On the characterization of protein interaction interfaces with computational approaches and experimental validations

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

Der Johannes Gutenberg-Universität Mainz

Chop Yan Lee

geb. am 21.04.1996 in Pulau Pinang, Malaysia

Mainz, Januar 2024

Dekan: Prof. Dr. Eckhard Thines

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung: 09.04.2024

Abstract

Proteins play crucial roles in virtually all cellular processes, and their functions are often realized through interactions with other proteins via different interfaces. Thus, for a comprehensive understanding of protein functions, it is important to investigate the molecular mechanisms of protein-protein interactions (PPIs) by characterizing their interaction interfaces. Unfortunately, due to most high-throughput assays detecting only PPIs and not their interfaces, mechanistic information is scarce for most PPI data.

The current thesis presents a collection of studies exploring the characterization of different types of PPI interfaces in the human protein interactome. Domains and motifs are two types of conserved functional modules that enable PPIs by forming domain-domain interfaces (DDIs) and domain-motif interfaces (DMIs). Chapter 2 and Article I delved into the characterization of DMIs and DDIs in PPIs by leveraging information from existing databases. Chapter 2 centered on automating the detection and scoring of DMIs in PPIs through a computer program. Article I focused on evaluating the DDIs annotated in the 3did database by manually curating a reference dataset of DDIs and identifying useful features to further score them. The program developed in Chapter 2 and the high-confidence DDIs from 3did were applied to PPIs detected in the human protein interactome to characterize their interfaces.

AlphaFold-Multimer (AF-MM) is an artificial intelligence (AI)-based tool for predicting the structures of protein complexes. Article II and III investigated the use of AF-MM to predict novel PPI interfaces. As there is a lack of a comprehensive assessment of AF-MM and its metrics, Article II systematically benchmarked AF-MM's ability to predict different types of interfaces and established criteria for discriminating good from bad structural models. Testing AF-MM using sequences longer than minimal interacting regions revealed that they are detrimental to AF-MM's prediction performance, prompting the development of a fragmentation-based approach to enhance AF-MM's sensitivity. The approach was applied to PPIs detected in the human protein interactome to predict their interfaces, and some predicted interfaces were experimentally validated. Similarly, article III applied the fragmentation approach on a protein pair whose interaction is important for piRNA amplification. Subsequent experimental validation also confirmed the interaction interface predicted by AF-MM.

This thesis provides various approaches to leverage existing knowledge on different PPI interface types to predict PPI interfaces in the human protein interactome, paving the way towards a mechanistically annotated human protein interactome.

Zusammenfassung

Proteine spielen eine zentrale Rolle in nahezu allen zellulären Prozessen. Ihre Funktionen werden dabei oft durch Interaktionen mit anderen Proteinen vermittelt, die wiederum an verschiedenen Interaktionsflächen stattfinden. Für ein umfassendes Verständnis der Proteinfunktionen ist es demnach essentiell, diese Interaktionsflächen zu analysieren, um die molekularen Mechanismen von Protein-Protein-Interaktionen (PPIs) zu charakterisieren. Da die die meisten Hochdurchsatz-Assays nur der Detektion von PPIs dienen und nicht auch der ihrer Interaktionsflächen dienen, sind mechanistische Informationen für die meisten PPIs nicht verfügbar.

Diese Dissertation unfasst eine Sammlung von Studien, die eine Charakterisierung verschiedener Arten von PPI-Interaktionsflächen im menschlichen Interaktom anstreben. Domänen und Motive sind zwei Arten von konservierten funktionalen Modulen, die PPIs durch die Bildung von Domäne-Domäne-Interaktionsflächen (DDIs) und Domäne-Motiv-Schnittstellen (DMIs) vermitteln. Kapitel 2 und Artikel I beschäftigen sich mit der Charakterisierung von DMIs und DDIs in PPIs unter der Verwendung von Informationen aus bestehenden Datenbanken. Dabei konzentriert sich Kapitel 2 auf die automatisierte Detektion und Bewertung von DMIs in PPIs durch ein Computerprogramm. Artikel I setzt sich mit der Bewertung der in der 3did-Datenbank annotierten DDIs auseinander, wobei ein Referenzdatensatz von DDIs manuell kuratiert und nützliche Merkmale zur weiteren Evaluierung identifiziert wurden. Das in Kapitel 2 entwickelte Programm und die als qualitativ hochwertig evaluierten DDIs aus 3did wurden im weiteren Verluaf genutzt um die Interaktionsflächen von PPIs im menschlichen Interaktom zu charakterisieren.

AlphaFold-Multimer (AF-MM) ist ein auf kuünstlicher Intelligenz (KI) beruhendes Programm zur Vorhersage der Strukturen von Proteinkomplexen. Artikel II und III untersuchen die Verwendung von AF-MM zur Vorhersage neuer PPI-Interaktionsflächen. In Anbetracht der fehlenden umfassenden Bewertung von AF-MM und seinen Messvariablen, hat Artikel II die Fähigkeit von AF-MM hinsichtlich der Vorhersage verschiedener Interaktionsflächentypen systematisch bewertet und Kriterien zur Unterscheidung von guten und schlechten Struktur-modellen identifiziert. Tests von AF-MM mit Sequenzen, die mehr als die minimalen Interaktionsregionen umfassten, ergaben, dass solche nachteilig für die Vorhersageleistung von AF-MM sind, was zur Entwicklung eines fragmentierungsbasierten Ansatzes führte, um so die Sensitivität von AF-MM zu erhöhen. Dieser Ansatz wurde auf die im menschlichen Interaktom detektierten PPIs angewendet, um deren Interaktionsflächen vorherzusagen. Einige dieser vorhergesagten Interaktionsflächen wurden gleichfalls experimentell validiert. Artikel III wendet den entwickelten Fragmentierungsansatz auf ein Proteinpaar an, dessen Interaktion essentiell für die piRNA-Amplifikation ist. Die anschließende experimentelle Validierung konnte ebenso die von AF-MM vorhergesagte Interaktionsfläche bestätigen.

Diese Dissertation zeigt verschiedene Ansätze auf, um bestehendes Wissen über verschiedene PPI-Interaktionsflächentypen zu nutzen, um schließlich PPI-Schnittstellen im menschlichen Interaktom vorherzusagen, und ebnet so den Weg zu einem mechanistisch annotierten menschlichen Interaktom.

Acknowledgement

Acronyms

3did 3D Interaction Domains.

AF AlphaFold.

AF-MM AlphaFold-Multimer.

AI Artificial intelligence.

AP-MS Affinity Purification-Mass Spectrometry.

BERT Bidirectional encoder representations from transformers.

BRET Bioluminescence resonance energy transfer.

CASP14 14th Critical Assessment of Structure Prediction.

cryo-EM cryogenic electron microscopy.

DDI Domain-domain interface.

DMI Domain-motif interface.

DNA Deoxyribonucleic acid.

ELM Eukaryotic Linear Motif.

ER Endoplasmic reticulum.

FHA Forkhead-associated.

FRET Förster resonance energy transfer.

GBD GTPase-binding domain.

GO Gene ontology.

GPT Generative pre-trained transformer.

HMM Hidden Markov Model.

HuRI Human Reference Interactome.

IDP Intrinsically disordered protein.

IDR Intrinsically disordered region.

iELM interactions of Eukaryotic Linear Motif.

KISS Kinase Substrate Sensor.

MAPPIT Mammalian Protein-protein Interaction Trap.

ML Machine learning.

MS Mass Spectrometry.

MSA Multiple sequence alignment.

NLP Natural language processing.

NMR Nuclear magnetic resonance.

ORF Open reading frame.

PDB Protein Data Bank.

Pfam Protein families.

piRNA Piwi-interacting RNA.

pLDDT predicted Local Distance Difference Test.

PPI Protein-protein interaction.

PR Precision recall.

PRS Positive reference set.

PSSM Position specific scoring matrix.

PTM Post-translational modification.

QFO Quest for Orthologs.

RF Random forest.

RLC Relative Local Conservation.

RNA Ribonucleic acid.

ROC Receiver operating characteristics.

RRS Random reference set.

SASA solvent accessible surface area.

SCOP Structural Classification of Proteins.

SH2 Src homology 2.

SLiM Short linear motif.

SMART Simple Modular Architecture Research Tool.

TIR Toll/interleukin-1 receptor.

Y2H Yeast two-hybrid.

Contents

Chapter 1

Introduction

Given the massive amount of proteins encoded in any given genome, probing the functions of proteins has been an objective of utmost importance in the field of molecular biology for decades. As proteins interact with other proteins to mediate their cellular functions, mapping the interactions between proteins have been a common approach to study the functions of proteins. With the development of interaction assays for high-throughput screening, the last decades have witnessed tremendous progress in the mapping of protein-protein interactions (PPIs). For example, the yeast two-hybrid (Y2H) assay is a simple and robust method for PPIs detection that was developed for high-throughput screening during the early 1990s (Chien et al., 1991; Fields & Song, 1989). Since its development, the assay has been employed by numerous researchers to screen for PPIs across various proteomes (McCraith et al., 2000; Stelzl et al., 2005; Uetz et al., 2000, 2006). The human reference interactions (HuRI) is the largest binary PPI screen for the human proteome so far, detecting over 60,000 PPIs, with many of them being mediated by poorly studied proteins (Luck et al., 2020). A similar trend has also been observed in the advancement of mass spectrometry (MS) technologies. MS-based approaches have also been utilized to study the co-complex associations between human proteins (Huttlin et al., 2021) and yeast proteins (Gavin et al., 2002; Krogan et al., 2006; Michaelis et al., 2023). These PPI networks are collectively termed the interactome.

While the systematic nature of these PPI mapping efforts enabled the discovery of PPIs mediated by many understudied proteins, a mechanistic description of the interactome is currently lacking. To put this in context, only approximately 4% of the interactions detected in HuRI have a structure solved at atomic resolution (Luck et al., 2020). A mechanistic description of a PPI entails examining the interaction interface between an interacting protein pair to determine the residues that are crucial for the interaction. Such mechanistic insights into PPIs will allow one to decipher the co-existence or mutual exclusivity of partner proteins in protein complexes and, subsequently, how one can modulate them (Fuller et al., 2009; H. Lu et al., 2020; Nim et al., 2016; Wells & McClendon, 2007). Furthermore, the integration of mechanistic information into PPIs will also shed light on the molecular functions of the PPIs, as well as the molecular mechanisms of disease caused by pathogenic variants. That said, a mechanistic description of the interactome is a prerequisite for a comprehensive understanding of biological systems.

Various experimental approaches, such as mutational scanning, crystallography, and protein fragmentation, exist to detect PPI interfaces at different resolutions (Araya & Fowler, 2011; Green et al., 1954). Nonetheless, these methods are laborious and resource-intensive. To exemplify this, the common crystallography approach to study PPI interfaces involves dissolving proteins in an aqueous environment and saturating the protein solution using different methods until it reaches supersaturated state. Once supersaturated, the protein crystals formed can be used to determine the macromolecular structures of the protein using various methods, such as X-ray diffraction and cryogenic electron microscopy (cryo-EM) (McPherson & Gavira, 2014; Milne et al., 2013). As of 2018, 89.5% of three-dimensional structures deposited in the Protein Data Bank (PBD) were solved using crystallography, 8.5% using nuclear magnetic resonance (NMR) technique, and 1.6% by 3D electron microscopy. Despite significant growth in structures deposited in the PBD in recent years, only 0.02% of the unique entries in the UniProtKB database have been linked to at least one structure in the PDB (wwPDB consortium, 2019).

Owing to the complexity and difficulty of experimental approaches, computational methods to predict PPI interfaces have become a field of active research in recent years. There are different approaches to predict PPI interfaces, and they can be largely classified as sequence-based or structure-based. Sequence-based approach involves scanning the primary sequences of a protein pair to detect potential interacting regions while structure-based approach utilizes the three-dimensional structure of the protein pair to find potential interacting sites. More recently, artificial intelligence (AI) algorithms have been incorporated into many bioinformatic tools, including those for structural modeling. These AI-powered tools can also be used to predict PPI interfaces. These computational methods will be discussed in more detail in section 1.3. To validate the predicted PPI interfaces, different assays can be used to probe PPI interfaces at different resolutions. These experimental methods will be discussed in more detail in section 1.4.

1.1 Modular architecture of proteins

Interactions between proteins are made possible through conserved functional modules embedded within their sequences. Intriguingly, many proteins possess multiple conserved modules that enable them to interact with multiple partners. These modules can be broadly classified into those with defined structure and those that lack a defined structure. The former is commonly known as domains, which refers to a region in a polypeptide chain that is capable of folding, often independently, into a tertiary structure. The latter is widely known as intrinsically disordered regions (IDRs) where short linear motifs (SLiMs) (hereafter referred to as motifs) predominantly reside. Motifs primarily exist in the cellular environment as flexible stretches of amino acids that occasionally fold into a secondary structure upon binding to a globular domain. These two types of functional modules will be discussed in more detail in the next sections.

In eukaryotes, 65-70% of their proteomes are made up of proteins that consist of multiple modules (Ekman et al., 2005; Han et al., 2007). The interplay of these modules between proteins give rise to complex genotype-to-phenotype relationship. To exemplify this, the well-characterized protein, p53, plays critical roles in many important cellular processes, such as cell cycle regulation and progression, apoptosis, and genomic stability. The 393 amino acid-long protein is composed of two transactivation domains at the N terminus, followed by a DNA-binding domain, and a tetramerization domain at the C terminus. The N terminal region is largely disordered with many known motifs involved in degradation, phosphorylation, and docking scattered throughout this disordered region (Canman et al., 1998; Sakaguchi et al., 2000; Shin et al., 2015; Turenne & Price, 2001; Zacchi et al., 2002). Additionally, various nuclear localization signals and docking motifs are also found in the C terminal disordered region (Gostissa et al., 1999; Liang & Clarke, 1999; Luciani et al., 2000; O'Keefe et al., 2003; Sheng et al., 2006). While p53 represents a very well-studied protein as it has been implicated in a broad range of cancers (Mantovani et al., 2019), it showcases the modularity of proteins and the complex phenotypes that can arise from the intricate interplay of functional modules (Figure 1.1).

The functional consequences of genetic variants can be pleiotropic, meaning that the genetic variants of the same gene can cause different phenotypes (Gratten & Visscher, 2016). In view of the modular nature of proteins, the pleiotropic effects of genetic variants can be, at least partially, attributed to the specific regions of a protein that they impact. In accordance, 50% of pathogenic variants were reported to affect only a subset of a protein's interactions while leaving other interactions intact, thereby altering only a subset of a protein's functions (Sahni et al., 2015). By modeling the PPI interfaces of a filtered set of PPIs based on homology, Wang et al. (2012) also reported a significant enrichment of in-frame disease mutations on PPI interfaces, sug-

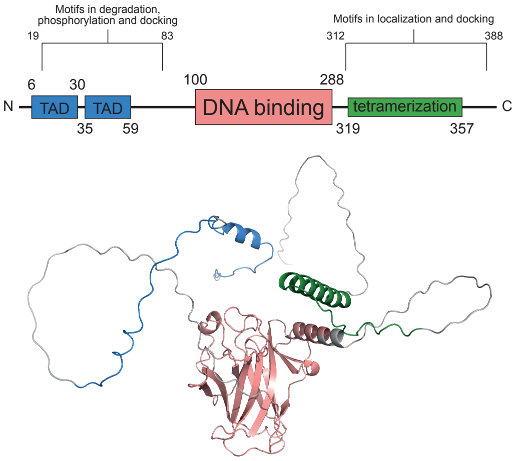

Motifs in degradation, phosphorylation and docking
19
83
Motifs in localization and docking
312
388
6
30
100
288
N
TAD
TAD
DNA binding
tetramerization
C
35
59
319
357

Figure 1.1: Domains and motifs as functional modules exemplified in p53 (not drawn to scale). The numbers above and below the boxes denote the boundaries of the domains. The top panel shows the modular architecture of p53. The bottom panel shows the full-length structure of p53 predicted by AlphaFold. Domains in the structure are colored according to their colors in the top panel. TAD stands for transactivation domain.

gesting that many mutations caused disease by perturbing PPIs. The same study also compared in-frame disease mutation pairs that impact either the same or different interaction interfaces, and revealed that mutation pairs affecting different interfaces are more than twice as likely to cause distinct disorders. Furthermore, it was also shown that pathogenic variants of neurodevelopmental genes are significantly enriched in protein regions that fold into domains or function as motifs (Iqbal et al., 2023). Together, these findings underpin the importance of deciphering the molecular mechanism of PPIs as they provide mechanistic insights into a protein's function. These insights are key to understanding and characterizing the functional effect of genetic variants.

1.1.1 Folded domains

Biological roles

The concept of a protein domain was first proposed in 1973 by Wetlaufer and Ristow after reviewing X-ray crystallographic studies of enzymes and immunoglobulins (Wetlaufer & Ristow, 1973). Since then, the terminology has been used primarily to refer to regions spanning 50 to 250 amino acids in a polypeptide chain that are usually capable of folding independently into a tertiary structure from the rest of the polypeptide chain. The process of folding can be understood as the search within the conformational space to find the conformation with the lowest energy. This usually refers to the sequestration of hydrophobic residues in a compact core while exposing hydrophilic residues to the aqueous solvent. While most domains are self-stabilizing, for some smaller domains, such as zinc fingers, coordination with metal ions is necessary to stabilize the fold. Domains form the functional units of a protein that enable the protein to perform its cellular functions. For instance, the protein PNKP is a key player in many DNA repair mechanisms such as non-homologous end-joining and base excision repair (Weinfeld et al., 2011). PNKP is composed of three domains that are linked by flexible linkers: a forkhead-associated (FHA) domain, a phosphatase domain, and a kinase domain. The FHA domain recruits PNKP to DNA damage sites, while the kinase and phosphatase domains catalyze the phosphorylation and dephosphorylation of damaged DNA termini to ensure that they are compatible with downstream DNA repair processes. In spite of these domains performing dissimilar molecular functions, when linked together in a single polypeptide chain, they allow a protein, in this case PNKP, to perform its important function in DNA repair pathways (Figure 1.2).

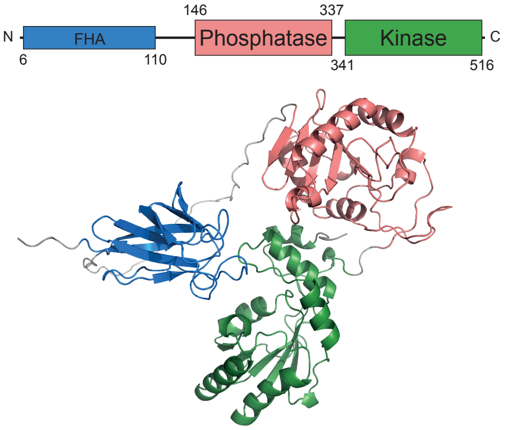


Figure 1.2: Domain architecture of PNKP. The top panel shows the domain architecture of PNKP (not drawn to scale). The numbers above and below the boxes denote the boundaries of the domains. The bottom panel shows the full-length structure of PNKP as predicted by AlphaFold. Domains in the structure are colored according to their colors in the top panel. FHA stands for forkhead-associated domain.

Being functional units with distinct molecular roles in the cell, domains also serve as building blocks that natural selection acts on to create diverse molecular machinery. Gene duplication and recombination events that occur over the course of evolution lead to similar domains occurring in a wide range of seemingly unrelated proteins. The shuffling and concatenation of domains in various permutations in turn allow proteins to specialize in specific cellular function by combining the functionalities conferred by different domains (Bagowski et al., 2010). For example, the Src homology 2 (SH2) domain is a structurally conserved domain that binds phosphorylated tyrosine, and it is found in a vast array of proteins that are involved in diverse signaling pathways. Albeit involved in distinct cellular functions, the SH2 domains in these proteins essentially perform the same role: linking and recruiting other signaling factors to form signal transduction complexes. Another interesting example is the zinc finger domain, a structural motif characterized by its coordination with one or more zinc ions to stabilize its fold. Owing to its ability to bind nucleic acids, the zinc finger domain recurs in many transcription factors. While retaining its overall structural similarity, the sequences of many zinc finger domains have diverged substantially over time to accommodate functions beyond nucleic acid binding. For example, unlike other zinc finger domains that bind to nucleic acids, the RING finger domain, commonly found in ubiquitin ligases, binds instead to other proteins (Cassandri et al., 2017).

Databases of domains and tools to predict domains

With their relatively limited sizes and their ability to fold independently, domains are readily identifiable from amino acid sequences using Hidden Markov models (HMMs). HMMs are statistical representations that can be used to model multiple sequence alignments (MSAs) and, therefore, detect sequence homology (Bystroff & Krogh, 2008). The versatility of HMMs lies in their ability to model entire alignments, including divergent regions with insertions and deletions. This property of HMMs makes them ideal to capture the evolutionarily conserved fingerprints in domain alignments.

There are several databases that make use of HMMs to identify domains in amino acid sequences, but they differ in their approaches to prepare domain alignments for HMMs construction. The CATH database provides a systematic and hierarchical structural classification of protein structures that are deposited in PBD (Sillitoe et al., 2020). Apart from structurally classifying domains, the database also generates domain alignments for HMM building using representative structures of domains and their sequences. Another database that performs a similar classification task on domains is Structural Classification of Proteins (SCOP) (Andreeva et al., 2014, 2020). In addition to structural information, SCOP also takes into account evolutionary relationships of domains in its classification. It is worth noting that the classification of protein structures by SCOP was largely manual with an emphasis on expert knowledge, while CATH uses a semi-automated approach. The SUPERFAMILY database uses the sequences of domains classified at the SCOP superfamily level to create domain alignments for HMM construction (Gough et al., 2001).

On the other hand, the Protein families (Pfam) and Simple Modular Architecture Research Tool (SMART) databases do not rely on protein structures but focus instead on protein and domain families that are defined based on sequence similarity (Letunic et al., 2021; Mistry et al., 2021). However, while Pfam has a broad coverage thanks to its automated sequence alignment building and HMM generation, SMART has a much smaller coverage due to manual curation of its HMMs. Pfam was recently merged with InterPro, and more than 21,000 Pfam HMMs are available in the InterPro database (Paysan-Lafosse et al., 2023). Thanks to its wide coverage and comprehensible naming convention, using HMMs to detect the occurrences of domains in amino acid sequence is now an important tool for molecular biologists to study the domain architecture of proteins.

1.1.2 Intrinsically disordered regions

Biological roles

Due to the long-standing structure-function paradigm in structural biology, the involvement of IDRs in biological processes was largely ignored until recent decades. The paradigm originated from the studies of enzymes in the 1930s where scientists observed that enzyme denaturation correlated with the loss of enzymatic activity (Northrop, 1930). This led to the parsimonious assumption that the three-dimensional structure of a protein is a prerequisite for its function. Five decades later, a shift in the paradigm was initiated with the discovery of protein segments that are essential for the function of the protein but yielded no electron density for them to be observed in X-ray crystallography studies. Substantiated by later NMR studies, the functionality of IDRs in proteins is now widely accepted in the field of molecular biology (Dunker et al., 2001). Further work also revealed that there exist also proteins that are fully disordered, and they are termed intrinsically disordered proteins (IDPs).

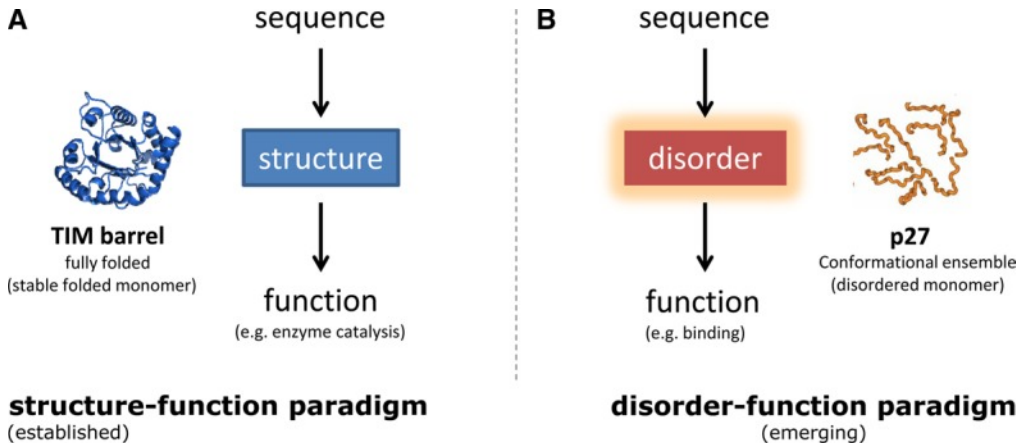


Figure 1.3: The established and emerging paradigms of structure-function and disorder-function. From (Babu, 2016).

Contrary to previous assumptions that IDRs are merely linkers between domains, they are involved in a wide range of molecular processes. The processes span from protein folding (To et al., 2023) to molecular recognition and drivers of subcellular organization (Holehouse & Kragelund, 2023; Tompa, 2005). IDRs are prevalent in eukaryotic proteomes, with 30-40% of residues in eukaryotic proteomes being in IDRs (Holehouse & Kragelund, 2023). A comparative study between the amino acid composition of eukaryotic and prokaryotic IDRs further revealed that serine is specifically enriched in eukaryotic IDRs. Since serine phosphorylation is crucial in the regulatory pathways in eukaryotes, its enrichment in eukaryotic IDRs hints at the regulatory role of IDRs in eukaryotes (Basile et al., 2019; Iakoucheva et al., 2004). IDRs are also a key feature of some scaffold proteins that serve to spatio-temporally coordinate interaction partners in signaling events. The flexibility of IDRs is often thought to be one of the key factors that enables proteins to interact with multiple partners (Cortese et al., 2008). Indeed, a study on a subset of *Caenorhabditis elegans* proteins revealed a positive correlation between the presence of IDRs and the number of interacting partners (Schlessinger et al., 2007). Besides, the extended nature of IDRs also exposes more surface area per residue to the environment and, therefore, facilitates interaction with more partners compared to ordered regions (Gunasekaran et al., 2003). The diverse role of IDRs can be attributed to the plethora of motifs found in IDRs, and they will be discussed in detail in the next section.

While the biological relevance of IDRs is becoming increasingly accepted, the experimental study of IDRs remains difficult for many reasons. Owing to their lack of structure, direct measurement of their dynamic behavior requires sophisticated experimental setups such as NMR. which demands high level of technical expertise. Moreover, many IDRs are context-dependent, with their propensity to be disordered dependent on external conditions, such as pH, presence of post-translational modifications (PTM), among others (Mohan et al., 2009). As a result, various computational tools have been developed to predict IDRs in proteins, and several databases have been set up to contain experimentally verified and computationally predicted IDRs.

Databases of disordered proteins and regions, and tools to predict protein disorder

DisProt is a database that contains standardized and easily accessible experimental evidence for IDPs and IDRs. Besides disorder annotation, molecular functions and structural transitions of IDPs and IDRs are also curated by experts through literature survey. Currently, over 2,000 and 6,000 eukaryotic IDPs and IDRs are annotated in DisProt.

Characterized by the absence of a compact core, IDRs are enriched with polar amino acids and depleted of hydrophobic amino acids (Romero et al., 2001). The difference in amino acid composition, as well as other factors such as flexibility, have been leveraged to develop predictors of IDRs in protein sequences. For example, the disorder predictor tool IUPred utilizes statistical potentials to quantify the tendencies of amino acid pairs to form contacts that are observed in a collection of globular protein structures. The derived statistical potentials are then applied onto the individual residues in an amino acid sequence to estimate their energies. Smoothed within a moving window centered on the calculated residue, residues that have favorable energies are predicted to be ordered while those with unfavorable energies are predicted to be disordered (Mészáros et al., 2009, 2018).

Recently, the structure prediction tool AlphaFold2 (AF) achieved prediction accuracies that are on par with experimental structures. As AF generates structural models for full-length proteins, it also outputs a confidence score known as predicted Local Distance Difference Test (pLDDT) for each amino acid in the predicted proteins. The score ranges from 0 to 100, where a higher score indicates greater confidence in the accuracy of a residue's predicted position. Interestingly, the pLDDT score has been shown to be highly correlated with the disorder propensity of residues, and therefore it can serve as an excellent discriminator of residue-wise disorder (Wilson et al., 2022). Similarly, the solvent accessible surface area (SASA) of each residue in AF-predicted models is also highly correlated with the disorder propensity of residues. Predicting residue disorder using pLDDT and SASA smoothed over a 20-residue window even outperformed IUPred2A (the latest version of IUPred) in a benchmark study (Akdel et al., 2022). AF has also been used for other applications, and they will be discussed in later sections.

1.1.3 Short linear motifs

Biological roles

Motifs are short and linear stretches of amino acids that are typically around 10 residues long, and they are predominantly found in the IDRs of proteins. When bound to its partner, some motifs adopt a secondary structure. An example of this type of motif is the degradation signal with the pattern PxAxVxP (x denotes wildcard positions where any amino acid is permitted) that is found in the degradation target of the E3 ubiquitin-ligase, SIAH1. Known as the beta augmentation mechanism, the degradation signal binds to SIAH1 by forming a beta strand parallel to the eight-stranded beta sandwich fold of the substrate binding domain of SIAH1 (Figure 1.4) (Santelli et al., 2005). Motifs are known to be degenerate, meaning that only three to four residues are conserved in the motifs, such as the PxAxVxP motif in the previous example, with solvent-facing residues being highly variable (Tompa et al., 2014; Van Roey et al., 2014). The conserved residues contribute to most of the free energy of binding and, therefore, determine the affinity and specificity of the interaction. Conversely, for some motifs, certain amino acids are prohibited at particular positions due to steric hindrance, while, for others, the motifs must be positioned at the termini to be functional.

It is estimated that more than 100,000 binding motifs exist in the human proteome alone, with many being so far uncharacterized (Tompa et al., 2014). The ones that have been characterized so far are classified into six classes that reflect their bi-

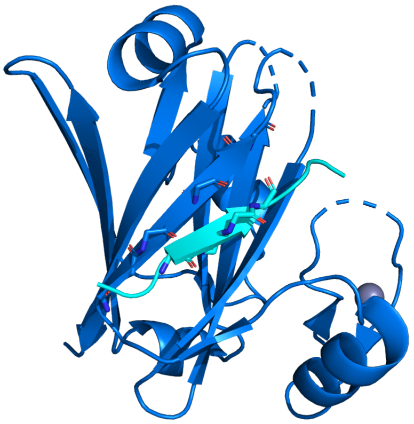


Figure 1.4: The binding of degradation signal PxAxVxP to E3 ubiquitin ligase, SIAH1, through beta augmentation mechanism (PDB ID: 2A25). The degradation signal is colored in cyan and the substrate binding domain of SIAH1 is colored as blue. The backbones of the residues at the interface are shown in stick. Dashed lines represent unsolved residues in the structure.

ological functions. The classes are modification sites, ligand binding sites, targeting signals, degrons, docking sites, and cleavage sites. Examples of modification sites include PTM sites like phosphorylation and SUMOylation sites where the motifs are subject to conjugation of moieties. Ligand binding sites refer to motifs that function to recruit their binding partners to the motif-containing proteins. This binding event often acts as an intermediate step in the formation of transient signaling complexes that are formed by multiple proteins. Targeting signals that enable the proteins to be trafficked to different cellular compartments are also a class of motif. They include nuclear localization and export signals, as well as Golgi and endoplasmic reticulum trafficking signals, among others. Another class of motif that functions as degradation signal is known as degrons, and they are sometimes found at the termini of a protein. Degrons are recognized by ubiquitin ligases. Upon recognition, the degron-containing protein is tagged with ubiquitin and subsequently degraded by proteostasis machinery. Docking sites, akin to degrons, facilitate substrate recognition by their respective enzymes. However, unlike degrons, the binding of docking site does not result in protein degradation. Docking sites typically bind to enzymes at a location distinct from the active site of the enzymes, allowing for the recruitment of enzymes without interfering with the enzymatic reaction. Finally, the cleavage sites are also a class of motif which is recognized and cleaved by proteases (Van Roey et al., 2014).

The fact that motifs are short and degenerate makes it possible that multiple motifs overlap in a protein region, where they act as molecular switches in biological pathways. Molecular switches are especially prevalent in signaling pathways, where the signaling cascade must be tightly regulated (Van Roey et al., 2013). A classic example of molecular switching mechanisms is the phospho-degrons. For instance, phospho-degrons are found in the protein IKBA, an important regulator of the TNF pathway. Under physiological conditions, the master regulator of immune response, a family of transcription factors named NFKB, is sequestered in the cytoplasm by IKBA. Upon infection, the TNF pathway is activated, bringing about a cascade of signaling event that eventually leads to the phosphorylation of IKBA. The phosphorylation of IKBA at residues serine 32 and serine 36 creates a degradation signal that in turn leads to the degradation of IKBA and the liberation of NFKB (Winston et al., 1999). The unbound NFKB then translocates to the nucleus and activate the transcription of immune response genes. Perhaps not so surprisingly, phospho-degron is also present in NFKB that regulates its own activity (Christian et al., 2016). Along the same line, targeting signals can also be occluded by phosphorylation-dependent binding. Such is the case in the protein ADAM22, where the phosphorylation of its endoplasmic reticulum (ER) retention signal leads to the binding of 14-3-3 proteins, thereby occluding the ER retention signal. Phosphorylation of the targeting signal therefore prohibits the protein from being retained in the ER, and instead, the protein is transported to the cell surface (Gödde et al., 2006).

There are also molecular switching mechanisms that happen before translation, where alternative splicing regulates the production of isoforms that lack certain motifs. For example, the serine/threonine-protein kinase NEK2 involved in cell division has an alternative isoform where the C terminal degron is removed. As opposed to the canonical isoform that is degraded in early mitosis, the absence of the degron allows the alternative isoform to persist until late mitosis (Hames et al., 2001). While motifs tend to bind weakly to their targets, multiple repeats of similar motifs can occur in the disordered region of a protein, and collectively, they give rise to multivalent interactions. Multivalency is a common mechanism that drives the formation of biomolecular condensates in the size of micrometers via phase separation. The role of repeated motifs in phase separation is extensively reviewed in Mitrea and Kriwacki (2016).

