was added for a 5 min chase. Subsequently, the plates were placed on ice, cells were washed twice with ice-cold PBS, scraped off the plates in 2 mL each of ice-cold isolation buffer, and pooled. Cell suspensions from four plates each were pooled, dounced, and centrifuged for 10 min at 600×*g*, 4 °C. The PNS was transferred to a new tube, and after additional douncing of the pellet and centrifugation of the solution both PNS fractions were combined. Early endosomes were enriched using LS columns in combination with a QuadroMACS magnet. The eluate fractions of individual endosome enrichments were pooled, and FLOT1-FLAG/FLOT2-FLAG-positive endosomes were further enriched by magnetic anti-FLAG beads (80 μL). After the addition of beads, eluate fractions were incubated on an end-over-end rotator at 4 °C for 4 h, beads were separated from samples by magnetic force (DynaMag 2 magnet), and supernatants were transferred to new tubes. The beads were washed three times with 500 μL PBS, and FLOT1-FLAG/FLOT2-FLAG-positive endosomes eluted by incubation of the beads with 150 μL of 150 ng/μL 3× FLAG peptide in 1× TRIS (hydroxymethyl) aminomethane buffered saline (TBS, pH 7.6) for 30 min at 600×*g*, 4 °C in a thermomixer. Subsequently, beads were separated from samples by magnetic force and eluate fractions transferred to a new tube. Protein concentrations were determined using the DC protein assay. For the generation of FLOT1/FLOT2 deletion constructs (Δ 100), 100 amino acids between residue 200 and 300 were deleted by PCR using the Q5 Site-Directed Mutagenesis Kit (New England Biolabs). The following primers were used: FLOT1_FWD: AGACACCTTTTCCTGCTTGG, FLOT1_REV: GAGAAGTCCCAACTAATTATGCAGG, FLOT2_FWD: CTCCT TCTTGCACTCAGCTTCC, FLOT2_REV: GCCGAGGGTGAAAAGGTGAA. The successful deletion was verified by Sanger sequencing.

Cross-linking of samples

For lysosome- and early endosome-enriched fractions, a portion of the eluate containing 500 μg and 200 μg protein, respectively, was transferred to a new tube. Intact organelles were pelleted by centrifugation for 30 min at 20,000×*g*, 4 °C, the supernatant discarded, and the pellets resuspended in isolation buffer at a protein concentration of 1 mg/mL. Lysosomal samples were cross-linked in two states (intact and disrupted) while endosomal samples were cross-linked only in the intact state. For disruption of lysosomes, resuspended samples were lyzed with a sonicator (Bioruptor Plus) at an amplitude of 40 with three cycles of 30 s each. All samples were cross-linked at final DSSO concentrations of 5, 2, 1, 0.5, 0.25 mM for investigation of the optimal cross-linker concentration (DSSO titration). For the XL-LC-MS/MS experiments of both lysosome- as well as endosome-enriched fractions, a final concentration of 0.25 mM DSSO was applied. After the addition of DSSO, the cross-link reaction was allowed to proceed for 30 min at room temperature, and quenched by the addition of TRIS-HCl pH 8.0 (20 mM final concentration). Subsequently, proteins were precipitated by the addition of acetone at a ratio of 4:1 (v/v) and incubation overnight at −20 °C. The next day, samples were centrifuged for 20 min at 20,000×*g*, 4 °C, the supernatant was discarded, the pellet was washed twice with ice-cold acetone, air-dried, and stored at −80 °C until further use.

Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE)

Polyacrylamide gels were prepared in-house. Both running and stacking gels were prepared with 10% (w/v) SDS, 40% (v/v) acrylamide, 10% (w/v) ammonium persulfate (APS), and 1% (v/v) tetramethylethylenediamine (TEMED), while 1.5 M TRIS-HCl pH 8.8 and 0.5 M TRIS-HCl pH 6.8 was used for running and stacking gels, respectively. Laemmli buffer[86] (4× stock, 240 mM TRIS-HCl pH 7.4, 4% (v/v) β-mercaptoethanol, 8% (w/v) SDS, 40% (v/v) glycerol, 4% (w/v) bromophenol blue) was added to samples (1× final concentration) followed by incubation for 10 min at 56 °C. Gel electrophoresis was performed at 80–140 V for up to 1.5 h. Gels were either stained overnight with Coomassie Brilliant Blue G-250 or with a Silver Stain Kit.

Antibodies

The following primary antibodies were used in this study: goat anti LIMP2 (Cat# AF1966-SP,1:2000), mouse anti ACT2 (Cat# A5316, 1:4000), and rabbit anti LAMTOR1 (Cat # HPA002997, 1:1000) from Sigma-Aldrich; mouse anti CANX (Cat# 66903-1-AP, 1:20,000), mouse anti FLOT2 (Cat# 66881-1-Ig, 1:1500), mouse anti FZD9 (Cat# 67023-1-Ig, 1:1500), rabbit anti ATP6V1D (Cat# 14920-1-AP, 1:1000), rabbit anti CTSD (Cat# # 21327-1-AP, 1:1000), rabbit anti FLOT1 (Cat# 15571-1-AP, 1:2000), rabbit anti GNB4 (Cat# 11978-2-AP, 1:2000), and rabbit anti SDHA (Cat# 14865-1-AP, 1:800) from Proteintech; goat anti LIMP2 (Cat# AF1966-SP, 1:2000) from R&D system; mouse anti ATP6V1B2 (Cat# SC166045, 1:1000) from Santa Cruz; mouse anti FLOT1 (Cat# 610821, 1:200), mouse anti FLOT2 (Cat# 610383, 1:200), and mouse anti GM130 (Cat# 610822) from BD Biosciences; mouse anti GAPDH (Cat# 5174, 1:2500), rabbit anti EEA1 (Cat# 2411, 1:200), rabbit anti FLOT1 (Cat# 18634, 1:200), and rabbit anti RRAGA (Cat# 4357, 1:1000) from Cell signaling; rabbit anti LPHN2 (Cat# NBP2-58704, 1:100) and rabbit anti LPHN3 (Cat# NLS1138, 1:200) from Novus Biologicals; goat anti-mouse IgG (H + L)-Alexa Fluor 488 (Cat# A-11029, 1:400), rabbit anti ATP6V1A1 (Cat# PA5-29191,1:2000), and rabbit anti LPHN1 (Cat # PA5-77475, 1:200) from Thermo Fisher Scientific; rabbit anti TUBA (Cat# 600-401-880, 1:2000) from Rockland; mouse anti LAMP2 (Cat# H4B4, 1:1000 for WB, 1:100 for IF) from Hybridoma Bank. rabbit anti DSSO (Self-made[35], 1:5000). The following secondary antibodies were used in this study: rat anti-FLAG (Cat# SAB4200119,1:10,000) from Sigma-Aldrich; HRP-coupled goat anti-mouse IgG (Cat# 115-035-044, 1:5000), Cy3-coupled goat anti rabbit IgG (H + L) (Cat# 111-165-144, 1:400), and HRP-coupled goat anti rabbit IgG (Cat# 111-035-003, 1:5000) from Dianova.

Immunoprecipitation of proteins

HEK293 cells were seeded at a density of 6×10^6 cells/10-cm plate and cultivated for 48 h in full medium. Cells were washed twice with ice-cold PBS, scraped off the plate in 1 mL of ice-cold PBS, transferred to a microtube, and centrifuged for 10 min at 600×*g*, 4 °C. The supernatant was discarded and the pellet resuspended in either 300 μL of RIPA lysis buffer (50 mM TRIS-HCl pH 7.4, 150 mM NaCl, 1% Triton X-100, 0.1% (w/v) SDS, 0.5% (w/v) sodium deoxycholate (SDC), 1× cOmplete EDTA-free protease-inhibitor cocktail, 1 mM EDTA) or 300 μL of lysis buffer (50 mM TRIS-HCl pH 7.4, 150 mM NaCl, 1% NP-40, 1× cOmplete EDTA-free protease-inhibitor cocktail, 1 mM EDTA). Samples were incubated on ice for 30 min and passed through a 25 Gauge needle every 10 min. Subsequently, the lysate was cleared by centrifugation for 15 min at 20,000×*g*, 4 °C, transferred to a new pre-cooled microtube, and protein concentrations were determined using the DC protein assay. For each sample, lysate containing 1.6 mg of protein was incubated with 3 μg of antibody overnight by end-over-end rotation at 4 °C. The next morning, 60 μL of Protein A beads were added to each sample, followed by end-over-end incubation for 1 h at 4 °C. Beads were pelleted by centrifugation for 5 min at 1000×*g*, 4 °C, supernatants transferred

to new tubes, and beads were washed three times with 500 µL of ice-cold PBS. Proteins were eluted from the beads by incubation in 2× Laemmli buffer for 30 min at 45 °C.

Blue native polyacrylamide gel electrophoresis (BN-PAGE)

Samples were supplemented with solubilization buffer (10 mM HEPES-NaOH pH 7.4, 1% (v/v) digitonin, 2 mM EDTA, 50 mM NaCl, 10% (v/v) glycerol, 1 mM phenylmethylsulfonyl fluoride (PMSF)) and loading dye (10 mM Bis-TRIS-HCl pH 7.0, 0.5% (w/v) Coomassie Brilliant Blue G-250, 50 mM ε-amino n-caproic acid), and loaded to self-cast blue native gradient gels (4–12% (w/v) acrylamide, 0.19–0.40% (w/v) bis-acrylamide, 67 mM ε-amino n-caproic acid, 50 mM Bis-Tris/HCl pH 7.0). The anode buffer was 50 mM Bis-TRIS-HCL pH 7.0, and the cathode buffer 50 mM Tricine pH 7.0, 15 mM Bis-TRIS-HCl pH 7.0, 0.2 % (w/v) Coomassie Brilliant Blue G-250. The temperature of the gel chamber was maintained at 4 °C and electrophoresis was performed at 50 V for 20 h.

Western blotting

Proteins were transferred to nitrocellulose or polyvinylidene fluoride (PVDF) membranes using a semi-dry or wet electro blotter for 1 h or 2 h at 200 mA/membrane. Membranes were blocked in 5% nonfat dry milk in TBS containing 0.05% (v/v) Tween 20 (TBS-T) for 1 h at RT followed by incubation with primary antibodies overnight at 4 °C. The next day, membranes were washed three times with TBS-T for 10 min at RT followed by incubation with secondary antibody for 60 min at RT. Subsequently, membranes were washed three times for 10 min at RT with TBS-T, and the blots were developed. Protein expression signals were detected using the enhanced chemiluminescence (ECL) kit, visualized with the FUSION SOLO 4 M system, and analyzed by the FusionCapt advance software. Uncropped and unprocessed scans can be found in the Source Data File in the Supplementary Information. Individual images present a merged image of the marker and the respective chemiluminescence protein signals.

Immunofluorescence microscopy and image analysis

HeLa and HEK293 cells were seeded in 12-well plates at a density of 2×10^4 and 4×10^5 cells per well, respectively. For HEK293 cells, glass coverslips were coated using PLL (0.5 mg/mL). Cells were cultured for 36 h after seeding and transfected with 1 µg of plasmid (1:1 mixture of FLOT1-GFP/FLOT2-FLAG and FLOT2-mCherry/FLOT1-FLAG) using TurboFect. After 4 h, fresh medium was supplemented and cells were incubated for 48 h. For staining, cells were washed using PBS and fixed with ice-cold methanol at −20 °C for 20 min. After fixation, cells were washed twice with PBS and blocked with 2% BSA in PBS for 1 h at RT. Blocked cells were stained with primary antibodies overnight at 4 °C in a humid chamber. Subsequently, cells were washed three times with TBS for 5 min each and incubated with secondary antibodies for 1 h at RT in the dark. Coverslips were washed three times with TBS for 5 min each, rinsed once with distilled water, and mounted on specimen slides using ROTI Mount FluorCare DAPI. Images were captured using Leica SP5 AOBS with SMD confocal microscope equipped with an HCX PL APO ×63/oil objective and 2× single-photon avalanche diode detector. Images were acquired using the LAS AF software. Images were annotated using Fiji software. Colocalization analyses and determination of Mander's coefficients (MCOE) were performed with the JACoP plugin[87].

Mass spectrometry sample preparation

Precipitated proteins were resuspended in 100 µL of freshly prepared 8 M urea/100 mM triethylammonium bicarbonate (TEAB) and incubated for 45 min at 600 rpm, 37 °C. Disulfide bridges were reduced with 4 mM DTT (final concentration) at 56 °C for 30 min, alkylated with 8 mM chloroacetamide (final concentration) at RT for 30 min[88], and the reaction was quenched by the addition of 4 mM DTT. Subsequently, samples were diluted 1:1 with 100 mM TEAB, rLysC was added at an enzyme-to-protein ratio of 1:100 (w/w), and proteolytic digestion was performed at 37 °C overnight. The following day, the urea concentration was reduced to 1.6 M by the addition of 100 mM TEAB, trypsin was added at an enzyme-to-protein ratio of 1:100 (w/w), and the samples were incubated at 37 °C for 8 h. The resulting peptides were desalted using 50 mg Sep-Pak C_{18} cartridges, dried using a vacuum centrifuge, and stored at −80 °C until further use.

Strong cation-exchange (SCX) chromatography fractionation

SCX fractionation was performed with an UltiMate 3000 RSLC HPLC chromatography system in combination with a PolySULFOETHYL A column (150 mm × 1 mm, 5-µm particle size). Desalted peptides (500 µg each) were reconstituted in 20 µL of SCX solvent A (20% acetonitrile (ACN), 10 mM KH_2PO_4 pH 2.7), loaded to the analytical column with 100% SCX solvent A, and eluted with increasing amounts of SCX solvent B (500 mM KCl, 20% ACN, 10 mM KH_2PO_4 pH 2.7) at a flow rate of 50 µL/min. The gradient was as follows (adapted from[38]) 0–42 min: 0–2% B; 42–50 min: 2–3% B; 50–60 min: 3–8% B; 60–70 min: 8–20% B; 70–80 min: 20–40% B; 80–86 min: 40–90% B; 86–90 min: 90% B; 90–91 min: 0% B; 91–120 min: 0% B. Eluting peptides were collected with a fraction collector, individual fractions dried using a vacuum centrifuge, resuspended in 20 µL of 5% ACN, 5% formic acid (FA), and desalted using C_{18} STAGE-tips[89]. STAGE tip eluates were dried using a vacuum centrifuge, and resuspended in 5% ACN, 5% FA, and peptide concentrations were determined with a quantitative fluorometric peptide assay.

Liquid chromatography-electrospray ionization tandem mass spectrometry (LC-ESI-MS/MS)

Dried peptides were resuspended in 5% ACN, 5% FA and analyzed using an UltiMate 3000 RSLCnano UHPLC system coupled to an Orbitrap Fusion Lumos mass spectrometer (Data acquisition with Orbitrap Fusion Lumos Tune Application (3.4) and Xcalibur (4.2)). From each sample (25% of the total amount for individual SCX fractions, 1 µg of non-cross-linked samples) were loaded on a 50 cm C_{18} reversed-phase analytical column at a flow rate of 600 nL/min using 100% solvent A (0.1% FA in water). For analytical column preparation, fused silica capillaries (360 µm outer diameter, 100-µm inner diameter) were used to generate spray tips using a P-2000 laser puller. Tips were packed with 1.9 µm ReproSil-Pur AQ C_{18} particles to a length of 50 cm. Peptide separation was performed with 120 min (SCX fractions) and 240 min (non-cross-linked samples) linear gradients from 5–35% solvent B (90 % ACN, 0.1% FA) at a flow rate of 300 nL/min. MS1 spectra were acquired in the Orbitrap mass analyzer from *m/z* 375 to 1575 at a resolution of 60,000. For peptide fragmentation, charge states from 3+ to 8+ for cross-linked samples and 2+ to 5+ for non-cross-linked samples were selected, and dynamic exclusion was defined as 60 sec and 120 sec for 120 min and 240 min gradients, respectively. Cross-linked samples were either analyzed with an MS2-MS3-MS2 strategy[90] or with a stepped collision energy approach[91], where ions with the highest charge state were prioritized for fragmentation. For both methods, MS2 scans were acquired at a resolution of 30,000 in the Orbitrap analyzer with a dynamic mass window. In case of stepped collision energy, peptides were fragmented using higher collision dissociation (HCD) with 21, 26, and 31% normalized collision energy (NCE). For the MS2-MS3-MS2 fragmentation method, sequential collision-induced dissociation (CID) and electron-transfer dissociation (ETD) spectra were acquired for each precursor. The precursor isolation width was set to *m/z* 1.6 with standard automatic gain control and automatic maximum injection time. The NCE for CID-MS2 scans were set to 25%, and calibrated charge-dependent ETD parameters enabled. MS3 scans were triggered by a targeted mass difference of 31.9721 detected in the MS2 scan. The MS3 scan was performed in the ion trap part of the instrument with CID at 35% NCE with normalized automatic gain control (AGC) target of 300%. DIA analyses were performed with the same instrumental setup as described above. For each sample, 1 µg of peptides were loaded

directly on a reversed-phase analytical column packed with 3-µm Reprosil-Pur AQ C_{18} particles to a length of 40 cm and separated with 120 min linear gradients. After the acquisition of one MS1 scan 24 static windows, DIA MS2 scans were performed. MS1 scans were acquired in the Orbitrap analyzer from *m/z* 350 to 1200 at a resolution of 120,000 with a maximum injection time of 20 ms and an AGC target setting of 5×10^5. MS2 scans were defined to cover the MS1 scan range with 36 scan windows of 24.1 *m/z* each, resulting in an overlap of 0.5 m/z and a cycle time of 3.44 s. Peptides were fragmented by HCD with an NCE of 27%, and spectra were acquired in the Orbitrap analyzer with a resolution of 30,000, a maximum injection time of 60 msec, and an AGC target setting of 1×10^6.

Proteome discoverer analysis

Thermo *.raw files from cross-linked samples were analyzed using Proteome Discoverer 2.4, utilizing Mascot 2.5.1 and the XlinkX node for peptide identification. The following settings were used for both algorithms: precursor ion mass tolerance: 10 ppm; Orbitrap fragment ion mass tolerance: 20 ppm; ion trap fragment ion mass tolerance: 0.5 Da; fixed modification: carbamidomethylation at cysteine; variable modification: oxidation at methionine; enzyme: trypsin; the number of allowed missed cleavage sites: 2; minimum peptide length: five amino acids; cross-linking site: lysine (K) and N-terminus of proteins. Data were searched against UniProt *Homo sapiens* (entries: 20,365, release date: 05/2020) in combination with the common repository of adventitious proteins (cRAP) database containing common contaminants. The Proteome Discoverer workflow was split into two branches with a cross-link and standard peptide search. MS2 spectra containing DSSO reporter ions were analyzed with pre-defined "MS2-MS2-MS3" and "MS2" search options using XlinkX. Peptide identifications were accepted with a minimum XlinkX score of 40 and filtered at FDRs of 1 and 5% at the cross-linked peptide spectrum level. Cross-links were exported. Spectra, which did not contain reporter ions were searched using Mascot. Identified peptides were filtered at 1% FDR on the peptide level using Percolator and proteins exported at 1% FDR. Only high-confidence peptide identifications were considered and data exported. Data from both algorithms were further analyzed applying different software packages (R, Excel 2019, GraphPad Prism 6.01, STRING 11.0, Cytoscape 3.8.0, xiVIEW (07/2019), TopoLink (06/2019), PSIPRED 4.0, PCOILS (07/2018), and PyMol 2.3).

MaxQuant analysis

Thermo *.raw files from non-cross-linked samples were analyzed using MaxQuant 2.0.3[92] for determining iBAQ values[37]. The following settings were used: precursor ion mass tolerance: 4.5 ppm; Orbitrap fragment ion mass tolerance: 20 ppm; fixed modification: carbamidomethylation at cysteine; variable modifications: oxidation at methionine, acetylation at protein N-terminus, and deamidation at asparagine (N) as well as glutamine (Q); enzyme: trypsin; the number of allowed missed cleavage sites: 2; minimum peptide length: 5 amino acids. Data were searched against UniProt *Homo sapiens* (Entries: 20,365) in combination with the cRAP database containing common contaminants. Data were filtered at 1% FDR on the peptide level and protein level and exported, followed by analysis with different software packages (Excel and GraphPad Prism).

Spectronaut analysis

Thermo *.raw DIA files from FLOT1-FLAG + FLOT2-FLAG-transfected HEK293 cells were analyzed using Spectronaut 14.7.20. Initially, hybrid spectral libraries were generated from both DDA and DIA files with the Pulsar search engine integrated into Spectronaut, applying the following parameters: precursor ion mass tolerance: dynamic; Orbitrap fragment ion mass tolerance: dynamic; fixed modification: carbamidomethylation at cysteine; variable modifications: oxidation at methionine, acetylation at protein N-terminus and deamidation at asparagine (N) as well as glutamine (Q); enzyme: trypsin; the number of allowed missed cleavage sites: 2; minimum peptide length: five amino acids. Data were searched against UniProt *Homo sapiens* (entries: 20,365, release date: 05/2020) in combination with the cRAP database containing common contaminants. For each peptide, the 3–6 most abundant b/y ions were selected for library generation, dependent on their signal intensity. Dynamic retention time alignment was performed based on the high-precision indexed retention time (iRT) concept[93]. Mass tolerances (precursor and fragment ions), as well as peak extraction windows, were defined automatically by Spectronaut. Normalization was disabled, and data filtered at 1% FDR on the peptide and protein level (q-value <0.01). High-confidence identifications were exported, followed by analysis with different software packages (R, Excel, STRING, Cytoscape, and GraphPad Prism).