Databases of motifs and tools to predict motifs

The Eukaryotic Linear Motifs (ELM) database provides comprehensive information on motifs by manually curating their experimental evidence (Kumar et al., 2024). In addition to the previously mentioned six classes that reflect the biological functions of motifs, ELM further classifies motifs based on their specific sequence characteristics, functional properties, or binding domains. For example, among ligand-binding motifs (LIG class), they are further classified into different types based on their binding domains, such as SH2 domain-binding motifs. ELM refers to these different types of motifs as ELM classes, and individual motif instances are referred to as ELM instances. Such classification of motifs also allows ELM to compile regular expressions for each motif type that it curates. While these regular expressions reflect the specificity determinants of motifs, they also capture the conservation pattern of different motif types. Therefore, these regular expressions can also be used to scan amino acid sequences for putative motifs.

Another approach to capture the conserved patterns of similar motifs is through the use of position-specific scoring matrices (PSSMs). The approach makes use of a motif alignment to statistically model the specificity determinants of the motif by weighing the observations of residues at each position to produce a PSSM. The PSSM can then be used to find and score PSSM matches by quantifying the similarity between the matches and the provided motif alignment. Each approach has its own pros and cons. While PSSM allows matches that are similar to a well-studied motif consensus and further scores the matches, the regular expression approach is more rigid and treats all the matches as equal. Conversely, regular expressions can detect motifs of variable length, but the same can only be achieved with PSSM when multiple PSSMs of different lengths are used (Krystkowiak et al., 2018).

While the use of regular expressions or PSSMs is simple and straightforward, relying solely on them to search for motif matches in the proteome often results in the identification of non-functional matches, such as those occurring in folded domains. This is particularly the case for phosphorylation sites where their regular expressions tend to be so short that only the phosphorylated residue is invariant in the motif. Although this reflects the general biology of PTM sites where the specificity is mainly driven by docking motifs located elsewhere in the protein sequence, it significantly complicates the computational detection of motifs in protein sequences. To overcome this limitation, considerable effort has been invested in improving the accuracy of motif detection algorithms.

Subject to selective pressure to preserve their functionality in biological systems, motifs do not occur at random, and their evolutionary and physicochemical constraints can be exploited to better detect functional motifs. Perhaps the most intuitive property of functional motifs is their conservation across orthologs. With the Relative Local Conservation (RLC) measurement developed by Davey et al. (2012), the conservation of a putative motif can be quantified by scoring how conserved the individual residues in the putative motif are relative to a window of flanking residues using an ortholog alignment.

However, not all sequences have sufficient orthologs for alignment, and aligning disordered regions can be challenging as they exhibit higher evolutionary flexibility than ordered regions (Brown et al., 2002). Therefore, other physicochemical features of putative motifs that do not require sequence alignment can further aid in detecting functional motifs. To this end, as some motifs transition to a secondary structure upon binding to a partner, the tool ANCHOR, developed by the same group as the disorder predictor IUPred, can be used to identify disordered regions with such propensity. Operating on principle similar to that of IUPred, ANCHOR estimates the energy for each residue based on its amino acid type and the amino acid composition of the local sequential neighbourhood. The estimated energies are used to identify residues that only form favorable interaction with their local sequential neighbors upon binding to a globular partner (Dosztányi et al., 2009; Mészáros et al., 2009).

Another tool that serves similar purpose as ANCHOR is known as MoRFpred. Instead of relying solely on energy estimation, this tool computes various features representing the biochemical and physicochemical properties of each residue, such as disorder, solvent accessibility, and flexibility. The computed features are then leveraged to train a support vector machine classifier to predict the likelihood of each residue forming secondary structure upon binding to a partner (Disfani et al., 2012). Many tools have included the aforementioned features in their motif detection algorithms. For example, the webserver SLiMSearch utilizes motif consensus in the form of regular expression to scan a proteome specified by a user for occurrences of the motif. The motif matches are subsequently annotated with different attributes that relate to common properties of known motifs, such as conservation and accessibility. The motif matches, along with their annotations, are then displayed on a web interface for users to inspect (Krystkowiak & Davey, 2017). Simiarly, the webserver PSSMSearch serves a purpose akin to SLiMSearch but utilizes PSSMs instead of regular expressions.

Advancing motif discovery through experiments and bioinformatics

With the structure-function paradigm coming under scrutiny in recent years, it is estimated that more than 100,000 motifs exist in the human proteome, many of which remain to be discovered (Tompa et al., 2014). Comparing the 340 motif types currently annotated in ELM to the more than 21,000 domain families described in InterPro, based on Pfam HMMs, it is evident that we have only touched the surface of motif biology with what we know so far compared to folded domains.

On the discovery of new motifs, high-throughput screens aimed at expanding the repertoire of known motifs or identifying novel motifs can help to refine or define the sequence patterns of the screened motifs. For example, Benz et al. (2022) utilizes phase display library method to systematically screen for binding peptides in the unstructured regions of the human proteome against a selection of motif-binding domains. In spite of clear limitations of the experimental technique, which is *in vitro* and only protein fragments used, the results helped refine the sequence patterns of the screened motifs and unveiled more than 2,000 human PPIs mediated by motifs, among which close to 500 are mediated by poorly studied proteins.

Besides, given the challenges associated with experimenting on IDRs, the discovery of motifs would also greatly benefit from a combination of computational and experimental approaches. Take the identified motifs from the study of Benz et al. (2022) for example, apart from pattern refinement, the identified motifs can also be used to train machine learning algorithms that can further improve motif prediction accuracy. With the growing awareness of the functional role of IDRs, the disorder-function paradigm has emerged to explain the functionality of protein regions that lack a defined structure. Needless to say, with the upscaling of experimental methods to higher throughput and the incorporation of computational tools to detect motif occurrence, the field of motif biology will continue to expand in the forthcoming years.

1.2 Protein-protein interaction interfaces

Earlier research on PPI interfaces discovered that while hydrophobicity stabilizes PPIs, shape and charge complementarity are crucial for the specificity of PPIs (Chothia & Janin, 1975). Hydrophobicity contributes to the majority of the binding free energy, and its effect is mediated through the exclusion of water upon the binding of proteins. Upon the binding of two proteins, the entropy gained by water resulting from a reduced accessible protein surface area is sufficient to compensate for the entropy lost by proteins. As the residues at PPI surfaces are as tightly packed as those inside a protein, the sidechains of surface residues creates protrusions on the surface. With these protrusions being matched by depressions on the binding partner, the shape complementarity of PPIs contributes to the binding specificities of PPIs. On a more fine-grained scale, the charged and polar residues at PPI interfaces also play a similar role as shape complementarity as they ensure the proper formation of salt bridges and hydrogen bonds between the proteins.

As mentioned previously, the conserved modules in proteins are key players in mediating PPIs, and thus they form PPI interfaces. Of note, many domains, especially those found in hub proteins, have multiple interfacial surfaces that enable them to bind to different partners (W. K. Kim et al., 2006). The interfaces formed by interaction between two domains or interaction between a domain and a motif are termed as domain-domain interfaces (DDIs) and domain-motif interfaces (DMIs), respectively. The properties of these two types of PPI interfaces, as well as available databases to study them, are discussed in detail in the next sections. Other types of interfaces also exist, and are briefly discussed in the section that follows. Different computational methods have been developed to detect DDIs and DMIs in interacting proteins, and they are discussed in the section section 1.3.

1.2.1 Domain-domain interfaces

Domains can interact with other domains or another copy of itself to form DDIs. DDIs commonly involve multiple specific contacts that mediate interactions of relatively high affinity, such as those found in stable protein complexes. A comparison between amino acid frequencies at domain interfaces and domain cores has revealed that, while they share physical properties such as charge and hydrophobicity, they differ in structural composition and packing (Hadarovich et al., 2021). In other words, the specificity determinants of DDIs mainly lie in the structural composition and packing of the domains. Indeed, the packing of the indole sidechain of a tryptophan at the DDI of the dimeric glutathione transferase is essential for the interaction and the catalytic function of the transferase (Wallace et al., 2000).

The higher-order assemblies formed by DDIs are important for signaling processes.

For example, the Toll/interleukin-1 receptor (TIR) domains are capable of interacting with another copy of themselves. Upon the detection of pathogens, Toll-like receptors (TLR) dimerize through their TIR domains. The dimerized TIR domains further recruit more TIR domain-containing adaptor proteins, such as MyD88 and MAL, as shown in Figure 1.5. Following the formation of the signaling complex, more factors are recruited and eventually pro-inflammatory genes are activated (Clabbers et al., 2021). Apart from mediating complex formation, DDIs are also involved in other processes. For example, the junctions of DDIs can serve as catalytic sites for oligomeric enzymes such as glutamine synthetase (Joo et al., 2018) and catalase (Zámocký & Koller, 1999). DDI can also be important for substrate recognition, as in the case of the transcription factor NarL where its dimerization mediated by a DDI is necessary for DNA binding (Maris et al., 2002). In the case of multi-domain proteins, domains that are tethered together in a single polypeptide chain by a flexible linker can also form intra-chain DDIs. The DDIs formed by the neighbouring domains can alter the folding pathways of the domains and therefore impact the stability of the protein (Dias et al., 2023; Flaugh et al., 2005; Levy, 2017).

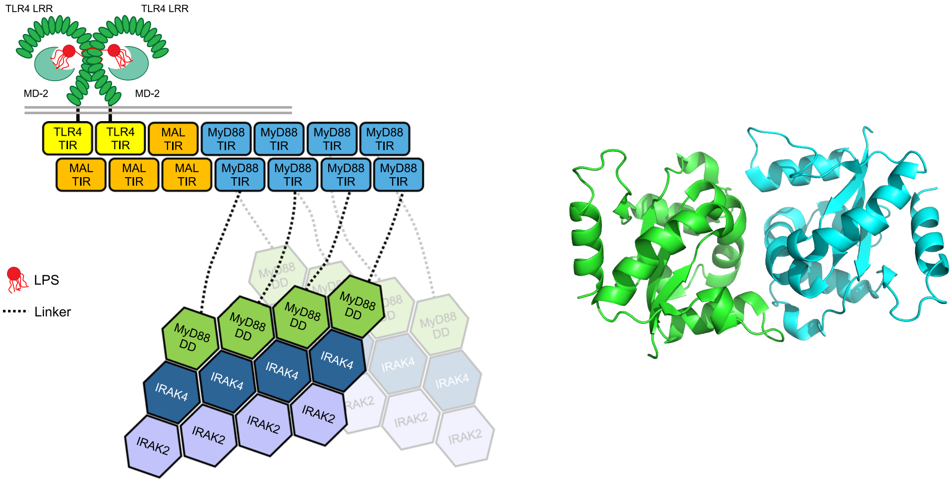


Figure 1.5: Higher-order assembly initiated and formed by the interaction between TIR domains of different proteins upon the binding of pathogen-associated molecular patterns to the LRR domain of Toll-like receptors. LRR, leucine-rich repeat domain; LPS, lipopolysaccharide; TIR, Toll/interleukin-1 receptor domain; DD, death domain. From (Clabbers et al., 2021). The right panel displays a DDI formed by TIR domains from Toll-like receptor 6 (PDB ID: 4OM7). Individual TIR domains are colored in green and cyan.

Databases of domain-domain interfaces

Owing to the stable interactions mediated by DDIs and the folded nature of domains, three-dimensional structures of many protein complexes formed through interactions between domains have been solved and deposited in the PDB. The contacts observed between the domains solved in these protein complexes have thus far served as a valuable source of experimental evidence for many DDIs. Several databases have utlized this evidence to annotate and catalog DDIs. The iPfam database archives interactions between Pfam domains based on known three-dimensional structures deposited in the PDB. The database first annotates the protein sequences of PDB chains with Pfam domains using HMMs. Subsequently, the database analyzes the contacts between different domains in the PDB chain to define the types of bonds between the annotated domains. The types of bonds include covalent bonds, electrostatic bonds or salt bridges, hydrogen bonds and van der Waals interactions, all of which are based on the geometric and chemical properties of the atoms involved. Of note, an interaction between two domains are defined as the presence of at least one bond between the two domains. With all this information stored in the database, iPfam allows users to inspect specific interactions between domains by visualizing them in two-dimensional graphical displays or three-dimensional PBD structures. Alternatively, users can submit a protein sequence to the database to search for potential DDIs that can be mediated by the domains in the protein. To achieve this, the server first annotates user-provided protein sequence with Pfam domains, followed by conducting a search within the database to look for DDIs mediated by the detected domains. The results are then displayed on a webpage for users to inspect. The last update of iPfam was in early 2014, and as of the time of writing this thesis, the database appears to be outdated (Finn et al., 2014).

Another database that operates o a similar principle is the database of 3D Interaction Domains (3did) (Mosca et al., 2014). In addition to contact annotation, 3did also scores the contacts by means of empirical potentials derived from a benchmark dataset that consists of enzyme inhibitors, signaling complexes, cytokine receptors, antibody/antigen complexes, and MHC complexes. Furthermore, statistical significance for the annotated contacts is estimated based on a background of random sequences (Aloy & Russell, 2003). it is worth noting that, unlike iPfam, 3did requires a minimum of five contacts detected between a domain pair to consider the structure as evidence for a DDI. Beyond DDIs, 3did also hosts a collection of DMIs with structural evidence by implementing an elaborate DMI discovery pipeline to annotate structurally-solved DMIs (Stein & Aloy, 2010). At the time of writing this thesis, 3did continues to receive routine updates and has cataloged more than 15,000 types of DDI, with DDI contact annotations derived from over 700,000 structures.

Indeed, the abundance of DDIs annotated in 3did represents a valuable resource for predicting DDIs in PPIs, as demonstrated by numerous studies (Karan et al., 2023; Luther et al., 2023; Mosca et al., 2013; Wang et al., 2012). Nonetheless, a closer examination on 3did revealed that many DDIs are supported exclusively by intrachain evidence, meaning that the contacts between the domains are observed only when the two domains are linked together in a single protein chain. While there are studies supporting the notion that intrachain and interchain DDIs share similarities in terms of structure and physicochemical properties (Jones et al., 2000; Verma & Pandit, 2019), domains tethered by a linker are more likely to associate with each other due to an increase in the effective concentration of the linked domains (Dyla & Kjaergaard, 2020; Sørensen & Kjaergaard, 2019). That said, DDIs with only intrachain evidence may need additional evidence to assess their ability to mediate PPIs under physiological conditions. Furthermore, a substantial amount of DDIs are observed to be in contact only in homodimers. The symmetry of homodimers in crystal structures makes it challenging to determine whether an observed interface is a genuine dimerization interface or a non-specific interface induced by crystallization. Therefore, interpreting DDIs from homodimers may require additional information such as sequence conservation at a given interface and interface size to distinguish between crystal artifacts and biologically relevant interface (Xu & Dunbrack, 2019). While 3did undoubtedly offers a wealth of valuable information on potential DDIs for biologists, it is important to remain mindful of the above-mentioned caveats.

Apart from annotating observed contacts in PDB structures to catalog DDIs, many computational approaches have been developed to predict DDIs using various sources of data, and they will be discussed in the later section **Sequence-based detection of binding regions**. With all these DDIs predicted from various data sources, there is a need to store them in a way that allows public users to easily assess them. The DOMINE database was created to collect both experimentally supported DDIs and those predicted by computational methods (Yellaboina et al., 2011). Besides gathering DDIs with experimental evidence from iPfam and 3did, the database also includes predicted DDIs from 12 other sources. Through a friendly web interface, users can browse the collection of DDIs in the database by querying with a Pfam domain name or a GO term. DOMINE was last updated in 2010 and currently contains a total of 26,219 DDIs mediated by 5,410 domains, with over 21,000 predicted by at least one computational method.

1.2.2 Domain-motif interfaces

Most motifs mediate their functions by binding to globular domains found within the same protein chain (intramolecular) or another protein chain (intermolecular), forming the so-called DMIs. While interactions between domains often result in the formation of stable complexes, interactions involving domains and motifs tend to lead to transient complexes that play a pivotal role in regulatory functions. Since motifs exist as flexible stretches of amino acid in their unbound state, their binding to a partner often results in a significant entropy loss. The significant entropic penalty incurred upon binding to a partner makes the interactions that they mediate transient and reversible, both of which are crucial in the regulatory network of a cell (Flock et al., 2014).

Intramolecular DMIs commonly function as an auto-inhibitory mechanism to regulate a protein's activity. An example of an intramolecular DMI can be found in the WASP protein. The WASP protein contains a GTPase-binding domain (GBD) at the N terminus, which is important to initiate actin nucleation, and a C terminal motif crucial for regulating the activity of WASP. Under basal conditions, the C terminal motif binds to the GBD domain, closing the protein in an end-to-end manner, thereby inhibiting its function in promoting actin nucleation (A. S. Kim et al., 2000). In contrast, intermolecular DMIs are involved in various biological functions, as mentioned in the **Short Linear Motifs** section.

Domain-motif binding relationships can be promiscuous, with many domains exhibiting a one-to-many binding relationship with various motifs. Many of these domains originate from large domain families, and they possess capacity to bind multiple types of motifs with varying degrees of binding preference. Considering that DMIs mediate interactions that are tightly regulated in signaling events, the seemingly promiscuous binding tendencies of these motif-binding domains raise questions about their specificity. One way to ensure binding specificity is by using different motif-binding pockets. Due to the folded nature of domains, some domains expose different motif-binding pockets on its exterior, permitting different motifs to bind at distinct sites. For example, the activator protein CDH1, which regulates the ubiquitin ligase activity and substrate specificity of the anaphase-promoting complex, APC/C, possesses a WD40 repeat domain that can bind to different types of degron motifs through distinct binding pockets (Figure 1.6). The degron motifs include the KEN-box, the D-box and the ABBA motif (Di Fiore et al., 2015; He et al., 2013; D. Lu et al., 2014).

Although the use of different binding pockets explains motif-binding specificity of the WD40 repeat domain, in the case of many other domains, such as kinase and SH2 domains, the same binding pockets are used to bind different types of motifs (Huang et al., 2008; Mok et al., 2010). This hints at the presence of additional factors that determine the motif-binding specificity of these domains. The key to their substrate specificity can be found in the physicochemical composition of their binding pockets. For instance, the sequence composition of the binding pockets in SH2 domains has undergone substantial divergence through gene duplication events. Mutations accumulating in the binding pockets of these duplicated copies of SH2 domains can help fine-tune their substrate specificity. Over time, this evolutionary process refines the individual duplicated copies of SH2 domains and enables them to specialize in binding phosphorylated tyrosines flanked by distinct amino acids (Marasco & Carlomagno, 2020). Similarly, kinase domains have also evolved different catalytic clefts to accommodate either bulkier tyrosine substrates or smaller serine/threonine substrates

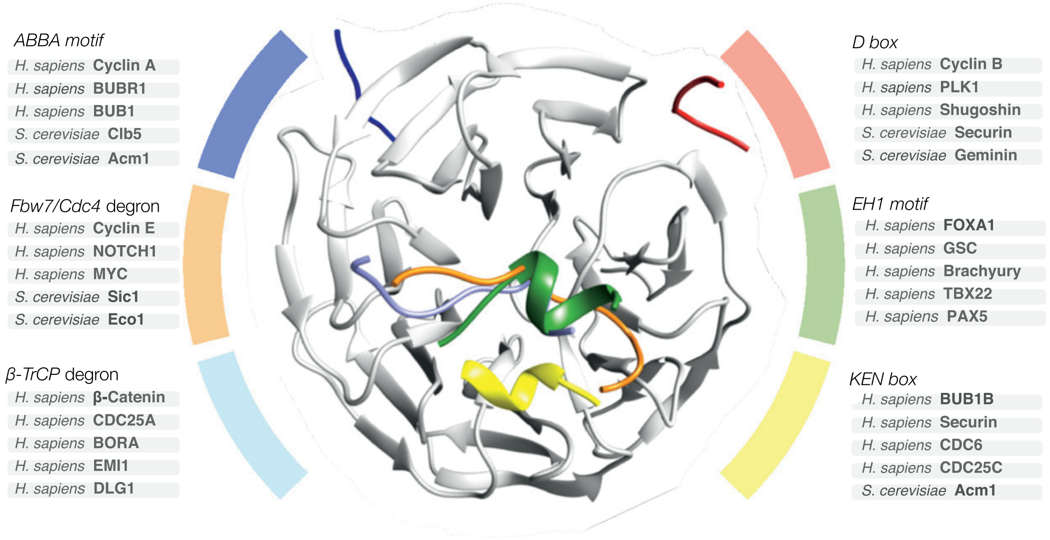


Figure 1.6: The binding of the WD40 domain of CDH1 to different motifs occurs through distinct surfaces. Representative examples of motifs are listed on the left and right, and the motifs in the structure are colored to match the colors of the listed motifs. From (Davey et al., 2015).

(Ubersax & Ferrell Jr, 2007). Subsequent gene duplications, followed by diversification of these kinase domains, resulted in increased specificities toward substrates with even more nuanced sequence characteristics (Howard et al., 2014). Factors external to a DMI can also enhance substrate specificity, such as the presence of docking motifs in the substrate that helps increase the local concentration of the substrate in the vicinity of a motif-binding domain. This is exemplified by the binding of CDK2-cyclinA complex to its substrate, where the binding of the RxL docking motif precedes the binding of the phosphorylation motif. This sequential binding process ensures the specificity of the phosphorylation (Ubersax & Ferrell Jr, 2007).

Since motifs interact with domains on a relatively small interface, their binding can be very prone to subtle changes in the amino acid composition or the shape of the motif-binding pockets. As a result, motifs also co-evolve with domains under the same selective pressure for specificity. For example, phosphorylation motifs targeted by different mitotic kinases have been observed to be under opposing selective pressures on specificity-determining residues. These opposing selective pressures ensure that phosphorylation by different mitotic kinases remains exclusive to different phosphorylation motifs (Alexander et al., 2011). The co-evolution of domain and motif is particularly salient in the case of the GYF domain binding to proline-rich sequences. Although both proteins, CD2B2 and GIGYF1, possess a copy of the GYF domain, they are found to bind to proline-rich sequences (PPPG-) that are followed by either a hydrophilic or hydrophobic residue, respectively. Closer inspection of their solved structures bound to proline-rich sequences, 1L2Z for CD2B2 and 7RUQ for GIGYF1, revealed that, albeit similar in overall structure, their GYF domains differ by a crucial hydrophobic cavity that dictates their ability to bind proline-rich sequences followed by a hydrophobic residue (Figure 1.7) (Freund et al., 2002; Sobti et al., 2023).

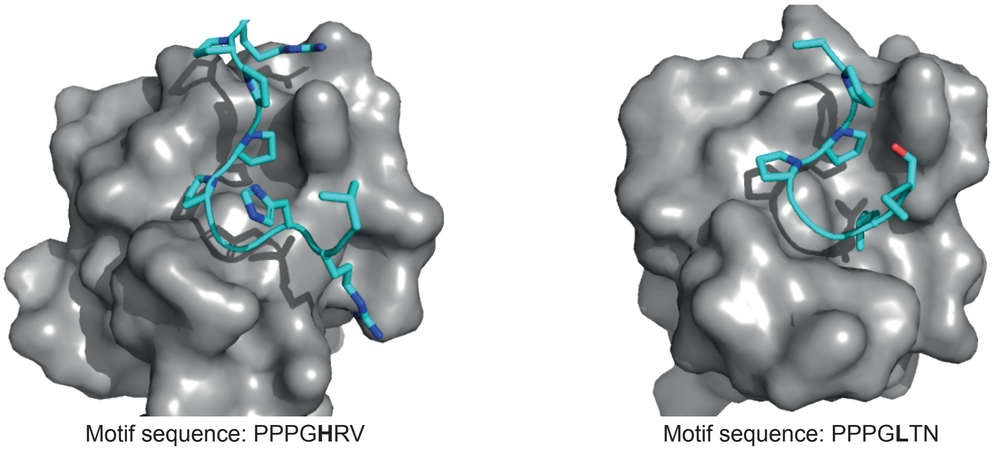


Figure 1.7: The binding of the GYF domains of CD2B2 (left) and GIGYF1 (right) to proline-rich sequences (PDB IDs: 1L2Z for CD2B2 and 7RUQ for GIGYF1). The domains are shown in grey, and the sidechains of the motifs are displayed in sticks. Notice in 7RUQ (right) that the leucine in the motif fits into a cavity in the GYF domain of GIGYF1, which is not present in that of CD2B2.

Databases of domain-motif interfaces

In the adaptation of cellular machinery to increasing complexity in eukaryotes, the degenerate nature of motifs and the large variety of motif-binding domains in the proteome serve as the sandbox for evolution to tweak and fine-tune regulatory networks. Indeed, many interfaces involving motifs binding to folded domains have been discovered, and they are annotated in the ELM database mentioned earlier in the **Short Linear Motifs** section. However, in comparison to the vast amount of potential DDIs deposited in DDI databases, the number of annotated DMIs in the ELM database is much more limited, with almost 4,300 DMIs annotated so far. Regardless, the ELM database represents a highly comprehensive DMI database because the annotated DMIs are manually curated by experts who corroborate the validity of evidence that often comes from different experimental methods. The majority of DMI instances annotated in the ELM database lack structural information, and mutation analyses constitute the bulk of the experimental evidence. The absence of structural information can be attributed to the transient nature of interactions mediated by DMIs and the disordered nature of motifs, both of which make crystallographic studies on DMIs an arduous task.

Since the binding of motifs is particularly contingent on the shape complementarity and physicochemical composition of the motif-binding pockets, structural information on DMIs can provide important mechanistic insights into DMI-mediated interactions. Furthermore, as motifs of the same type (or of the same ELM class) tend to bind to their binding domains in a similar manner, structural insights from one DMI can be readily transferred to another DMI formed by a different motif of the same type. That said, using modeling software to leverage previously solved DMI structures to predict the binding interfaces of new but similar DMIs can help overcome this challenge. Becuase DMI-mediated interactions frequently participate in transient signaling events, the investigation of DMIs will benefit from *in vivo* experimental methodologies capable of detecting weak interactions in real-time. In section 1.4, I will discuss experimental methods available to detect PPI interfaces mediated by DMIs.

1.2.3 Other types of interfaces

Other types of interface exist, such as disorder-disorder interfaces and coiled-coil interfaces. Examples of disorder-disorder interfaces includes the binding between the linker histone protein H1, a highly disordered and positively charged protein with one folded domain flanked by disordered regions, and the nuclear protein prothymosin-alpha, a fully disordered and highly negatively charged protein. Mediated by their oppositely charged disordered regions, the two proteins interact with high affinity and retain their structural disorder even during the interaction (Borgia et al., 2018). Unfortunately, current knowledge about interfaces formed exclusively by disordered regions is limited compared to DDIs and DMIs due to the challenges in experimentally studying IDPs and IDRs.

Coiled-coil interfaces are formed by bundles of amphipathic alpha helices that wind into superhelical structures to bury their hydrophobic faces while exposing their hydrophilic faces. Unlike regular alpha-helices with a turn periodicity of 3.6 amino acids, helices capable of forming coiled-coil interfaces have a turn periodicity of 3.5 amino acids. These helices are formed from heptad repeats, where the first and the fourth amino acids are hydrophobic while the others are polar. The length of coiled-coils plays a crucial role in their molecular functions, such as serving as spacers to separate functional domains or position substrates at a defined distance for enzymatic reactions. Beyond serving as molecular spacers, coiled-coils are also important for various molecular processes, including oligomerization, DNA binding and cleavage, and vesicle tethering (Truebestein & Leonard, 2016). While coiled-coil interfaces are involved in a significant number of molecular processes, crystallographic studies on them are often hampered by their natural tendency to pack together. This tendency complicates the interpretation of whether the packing is biological or induced by crystallization.

In the next section, I will discuss different strategies for computationally characterizing PPI interfaces.

1.3 Computational strategies to predict protein interaction interfaces

Many proteins interact with other proteins in their full-length forms to fulfill their functional roles. Moreover, being the products of gene duplication and fusion events, many proteins consist of multiple functional modules that act synergistically to perform their functions. Therefore, in an ideal scenario, the structural determination of PPIs should be conducted using full-length proteins, as this approach would yield the most comprehensive mechanistic information on the PPIs. However, the IDRs of proteins are often too flexible to be crystallized for structural determination. As a result, many protein complexes have been solved by fragmenting the proteins down to their minimal interacting regions. This is particularly the case for DMIs as the complexation of disordered peptides with their binding domains can help anchor and fix the peptides for structural determination using crystallography. Such a fragmentation approach is rather reductionistic as it overlooks the synergistic effects of the functional modules. Thus, for a more holistic understanding of a protein's function and to more efficiently characterize the interaction interface between PPIs, it is advantageous to employ a combined approach of sequence-based analysis and structural modeling.

The combined approach consists of two steps: 1) using a sequence-based interaction interface prediction tool to detect potential interacting regions at the residue level between a given pair of interacting proteins, 2) structurally modeling of the potential interacting regions to pinpoint important inter-atomic interactions between residues at the interface. To validate the predicted interaction interface, the modeled interface can guide the design of mutations intended to perturb key contacts between the interacting regions. A reduction in binding strength or the absence of binding between the mutated proteins relative to the wild-type proteins is then indicative of the modeled interface being responsible, or at least partially responsible, for mediating the interaction. This approach effectively characterizes PPI interfaces without the need to solve the three-dimensional structures of PPIs.

Databases like 3did and ELM annotate information about known interfaces. To achieve step 1, these known interfaces can be leveraged to predict their occurrences in interacting proteins In the following sections, I will discuss a method developed by others to achieve step 1. I have also developed methods to achieve step 1, and they are showcased in Chapter 2 and Chapter 3, Article I. Following step 1, step 2 consists of structurally modeling the predicted interfaces. Some options available to achieve that purpose will be discussed. As mentioned earlier, many PPI interfaces await characterization due to various reasons. To this end, I will discuss the use of AlphaFold-Multimer (AF-MM), an AI-based structural modeling tool, to predict novel PPI interfaces in interacting proteins. This strategy effectively achieves both steps 1 and 2 concomitantly.

1.3.1 Sequence-based prediction of binding regions

Sequence-based detection of domain-motif interfaces

The detection of DMIs largely depends on accurately predicting the occurrence of functional motifs. As discussed in the section **Short linear motifs**, much work in the field of motif discovery has been dedicated to improving the accuracy of motif detection algorithms. Although these properties of motifs are instrumental in assessing the confidence of a motif match, it is important to note that they are not restricted to specific interactions. To find DMIs in specific interactions, the domains detected in the interacting partner need to be considered.

The interactions of Eukaryotic Linear Motif (iELM) webserver allows user to explore the PPIs of a user-submitted protein that are potentially mediated by DMIs. iELM achieves this by querying the PPI database STRING with the user-submitted protein and performing a DMI search on the returned PPIs. Alternatively, users can also provide the webserver with a set of PPIs, and potential DMIs are searched within the provided set of PPIs. iELM tackles the challenge of detecting DMIs using a two-pronged approach. First, to better detect motif-binding domains, experimentally annotated motif-binding domains are aligned with their orthologs to produce HMMs specific to the motif-binding domains. Second, putative motifs are detected by combining the motif discovery tool SLiMSearch (Krystkowiak & Davey, 2017) and the disorder predictor IUPred (Dosztányi et al., 2005). The E-value returned from the HMM match of motif-binding domains, along with the conservation score and disorder score of detected motifs, is used to train a support vector machine to evaluate the detected DMIs. Detected DMIs with template available in the PDB database are further subjected to structural modeling by PepSite (Petsalaki et al., 2009). The modelled DMIs are further used by PepSite to score the biophysical feasibility of the detected DMIs. Of note, the iELM algorithm terminates further DMI search if potential DDIs from 3did are found between a given pair of interacting proteins. The authors of iELM reported an impressive performance: 84.8% sensitivity and 86.5% specificity on its test set. Nonetheless, the benchmark could be overly optimistic, as the negative data points in the test set outnumbers the positive data points by almost 30-fold. In this scenario, the precision of iELM would be more informative, yet it was not reported in the publication (Weatheritt, Luck, et al., 2012; Weatheritt, Jehl, et al., 2012). As of the time of writing this thesis, the iELM webserver is no longer in use.

Sequence-based detection of domain-domain interfaces

PPI data, apart from its purpose to study biological functions of proteins, have also been used to infer DDIs. The underlying assumption is that if two domains are able to interact and mediate PPI, then there should be an enrichment of PPIs containing the domain pair in PPI networks. This is also known as the association method developed by Sprinzak and Margalit (2001). The method defines interaction between two domains as the fraction of PPIs containing the domain pair among all PPIs, with overrepresented domain pairs indicating putative DDIs. To identify overrepresented domain pairs, the ratio between the observed and expected frequencies of PPIs with a given domain pair is computed and expressed as a log-odds value. While this is a simple and sound approach to identify putative DDIs, the high log-odds values of domain pairs involving domains that are rare in the PPI network should be interpreted with caution. This is because their high log-odds values could simply be due to the very low expected frequencies of the rare domains.

Several other studies have also investigated the idea of inferring DDIs through PPI networks using different approaches. Deng et al. (2002), for example, used a probabilistic model to infer DDIs using PPI data from yeast proteins. They first annotated the proteins in the PPI data with Pfam domains, followed by applying the maximum likelihood approach to estimate the probability of DDIs from the PPI data. Owing to the lack of known DDIs at the time of publication, their approach to evaluate the model was to use the DDIs inferred by the model to predict PPIs, and the accuracy of model was evaluated as the accuracy of PPI prediction. To supplement more PPI data, likely PPIs were generated by searching for proteins with similar expression profiles. Nonetheless, the probabilistic model to infer DDI from PPI data suffered from low specificity, possibly due to limited amount of PPI data, as reported by the authors. Through a machine learning approach, other works have also leveraged PPI data to identify potential DDIs (X.-W. Chen & Liu, 2005; X.-M. Zhao et al., 2010). In brief, a set of PPIs was used as positive samples and a set of randomly generated PPIs was used as negative samples. Following that, all possible domain combinations were treated as features to discriminate between PPIs and non-PPIs. The likelihood of a domain combination being a putative DDI was then assessed based on its contributions to the discrimination of PPIs and non-PPIs. In essence, the inference of putative DDIs was formulated as a feature selection task to predict PPIs. While X.-W. Chen and Liu (2005) utilized a random forest (RF)-based method for this purpose, X.-M. Zhao et al. (2010) developed a simple classifier that additionally downweighted DDIs occurring in non-PPIs, as they were more likely to be false DDIs.

Other data sources, such as the coevolution of sequences of interacting proteins (Kann et al., 2007), gene ontology (GO) (Liu et al., 2009) and phylogenetic profiling

(Pagel et al., 2004), have also been utilized for predicting DDIs.

1.3.2 Structure-based prediction of protein interaction interfaces

Since many domains interact with their partners through amino acids that are not contiguous in the protein chain, the application of sequence-based approaches is limited when it comes to capturing the detailed mechanisms of PPIs. A detailed mechanistic description of PPIs entails uncovering information like the binding surface on a predicted domain and the sidechain orientation of interfacial residues, both of which can only be revealed through structural modeling. Furthermore, as DMIs are especially prone to subtle differences in binding pockets, structural models of DMIs can also greatly aid in inspecting the validity of DMIs predicted from sequence-based approaches. In the following sections, I will discuss different options available to structurally model PPI interfaces.

Protein docking interface prediction

Analogous to the folding of a protein, packing of secondary structures between proteins that lead to PPIs are also governed by physical forces. There exist tools that can model PPI interfaces from the monomeric structures of the interacting proteins based on similar physical principles. Known as protein docking, many computer programs have been developed to optimally 'piece' proteins together as a means to study their interaction interface. The idea of protein docking traces back to as early as 1978 (Wodak & Janin, 1978). The most common approach to docking two proteins is referred to as the rigid-body docking approach. This approach involves fixing one protein while rotating and translating the partner around the fixed protein repeatedly to generate putative protein complexes. Simultaneously, physical-chemical force fields are applied to identify favorable interaction sites on the surfaces of the proteins. The putative protein complexes are then scored using a variety of metrics that generally reflect the shape and physicochemical complementarity of the proteins. The iterative process of rigid-body docking makes the method time-consuming and difficult to scale up for higher throughput. Furthermore, the accuracy of rigid-body docking approach is largely hindered by the dynamic nature of proteins, particularly the conformational changes in proteins induced by their binding partners. While most docking programs nowadays can account for protein sidechain flexibility to some extent, perhaps the most practical use of docking algorithms is on the conformational ensemble of proteins obtained from NMR studies (Kaczor et al., 2018).

Homology-aided interface prediction

Given that most proteins arise from the concatenation and recombination of conserved functional modules in the genome, the idea of inferring the three-dimensional structure of a PPI from available structures of homologous proteins is not new to the field of computational structural biology. However, the lack of experimentally solved structures was a significant stumbling block to comparative modeling. With the growing number of experimentally solved protein complexes, only in recent years has the concept of comparative modeling for PPIs become more feasible. To transfer structural information from one PPI to its homologs, the key question is to what extent the interaction interface is conserved among homologs with varying degrees of sequence similarity. Aloy et al. (2003) used an extensive set of domain-domain interaction structures to study the relationship between sequence similarity and interaction conservation in proteins. It was shown that proteins with more than 30-40% sequence identity tend to interact in the same way. Importantly, given that a binary interaction query leads to two sequence comparisons to their homologs that returns two sequence identities, the reported sequence identity refers to the minimum sequence identity among the two sequence comparisons. This crucial insight paved the way for comparatively modeling.

Building on this discovery, Interactome3D, a tool that models the three-dimensional structure of a given PPI based on solved structures of homologous proteins was developed (Mosca et al., 2013). The automated homology modeling pipeline begins with querying several databases to collect structures for the individual proteins in a user-provided interaction, followed by structures of interactions that share over 30% sequence identity with the user-provided protein pair to serve as templates. The pipeline then evaluates the retrieved structures based on sequence similarity and coverage, as well as the resolution of the structures. This assessment identifies the most suitable templates for subsequent modeling by Modeller (Sali & Blundell, 1993). Of note, if no experimental structures or modeling templates are found for an interaction, the pipeline resorts to detecting domain matches by Pfam HMMs in the interactors and identifying potential DDIs between the interactors that are annotated in 3did. The structures of the potential DDIs then serve as the templates for the modeling of the interaction. The tool is accessible to the public via a web interface and has the capacity to model up to 3,000 interactions per day.

Machine learning-boosted interface prediction

While there are tools that model the interface between an interaction, there are also tools that predict interfacial residues on the individual interactors. The latter is the objective behind the development of the tool PredUs which relies on structural neighbors to predict interfacial residues on monomeric protein structure (Zhang et al., 2011).

Structural neighbours are defined as those that belong to the same family, superfamily, or fold, based on their classifications in the SCOP database (Andreeva et al., 2008). In addition to structural neighbours defined by SCOP, structural neighbours detected by means of protein structure distance calculated from the structural alignment program Ska are also included (Yang & Honig, 2000). In brief, given a protein structure submitted by users, the tool identifies the interfacial residues of the protein structure by searching for structural neighbours of the query protein involved in a complex and subsequently mapping the interfacial residues of the neighbours onto the query protein. For each structural neighbour, the mapped interfacial residues are used to generate a contact map. Summing the individual contact maps produces a contact frequency heatmap where higher contact frequency alludes to more conserved interface (Zhang et al., 2010). The webserver of PredUs includes a support vector machine that uses the contact frequency and solvent accessible surface areas of a given residue to further score the likelihood of the residue being at an interface. A few years later, PredUs was upgraded to PredUs 2.0 with higher accuracy. The update incorporated additional features that relate to surface patches into the program using a Bayesian network (Hwang et al., 2016).

Despite originally intended for PPI prediction and not interface prediction at the residue level, the tool PrePPI has been shown to be able to predict PPI interfaces (Zhang et al., 2012). To predict the interaction between a pair of user-submitted proteins, PrePPI operates on a principle akin to that of Interactome3D, which models the interface between the protein pair based on homology. The server first searches for representative structures of the input sequences in PDB, including, if available, the structures of the input sequences, and the structures of their homologs. Then interacting chains showing the structural neighbours of the representative structures are also searched in PDB to serve as templates for the modeling of the hypothetical structure of the input sequences in complex. The modeling is accomplished by superimposing the representative structures on their corresponding structural neighbours in the templates. Different scores are extracted from the modelled structure, including the aforementioned PredUs score. These scores are then combined with other metrics such as co-expression and functional similarity, among others, to predict the likelihood of interaction between the protein pair. Although in the pipeline of PrePPI, interface modeling primarily serves as a proxy to predict the likelihood of two proteins interacting, the approach illustrates how existing structures in PDB can be leveraged to model PPI interfaces or approximate interfacial residues. Indeed, the publication also reported the prediction of a crucial residue at a modelled interface, which was later verified in an experimental setting. The webserver of PrePPI allows user to inspect the modelled interface of a query sequence pair. However, only the highest-scoring interface is displayed, which limits the exploration of PPIs mediated by multiple interfaces.

The first implementation of PrePPI excluded interfaces involving unstructured regions owing to the lack of structural data for this type of interface. Subsequent updates to the tool incorporated this type of interface as well as the use of structures predicted by AF in the interface modeling pipeline (T. S. Chen et al., 2015; Petrey et al., 2023).

The field of AI has been making rapid strides in recent years, with algorithms that surpass even human capability being developed at an unprecedented rate. With its ability to identify patterns in existing data and applying them on unseen data to make predictions, it was only a matter of time before AI extended its reach into structural biology, a realm where data is plentiful. Though not the first tool to employ AI in its workflow, the year 2020 marks an important breakthrough in the field of computational structural biology as the tool AF from Google Deepmind won the 14th Critical Assessment of Structure Prediction (CASP14) competition by achieving prediction accuracies that are on a par with experimentally solved structures (Jumper et al., 2021; Pereira et al., 2021). The success of AF lies in its neural network architecture that is designed to capture the evolutionary, physical, and geometric constraints of protein structures. Trained on a vast amount of experimentally solved structures, the network is capable of predicting the three-dimensional structures of proteins using solely their amino acid sequences. In an effort to make the predicted structures available to the public, a database has been set up that contains more than 360,000 predicted structures across 21 model organism proteomes, with the structural prediction of many other proteins still ongoing (Varadi et al., 2022). Undoubtedly, the outstanding performance of AF in predicting the monomeric structures of proteins has garnered a significant amount of attention in the structural biology community, and many are eager to explore further use of AF in areas beyond the structural prediction of protein monomers.

The success of AF in accurately predicting the tertiary structures of proteins suggests that AF has learned to emulate the process by which fundamental physical forces drive protein folding. Since complex formation is merely the packing of secondary structures that is similarly governed by physical forces, it is logical to think that AF may also be capable of predicting the structures of protein complexes. Indeed, it did not take long until researchers devised a way to hijack AF for complex structure prediction. This is achieved by concatenating protein sequences using poly-glycine linker and pairing the MSAs of the individual sequences. This approach enables AF to extract interchain co-evolutionary information for the structural prediction of the proteins in complex (Bryant et al., 2022; Mirdita et al., 2022; Tsaban et al., 2022). Using this approach, another group of researchers compared the accuracy of AF with a traditional shape complementarity-based docking method on complexes that involved DDIs and found that AF significantly outperforms the docking method. Additionally, they also performed similar test on 14 complexes that involved disordered regions and found that AF generally performed well at predicting these complexes. However, among these complexes, they noted that successful cases tend to be those driven by largely hydrophobic interactions. On the contrary, the symmetric nature of the disordered regions in these complexes has a detrimental effect on the performance of AF (Akdel et al., 2022).

One year after the development of AF, Google Deepmind introduced a spin-off version called AF-MM, that specializes in protein complex prediction (Evans et al., 2022). To facilitate the extraction of interchain evolutionary information, AF-MM adopts an approach similar to that of Bryant et al. (2022), deemed FoldDock, which pairs the MSAs of the individual chains. On top of that, AF-MM was specifically trained on protein complexes to further enhance its ability to predict the structures of protein complexes. Indeed, Evans et al. (2022) reported that the prediction accuracy of AF-MM stood out as the highest among other AF-based complex prediction methods, including the method that concatenates the individual chains with a linker for complex prediction using AF ((Mirdita et al., 2022)) and the method that docks AF-predicted monomeric structures using ClusPro ((Kozakov et al., 2017)). While the comparison of prediction accuracy between AF-MM and FoldDock would be intriguing as they employ a similar approach for complex prediction, a direct comparison between the two is challenging as AF-MM was developed using the dataset on which FoldDock was tested (Bryant et al., 2022).

While the developers of AF-MM reported impressive performance, a closer inspection of the benchmark dataset revealed that it was composed of only complexes mediated by DDIs. Considering the observation noted by Akdel et al. (2022) regarding the influence of the nature of the interface on AF's performance, it is crucial to systematically benchmark the performance of AF-MM using different types of interfaces, such as DDIs and DMIs. The rationale behind conducting a systematic benchmark is twofold. First, since AF-MM was trained on structures deposited in PDB, DDI structures are likely overrepresented in its training data. To illustrate such imbalance of interface types in PDB structures, 3did contains DDI annotations from over 700,000 structures and DMIs from over 12,000 structures (Mosca et al., 2014). Second, although different studies have tested the ability of AF and AF-MM to predict interfaces involving disordered regions, there lacks a comprehensive assessment of AF-MM's sensitivity, specificity and potential biases for the prediction of complex structures. Moreover, as AF-MM always produces structural models for a pair of sequences (or more, as AF-MM is not limited to binary interactions), there lacks clear criteria to differentiate between high and low-quality models. For instance, in spite of AF-MM outputting various metrics to evaluate the quality of the generated models, there is no threshold on the metrics to define a good model. Setting a threshold on the metrics to discriminate high from low quality models is paramount for the application of AF-MM.

1.4 Experimental methods to validate protein interaction interfaces

A common approach to study PPI interfaces is by solving their three-dimensional structures through techniques like crystallography and NMR. While these structural determination methods can provide atomic-level details on PPI interfaces, they often require extensive technical expertise and can be challenging to scale up for higher throughput. Other methods, such as mutational scanning, can also elucidate mechanistic details of PPI at the residue level. The mutational scanning approach is used to identify residues at PPI interfaces that contribute significantly to binding, which are known as hotspot residues. Hotspot residues are identified through scanning mutagenesis, an experimental method that systematically substitutes the amino acids of a protein chain or a region of interest with alanine or glycine. This substitution tests the effect of removing the sidechain atoms beyond the alpha-carbon on the binding free energy. With every substitution, binding free energy is measured to calculate the change in binding free energy relative to the unsubstituted polypeptide chain, and the positions where substitutions lead to an increase in binding free energy of 2.0 kcal/mol are classified as hotspots (Moreira et al., 2007).

The concept of mutational scanning can be incorporated into interaction assays to validate PPI interfaces without the need to determine their structures. To accomplish this, a quantitative interaction assay can be employed to probe the binding strength of a PPI using wild-type proteins. Subsequently, mutations are introduced at a predicted PPI interface, and the binding strength between the mutated protein and its wild-type partner is re-quantified. A reduction in binding strength compared to that of the wild-type proteins, or the absence of binding between the mutated protein and its wild-type partner, thus indicates that the predicted interface is responsible, or at least partially responsible, for mediating the interaction. As interaction assays demand less technical expertise than structural determination methods, this approach is also more scalable. However, because the change in binding strength can be subtle, especially in the case of point mutations, it is important that the interaction assay produces a quantitative read-out that correlates with binding strength. While this strategy is effective in circumventing many challenges associated with structural determination, several concerns need to be addressed. Apart from interface disruption, introduced mutations can also unfold, misfold or mislocalize a protein, all of which can lead to the degradation of the protein. In this scenario, the absence of PPI is caused by the reduced expression of the mutated protein, rather than the mutation perturbing the interaction interface. Hence, the expression of mutated proteins needs to be measured and monitored when they are used to validate predicted interfaces.

In the following sections, I will discuss different approaches to detect PPIs *in vivo*.

Following that, specific interaction assays and their suitability for the aforementioned strategy will also be discussed.

1.4.1 Binary methods vs co-complex methods

Numerous methods have been developed to detect PPIs in various approaches. These methods can be broadly classified into binary methods or co-complex methods. Binary methods rely on the co-expression of genetically tagged proteins to detect PPIs. When the genetically tagged proteins interact, thereby bringing their genetic tags into close proximity, different read-outs are produced to signal the presence of a PPI. Typical read-outs include the reconstitution, activation or expression of reporter proteins, as well as the emission of measurable wavelengths like fluorescence. The Y2H assay is an example of binary methods (Chien et al., 1991; Fields & Song, 1989). This assay makes use of a transcription factor that is physically split into its DNA-binding domain and its transcription activation domain. The DNA-binding domain that is capable of binding to the promoter region of a reporter gene is fused to a bait protein, while the transcription activation domain is fused to a prey protein. Upon interaction of the bait and prey proteins, the transcription factor is reconstituted and drives the expression of a reporter gene. In a conventional Y2H setup, the presence of interaction between the bait and prey is indicated by the requirement for the reporter gene for growth, making yeast growth a read-out. Y2H serves as the conceptual predecessor to many more potent binary PPI technologies that are subsequently developed.

On the other hand, co-complex methods typically involve either a single genetically tagged protein or no tagged proteins at all. These methods generally include at least one purification step, followed by the identification of co-purifying proteins using MS. Affinity Purification—Mass Spectrometry (AP-MS) is one of the most common co-complex methods, and it detects PPI by using a bait protein that is fused to an affinity tag like the FLAG-tag or HA-tag (Ho et al., 2002). The affinity tag is used to purify the bait protein from cell lysates, and other proteins that are co-purified in the purification process are digested into peptides and subsequently identified using MS. Nonetheless, as co-complex methods detect interactors of a bait protein by means of co-purification. The identified interactors may not necessarily interact directly with the bait protein, as their associations with the bait protein can be mediated by a third interactor. Moreover, the cell lysis step in these methods may result in false negatives by disrupting weaker interactions and false positives by bringing into contact proteins that would not naturally encounter each other under physiological conditions due to different subcellular localization.

Considering these limitations, co-complex methods, albeit quantitative and highly sensitive, are less suitable for probing the interaction interface between proteins. Binary methods, which are better at detecting PPIs mediated by direct contacts, are more appropriate for the investigation of PPI interfaces. Both binary and co-complex methods are extensively reviewed in Titeca et al. (2019).

Binary methods to probe protein interaction interfaces

Owing to its robustness and simplicity, Y2H has been employed by researchers to study the interaction interface between proteins (Wang et al., 2012). However, there are some clear disadvantages associated with the method. As the assay requires the reconstitution of a transcription factor to produce a read-out, trafficking the proteins of interest into the nucleus is thus necessary, and this is achieved by the nuclear localization signals that are found in the DNA-binding and transcription activation domains that are fused to the proteins of interest (Chien et al., 1991; Fields & Song, 1989). However, such localization is often artificial for many proteins as they are sequestered in the cell in different subcellular compartments under physiological conditions. That said, the PPIs detected by Y2H should be carefully interpreted as a positive readout merely reflects that the tested proteins are capable of interacting on the basis of physicochemical principles, such as shape complementarity and charge distribution. The need for nuclear localization of the bait and prey proteins also dictates that the assay is not suitable for the detection of proteins that are toxic to the cell if accumulated in the nucleus. Besides, auto-activation is also a well-known issue in Y2H. Common sources of auto-activation include transcription factors that natively have transactivation function or non-transcription factors that possess cryptic transactivation domain (Shivhare et al., 2021). Testing PPIs between non-yeast proteins using Y2H can also lead to false negatives. This is because, depending on the origin of the proteins of interest, their interaction may never take place in a yeast setting due to the absence of PTM or co-factors that enables their interaction. The quantification of the binding strength of a PPI using Y2H is also often indirect, such as by measuring the transcript level of the reporter gene (Maier et al., 2012). As Y2H is not equipped with a direct way to monitor the expression of bait and prey, testing the effect of mutation on the expression level of a protein is often achieved through other techniques. For example, Wang et al. (2012) separately transfected mammalian cells with the open reading frames (ORFs) of the mutated proteins to quantify their expressions using the western blot technique. Taken together, while Y2H is a fairly straightforward assay to test for interactions between proteins, its application in characterizing the interaction interface between proteins is hindered by several factors, with the use of yeast background being one of the most significant caveats. Nevertheless, considering that the biology of many proteins remains uncharted, having a reference for physically possible PPIs is vital for future research, and the PPIs detected by Y2H can serve as a valuable foundation for subsequent studies.

Following principles similar to Y2H, many binary methods have been subsequently developed to circumvent the shortcomings of Y2H. Examples of these methods include the Mammalian Protein-protein Interaction Trap (MAPPIT) assay and the Kinase Substrate Sensor (KISS) assay. Unlike Y2H that physically splits up a transcription factor, MAPPIT and KISS functionally split up the JAK-STAT signal transduction pathway required for reporter protein transcription upon cytokine administration. These methods have the advantage of operating in intact mammalian cells, therefore providing a more optimal cellular context for testing PPIs that involve mammalian proteins. MAPPIT makes use of a mutated cytokine receptor, such as Leptin receptor, that can no longer recruit the STAT protein, rendering it deficient in signal transduction. To use the reconstitution of the signal transduction as read-outs for PPIs, the bait protein is fused to the mutated cytokine receptor, and the prey protein is fused to a cytokine receptor fragment with functional STAT docking sites. This way, under the conditions that a ligand like erythropoietin or leptin is administered and an interaction occurs between the bait and prey, JAK is recruited to the mutated cytokine receptor to phosphorylate the docking sites that are fused to the prey, which in turn leads to the recruitment of STATs. The recruited STATs are subsequently phosphorylated and activated by JAK. The activated STATs then homodimerize and migrate to the nucleus to activate the transcription of a reporter gene (Eyckerman et al., 2001; Lievens et al., 2011). While the system is elegant in its design, as JAK is anchored to the plasma membrane, it restricts PPI detection to those happening near the plasma membrane. Furthermore, with the bait being fused to a mutated cytokine receptor, a relatively large transmembranal protein of over 1,000 amino acids, the interaction between the bait and the prey can potentially be blocked due to steric hindrance.

KISS solves this by fusing the bait to the kinase-containing portion of TYK2, while keeping the prey similarly fused to the cytokine receptor fragment. When the bait and prey interact, the kinase domain of TYK2 phosphorylates the cytokine receptor fragment of the prey, leading to the recruitment of STATs and its phosphorylation by the kinase domain of TYK2 (Lievens et al., 2014). The downstream signaling cascade resembles that of MAPPIT, where a reporter gene is eventually expressed. The use of only the kinase-containing portion of TYK2 effectively eradicates the need for the interaction between the bait and the prey to happen near the plasma membrane for it to be detected. Moreover, KISS also uses a smaller tag which poses less risk of false negatives caused by steric hindrance. The MAPPIT method has been used to map PPI interfaces by using luciferase as the reporter gene and measuring luciferase activity to probe the effect of mutation on the PPIs (Vyncke et al., 2019).

While earlier methods rely on reconstituting a signaling pathway and subsequent reporter gene expression for read-outs, other methods offer more direct read-outs. Examples of these methods include the Förster resonance energy transfer (FRET) assay and the Bioluminescence resonance energy transfer (BRET) assay that detect PPIs based on physical distance. As their names suggest, these methods detect PPIs through non-radiative (without emitting photons) energy transfer from a donor molecule to an acceptor molecule. Since this form of energy transfer only occurs when the donor-acceptor pair is within a close distance, it serves as an effective probe for determining the proximity of the donor-acceptor pair. The theoretical range for energy resonance is 10Å to 100Å, with the actual working range dependent on the type of donor-acceptor pair and its *in vivo* applications (Weihs et al., 2020). Both FRET and BRET assays incorporate this principle into their approach to detect PPIs.

This is achieved in FRET by fusing the bait and prey with a donor fluorophore and an acceptor fluorophore, each having different excitation and emission wavelengths. These fused constructs are then expressed in intact cells to test interactions between the bait and prey. Upon interaction between the bait and prey, the energy emitted by the donor fluorophore, following its excitation with light from the appropriate wavelength, is transferred to the acceptor fluorophore. This energy transfer yields a fluorescent signal at a different wavelength, serving as a read-out that indicates the presence of interaction between the bait and prey. Conversely, when the bait and prey do not interact, only the donor fluorophore emits fluorescence (Sekar & Periasamy, 2003). Operating on the same principle, BRET differs from FRET by fusing the bait with a donor luciferase. The energy emitted by luciferase can similarly be transferred to an acceptor fluorophore when in close proximity. The distance limit for energy transfer in BRET is the same as FRET (Pfleger & Eidne, 2006). Since luciferase oxidizes a substrate like coelenterazine to produce luminescence, the use of luciferase as donor removes the need for external illumination.

Both FRET and BRET offer excellent time resolution for real-time study of transient PPIs in intact mammalian cells. As emphasized earlier, determining the expression levels of the bait and prey proteins is important when it comes to characterizing their interaction interface(s) in order to understand the reason behind the absence of an interaction. In FRET and BRET, direct and independent monitoring of bait and prey expression levels is made possible thanks to the use of donors and acceptors that have different emission wavelengths. In terms of practicality, BRET offers the advantage of decoupling luminescence and fluorescence signals for the separate measurement of bait and prey expression levels. In contrast, for the same purpose, FRET requires reasonable separation of emission spectra between the donor and acceptor fluorophores (Sekar & Periasamy, 2003). Furthermore, the use of BRET can also bypass certain practical issues associated with fluorescence-based technologies, such as photobleaching and autofluorescence.

While the strategy of detecting PPIs through the spatially-restricted transfer of resonance energy is ingenious, the major drawback of the strategy also lies in its stringent limitation in distance. Consequently, these methods are inherently sensitive to the distance between and the relative orientation of the tags. Their sensitivity to the aforementioned factors carries significant implications for their application. On the one hand, these methods are particularly well-suited for the detection of direct PPI. On the other hand, they might miss PPIs that involve larger proteins or multiple partners as these can sterically hinder the transfer of energy, resulting in false negatives (Titeca et al., 2019). Beyond the detection of PPIs, these methods have also been applied to study the subcellular localization of detected interactions (Coulon et al., 2008; Del Pozo et al., 2002).

The BRET assay has been used extensively in this thesis to validate the interaction interfaces predicted by AF-MM, and it is included in the article presented in Chapter 4, Article II.

1.5 Aims of the thesis

Many proteins interact with other proteins to perform their cellular functions. Many PPIs have so far been detected using high-throughput interaction assays to study their biological functions. However, these assays do not provide mechanistic information on the detected PPIs, impeding our ability to study their molecular mechanisms. Domains and motifs are important functional modules that enable proteins to interact with others by forming DMIs and DDIs. Different databases annotate DMIs and DDIs using diverse sources of experimental evidence, providing a way to map these interfaces onto PPIs to generate mechanistic insights. Even though these annotated interfaces can be readily mapped onto PPIs, it is important to note that the mapped interfaces are not necessarily accurate, and precise detection of DMIs and DDIs poses distinct challenges.

Owing to the short and degenerate nature of motifs, detecting them in protein sequences often leads to stochastic matches. Consequently, reliable detection of motifs requires additional information to assess their likelihood to be functional. The first objective of this thesis was to develop a computer program to automate the detection and scoring of DMIs in the sequences of interacting protein pairs. To achieve this, a list of DMI types from ELM database was used for DMI matching and an machine learning (ML) algorithm was trained to score DMI matches.

3did is a database that hosts a collection of DDI types by employing an automatic pipeline to annotate DDIs based on observed contacts in PDB structures. While the detection of domains with HMMs is generally accurate and DDI types from 3did can be readily mapped onto PPIs, there is a need for a reference set to assess the accuracy of DDIs mapped through this approach. The second objective was to identify a very high-confidence subset of DDI types from 3did by manually curating a randomly selected subset of DDI types. The reference set produced from this manual curation effort was further utilized to identify features that can predict high-confidence DDI types.

The development of AF-MM has brought forth exciting possibilities for the characterization of novel PPI interfaces through structural modeling. Although many studies have tested the ability of AF-MM to predict different interface types, there is a need for a comprehensive evaluation on AF-MM and its metrics for their sensitivity, specificity and potential biases. The third objective was to comprehensively evaluate AF-MM's predictive capability and apply AF-MM to discover novel interaction interfaces between proteins.

Chapter 2

Development of a domain-motif interface predictor

2.1 Introduction

Most PPIs in the human protein interactome lack interface information. The lack of mechanistic information on these PPIs hinders our ability to investigate their molecular functions. One approach to generate mechanistic insights into these PPIs is through the use of structural modelling tool to model their interfaces. While this approach can be fruitful for interfaces involving exclusively domains since a significant number of DDIs have been solved so far, it is not as applicable to DMIs as many DMIs lack structural information. The detection of DMIs in PPIs will thus benefit from a sequence-based approach where the occurrences of known DMI types are searched within the sequences of interacting protein pairs. The ELM database contains a list of high-quality DMI types that are curated based on experimental evidence of motifs. Moreover, the database provides useful information that can be leveraged to detect the occurrences of annotated DMI types, such as the regular expressions of curated motif types and, in most cases, the HMMs of their corresponding binding domains.

iELM is a bioinformatic tool that detects DMIs in user-submitted PPIs, and its working principle has been previously discussed in the section **Sequence-based detection of domain-motif interfaces**. Drawing inspiration from iELM, I developed a DMI predictor program that similarly automates the search of potential DMIs in the sequences of interacting protein pairs and scores the returned DMI matches using a machine learning algorithm (Weatheritt, Luck, et al., 2012; Weatheritt, Jehl, et al., 2012). While there exists overlap between the workflow of the developed DMI predictor and iELM, I have designed the program to operate differently in several aspects to overcome the limitations of iELM. First, iELM discontinues further DMI search if it can map any DDI type from 3did to a user-provided PPI. As the prediction of DMIs and DDIs should not be mutually exclusive, since they can act synergistically to mediate an interaction, I have designed the DMI predictor to be agnostic to potential DDIs. Consequently, the DMI predictor does not discontinue the search of DMIs in case of detecting DDIs. Second, iELM builds HMMs specific to motif-binding domains by using hand-curated sets of known motif-binding domain sequences. Such an approach poses the risk of overfitting HMMs to a small and specific set of domains. Based on observations from manual checking of DMI types and their instances available in the ELM database, I discovered that HMMs from SMART detects motif-binding domains more reliably than those of Pfam. With SMART creating HMMs not specifically for the detection of motif-binding domains, this approach poses a lesser risk of overfitting. Hence, SMART HMMs are opted for over Pfam HMMs when available. Lastly, the evaluation of iELM was done on an imbalanced dataset, with the number of negative data points outnumbering the positive data points by almost 30-fold. To ensure a robust evaluation of the DMI predictor, the training as well as testing of the machine learning algorithm were done on a balanced dataset.

To aid in scoring DMI matches, I also engineered new features using systematic PPI networks and incorporated them into the machine learning model. The underlying assumption is that if a motif match is functional, its binding domain would be significantly enriched among the interaction partners of the motif-bearing protein. Such enrichment of binding domains among interaction partners can be quantified and used for DMI match scoring.

The DMI predictor requires only the UniProt accession identifiers of human interacting protein pairs to run. The predictor has been applied to PPIs from HuRI to characterize their interaction interfaces.

2.2 Methods

Manual curation of domain-motif interface types

A list of DMI types was downloaded from the ELM DB (version 1.4, accessed on 29.09.2020) and checked manually for correct annotation of interacting domain(s) by manual inspection of their solved structures, if available, to confirm their interactions. Special cases exist and extra caution was taken when checking them. For example, some DMI types require two domains to come together to form an interface where the motif binds, and these cases were annotated accordingly. HMMs are widely used to computationally predict the occurrence of domains in protein sequences. Since ELM DB also annotates HMMs of interacting domains, prediction accuracies of HMMs from SMART and Pfam database were compared, and the HMMs with better prediction performance were selected for subsequent development of the DMI predictor. Furthermore, some domains are formed by tandem repeats, and most of these domains are detected by SMART and Pfam as individual tandem repeats rather than the complete domain. In these cases, the stoichiometry of tandem repeats required to form the interacting domain of a given DMI type was annotated. The resulting list of 290 DMI types was used for subsequent development of the DMI predictor.

Curation of domain-motif interface instances for positive reference set

A list of 20,396 reviewed human protein sequences was downloaded from SwissProt (16.02.2021). A list of experimentally validated DMI instances was downloaded from the ELM database, and the list served as the basis for the making of the positive reference set (PRS). The list was filtered for interactions involving only human proteins from SwissProt. The DMI instances in the resulting list were then subject to further processing by checking the annotated positions of motif and domain on the protein pair. This was done by running pattern search in the sequence of the annotated protein using the regular expression of the motif type to confirm the motif's positions in the protein sequence. Similar step was performed on annotated domains by running HMM search on the annotated protein using the HMM selected during the curation of the DMI types list. If the reported positions of motif did not align with the match returned from pattern search, the DMI instance would be deleted. Similarly, should the reported domain not occur on the annotated protein through HMM search, the DMI instance would be deleted as well. DMI instances formed by alternative isoforms were converted to their canonical isoforms and manually checked to ensure that the annotated motif sequence and match positions were still the same in canonical isoforms and the annotated domains were still detected. Additional DMI instances were obtained from our collaborator, Dr. Norman Davey, who curated DMI instances from available crystal structure on PDB. Similar filtering processes were applied, and 98 DMI instances were subsequently included in the PRS. The resulting list contained 662 proteins that formed 898 DMI instances from 106 DMI types. The list served as the PRS for the development of the DMI predictor.

Generation of random reference sets

Random reference set (RRS) was generated by randomly pairing proteins from the downloaded list of SwissProt human proteins, while excluding known interacting protein pairs. Known interacting protein pairs were downloaded from IntAct on 24.02.2021 that consisted of 145,468 interactions from human proteins (Orchard et al., 2014). Motif matches and domain matches were detected in randomly paired proteins by running regular expressions and HMMs search on the paired proteins. Potential DMIs were then matched among the paired proteins to form DMI instances for RRS. As the PRS was made up of a relatively small pool of proteins, different strategies to sample random protein pairs were tested to check for biases that could be learned by the predictor from some inherent properties in the PRS. To this end, random protein pairs were either sampled from the PRS protein pool or the whole human proteome to allow for the comparison of prediction performance using different RRSs. Furthermore, potential biases could also come from the limited representation of DMI types in the PRS. On the one hand, to ensure that prediction performance was measured across many DMI types, one way to sample RRS DMI instances was to sample in a uniform way to guarantee DMI type representation to the best possible extent in the RRS. To sample DMI types uniformly, 5 DMI instances per DMI type were sampled when PRS protein pool was used, and 3 DMI instances per DMI type were sampled when the whole human proteome was used. On the other hand, the prediction of DMI instances depends on the frequency of domains and regular expression matches, and sampling DMI instances uniformly would not allow these features to contribute to the outcome of the predictor. To this end, 1,000 DMI instances were sampled randomly regardless of DMI types. This step ensured that the number of DMI instances in RRS was roughly the same as that of PRS. Four versions of RRS, with triplicates of each version, were generated in total by combining the strategies of sampling proteins and DMI instances.