Structural analysis

Protein sequences were obtained from UniProt and searched using BLAST 2.9.0 against protein database (PDB) entries with an E-value of 0.0001. In case no reference structure for *Homo sapiens* was available, structures from other organisms were obtained from the SWISS-MODEL repository (09/2021)[46], and/or predicted structures from the AlphaFold2 protein structure database were used[47]. Amino acid numbering was adjusted to UniProt entries, identified cross-links were mapped and topologically evaluated with TopoLink (06/2019)[45], and visualized using PyMol 2.3.

Molecular docking

Protein structure perturbation and optimization was performed with SCWRL 4.0[94] and restraint-based docking with HADDOCK2.2[95] as well as CNS 1.3[96]. Distance constraints of identified cross-links (20 ± 10 Å) were used to limit the possible interaction search space, applying unambiguous restraint distances on C-beta (except for glycine, for which C-alpha was used). In line with the default HADDOCK protocol, 500 initial restraints-based complex models were generated, followed by their rigid-body energy minimization. For the best 100 models, semi-flexible refinement in torsion angle space was performed, followed by molecular dynamics refinement in explicit water. Generated models were evaluated based on the weighted sum of electrostatic and van der Waals energies, complemented by the empirical desolvation energy. Based on these parameters, the ten best-scoring models were reported. Finally, models were further clustered within a 5 Å pairwise root mean square deviation, and the lowest energy model of each cluster reported. Results were visualized using PyMol 2.3.

Identification of putative FLOT-endosome cargo by statistical analysis

Spectronaut results were analyzed using R (R Core Team, 2020) applying the integrated development environment RStudio 1.3.1056. For data importing, tidying, transforming and statistical modeling Tidyverse 1.3.0[97] and R base packages were used. Results were exported using Openxlsx 4.1.5 and visualized by Ggplot 3.3.2 and Viridis 0.5.1. Protein signal intensities were initially log2-transformed for data quality assessment and visualization. Missing values were replaced by "NA", while imputation of missing values was omitted. Subsequently, data from individual replicates of the experimental conditions (GFP, SPIONs, and SPIONs+IP) were categorized into three populations as follows: background (three valid values in all three conditions), SPIONs-specific (no valid GFP value and ≥2 valid values in SPIONs or SPIONs+IP) and SPIONs "on/off" (≥2 valid values in only one SPIONs condition and no valid value in other conditions). Subsequently, non-logarithmic data of the individual datasets were normalized on the signal intensities of FLOT1 and FLOT2 in the respective datasets, while the GFP sample was not normalized, followed by log2 transformation of all datasets. Proteins with ≥2 valid values in each dataset were compared using a two-sided unpaired t test. On/off proteins were

defined as significant and their *P* values set to zero; while *P* values of proteins not matching any of the two conditions were set to one. *P* value adjustment was performed according to Benjamini–Hochberg[98] and proteins with a *q*-value ≤0.05 were considered differently enriched. Significantly enriched proteins were submitted to protein–protein interaction and enrichment analysis with STRING 11.0[99] while the entire list of proteins found in the experiment was defined as a background gene set. Networks (type: full) were generated, in which the edges indicate both functional and physical protein associations.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw MS-data data generated in this study have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository database and are publicly available as of the date of publication under the accession code PXD030532 (XL/DIA-LC-MSMS and LC-MSMS data of lysosome-/early endosome-enriched fractions). The integrative cross-link models for PPT1 and FLOT1/FLOT2, as well as all docking input/output parameters generated in this study, are provided in Supplementary Data 8 and 9. Protein structure data used in this study are available from the PDB database. List of PDB accessions: 6WLW (ATP6V0B), 6WLZ (ATP6V1A), 6B9X (LAMTOR3), 5Y39 (LAMTOR5), 5TDH (GNAI1), 5FWK (HSP90AB1), 5ULS (HSP90B1), 3WGD (TXNDC5), 3TJB (PRDX4), 4XCS (PRDX1), 3GRO (PPT1), 1W7B (ANXA2), 1R46 (GLA), 3WEZ (GLB1), 6CES (NPRL3), 3PDF (CTSC), 5UX4 (CTSD), 1GL2 (STX7), 6ASY (HSPA5), 3C7N[] (SSE1), 6WM2 (ATP6V1E1), 5Y3A (LAMTOR1-5), 5H64 (MTOR, RPTOR, MLST8), 5HE0 (GNB1, GNG2, GRK2), 3LMY (HEXB), 2GJX (HEXA), 6JMQ (SLC7A5). Source data are provided with this paper.

References

1. Fraldi, A., Klein, A. D., Medina, D. L. & Settembre, C. Brain disorders due to lysosomal dysfunction. *Annu. Rev. Neurosci.* **39**, 277–295 (2016).
2. Platt, F. M., d'Azzo, A. L., Davidson, B. L., Neufeld, E. F. & Tifft, C. J. Lysosomal storage diseases. *Nat. Rev. Dis. Prim.* **4**, 1–25 (2018).
3. Shin, H. R. & Zoncu, R. The lysosome at the intersection of cellular growth and destruction. *Dev. Cell* **54**, 226–238 (2020).
4. Ballabio, A. & Bonifacino, J. S. Lysosomes as dynamic regulators of cell and organismal homeostasis. *Nat. Rev. Mol. Cell Biol.* **21**, 101–118 (2019).
5. Liu, G. Y. & Sabatini, D. M. mTOR at the nexus of nutrition, growth, ageing and disease. *Nat. Rev. Mol. Cell Biol.* **21**, 183–203 (2020).
6. Cabukusta, B. & Neefjes, J. Mechanisms of lysosomal positioning and movement. *Traffic* **19**, 761–769 (2018).
7. Wang, L., Wu, D., Robinson, C. V., Wu, H. & Fu, T. M. Structures of a complete human V-ATPase reveal mechanisms of its assembly. *Mol. Cell* **80**, 501–511 (2020).
8. Garg, S. et al. Lysosomal trafficking, antigen presentation, and microbial killing are controlled by the Arf-like GTPase Arl8b. *Immunity* **35**, 182–193 (2011).
9. Balderhaar, H. J. & Ungermann, C. CORVET and HOPS tethering complexes—coordinators of endosome and lysosome fusion. *J. Cell Sci.* **126**, 1307–1316 (2013).
10. Skowyra, M. L., Schlesinger, P. H., Naismith, T. V. & Hanson, P. I. Triggered recruitment of ESCRT machinery promotes endolysosomal repair. *Science* **360**, eaar5078 (2018).
11. Reddy, A., Caler, E. V. & Andrews, N. W. Plasma membrane repair is mediated by Ca(2+)-regulated exocytosis of lysosomes. *Cell* **106**, 157–169 (2001).
12. Levin-Konigsberg, R. et al. Phagolysosome resolution requires contacts with the endoplasmic reticulum and phosphatidylinositol-4-phosphate signalling. *Nat. Cell Biol.* **21**, 1234–1247 (2019).
13. Hao, F. et al. Rheb localized on the Golgi membrane activates lysosome-localized mTORC1 at the Golgi-lysosome contact site. *J. Cell Sci.* **131**, jcs208017 (2018).
14. Chu, B. B. et al. Cholesterol transport through lysosome-peroxisome membrane contacts. *Cell* **161**, 291–306 (2015).
15. Liao, Y. C. et al. RNA granules hitchhike on lysosomes for long-distance transport, using annexin A11 as a molecular tether. *Cell* **179**, 147–164.e120 (2019).
16. Wong, Y. C., Ysselstein, D. & Krainc, D. Mitochondria-lysosome contacts regulate mitochondrial fission via RAB7 GTP hydrolysis. *Nature* **554**, 382–386 (2018).
17. Niedzialkowska, E. et al. Protein purification and crystallization artifacts: the tale usually not told. *Protein Sci.* **25**, 720–733 (2016).
18. O'Reilly, F. J. & Rappsilber, J. Cross-linking mass spectrometry: methods and applications in structural, molecular and systems biology. *Nat. Struct. Mol. Biol.* **25**, 1000–1008 (2018).
19. Yu, C. & Huang, L. Cross-linking mass spectrometry: an emerging technology for interactomics and structural biology. *Anal. Chem.* **90**, 144–165 (2018).
20. Kim, D. I. & Roux, K. J. Filling the void: proximity-based labeling of proteins in living cells. *Trends Cell Biol.* **26**, 804–817 (2016).
21. Muthukottiappan, P. & Winter, D. A proteomic view on lysosomes. *Mol. Omics* **17**, 842–859 (2021).
22. Barysz, H. M. & Malmström, J. Development of large-scale cross-linking mass spectrometry. *Mol. Cell. Proteom.* **17**, 1055–1066 (2018).
23. Chen, Z. A. et al. Architecture of the RNA polymerase II-TFIIF complex revealed by cross-linking and mass spectrometry. *EMBO J.* **29**, 717–726 (2010).
24. Albanese, P., Tamara, S., Saracco, G., Scheltema, R. A. & Pagliano, C. How paired PSII-LHCII supercomplexes mediate the stacking of plant thylakoid membranes unveiled by structural mass-spectrometry. *Nat. Commun.* **11**, 1–14 (2020).
25. O'Reilly, F. J. et al. In-cell architecture of an actively transcribing-translating expressome. *Science* **369**, 554–557 (2020).
26. Schweppe, D. K. et al. Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. *Proc. Natl Acad. Sci. USA* **114**, 1732–1737 (2017).
27. Fasci, D., van Ingen, H., Scheltema, R. A., Heck, A. J. Histone interaction landscapes visualized by crosslinking mass spectrometry in intact cell nuclei. *Mol. Cell. Proteomics* **17**, 2018–2033 (2018).
28. Chavez, J. D., Keller, A., Zhou, B., Tian, R. & Bruce, J. E. Cellular interactome dynamics during paclitaxel treatment. *Cell Rep.* **29**, 2371–2383 (2019).
29. Liu, F., Rijkers, D. T. S., Post, H. & Heck, A. J. R. Proteome-wide profiling of protein assemblies by cross-linking mass spectrometry. *Nat. Methods* **12**, 1179–1184 (2015).
30. Itzhak, D. N., Tyanova, S., Cox, J. & Borner, G. H. Global, quantitative and dynamic mapping of protein subcellular localization. *eLife* **5**, e16950 (2016).
31. Singh, J. et al. Systematic comparison of strategies for the enrichment of lysosomes by data independent acquisition. *J. Proteome Res.* **19**, 371–381 (2020).
32. Walker, M. W. & Lloyd-Evans, E. A rapid method for the preparation of ultrapure, functional lysosomes using functionalized superparamagnetic iron oxide nanoparticles. *Methods Cell Biol.* **126**, 21–43 (2015).
33. Tharkeshwar, A. K. et al. A novel approach to analyze lysosomal dysfunctions through subcellular proteomics and lipidomics: the case of NPC1 deficiency. *Sci. Rep.* **7**, 41408 (2017).

34. Kao, A. et al. Development of a novel cross-linking strategy for fast and accurate identification of cross-linked peptides of protein complexes. *Mol. Cell. Proteomics* **10**, M110.002212 (2011).
35. Singh, J., Ponnaiyan, S., Gieselmann, V. & Winter, D. Generation of antibodies targeting cleavable cross-linkers. *Anal. Chem.* **93**, 3762–3769 (2021).
36. Akter, F. et al. Multi cell line analysis of lysosomal proteomes reveals unique features and novel lysosomal proteins. Preprint at https://www.biorxiv.org/content/10.1101/2020.12.21.423747v1 (2020).
37. Schwanhausser, B. et al. Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
38. Klykov, O. et al. Efficient and robust proteome-wide approaches for cross-linking mass spectrometry. *Nat. Protoc.* **13**, 2964–2990 (2018).
39. Maxson, M. E. & Grinstein, S. The vacuolar-type H^+-ATPase at a glance—more than a proton pump. *J. cell Sci.* **127**, 4987–4993 (2014).
40. Grainger, S. et al. EGFR is required for Wnt9a-Fzd9b signalling specificity in haematopoietic stem cells. *Nat. Cell Biol.* **21**, 721–730 (2019).
41. Cruciat, C.-M. et al. Requirement of prorenin receptor and vacuolar H+-ATPase-mediated acidification for Wnt signaling. *Science* **327**, 459–463 (2010).
42. Babuke, T. et al. Hetero-oligomerization of Reggie-1/flotillin-2 and Reggie-2/flotillin-1 is required for their endocytosis. *Cell Signal* **21**, 1287–1297 (2009).
43. Rivera-Milla, E., Stuermer, C. A. O. & Málaga-Trillo, E. Ancient origin of Reggie (flotillin), Reggie-like, and other lipid-raft proteins: convergent evolution of the SPFH domain. *Cell. Mol. Life Sci.* **63**, 343–357 (2006).
44. Kastritis, P. L. et al. Capturing protein communities by structural proteomics in a thermophilic eukaryote. *Mol. Syst. Biol.* **13**, 936 (2017).
45. Ferrari, A. J. R. et al. TopoLink: evaluation of structural models using chemical crosslinking distance constraints. *Bioinformatics* **35**, 3169–3170 (2019).
46. Waterhouse, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
47. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
48. De Araujo, M. E. et al. Crystal structure of the human lysosomal mTORC1 scaffold complex and its impact on signaling. *Science* **358**, 377–381 (2017).
49. Su, M. Y. et al. Hybrid structure of the RagA/C-Ragulator mTORC1 activation complex. *Mol. Cell* **68**, 835–846.e833 (2017).
50. Lyly, A. et al. Glycosylation, transport, and complex formation of palmitoyl protein thioesterase 1 (PPT1)–distinct characteristics in neurons. *BMC Cell Biol.* **8**, 22 (2007).
51. Kaushik, S., Massey, A. C. & Cuervo, A. M. Lysosome membrane lipid microdomains: novel regulators of chaperone-mediated autophagy. *EMBO J.* **25**, 3921–3933 (2006).
52. Kokubo, H. et al. Ultrastructural localization of flotillin-1 to cholesterol-rich membrane microdomains, rafts, in rat brain tissue. *Brain Res.* **965**, 83–90 (2003).
53. Frick, M. et al. Coassembly of flotillins induces formation of membrane microdomains, membrane curvature, and vesicle budding. *Curr. Biol.* **17**, 1151–1156 (2007).
54. Dempwolff, F. et al. Super resolution fluorescence microscopy and tracking of bacterial flotillin (Reggie) paralogs provide evidence for defined-sized protein microdomains within the bacterial membrane but absence of clusters containing detergent-resistant proteins. *PLoS Genet.* **12**, e1006116 (2016).
55. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
56. Zimmermann, L. et al. A completely reimplemented MPI bioinformatics toolkit with a New HHpred Server at its core. *J. Mol. Biol.* **430**, 2237–2243 (2018).
57. Mirdita, M., Schuetze K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
58. Neumann-Giesen, C. et al. Membrane and raft association of Reggie-1/flotillin-2: role of myristoylation, palmitoylation and oligomerization and induction of filopodia by overexpression. *Biochem. J.* **378**, 509–518 (2004).
59. Riento, K., Frick, M., Schafer, I. & Nichols, B. J. Endocytosis of flotillin-1 and flotillin-2 is regulated by Fyn kinase. *J. Cell Sci.* **122**, 912–918 (2009).
60. Bach, J. N. & Bramkamp, M. Dissecting the molecular properties of prokaryotic flotillins. *PLoS ONE* **10**, e0116750 (2015).
61. Solis, G. P. et al. Reggie/flotillin proteins are organized into stable tetramers in membrane microdomains. *Biochem. J.* **403**, 313–322 (2007).
62. Kwiatkowska, K., Matveichuk, O. V., Fronk, J. & Ciesielska, A. Flotillins: at the intersection of protein S-palmitoylation and lipid-mediated signaling. *Int. J. Mol. Sci.* **21**, 2283 (2020).
63. Stuermer, C. A. O. et al. Glycosylphosphatidyl inositol-anchored proteins and Fyn kinase assemble in noncaveolar plasma membrane microdomains defined by Reggie-1 and −2. *Mol. Biol. Cell* **12**, 3031–3045 (2001).
64. Gorbea, C. et al. A protein interaction network for Ecm29 links the 26 S proteasome to molecular motors and endosomal components. *J. Biol. Chem.* **285**, 31616–31633 (2010).
65. Glebov, O. O., Bright, N. A. & Nichols, B. J. Flotillin-1 defines a clathrin-independent endocytic pathway in mammalian cells. *Nat. Cell Biol.* **8**, 46–54 (2006).
66. Meister, M. & Tikkanen, R. Endocytic trafficking of membrane-bound cargo: a flotillin point of view. *Membranes* **4**, 356–371 (2014).
67. Ge, L. et al. Flotillins play an essential role in Niemann-Pick C1-like 1-mediated cholesterol uptake. *Proc. Natl Acad. Sci. USA* **108**, 551–556 (2011).
68. Thelen, M., Winter, D., Braulke, T. & Gieselmann, V. SILAC-based comparative proteomic analysis of lysosomes from mammalian cells using LC-MS/MS. *Methods Mol. Biol.* **1594**, 1–18 (2017).
69. Eskelinen, E. L. et al. Disturbed cholesterol traffic but normal proteolytic function in LAMP-1/LAMP-2 double-deficient fibroblasts. *Mol. Biol. Cell* **15**, 3132–3145 (2004).
70. Go, C. D. et al. A proximity-dependent biotinylation map of a human cell. *Nature* **595**, 120–124 (2021).
71. Karasawa, T., Yokokura, H., Kitajewski, J. & Lombroso, P. J. Frizzled-9 is activated by Wnt-2 and functions in Wnt/beta-catenin signaling. *J. Biol. Chem.* **277**, 37479–37486 (2002).
72. Hermle, T., Saltukoglu, D., Grünewald, J., Walz, G. & Simons, M. Regulation of frizzled-dependent planar polarity signaling by a V-ATPase subunit. *Curr. Biol.* **20**, 1269–1276 (2010).
73. Zhang, T. et al. Structural basis for Ragulator functioning as a scaffold in membrane-anchoring of Rag GTPases and mTORC1. *Nat. Commun.* **8**, 1394 (2017).
74. Bellizzi, J. J. et al. The crystal structure of palmitoyl protein thioesterase 1 and the molecular basis of infantile neuronal ceroid lipofuscinosis. *Proc. Natl Acad. Sci. USA* **97**, 4573–4578 (2000).
75. Bodin, S., Planchon, D., Rios Morris, E., Comunale, F. & Gauthier-Rouvière, C. Flotillins in intercellular adhesion—from cellular physiology to human diseases. *J. Cell Sci.* **127**, 5139–5147 (2014).
76. Fan, W. L. et al. Flotillin-mediated endocytosis and ALIX-syntenin-1-mediated exocytosis protect the cell membrane from damage caused by necroptosis. *Sci. Signal.* **12**, 10 (2019).

77. Ruiz-Velasco, V., Ikeda, S. R. & Puhl, H. L. Cloning, tissue distribution, and functional expression of the human G protein β4-subunit. *Physiol. Genomics* **8**, 41–50 (2002).
78. Soong, B. W. et al. Exome sequencing identifies GNB4 mutations as a cause of dominant intermediate Charcot-Marie-tooth disease. *Am. J. Hum. Genet.* **92**, 422–430 (2013).
79. Pawlinski, L. et al. Gene expression with corresponding pathways analysis in Gaucher disease. *Exp. Mol. Pathol.* **123**, 104679 (2021).
80. Schroder, B. et al. Integral and associated lysosomal membrane proteins. *Traffic* **8**, 1676–1686 (2007).
81. Haenig, C. et al. Interactome mapping provides a network of neurodegenerative disease proteins and uncovers widespread protein aggregation in affected brains. *Cell Rep.* **32**, 108050 (2020).
82. Hornbeck, P. V. et al. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res.* **43**, D512–D520 (2015).
83. Sugawara, Y. et al. The lipid raft proteins flotillins/reggies interact with Galphaq and are involved in Gq-mediated p38 mitogen-activated protein kinase activation through tyrosine kinase. *Cell. Signal.* **19**, 1301–1308 (2007).
84. Langenhan, T., Piao, X. & Monk, K. R. Adhesion G protein-coupled receptors in nervous system development and disease. *Nat. Rev. Neurosci.* **17**, 550–561 (2016).
85. Wendeler, M. & Sandhoff, K. Hexosaminidase assays. *Glycoconj. J.* **26**, 945–952 (2009).
86. Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685 (1970).
87. Bolte, S. & Cordelieres, F. P. A guided tour into subcellular colocalization analysis in light microscopy. *J. Microsc.* **224**, 213–232 (2006).
88. Müller, T. & Winter, D. Systematic evaluation of protein reduction and alkylation reveals massive unspecific side effects by iodine-containing reagents. *Mol. Cell. Proteom.* **16**, 1173–1187 (2017).
89. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).
90. Liu, F., Lossl, P., Scheltema, R., Viner, R. & Heck, A. J. R. Optimized fragmentation schemes and data analysis strategies for proteome-wide cross-link identification. *Nat. Commun.* **8**, 1–8 (2017).
91. Stieger, C. E., Doppler, P. & Mechtler, K. Optimized fragmentation improves the identification of peptides cross-linked by MS-cleavable reagents. *J. Proteome Res.* **18**, 1363–1370 (2019).
92. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
93. Bruderer, R., Bernhardt, O. M., Gandhi, T. & Reiter, L. High-precision iRT prediction in the targeted analysis of data-independent acquisition and its impact on identification and quantitation. *Proteomics* **16**, 2246–2256 (2016).
94. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins Struct. Funct. Bioinf.* **77**, 778–795 (2009).
95. Van Zundert, G. C. P. et al. The HADDOCK2.2 web server: user-friendly integrative modeling of biomolecular complexes. *J. Mol. Biol.* **428**, 720–725 (2016).
96. Brunger, A. T. Version 1.2 of the crystallography and NMR system. *Nat. Protoc.* **2**, 2728–2733 (2007).
97. Wickham, H. et al. Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
98. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
99. Szklarczyk, D. et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

Acknowledgements

The authors would like to thank Stephane Bodin and Cecile Gauthier for sharing of plasmids containing GFP/RFP-tagged FLOT1/FLOT2 as well as Norbert Rösel and Nadine Weiberg for help with experiments. J.S. received funding by the Studienstiftung des Deutschen Volkes, P.M. by the Jürgen Manchot Stiftung, and H.E. by the IMPRS "From Molecules to Organisms". M.S. and M.E. were funded by the de.NBI project (FKZ 031A 534A) of the Bundesministerium für Bildung und Forschung (BMBF) and D.W. received funding from the FOR2625 of the DFG (Deutsche Forschungsgemeinschaft, German Research Foundation). We would like to thank the Microscopy Core Facility of the Medical Faculty at the University of Bonn for providing help, services, and devices funded by the DFG project number 169331223.