Annotation of domain and motif features

Motif probability

The motif probability quantifies the degeneracy of a motif type by calculating the likelihood of a regular expression matching in the disordered regions of proteins. The motif probability of every motif type was retrieved from ELM.

Propensity to disorder and undergo secondary structure transition of motifs

The disorderedness of a protein region can be predicted using the IUPred tool (Mészáros et al., 2018). The propensity of a motif match to undergo secondary structure transition can also be predicted using the ANCHOR2 tool (Mészáros et al., 2018). The IUPred and ANCHOR2 scores of a motif match were quantified by averaging the IUPred scores (short option) and ANCHOR2 scores across the amino acids of the motif match.

Relative Local Conservation and variance of conservation of motifs

The RLC of motif matches was calculated following the $\text{Sig}_{\text{motif}}$ formula from SLiMPrints (Davey et al., 2012) using four types of conservation scores: conservation calculated across orthologues from Quest for Orthologs (QFO), conservation on metazoan level, conservation on the level of class (mammalia), and subphylum (vertebrata). The variance of conservation scores in a motif match was also calculated using the following equation:

$$Variance = \frac{\sum_{i=1}^{n}(x_i - \mu)}{n}$$

where n is the number of defined positions in the motif, and x_i is the conservation score of i-th defined positions in the motif match. μ is the mean of conservation across the defined positions in the motif match.

Enrichment of the motif-binding domain among interaction partners

PPI networks of motif-containing proteins were leveraged to probe the likelihood of a motif match being functional in a protein. The idea is that if a motif match is functional in a protein, then its binding domain should also be enriched among the interaction partners of the protein compared to the background distribution of the domain in the complete PPI network. To this end, HuRI and literature-curated binary PPI networks were downloaded from Luck et al. (2020) and combined to produce a PPI network consisting of 13,722 proteins and 101,219 binary interactions. To account for domain frequency and protein degree in the network, 1,000 random networks were generated from the real network while maintaining the degree of the proteins observed in the real network. The enrichment of a motif-binding domain among the interaction partners of a motif-containing protein was first quantified by counting the number of interaction partners having the motif-binding domain. To compute the significance of the observed domain enrichment, the generated random networks were used to count the number of random interaction partners having the motif-binding domain, thereby producing a background distribution for the domain enrichment. The p-value of the observed domain enrichment was thus computed by calculating the fraction of random networks having greater than or equal to the domain enrichment observed in the real network. The z-score of the observed domain enrichment was also computed to quantify the extent of enrichment observed in the real network. The z-score of domain enrichment was calculated by subtracting the mean number of interaction partners with the domain in the random networks from that observed in the real network and dividing the result by the standard deviation of the number of interaction partners with the domain in the random networks. In cases where the standard deviation in the random networks is zero, the absolute difference between the number of interaction partners with the domain observed in the real network and the mean number of interaction partners with the domain in the random networks was calculated as the z-score. During the training of the DMI predictor, interacting protein pairs in the PRS were removed from the protein's real network to avoid circularity in the calculation of domain enrichment.

Overlap of motifs with domains

Some motif matches returned from regular expression search are found in regions that are detected by HMMs to fold into domains, and these motif matches are unlikely to be functional.

The domain overlap of a motif match was quantified by calculating the fraction of residues in the motif match that overlap with a domain detected by HMMs.

Domain frequencies

As multiple occurrences of a domain can be detected in a given protein sequence, two methods were employed to quantify the frequencies of motif-binding domains in the human proteome from SwissProt. The first method quantifies domain frequencies by calculating the fraction of proteins with at least one HMM match of the domain. This feature was named domain frequency by protein. The second method quantifies domain frequencies by calculating the number of domain HMMs matching in human proteins divided by the total number of proteins in the human proteome. The second method accounts for the domains that are formed by multiple tandem repeats as their HMMs were produced for individual tandem repeats rather than the complete domain.

This feature was named domain frequency in proteome. For DMI types that require two domains to form an interface where the motif binds, the domain frequencies of the two domains were averaged and used as features for the DMI predictor.

All features were calculated for every DMI instance in the PRS and RRS. Data points with missing values were removed. These removed data points were either without conservation score or PPI network information. After listwise deletion of data points with missing values, 830 DMI instances were retained in the PRS. Similar processing was applied to all versions of RRS, and the number of DMI instances retained in the triplicates of the RRS versions ranged from 870 to 985, making a fairly balanced dataset for the training of the RF model.

Model training

Each replicate of the RRS versions was combined with the PRS to train a RF model. The combined dataset was stratified by labels (known and random DMI instances) and split into a training and a test set with a ratio of 3:1. Subsequently, a RF model with 1,000 decision trees was fitted on the training set, and the fitted RF's accuracy was then evaluated on the test set. Random seed was set at 0 for model building for reproducibility. Predictive performances of fitted RFs were further evaluated by means of ROC curve and PR curve. For each RRS version, ROC and PR curves averaged across the triplicates of the RRS version were plotted by interpolation. Feature importance was quantified as mean decrease in impurity by extracting the values from the attribute 'feature_importances_' from the fitted RF model. Similarly, feature importance was also plotted by averaging feature importance of each model built with the triplicates of each RRS version.

For the final model, the whole dataset was used to train an RF with 1,000 decision trees with random seed set at 0. A median imputer was also fitted on the whole dataset so that missing values in user input protein pairs can be imputed with the median of the corresponding feature in the training data. The effect of imputation on the accuracy of the predictor was tested by imputing masked values in the test set, and the effect was found to be negligible. The dataset used to train the final model was PRS combined with RRS generated by sampling 1000 DMI instances from the human proteome.

Softwares and packages used

The software PyMOL (TM) Molecular Graphics System, Version 2.5.0. Copyright (c) Schrodinger, LLC., was used for the visualization of solved DMI structures. All codes were written in Python version 3.9 and analyses were done using Jupyter notebooks. The following Python libraries were used: re package for regular expression search, pandas (McKinney, 2010) for data analysis, and Matplotlib (Hunter, 2007) and seaborn (Waskom, 2021) for data visualization, scikit-learn for model training, feature importance analysis, and ROC and PR statistics (Pedregosa et al., 2011), scipy for RLC calculations (Virtanen et al., 2020), igraph for network randomization (“Igraph – Network Analysis Software”, n.d.).

2.3 Results

Improving the binding domain annotations from ELM

While ELM provides binding domain information for almost all the motif types curated in the database, there are some motif types for which such information is missing. To thoroughly assess the binding relationship between motifs and domains, a list of motif types and their binding domains (DMI types) was downloaded from ELM. I manually checked through the list of DMI types to ensure that the correct domains were annotated in the list. Throughout the manual checking process, I noticed the HMMs of SMART were more accurate at detecting motif-binding domains than the HMMs from Pfam. As a result, despite Pfam HMMs being the default in ELM, the HMMs used in the subsequent development of the DMI predictor were defaulted to SMART HMMs. Among the 290 DMI types that I checked, 63 had their domain annotations changed (Figure 2.1A).

The reasons of change in annotation can be classified into four categories. The majority of the changed annotations comes from improved annotation category where the switch from Pfam HMMs to SMART HMMs results in better detection of the binding domains. For example, Pfam often creates different HMMs for different subtypes of a domain, such as Efhand and SH3 domains, while SMART creates HMMs only on the domain level and not domain subtypes. While these Pfam HMMs that are created for domain subtypes have higher specificity at the cost of sensitivity, SMART HMMs that are less specific but more sensitive were opted over Pfam HMMs to ensure that all motif-binding domains can be detected in the initial step of the pipeline for subsequent DMI matching and scoring. Some motifs require two domains to first bind to form a groove where the motifs fit into, and these cases are not always annotated in ELM. These annotations were improved by including the second domain as an additional binding domain, and they represent the trimeric annotation category. Five motifs did not have binding domain information annotated in ELM. For these motifs, the sequences of their binding partners were used to detect domain occurrences, and their binding domain information was inferred by checking for enrichment of specific domains in the binding partners. Three motifs had multiple domains annotated as their binding domains, but some of them were not detected in the sequences of their binding partners. For instance, the ELM class LIG_Actin_RPEL_3 has the HMM of RPEL repeat (PF02755) annotated as one of the binding domain. This represents as a wrong annotation because the annotated HMM is for the detection of RPEL repeat and not the binding domain. The erroneous annotations were subsequently removed. For more details on the manual checking, please refer to the Method section.

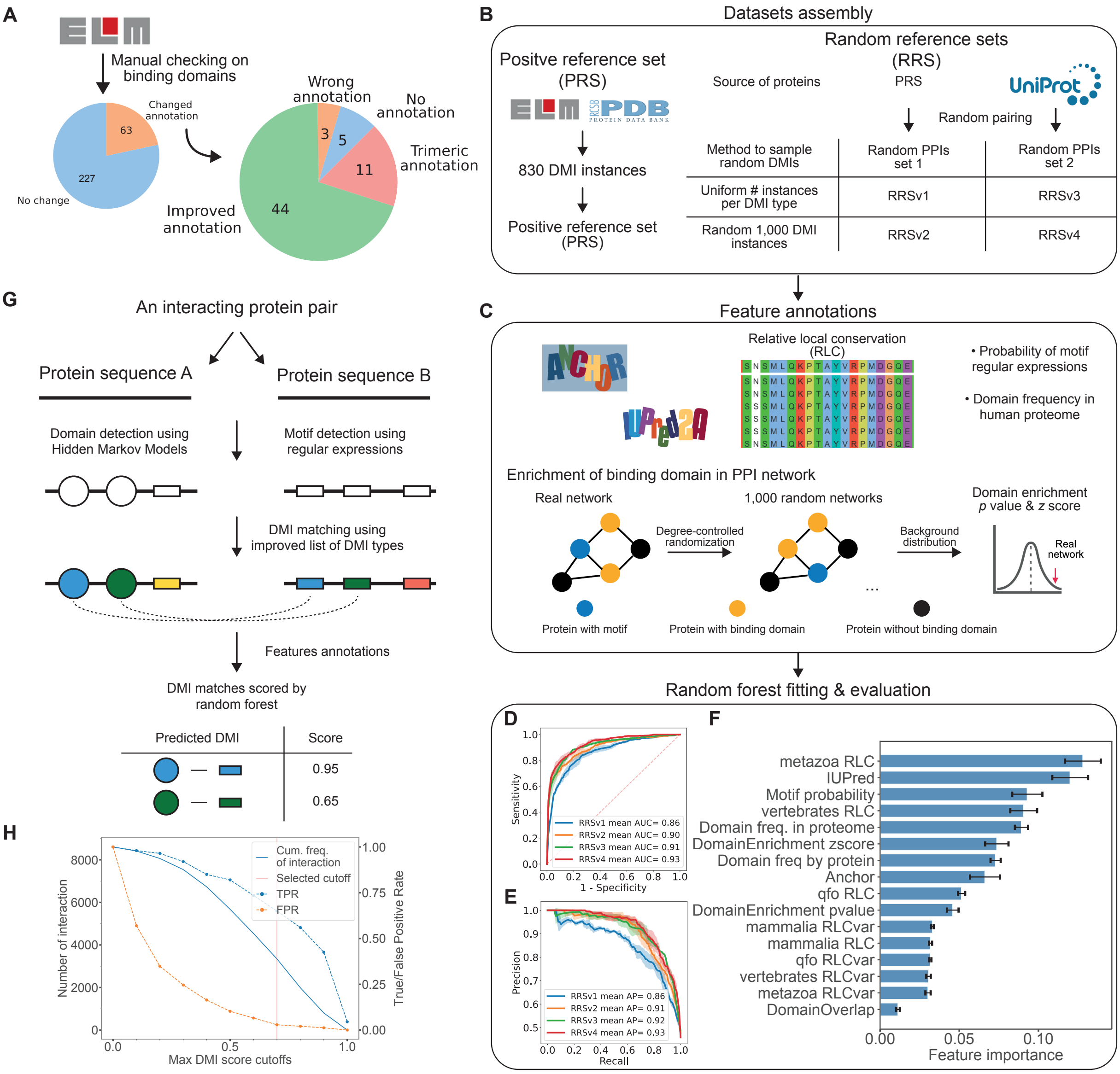


Figure 2.1: The development of DMI predictor and its application on HuRI. **A**) Outcome of manual checking performed on the list of DMI types downloaded from ELM. **B**) Schematic illustrating the assembly of PRS and different versions of RRS. **C**) Annotation of features on the PRS and RRSs. The schematic at the bottom illustrates the calculation of domain enrichment within the interaction partners of a motif-bearing protein as additional features. **D**) ROC curve of RF models trained using different sets of RRS. For each RRS version, ROC and PR curves averaged across the triplicates of the RRS version were plotted by interpolation. **E**) Same as **D** but for PR curve of the RF models. **F**) The importance of different features to the RF trained using the PRS combined with the RRSv4 as quantified using mean decrease in impurity. **G**) Schematic showcasing the workflow of the developed DMI predictor where the improved list of DMI types and the trained RF model are incorporated into the DMI detection pipeline. The output of the pipeline presents the DMI matches and their likelihoods to be functional in a tabular format. The predicted likelihoods are within the range of 0 to 1, where 1 indicates the highest likelihood to be functional. Circle represents domain matches and rectangle represents motif matches. Same color of circles and rectangles represents DMI matches. **H**) The developed DMI predictor was applied on PPIs that are detected in HuRI, and the scores of the predicted DMIs are titrated over increasing cutoffs. Only the highest-scoring DMIs are considered in this analysis. The dashed lines refer to the right y-axis, while the filled line refers to the left y-axis. The red vertical line indicates the cutoff of 0.7 applied on the DMI scores to call a predicted DMI of high-confidence.

Datasets preparation and feature annotations for model training and evaluation

To score the DMI matches detected in the pipeline, an RF model was trained and evaluated for its ability to score DMI matches (Figure 2.1B). To train and evaluate an RF model, a PRS and an RRS were needed. The PRS was constructed using known DMI instances that are deposited in ELM, consisting of 830 DMI instances. As negative data are scarce in biology, random DMI instances were computationally generated by randomly pairing proteins from the PRS and scanning for DMI occurrences to serve as the RRS. However, as the PRS consists of a limited number of experimentally validated DMI instances – they pertain to only a subset of DMI types found in a limited number of proteins – generating the RRS by randomizing protein pairs from the PRS would thus also propagate the potential bias from the PRS to the generated RRS. To address these potential biases, different RRS versions were created. Briefly, random protein pairs were generated by either randomizing protein pairs from the PRS protein pool or from the whole human proteome in UniProt. Next, the DMI instances detected in these randomized proteins were sampled either randomly, i.e., irrespective of DMI types, or by controlling for a uniform number of DMI instances across all DMI types. Combining the two ways of randomizing protein pairs and sampling for DMI instances resulted in four RRS versions (Figure 2.1B). Each RRS version was generated three times to produce triplicates for quantification of statistical significance. The number of random DMI instances in the different versions of RRS ranged from 870 to 985, making the number of positive and negative data points for model training and evaluation fairly balanced.

Different motif-specific features and domain-specific features were annotated for the DMIs from the PRS and different versions of the RRS. The ANCHOR and IUPred scores were used to quantify the propensity of motifs to be disordered or undergo secondary structure transition upon binding to a partner. The RLC scores and their variance were quantified to assess the degree of motif conservation across orthologs. The motif probability retrieved from ELM was used to quantify the degeneracy of motif types based on their regular expressions. Two additional PPI network-based features were also calculated to assess the likelihood of a motif being functional. The underlying assumption is that if a motif match is functional, its binding domain would be significantly enriched in the interaction partners of the motif-bearing protein. To this end, for a given motif matching in a protein, I quantified the enrichment of its binding domain in the interaction partners of the protein by means of calculating the z-score and p-value of the observed enrichment. The frequencies of motif-binding domains were also used as features (Figure 2.1C).

The PRS was combined with each of the four different versions of RRS and their features were used to train four RF models. The combined dataset was stratified by label and split into a training and a test set with a ratio of 3:1. The performance of the trained models, as evaluated on their respective test sets, is summarized in Figure 2.1D and E. While similar performance was obtained using models trained on different RRS versions, RRS version 4, generated through randomly sampling DMI instances from random protein pairs from the human proteome, achieved the best performance (Area under Receiver Operating Characteristics (ROC) and precision recall (PR) curves 0.93). The contribution of different features to the performance of the model trained using RRS version 4 is quantified as mean decrease in impurity (Figure 2.1F). The conservation and disorder propensity of motifs ranked highest in their contribution towards prediction outcome. The enrichment of the binding domain of a motif as a feature ranked sixth in terms of importance towards prediction outcome.

Since RF models output a value corresponding to the number of trees within the ensemble that predicts '1' for a given data point, this value can essentially be used as a score to quantify the likelihood of a DMI match being functional. As RRS version 4 achieved the best performance, I established a cutoff score of 0.7 based on its test set to call DMI matches of high-confidence. The cutoff corresponds to a sensitivity of 66.3% and a specificity of 97.2%.

Workflow of DMI predictor and its application on HuRI

As RRS version 4 demonstrated the best performance, I retrained an RF model using all the data points in the PRS and the RRS version 4 for deployment in the DMI predictor pipeline (Figure 2.1G). Briefly, the pipeline operates on the UniProt identifiers of a pair of interacting proteins. HMMs of motif-binding domains and regular expressions of motifs are used to detect occurrence of motif-binding domains and motifs in the sequences of the interacting proteins. Based on the improved list of DMI types, the detected domains and motifs are matched to form DMI matches. The DMI matches are then subject to feature annotations, followed by scoring by the RF model. The DMI matches as well as their scores are returned by the pipeline, with higher score indicating a higher likelihood to be functional.

The developed DMI predictor was applied onto HuRI to detect PPIs that are potentially mediated by DMIs (Figure 2.1H). As expected by the degeneracy of motifs, many DMI matches were detected in the PPIs from HuRI. Nonetheless, applying the cutoff of 0.7 on the scores returned 13,406 high-confidence DMI matches that are found in 3,195 interactions. Among the 3,195 interactions with at least one high-confidence DMI match, more than half (54%) have their top-ranked DMI matches from the ligand (LIG) class, and an additional 19.5% have their top-ranked DMI matches from the modification (MOD). The enrichment of ligand-binding motifs and PTM sites among the interactions is consistent with previously published observations that Y2H-based screens primarily detect regulatory and signalling-related PPIs (Lambourne et al., 2022; Rolland et al., 2014). The DMI predictor is executed using an in-house script and will be made available on GitHub in the near future.

As the predicted DMIs can elucidate the molecular mechanisms of disease mutations, a colleague is currently investigating the functional consequences of genetic variants using these predictions.

2.4 Discussion

In this study, I developed a DMI predictor where the sequences of an interacting protein pair are used to find occurrences of known DMIs. To reliably detect DMIs, I used a list of DMI types annotated in ELM and further improved the list by manual checking. Additionally, I also trained an RF model to score DMI matches, thereby providing a quantitative assessment to the detected DMIs. Because negative data are scarce in biology, a common approach is to randomize positive data to simulate 'negative' data for the training of supervised machine learning models. Care was taken while generating the 'negative' data as potential biases could be propagated from the positive data to the 'negative' data. This study is unique in that potential biases were addressed through the use of 'negative' datasets that were generated in different ways. Moreover, the training and evaluation of the models were performed using a fairly balanced dataset and the precision of the models was also reported, making the assessment more robust and comprehensive.

iELM reported a sensitivity of 84.8% and a specificity of 86.5% without reporting the associated cutoff. Furthermore, the publication of iELM did not report the precision of its model, and the above evaluation was determined from a test set that was heavily imbalanced, with the negative data points outnumbering the positive ones by almost 30-fold. The developed DMI predictor achieved a similar specificity at around 85% sensitivity (Figure 2.1H), and it can recall around 70% of the known DMIs with nearly 95% precision (Figure 2.1E). Notably, the DMI predictor managed to achieve similar performance as iELM without using specific HMMs to detect motif-binding domains. Switching from Pfam HMMs to SMART HMMs as well as using an RF model likely significantly contributed to the excellent performance of the DMI predictor.

As it takes a motif and one or more domains to form a DMI, a clear limitation of the study is the lack of domain-specific features. I explored the use of E values that are returned from HMM matches as a feature but to no avail as the E values can differ greatly between domains. Since the binding of motifs is highly contingent on the physicochemical property of the motif-binding pockets on domains, the structures of domains can provide useful information to the model. Nonetheless, since the DMI predictor was developed before the release of AF, structural information on binding domains was not always available during its development. With the release of AF and its predicted structures, additional domain-specific features can now be incorporated into the DMI predictor. To this end, structural models predicted by AF can be transformed into structural embeddings, and these vector representations can be used as another set of features. Along the same line, given that AF-predicted monomeric structures are nowadays publicly available, the AF-predicted structures of the domains in the interaction partner of a motif-containing protein can also be superimposed to the structure of a domain known to bind to the detected motif type. This approach effectively quantifies the structural similarity between a predicted motif-binding domain and a known motif-binding domain, and this structural similarity score can be included as a domain-specific feature. Alternatively, detected DMI matches can also be modelled by AF-MM, and the confidence metrics derived from the predicted models can be used as features too. This method, however, may be impractical due to the time-consuming nature of AF-MM modelling process. Perhaps a more pragmatic application of AF-MM would be to model the structures of high-confidence DMI matches for subsequent experimental validation. In terms of motif detection, the use of regular expressions could also be overly strict, thereby leading to false negatives. An alternative approach is to use PSSMs that are less rigid for motif detection. As PSSMs allow the quantification of the similarity between a match and the PSSM, this similarity score can also be used as an additional feature. While ELM provides a probability that reflects the degeneracy of a motif type, the probability represents the likelihood of a regular expression matching in a disordered region, and not how similar a regular expression match is to other known instances of the motif. Combining these two motif detection approaches could potentially yield the highest number of predicted motifs that can be subsequently matched with their binding domains and evaluated by the downstream RF model.

Despite these limitations, the developed DMI predictor was able to achieve an excellent performance. With this automated DMI prediction pipeline, one can gain further mechanistic insights into interaction networks and formulate testable hypotheses. For example, the high-confidence DMI matches can be experimentally validated, thereby characterizing the interaction interfaces between interacting proteins. The functional effects of genetic variants can also be more reliably predicted should they impact regions where high-confidence DMI matches are detected by the pipeline. As the detection of motifs is often plagued with stochastic matches, a strict cutoff was selected to ensure high specificity. Depending on the application, the cutoff can be adjusted to achieve a higher sensitivity. However, as the current prediction pipeline operates only on human canonical sequences, future work to include sequences of alternative isoforms as well as proteins from other proteomes is warranted. In summary, this study presents a DMI predictor that can be used to detect DMI matches at large scale. This study furthermore laid out clear limitations and future directions for sequence-based approaches in characterizing the interfaces of PPIs detected in the human protein interactome.

Chapter 3

Evaluation of the 3did database

3.1 Article I: A reference set of hand-curated domain-domain interface types for high confident interface predictions in protein interaction networks

Manuscript submitted to Bioinformatics Oxford Journal.

Summary

This project focused on producing a reference set of DDI types to evaluate the accuracy of DDIs predicted in protein interaction networks.

The 3did database currently hosts a large collection of DDI types by annotating DDIs based on their observed contacts in PDB structures. While 3did provides a direct means to map DDIs onto protein interaction networks, there is currently no reference set available to assess the accuracy of DDIs predicted using this approach. To evaluate the DDI types annotated in the 3did database, a randomly selected subset of 80 DDI types was manually inspected, resulting in 60% of the DDIs being approved. The main reason for non-approval was the inaccuracies in the boundaries of domains detected by HMMs. The manually inspected dataset also served as the training data to build a logistic regression model. Various features derived from the dataset were tested for their abilities to discriminate between approved and non-approved DDI types, and the 3did z-score was found to be the most predictive feature in our analysis. Nonetheless, our result revealed that a stricter cutoff on the 3did z-score than that reported by 3did was necessary to achieve a satisfactory approval rate. Applying the stricter cutoff on the DDI types from 3did, we subsequently mapped them to the PPIs in HuRI in an attempt to characterize their interaction interfaces. In total, 342 DDI types were mapped to HuRI, with 1,533 PPIs having at least one DDI types mapped to them.

In summary, this project resulted in a valuable reference set of DDIs and provided an accuracy estimate on the DDIs mapped onto protein interaction networks using DDI types annotated in 3did.

Statement of contribution

I was in charge of directly supervising [redacted] for the manual curation of DDIs and data analysis of the project. Additionally, I computed some features for the analysis and made minor contributions to editing the manuscript.

Supervisor confirmation

A reference set of hand-curated domain-domain interface types for high confident interface predictions in protein interaction networks

Johanna Lena Geist[1#], Chop Yan Lee[1#], José de Jesús Naveja[1,2], Katja Luck[1]

[1]Institute of Molecular Biology (IMB) gGmbH, 55128 Mainz, Germany.
[2]University Cancer Center (UCT), University Medical Center, 55131 Mainz, Germany.
[#]These authors contributed equally.

Abstract

Motivation: While the scientific community can map at increasing pace protein-protein interactions (PPIs), we lack experimental methods at similar scale to provide structural information on how these proteins interact with each other. Many PPIs are mediated by folded domains binding each other. Various computational resources, such as 3did, infer interacting domain types from observed contacts in protein structures. These domain-domain interface (DDI) types can be used to predict DDI occurrences in protein interactions. However, assessing the accuracy of these predictions is difficult because we lack hand-curated reference sets of verified domain interactions and DDI types.

Results: To generate a reference set of DDI types, we manually inspected 80 randomly selected DDI types from the 3did resource and approved 60% of them. We found inaccuracies from Pfam Hidden Markov Model (HMM) domain matches often to be causative for DDI type non-approval. Using this reference dataset and machine learning, we predicted a subset of 2411 out of 5724 considered DDI types in 3did to be of high confidence with an estimated approval rate and sensitivity of 80% and specificity of 73%. Applying this subset of DDI types for DDI prediction to a dataset of 53000 human PPIs revealed predicted interfaces for 1533 interactions functioning in proteostasis, gene regulation, kinase and small G-protein signaling to name a few.

Contact: J.NavejaRomero@imb-mainz.de, K.Luck@imb-mainz.de

Keywords

protein interaction interface prediction, domain-domain interfaces, reference dataset, manual curation

Introduction

Protein-protein interactions (PPIs) shape complex communication networks that modulate essentially all molecular pathways within the cell (Braun & Gingras, 2012). A comprehensive structural and functional understanding of these molecular interaction networks is thus a prerequisite to study cellular mechanisms in health and disease (Vidal et al., 2011). While the systematic mapping of protein interactomes is well advanced for many model organisms including human, experimental methods lack in throughput to provide a structural resolution of protein interactomes (Aloy et al., 2004; Huttlin et al., 2021; Luck et al., 2020; Michaelis et al., 2023). Various computational approaches have been developed to predict interfaces of known protein interactions. These include the collection of known interface types followed by searching for the occurrence of these interface types in PPIs using Hidden Markov Models (HMMs) or regular expressions (Mosca et al., 2013; Weatheritt et al., 2012). Alternatively, enrichment for interaction modules such as folded domains or short linear motifs among the interaction partners of a protein can be used to infer possible modes of binding (Huttlin et al., 2015; Rolland et al., 2014). More recently, deep learning-based tools such as AlphaFold-Multimer promise to predict occurrences of known as well as novel modes of protein binding at a structural resolution (Evans et al., 2022). Of note, for any of these approaches, a collection of bona fide interface types is essential, either for the prediction of interfaces, the benchmarking of prediction approaches, or the determination whether a predicted interface corresponds to a known or novel mode of binding.

A common mode of protein binding is mediated by folded domains binding each other (Liddington, 2004; Wodak & Janin, 2002). Protein domains are independent folding units with well-conserved sequences and structures. Typically, they are compact regions of around 100 to 250 residues with a unique tertiary fold and a hydrophobic core (Janin &

Wodak, 1983). Hundreds of different domain folds/types have emerged through the course of evolution. Occurrences of the same domain type share specific sequence signatures, which can be captured in HMMs. These HMMs can then be used to reliably predict domain occurrences in protein sequences. The most comprehensive collection of HMMs for protein domains is the Pfam database (recently merged with InterPro) (Finn et al., 2014; Paysan-Lafosse et al., 2023). In an effort to generate a collection of domain-domain interface (DDI) types, researchers used Pfam HMMs to scan protein sequences in structures deposited in the PDB to identify domain occurrences and atomic contacts between them, resulting in the 3did (3D interacting domains database) resource that currently hosts 14972 putative DDI types (Mosca et al., 2014). The definition of domain-domain contacts in 3did is distance-based: A domain pair must exhibit at least five non-covalent contacts such as hydrogen bonds, salt bridges, or van der Waals interactions to be considered a possible DDI type (Stein et al., 2011). The accuracy of DDI type predictions using this approach strongly depends on HMMs to accurately capture folded domains and their boundaries. However, automatic HMM generation in Pfam is primarily based on protein sequence conservation, likely limiting its ability to specifically predict folded domains with accurate boundaries.

Resources like DOMINE combined DDIs from 3did with DDI predictions from a variety of different tools resulting in larger predicted DDI collections among which DDIs derived from 3did were deemed as the highest confidence (Yellaboina et al., 2011). Many computational studies have also used 3did content to develop more sophisticated tools to predict interfaces between proteins (Karan et al., 2022; Luther et al., 2023; Zhang et al., 2016; Zheng et al., 2021). While there is no doubt that 3did and other DDI resources are very useful for the scientific community, we lack understanding about the accuracy of DDI predictions in protein interaction networks using these resources because we lack reference datasets of manually curated domain-domain interactions and DDI types. However, reliable accuracy estimates and low false positive rates are a prerequisite for experimentalists to trust in interface predictions and embark on their experimental validation.

To address this need and to identify a very high confidence subset of DDI types, we manually curated 80 randomly selected DDI types from 3did and were able to approve 60%. A considerable fraction of non-approved DDI types indeed resulted from incorrectly predicted domain boundaries and HMMs not corresponding to folded domains. Exploring further reasons for non-approval enabled identification of features that we used in a logistic regression model to predict a subset of high confident DDI types in 3did. Applying this subset of DDI types to a human protein interactome dataset for DDI prediction, we obtained statistically more significant results compared to full 3did and annotated 1533 PPIs with predicted DDIs of high confidence.

Results

Generation of a reference dataset of DDI types

To generate a manually curated reference set of DDI types, we aimed to randomly select a subset of DDI types from 3did and subject the structures and corresponding publications to manual inspection. In 3did residue contacts are computed between Pfam HMM matches occurring in the same or different protein chains. 2220 DDI types in 3did (15%) have only intrachain evidence. While there is evidence that some domain types found to be in contact with each other within a protein chain can mediate a protein interaction when situated in distinct protein chains (see also Discussion), it is impossible to evaluate intrachain DDI types for this ability using available intrachain structures and corresponding publications. Given limited manpower and time, we decided not to extend our manual curation to experimental evidence beyond the solved structures and therefore opted to restrict the generation of a reference dataset of DDI types to those with interchain evidence as reported in 3did.

One of the challenges in crystallography, the method that was used to determine the vast majority of protein structures deposited in the PDB, is the discrimination of biological contacts between protein molecules from those that are solely driven by the crystallization process. While so-called crystal contacts can hint at interfaces between proteins occurring under high concentration conditions, for the generation of a high confidence set of DDI types, we opted to exclude DDI types that are likely based on crystal contacts alone. Discrimination between biological assemblies and crystal contacts is particularly difficult for contacts observed between molecules of the same protein, i.e. homodimers, a task with which even structural biology experts struggle (Elez et al., 2020; Schweke et al., 2023). We found that 60% (7028) of the DDI types in 3did with at least one interchain interface in a PDB structure are only derived from structures of homodimers, which we felt unconfident to evaluate. This leaves us with 5724 DDI types with interchain evidence and at least one structure with a heterodimeric interface, which we considered further for the generation of a reference dataset of DDI types (Fig. 1A).

We randomly selected 80 hetero-protein DDI types with interchain evidence for manual inspection. We approved a DDI type if there was at least one structure supporting that the given PPI was mediated by contacts between both domains. In order to evaluate a DDI type, we assessed the publication (if available) reporting a structure for evidence that the two domains are mediating the PPI. We also inspected the residues observed to be in contact with each other and their overlap with the corresponding Pfam HMM matches. We further evaluated to which extent the HMM matches corresponded to folded domains and if most contacts were mediated by residues belonging to these folded regions. Lastly, we also evaluated to which extent the reported interface was observed in other structures as reported in the ProtCID database (Fig. 1B, see also Methods).

3did computes a score and its significance as a z-score for every DDI type. These scores reflect the significance of the interface in terms of type and number of contacts observed between residues of both Pfam HMM matches in comparison to random observations. The higher the score and z-score, the more reliable the DDI type. For each of the 80 selected DDI types, we ranked all structures from the highest to the lowest 3did score and inspected the top-ranked structures. We quickly realized that either the top-ranked structure was approved or none of the structures were. We therefore evaluated at most the top two ranked structures for every DDI type. In total, 95 structures were manually inspected with an average of 2 hours spent per structure. This work resulted in the approval of 48 of 80 potential DDI types (60%) (Fig. 1C). Of note, non-approval of a DDI type does not necessarily mean that this DDI type cannot mediate a PPI, but the available structural data currently does not provide enough evidence. Detailed reasoning for approval and non-approval is provided for every DDI type and inspected structure in Table S1.

Most often a DDI type was not approved because the observed domain-domain contacts were considered too few to mediate the PPI (12 DDI types, Fig. 1D). In these cases, we observed contacts between other domains or neighboring linear motifs in the same structure that were much more likely to mediate the PPI. For example, a structure between the two *Bombyx mori* proteins SP2 and SP3 shows a strong interaction between the Hemocyanin_M domains of both proteins, which is also discussed in detail in the accompanying study (Hou et al., 2014). The contact between the Hemocyanin_M domain in SP3 and the Hemocyanin_C domain in SP2 is, however, very small (Fig. 1E). This interface is not further discussed in the corresponding publication but is reported as a DDI type in 3did. In another case, 3did reports a DDI type between the Plug domain of BtuB and the TonB_C domain of TonB from *Escherichia coli*. The contact between both domains is minimal and involves part of a disordered region, which is unlikely to belong to the fold of the Plug domain. The main contact, which is well described in the literature, is established between a linear motif called the Ton-box located N-terminal to the Plug domain, binding in beta-strand augmentation to TonB_C (Fig. 1F) (Shultis et al., 2006). For nine other DDI types, the observed domain-domain contacts were likely a result from crystallization only (Fig 1D). For example, for DDI types such as PF00019_PF00041 and PF12139_PF02910 the authors of the respective publications studied the corresponding interfaces and dismissed their biological relevance (Fritz et al., 2002; Healey et al., 2015).

For five additional DDI types, reported contacts by 3did between protein fragments indeed likely mediate the interaction observed in the structure, however, at least one of the two interacting protein fragments does not correspond to a tertiary folded domain. For example, the Pfam HMM Proteasome_A_N (PF10584) matches in the N-terminal disordered region of PSMA4, which mediates binding to the AAA domain (PF00004) in protein PSMC1 (Fig. 1G). In another case, the Pfam HMM corresponds to the known prolyl hydroxylation short linear motif in hypoxia-inducible factors (HIFs) (Fig. 1H), which is recognized by VHL leading to the ubiquitination and degradation of HIF (see also the entry DEG_ODPH_VHL_1 in the ELM DB). The motif sits in a disordered region of HIFs. Of note, the Proteasome_A_N and HIF-1 HMMs are classified as "family" and "domain" HMM types by Pfam, respectively. We also found a case where the Pfam HMM betaPIX_CC (PF16523) comprises a coiled-coil region and a C-terminal PDZ-binding motif. The DDI reported by 3did is clearly mediated by the linear motif in the protein ARHGEF7 binding in beta-strand augmentation to the PDZ domain of SHANK1 (Fig. 1I). For five of the inspected DDI types we found that existing folds were further engineered to bind new targets. Ankyrin repeat domains are commonly used to develop DARPINs while V-set domains are used in antibodies and nanobodies to bind antigens. DDI types PF00023_PF07686 and PF00571_PF07686 are two examples of such synthetic interfaces, which should not be used to predict DDIs in physiological PPIs. Among the 32 non-approved DDI types was also one instance (PF11831_PF08231) where reported Pfam HMM and PDB chain annotations in 3did were wrong.

Feature annotation for machine learning

Given the considerable number of non-approved DDI types in our curation dataset, the time needed to manually check a DDI type, and the considerable number of hetero-domain DDI types with interchain structures in 3did, a clear need arises to computationally identify high confident DDI types. To train a classifier for this task, we set out to determine features for DDI types that would serve as input for the classifier. In our curation effort, the most important feature supporting a DDI type was its description in an accompanying publication of the structure. We employed text mining to identify the co-occurrence of domain names in abstracts but realized that available domain names from the Pfam database are rarely similar (and essentially never identical) to domain names used by experimentalists studying these domains (Table 1). We therefore had to exclude text mining of abstracts as a feature to predict approval of DDI types.

In the manual curation process we found information provided by 3did and ProtCID to be useful for the annotation of DDI types. From 3did we used as feature the 3did score and z-score for each DDI type, the fraction of all reported structures for a given DDI type with interchain evidence as well as the total number of reported structures and the number of identified residue-residue contact pairs as defined by 3did. From ProtCID we extracted for a given DDI type and corresponding ProtCID clusters the maximum number of distinct proteins in a cluster, the minimal sequence identity between proteins from different structures in the cluster, and the maximum number of crystal forms. All these ProtCID features can be indicative of how often this type of DDI was observed between diverse protein sequences that were crystallized under different conditions. The hypothesis is that the more often this DDI was found in diverse structures, the less likely it is to be an artifact of the crystallization process. Of these features, the 3did score and z-score, the fraction of structures with interchain interfaces and the number of residue-residue contacts significantly differed in their distribution between approved and non-approved DDI types (Fig S1A-H).

Previous studies indicated the potential use of PPI data for the identification of DDI types by scoring for an enrichment of domain pairs in interacting proteins (Huttlin et al., 2015; Rolland et al., 2014). To avoid study bias, we referred to two systematically generated human protein interactome datasets either obtained from yeast two-hybrid screens (hereafter referred to as the HuRI dataset) or from affinity purification coupled to mass spectrometry experiments (hereafter referred to as the BioPlex dataset) (Huttlin et al., 2017; Luck et al., 2020). We computed for HuRI and BioPlex separately for each DDI type the number of interactions between proteins with matches of the respective domain types and compared this count to counts obtained from degree-controlled randomized networks. We determined z-scores, which were used as a feature. Given the incompleteness of both networks and the restriction to human PPIs, we were not able to compute these features for 16 out of 80 curated DDI types because no proteins with matches of either of the domain types occurred in the network. For the remaining 64 DDI types we observed that the z-score lacked significant discriminatory power to distinguish between approved and non-approved DDI types (Fig. S1I-J). Because of this lack of signal and because this feature was not available for a considerable fraction of our DDI benchmark dataset, we decided to exclude this feature.

Numerous studies, including from our own lab, reported on the ability of AlphaFold-Multimer (AF) to predict with reasonable accuracy the structure of interacting proteins (Akdel et al., 2022; Bryant et al., 2022; Lee et al., 2023; Tsaban et al., 2022). We hypothesized that AF's ability to accurately predict the structure of a DDI might depend on the “strength” or likeliness of that DDI to be able to mediate a PPI. Using AF, we generated a model for the top ranked structure of a given DDI type and computed the DockQ score between the model and actual structure to assess their similarity. The DockQ score and model confidence computed by AF significantly differed between approved and non-approved DDI types (Fig. S1K-L) and were thus selected as features.

To identify DDI types with structures whose interface is partially or fully mediated by disordered regions, we computed the average of the IUPred values of the interface residues using either full protein sequences or those resolved in the structures. We obtained mixed statistical significances (Fig. S1M-N) and selected both as features. Of note, for the aforementioned feature analyses as well as for the downstream modeling we removed all DDI types from the benchmark dataset (5 in total) that involved an Ankyrin or V-set domain since they are commonly used in protein design and thus most often, will not correspond to naturally occurring DDIs. The benchmark dataset with all feature annotations is available in Table S2.

Given the small size of our DDI benchmark dataset in comparison to the large number of hetero-domain DDI types in 3did, we assessed to which extent the mean of the annotated features differed between both sets of DDI types and found that all features apart from the 3did score, 3did z-score, and number of residue-residue contacts did not significantly deviate in their mean (Fig. S2).

Feature selection, training, and testing of logistic regression model

We analyzed the correlations among the features to sort out redundant variables (Fig. S3A-B). Given the small training dataset, we opted to use logistic regression to predict DDI approval. We refer to this model as the full model, because it includes the largest set of variables (Table S3, Fig. S3B). To simplify the model we used backwards stepwise feature selection to keep only the most informative variables. We also excluded the number of intrachain structures from the model as its regression coefficient had a very large standard error, its p-value was very close to 1, and it caused convergence issues (Table S3). The resulting reduced model consisted of the 3did z-score, AF DockQ, and the IUPred score (Table S3, Fig. S3C). We also tested a minimal model consisting only of the 3did z-score to assess if addition of any of the features improved DDI approval prediction compared to using the 3did z-score alone (Table S3). The full model fit significantly better than the minimal model (p=0.007; chi-square test), whereas the reduced model was not significantly inferior to the full one (p=0.336; chi-square test, Fig. 2A-C). We assessed the predictive power of the models through ROC curve analysis using leave-one-out cross-validation (Fig. 2D). A bootstrap test failed to identify significant differences in performance among the models (minimal vs. full: p=0.58; minimal vs. reduced: p=0.32; full vs. reduced: p=0.61). The area under the curve was 0.795, 0.807, and 0.772 for the full, reduced, and minimal model, respectively. This suggests that, although the reduced model fits better the training data than the minimal model, at least with the current benchmark dataset we could not demonstrate a significant improvement in prediction accuracy by the addition of further variables. The overall predictive power of all three models when considering the optimal threshold was found to be in the range of 78 to 81% sensitivity and 73 to 75% specificity.

The 3did z-score is a robust predictive feature in our analysis. The minimal model indicates that every increase of the 3did z-score by one unit roughly doubles the odds of approval for a DDI type (Table S3). The authors of 3did suggest using a 3did z-score of 2.3, which, however, is not applied by default to 3did content. We found that a z-score cutoff of 2.3 results in an expected DDI type approval probability of 33%. Only much higher z-score cutoffs of 4.47 and 5.30 can reach a desired approval probability of 80 and 90%, respectively. Applying a 3did z-score cutoff of 2.3, 4.47, or 5.3 retains 87%, 46%, or 32% of the hetero-protein DDI types with interchain evidence in 3did (Fig. 2E, Table S4).

Prediction of DDIs in the human protein interactome

The DDI type collection in 3did can be used to predict interfaces in protein interactions by scanning the protein sequences of two interacting proteins for the presence of HMM matches that correspond to domain types previously observed to interact with each other. We applied this prediction strategy to the HuRI dataset and found that more stringent 3did z-score cutoffs resulted in larger differences between the number of PPIs with a predicted DDI in HuRI versus degree-controlled randomized networks (Fig. S4). This suggests that higher 3did z-score cutoffs enrich for functional DDI types. Using a 3did z-score cutoff of 4.47 we predicted at least one DDI for 1533 PPIs in HuRI (Fig. 2F, Table S5). For 135 of these PPIs, more than one possible DDI was predicted (Fig. 2G). In total, 342 DDI types resulted in at least one prediction in HuRI and on average a DDI type was predicted to mediate about 2 to 5 PPIs (Fig 2H). Focusing on the most commonly recurring DDI types predicted in HuRI (Fig. 2F) we observe DDI types primarily associated with regulatory and signaling interactions involving ubiquitin-like, BTB, Fbox, and Cullin domain interactions functioning in proteostasis, LSM domain interactions functioning in mRNA splicing, helix-loop-helix, nuclear, and hormone receptor domain interactions functioning in gene expression, as well as SH3, L27, Ras, kinase, and phosphatase-related domains functioning more broadly in signal transduction. There were no PPIs with predicted DDIs known to be part of stable protein complexes, which are more often thought to be mediated by DDIs compared to other modes of protein binding. This along with previously published observations indicates a primary detection of regulatory and signaling-related PPIs in Y2H-based screens (Lambourne et al., 2022; Rolland et al., 2014).

Discussion

Protein interaction interface prediction involving those mediated by folded domains is a challenge in computational structural biology that many scientists are trying to solve. To aid in the prediction of highly confident DDIs in protein interactions, we established a reference set of manually curated DDI types and used it to train a logistic regression model that enabled prediction of likely high confidence DDI types in 3did. Of note, we based our curation on DDI types with interchain evidence and contacts observed between heterodimers. Among this subset, we predicted 2411 DDI types (46%) to be of high confidence. Application of this filtered set of DDI types for DDI prediction on the HuRI protein interactome resulted in a DDI prediction for 1533 PPIs. We had excluded all DDI types from 3did that were derived from structurally resolved homodimers due to particular difficulties in distinguishing biological assemblies from crystal contacts for this class of PPIs. In the future, specific models trained for this task could be incorporated to curate and rank this class of DDIs (Schweke et al., 2023). We also systematically excluded DDI types that only have structural evidence from intrachain contacts. We are aware of many intramolecular modes of binding that, for example, serve functionally important autoinhibitory roles. There is also evidence from studying mainly unicellular organisms that interacting domains from distinct proteins in some species evolved via protein fusion to become part of a single protein chain in other species. Furthermore, artificial protein fusions are sometimes used in crystallography to enhance crystallization. These types of evidence can be obtained by extending the manual curation process to experimental studies beyond the solved structures, enabling addition of a fraction of intrachain DDI types to the set of high confident DDI types in the future. Future DDI curation and prediction efforts may also benefit from incorporation of protein-protein interface features computed by PDBePISA (Krissinel & Henrick, 2007).

We acknowledge that our DDI benchmark dataset is small due to limited time and manpower available to curate DDI types. It is possible that because of the limited size of the training and test dataset, significant differences between the trained models could not be observed. This should also be taken into account when interpreting computed prediction performances. Metrics derived from AF were among the few features that clearly separated approved from non-approved DDI types. Due to limited computational resources we were not able at this point to compute this feature across all DDI types in 3did but this is feasible for future applications. The ability of AF to more accurately model protein interaction interfaces, if they corresponded to biological assemblies, was also more recently reported elsewhere providing further evidence for the use of AF in the scoring of interfaces deposited in the PDB (Schweke et al., 2023). We noted in the curation process that the strongest evidence for supporting a DDI type was obtained from publications accompanying the released structures. However, efforts for the use of text mining to automatically extract such information for incorporation into the logistic regression model failed because there is essentially no consensus on the naming of folded domains. This analysis urges the scientific community to develop standards for the naming of individual types of functional modules in proteins.

A considerable fraction of DDI types were not approved during manual curation because a HMM match did not correspond to a folded domain but rather a conserved disordered region. In other cases the HMM match comprised next to a folded domain a short disordered region that often corresponded to a short linear motif, which was actually mediating the interaction. Pfam HMMs are classified into family, domain, repeat, and signal types, all of which are considered in the search for domain-domain contacts in 3did to the best of our understanding. However, the Pfam HMMs that we found matching to disordered regions are classified as family or domain type meaning that a restriction to these HMM types would not have helped identifying these cases. These observations highlight often neglected limitations from Pfam HMMs, which sometimes do not correspond to folded domains and whose match boundaries often do not correspond to the actual boundaries of the folded regions. Now, with a high confidence structural model from AF at hand for most protein sequences in UniProt, the definition and classification of Pfam HMMs should be revisited and could be aligned with structural annotations provided by CATH for example, to build a next generation of domain type classifications and search algorithms (Bordin et al., 2023; Sillitoe et al., 2021; van Kempen et al., 2023). Ultimately, such resources can be combined with a systematic classification of interface types, which is ever more needed in the advent of interactome-wide protein complex predictions thanks to advances in deep learning. Our study highlighted limitations in currently available sets of DDI types and provided a short-term solution as well as long-term perspectives towards a comprehensive and high quality classification of interface types.

Methods

Computer code for the processing and analysis of all data apart from the modeling was written in Python3 (version 3.9.5) in a Jupyter Notebook environment using the pandas, numpy, matplotlib, seaborn and scipy libraries (Harris et al., 2020; Hunter, 2007; McKinney, 2010; Virtanen et al., 2020; Waskom, 2021).

Processing of 3did content

3did content was downloaded from 3did.irbbarcelona.org/download/current/3did_flat.gz. The downloaded file contains DDI interfaces identified in 3D structures of the PDB. Multiple interfaces can be related to the same DDI type, if the same two domains are found to be in contact with each other. Each interface is uniquely identified by the Pfam-Pfam combination of the interacting domains, the PDB structure it was detected in as well as the ID(s) of the chain(s) the domains were detected in. Furthermore, 3did provides their 3did score and z-score for each interface, as well as the interacting residues for both domains involved in the DDI type. The number of residue-residue contacts per DDI type was derived from the list of contacting residues of the domains mediating the interaction for the interface with the highest 3did score. Intrachain DDI types were defined as being supported only by structures exhibiting intrachain interfaces. If both domains mediating the interaction had the same chain ID, the interface was classified as intrachain. Homo-protein DDI types were defined to be only supported by homodimeric interfaces between the same protein UniProt ID. To this end, we retrieved the UniProt IDs of the proteins present in structures with interfaces formed by a certain DDI type along with the respective chain IDs using the GraphQL-based API of the RCSB PDB. If multiple interfaces of the same DDI type are present in one structure, we selected the interchain interface with the highest 3did score of this structure. If no interchain interface was present in a structure, the intrachain interface with the highest 3did score was kept as the representative interface for the structure.

Development of a manual curation standard and manual curation

For the manual curation of the 80 randomly selected DDI types, we primarily relied on assessing the publication(s) of the structures exhibiting a DDI interface (if they were available). This was conducted by searching for a description of the DDI, either affirming the interaction between the respective regions or identifying the contact as a crystallographic artifact. We did not require the authors to refer to the specific Pfam domains but accepted the description of an interaction between the regions corresponding to the Pfam domains. Furthermore, we evaluated information provided in the publication about interaction stoichiometries, the localization of both proteins, and other functional background of the interaction. This information was especially helpful to evaluate the DDI type, if the DDI itself was not described in the publication. We also investigated whether the interacting proteins were the result of synthetically engineered protein scaffolds or whether the structure contained a specifically designed antibody binding its target protein. If the DDI was described as a functional interface meditating the PPI within the publication, we considered this as sufficient to evaluate the DDI type as approved. Following the assessment of the publication(s), we continued the curation process by inspecting the residue-residue contacts between the domains of the DDI using the PyMOL Molecular Graphics System (version 2.5.0). The ProtCID database provides metrics for interaction interfaces to aid in the identification of protein contacts that likely originated from crystallization. ProtCID is based on the assumption that conservation of biological assemblies among homologous proteins identified in different crystal forms is unlikely to be the result of crystallization alone. They provide clusters of interfaces present in PDB structures in at least two different crystal forms, with further annotations including the PDB structures contributing to the cluster, the UniProt IDs of the proteins in the structures, the minimal sequence identity of the proteins in the cluster or the crystal forms of the PDB structures in the cluster. As the final step in our manual curation, we queried the web server of the ProtCID database (http://dunbrack2.fccc.edu/ProtCiD/default.aspx) with the Pfam-Pfam combination of a given DDI type and checked for available clusters and derived the number of different crystal forms as well as the minimal sequence identity. If there was a cluster existing that included the curated structure(s) of the DDI type that had a minimal sequence identity < 80% and showed at least 5 different crystal forms, we approved the cluster as supporting the DDI type. An interface present in a PDB structure supporting a specific DDI type was finally approved, if the related publication(s) described the interface to be able to mediate the respective protein-protein interaction. In case the available literature did not refer to the interface at all, we approved an interface, if the detected residue-residue contacts between the Pfam domains agreed with known stoichiometries and subcellular localizations of both proteins (i.e. interfaces between extracellular hormones and intracellular domains of receptors were not approved) and the contact between the domains provided a major interface for the interacting proteins. Furthermore, we included ProtCID information, if available, to assess whether the interface represents a crystal contact or a biologically relevant interaction. If there was a related publication available that provided evidence for the interface to be the product of crystallization or a synthetically engineered complex, we did not approve the interface.

Feature annotation for machine learning

All features for the trained classifier derived from 3did were obtained from the 3did download mentioned earlier. ProtCID cluster data for Pfam-Pfam interactions that was used as features for model training was kindly provided by the Dunbrack lab. We downloaded the HuRI and BioPlex dataset from the supplementary information from Luck et al. (2020) and removed all homodimers prior to network randomization. 1000 degree-controlled randomized PPI networks of each of the HuRI and BioPlex network were generated using the igraph library degree_sequence() function in python. As the networks comprise only Ensembl gene IDs of the interacting protein pairs, we used Bioconductor BiomaRt to map the Ensembl gene IDs to their respective UniProt IDs. This was necessary to obtain annotations for proteins from other resources such as UniProt. We downloaded the Pfam HMM matches for all Swiss-Prot reviewed human proteins from Interpro (Paysan-Lafosse et al., 2023) along with their UniProt IDs in a .json file, so that we could annotate the Pfam domains for every protein in the HuRI and BioPlex networks based on their UniProt IDs. The z-scores for a given DDI type for the HuRI and BioPlex network were computed separately by first determining the number of PPIs in each respective network where both proteins matched the Pfam HMMs of a given DDI type, followed by subtracting from this count the mean number of PPIs with the

DDI observed in the randomized network. Then, the difference was divided by the standard deviation of the PPI count from the randomized networks to derive the z-scores.

AlphaFold Multimer version 2.2.0 was used to generate models of the interaction between the domains of a given DDI type. To this end, we used PyMOL (version 2.5.0) to select the sequences in the structure that corresponded to both Pfam HMM matches for export. If the HMM matches did not fully cover the folded regions of either domain then the sequence(s) were extended, respectively, and an additional 10 residues added to the N- and C-terminus of the domains prior to export. This was done to ensure that all residues important for the fold of the domain were included as prior experience showed that AlphaFold models for interacting proteins are hugely misled if the folds of the domains are not complete. These sequences were used as input for model prediction by AlphaFold Multimer version 2.2.0. For running AlphaFold, we used the default query databases provided by AlphaFold on its GitHub page. The predictions were done using the full_dbs option and the use of template was allowed by setting the flag max_template_date to 2020-05-14. We used CPUs to relax the predicted models by toggling the flag use_gpu_relax to False. Five models were predicted with a single seed per model by setting the flag num_multimer_predictions_per_model to 1. The calculation of the DockQ score was done as described by Basu & Waller (2016). For the computation of the fraction of disordered residues at the interface, we used the structure of a DDI type with the highest 3did score and retrieved IUPred2A values for the protein sequence used for crystallization as well as for the full-length sequence of the proteins containing the domains forming a DDI using a local installation of IUPred2A and the short disorder prediction mode (Mészáros et al., 2018). We retrieved the sequences used for crystallization using PyMOL (version 2.5.0) and its sequence export function. The full-length sequences were retrieved using the UniProt API. We extracted the IUPred2A scores for the residues in both proteins that were listed as contacting residues according to 3did and calculated the average IUPred2A score of those residues to use as a feature, either using the sequences used for crystallization or the full-length sequences. We used both sequence versions because sequence context can influence disorder prediction by IUPred. For assessment of statistical significance of the individual features, a Mann-Whitney-U test was performed for the approved and non-approved subsets of DDI types in the manual curation set.

Statistical modeling

Model training and analysis was performed with code written in R version 4.3.0. The logistic regression and leave-one-out cross validation were performed using R base functions. Odds ratios are obtained from the logistic model estimates through the function $f(x) = e^{x}$, where $e \approx 2.72$, and x is the estimate from the logistic model (Agresti, 2012). ROC curves and optimal thresholds for calculating sensitivity and specificity (determined using Youden's criterion) were computed using the package pROC version 1.18.2. The package glmtoolbox version 0.1.7 was used to compute Hosmer-Lemeshow goodness-of-fit tests. All code is available at https://github.com/KatjaLuckLab/DDI_manuscript including the feature table for model training.

Acknowledgements

We thank all members of the Luck lab for fruitful discussions of results from this manuscript. We thank Milena Djokic for help with the generation of randomized networks and Toby Gibson for helpful feedback on the manuscript.

Funding

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) [Project-IDs LU 2568/1-1, SFB1551 464588647] awarded to K.L. and a postdoctoral stipend from the Alexander von Humboldt foundation awarded to J.J.N.

References

Agresti, A. (2012). *Categorical Data Analysis*. John Wiley & Sons.
Akdel, M., Pires, D. E. V., Pardo, E. P., Jänes, J., Zalevsky, A. O., Mészáros, B., Bryant, P.,

Good, L. L., Laskowski, R. A., Pozzati, G., Shenoy, A., Zhu, W., Kundrotas, P., Serra, V. R., Rodrigues, C. H. M., Dunham, A. S., Burke, D., Borkakoti, N., Velankar, S., … Beltrao, P. (2022). A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, *29*(11), 1056–1067. https://doi.org/10.1038/s41594-022-00849-w

Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.-C., Bork, P., Superti-Furga, G., Serrano, L., & Russell, R. B. (2004). Structure-based assembly of protein complexes in yeast. *Science (New York, N.Y.)*, *303*(5666), 2026–2029. https://doi.org/10.1126/science.1092645

Basu, S., & Wallner, B. (2016). DockQ: A Quality Measure for Protein-Protein Docking Models. *PloS One*, *11*(8), e0161879. https://doi.org/10.1371/journal.pone.0161879

Bordin, N., Sillitoe, I., Nallapareddy, V., Rauer, C., Lam, S. D., Waman, V. P., Sen, N., Heinzinger, M., Littmann, M., Kim, S., Velankar, S., Steinegger, M., Rost, B., & Orengo, C. (2023). AlphaFold2 reveals commonalities and novelties in protein structure space for 21 model organisms. *Communications Biology*, *6*(1), 160. https://doi.org/10.1038/s42003-023-04488-9

Braun, P., & Gingras, A.-C. (2012). History of protein-protein interactions: From egg-white to complex networks. *Proteomics*, *12*(10), 1478–1498. https://doi.org/10.1002/pmic.201100563

Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-28865-w

Elez, K., Bonvin, A. M. J. J., & Vangone, A. (2020). Biological vs. Crystallographic Protein Interfaces: An Overview of Computational Approaches for Their Classification. *Crystals*, *10*(2), Article 2. https://doi.org/10.3390/cryst10020114

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., … Hassabis, D. (2022). *Protein complex prediction with AlphaFold-Multimer* (p. 2021.10.04.463034). bioRxiv. https://doi.org/10.1101/2021.10.04.463034

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Research*, *42*(Database issue), D222–D230. https://doi.org/10.1093/nar/gkt1223

Fritz, G., Roth, A., Schiffer, A., Büchert, T., Bourenkov, G., Bartunik, H. D., Huber, H., Stetter, K. O., Kroneck, P. M. H., & Ermler, U. (2002). Structure of adenylylsulfate reductase from the hyperthermophilic Archaeoglobus fulgidus at 1.6-A resolution. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(4), 1836–1841. https://doi.org/10.1073/pnas.042664399

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., … Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), Article 7825. https://doi.org/10.1038/s41586-020-2649-2

Healey, E. G., Bishop, B., Elegheert, J., Bell, C. H., Padilla-Parra, S., & Siebold, C. (2015). Repulsive guidance molecule is a structural bridge between neogenin and bone morphogenetic protein. *Nature Structural & Molecular Biology*, *22*(6), 458–465. https://doi.org/10.1038/nsmb.3016

Hou, Y., Li, J., Li, Y., Dong, Z., Xia, Q., & Yuan, Y. A. (2014). Crystal structure of Bombyx mori arylphorins reveals a 3:3 heterohexamer with multiple papain cleavage sites. *Protein Science*, *23*(6), 735–746. https://doi.org/10.1002/pro.2457

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., Gygi, M. P., Thornock, A., Zarraga, G., Tam, S., Szpyt, J., Gassaway, B. M., Panov, A., Parzen, H., Fu, S., Golbazi, A., Maenpaa, E., Stricker, K., Guha Thakurta, S., … Gygi, S. P. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, *184*(11), 3022-3040.e28. https://doi.org/10.1016/j.cell.2021.04.011

Huttlin, E. L., Bruckner, R. J., Paulo, J. A., Cannon, J. R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M. P., Parzen, H., Szpyt, J., Tam, S., Zarraga, G., Pontano-Vaites, L., Swarup, S., White, A. E., Schweppe, D. K., Rad, R., Erickson, B. K., … Harper, J. W. (2017). Architecture of the human interactome defines protein communities and disease networks. *Nature*, *545*(7655), 505–509. https://doi.org/10.1038/nature22366

Huttlin, E. L., Ting, L., Bruckner, R. J., Gebreab, F., Gygi, M. P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., Dong, R., Guarani, V., Vaites, L. P., Ordureau, A., Rad, R., Erickson, B. K., Wühr, M., Chick, J., Zhai, B., … Gygi, S. P. (2015). The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell*, *162*(2), 425–440. https://doi.org/10.1016/j.cell.2015.06.043

Janin, J., & Wodak, S. J. (1983). Structural domains in proteins and their role in the dynamics of protein function. *Progress in Biophysics and Molecular Biology*, *42*(1), 21–78. https://doi.org/10.1016/0079-6107(83)90003-2

Karan, B., Mahapatra, S., Sahu, S. S., Pandey, D. M., & Chakravarty, S. (2022). Computational models for prediction of protein-protein interaction in rice and Magnaporthe grisea. *Frontiers in Plant Science*, *13*, 1046209. https://doi.org/10.3389/fpls.2022.1046209

Krissinel, E., & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of Molecular Biology*, *372*(3), 774–797. https://doi.org/10.1016/j.jmb.2007.05.022

Lambourne, L., Yadav, A., Wang, Y., Desbuleux, A., Kim, D.-K., Cafarelli, T., Pons, C., Kovács, I. A., Jailkhani, N., Schlabach, S., Ridder, D. D., Luck, K., Bian, W., Shen, Y., Yang, Z., Mee, M. W., Helmy, M., Jacob, Y., Lemmens, I., … Vidal, M. (2022). *Binary interactome models of inner- versus outer-complexome organisation* (p. 2021.03.16.435663). bioRxiv. https://doi.org/10.1101/2021.03.16.435663

Lee, C. Y., Hubrich, D., Varga, J. K., Schäfer, C., Welzel, M., Schumbera, E., Đokić, M., Strom, J. M., Schönfeld, J., Geist, J. L., Polat, F., Gibson, T. J., Valsecchi, C. I. K., Kumar, M., Schueler-Furman, O., & Luck, K. (2023). *Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation* (p. 2023.08.07.552219). bioRxiv. https://doi.org/10.1101/2023.08.07.552219

Liddington, R. C. (2004). Structural Basis of Protein-Protein Interactions. In H. Fu (Ed.), *Protein-Protein Interactions: Methods and Applications* (pp. 3–14). Humana Press. https://doi.org/10.1385/1-59259-762-9:003

Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B., Choi, D., Coté, A. G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., … Calderwood, M. A. (2020). A reference map of the human binary protein interactome.

Nature, *580*(7803), 402–408. https://doi.org/10.1038/s41586-020-2188-x

Luther, C. H., Brandt, P., Vylkova, S., Dandekar, T., Müller, T., & Dittrich, M. (2023). Integrated analysis of SR-like protein kinases Sky1 and Sky2 links signaling networks with transcriptional regulation in Candida albicans. *Frontiers in Cellular and Infection Microbiology*, *13*, 1108235. https://doi.org/10.3389/fcimb.2023.1108235

McKinney, W. (2010). *Data Structures for Statistical Computing in Python*. 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Mészáros, B., Erdős, G., & Dosztányi, Z. (2018). IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research*, *46*(W1), W329–W337. https://doi.org/10.1093/nar/gky384

Michaelis, A. C., Brunner, A.-D., Zwiebel, M., Meier, F., Strauss, M. T., Bludau, I., & Mann, M. (2023). The social and structural architecture of the yeast protein interactome. *Nature*, *624*(7990), 192–200. https://doi.org/10.1038/s41586-023-06739-5

Mosca, R., Céol, A., & Aloy, P. (2013). Interactome3D: Adding structural details to protein networks. *Nature Methods*, *10*(1), 47–53. https://doi.org/10.1038/nmeth.2289

Mosca, R., Céol, A., Stein, A., Olivella, R., & Aloy, P. (2014). 3did: A catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, *42*(D1), D374–D379. https://doi.org/10.1093/nar/gkt887

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Rivoire, C., … Bateman, A. (2023). InterPro in 2022. *Nucleic Acids Research*, *51*(D1), D418–D427. https://doi.org/10.1093/nar/gkac993

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., … Vidal, M. (2014). A proteome-scale map of the human interactome network. *Cell*, *159*(5), 1212–1226. https://doi.org/10.1016/j.cell.2014.10.050

Schweke, H., Xu, Q., Tauriello, G., Pantolini, L., Schwede, T., Cazals, F., Lhéritier, A., Fernandez-Recio, J., Rodríguez-Lumbreras, L. A., Schueler-Furman, O., Varga, J. K., Jiménez-García, B., Réau, M. F., Bonvin, A. M. J. J., Savojardo, C., Martelli, P.-L., Casadio, R., Tubiana, J., Wolfson, H. J., … Wodak, S. J. (2023). Discriminating physiological from non-physiological interfaces in structures of protein complexes: A community-wide study. *Proteomics*, *23*(17), e2200323. https://doi.org/10.1002/pmic.202200323

Shultis, D. D., Purdy, M. D., Banchs, C. N., & Wiener, M. C. (2006). Outer membrane active transport: Structure of the BtuB:TonB complex. *Science (New York, N.Y.)*, *312*(5778), 1396–1399. https://doi.org/10.1126/science.1127694

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J., & Orengo, C. A. (2021). CATH: Increased structural coverage of functional space. *Nucleic Acids Research*, *49*(D1), D266–D273. https://doi.org/10.1093/nar/gkaa1079

Stein, A., Céol, A., & Aloy, P. (2011). 3did: Identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, *39*(Database issue), D718-723. https://doi.org/10.1093/nar/gkq962

Tsaban, T., Varga, J. K., Avraham, O., Ben-Aharon, Z., Khramushin, A., & Schueler-Furman, O. (2022). Harnessing protein folding neural networks for peptide–protein docking.