Author contributions

Conceptualization: J.S. and D.W.; methodology: J.S., V.G., D.W.; bioinformatic analysis and structural modeling: J.S., H.E., M.S., and O.K.; investigation: J.S. and P.M.; writing—original draft: J.S. and D.W.; writing—review & editing: J.S., H.E., P.M., M.E., O.K., V.G., and D.W.; visualization: J.S., H.E., and P.M.; funding acquisition: J.S. and D.W.; resources: V.G. and D.W.; supervision: D.W. All authors contributed to the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-33951-0.

Correspondence and requests for materials should be addressed to Dominic Winter.

Peer review information *Nature Communications* thanks Cecile GAUTHIER-ROUVIERE, and the other, anonymous, reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

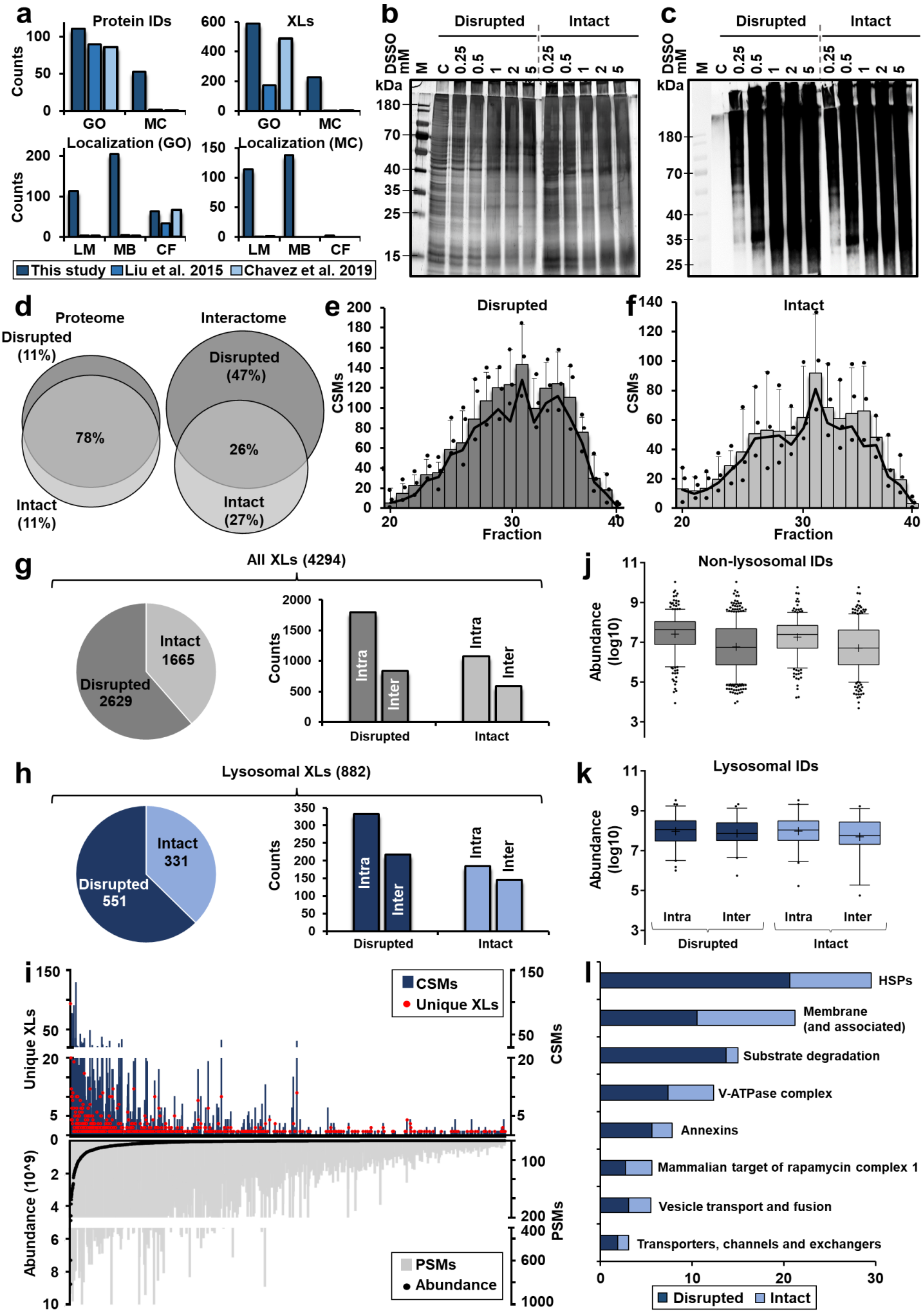
Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

Supplementary Information

Cross-linking of the Endolysosomal System Reveals Potential Flotillin Structures and Cargo

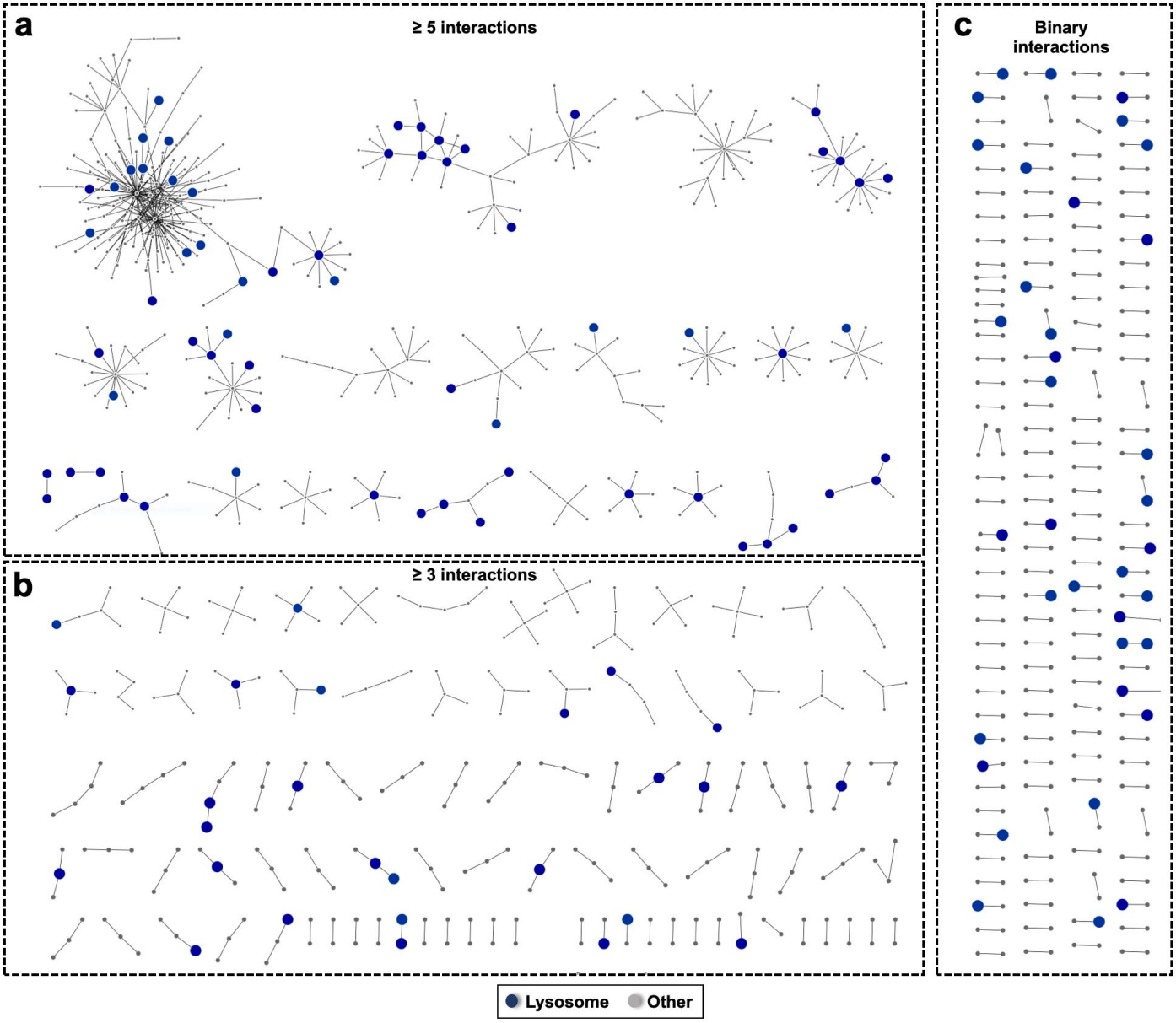
Jasjot Singh, Hadeer Elhabashy, Pathma Muthukottiappan, Markus Stepath, Martin Eisenacher, Oliver Kohlbacher, Volkmar Gieselmann, and Dominic Winter


a
Protein IDs
XLs
Counts
GO
MC
Localization (GO)
Localization (MC)
LM
MB
CF
This study
Liu et al. 2015
Chavez et al. 2019
b
Disrupted
Intact
DSSO mM
M
C
0.25
0.5
1
2
5
kDa
c
d
Proteome
Interactome
Disrupted (11%)
78%
Intact (11%)
Disrupted (47%)
26%
Intact (27%)
e
Disrupted
CSMs
Fraction
f
Intact
g
All XLs (4294)
Intact 1665
Disrupted 2629
Intra
Inter
h
Lysosomal XLs (882)
Intact 331
Disrupted 551
i
Unique XLs
CSMs
Abundance (10^9)
PSMs
Abundance
j
Non-lysosomal IDs
Abundance (log10)
k
Lysosomal IDs
l
HSPs
Membrane (and associated)
Substrate degradation
V-ATPase complex
Annexins
Mammalian target of rapamycin complex 1
Vesicle transport and fusion
Transporters, channels and exchangers

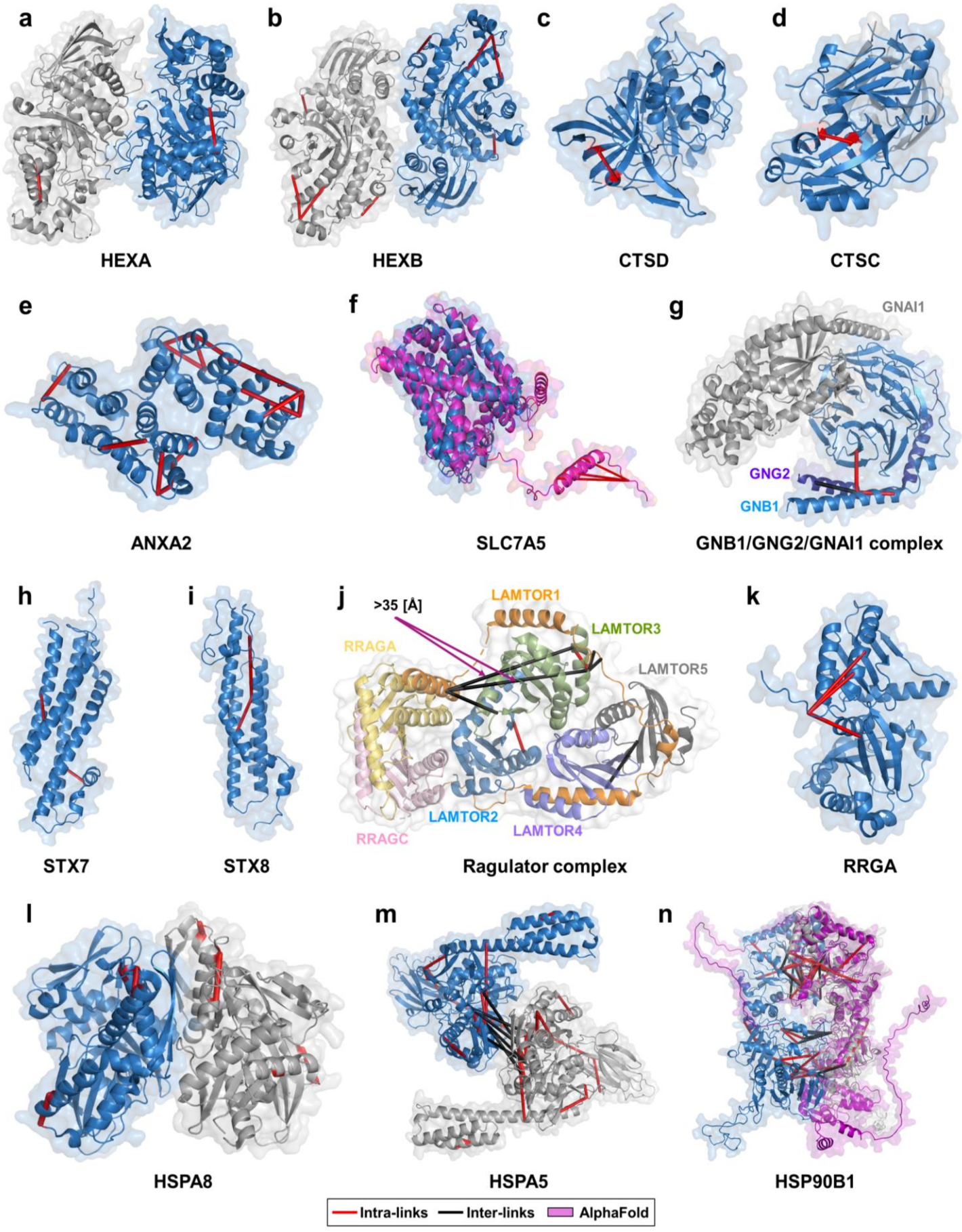
Supplementary Fig. 1: Cross-linking mass spectrometry analysis of HEK293 lysosome-enriched fractions.

a Identified cross-links for lysosomal and lysosome-related proteins. Proteins are categorized in two groups: GO: proteins which are related to lysosomes based on their categorization in gene ontology databases, also including proteins which are not important for lysosomal function; MC manually curated list of proteins which are localized in/at lysosomes [1] and known to play a role in lysosomal function. Numbers from published whole proteome cross-linking studies were extracted from the respective supplementary tables. Cross-links assigned to both GO and MC were further subcategorized based on the protein's location at the lysosome (lumen, membrane, and cytosolic face). **b** Silver-stained SDS-PAGE gel of lysosome-enriched fractions cross-linked with the indicated amounts of DSSO in a disrupted and intact state at a protein concentration of 1 µg/µL (n=2). For each sample, 2.5 µg of protein were loaded to the gel. Control samples were treated with DMSO. Based on the band pattern, the optimal concentration of DSSO was determined to be 0.25 mM (final concentration), as the application of higher amounts resulted in a shift of band patterns. Especially with respect to the lower molecular weight region, less bands were observed, while the higher molecular weight region (containing cross-linked proteins) increased in intensity, which is indicative for protein aggregates resulting from over-cross-linking. Shown are two separate SDS-PAGE analyses (performed under the same conditions) for samples cross-linked in the intact and the disrupted state. Individual gels are indicated by a dashed grey line. **c** Western blot analysis of samples described in (b) for the visualization of cross-linker-modified proteins using an in-house produced antibody which shows immunoreactivity against water-quenched DSSO[2] (n=2). For each sample, 1 µg of protein was loaded to the gel. Shown are two separate western blot analyses (performed under the same conditions) for samples cross-linked in the intact and the disrupted state. Individual gels are indicated by a dashed grey line. **d** Overlap in proteins, which are not categorized as lysosomal based on their GO term, identified in mass spectrometric analyses of lysosome-enriched fractions in the disrupted and the intact state. Proteome: proteins identified in analyses of non-cross-linked samples; Interactome: unique cross-links identified in analyses of cross-linked samples. **e, f** Distribution of identified CSMs in LC-MS/MS analyses of SCX-fractionated tryptic digests of lysosome-enriched fractions cross-linked in a disrupted (e) and intact (f) state. Since later eluting SCX fractions contain the majority of higher charged cross-linked tryptic peptides, only SCX fractions 20-40 were analyzed by LC-MS/MS. Data are presented as mean values +SD (n=3, biologically independent samples over 3 independent experiments). **g, h** Distribution of identified cross-links from the analysis of lysosome-enriched fractions cross-linked in the disrupted and the intact state. Shown are numbers identified in the disrupted/intact state, and how many cross-links were assigned as intra- or inter-links, for cross-links identified in the whole dataset (g) and such assigned to lysosomal proteins (h). **i** Correlation of cross-link identification and protein abundance for all non-lysosomal proteins. Proteins are sorted based on their mean iBAQ abundance[3] in the LC-MS/MS analysis of non-cross-linked samples. Mean iBAQ abundances as well as total numbers of unique cross-links, CSMs, and PSMs are shown. Values are based on the analysis of lysosome-enriched fractions in the intact and disrupted state (n=3 each). **j, k** Distribution of iBAQ abundances for all (j) and lysosomal (k) proteins for which intra- or inter-links were detected. Values are based on the LC-MS/MS analysis of non-cross-linked lysosome-enriched fractions in the disrupted and the intact state. Only proteins for which cross-links were identified in the respective state were considered. Data are presented as box-whisker plots (n=3, biologically independent samples over 3 independent experiments). Each box depicts the interquartile range (IQR, the range between the 25th and 75th percentile, 2.5% +/- whisker, median=line, mean=+). **i** Categorization of the 1,415 CSMs identified for 68 lysosomal proteins. Proteins were assigned to previously defined categories of lysosomal proteins [4].