Nature Communications, *13*(1), Article 1. https://doi.org/10.1038/s41467-021-27838-9

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

Vidal, M., Cusick, M. E., & Barabási, A.-L. (2011). Interactome networks and human disease. *Cell*, *144*(6), 986–998. https://doi.org/10.1016/j.cell.2011.02.016

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Weatheritt, R. J., Luck, K., Petsalaki, E., Davey, N. E., & Gibson, T. J. (2012). The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics (Oxford, England)*, *28*(7), 976–982. https://doi.org/10.1093/bioinformatics/bts072

Wodak, S. J., & Janin, J. (2002). Structural basis of macromolecular recognition. *Advances in Protein Chemistry*, *61*, 9–73. https://doi.org/10.1016/s0065-3233(02)61001-0

Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., & Jothi, R. (2011). DOMINE: A comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Research*, *39*(Database issue), D730-735. https://doi.org/10.1093/nar/gkq1229

Zhang, X., Jiao, X., Song, J., & Chang, S. (2016). Prediction of human protein-protein interaction by a domain-based approach. *Journal of Theoretical Biology*, *396*, 144–153. https://doi.org/10.1016/j.jtbi.2016.02.026

Zheng, C., Liu, Y., Sun, F., Zhao, L., & Zhang, L. (2021). Predicting Protein-Protein Interactions Between Rice and Blast Fungus Using Structure-Based Approaches. *Frontiers in Plant Science*, *12*, 690124. https://doi.org/10.3389/fpls.2021.690124

Tables

DDI type	Pfam domain names	Pfam short names	Domain names used in publications
PF00059_PF00041	Lectin C-type domain; Fibronectin type III domain	Lectin_C; fn3	C-type lectin (CLD) subdomains; fibronectin type III repeats 3–5
PF12162_PF02135	STAT1 TAZ2 binding domain; TAZ zinc finger	STAT1_TAZ2bind; zf-TAZ	transactivation domains (TADs) of STAT2 and STAT1; TAZ1
PF00514_PF00104	Armadillo/beta-catenin-like repeat;	Arm; Hormone_recep	β-catenin armadillo repeat; (LRH-1) ligand binding domain
PF01298_PF00405	Ligand-binding domain of nuclear hormone receptor; Transferrin	TbpB_B_D; Transferrin	N-lobe β-barrel domain; C1/C2 subdomain
PF00026_PF06394	Eukaryotic aspartyl protease; Pepsin inhibitor-3-like repeated domain	Asp; Pepsin-I3	Pepsin; N-terminal domain of PI-3

Table 1. Examples of domain names in Pfam/InterPro and corresponding domain names used in publications for selected DDI types in 3did indicated with their Pfam identifiers.

Figures

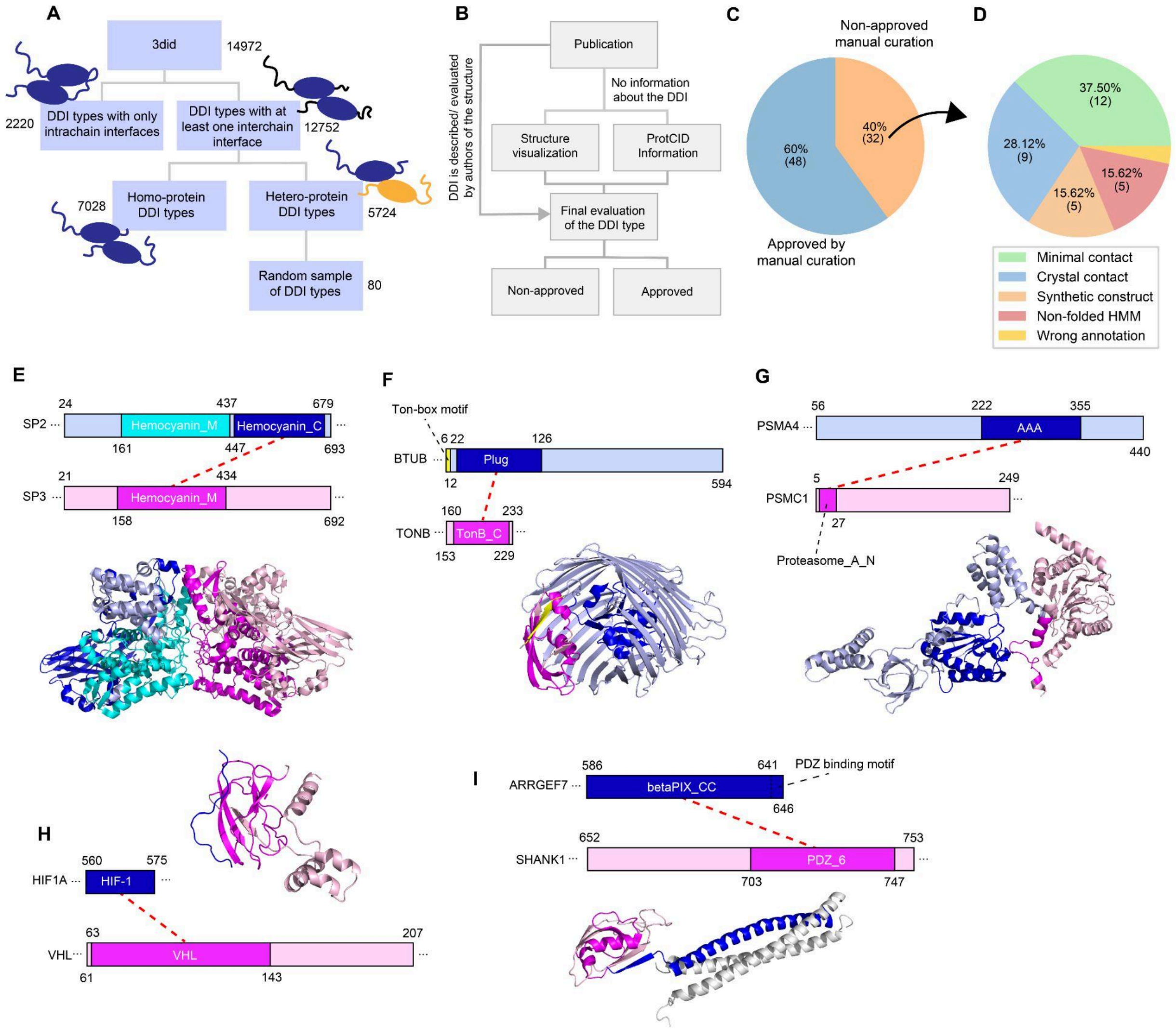


Fig. 1. Classification of DDI types in the 3did database and manual curation. **A**) Schematic describing the classification of DDI types from 3did and from which subset the sample of DDI types were drawn for manual curation. **B**) Schematic describing the manual curation procedure. **C**) Pie chart showing the fraction of approved and non-approved DDI types from manual curation. **D**) Pie chart showing categories and fractions of non-approved DDI types. **E-I**) Exemplary structures and domain architectures of protein pairs from non-approved DDI types. Red lines mark the DDI as reported in 3did. Structures were obtained from PDB IDs 3wjm, 2gsk, 6epf, 6gfx, and 3l4f.

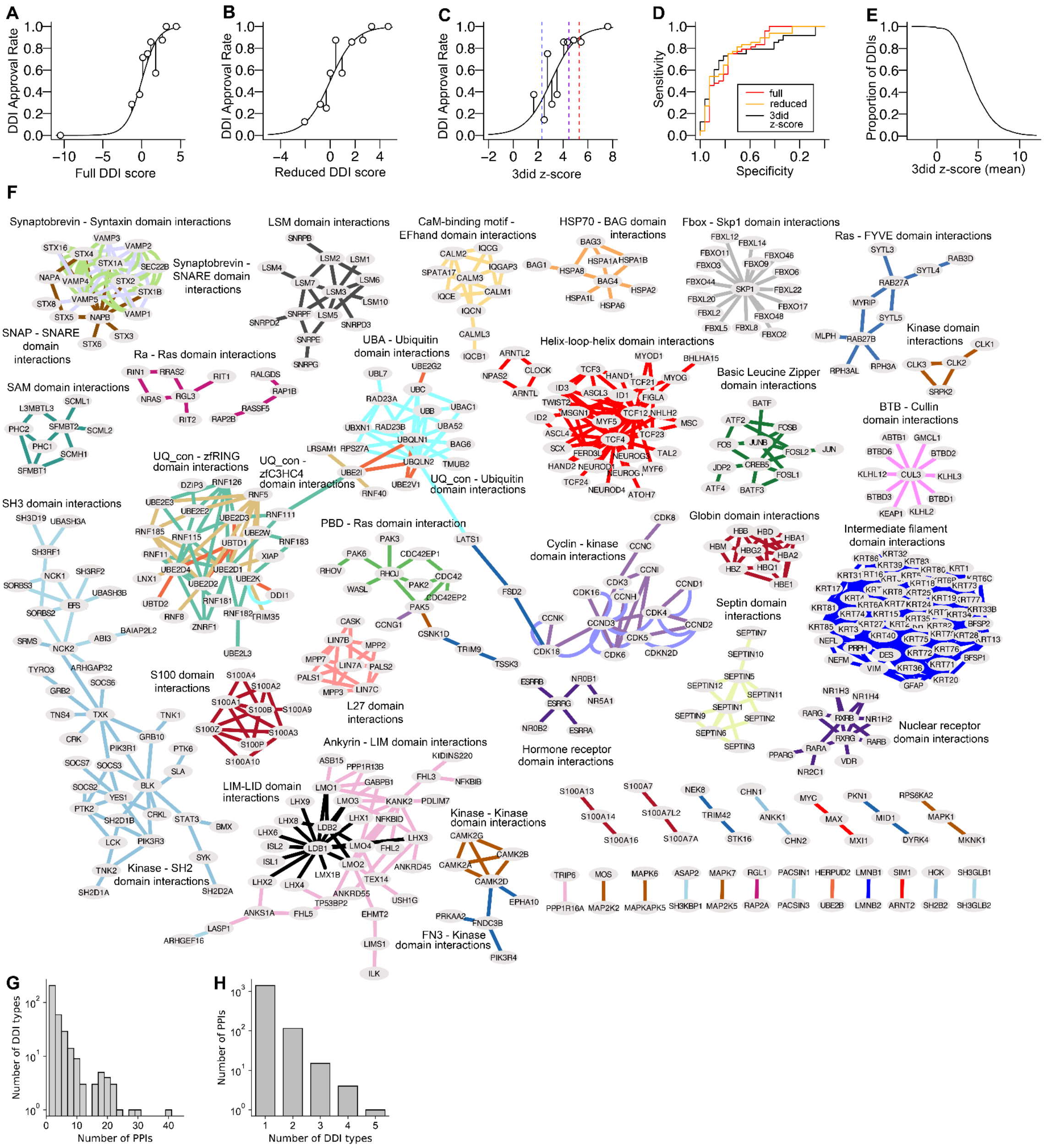


Fig. 2. Model training and DDI prediction in HuRI. **A-C**) Empirical fit assessment is shown for the full (A), reduced (B), and minimal model (C). The DDI training dataset was split in 10 groups by deciles and for each group the proportion of approved DDI types is shown. The z-score threshold of 2.3 (y=0.31), 4.47 (y=0.8), and 5.30 (y=0.9) are plotted in blue, purple, and red dashed lines, respectively in panel C. **D**) ROC curves for all three models computed using leave-one-out cross-validation. **E**) Proportion of heteroprotein DDI types with interchain evidence in 3did retained based on increasing 3did z-score cutoffs. **F**) Network of PPIs from HuRI with predicted DDIs. Shown are PPIs with the most commonly predicted DDI types. Edges are colored based on DDI type predictions. Predicted DDI types are indicated. Nodes are labeled with gene symbols. **G**) Histogram showing the number of PPIs based on how many different DDI types were predicted to occur in any given PPI (outliers excluded from the plot). **H**) Histogram showing how many DDI types were predicted to occur in how many PPIs.

Supplementary Information

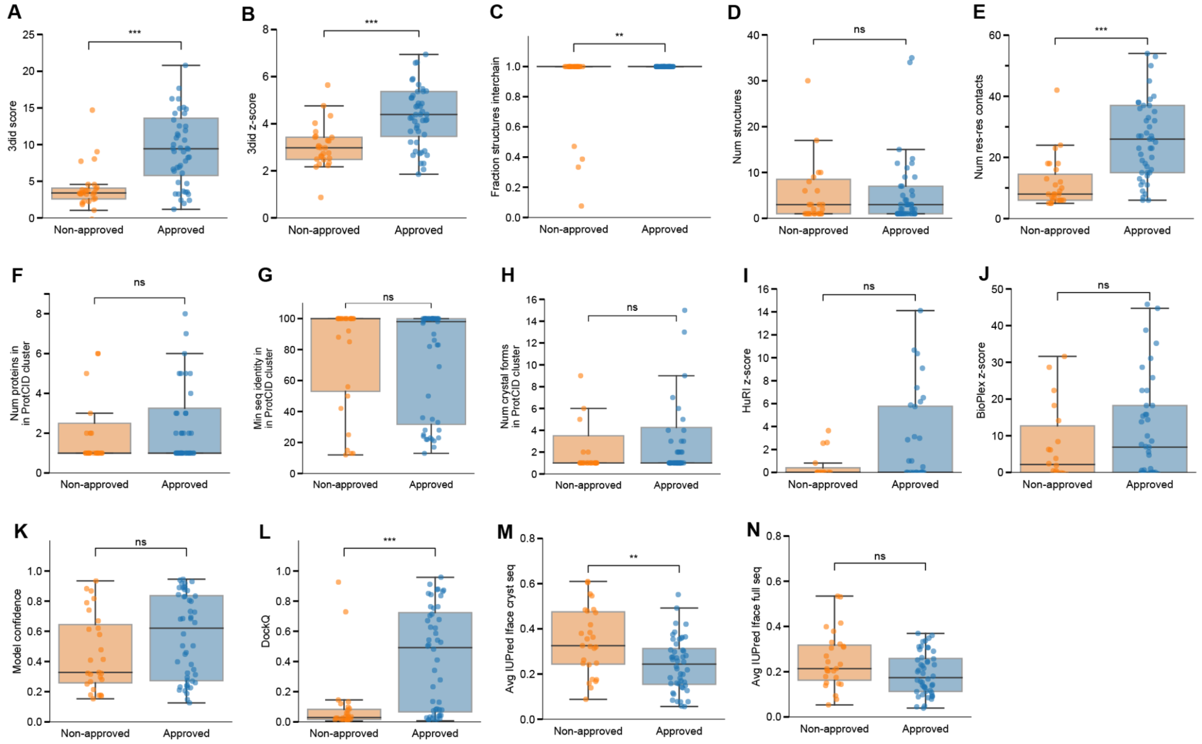


Fig. S1. Feature exploration to discriminate approved from non-approved DDI types. **A-N**) Boxplots showing the distribution of approved and non-approved DDI types for different features as indicated on the y-axis. Significances were computed using the Mann-Whitney-U test.

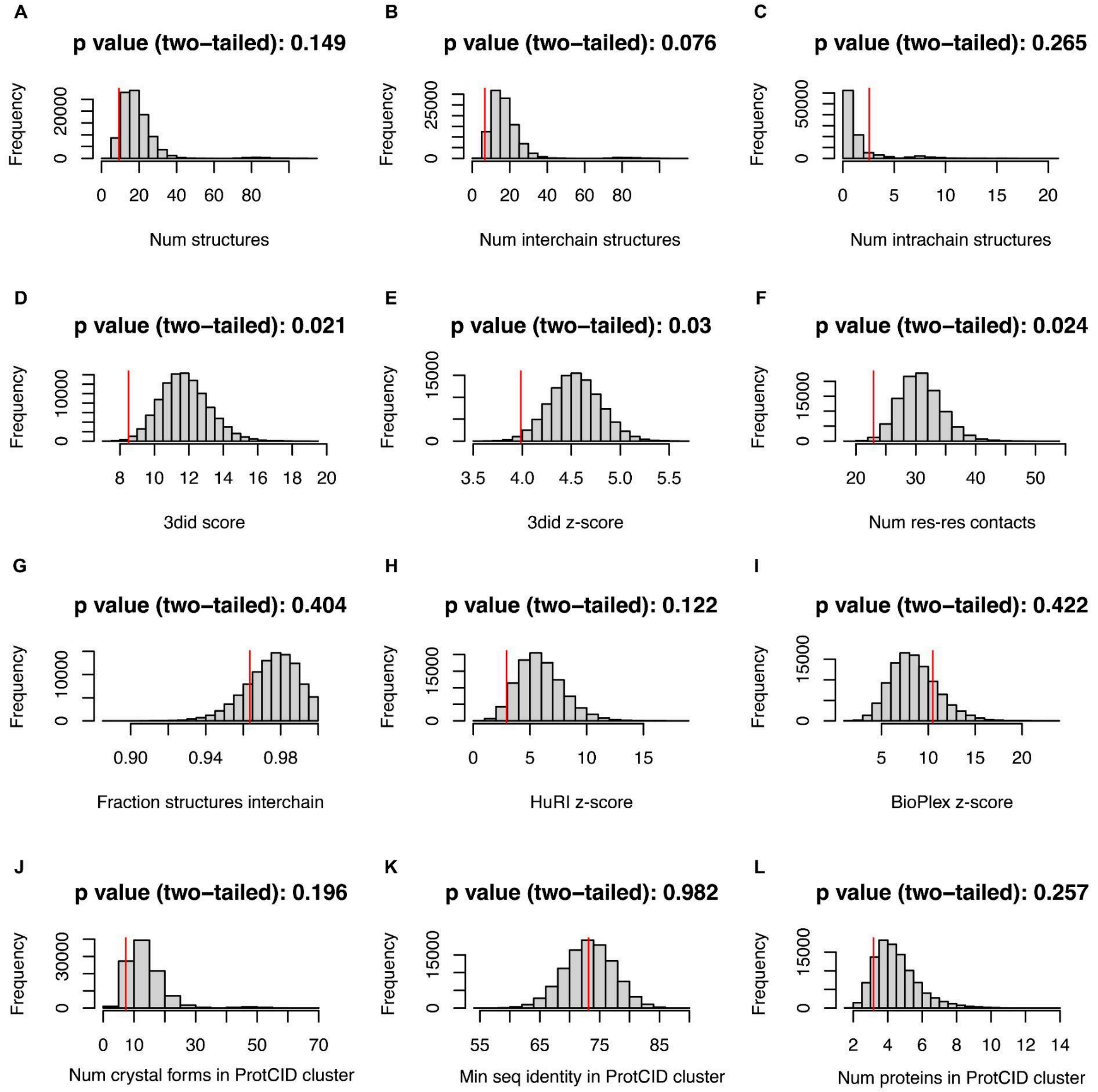


Fig. S2. Representativity analysis of features in manual curation DDI dataset in comparison to the subset of hetero-protein DDI types with interchain evidence. **A-L**) The average value in the different statistics for the curation dataset is shown as a red line overlaying the distribution obtained by random sampling of the 3did DDI types with n=75.

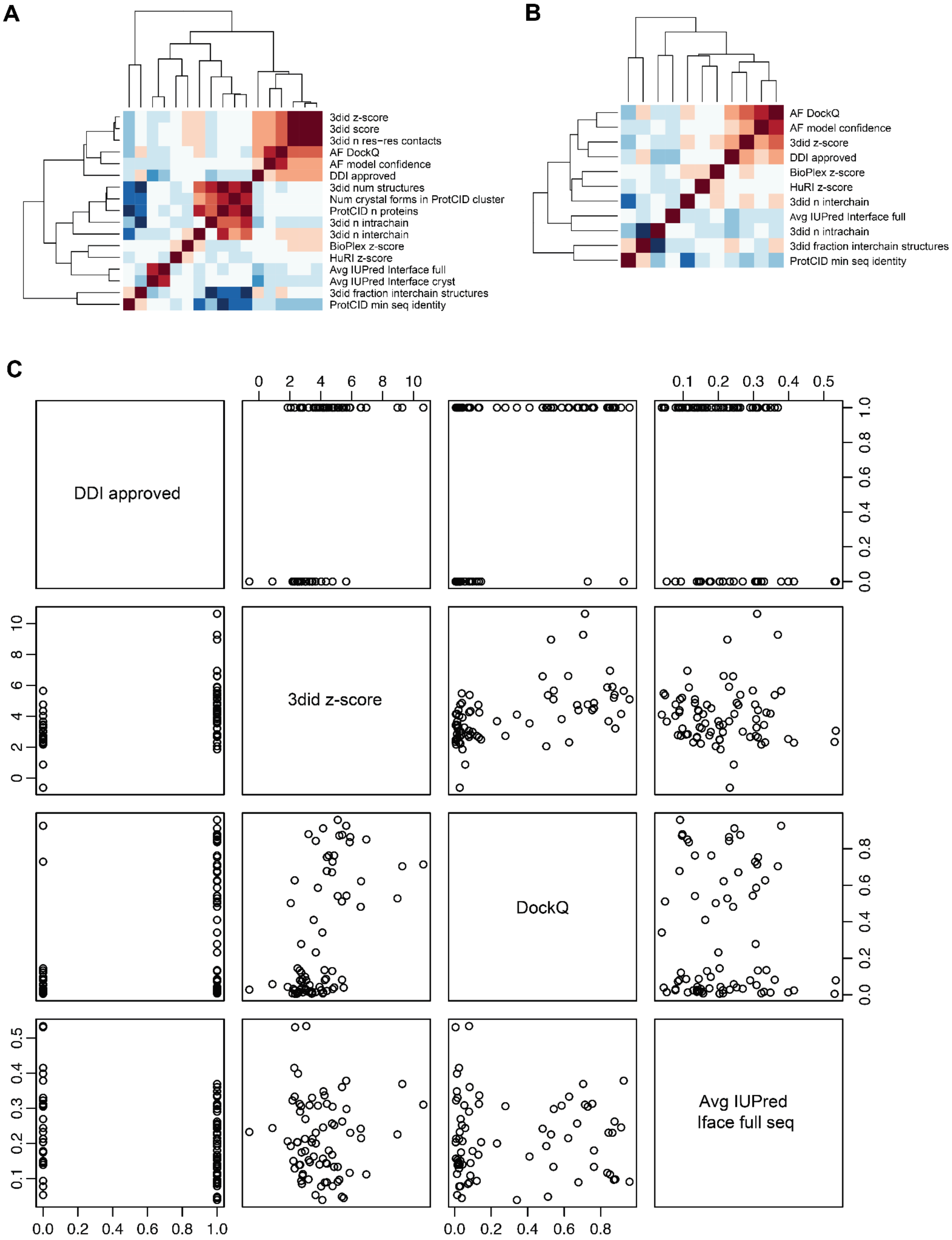


Fig. S3. Correlation matrices between features. **A**) Correlation matrix and hierarchical clustering of all features considered in this study. To compute correlations, pairwise complete data were used. **B**) Correlation matrix of the features used for model training. **C**) Pairwise scatterplot of the features used in the reduced model.

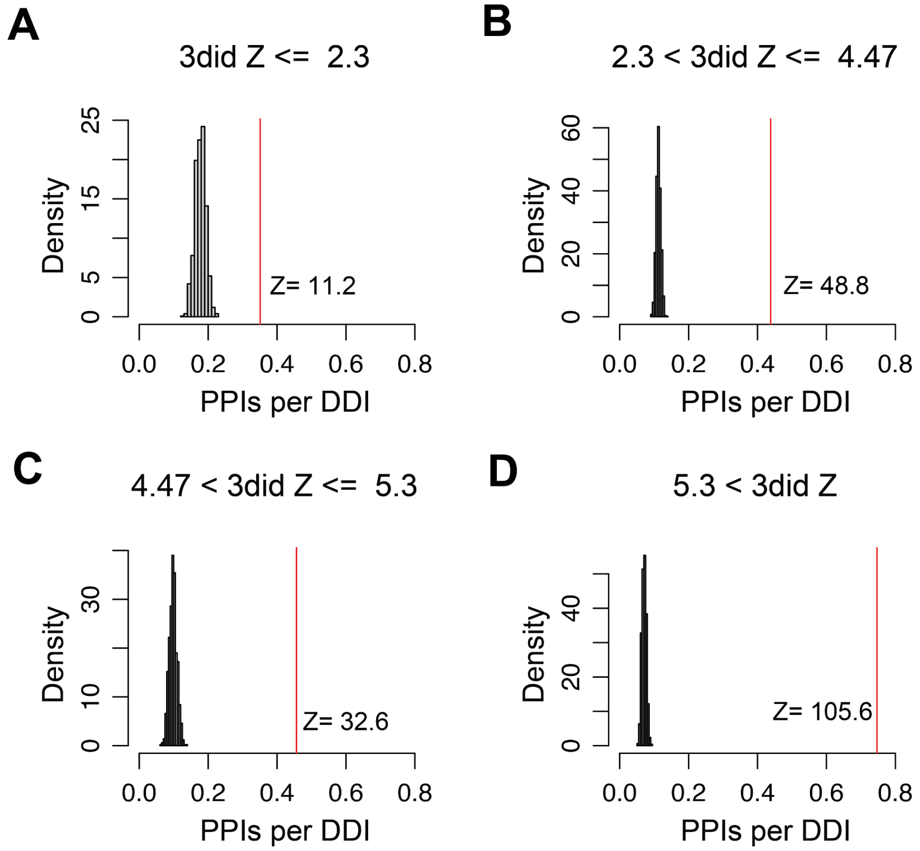


Fig. S4. Enrichment analysis for the number of PPIs in HuRI predicted with a DDI using different subsets of DDI types based on increasing confidence as defined by indicated 3did z-score ranges (**A-D**). Z-scores indicated in the plot represent enrichment over background distribution. The number of PPIs with a predicted DDI were normalized by the number of DDI types available in the corresponding subset.

Chapter 4

Benchmarking AlphaFold-Multimer and its application in characterizing novel interfaces

4.1 Article II: Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation

Summary

This project focused on benchmarking AF-MM and its metrics, and on applying it to predict novel interaction interfaces between proteins.

AF-MM is an AI-based tool for the structural prediction of protein complexes. Numerous studies have tested the ability of AF-MM to predict different interface types. However, there is a lack of a comprehensive assessment of the sensitivity, specificity, and potential biases of AF-MM and its metrics. Such an assessment of AF-MM is paramount for its application in predicting PPI interfaces. To systematically benchmark AF-MM's ability to predict the structures of DMIs, we used a list of DMIs whose structures were available, and their minimal interacting regions (a domain and a motif) were subject to structure prediction by AF-MM. The predicted models were evaluated in reference to the solved structures. We found that 35% of the DMIs were predicted so accurately that even the sidechains of the motifs were correctly positioned, and for another 32%, only the backbones but not the sidechains of the motifs were correctly predicted.

Despite AF-MM outputting various metrics that evaluate the quality of the predicted models, no clear criteria are provided by AF-MM to deem a model to be of high confidence. Different metrics were assessed for their abilities to discriminate between known DMIs and randomly paired DMIs. Cutoff values on the best-performing metrics were subsequently derived to achieve the optimal separation of good and bad models. We then applied AF-MM on DMI instances annotated in ELM that lack structural information, and the predicted models were evaluated using the cutoff values. Experimental validation was subsequently performed to verify one of the predicted interfaces.

We further examined the effect of sequence length on prediction performance of AF-MM, and found that sequence length had a detrimental effect on prediction accuracy. This result led us to devise a strategy that fragments protein sequences into short and overlapping fragments for AF-MM predictions. While this approach boosted the sensitivity of AF-MM at detecting potential interaction interfaces between interacting proteins, we also observed a decrease in AF-MM's specificity. In spite of filtering the models with the derived cutoff values, manual inspection of the high-scoring models remained necessary to gauge their validity. Some of the high-scoring models were also experimentally validated.

In summary, this project provided a thorough assessment of AF-MM and its metrics, explored the use of AF-MM to predict novel PPI interfaces, and highlighted potential caveats in its application.

Statement of contribution

This is a collaborative project in which I conducted the following aspects of the study: benchmarking of AF-MM and its metrics, applying AF-MM to predict PPI interfaces, certain parts of the experimental validation of interfaces predicted by AF-MM, and the writing of the manuscript. I was in charge of processing, organizing, storing and analyzing AF-MM predictions. I also took part in the cloning of open reading frames (ORFs) and their mutants, as well as in testing interactions between proteins. I assembled some figures and tables for the manuscript and participated in the writing and revision of the manuscript.

Supervisor confirmation

Article

SOURCE DATA TRANSPARENT PROCESS OPEN ACCESS CHECK FOR UPDATES

molecular systems biology

Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation

Chop Yan Lee[1,5], Dalmira Hubrich[1,5], Julia K Varga[2,5], Christian Schäfer[1], Mareen Welzel[1], Eric Schumbera[1,4], Milena Djokic[1], Joelle M Strom[1], Jonas Schönfeld[1], Johanna L Geist[1], Feyza Polat[1], Toby J Gibson[3], Claudia Isabelle Keller Valsecchi[1], Manjeet Kumar[3], Ora Schueler-Furman[2]✉ & Katja Luck[1]✉

Abstract

Structural resolution of protein interactions enables mechanistic and functional studies as well as interpretation of disease variants. However, structural data is still missing for most protein interactions because we lack computational and experimental tools at scale. This is particularly true for interactions mediated by short linear motifs occurring in disordered regions of proteins. We find that AlphaFold-Multimer predicts with high sensitivity but limited specificity structures of domain-motif interactions when using small protein fragments as input. Sensitivity decreased substantially when using long protein fragments or full length proteins. We delineated a protein fragmentation strategy particularly suited for the prediction of domain-motif interfaces and applied it to interactions between human proteins associated with neurodevelopmental disorders. This enabled the prediction of highly confident and likely disease-related novel interfaces, which we further experimentally corroborated for FBXO23-STX1B, STX1B-VAMP2, ESRRG-PSMC5, PEX3-PEX19, PEX3-PEX16, and SNRPB-GIGYF1 providing novel molecular insights for diverse biological processes. Our work highlights exciting perspectives, but also reveals clear limitations and the need for future developments to maximize the power of Alphafold-Multimer for interface predictions.

Keywords AlphaFold; Protein Interaction Interface Prediction; Linear Motifs; Benchmarking; Experimental Validation

Subject Categories Computational Biology; Structural Biology

https://doi.org/10.1038/s44320-023-00005-6

Received 3 August 2023; Revised 4 December 2023; Accepted 5 December 2023

Published online: 15 January 2024

Introduction

Protein-protein interactions (PPIs) are essential for the proper functioning of essentially all cellular processes. The last decade has seen tremendous progress in the systematic mapping of human protein interactions enabling gene function prediction and the study of genotype-to-phenotype relationships (Luck et al, 2020; Drew et al, 2017; Huttlin et al, 2021). However, to understand the molecular function of individual PPIs, co-existence or mutual exclusivity of partner proteins in protein complexes, and the effect of mutations on protein function, structural information on how these proteins interact with each other is required. Unfortunately, a structure at atomic resolution is only available for ~4% of known human PPIs (Luck et al, 2020). Modular proteins interact with each other using a variety of different functional elements such as stably folded domains, intrinsically disordered polypeptide regions, short linear motifs (hereafter referred to as motifs), or coiled-coil helices forming domain-domain, domain-motif, disorder-disorder, or coiled-coil interfaces for example. Resources such as 3did (Mosca et al, 2014) or the ELM database (ELM DB) (Kumar et al, 2022) collect observed contacts between domain types and between domains and motifs, respectively. Such interface type collections can be used to predict occurrences of known interface types in protein interactions (Weatheritt et al, 2012; Mosca et al, 2013). However, it is reasonable to expect that many more protein interface types remain to be discovered. This is likely particularly true for motif-mediated PPIs, which are anticipated to number in the hundreds of thousands or millions (Tompa et al, 2014). Motifs are short stretches of amino acids in disordered regions of proteins that usually adopt a more rigid structure upon binding to folded domains in interaction partners (Davey et al, 2012). Motif-mediated interactions are of moderate binding affinity and thus, are particularly suited to mediate dynamic cell regulatory and signaling events (Van Roey et al, 2012). However, due to the transient nature of their interactions and the disorderliness of motif-containing proteins, this mode of binding is also expected to be highly understudied. Systematically generated human protein interactome maps (Luck et al, 2020; Huttlin et al, 2021) are likely a treasure trove for the discovery of novel interface types, yet no good experimental or computational methods exist to systematically map or predict protein interaction interfaces at scale.

[1]Institute of Molecular Biology (IMB) gGmbH, 55128 Mainz, Germany. [2]Department of Microbiology and Molecular Genetics, Institute for Biomedical Research Israel-Canada, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 9112001, Israel. [3]Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg 69117, Germany. [4]Present address: Computational Biology and Data Mining Group Biozentrum I, 55128 Mainz, Germany. [5]These authors contributed equally: Chop Yan Lee, Dalmira Hubrich, Julia K Varga. ✉E-mail: ora.furman-schueler@mail.huji.ac.il; k.luck@imb-mainz.de

The release of the neural network-based software AlphaFold (AF) was not only a breakthrough for the prediction of monomeric structures of proteins (Jumper et al, 2021) but multiple studies published shortly thereafter also suggested the ability of AF to predict structures of pairwise protein interactions and complexes. Sensitivities of around 70% were reported using benchmark datasets of structurally resolved protein interactions originally developed to evaluate docking methods (Akdel et al, 2022; Bryant et al, 2022; Johansson-Åkhe et al, 2021; preprint:Evans et al, 2021). Other studies focused on structures of domain-motif interfaces to specifically evaluate AF's ability to predict structures for this mode of binding, reporting similar success rates (Akdel et al, 2022; Johansson-Åkhe et al, 2021; Tsaban et al, 2022). Only a few studies have also evaluated AF's specificity for the prediction of interface structures using controls such as random protein pairs or mutation of motifs to poly-alanine stretches (Akdel et al, 2022; Johansson-Åkhe et al, 2021; Tsaban et al, 2022). Different benchmarking studies used different versions of AF and reported on different metrics for their ability to distinguish good from bad structural models (Bryant et al, 2022; O'Reilly et al, 2023; Tsaban et al, 2022; preprint:Evans et al, 2021; Teufel et al, 2023). We generally lack a comprehensive assessment of the latest AF releases and metrics across different types of PPI interfaces for their sensitivity, specificity, and potential biases for the prediction of complex structures.

In a landmark study, researchers applied AF onto 65,000 human PPIs derived from a yeast two-hybrid-based interactome map (hereafter referred to as HuRI) and highly confident co-complex associations to structurally annotate the human interactome with AF-derived models. High confidence models were obtained for about 3000 PPIs (Burke et al, 2023). The authors noted a smaller fraction of highly confident structural models obtained for PPIs from the HuRI dataset compared to the co-complex dataset and reported that proteins in HuRI contain more intrinsic disorder and are less conserved compared to proteins from co-complex datasets. AF model confidence scores also increased for PPIs with proteins that are less disordered and more conserved, indicating that AF predictions work less well for PPIs mediated by interfaces involving disordered regions such as domain-motif interfaces, which likely dominate the human interactome (Tompa et al, 2014). However, AF benchmarking studies reported similarly high success rates for domain-motif interfaces compared to general docking benchmark datasets (Tsaban et al, 2022; Akdel et al, 2022). These discrepancies in sensitivities could be a result of two possible factors. First, they might point to differences in AF performance if small interacting fragments are used for interface prediction, as done in the benchmark studies, versus full length sequences used for structure prediction in (Burke et al, 2023). Second, these discrepancies could also point to difficulties of AF to predict structures of interface types involving disordered regions that have not been solved before, of which there are likely many in HuRI. It remains to be addressed to what extent these two possible factors contribute to the challenges encountered specifically for domain-motif interface modeling.

Determination of accuracies of novel predicted interface structures by AF ultimately requires experimentation. AF interface predictions for individual PPIs have occasionally been experimentally corroborated (Mishra et al, 2023; Bronkhorst et al, 2023). A more systematic experimental confirmation of AF interface models has been conducted using crosslinking mass spectrometry (XL-MS) (Burke et al, 2023; O'Reilly et al, 2023). While in-cell XL-MS is a very elegant approach to obtain experimental information on PPI interfaces in unperturbed settings, it is still a method that is only accessible to few experts in the field. Other experimental approaches are needed, which can, ideally at high throughput, confirm predicted interfaces for PPIs. In this study, we thoroughly benchmarked the two most recent versions of AlphaFold-Multimer (hereafter referred to as AF) for their ability to predict domain-domain and domain-motif interfaces (DDIs and DMIs). We found that prediction accuracies drop when using longer protein fragments or full length proteins for interface predictions and developed a strategy particularly suited for the prediction of novel domain-motif interfaces in human PPIs. We applied this strategy to 62 PPIs from HuRI that connect disease-associated proteins and experimentally assessed the obtained interface predictions for seven PPIs using a plate-based bioluminescence resonance energy transfer (BRET) assay (Trepte et al, 2018) combined with site-directed mutagenesis. We identify novel interface types and report on important limitations and sources of errors in AF-derived structural models, which pave the way for future improvements in the field.

Results

Evaluating AlphaFold's accuracy for predicting domain-motif interfaces

To thoroughly assess the ability of AF to predict structures of binary protein complexes that are formed by a DMI, we extracted information on annotated DMI structures from the ELM DB (Kumar et al, 2022). We selected one representative structure per motif class (136 structures in total), manually defined the minimal domain and motif boundaries, and submitted the corresponding protein sequence fragments for interface prediction to AF (Fig. 1A; Dataset EV1). The domain sequences from this benchmark dataset mostly shared 20–30% sequence identity (Appendix Fig. S1A). To evaluate the accuracy of the predicted structural models, we superimposed the actual structure and predicted model on their domains and based on this superimposition, we computed the all atom RMSD between the motif of the predicted model and the actual structure (Fig. 1A). We found that 35% of the structural models were so accurately predicted that even the side chains of the motif were correctly positioned while for another 32% the backbone but not the side chains of the motif were accurately predicted. For 26% of the structures the motif was modeled into the correct pocket, but in a wrong conformation, while, for the remainder of the structures, AF failed to identify the right pocket (Fig. 1A; Dataset EV1). A similar performance was obtained when using the DockQ metric (Appendix Fig. S1B,C; Dataset EV1). This performance is unaltered when using or switching off AF's template function (Fig. S1D,E). The use of DMI structures annotated by the ELM DB enables us to explore potential differences in AF's performance regarding motif properties. We find no significant differences in average model accuracy between different categories of motif classes (two-sided Mann–Whitney test on all pairwise combinations, n: DEG = 10, DOC = 21, LIG = 94, TRG = 9, MOD = 2, $\alpha = 0.05$, test statistics of all pairwise combinations between 15 and 852, Appendix Fig. S1F), although the variance in model accuracy appears to differ between the motif classes. Similarly, we found no significant difference in prediction accuracy when

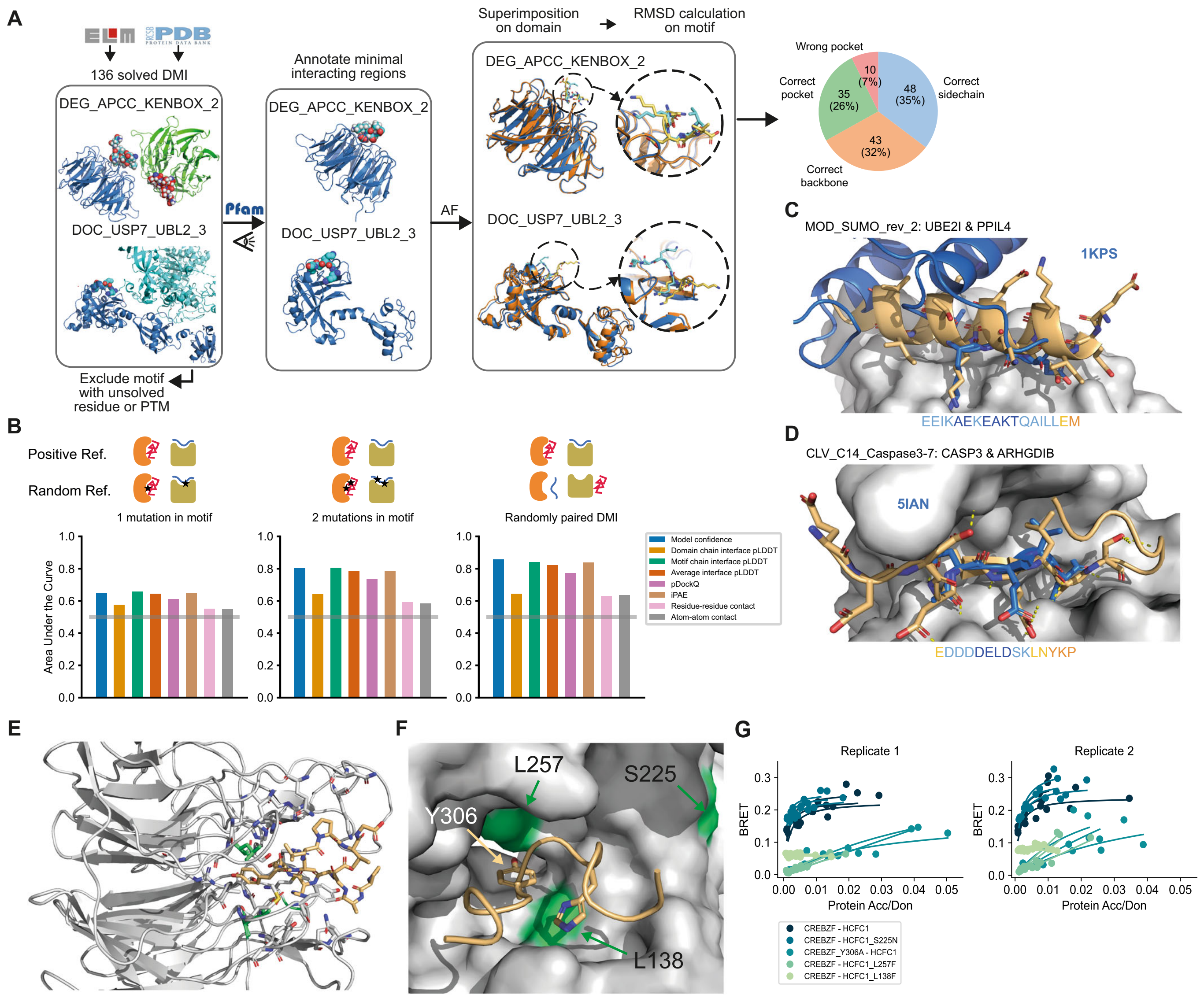


Figure 1. Benchmarking and application of AF for DMI interface prediction using minimal interacting fragments.

(**A**) Schematic illustrating the assembly of the DMI positive reference dataset and evaluation of AF prediction accuracies by superimposition of the solved and modeled structures. Blue and cyan indicate the domain and motif in the native structure, respectively. Orange and yellow indicate the domain and motif in the modeled structure, respectively. Proportion of structures of DMIs predicted by AF to different levels of accuracy is shown on the right. (**B**) Area under the Receiver Operating Characteristics Curve (AUROC) for different metrics using the DMI benchmark dataset as positive reference and the following different random reference sets: Left, 1 mutation introduced in conserved motif position; middle, 2 mutations introduced in conserved motif positions; right, random reshuffling of domain-motif pairs. Gray horizontal line indicates the AUROC of a random predictor. (**C**) Superimposition of AF structural model for motif class MOD_SUMO_rev_2 (orange) with homologous solved structure (PDB:1KPS) from motif class MOD_SUMO_for_1 (blue). The motif sequence used for prediction is indicated at the bottom, colored by pLDDT (dark blue=highest pLDDT). (**D**) Superimposition of AF structural model for motif class CLV_C14_Caspase3-7 (orange) with homologous structure (PDB:5IAN) solved with a peptide-like inhibitor (blue). The motif sequence used for prediction is indicated at the bottom, colored by pLDDT (dark blue=highest pLDDT). (**E**) AF prediction of a LIG_HCF-1_HBM_1 motif in CREBZF (orange) binding to the beta-propeller Kelch domain of HCFC1 (gray). Mutated domain residues for experimental testing are colored in green. (**F**) Close up on the interface shown between CREBZF and HCFC1 from (**E**). Coloring is the same as in (**E**). Key conserved motif residues are drawn as sticks. Mutated residues in the domain and motif for experimental testing are labeled. (**G**) BRET titration curves are shown for wildtype interactions and mutant constructs for CREBZF-HCFC1 pairs for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the *x*-axis determined from fluorescence and luminescence measurements, respectively.

stratifying by the secondary structure elements adopted by the motifs (two-sided Mann–Whitney test on all pairwise combinations, *n*: helix = 42, strand = 7, loop = 87, $\alpha = 0.05$, test statistics of all pairwise combinations between 184 and 2029, Appendix Fig. S1G), nor by how hydrophobic, symmetric, or degenerate the motif sequence is (Pearson r < abs(0.08), $\alpha = 0.05$ Appendix Fig. S1H–J). AF models display significantly more differences to structures solved by other methods, i.e., NMR, than X-ray crystallography (two-sided Mann–Whitney test, *n*: X-ray = 115, Others = 21, $p < 0.01$, test statistics = 811, Appendix Fig. S1K) possibly because

NMR structures better represent structural dynamics that AF cannot capture, since it was trained to predict the crystallized forms of proteins.

The all-atom motif RMSD significantly anti-correlates with various AF-derived metrics (Pearson $r = -0.55$, p-value < 0.05 Appendix Fig. S1L,M; Dataset EV1) suggesting that these metrics are indicative of good versus bad structural models and can be used for de novo interface predictions. To evaluate AF's ability to identify high confident structural models of DMIs, we generated three different random DMI datasets. First, we randomly paired domain and motif sequences from the positive reference dataset taking into account that no motif sequence was paired with a domain sequence from the domain type that the motif is known to interact with. Second and third, we mutated one and two key motif residues, respectively, to residues of opposite chemico-physical properties. Based on the conservation of these key motif residues, we assume that the mutations would be disruptive to binding, at least when experimentally tested using minimal interacting protein fragments. Receiver operating characteristic (ROC) and precision-recall (PR) curves using the positive and random datasets (Fig. 1B; Appendix Fig. S2A,B; Dataset EV2) show that the domain interface residue pLDDT (for all metric definitions, see Methods) or the number of atoms or residues predicted to be in contact with each other, discriminated poorly between all reference datasets (AUC around 0.64). Furthermore, we observed that all tested metrics failed to discriminate interacting from non-interacting interfaces when mutating one motif residue (max AUC 0.66). However, the AF-derived metrics model confidence (preprint:Evans et al, 2021), average interface residue pLDDT, average motif interface residue pLDDT, pDockQ (Bryant et al, 2022), and iPAE (Teufel et al, 2023) discriminated well between both reference datasets when randomizing domain-motif pairs or introducing two motif mutations (max AUC 0.86, ROC statistics and ideal cutoffs can be found in Dataset EV2). We also evaluated whether the top 5 reported models by AF tend to be more similar to each other when corresponding to a correct structural model (Pozzati et al, 2022) and found that this feature has moderate predictive power (Appendix Fig. S2C).

Application of AlphaFold for providing structural models for motif classes without available structural data

After evaluating the accuracy of AF to predict DMIs using minimal interacting regions, we aimed to use this setup for the prediction of structural models for motif classes in the ELM DB for which no structure of a complex has been solved yet. We identified 125 such motif classes based on ELM DB annotations. Of those, we selected all domain-motif instances where both the motif and the domain were derived from human or mouse proteins and submitted the corresponding domain and motif sequences for structure prediction to AF (Dataset EV3). Using a motif chain pLDDT cutoff of > 70, we obtained confident structural models for 21 motif classes. We manually inspected the structural models and noticed that even though these ELM classes have no annotations with structures, solved structures for an exact ELM instance or a very likely new instance for the ELM class are available for 11 out of the 21 cases. For most others, a close homolog structure had been solved, i.e., for LIG_MYND_3 and LIG_MYND_1, a structure solved by NMR for a LIG_MYND_2 interaction is available (Appendix Fig. S2D,E). For MOD_SUMO_rev_2, a structure of a reversed motif is available (and annotated as such in the MOD_SUMO_for_1 class). Here it is interesting to see how very dissimilar binding modes (flexible for MOD_SUMO_for_1, helical for MOD_SUMO_rev_2), are still able to place the important binding residues in the same pockets (Fig. 1C). For CLV_C14_Caspase3-7, the structure of the caspase bound to peptide-like inhibitors has been solved (e.g. PDB:1F1J, PDB:5IAN, PDB:6KMZ), and structures of more distant caspases bound to a cleaved peptide substrate are also available. For proteases, one great advantage of AF is the ability to model both the catalytically active enzyme and an uncleaved substrate, which is practically impossible to solve experimentally (Fig. 1D).

Finally, for LIG_HCF-1_HBM_1 we were not able to identify a homologous structure in the PDB, hence, our AF-derived structural models for this motif class are likely novel. Motifs of this class are bound by the N-terminal beta-propeller Kelch domain of HCFC1 consisting of six Kelch repeats. Kelch domains have been shown to bind to motifs at a number of different sites, and thus, without prior knowledge, it is difficult to determine where the HCFC1-binding motif (HBM) would bind. HCFC1 is a transcription factor that associates with other transcription factors (Lu et al, 1997), splice factors (Ajuh et al, 2002), and cell cycle regulators (Freiman and Herr, 1997; Machida et al, 2009). We generated AF models of high confidence for the HCFC1 Kelch domain interacting with multiple motif instances that are annotated in the ELM DB. All complexes show the tyrosine of the motif docked into a deep pocket at the bottom/top of the Kelch domain (Fig. 1E,F; Appendix Fig. S2F–H), with slight variations in how the tyrosine is exactly positioned in the pocket (Fig. S2F–H). Based on clone availability we selected the structural model between HCFC1 and CREBZF for experimental validation. For this purpose, we used a BRET protein interaction assay that is based on transient overexpression of two proteins in HEK293 cells (Trepte et al, 2018). Both proteins are expressed as fusion constructs either to the Nanoluc luciferase (the donor) or mCitrine (the acceptor). Interaction of both proteins results in a BRET from the oxidized substrate of the donor to the acceptor molecule, if both are close enough to each other for the BRET to occur (see Methods for details). We observed significant binding and BRET saturation when assaying wildtype CREBZF and HCFC1 proteins (Fig. 1G; Appendix Fig. S2I,J). Mutation of the [DE]H.Y motif tyrosine to alanine (Y306A) or mutation of two residues in the Kelch domain pocket (L257F, L138F), which are modeled to be in contact with the motif tyrosine or histidine residue (Fig. 1F), strongly reduced BRET signals indicating weakening or loss of binding (Fig. 1G; Appendix Fig. S2I,J). A pathogenic mutation (S225N, source ClinVar (Henrie et al, 2018)) close to the pocket slightly reduced expression levels of HCFC1 but did not result in loss of binding (Fig. 1F,G; Appendix Fig. S2I,J). Our experiments suggest that a potential pathogenic mechanism of this mutation is not mediated via perturbed binding of partners to the Kelch repeat domain pocket of HCFC1 that we identified in this study. Unfortunately, no assertion criteria for the annotation of this mutation to be pathogenic is provided by ClinVar meaning that the mutation is either not pathogenic after all or its pathogenicity is mediated via another perturbed function not tested in this study. Collectively, these experimental results support the structural models of the HCFC1 Kelch domain pocket - motif interaction and overall provide highly confident structural models for multiple motif classes of the ELM DB without available structural information (Dataset EV4).

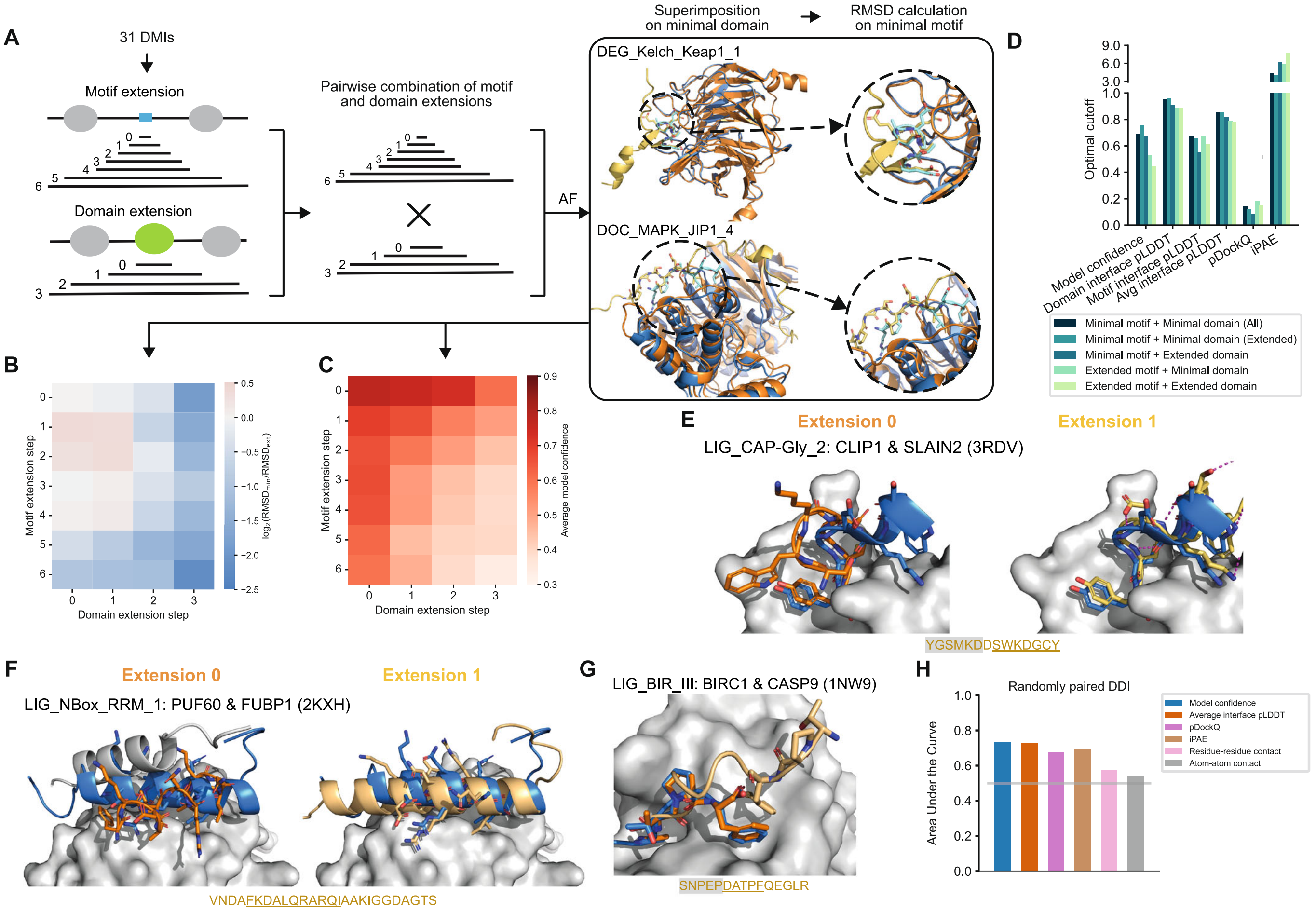


Figure 2. Effect of protein fragment extensions on the accuracy of AF predictions.

(**A**) Workflow established to assess changes in AF performance upon protein fragment extension. Blue and cyan indicate the domain and motif in the native structure, respectively. Orange and yellow indicate the domain and motif in the modeled structure, respectively. (**B**) Heatmap showing the fold change in motif RMSD before and after extension where positive values indicate improved predictions from extension and negative values indicate worse prediction outcomes upon extension. (**C**) Heatmap of the average model confidence for combinations of different motif and domain sequence extensions. (**D**) Optimal cutoffs derived for different metrics from ROC analysis benchmarking AF different motif and domain extensions from the reference dataset used in A and random pairings of domain and motif sequences. pLDDT-related metrics were divided by 100 for visualization purposes. (**E**, **F**) Superimposition of the structural model of the minimal (left, orange) or extended (right, yellow) motif sequence with the solved structure (motif in blue) for two different motif classes as indicated on the top of each panel. The motif sequence from the solved structure is indicated at the bottom. Motif residues are underlined, motif residues not resolved in the structure have a gray background. Sticks indicate the motif residues, domain surfaces are shown in gray based on experimental structures. (**G**) Superimposition of the structural model of the minimal (orange) and extended (yellow) motif sequence with the solved structure (motif in blue) for a motif instance from the motif class LIG_BIR_III. Motif sequence indicated as in (**E**). (**H**) Area under the Receiver Operating Characteristics Curve (AUROC) for different metrics using the DDI benchmark dataset as positive reference and randomly shuffled domain-domain pairs as random reference. Gray horizontal line indicates the AUROC of a random predictor.

Evaluation of AlphaFold's ability to predict interfaces in full length proteins

Most PPIs known to date have been identified using full length protein sequences in systematic interactome mapping efforts. For the vast majority of these PPIs, no fragment or interface information is available. Thus, the question emerges how AF would perform on DMI predictions when longer protein sequences or full length proteins are submitted. To answer this question we selected 31 DMI structures from the positive reference dataset used above and generated random domain-motif pairs of those as negative control. The selected structures were sampled from different prediction accuracy categories (Fig. 1A; Dataset EV5). We then gradually extended the motif and domain sequences by first adding flanking disordered regions, then neighboring folded domains before using the full length sequences (Fig. 2A). Comparison of the motif RMSD computed for extended versus minimal domain-motif pairs from the positive reference dataset revealed that the addition of flanking disordered regions on the motif or domain side sometimes slightly improved prediction accuracies while the addition of neighboring structured domains or the use of full length sequences led to a significant worsening of model accuracies (Fig. 2B; Dataset EV5). Interestingly, despite the fact that, for smaller extensions, model accuracies remained the same or slightly improved as determined by motif RMSD, AF-derived metrics such as the model confidence or average motif interface residue pLDDT gradually dropped with increasing fragment length (Fig. 2C; Appendix Fig. S3A-C). ROC plots of predictions for a benchmark consisting of the positive and random domain-motif pairs revealed that, upon extension, the optimal cutoff of model confidence and iPAE considerably changed as well (Fig. 2D; Appendix Figs. S3D,E, S4A; Dataset EV6). This means that different model confidence or iPAE cutoffs are to be used depending on the length of the submitted protein sequences, which is rather impractical and thus disfavors both metrics for DMI predictions. The average motif interface residue pLDDT metric appeared to be more robust with respect to fragment length. Based on these results we chose this as the main metric and a cutoff of 70 to discriminate good from bad AF-generated DMI models regardless of fragment length.

Extending motif sequences for interface prediction with AlphaFold reveals important motif sequence context

Various studies have highlighted that flanking sequences of motifs can influence binding affinities and specificities (Luck et al, 2012; Bugge et al, 2020). Motif annotations in the ELM DB usually refer to the core sequence of the motif, often because information on putative roles of flanking sequences is missing. In the previous section, we observed that some motif extensions notably improved AF prediction accuracies. In the hope that these cases would point to motifs with important sequence context, we manually inspected eight predictions for which the motif RMSD decreased by more than 1 Å when extending the minimal motif sequence once to the left and right by the length of the motif (extension step 1 in Fig. 2A; Appendix Fig. S4B).

By doing so interesting patterns emerged: The most prevalent contribution to increased prediction accuracies is the stabilization of the secondary structure of the motif contributed by both sidechain and backbone atoms in the flanking regions, as shown for the interaction involving the motif LIG_CAP-Gly_2 (Fig. 2E; Appendix Fig. S4C). For the LIG_NBox_RRM_1 motif, AF placed a part of the domain into the binding pocket rather than the motif, although the motif had the correct helical conformation. Elongation of the motif extended this helix, thereby increasing the interaction surface and eventually pushing out the domain's tail from the pocket (Fig. 2F). This fits with other reports where AF has been shown to predict preferential binding of competing motifs (Chang and Perez, 2023). For the LIG_HOMEOBOX class prediction, the motif is positioned in the wrong pocket unless flanking regions are included (Appendix Fig. S4C). For DOC_MAPK_JIP1_4, motif extension results in an extended motif conformation and consequently in a structural model with lower overall RMSD (Appendix Fig. S4C). For the LIG_GYF class, most models converge into an inverse orientation of the backbone except for one of the extended motifs, which lies in the binding pocket in the correct orientation (Appendix Fig. S4C). In summary, these analyses point to motif classes whose sequence boundaries could be refined.

Interestingly, for a motif instance from the LIG_BIR_III_2 class, slight motif extensions actually led to a substantial decrease in prediction accuracy. In this case, the motif is located at a neo-N-terminus that is only revealed after cleavage of the protein by a caspase (Fig. 2G). When the motif is extended in the context of the full length protein, the residues now upstream of the previous neo-N-terminus likely impede binding of the motif into the pocket due to steric clashes. AF predicts the extended motif to bind in reversed orientation and it is mostly pushed out of the pocket. This highlights the importance of not only incorporating sequence context but also knowledge about the biological context, wherever possible, into AF modeling and model interpretation.

Evaluating AlphaFold's performance for the prediction of domain-domain interfaces

Folded domains can not only interact with motifs but also with other folded domains forming so-called domain-domain interfaces (DDIs). To enable simultaneous prediction of DDIs and DMIs in a given protein interaction, we set out to evaluate AlphaFold's performance on DDI predictions using a reference dataset of 48 DDI structures that we manually curated out of random selections of domain-domain contact pairs extracted from 3did (Mosca et al, 2014). As a negative dataset, we randomized the pairing of these domains. Using ROC and PR statistics we found that AlphaFold performed slightly worse on this DDI benchmark dataset compared to its performance on DMIs (max AUC 0.73 vs. 0.86) (Fig. 2H; Appendix Fig. S4D–F; Dataset EV7) but still showed significant discriminative power. Interestingly, the best performing metric for DDI predictions was the average interface pLDDT score with an optimal cutoff of 75, which ranked fourth for DMI predictions.

Comparison of AlphaFold v2.2 with v2.3

During the course of our work, AF multimer version 2.3 was released. To determine whether the new release improved DMI and DDI prediction accuracies, we repeated all benchmarking with AF v2.3 and found that motif RMSDs and other AF-derived metrics on average improved compared to AF v2.2 when using minimal interacting fragments (Appendix Fig. S5A–D; Dataset EV1, two-sided Wilcoxon signed-rank test on motif all atom RMSD: $n = 136$, $W = 2413$, $p < 0.0001$). AF v2.3 still showed a decrease in prediction accuracy when using extended protein fragments but this decrease was less pronounced compared to the corresponding decrease for v2.2 (Appendix Fig. S5E,F; Dataset EV5). Despite these improvements on the sensitivity side of AF, when benchmarked against random datasets, overall prediction accuracies only slightly improved compared to v2.2 (Appendix Fig. S5G,H; Appendix Fig. S6A–C; Dataset EV2, EV6, EV7, EV8).

Application of AlphaFold for the discovery of novel interfaces in protein interactions without any a priori interface information

Since the use of larger or full length protein sequences leads to a poor sensitivity for DMI predictions by AF, we devised the following strategy for the use of AF for interface predictions for known protein interactions: Using AF models of the full length monomeric structures of both interacting proteins, we decided on boundaries between structured domains and disordered regions based on manual inspection (see Methods). We then fragmented the disordered regions by designing overlapping fragments varying in length from ten residues up to the length of the respective disordered region (Fig. 3A). We then paired disordered with ordered, and ordered with ordered fragments for interface prediction by AF (Fig. 3A). To assess to which extent this

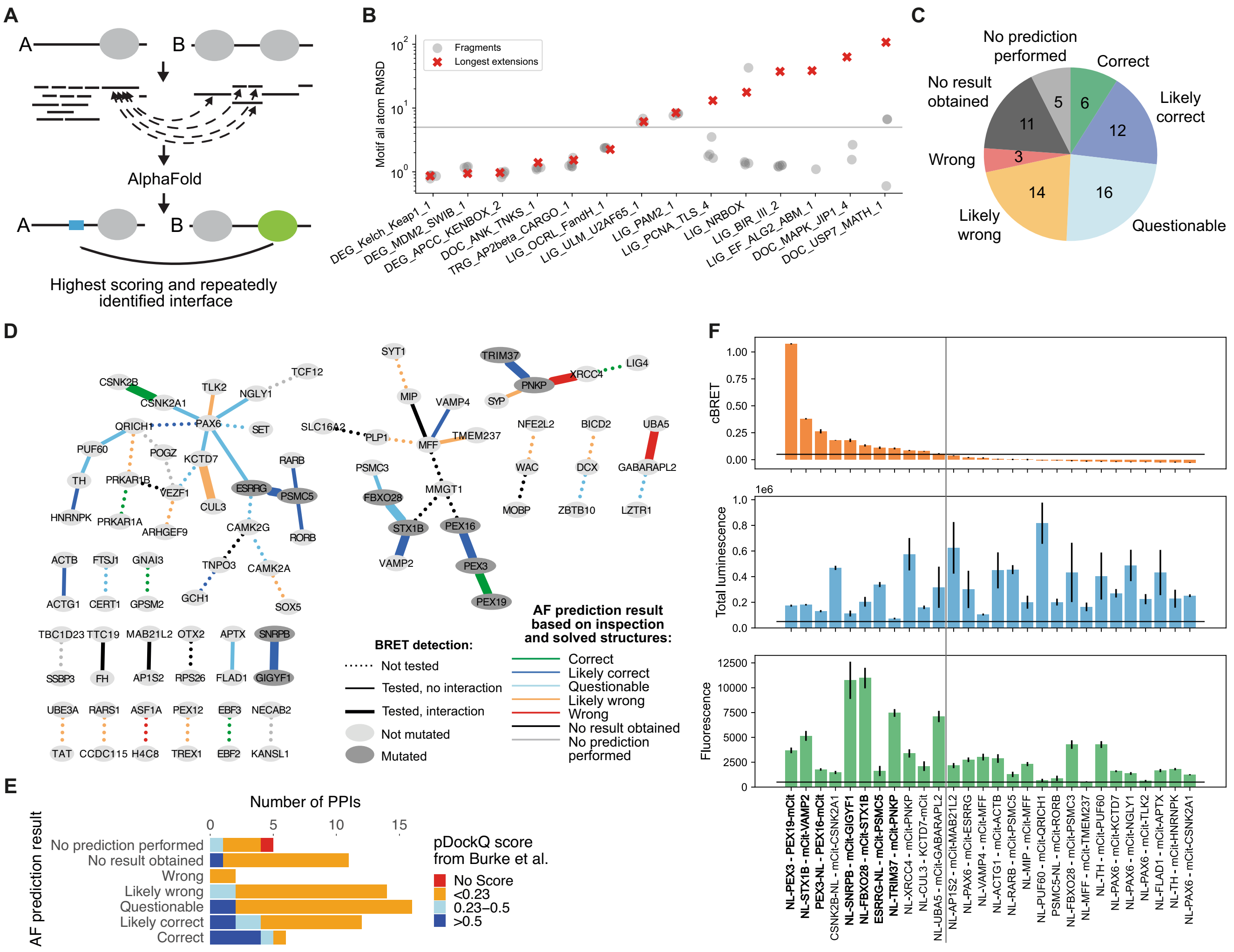


Figure 3. AF prediction and experiments on PPIs connecting NDD proteins.