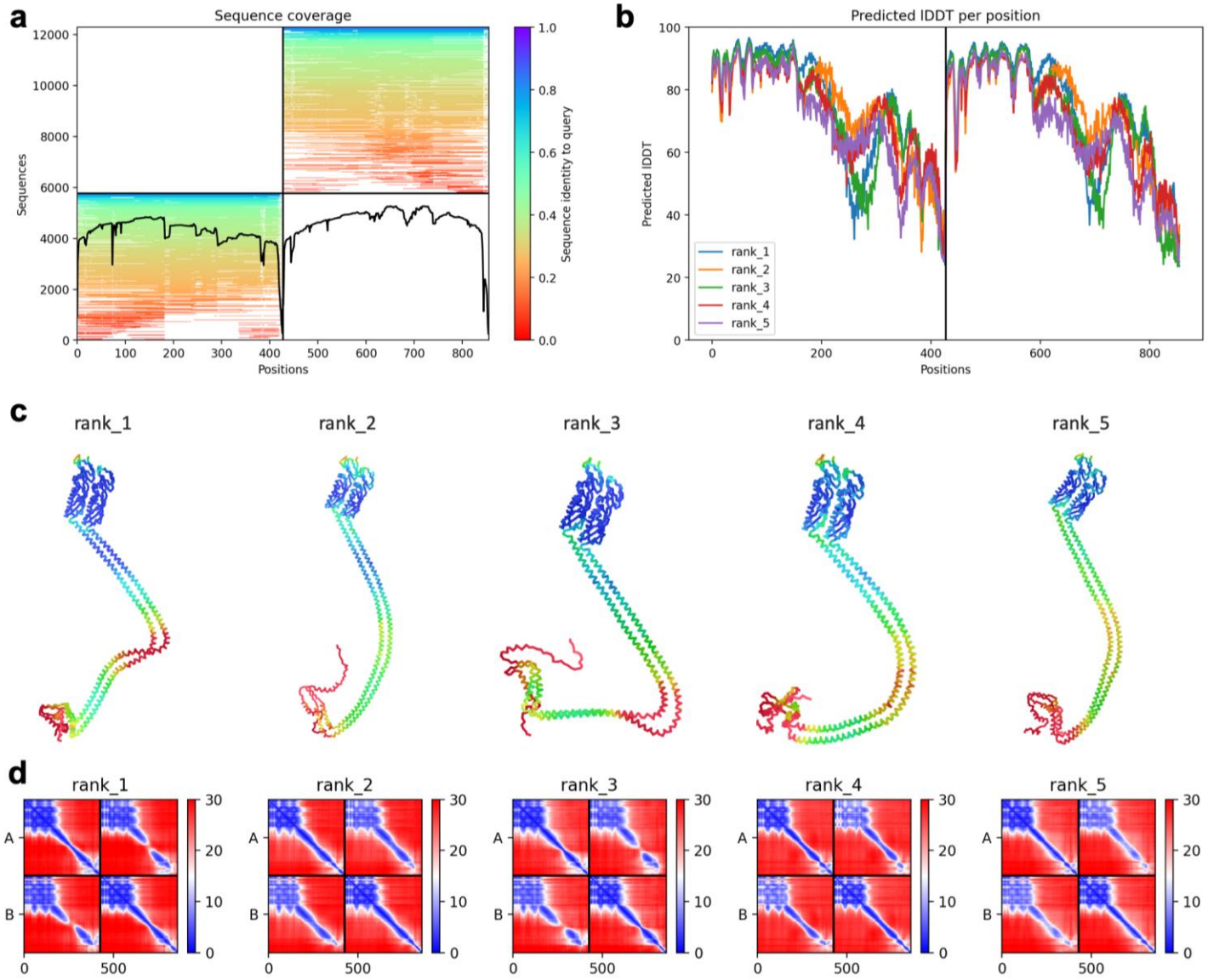
IDs: identifications; LM: lumen; MB: membrane; CF: cytosolic face; GO: gene ontology; MC: manually curated; DSSO: disuccinimidyl sulfoxide; DMSO: dimethyl sulfoxide; C: control (no DSSO); M: protein marker; iBAQ: intensity based absolute quantification; SCX: strong cation exchange; XL: cross-link; PSMs: peptide spectral matches; CSMs: cross-link spectral matches; HSPs: heat shock proteins.


a
≥ 5 interactions
b
≥ 3 interactions
c
Binary interactions
Lysosome
Other
d
MTOR
RPTOR
RRAGA
LAMTOR4
LAMTOR3
LAMTOR5
LAMTOR1
e
STX7
LS
NC
PM
CS
ER
f
Disrupted
R1
626 (34.1%)
167
(9.3%)
79
(4.9%)
189
(10.5%)
R2
551 (24.4%)
90
(5%)
R3
213 (11.8%)
Intact
R1
389 (30%)
110
(8.5%)
25
(1.9%)
69
(5.3%)
R3
127 (9.7%)
93
(7.2%)
R2
486 (37.4%)

Supplementary Fig. 2: Analysis of protein-protein interactions in lysosome-enriched fractions. **a-c** Protein-protein interaction networks based on cross-links identified between two different proteins (inter-links, 1,023 in total). Lysosomal proteins are highlighted as large blue filled circles while non-lysosomal proteins are depicted as small grey dots. Interactions were extracted from the entire XL-LC-MS/MS dataset from disrupted and intact lysosomes. Based on the number of interactors involved in a particular network, data were grouped into different classes: a) ≥5 interactors, b) ≥3 interactors, c) binary interactions. The networks were generated using Cytoscape. **d** Protein-protein interaction networks for proteins related to mTORC1; subcellular localization of interactors is indicated by color. The networks were generated using Cytoscape. **e** Protein-protein interactions identified for Syntaxin 7 (STX7); subcellular localizations of interactors are indicated by color. The network was generated using Cytoscape. **f** Overlap of identified cross-links for the three individual biological replicates of cross-linked lysosome enriched fractions in a disrupted and intact state.
LS: lysosome; NC: nucleus; PM: plasma membrane; CS: cytoskeleton; ER: endoplasmic reticulum.


a
HEXA
b
HEXB
c
CTSD
d
CTSC
e
ANXA2
f
SLC7A5
g
GNAI1
GNG2
GNB1
GNB1/GNG2/GNAI1 complex
h
STX7
i
STX8
j
>35 [Å]
LAMTOR1
LAMTOR3
RRAGA
LAMTOR5
LAMTOR2
LAMTOR4
RRAGC
Ragulator complex
k
RRGA
l
HSPA8
m
HSPA5
n
HSP90B1
Intra-links
Inter-links
AlphaFold

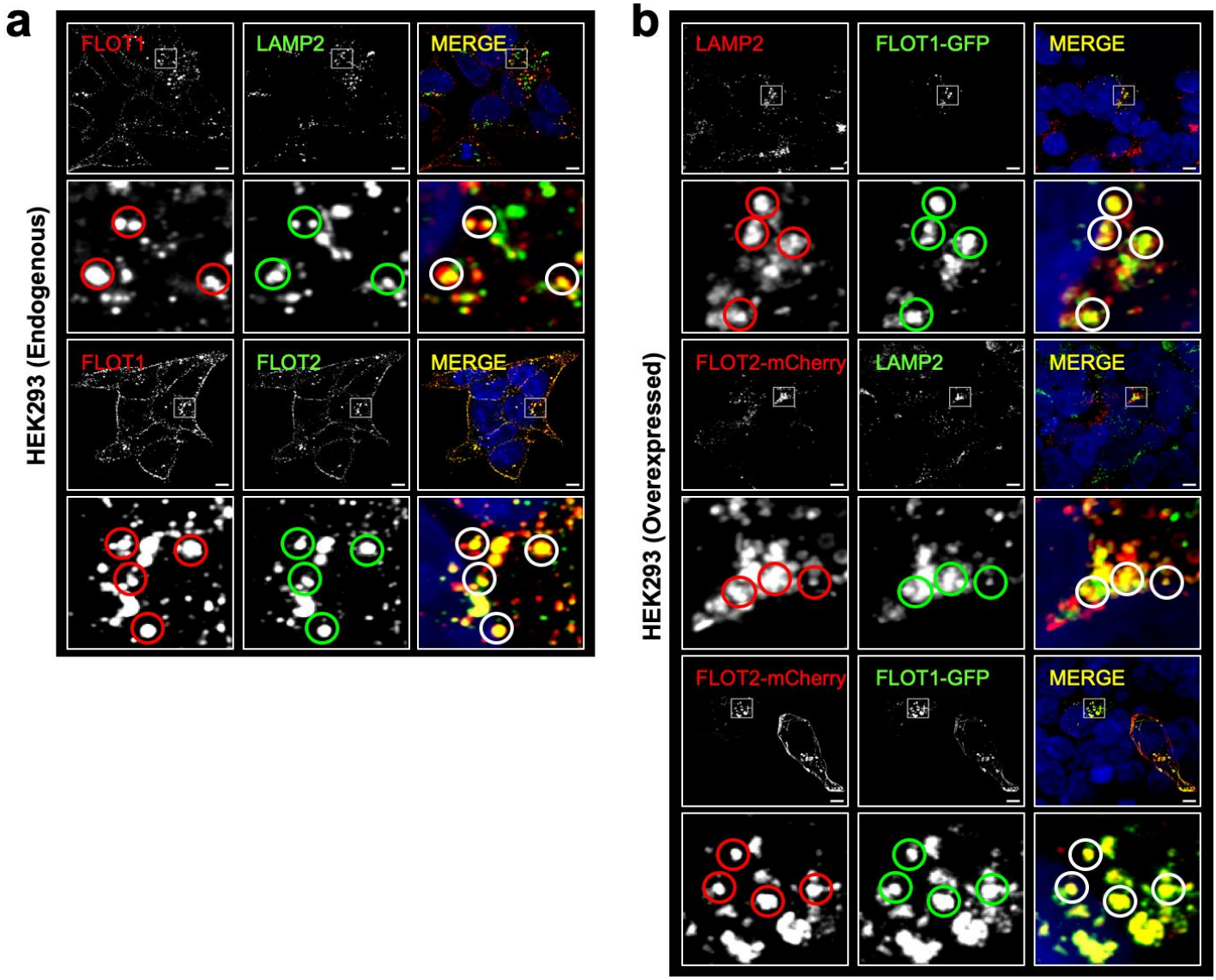
Supplementary Fig. 3: Integration of cross-links into resolved structures and AlphaFold2 models of lysosomal and lysosome-associated proteins.
a Integration of cross-links into the RS of beta-hexosaminidase subunit alpha (HEXA). **b** Integration of cross-links into the RS of Beta-hexosaminidase subunit beta (HEXB). **c** Integration of cross-links into the RS of cathepsin D (CTSD). **d** Integration of cross-links into the RS of cathepsin C (CTSC).
e Integration of cross-links into the RS of annexin A2 (ANXA2). **f** Integration of cross-links into the AF [5] model of the large neutral amino acids transporter small subunit 1 (SLC7A5). **g** Integration of cross-links into the RS of the complex of GNB1, GNG2, and GNAI1. **h** Integration of cross-links into the homology model of syntaxin 7 (STX7). **i** Integration of cross-links into the homology model of syntaxin 8 (STX8). **j** Integration of cross-links into the RS of the Ragulator complex (LAMTOR1-5) in the presence of RRAGA and RRAGC. **k** Integration of cross-links into the AF model of Ras-related GTP-binding protein A (RRGA). **l** Integration of cross-links into the RS of heatshock protein 8 (HSPA8). **m** Integration of cross-links into the RS of heatshock protein 5 (HSPA5). **n** Integration of cross-links into the mixed model based on the RS and the AF model for heatshock protein 90B1 (HSP90B1). Magenta structures were extracted from AlphaFold while grey/blue structures are based on crystal or cryo-EM structures retrieved from PDB.
Inter-links (cross-links between two different proteins or two subunits of the same protein) are shown in black and intra-links (cross-links within the same protein) are shown in red. Overlength cross-links (>35 Å) are indicated.
RS: resolved structure; AF: AlphaFold2.



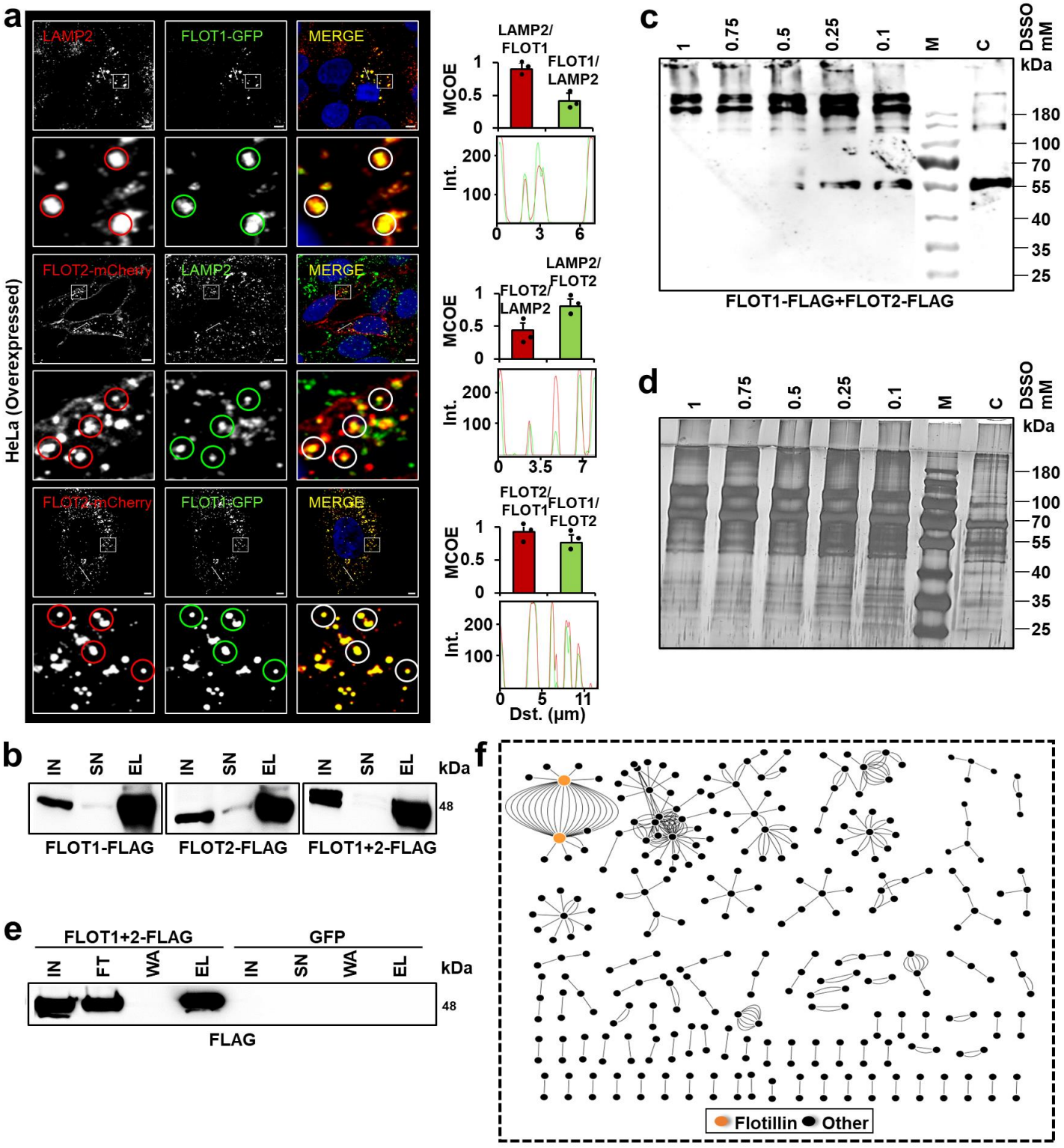
Supplementary Fig. 4: Quality assessment of different AlphaFold2-based ColabFold models of the FLOT1/FLOT2 heterodimeric complex.

a Visualization of depth and diversity of the multiple sequence alignment (MSA). **b** Predicted local distance difference test (plDDT) assessing local structural confidence of the different models. LDDT is a metric that evaluates local distance differences between all heavy atoms in a model, including validation of stereochemical plausibility. It ranges from 0 to 100, where 100 is the most confident. **c** Five highest scoring ColabFold models for the FLOT1-FLOT2 heterodimer. Color coding of individual structural features is based on their plDDTs per residue. **d** Inter-chain predicted aligned error (PAE), which aims to evaluate the position error at residue x, if the predicted and the true structures were aligned on residue y. It indicates the pairwise confidence of the respective prediction and ranks the models.

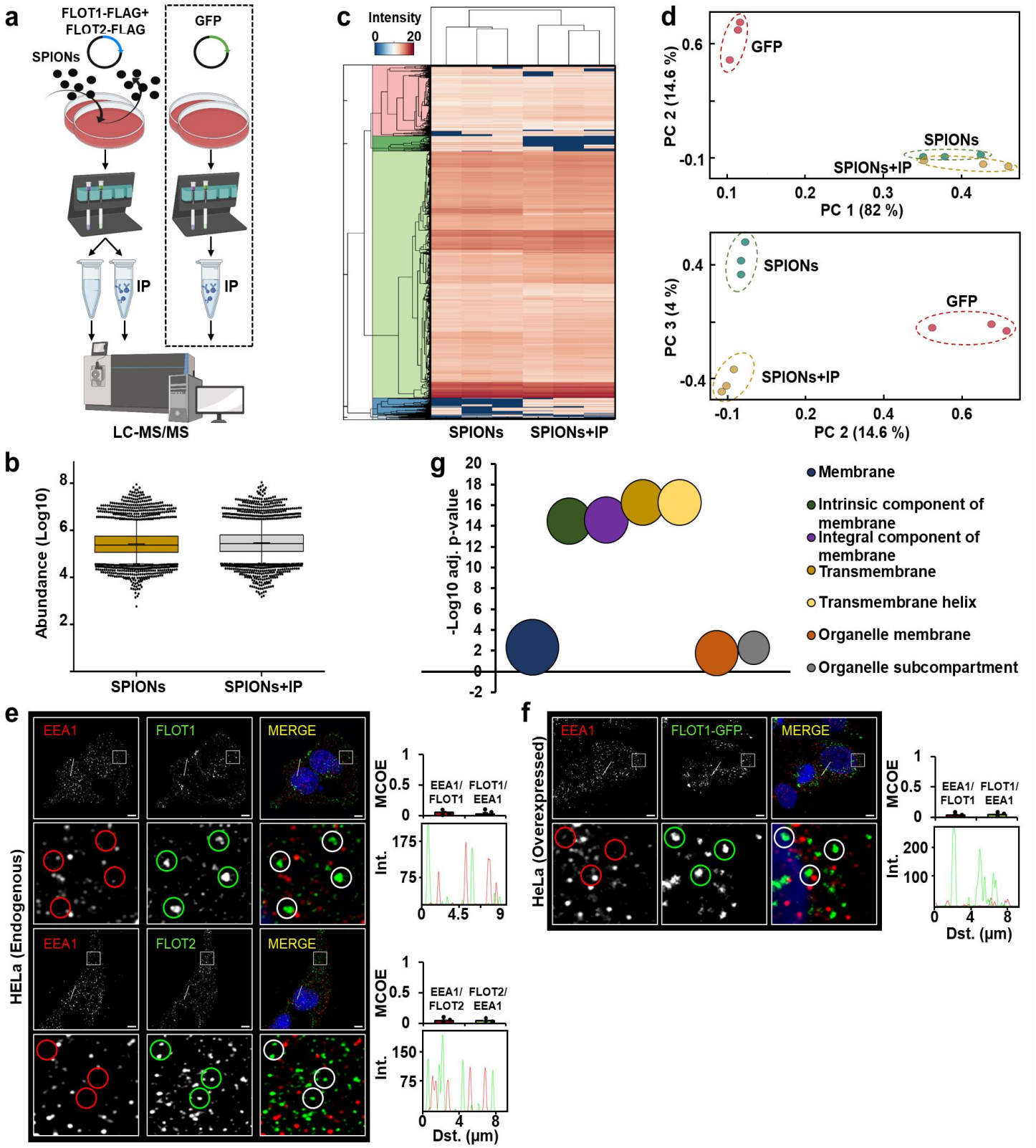
MSA: multiple sequence alignment; plDDT: predicted local distance difference test; PAE: predicted aligned error.



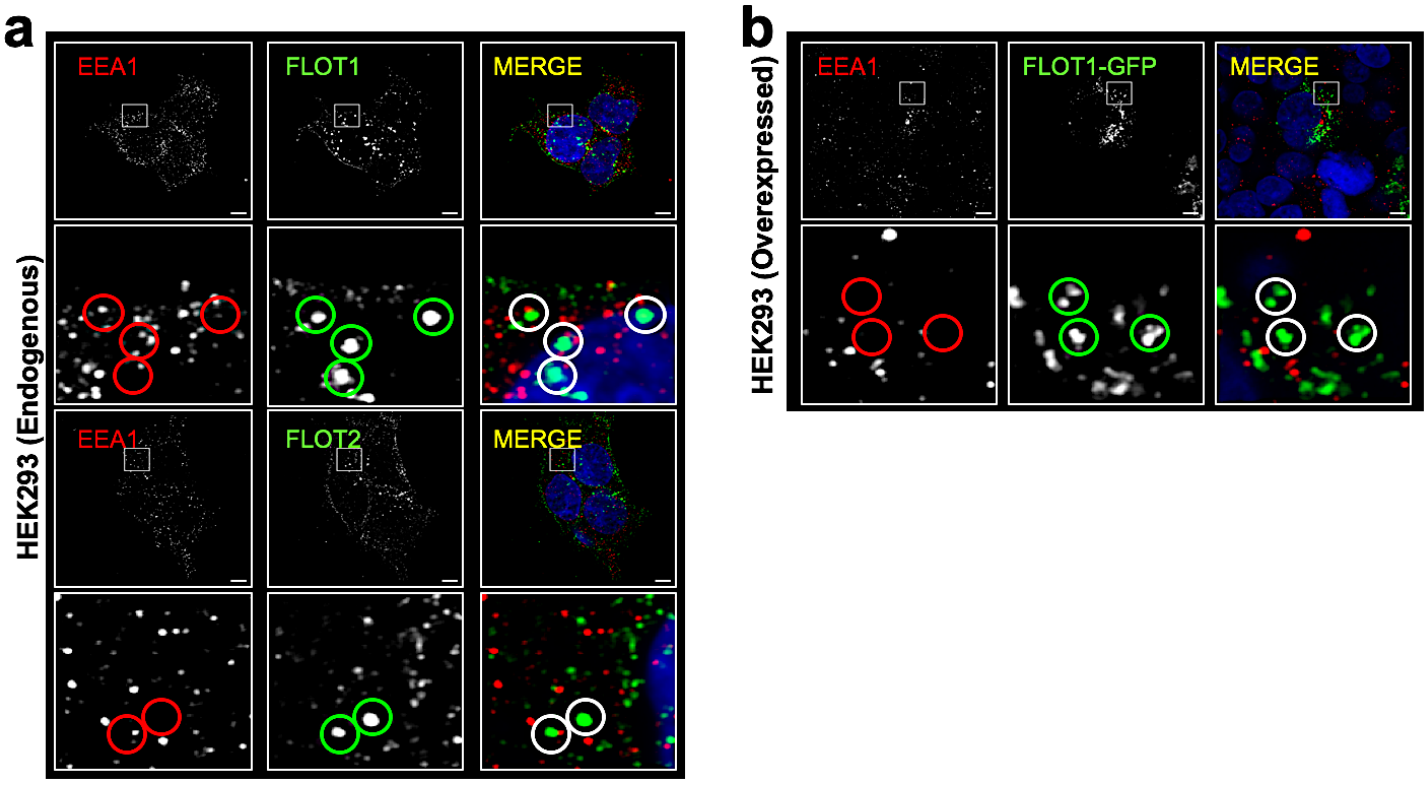
Supplementary Fig. 5: Investigation of FLOT1/FLOT2 distribution in HEK293 cells.
a Immunostaining of HEK293 cells for endogenous FLOT1, FLOT2, and the lysosomal marker LAMP2 (n=3). **b** Immunostaining of HEK293 cells transfected with FLOT1-GFP and/or FLOT2-mCherry in combination with staining for endogenous LAMP2. Images were acquired by confocal microscopy (n=2). The magnified region in the lower panel is indicated in the full-sized image. Scale bar = 5 µM.


a
HeLa (Overexpressed)
LAMP2
FLOT1-GFP
MERGE
FLOT2-mCherry
LAMP2
MERGE
FLOT2-mCherry
FLOT1-GFP
MERGE
LAMP2/ FLOT1
FLOT1/ LAMP2
MCOE
Int.
LAMP2/ FLOT2
FLOT2/ LAMP2
FLOT2/ FLOT1
FLOT1/ FLOT2
Dst. (μm)
b
IN SN EL IN SN EL IN SN EL kDa
48
FLOT1-FLAG
FLOT2-FLAG
FLOT1+2-FLAG
c
1 0.75 0.5 0.25 0.1 M C DSSO mM
kDa
180
100
70
55
40
35
25
FLOT1-FLAG+FLOT2-FLAG
d
1 0.75 0.5 0.25 0.1 M C DSSO mM
kDa
180
100
70
55
40
35
25
e
FLOT1+2-FLAG
GFP
IN FT WA EL IN SN WA EL kDa
48
FLAG
f
Flotillin
Other

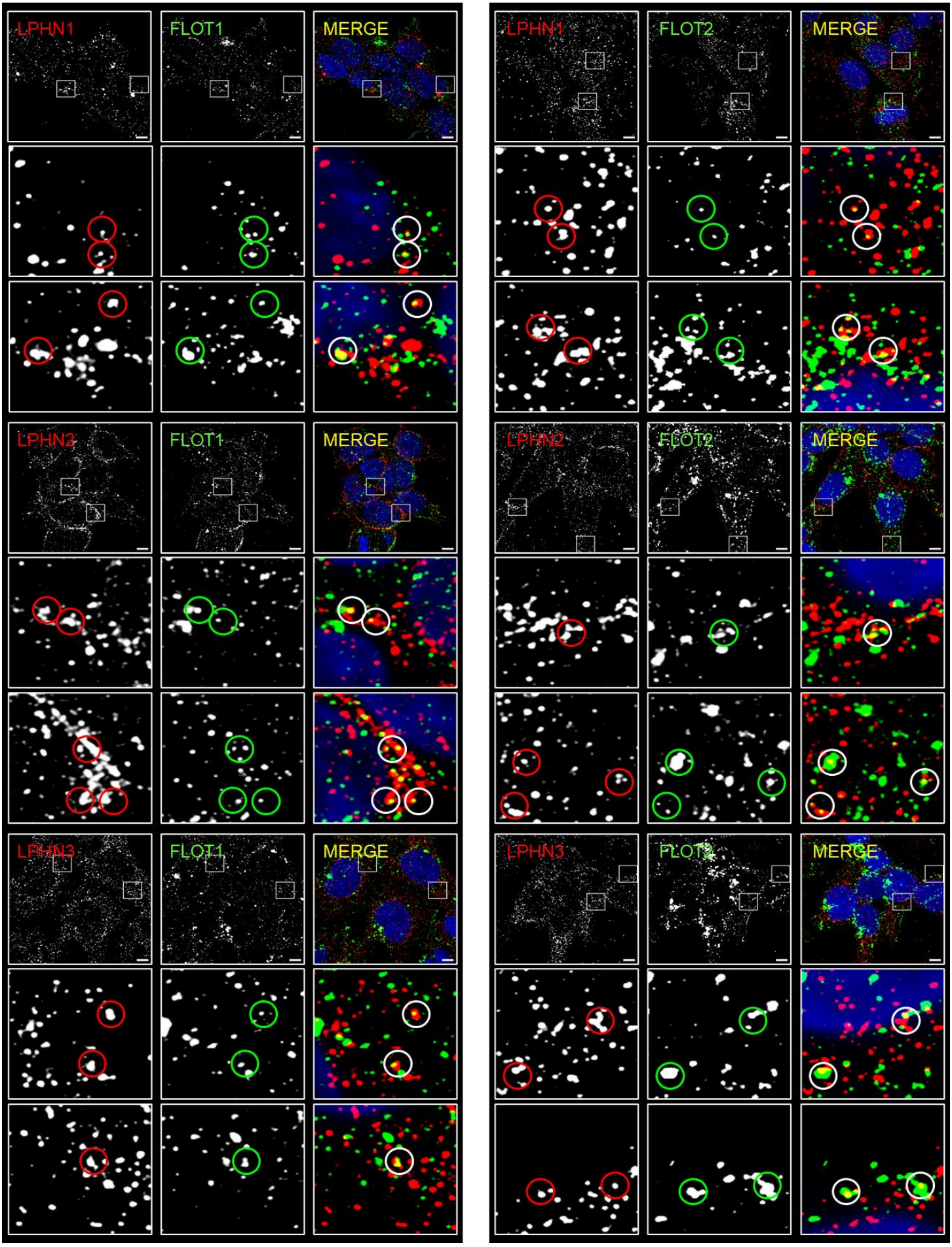
Supplementary Fig. 6: Investigation of higher order FLOT1/FLOT2 assemblies and XL-LC-MS/MS analysis of early endosomes.
a Immunostaining analysis of HeLa cells transiently transfected with FLOT1-GFP and/or FLOT2-mCherry constructs in combination with endogenous LAMP2. The magnified region in the lower panel is indicated in the full-size image. Mander's coefficients show the average degree of overlap of one protein population relative to another. Data are presented as mean values +SD (n=3, independent samples examined over one experiment). Profile plots indicate the degree of co-localization for individual vesicles. Scale bar = 5 µM. **b** Western blot analysis for verification of FLOT1-FLAG, FLOT2-FLAG, and FLOT1-FLAG+FLOT2-FLAG expression and enrichment by FLAG-IP (n=6). **c** Western blot analysis for determination of the optimal DSSO concentration for the cross-linking of SPIONs-enriched early endosomes from FLOT1-FLAG+FLOT2-FLAG transfected HEK293 cells (n=2). Early endosome-enriched fractions were cross-linked with the indicated amounts of DSSO at a protein concentration of 1 µg/µL followed by SDS-PAGE and anti-FLAG western blot. For each sample, 5 µg of protein were loaded to the gel. The non-cross-linked control sample (c) was treated with DMSO. Efficient cross-linking of FLOT1/FLOT2 complexes was judged by detection of the anti-FLAG band pattern. These analyses revealed that application of > 0.25 mM DSSO resulted in a near-quantitative loss of the FLOT1/FLOT2 monomers (~55 kDa) with almost exclusive detection of higher molecular weight assemblies (>180 kDa) and increasing amounts of protein aggregates which did not enter the separation gel. **d** Silver staining-based determination of the optimal DSSO concentration for the cross-linking of SPIONs-enriched early endosomes from FLOT1-FLAG+FLOT2-FLAG transfected HEK293 cells (n=1). Samples were generated and cross-linked as described in (c) followed by SDS-PAGE and silver staining of the gel. For each sample, 2.5 µg were loaded. The optimal DSSO concentration was determined to be 0.25 mM (final concentration), as the application of higher DSSO amounts resulted in reduced intensities in the low molecular weight region, indicating possible over cross-linking which could result in protein aggregates. Therefore, in accordance to the western blot results (see c), 0.25 mM DSSO was chosen. **e** Verification of presence of FLOT1 and FLOT2 in SPIONs-enriched early endosomes from FLOT1-FLAG+FLOT2-FLAG overexpressing HEK293 cells prior to the XL-LC-MS/MS experiment (n=6). Loading amount: 10 % of each fraction. **f** Interaction networks based on cross-links identified in the XL-LC-MS/MS dataset for DSSO-treated early endosome-enriched fractions from FLOT1-FLAG+FLOT2-FLAG-overexpressing HEK293 cells. In total, 324 inter-links were identified. FLOT1 and FLOT2 are highlighted in orange. The network was generated using Cytoscape. DSSO: disuccinimidyl sulfoxide; DMSO: dimethyl sulfoxide; PNS: post nuclear supernatant; M: marker; IN: input; SN: supernatant; FT: flow through; WA: wash; EL: eluate; C: control (no DSSO); IP: immunoprecipitation; BN-PAGE: blue native polyacrylamide gel electrophoresis; GFP: green fluorescent protein.


a
FLOT1-FLAG+
FLOT2-FLAG
SPIONs
GFP
IP
IP
LC-MS/MS
c
Intensity
0
20
SPIONs
SPIONs+IP
d
GFP
SPIONs
SPIONs+IP
PC 2 (14.6 %)
PC 1 (82 %)
SPIONs
GFP
SPIONs+IP
PC 3 (4 %)
PC 2 (14.6 %)
b
Abundance (Log10)
SPIONs
SPIONs+IP
g
-Log10 adj. p-value
Membrane
Intrinsic component of membrane
Integral component of membrane
Transmembrane
Transmembrane helix
Organelle membrane
Organelle subcompartment
e
HELa (Endogenous)
EEA1
FLOT1
MERGE
EEA1
FLOT2
MERGE
MCOE
EEA1/ FLOT1
FLOT1/ EEA1
EEA1/ FLOT2
FLOT2/ EEA1
Int.
Dst. (µm)
f
HeLa (Overexpressed)
EEA1
FLOT1-GFP
MERGE
MCOE
EEA1/ FLOT1
FLOT1/ EEA1
Int.
Dst. (µm)

Supplementary Fig. 7: Determination of flotillin vesicular cargo.
a Workflow for the enrichment of early endosomes by SPIONs and the subsequent enrichment of FLOT1-/FLOT2-positive early endosomal subpopulations by FLAG-IP. HEK293 cells were double-transfected with FLOT1-FLAG+FLOT2-FLAG, control cells with GFP. All samples were analyzed by LC-MS/MS using DIA. **b** Analysis of protein abundances in DIA datasets of SPIONs-enriched early endosomes and FLAG-IP-enriched FLOT1-/FLOT2-positive early endosomes. Data are presented as box-whisker plots (n=3, biologically independent samples over 3 independent experiments). Each box depicts the interquartile range (IQR, the range between the 25th and 75th percentile, 2.5% +/- whisker, median=line, average=+). **c** Unsupervised clustering of protein abundances (n=3, biologically independent samples over 3 independent experiments) for SPIONs-enriched early endosomes and FLAG-IP-enriched FLOT1-/FLOT2-positive early endosomes. Color coding correlates with the intensity of individual proteins. **d** PCA for all SPIONs-enriched early endosomes, FLAG-IP-enriched FLOT1-/FLOT2-positive early endosomes, and GFP-transfected control cells. Four individual principal components were used (PC1 and PC2, as well as PC3 and PC4). **e** Staining of endogenous levels of FLOT1, FLOT2, and EEA1 by confocal microscopy and investigation of co-localization. **f** Transient transfection of HeLa cells using FLOT1-GFP and co-staining with endogenous EEA1. **e/f** Images were acquired using confocal microscopy. The magnified region in the lower panel is indicated in the full-sized image. Mander's coefficients show the average degree of overlap of one protein population relative to another. Data are presented as mean values +SD (n=3, independent samples examined over one experiment). Profile plots indicate the degree of co-localization for individual vesicles. Scale bar = 5 µM. **g** GO enrichment analysis applying STRING for proteins which are overrepresented in FLOT1-/FLOT2-positive early endosomes. Results from the GO-category cellular component, as well as the UniProt keywords transmembrane and transmembrane helix are shown. Bubble size correlates with the number of proteins assigned to an individual category. Shown are p-values corrected for multiple testing within each category using the Benjamini–Hochberg procedure (cut off < 0.05) representing the significance of enrichment (n=3, biologically independent samples over 3 independent experiments).
SPIONs: superparamagnetic iron oxide nanoparticles; DIA: data independent acquisition; IP: immunoprecipitation; PCA: principal component analysis; adj.: adjusted; GO: gene ontology.



Supplementary Fig. 8: Investigation of EEA1 colocalization with FLOT1/FLOT2 in HEK293 cells. **a** Endogenous levels of FLOT1, FLOT2, and EEA1 were analyzed. The images were acquired using confocal microscopy (n=3). **b** HEK293 cells were transiently transfected using FLOT1-GFP and analyzed in combination with endogenous EEA1 staining. The images were acquired using confocal microscopy (n=2). The magnified images shown in the lower panel are indicated in the full-sized image. Scale bar = 5 µM.


LPHN1
FLOT1
MERGE
LPHN1
FLOT2
MERGE
LPHN2
FLOT1
MERGE
LPHN2
FLOT2
MERGE
LPHN3
FLOT1
MERGE
LPHN3
FLOT2
MERGE

Supplementary Fig. 9: Latrophillins co-localize with FLOT1 and FLOT2 in punctate stainings. HEK293 cells were stained for FLOT1 and FLOT2 in combination with LPHN1, LPHN2, or LPHN3. Images were acquired using confocal microscopy (n=3). The magnified images shown in the lower panel are indicated in the full-sized image. Scale bar = 5 µM.

Supplementary References

1. Thelen, M., Winter, D., Braulke, T., Gieselmann, V. SILAC-Based Comparative Proteomic Analysis of Lysosomes from Mammalian Cells Using LC-MS/MS. *Methods Mol. Biol.* **1594**, 1–18 (2017).
2. Singh, J., Ponnaiyan, S., Gieselmann, V. & Winter, D. Generation of Antibodies Targeting Cleavable Cross-Linkers. *Anal. Chem.* **93**, 3762-3769 (2021).
3. Schwanhausser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337-342 (2011).
4. Akter, F., Ponnaiyan, S., Kögler-Mohrbacher, Bi., Bleibaum, F., Damme, M., Renard, B.Y., Winter, D. Multi cell line analysis of lysosomal proteomes reveals unique features and novel lysosomal proteins. *bioRxiv* (2020).
5. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

Appendix D

Paper Four: XLEC: Large-scale prediction and modeling of protein-protein interactions by combining sequence co-evolution and cross-linking data

XLEC: Large-scale prediction and modeling of protein-protein interactions by combining sequence co-evolution and cross-linking data

Hadeer Elhabashy [a,b,c*], Eugen Netz[b,c], Oliver Kohlbacher [b,c,d*]

[a] Department of Protein Evolution, Max Planck Institute for Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany
[b] Applied Bioinformatics, Department of Computer Science, University of Tübingen, Sand 14, 72076 Tübingen, Germany
[c] Institute for Bioinformatics and Medical Informatics, University of Tübingen, Sand 14, 72076 Tübingen, Germany
[D] Institute for Translational Bioinformatics, University Hospital Tübingen, Schaffhausenstraße 77, 72076 Tübingen, Germany

Abstract

Protein-protein interactions (PPIs) mediate almost all biological processes. The majority of large-scale experimental and computational approaches for PPI prediction do not provide structural information and suffer from low accuracy. Here we introduce an integrative approach, XLEC, combining data from cross-linking mass spectrometry (XL-MS) and evolutionary coupling (EC) for the prediction and modeling of PPIs. While EC derived from multiple sequence alignments primarily yield information on direct contacts between proteins across the interface, XL-MS data preferentially captures longer-range interactions, hence these methods contain complementary information. XLEC integrates information from both approaches in a machine learning-based model and subsequent restraint-based modeling of the complex structure. We applied XLEC to data from murine mitochondrial proteomes and compared its performance to that of XL-MS and EC separately. Our assessment suggests that XLEC outperforms XL-MS or EC-based identification of PPIs. Furthermore, XLEC achieved higher rates in generating models within the acceptable, medium and highly accurate CAPRI categories. Using XLEC, we generated more than 400 *de novo* PPI models revealing novel insights into the mitochondrial interactome.

Keywords: Protein-protein interaction, interaction prediction, interaction modeling, cross-linking mass spectrometry, sequence co-evolution, evolutionary couplings.

* Corresponding author
Email addresses: hadeer.elhabashy@tuebingen.mpg.de (Hadeer Elhabashy), eugen.netz@uni-tuebingen.de (Eugen Netz) , oliver.kohlbacher@uni-tuebingen.de (Oliver Kohlbacher).

Introduction

Cellular function relies on molecular machinery formed by protein-protein interaction (PPI) and referred to as the interactome. Characterizing such cellular interactomes would be crucial to understand physiological function and to enable interventions with efficient therapeutics in the case of disease. Several high-throughput approaches for the characterization of PPIs have been developed under diverse functional principles and yet our knowledge of the cellular interactome is far from complete due to their high degree of heterogeneity.[1] Experimental approaches (such as two-hybrid[2–4], affinity purification mass spectrometry[5], protein microarrays[6,7], protein display[8,9] and cross-linking mass spectrometry[10–12]) have been essential in providing a large body of pairwise interaction data. These approaches are complemented by computational approaches, such as gene proximity[13] , gene fusion[14,15], gene co-expression[16–18], phylogenetic profiling[19,20], and co-evolution analysis. [21,22] Common to both computational and experimental methods is in general a low accuracy, which can be as low as 50%.[23] Among the many different approaches, Cross-linking Mass Spectrometry (XL-MS) and Evolutionary couplings (EC) are distinct in their ability to also reveal structural information about these interactions. Furthermore, it has been reported that the data generated by these approaches are complementary and have previously been exploited to improve protein fold determination. [24,25]

XL-MS provides information on PPIs by covalently cross-linking proteins under near-native conditions by analyzing the resulting cross-linked peptides (after digest) via mass spectrometry.[10–12] By identifying the exact position of the cross-links in the protein sequence, XL-MS can reveal which residues are in physical proximity in the complex and thus provide structural information at the residue level. XL-MS can be applied to a single purified protein complex, but also at the large scale of whole cells. The limited dynamic range of mass spectrometers is a limiting factor in the detection of cross-links and usually results in the detection of only the most abundant cross-links.[26] Also, the identification of the cross-links based on their mass spectra with sufficient statistical reliability requires sophisticated algorithms. The interpretation of the resulting data is complicated by the fact that physical proximity does not necessarily imply direct physical interaction.[27] For homo-oligomeric complexes, the identification of the correct complex topology is difficult due to the sequence identity between the interaction partners. The sparsity of the XL-MS data often results in ambiguity when reconstructing the complex structure.

Interacting proteins tend to co-evolve at the interaction sites via correlated, compensatory mutations to maintain the interaction. Evolutionary couplings (EC) is a computational approach that makes use of this concept by utilizing multiple sequence alignments to extract signals of correlated mutations from deep and divergent multiple sequence alignments. Numerous algorithms have been developed to enable efficient coupling calculation.[28–35] Strong co-evolution signals between proteins hint at direct physical interactions and reveal residue-residue contacts across the interface. Evolutionary couplings showed comparable performance to Yeast Two-Hybrid experiments plus the added feature of providing residue level resolution to the interface contacts.[22] The robustness and efficiency of current EC implementations enable proteome-wide screening of PPIs. Current limitations of EC-based PPI analysis are: I) PPIs involving orphan proteins are hard to infer due to lack of homologous sequences, II) not all interacting protein pairs show strong co-evolutionary signals (e.g., for transient PPIs), III) differentiating between intra-chain and inter-chain signals of co-evolution in the case of homo-oligomeric complexes is not straightforward.

Here we present XLEC (Cross-linking and Evolutionary Coupling), an integrative approach for large-scale prediction and structural modeling of PPIs. Combining data from both approaches boosts the confidence in discovering new protein interactions via evidence of

physical proximity within native conditions and signals of sequence co-evolution.

The complementarity between the identified restraints from XL-MS and EC reveals different structural aspects of the interaction ranging from the topological arrangement of the interacting domains to the key interacting residues across their physical interface. We systematically investigated residue co-evolution for a cross-linked interactome of murine mitochondria. We performed a comparative analysis between the integrative approach XLEC and the individual approaches (XL-MS and EC) to compare their abilities to predict and model PPIs. An iterative docking approach was used to eliminate false positives and enrich true interactions based on hybrid signal consistency and consequently enhance the quality of the generated interaction models. We used the structural features of the generated models to guide the interaction prediction on the geometrical consistency between the hybrid signals. Using XLEC, we generated more than 400 *de novo* PPI models containing novel insights into the mammalian mitochondrial interactome.

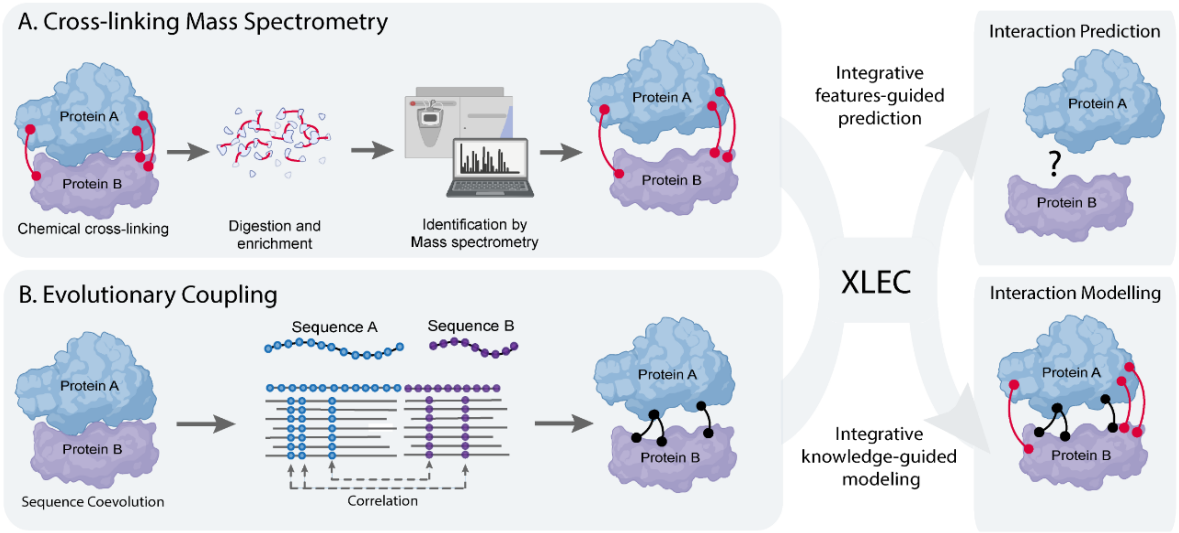


Figure 1: Schematic illustration of the XLEC approach.

XLEC integrates data from cross-linking and evolutionary coupling to provide consensus evidence for PPI discovery and a complementary set of distance restraints that can be effectively utilized by modeling approaches. A. Conceptual illustration of the generic XL-MS pipeline. B. Conceptual illustration of how interacting proteins exhibit patterns of sequence co-evolution that can be extracted using EC approaches from an MSA.

Results

We applied XLEC, as illustrated in Fig. 1, on the interactome of mouse mitochondria from two previously reported XL-MS experiments.[36,37] The combined dataset consists of 2,616 inter-protein cross-linked peptide pairs covering 324 proteins and 770 protein pairs. Supplementary Fig. S.1 shows the distribution of the number of cross-links per protein pair. The highest number of cross-links per complex in the dataset is 65, whereas many of the complexes (~42%) have only one identified cross-link. Out of 324 protein sequences, 238 have available three-dimensional structures including 41 PDB structures and 197 homology models. Protein sequences that did not have any structural hit or a homology modeling

template were not considered. Out of the 770 cross-linked protein pairs, 104 complexes were found to have experimentally determined complex structures and only 82 of these were found to have a physical interface of at least ten inter-chain contacts with less than 5 Å heavy atom distances.

From those complexes, we created a benchmark dataset of interactions with known complex structures that have sufficient multiple sequence alignments for evolutionary coupling calculations and at least one structurally mapped cross-link. This resulted in a benchmark dataset of 85 protein pairs including 67 PPIs with physical interfaces and 18 protein pairs in complexes but without direct interactions (i.e., proteins that are part of the same multimeric complex and thus in proximity, but not forming a sufficiently large interface). On our benchmarking dataset, we evaluated the Euclidean distances for both the top 0.001 percentile of evolutionary coupling scores and all the identified and structurally mapped cross-linking restraints. About 34% of the benchmark complexes with a physical interface have more than one EC with a distance below 10 Å, while 58% of all the mapped cross-links achieved a CA-CA distance of less than 35 Å. An additional prediction set of more than 400 protein pairs with sufficient multiple sequence alignments, cross-links, and 3D structures of their subunits, but not the complexes, was also created.

We modeled the benchmark dataset using multiple restraint-based docking protocols (XL-based docking, EC-based docking, and XLEC-based docking) using Haddock 2.2. For each docking procedure, we shortlisted and reported the top ten models based on the Haddock scoring function. We also clustered the resulting 100 models based on a 5 Å pairwise RMSD cut-off and reported the top model of each cluster. Our assessment of the used approaches of XL-MS, EC, and XLEC was based on two main criteria: the ability to model the interaction and the ability to predict the interaction in that order.

XLEC-based docking provides better structural models than XL-based and EC-based docking individually

For our assessment of model quality, we evaluated the overall geometric fit of the top predicted models for 67 benchmark complexes with direct interaction to their target structures using the CAPRI assessment protocol.[38] It uses the metrics Ligand Root Mean Square Deviation (L-RMSD), Interface RMSD (I-RMSD) and Fraction of Interface Native Contacts (F_{nat}). In addition, we used the DockQ[39] score which combines the three prior CAPRI metrics into a single score ranging between 0 and 1 and can recreate the CAPRI model quality categories (Table 3). For every complex, we used the model with the highest DockQ score to build our statistics.

The XL-based docking produced CAPRI-acceptable quality models for 3.0% and with an acceptable DockQ score for 4.5% of the benchmark interaction set with no models achieving medium or high quality (Table 1). This low performance can be attributed to the length of the cross-linkers (up to 40 Å), the sparse nature as well as the low quantity of identified cross-linking restraints per complex. Despite the usage of center-of-mass restraints, which pull the proteins' center of mass closer to achieve the interface, XL restraints still permit a high degree of freedom due to sparsity and length. One of the major issues appears to be the small number of experimentally identified cross-links due to the low sensitivity when

cross-linking a whole organelle. Further fractionation, enrichments for specific protein families, or advances in the instrumentations would certainly boost these rates. At the same time, even a small number of experimentally determined cross-links will boost confidence in otherwise computationally generated models.

EC-based docking has shown better performance than XL-based docking resulting in at least CAPRI acceptable quality models and DockQ acceptable models for 17.9% and CAPRI medium quality models for 9.0% and models with DockQ medium quality ($0.49 \leq$ DockQ < 0.8) for 6% of the interacting benchmark set and no models with high quality. Compared to the large-scale cross-linking experiment of whole organelles, this implies a roughly fourfold higher success rate to model PPIs, albeit without the benefit of experimental validation. Since EC restraints represent interface contacts, they also tend to be short-range restraints across the interface and can reduce the degrees of freedom available to docking, which in turn results in a higher likelihood of obtaining good structural models.

We then compared these results for the docking based on individual methods to the XLEC approach (as described in the Methods section). In terms of CAPRI categories, XLEC generated models of acceptable quality or better for 19.4% and medium quality or better for 11.9% of the physically interacting complexes in our benchmark set. In terms of DockQ categories, XLEC generated models with at least acceptable quality for 14.9% and at least medium quality for 9.0% over the same complexes set. As can be seen from Table 1, when compared to EC-based docking, XLEC achieved a higher rate (CAPRI: 33% and DockQ 50%) of medium to higher quality structures. Similar to EC- and XL-based docking, it was not possible to achieve a significant number of high-quality structures (CAPRI: 3.0% and DockQ: 1.5%).

Fig. 2. B, 2. C and 2. D show exemplary models from all three described protocols of XL-, EC-, and XLEC-based dockings compared to their known PDB structures; I) Interaction model between subunits 2 and 7 of cytochrome b-c1 complex with respect to its PDB structure (ID: 4D6T, chains O and F). II) Interaction between the 75 kDa subunit of NADH-ubiquinone oxidoreductase and NADH dehydrogenase 1 alpha subcomplex subunit 2 of the mitochondrial NADH dehydrogenase complex with its PDB structure (ID: 6G2J, chains G and S). II) Interaction between subunits 7B and 4 of cytochrome c oxidase complex with PDB structure (ID: 2Y69, chains K and D). In these three complexes, even though the XL restraints are all consistent with the known structure, XL-based docking struggled to achieve an acceptable quality due to their sparsity. While the short-range interface contacts from EC achieved acceptable to medium-quality models. XLEC-based models showed increased model quality between medium to high accuracy.

Table 1: Assessment of prediction quality
The performance of different interaction modeling approaches qualified as protein complex counts and percentages with respect to the interaction benchmark dataset in terms of CAPRI categories and DockQ scores.

		High quality	Medium quality or better	Acceptable quality or better
XL-based docking	CAPRI	0 (0%)	0 (0%)	2 (3.0%)
	DockQ	0 (0%)	0 (0%)	3 (4.5%)
EC-based docking	CAPRI	0 (0%)	6 (9.0%)	12 (17.9%)
	DockQ	0 (0%)	4 (6.0%)	12 (17.9%)
XLEC-based docking	CAPRI	2 (3.0%)	8 (11.9%)	13 (19.4%)
	DockQ	1 (1.5%)	6 (9.0%)	10 (14.9%)

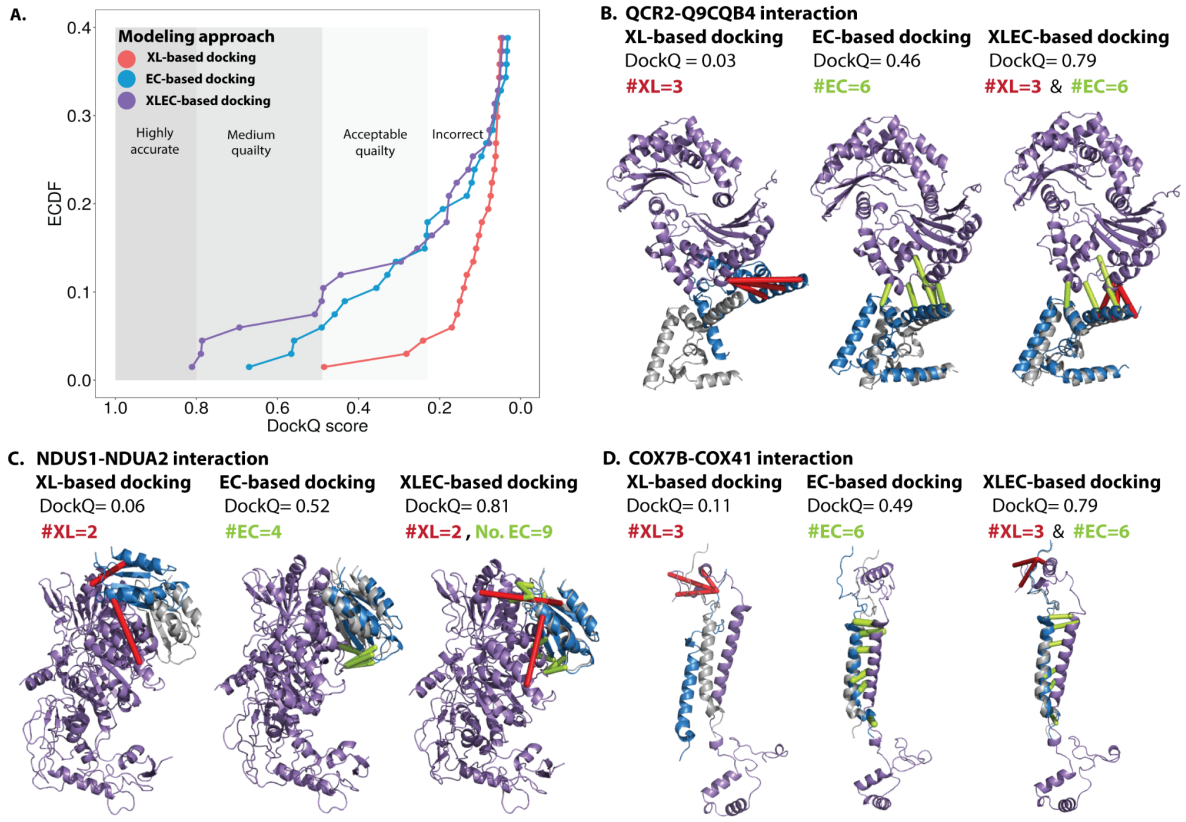


Figure 2: Benchmark set interaction models as predicted by XL-based docking, EC-based docking, and XLEC-based docking.

A. The panel shows the empirical cumulative distribution function (ECDF) versus DockQ score, illustrating the performance of the different modeling approaches on our benchmark set of complexes of known structures. B. Docked models of QCR2-Q9CQB4 interaction as compared to PDB structure (ID:4D6T, chains O and F). C. Docked models of NDUS1-NDUA2 interaction as compared to PDB structure (ID:6G2J, chains G and S). D. Docked models of COX7B-COX41 interaction as compared to PDB structure (ID:2Y69, chains K and D). For every model, we show the predicted location of the ligand subunit (in blue) compared to its equivalent subunit (in gray) from the known complex structure after the backbone atoms of receptor subunits (in purple) were superimposed. Cross-linking restraints

(XL) are shown in red and the evolutionary coupling restraints (EC) are shown in yellow.

XLEC can add confidence to AI-based structure prediction

With the recent advancements in ML-based structure prediction, there are also predictors for protein complexes, such as AlphaFold-Multimer.[40,41] Unlike the described docking protocols, AlphaFold-Multimer has access to all known structures for both training and template selection for the predictions. It is thus not a blind prediction and a fair comparison of its performance is difficult, but we tried to compare the results of the ML-based predictions to our approach to the extent possible. The results showed that AlphaFold-Multimer surpassed all the described docking protocols in modeling the benchmark dataset. Specifically, it modeled 56.7% of the complexes with acceptable quality, 49.3% with at least medium quality, and 20.9% with highly accurate quality as determined by CAPRI metrics. In terms of DockQ score, AlphaFold-Multimer generated 56.7% of the complexes with acceptable quality or better, 35.8% with at least medium quality, and 14.9% with highly accurate quality, (Fig. 3. A).

In cases where both XLEC-based docking and AlphaFold-Multimer achieve an accurate topology, AlphaFold-Multimer achieves higher DockQ scores due to the full-length flexible body modeling and side chain refinements. Fig. 3. C shows a highly accurate model (DockQ = 0.98) of interaction between alpha and beta subunits of Electron transfer flavoproteins (ETFA-ETFB) as generated by AlphaFold-Multimer, with respect to their PDB structure (ID:1EFV, chain A and B), alongside the acceptable quality model (DockQ = 0.44) generated by XLEC-based docking. Fig. 3.D shows models of the interaction between NADH dehydrogenase iron-sulfur protein 3 and 4 (NDUS3-E9QPX3) as generated by XLEC-based docking (DockQ = 0.66) and AlphaFold-Multimer (DockQ = 0.78) as compared to the PDB structure (ID: 6G2J, chains C and Q).

While AlphaFold-Multimer demonstrated excellent performance on the benchmark dataset, one cannot extrapolate these results to the interactions without 3D structures or templates in the database, as AlphaFold-Multimer models benefit from structural training and template-based modeling. This claim was substantiated by a decrease in its performance upon running AlphaFold-Multimer on the benchmark dataset without the use of templates but does not mitigate the fact that AlphaFold-Multimer was still aware of the complex structures due to its training. Based on CAPRI assessment, 49.3% of the complexes achieved acceptable quality or better, 43.3% achieved at least medium quality, and 11.7% achieved high accuracy. Based on DockQ categories, the results show 49.5% of benchmark complexes achieved acceptable quality better, 20.90% achieved at least medium quality, and 8.9% achieved highly accurate quality. In the absence of complex structures or templates, one can refer to AlphaFold-Multimer's intrinsic model quality metrics (e.g., pLDDT, pTM, ipTM, and PAE) to estimate the reliability of the resulting models.

We employed AlphaFold-Multimer to model our prediction set of more than 400 cross-linked protein pairs that have no 3D structures in the public databases. For only ~5% of the prediction dataset, AlphaFold-Multimer achieved a pTM-based confidence score of more than 0.5 in comparison to 40% in the case of the benchmark dataset. Fig. 3. B shows a

comparison between the distributions of an intrinsic model quality measure, the pTM-based confidence score, generated by AlphaFold-Multimer for both the benchmark set in which the complexes have known structures in the PDB and the prediction set with unknown complex structures at the time of preparing this manuscript. We would thus expect the performance of ML-based predictions to be significantly lower than the performance we observe using XLEC. In addition, the inclusion of XL-based data that are part of the XLEC approach would further validate the computationally generated models. While we cannot readily compare the prediction quality of XLEC and ML-based predictions, we show that their integration will turn out to be beneficial in the future.

From our prediction set, we generated more than 400 *de novo* PPI models revealing novel structural insights into the mitochondrial interactome. Fig. 5. A and Fig. 5. B show models of XLEC in agreement with models from AlphaFold-Multimer for the interaction between DNA topoisomerase I (TOP1M) and electrogenic aspartate/glutamate antiporter (SLC25A13) and the interaction between the heat shock chaperone (TRAP1) and the stress-70 chaperone (GRP75). Fig. 5. C and Fig. 5. D show different structural models between XLEC and AlphaFold-Multimer for the interaction between alanine aminotransferase 2 (ALAT2) and Hexokinase-1 (HXK1) and the interaction between cytochrome b-c1 complex subunit 2 (QCR2) and aconitate hydratase (ACON). Supplementary Fig.S. 3 shows further *de novo* interaction models generated using XLEC-based docking.

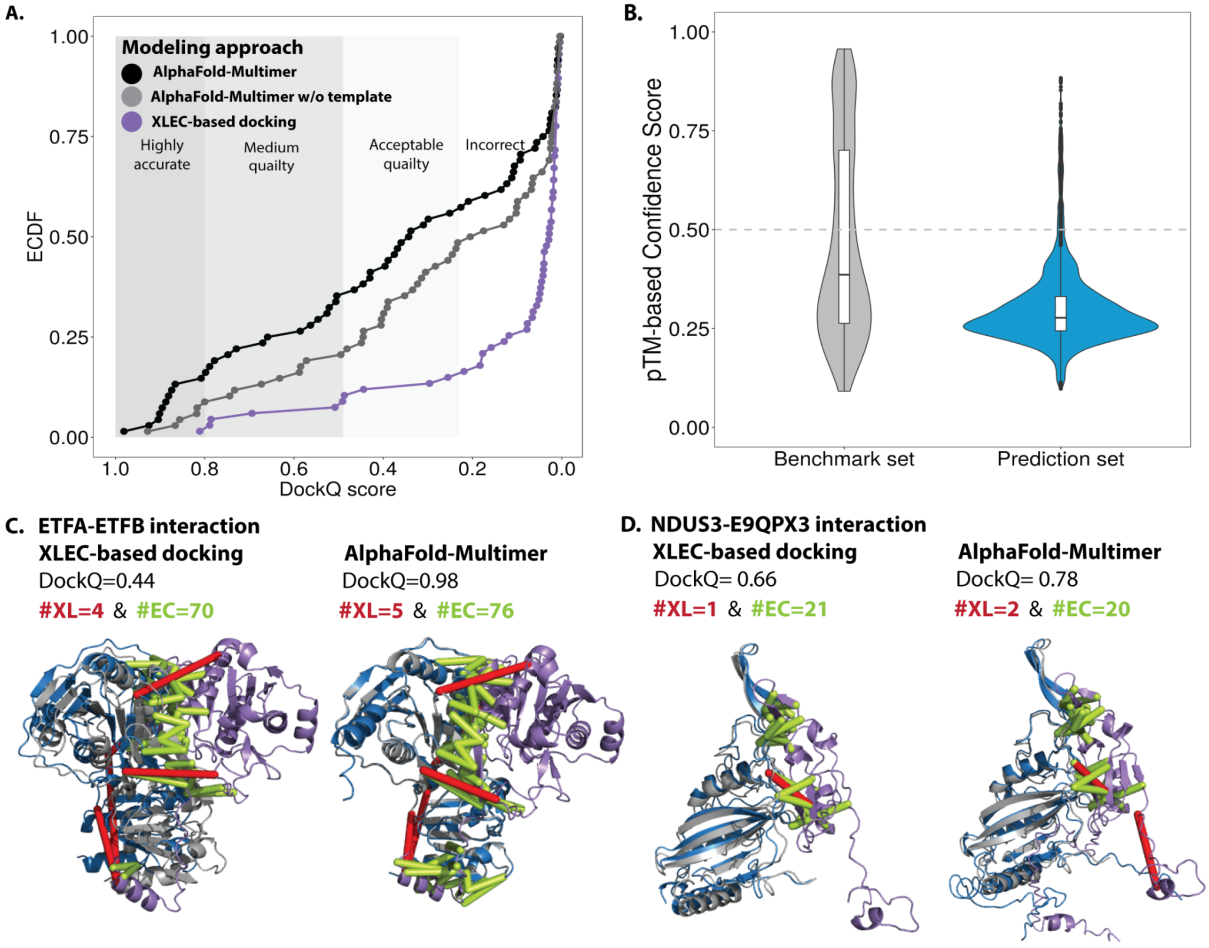


Figure 3: Comparison of XLEC to AlphaFold-Multimer predictions.
A shows the empirical cumulative distribution function (ECDF) versus DockQ score, illustrating the performance of the different modeling approaches including XLEC-based docking, AlphaFold-Multimer, and AlphaFold-Multimer without the usage of templates on our benchmark dataset with complexes of known structures. **B** shows a violin plot of the distribution of the pTM-based confidence scores evaluating the intrinsic qualities of the models generated by AlphaFold-Multimer of the benchmark set (in gray) and prediction set (in blue). **C.** Models of the interaction between the electron transfer flavoprotein alpha and beta (ETFA-ETFB) generated by XLEC-based docking and AlphaFold-Multimer as compared to PDB structure (ID 1EFV, chains A and B). **D.** Models of the interaction between the NADH dehydrogenase iron-sulfur proteins 3 and 4 (NDUS3-E9QPX3) generated but XLEC-based docking and AlphaFold-Multimer as compared to PDB structure (ID 6G2J, chains C and Q).

Integrating XL and EC data improves interaction discovery

Besides the prediction of 3D structures, XLEC could be employed to predict whether a pair of proteins interact at all, which is an essential question in molecular biology. Using the proximity between cross-linked proteins to imply direct physical interactions might lead to false-positive identification of biologically non-relevant physical interactions. That is because cross-linkers may often connect proximate proteins that are not necessarily physically interacting.To this end, XLEC includes an ML-based model to better classify physically interacting from non-interacting proteins in XL-MS data. Similar inferences are possible with XL-MS and EC individually, so we assessed the performance of an integrated approach (XLEC) to the individual approaches (XL-MS, EC) in this context as well. The prediction models are logistic regression models based on features from cross-linking and evolutionary couplings individually. For the XLEC model, these are combined with and complemented by structural features including the number of satisfied XL and EC restraints when mapped on the generated models and metrics of intrinsic model quality. Table 4 shows the features utilized by each model. Five-fold cross-validation was used to evaluate the performance of the models (Table 2).

Not surprisingly, we observed that the integrated approach as implemented in XLEC outperforms both XL- and EC-based predictors. With XL-MS data, the reason is often that cross-linking can also occur between - even temporarily - close proteins that are not directly interacting. In 22% percent of the structures in the benchmark dataset, we actually saw cases where the two proteins do not have a significant physical interface, but are still part of a larger molecular complex and thus in proximity. Other factors likely contributing to the low performance of XL-based predictions are the sensitivity in XL-MS and the accessibility of potential cross-link sites.

Similarly, not all interacting proteins show signals of co-evolution. Orphan proteins that do not belong to a functionally characterized protein family may lack the depth of the sequence alignment to extract reliable co-evolution signals. Interactions with fast evolutionary rates (e.g. transient interactions) tend to yield weaker co-evolution signals. Therefore, the EC-based predictor has a high true negative rate of unpredicted protein interactions. Combining evidence of interaction from both approaches of XL and EC alongside the structural features of the models results in slightly higher performance with an AUC of 0.73

compared to 0.64 (XL alone) and 0.68 (EC alone). These results clearly show that integrating XLEC and ML-based methods harnesses a synergy for enhanced prediction of PPIs.

Integration of XLEC with AlphaFold-Multimer further improves interaction prediction

Since XLEC modeling features appeared to improve our ability to predict interactions, we investigated whether incorporating AlphaFold-Multimer's structural features and intrinsic model quality would similarly enhance our predictions. For this we trained a logistic regression model on features from AlphaFold-Mutlimer combined with XLEC to classify physically interacting from non-interacting cross-linked proteins. The training features included pTM-based confidence score, number of identified XL restraints, maximum EC scores, EC statistics, and the numbers of EC and XL restraints that achieve a maximum of 10 Å and 30 Å respectively. We trained another model only on the pTM-based confidence scores of AlphaFold-Mutlimer to be used as a baseline model for comparison. Table 4 shows the features utilized by each model. AlphaFold-Multimer-based classifier slightly underperforms the XLEC classifier achieving ROC-AUC 0.72. The best predictor combined AlphaFold-Multimer as well as XLEC features achieving ROC-AUC of 0.84 (Fig. 4. B). The consensus between cross-linking restraints, evolutionary couplings features, and an accurate model that satisfies both of them, made a clear difference between interacting and non-interacting protein pairs.

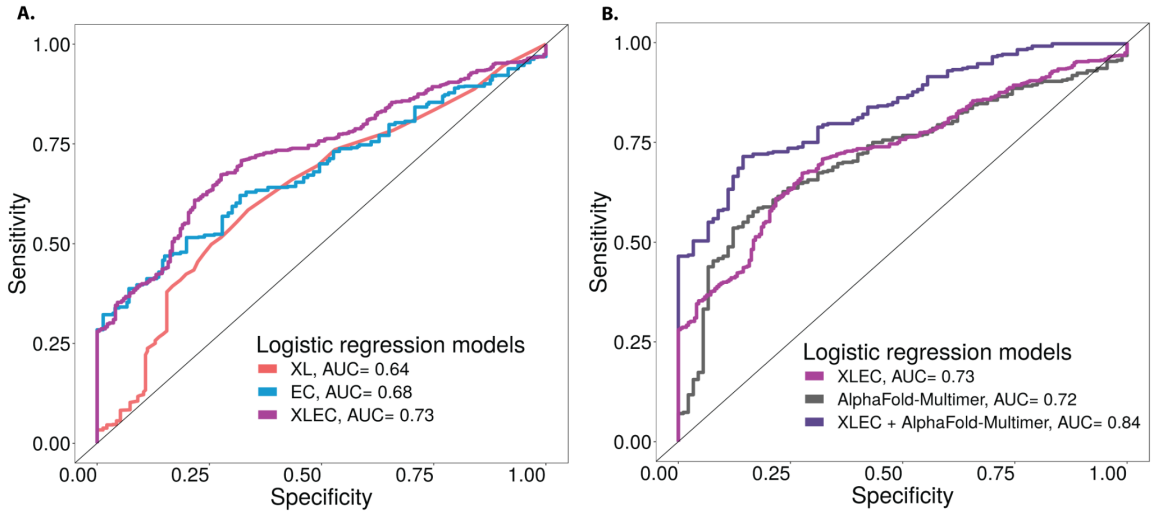


Figure 4: The performance of the trained interaction predictors
A. The Receiver Operating Characteristic (ROC) curves for the developed logistic regression classifier of protein interaction and their associated Area Under the Curve (AUC) for XL, EC and XLEC based predictors. **B.** shows the ROC curves with their corresponding AUC of XLEC, AlphaFold-Mulitmer, XLEC + AlphaFold-Mulitmer based predictors.

Table 2: Assessment of interaction prediction

The performance of different predictors to classify between interacting and non-interacting proteins in a five-fold cross-validation.

	XL	EC	XLEC	AlphaFold-Multimer	XLEC + AlphaFold-Multimer
ROC-AUC	0.64	0.68	0.73	0.72	0.84

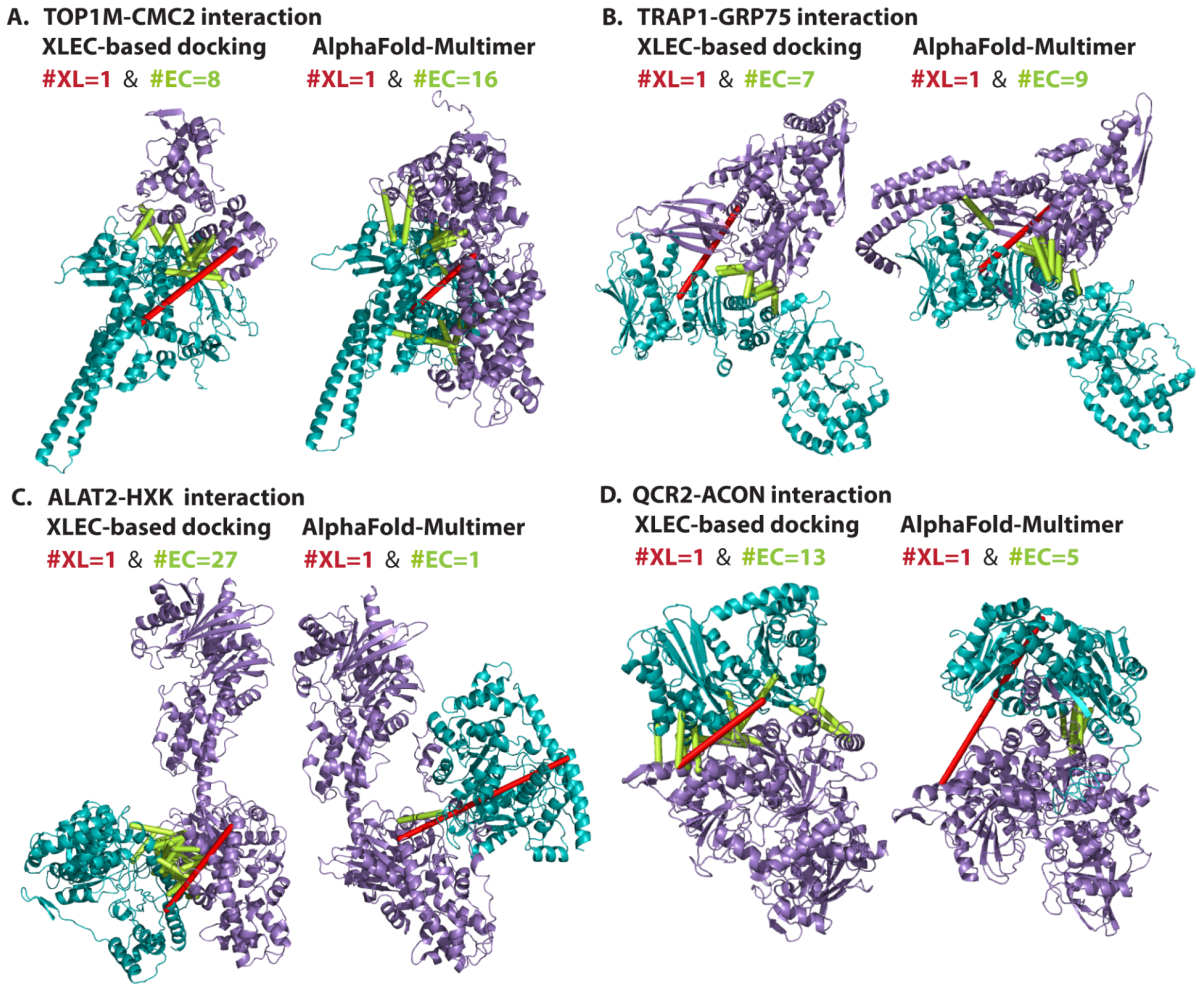


Figure 5: Prediction set interaction models as predicted by XLEC-based docking and AlphaFold-Multimer.

Interaction models from the prediction set as predicted by XLEC-based docking and AlphaFold-Multimer. For every model, we show the predicted location of ligand subunit (in teal) and receptor subunits (in purple). Cross-linking restraints (XL) are shown in red and the evolutionary coupling restraints (EC) are shown in yellow. **A.** Interaction model of the DNA topoisomerase I (TOP1M) and electrogenic aspartate/glutamate antiporter (SLC25A13/CMC2). **B.** Interaction model between the heat shock chaperone (TRAP1) and the stress-70 chaperone (GRP75). **C.** Interaction model between the alanine aminotransferase 2 (ALAT2) and Hexokinase-1 (HXK1). **D.** Interaction model between cytochrome b-c1 complex subunit 2 (QCR2) and aconitate hydratase (ACON).

Discussion

Integrating experimental data with computational predictions has a long-standing tradition in structural biology. With the advances in machine learning-driven approaches and the availability of large amounts of sequences, these computational predictions yield reliable structures by themselves but also enable the use of sparse experimental data for structural modeling. Combining experimental and computational approaches is also attractive since they capture complementary aspects of protein structures and interactions. We presented a novel approach, XLEC, that can improve the prediction of PPIs over individual experimental and computational approaches through such an integrative approach.

XLEC can steer PPI discovery towards higher confidence due to the consensus between native XL-MS and EC evidence. This performance further increases when using the geometrical information from structural models generated by knowledge-based docking or AlphaFold-Multimer. Either for the purpose of interaction prediction or modeling, XLEC can be implemented in two ways: By adding evolutionary couplings calculation to data generated by XL-MS, or by combining the results of two independent systematic interactome screens by the two approaches.

In the same way that XLEC combines the advantages of both XL-MS and EC, XLEC also inherits some of their limitations. The low sensitivity of XL-MS and the lack of sequence information for EC limit the applicability of integrated methods as well. Another limitation stems from the application of rigid body docking. This requires on one hand the availability of good structures of the individual subunits to be docked, on the other hand, assumes only limited structural changes upon binding.

In future approaches, we thus see ample room for improvement. For instance, we believe that the establishment of accessible databases of evolutionary couplings and cross-linking data would leverage data-driven prediction and modeling of protein-protein interaction and would enable further experimental validation of the discovered interaction. Also, experimental improvements will have an impact. In our data, a substantial number of protein pairs had only one cross-link indicating an interaction, but providing very little structural information. More sensitive XL-MS experiments will yield more structural information in the future.

While AlphaFold-Multimer has shown excellent performance in modeling complexes with known structures or templates, our analysis indicates that it has limitations for novel complex structures. Integrating the knowledge from sparse experimental data into machine learning predictions from the start might also be a promising route to not only improve predictions but also boost confidence in the computationally generated models.

Methods

Datasets

We generated a dataset of cross-linked protein pairs by combining the results of two previously reported studies on murine mitochondria.[36,37] The first study performed XL-MS on

functional murine mitochondria, using protein interaction reporter (PIR), and identified 2,437 cross-linked peptide pairs covering 327 mitochondrial proteins and 459 protein-protein interactions at false discovery rate (FDR) less than 5%.[36] The second study conducted XL-MS on intact mouse heart mitochondria, using a gas-phase cleavable cross-linker DSSO, identifying 2,040 cross-linked peptide pairs from 221 mitochondrial proteins and across 609 protein-protein interactions filtered to FDR less than 5%. [37] A non-redundant combination of these two studies resulted in a dataset of 2,616 non-redundant interchain cross-linked peptides from 324 unique proteins and across 770 unique protein pairs. Each cross-link in the dataset is represented with UniProt IDs of the cross-linked proteins, sequences of the cross-linked peptides, positions of the cross-linked residues, the study of origin, and the false discovery rate (FDR). Protein pairs that share a Pfam domain or sequence similarity of more than 30% were removed from our dataset because they often reflect intra-chain co-evolutionary signals between their similar regions. This reduced our dataset from 770 to 751 protein pairs (supplementary Table S1).

Protein and complex structures

Protein sequences were retrieved using the REST API service from the UniProt web server (release Jan. 2021).[42] Sequences were then searched against the Protein Data Bank (PDB)[43] (release Dec. 2019) using Blastp (2.9.0)[44] with an E-value of less than 0.00001. Protein structures were considered under two conditions: I) hits from the same species with more than 90% sequence identity and more than 50% alignment coverage to the query sequence and II) hits from different species with 100% sequence identity and more than 50% alignment coverage to the query sequence. Sequences with structural hits that did not achieve any of these criteria were modeled based on homology using SWISS-MODEL.[45–49] Protein sequences that did not have any 3D structures or homology model templates were not considered in this study to avoid further factors of uncertainty. Out of the 324 protein sequences, 238 have available 3D structures including 41 PDB structures and 197 homology models. A list of the accessions for the PDB structures and models is included in the supplementary Table S2.

For complex structures, we directly intersected the BlastP results of every pair in our dataset of 770 cross-linked protein pairs. About 104 protein pairs were found to have 3D structures within the same PDB file. To test if the cross-linked proteins are physically interacting, we calculated heavy atoms' Euclidean distances between all interchain permutations for each protein pair contained in the same PDB file. Protein pairs with the largest interface were then selected for further analysis and comparisons. Nevertheless, only 87 of 104 protein pairs were found to have a significant physical interface of more than ten residue-residue contacts with less than 5 Å between their nearest heavy atoms. The remaining pairs represent cross-linked protein pairs that are in the same complex but have no sufficiently large direct interface. These were used as negative examples for interaction prediction.

Benchmark dataset

We created a benchmark dataset of 85 protein pairs with 67 physically interacting proteins with more than ten heavy atom contacts across the interface (within 5 Å) and 18 protein pairs that have fewer or no interface contacts. Protein pairs of the benchmark dataset were selected based on four main criteria; I) Protein pairs must have 3D structural information in the PDB or the SWISS-MODEL for their individual protein. II) Protein pairs must have

sufficient multiple sequence alignments to be eligible for evolutionary coupling calculation. III) Protein pairs must have a minimum of 1 identified cross-link between them that can be structurally mapped on both proteins. Since cross-links between proteins that are mapped on missing regions can not be used for protein docking. IV) Protein pairs must have 3D information for their complex structure or must be found in the same PDB file in case they are not interacting (supplementary Table S3).

Prediction dataset

We created a prediction set of 417 protein pairs that satisfy the same criteria of the benchmark dataset except they have no three-dimensional structures for their complexes at the time of preparing this manuscript (supplementary Table S4).

Evolutionary couplings calculations

For evolutionary coupling calculation, we used EVcomplex 2.0.[22,50] The configuration of EVcomplex is given in supplementary File S1. In brief, EVcomplex searches each protein sequence using JackHMMER (release 3.1b2)[51] for five iterations against the UniProt database (release Jan. 2021).[52] For each sequence, different bit score thresholds are passed to JackHMMER. For each of these alignments, we picked a good (and still compact) MSA based on alignment coverage. To this end, we determined six quality categories as follows: I) N_{eff}/L is less than 100, N_{eff} is not less than 2,000, domain coverage is more than 0.2, and alignment coverage is more than 0.8. II) N_{eff}/L is less than 100, N_{eff} is not less than 2,000, domain coverage is more than 0.2, and alignment coverage is more than 0.6. III) N_{eff}/L is less than 100, N_{eff} is not less than 2,000, and domain coverage is more than 0.2. IV) N_{eff}/L is less than 100, and domain coverage is more than 0.2. V) Domain coverage is more than 0.2. VI) Domain coverage equal to 0.2. For each sequence pair, we then selected the alignment that achieved the highest (I through VI) quality category.

EVcomplex then annotates the sequences in each alignment with their origin species and reduces them to a single representative sequence per species. For this filter, the best reciprocal identity to the query sequence was the criteria after excluding closely related paralogs with more than 90% sequence identity. Each pair of protein alignments was then concatenated into one MSA by matching species. MSA columns that contain more than 50% gaps were filtered out from the alignment, and sequences that were more than 80% identical to each other were downweighed, as described in Hopf et al [50], to compute the effective number of sequences (N_{eff}). Evolutionary couplings were calculated using pseudo-likelihood maximization direct coupling analysis [33,35,53] as implemented in the EVCouplings Python package.[54] The Corrected Norm (CN) scores were used to rank the co-evolutionary behavior of interchain residue pairs. For more details about the pipeline of EVcomplex 2.0, please check the supplementary material of Green and Elhabashy et al.[22]

Solvent accessibility calculations

Unless the interaction happens upon a conformational change in the interacting subunits, interface residues are assumed to be solvent accessible.[55] For every protein with 3D information, we calculated absolute and relative solvent accessibility per residue by using Naccess V2.1.1.[56] Naccess uses the Lee & Richards[57] method to calculate the solvent

accessible surface (1.4 Å probe radius). The relative solvent accessibility is later used as a filter for evolutionary coupling restraints.

Structural modeling

We built 3D interaction models for the benchmark dataset using cross-linking (XL)-based docking and evolutionary coupling (EC)-based docking to establish the baseline performance of these approaches individually. We then geometrically evaluated the identified XL restraints on the shortlisted pool of the docked XL-based models and the EC restraints on the shortlisted pool of docked EC-based models. This evaluation enables cleaning the restraints by eliminating false positives that are unable to achieve an acceptable distance between the interacting subunits and are not consistent with the majority of satisfied restraints. This approach also offered enrichment for the evolutionary coupling signal by rescuing lower-scored EC restraints that show structural consistency with the top EC scores within the docked structural model. Using the updated set of restraints, we modeled the interaction within our benchmark dataset using our proposed integrative approach of cross-linking and evolutionary coupling (XLEC)-based docking. Finally, we used the machine learning-based approach AlphaFold-Multimer V2.1[40,41] to model the interactions within our benchmark set with the usage of templates, maximum template date 31-10-2021, and without it. For our prediction set, we generated *de novo* models using both the proposed XLEC-based docking and AlphaFold-Multimer V2.1.

For all the described restraints-based docking approaches, we used the default protocol of HADDOCK 2.2 (High Ambiguity Driven biomolecular DOCKing)[58] and CNS 1.3.[59] Every docking procedure started with rigid-body docking and energy minimization for 500 models. Of which, the best 100 models underwent semi-flexible refinement in torsion angle space followed by refinement in explicit water. For every complex, we ranked the generated 100 models using the HADDOCK scoring function and clustered them based on a 5 Å pairwise RMSD cut-off. The top ten of the docked models based on the HADDOCK scoring function and the top clusters' model as well as the resulting five models from AlphaFold-Multimer are available at (https://orinoco.cs.uni-tuebingen.de/Elhabashy/supplementary_models.zip). Specific details for every docking protocol are described in the following:

1. **XL-based docking**
   Cross-linked restraints were assigned as unambiguous restraints between C-beta atoms of the cross-linked residues, usually Lysine, with an effective distance of 15 Å with an upper and lower limit of 30 and 5 Å, respectively. Because of the long-range of the cross-links, we enabled the center of mass restraints that pull the subunits towards each other to ensure physical interaction and promote compactness of the docked solution while achieving the submitted restraints.

2. **EC-based docking**
   In this protocol, we examined the top ten EC and utilized only a maximum of the five highest-scoring EC that achieved these two criteria; I) Both residues in each restraint can be structurally mapped on the subunits, not on missing regions. II) Both residues have relative solvent accessibility of more than 20%. EC restraints were assigned as unambiguous restraints on C-beta atoms of their corresponding residues (except for

Glycine, for which C-alpha was used) with an effective distance of 5 Å with an upper and lower bound of 5 and 2 Å, respectively.

3. **XLEC-based docking**
   This protocol used three types of restraints; unambiguous XL restraints, unambiguous EC restraints, and ambiguous center-of-mass restraints. We used an updated set of XL and EC restraints upon their geometrical evaluation on the shortlisted pools of their corresponding docked models. XL restraints that achieved a distance of less than 35 Å in the XL-based model with the highest number of satisfied restraints were accepted. EC restraints, from the top 0.001 percentile of EC scores, that achieved a distance of less than 15 Å in the EC-based model with the highest number of satisfied restraints were accepted.

Assessment of model quality

To evaluate the overall geometric fit of the predicted models with respect to their target structures, we used measures and categories of CAPRI assessment protocol including the ligand Root Mean square deviation (L-RMSD), The interface RMSD (I-RMSD) and The fraction of interface native contacts, (F_{nat}).[38] We also assessed our models using DockQ score[39] which is a single score, ranging from 0-1, that can recreate CAPRI categories of model quality via a linear combination of L-RMSD, I-RMSD, and F_{nat} (Table 3). The intrinsic quality of AlphaFold-Multimer models was estimated using the pTM-based confidence score which equals a weighted linear combination of the model accuracy as a predicted TM-score (pTM) and interface accuracy as a predicted interface TM-score (ipTM).[41]

$$\text{Model confidence} = 0.8 \cdot \text{ipTM} + 0.2 \cdot \text{pTM} \qquad (1)$$

The model quality can be categorized as of high, medium, acceptable or incorrect based on their evaluation of three CAPRI measures of L-RMSD, I-RMSD, and F_{nat} or by a single DockQ score.

Table 3: Categories of model quality based on DockQ and CAPRI criteria.

Model Quality	**DockQ score**	**CAPRI criteria**
Highly	DockQ > 0.8	$F_{nat} \geq 0.5$ AND (LRMS ≤ 1.0 OR iRMS ≤ 1.0)
Medium	0.49 ≤ DockQ < 0.8	($F_{nat} \geq 0.3$ AND $F_{nat} < 0.5$) AND (LRMS ≤ 5.0 OR iRMS ≤ 2.0) OR ($F_{nat} \geq 0.5$ AND LRMS > 1.0 AND iRMS > 1.0)
Acceptable	0.23 ≤ DockQ < 0.49	($F_{nat} \geq 0.1$ AND $F_{nat} < 0.3$) AND (LRMS ≤ 10.0 OR iRMS ≤ 4.0) OR ($F_{nat} \geq 0.3$ AND LRMS > 5.0 AND iRMS > 2.0)
Incorrect	0 < DockQ < 0.23	$F_{nat} < 0.1$ OR (LRMS > 10.0 AND iRMS > 4.0)

Assessment of interaction prediction

We trained logistic regression classifiers (LRC) to differentiate between physically interacting proteins (class 1 = interface) and proximate proteins with no significant physical interface (class 0 = no interface). We trained the LRCs on our benchmark dataset of 85 cross-linked protein pairs of which, 67 protein pairs are physically interacting with at least ten heavy atom interface contacts within 5 Å and 18 protein pairs have lesser or no interface contacts. This imbalance in the dataset presumably reflects the interacting proteins to non-interacting but functionally associated proteins within a cross-linking dataset. Two baseline classifiers were trained on features of XL and EC separately. The XL classifier was on the number of the identified cross-links. While the EC classifier was trained on features including maximum EC score (EC_{max}), kurtosis and skewness of EC scores distribution, the effective number of sequences in the multiple sequence alignment (N_{eff}), and the percentage of MSA coverage by the EC calculations. The hybrid XLEC classifier was trained on the integration of these XL and EC features as well as the structural features of the generated models by XLEC-based docking. The number of satisfied cross-linking restraints (within 30 Å), the number of satisfied EC restraints (within 10 Å), the Haddock score, and the number of clusters per docking procedure. A fourth classifier that combines data from XL, and EC, as well as features extracted from AlphaFold-Multimer models, was developed. AlphaFold-Multimer features include the number of satisfied cross-linking restraints (within 30 Å), the number of satisfied EC restraints (within 10 Å), the pTM-based confidence score of AlphaFold-Multimer as calculated in Eq. 1 in the methods section. An AlphaFold-Multimer baseline classifier was trained only on the pTM-based confidence score of AlphaFold-Multimer. Table 4 shows the features used for every LRM. All the models were developed using the library(caret) in R.[60] We used five-fold cross-validation to evaluate the performance of the models. The models, the training sets, and the code are available under an MIT license at (https://github.com/KohlbacherLab/XLEC).

Table 4: Features of the logistic regression models.

Inference model	# XL	# XL < 30 Å	EC_{max}	# EC < 10 Å	MSA N_{eff}	MSA % coverage	EC Kurtosis & Skewness	AlphaFold pTM-based confidence score	Haddock score	# Haddock clusters
XL	✔	–	–	–	–	–	–	–	–	–
EC	–	–	✔	–	✔	✔	✔	–	–	–
XLEC	✔	✔	✔	✔	✔	✔	✔	–	✔	✔
XLEC + AlphaFold-Multimer	✔	✔	✔	✔	✔	✔	✔	✔	–	–
AlphaFold-Multimer	–	–	–	–	–	–	–	✔	–	–

Acknowledgements

Authors thank members Kohlbacher Lab and Lupas Lab in particular, Prof. Dr. Andrei Lupas, Dr. Timo Sachenberg, and Dr. Kyowon Jeong for discussion and feedback. Hadeer

Elhabashy was funded by the IMPRS “From Molecules to Organisms” and the Eberhard Karls University of Tübingen. Eugen Netz was supported by the Ministry of Science, Research and Arts Baden-Württemberg. Computational resources were provided by the Orchestra High Performance Compute Cluster at Harvard Medical School, which is funded by the NIH (NCRR 1S10RR028832-01) and the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A).

Author contributions

HE and OK developed the concept and outline for the study. HE wrote the code, performed the calculations, did the analysis and structural modeling. HE, EN and OK wrote and revised the manuscript. OK supervised the study. All authors read, discussed and edited the final version of this manuscript.

Data and code availability

All the generated interaction models are available at
(https://orinoco.cs.uni-tuebingen.de/Elhabashy/supplementary_models.zip)
XLEC is available under an MIT license at (https://github.com/KohlbacherLab/XLEC).

Competing Interests

The authors declare no competing financial interests.

References

1. Elhabashy, H., Merino, F., Alva, V., Kohlbacher, O., and Lupas, A.N. (2022). Exploring protein-protein interactions at the proteome level. Structure *30*, 462–475.

2. Fields, S., and Song, O. (1989). A novel genetic system to detect protein-protein interactions. Nature *340*, 245–246.

3. Van Criekinge, W., and Beyaert, R. (1999). Yeast Two-Hybrid: State of the Art. Biol. Proced. Online *2*, 1–38.

4. Brückner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. Int. J. Mol. Sci. *10*, 2763–2788.

5. Dunham, W.H., Mullin, M., and Gingras, A.-C. (2012). Affinity-purification coupled to mass spectrometry: basic principles and strategies. Proteomics *12*, 1576–1590.

6. Chang, T.W. (1983). Binding of cells to matrixes of distinct antibodies coated on solid surface. J. Immunol. Methods *65*, 217–223.

7. Hall, D.A., Ptacek, J., and Snyder, M. (2007). Protein microarray technology. Mech. Ageing Dev. *128*, 161–167.

8. Smith, G.P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science *228*, 1315–1317.

9. Sidhu, S.S., Fairbrother, W.J., and Deshayes, K. (2003). Exploring protein-protein interactions with phage display. Chembiochem *4*, 14–25.

10. Leitner, A. (2016). Cross-linking and other structural proteomics techniques: how chemistry is enabling mass spectrometry applications in structural biology. Chem. Sci. *7*, 4792–4803.

11. Rappsilber, J. (2011). The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. J. Struct. Biol. *173*, 530–540.

12. Tang, X., Wippel, H.H., Chavez, J.D., and Bruce, J.E. (2021). Crosslinking mass spectrometry: A link between structural biology and systems biology. Protein Sci. *30*, 773–784.

13. Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. *23*, 324–328.

14. Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature *402*, 86–90.

15. Chia, J.-M., and Kolatkar, P.R. (2004). Implications for domain fusion protein-protein interactions based on structural information. BMC Bioinformatics *5*, 161.

16. Kim, M.-D., Han, K.-C., Kang, H.-A., Rhee, S.-K., and Seo, J.-H. (2003). Coexpression of BiP increased antithrombotic hirudin production in recombinant Saccharomyces cerevisiae. J. Biotechnol. *101*, 81–87.

17. Kemmeren, P., van Berkum, N.L., Vilo, J., Bijma, T., Donders, R., Brazma, A., and Holstege, F.C.P. (2002). Protein interaction verification and functional annotation by integrated analysis of genome-scale data. Mol. Cell *9*, 1133–1143.

18. Shoemaker, B.A., and Panchenko, A.R. (2007). Deciphering protein-protein interactions. Part I. Experimental techniques and databases. PLoS Comput. Biol. *3*, e42.

19. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc. Natl. Acad. Sci. U. S. A. *96*, 4285–4288.

20. Škunca, N., and Dessimoz, C. (2015). Phylogenetic profiling: how much input data is enough? PLoS One *10*, e0114701.

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

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

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

24. Dos Santos, R.N., Bottino, G.F., Gozzo, F.C., Morcos, F., and Martínez, L. (2020). Structural complementarity of distance constraints obtained from chemical cross-linking and amino acid coevolution. Proteins *88*, 625–632.

25. Dos Santos, R.N., Ferrari, A.J.R., de Jesus, H.C.R., Gozzo, F.C., Morcos, F., and Martínez, L. (2018). Enhancing protein fold determination by exploring the complementary information of chemical cross-linking and coevolutionary signals. Bioinformatics *34*, 2201–2208.

26. Merkley, E.D., Cort, J.R., and Adkins, J.N. (2013). Cross-linking and mass spectrometry methodologies to facilitate structural biology: finding a path through the maze. J. Struct. Funct. Genomics *14*, 77–90.

27. Yu, C., and Huang, L. (2018). Cross-Linking Mass Spectrometry: An Emerging Technology for Interactomics and Structural Biology. Anal. Chem. *90*, 144–165.

28. Lockless, S.W., and Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. Science *286*, 295–299.

29. Gloor, G.B., Martin, L.C., Wahl, L.M., and Dunn, S.D. (2005). Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry *44*, 7156–7165.

30. Fodor, A.A., and Aldrich, R.W. (2004). On evolutionary conservation of thermodynamic coupling in proteins. J. Biol. Chem. *279*, 19046–19050.

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

32. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D.S., Sander, C., Zecchina, R., Onuchic, J.N., Hwa, T., and Weigt, M. (2011). Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. U. S. A. *108*, E1293–E1301.

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

34. Jones, D.T., Buchan, D.W.A., Cozzetto, D., and Pontil, M. (2012). PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics *28*, 184–190.

35. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M., and Aurell, E. (2013). Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *87*, 012707.

36. Schweppe, D.K., Chavez, J.D., Lee, C.F., Caudal, A., Kruse, S.E., Stuppard, R., Marcinek, D.J., Shadel, G.S., Tian, R., and Bruce, J.E. (2017). Mitochondrial protein interactome elucidated by chemical cross-linking mass spectrometry. Proc. Natl. Acad. Sci. U. S. A. *114*, 1732–1737.

37. Liu, F., Lössl, P., Rabbitts, B.M., Balaban, R.S., and Heck, A.J.R. (2018). The interactome of intact mitochondria by cross-linking mass spectrometry provides evidence for coexisting respiratory supercomplexes. Mol. Cell. Proteomics *17*, 216–232.

38. Lensink, M.F., Méndez, R., and Wodak, S.J. (2007). Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins *69*, 704–718.

39. Basu, S., and Wallner, B. (2016). DockQ: A Quality Measure for Protein-Protein Docking Models. PLoS One *11*, e0161879.

40. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

41. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. (2022). Protein complex prediction with AlphaFold-Multimer. bioRxiv, 2021.10.04.463034. 10.1101/2021.10.04.463034.

42. UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. *49*, D480–D489.

43. Burley, S.K., Bhikadiya, C., Bi, C., Bittrich, S., Chen, L., Crichlow, G.V., Christie, C.H., Dalenberg, K., Di Costanzo, L., Duarte, J.M., et al. (2021). RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. Nucleic Acids Res. *49*, D437–D451.

44. Boratyn, G.M., Schäffer, A.A., Agarwala, R., Altschul, S.F., Lipman, D.J., and Madden, T.L. (2012). Domain enhanced lookup time accelerated BLAST. Biol. Direct *7*, 12.

45. Guex, N., Peitsch, M.C., and Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. Electrophoresis *30 Suppl 1*, S162–S173.

46. Bienert, S., Waterhouse, A., de Beer, T.A.P., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T. (2017). The SWISS-MODEL Repository-new features and functionality. Nucleic Acids Res. *45*, D313–D319.

47. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., and Schwede, T. (2017). Modeling protein quaternary structure of homo- and hetero-oligomers beyond binary interactions by homology. Sci. Rep. *7*, 10480.

48. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. *46*, W296–W303.

49. Studer, G., Rempfer, C., Waterhouse, A.M., Gumienny, R., Haas, J., and Schwede, T. (2020). QMEANDisCo-distance constraints applied on model quality estimation. Bioinformatics *36*, 1765–1771.

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

51. Johnson, L.S., Eddy, S.R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. BMC Bioinformatics *11*, 431.

52. UniProt Consortium (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.

53. Weinreb, C., Riesselman, A.J., Ingraham, J.B., Gross, T., Sander, C., and Marks, D.S. (2016). 3D RNA and Functional Interactions from Evolutionary Couplings. Cell *165*, 963–975.

54. Hopf, T.A., Green, A.G., Schubert, B., Mersmann, S., Schärfe, C.P.I., Ingraham, J.B., Toth-Petroczy, A., Brock, K., Riesselman, A.J., Palmedo, P., et al. (2019). The

EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics *35*, 1582–1584.

55. Yan, C., Wu, F., Jernigan, R.L., Dobbs, D., and Honavar, V. (2008). Characterization of protein-protein interfaces. Protein J. *27*, 59–70.

56. Hubbard, S.J., and Thornton, J.M. naccess. Computer Program, Department of Biochemistry and.

57. Lee, B., and Richards, F.M. (1971). The interpretation of protein structures: estimation of static accessibility. J. Mol. Biol. *55*, 379–400.

58. van Zundert, G.C.P., Rodrigues, J.P.G.L.M., Trellet, M., Schmitz, C., Kastritis, P.L., Karaca, E., Melquiond, A.S.J., van Dijk, M., de Vries, S.J., and Bonvin, A.M.J.J. (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J. Mol. Biol. *428*, 720–725.

59. Brunger, A.T. (2007). Version 1.2 of the Crystallography and NMR system. Nat. Protoc. *2*, 2728–2733.

60. Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. J. Stat. Softw. *28*, 1–26.

Supplementary figures

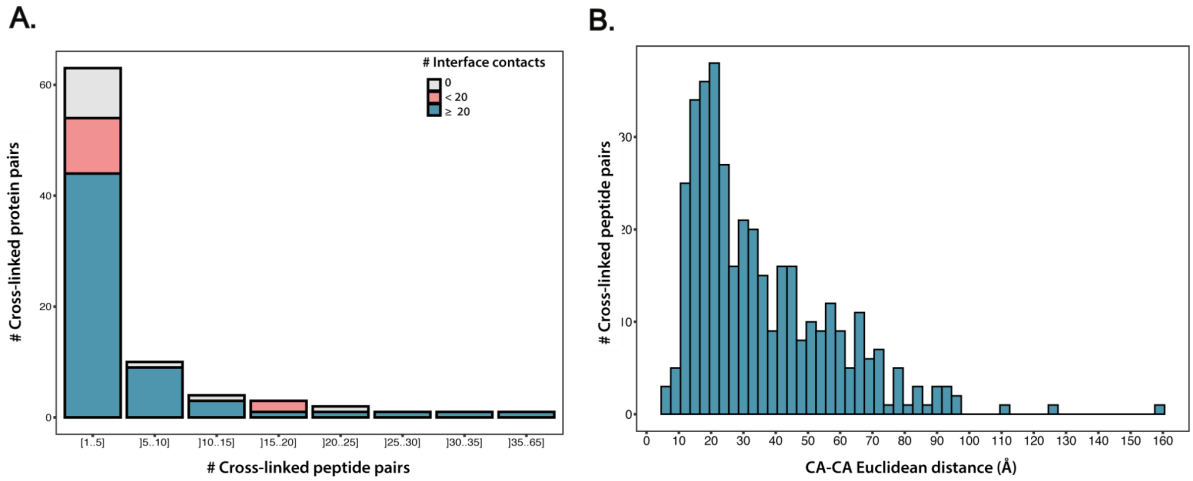
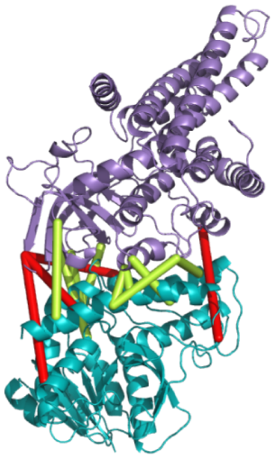


Figure S.1: Statistics of the XL-MS dataset
Panel A shows a bar plot between the number of identified cross-links per protein pair versus the count of protein pairs achieving this number. Panel B shows the distribution of the CA-CA Euclidean distances achieved by the identified cross-linked when mapping them on their corresponding structures, if they exist.

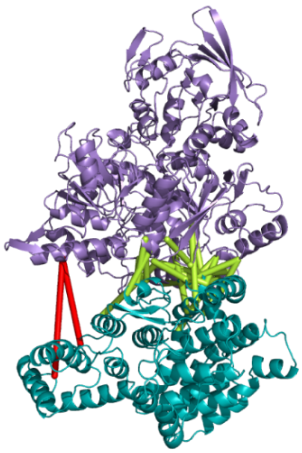
A. ATPA-HCDH interaction
#XL=4 & #EC=18



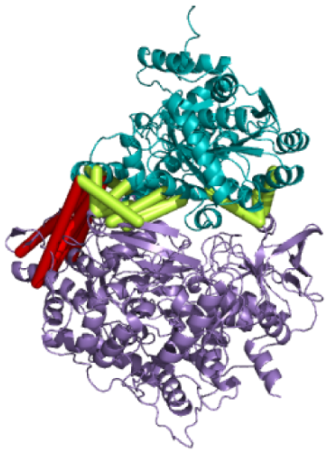
B. MDHM-ACADM interaction
#XL=4 & #EC=9



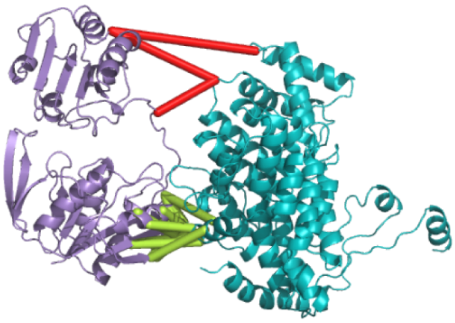
C. CISY-ACON interaction
#XL=2 & #EC=23



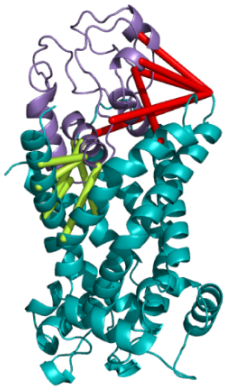
D. AATM-ACON interaction
#XL=4 & #EC=10



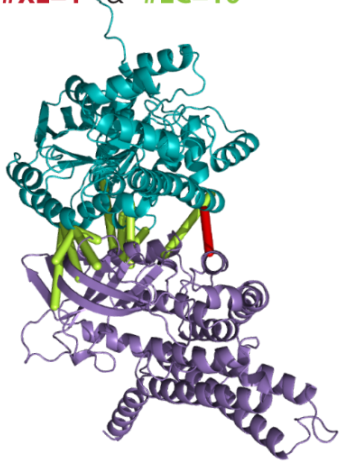
E. CISY-ETFA interaction
#XL=3 & #EC=17



F. ADT1-CYC interaction
#XL=6 & #EC=6



G. AATM-ACADM interaction
#XL=1 & #EC=10



H. FUMH-CISYV interaction
#XL=2 & #EC=10



I. CISY-QCR1 interaction
#XL=3 & #EC=13

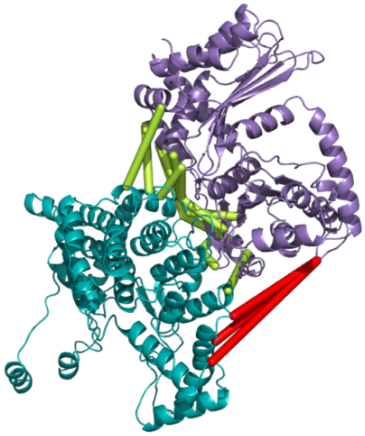


Figure S.2: *De novo* XLEC models

De novo interaction models (in teal and purple) predicted by XLEC-based docking including A. ATPA and HCDH interaction. B. MDHM and ACADM interaction. C. CISY and ACON interaction. D. AATM and ACON interaction. E. CISY and ETFA interaction. F. ADT1 and CYC interaction. G. AATM and ACADM interaction. H. FUMH and CISYV interaction. I. CISY and QCR1 interaction. Cross-linking restraints (XL) are shown in red and the evolutionary coupling restraints (EC) are shown in yellow.