(**A**) Schematic of the fragmentation approach applied on a pair of interacting proteins, A and B. Proteins are fragmented into folded and disordered regions based on manual inspection. Disordered regions are further fragmented. All disordered and folded fragments of one protein are paired with the folded regions of the other protein and vice versa for AF prediction. (**B**) Accuracy measured in motif RMSD compared to native structures for models obtained from fragmenting proteins from 20 DMIs from the positive reference dataset and comparison to model accuracy obtained when using (near) full length proteins for structure prediction (red crosses). Only models that meet the cutoff for identifying high confident models are shown. Six DMIs did not result in any such model. The gray horizontal line indicates the RMSD cutoff used to identify accurate models (see methods for details). (**C**) AF prediction outcome on 67 HuRI PPIs connecting NDD proteins. (**D**) PPI networks illustrating AF prediction outcomes and experimental retesting of PPIs in BRET assay. (**E**) Number of PPIs connecting NDD proteins with structural models at indicated pDockQ cutoffs from (Burke et al, 2023) grouped based on AF prediction outcomes using the fragmentation approach as shown in (**C**). (**F**) cBRET, total luminescence, and fluorescence for 28 PPIs connecting NDD proteins that were tested in the BRET assay. Luminescence and fluorescence measurements indicate expression levels of NL and mCit fusion proteins, respectively. Black horizontal lines indicate expression level and PPI detection cutoffs. The gray vertical line separates the detected (left) from undetected PPIs. Protein pairs in bold indicate those selected for interface validation via site-directed mutagenesis. Error bars indicate STD of three technical replicates. Source data are available online for this figure.

fragmentation approach would lead to an increase in sensitivity but also in false model predictions, we selected 20 out of the 31 DMI structures that were previously used to investigate the effect of fragment extension on prediction accuracies. We attempted model prediction with the full length sequences of these 20 DMI pairs and obtained a model for two of which only one met the motif interface pLDDT cutoff and corresponded to an accurate prediction (TRG_AP2beta_CARGO_1 in Fig. 3B; Dataset EV9, see methods for details). We then switched to using fragment extension step 5 for motifs and/or 2 for domains (Fig. 2A) and obtained accurate models for an additional 5 of the 20 DMI pairs. Applying the full fragmentation approach onto all 20 DMI pairs resulted in accurate model prediction for an additional 6 DMI pairs (Fig. 3B) representing an increase in sensitivity for full length vs fragments from 5 to 60%. We then shuffled the 20 DMI pairs to generate 20 random DMI pairs for which we performed the fragmentation approach. As expected from an earlier estimated 20% false positive rate (FPR) (Appendix Fig. S4A), 19 of the 20 random protein pairs had at least one fragment pair that produced a model above the motif interface pLDDT cutoff (Appendix Fig. S6D; Dataset EV9) indicating that predictions done using this fragmentation approach can substantially increase sensitivity while also producing a considerable number of false models using the established scoring metrics. This needs to be taken into account when modeling new interactions with this fragmentation strategy, as covered in the following section.

We selected PPIs from HuRI that connect proteins associated with neurodevelopmental disorders (NDDs) and subjected these to our AF fragmentation pipeline to predict putative DMIs and DDIs. For 51 out of 62 PPIs we obtained at least one structural model of significant confidence (Fig. 3C,D). In retrospect, manual inspection of the predictions obtained for these PPIs revealed that, for 9 PPIs, a solved structure of the interface was already available. Reassuringly, six out of these were accurately predicted by AF. For the remainder of the PPIs, 12, 16, and 14 resulted in a likely correct, questionable, or likely wrong prediction, respectively, based on manual inspection of the models (Fig. 3C,D; Dataset EV10). Likely wrong predictions were scored as such based on docking of the protein partner into nucleic acid or metal ion binding or catalytically active sites. We also considered structural models as likely wrong, if different protein fragments of the partner were predicted with similarly high scores to bind to the same pocket on the domain. More detailed information can be found in Methods and Appendix Text S1. Of note, for 8 of the 12 PPIs with a likely correct prediction, AF predictions performed using the full length proteins (Burke et al, 2023) did not result in a high confidence prediction (Fig. 3E). 28 of the 62 PPIs were in our hands amenable to experimental testing using the BRET assay introduced earlier (see Methods for details). Significant BRET signals were observed for 11 of these 28 PPIs (Fig. 3F). Of those, 7 PPIs were selected for validating the predicted interfaces (Fig. 3D,F). The remaining four PPIs were not further considered because for three of them a structure already exists (CSNK2B-CSNK2A1, PNKP-XRCC4, UBA5-GABRAPL2) and for the fourth interaction (KCTD7-CUL3) we classified the predicted interface as likely wrong. Next, we will first describe failures in validating predicted interfaces followed by the successes.

For the interaction between PNKP and TRIM37, we obtained high confident structural models involving two different interfaces. AF predicted the PNKP FHA domain to bind to several disordered stretches in TRIM37 (Fig. 4A) that are overall negatively charged. These short regions were predicted to bind to a pocket on the FHA domain that is known to bind phosphorylated threonines (Durocher et al, 2000), which led us to conclude that these predictions were likely wrong. AF also predicted the MATH domain of TRIM37 to bind to two separate disordered putative motifs located between the FHA domain and phosphatase domain in PNKP (Fig. 4A–C). However, none of the mutants aimed at disrupting the predicted interfaces (Fig. 4B) involving the MATH domain showed a decrease in BRET signal compared to wildtype (Fig. 4D; Appendix Fig. S7A) indicating that TRIM37 and PNKP do not interact with each other via this interface.

AF predicted with high confidence binding of PSMC5 to the hormone receptor domain of ESRRG via two distinct motifs (Fig. 4E–G) with similarity to LxxLL motifs known to bind this type of domain (LIG_NRBOX in ELM DB). We reproducibly found that none of the motif mutations in PSMC5 decreased binding to ESRRG compared to wildtype while both domain pocket mutations led to a remarkable reduction in BRET signal (Fig. 4H; Appendix Fig. S7B,C) indicating that PSMC5 might bind to ESRRG via this pocket but not with the predicted motifs.

AF predicted a coiled-coil interface between STX1B and VAMP2 of moderate confidence (Fig. 5A,B). STX1B is a close homolog to STX1A, which binds in a 4-helix bundle to VAMP2 together with SNAP25 in a 1:1:2 stoichiometry, respectively, as observed by crystallography (PDB:1N7S (Ernst and Brunger, 2003)). This structure together with our predictions suggest that STX1B might bind VAMP2 in a similar way. Indeed, removal of the single helical SNARE domain in STX1B led to complete loss of binding to VAMP2 (Fig. 5C; Appendix Fig. S8A,B). Interestingly, FBXO28 was predicted by AF to bind to STX1B via a similar coiled-coil interface involving an extended helix in FBXO28 and the SNARE domain in STX1B (Fig. 5A,D). Here, deletion of the SNARE domain in STX1B or of the extended helix in FBXO28 reproducibly reduced, but did not abolish the interaction between STX1B and FBXO28 (Fig. 5E; Appendix Fig. S8C,D). We identified three pathogenic or likely pathogenic mutations in the SNARE domain of STX1B in ClinVar of which V216E and G226R are associated with generalized epilepsy with febrile seizures plus, type 9. Testing all three mutations in the BRET assay we observed a drastic decrease in binding for STX1B V216E to FBXO28 (Fig. 5F; Appendix Fig. S8C,D). However, the measured effects of the mutations on the FBXO28-STX1B interaction do not correlate with their location at the predicted interface. V216E, for example, is not predicted to be in contact with residues of FBXO28 (Fig. 5D). This indicates that the actual predicted orientation of the two extended helices with respect to each other is likely incorrect.

The fact that the deletion of the extended helix in FBXO28 or the SNARE domain in STX1B reduced but did not abrogate binding of both proteins to each other (Fig. 5E) suggests that a secondary interface might exist. Indeed, AF predicted additional interfaces between FBXO28 and STX1B involving folded and disordered regions in both proteins (interfaces i and ii in Fig. 5A). Mutations designed to disrupt these interfaces partially confirmed the involvement of some of these regions in binding as assayed with BRET (Appendix Fig. S8E–H). In addition, the pathogenic mutation R348L in FBXO28 predicted to be at interface ii seemed to increase binding to STX1B (Appendix Fig. S8I–L). In summary, our experimental data indicate that multiple regions of FBXO28 and STX1B may be involved in the binding but the exact structural details of this interaction remain to be elucidated. In the following two sections, we will describe in more detail successful interface validations for interactions involving PEX3, PEX19, and PEX16 as well as SNRPB and GIGYF1.

PEX3, PEX19, and PEX16

The interaction interface between PEX19 and PEX3 has been structurally resolved before and consists of an interaction between an N-terminal motif in PEX19 that binds to the cytosolic alpha-helical domain of PEX3 (PDB:3MK4, (Schmidt et al, 2010)). Using corresponding protein fragments, AF predicted a structural model that is highly similar to the solved structure (Fig. 5G; Appendix Fig. S9A,B). We introduced mutations in the PEX19 motif and PEX3 pocket (Appendix Fig. S9A) and found that F29K in the motif weakened but clearly maintained BRET binding signals indicating the existence of a secondary binding site between both proteins (Fig. 5H; Appendix Fig. S9C,D). Indeed, AF predictions with other

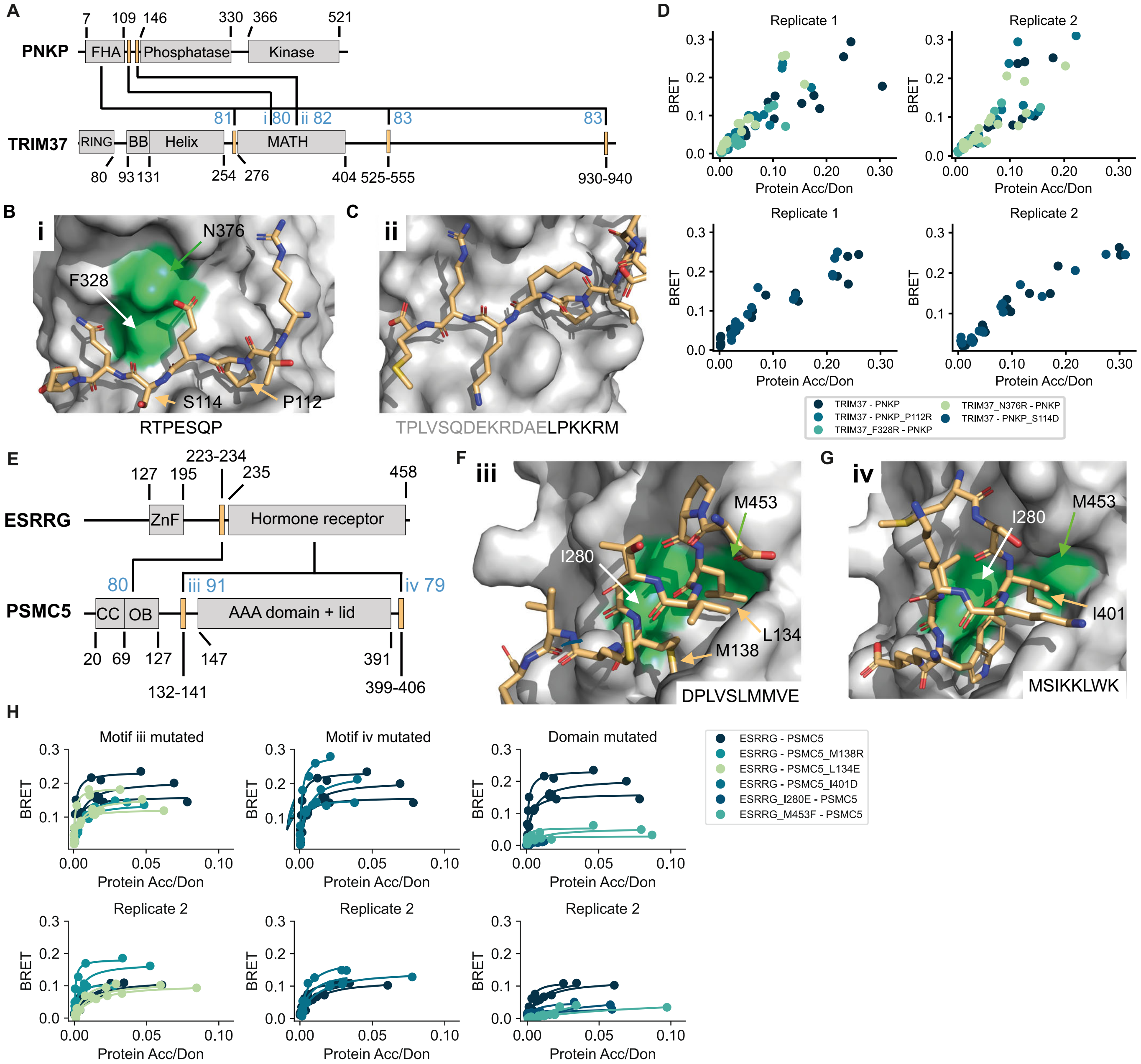
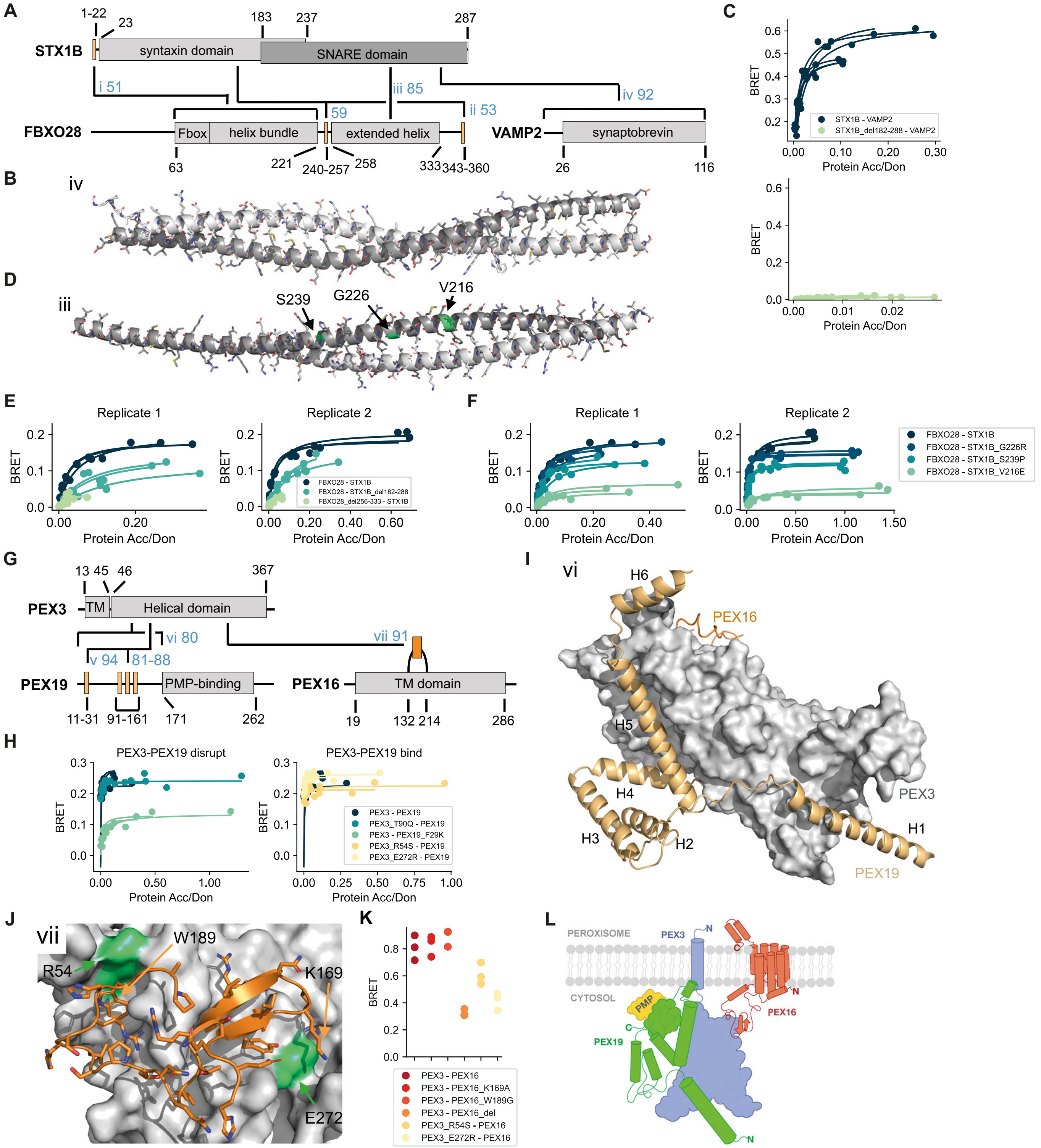


Figure 4. Verification of interface predictions for TRIM37-PNKP and ESRRG-PSMC5.

(**A**) Schematic of the domain architecture of PNKP and TRIM37 with indication of top predicted interfaces. Numbers in blue indicate the motif interface pLDDT for the respective interface. Roman numbering refers to structural models in (**B**) and (**C**). (**B**) Structural model of interface i shown in (**A**) with labeled residues that were mutated. (**C**) Structural model of interface ii shown in (**A**). (**D**) BRET titration curves are shown for wildtype interaction and mutants for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the *x*-axis determined from fluorescence and luminescence measurements, respectively. The BRET trajectory could not be fitted because of an unusual saturation behavior (see methods for details). (**E**) Schematic of the domain architecture of ESRRG and PSMC5 with indication of top predicted interfaces. Numbers in blue indicate the motif interface pLDDT for the respective interface. Roman numbering refers to structural models in (**F**) and (**G**). (**F**) Structural model of interface iii shown in (**E**) with labeled residues that were mutated. (**G**) Structural model of interface iv shown in (**E**). (**H**) BRET titration curves are shown for wildtype interaction and mutants of ESRRG-PSMC5 pairs for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the *x*-axis determined from fluorescence and luminescence measurements, respectively. In panels (**B**), (**C**), (**F**), and (**G**) motif sequences are indicated at the bottom. Gray letters indicate residues not predicted to bind. Source data are available online for this figure.



disordered fragments of PEX19 paired with the PEX3 domain resulted in highly confident models for interfaces involving a binding pocket on PEX3 that is distal to the pocket where the N-terminal PEX19 motif is known to bind. When using a protein fragment that spans the full disordered N-terminal region of PEX19 (1–170), AF predicts the known PEX3-binding motif and helix 4 and 5 to dock into the primary and secondary pocket, respectively (Fig. 5G,I), supporting simultaneous interaction via both interfaces.

While the interaction between PEX3 and PEX16 has been described before, little is known about how both proteins interact with each other. The monomeric AF model of PEX16 shows a helical fold, which could in its entirety be transmembrane (TM).

Figure 5. Verification of interface predictions for STX1B-FBXO28, STX1B-VAMP2, PEX3-PEX19, and PEX3-PEX16.

(**A**) Schematic of the domain architecture of STX1B, FBXO28, and VAMP2 with indication of top predicted interfaces. Numbers in blue indicate the motif interface pLDDT (for order-disorder fragment pairs) or average interface pLDDT (for ordered-ordered fragment pairs) for the respective interface. Roman numbering refers to structural models in (**B**), (**D**), Appendix Fig. S8E, and Appendix Fig. S8I. (**B**) Structural model of interface iv shown in (**A**). In panel (**B**) and (**D**), the chains are color-coded according to the colors of the domains in (**A**). (**C**) BRET titration curves are shown for wildtype interactions and deletion constructs for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the x-axis determined from fluorescence and luminescence measurements, respectively. (**D**) Structural model of interface iii shown in (**A**) with tested pathogenic mutations labeled and colored in green. (**E, F**) BRET titration curves are shown for wildtype interactions and deletion constructs for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the x-axis determined from fluorescence and luminescence measurements, respectively. (**G**) Schematic of the domain architecture of PEX3, PEX19, and PEX16 with indication of top predicted interfaces. Numbers in blue indicate the motif interface pLDDT for the respective interface. Roman numbering refers to structural models in (**I**), (**J**), and Appendix Fig. S9A. Region vi covers residues 1–170, which includes the previously reported N-terminal motif as well as three putative motifs suggested by the AF models. (**H**) BRET titration curves are shown for wildtype interaction and mutants of PEX3-PEX19 pairs for three technical replicates. Protein acceptor over protein donor expression levels are plotted on the *x*-axis determined from fluorescence and luminescence measurements, respectively. The left plot displays mutants aimed at disrupting binding between PEX3-PEX19 while the right plot displays mutants aimed at disrupting the PEX3-PEX16 PPI why binding between PEX3-PEX19 should not be altered. (**I**) Superimposition of structural models of interface vi (PEX3-PEX19) and vii (PEX3-PEX16) on the PEX3 domain. Note that modeling smaller fragments of PEX19 generates alternative interactions with the binding sites. (**J**) Structural model of interface vii shown in (**G**). (**K**) BRET values with subtracted bleedthrough for PEX3-PEX16 wildtype and various mutated constructs. Three technical replicates are shown. (**L**) Proposed model for how the trimeric complex of PEX3, PEX19, and PEX16 might assemble at the peroxisomal membrane. Source data are available online for this figure.

Between the putative TM helix 4 and 5 there is a large loop (132–214), which was predicted by AF with very high confidence to bind to a third pocket on the PEX3 domain, opposite to both binding sites mentioned earlier for PEX19 (Fig. 5G,I,J). Of note, different fragments of this loop as well as the entire PEX16 were repeatedly predicted to bind in similar modes to PEX3, further increasing the confidence in this prediction. Encouraged by these results, we submitted all three full length PEX sequences for complex prediction to AF and obtained a model that supports simultaneous binding of PEX16 and PEX19 to PEX3 (Appendix Fig. S9E). We individually mutated two residues in the PEX16 loop, deleted the loop in its entirety (del162-192), and mutated two residues on PEX3 (highlighted in Fig. 5J). Unfortunately, higher expression levels of PEX16 seem to trigger degradation of PEX3 (Appendix Fig. S9F), which we did not observe for the same constructs when co-expressed with PEX19 (Appendix Fig. S9G). As a consequence, we could not obtain titration curves and BRET50 estimates but obtained reliable BRET signals for lower PEX3-PEX16 DNA transfection ratios showing that the deletion as well as both PEX3 mutants significantly decreased binding to PEX16 (Fig. 5K; Appendix Fig. S9H). Of note, these PEX3 mutants (R54S and E272R) did not alter binding to PEX19, showing that the overall structural integrity of PEX3 was not perturbed by these mutations (Fig. 5H; Appendix Fig. S9D).

PEX3 and PEX19 are peroxin proteins that regulate peroxisome homeostasis. PEX16 is believed to serve as an integral membrane-bound receptor for PEX3 (Matsuzaki and Fujiki, 2008) while PEX3 is thought to serve as a docking site for PEX19 (Fujiki et al, 2006). PEX19 in turn is a cytosolic carrier for peroxisomal membrane proteins to the peroxisome (Fujiki et al, 2006). Combining results from previously published functional studies with the structural and experimental results obtained in this study, a model for a trimeric complex between PEX3, PEX19, and PEX16 emerges (Fig. 5L) where PEX16 fully inserts into the peroxisome membrane via a fold that consists of seven helices (residues 19-286) with its N-terminal end being cytosolic and its C-terminal end protruding into the peroxisome. The extended loop between TM helix 4 and 5 reaches into the cytosol and docks onto PEX3, which is further anchored into the peroxisomal membrane via its N-terminal TM helix (residues 13–45). PEX19 docks onto PEX3, opposite to where PEX16 is bound, via two interaction surfaces—one corresponding to the known PEX3-binding motif in PEX19 and a second one corresponding to a novel motif (residues 99–146) docking at a hitherto unknown second binding site on PEX3 for PEX19. This model explains how PEX3 is anchored to the peroxisomal membrane via PEX16 and how PEX3 can bind very tightly PEX19, which can then deliver PMPs to the peroxisome. Mutations in any of the three PEX proteins are associated with severe developmental phenotypes referred to as peroxisome biogenesis disorders (Fujiki et al, 2022). The vast majority of the around 150 mutations annotated for the three proteins are uncharacterized (Henrie et al, 2018), dozens of which fall into the predicted interfaces. The structural models obtained from this work can inform future studies aimed at characterizing the effects of these mutations.

SNRPB and GIGYF1

AF predicted two different types of interfaces with high confidence for the interaction between SNRPB and GIGYF1. The first interface involves the LSM domain of SNRPB which was predicted to bind to various fragments in the long disordered regions of GIGYF1 (Fig. 6A). These regions do not display any common sequence pattern. The structure of SNRPB has been resolved as part of the Sm ring complex that binds small nuclear RNA (PDB:4WZJ, (Leung et al, 2011)) showing that the surface on the LSM domain predicted to bind to disordered fragments of GIGYF1, is actually engaged in binding LSM domains of other Sm proteins within the complex (Fig. 6B). We thus conclude that these predictions are likely wrong. The second type of interface predicted by AF involves the GYF domain in GIGYF1 and multiple short disordered fragments in the C-terminal region of SNRPB, which repeatedly carry the sequence PPPGM(R) (Fig. 6A,C). We designed various deletion constructs of SNRPB that would gradually remove more and more of the repeated proline-rich motif. We observed, using the BRET assay, that these deletion constructs gradually decreased binding to GIGYF1 (Fig. 6D; Appendix Fig. S10A,B). We also mutated the GYF domain pocket and found that W498E but not L508F would decrease binding to SNRPB (Fig. 6D,E; Appendix Fig. S10A–D). To further corroborate these findings we performed a co-immunoprecipitation experiment, where endogenous GIGYF1 interacted with HA-tagged full length SNRPB (Fig. 6F). This

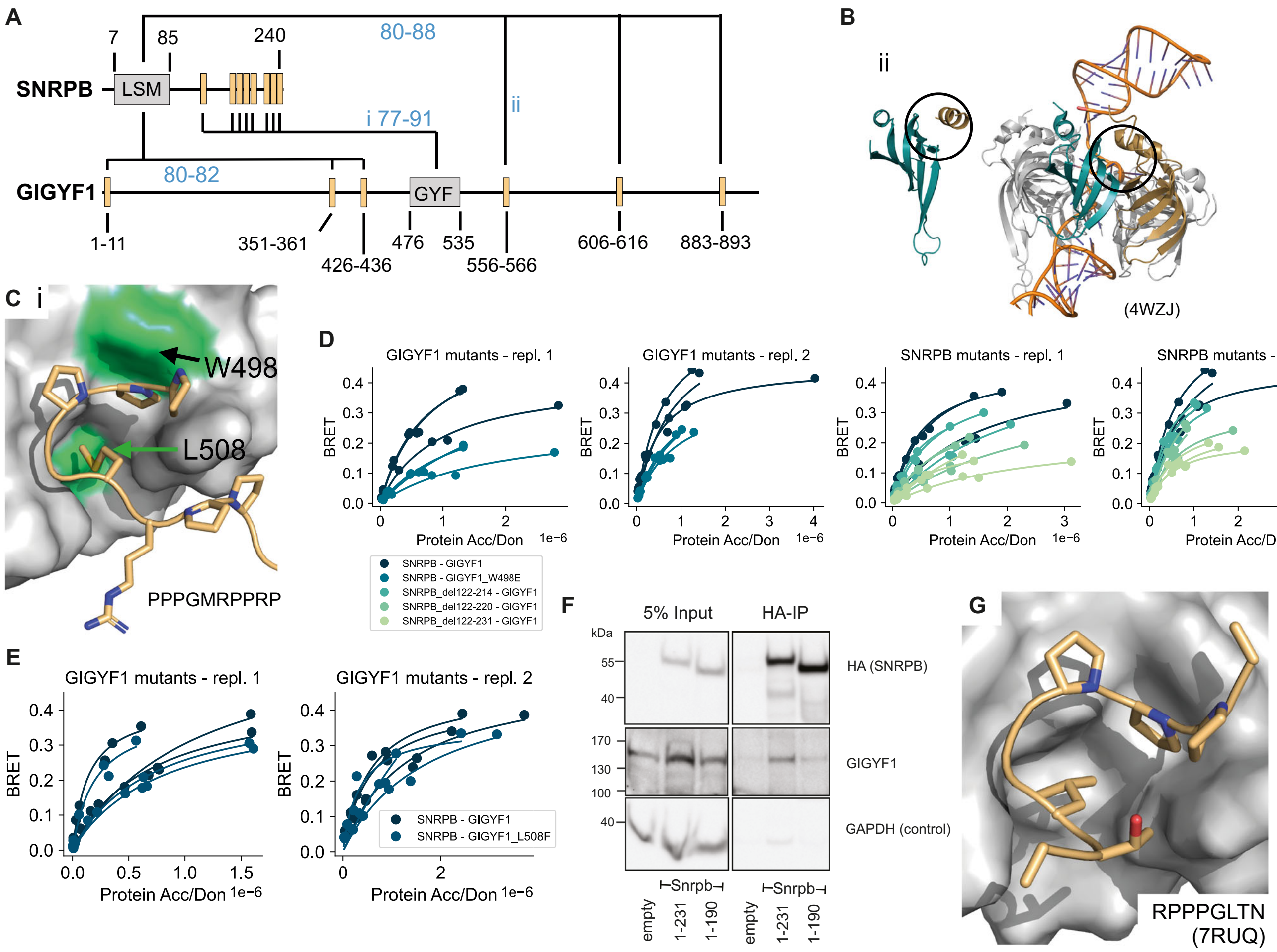


Figure 6. Verification of interface predictions for SNRPB-GIGYF1.

(**A**) Schematic of the domain architecture of SNRPB and GIGYF1 with indication of top predicted interfaces. Numbers in blue indicate the motif interface pLDDT for the respective interface. Roman numbering refers to structural models in (**B**) and (**C**). (**B**) Structural model of interface ii shown in (**A**) (left) and in comparison a solved structure (PDB:4WZJ) of the Sm ring complex (right) bound to RNA (orange). The LSM domain of SNRPB is shown in cyan. The position of the predicted motif (left) or neighboring LSM domain of SNRPD3 (right) are indicated in gold. Black circles indicate the predicted interface in the model and corresponding interface in the complex on the LSM domain of SNRPB. (**C**) Structural model of interface i shown in (**A**) with tested domain mutations labeled and colored green. The motif sequence is indicated at the bottom. (**D, E**) BRET titration curves are shown for wildtype interactions, deletion constructs of SNRPB, and single point mutants in GIGYF1 for two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the *x*-axis determined from fluorescence and luminescence measurements, respectively. (**F**) Cropped immunoblot of input (5%) and HA antibody immunoprecipitation (IP) performed in parental HEK cells (empty, untagged negative control), Snrpb(full-length, 1-231)-2xHA-mNeonGreen, Snrpb(1-190)-2xHA-mNeonGreen expressed from a single locus in Flp-In™ T-REx™ 293 Cell Lines. The HA antibody was used for detecting the immunoprecipitated Snrpb-proteins, endogenous GIGYF1 was detected with GIGYF1 antibody, GAPDH serves as a loading and negative-IP control. The experiment was performed twice with equivalent outcome, one representative experiment is shown. (**G**) Solved structure (PDB:7RUQ) of the GYF domain of GIGYF1 bound to a proline-rich motif in TNRC6C. The sequence of the motif in TNRC6C is indicated. Source data are available online for this figure.

interaction appeared less pronounced upon truncation of the C-terminal proline-containing region of SNRPB (Fig. 6F). This further suggests that both proteins interact with each other in cells and that this interaction is stabilized by the predicted interface.

During the course of these studies, a structure was published (PDB:7RUQ, Sobti et al, 2023) showing binding of the GYF domain of GIGYF1 to a motif of sequence PPPGL of the protein TNRC6C confirming the binding mode predicted by AF where a hydrophobic residue (M or L) inserts into a hydrophobic pocket and where the proline residues contact the surrounding domain surface (Fig. 6C,G). Interestingly, this hydrophobic pocket does not exist in the previously solved structure of the GYF domain of CDBP2 binding to a proline-rich peptide that is flanked by positively charged residues establishing important contacts with the domain (PDB:1L2Z, (Freund et al, 2002)). This structure formed the basis for the definition of the LIG_GYF motif class in the ELM DB. The recently resolved structure of the GYF domain of GIGYF1 together with our structural models and experimental validations argue for an extension of the existing motif definition or definition of a new motif subclass.

Discussion

AF has revolutionized the field of structural bioinformatics and has sparked much excitement about its potential to predict structures of interacting proteins and bringing us closer to a structurally resolved protein interactome. However, from existing studies it largely remained unclear whether AF's performance depends on the type of interfaces and the length of submitted protein chains for interface prediction, which metrics perform best in identifying likely correct structural models of interfaces, how specific AF predictions are, and to which extent highly confident structural models can be experimentally corroborated. In this study, we showed that AF performs similarly well for interfaces between folded domains and interfaces formed between a folded domain and a short linear motif. Using minimal interacting regions for interface prediction we reached sensitivities of up to 80% similar to previously published work (Tsaban et al, 2022; Johansson-Åkhe et al, 2021). We thoroughly investigated AF's FPR using random domain-motif pairs and found it to be around 20%. However, asking AF to discriminate binders from non-binders when motif sequences carried one disruptive mutation, we found that prediction accuracies were close to random. This points to an important limitation in AF's ability to predict binding specificities and is in line with previous reports on AF's inability to predict the effect of mutations (Buel and Walters, 2022). Comparison of different metrics to discriminate good from bad structural models using either minimal interacting fragments or extensions revealed the average interface pLDDT for DDI models and the motif interface pLDDT for DMI models to be the most robust and best performing metrics. However, when manually inspecting AF predictions we found it useful to also consider AF's model confidence, suggesting that in the future a combination of different metrics might be even more powerful to discriminate good from bad structural models. The alignment depth has been previously reported to somewhat influence model accuracy (Bryant et al, 2022). While this feature was not investigated here, it might serve as a pre-filter to identify PPIs of high conservation for which structural modeling will likely be more successful. Interestingly, the number of residues or atoms predicted to be in contact with each other was poorly predictive, in contrast to a previous report (Bryant et al, 2022), confirming our observations that the tested AF versions in this study will always put both chains in contact with each other to create atomic contacts, and from visual inspection alone it is very challenging to tell good from bad structural models apart. Of note, observed differences in AF performance across studies likely originate both from using different benchmark datasets and different AF versions. Our study is unique in that it assesses multiple metrics on two different classes of interfaces, DMIs and DDIs, using two different AF versions. More work is needed to develop benchmark datasets of coiled-coil and disorder-disorder interfaces to also evaluate AF's performance for these modes of binding. Of note, our benchmark datasets almost exclusively consisted of structures that AF has seen in the training process. Interestingly, benchmark studies done with unseen structures reported similar sensitivities (preprint:Bret et al, 2023) indicating that AF is not strongly biased towards structures it has seen before.

We extensively explored the influence of protein fragment length on AF's performance and found that slight extensions of minimal motif sequences can improve prediction accuracies. Inspection of individual cases revealed novel information on important motif sequence context that was so far missing in corresponding motif entries at the ELM DB. However, longer disordered fragments or fragments containing ordered and large disordered regions generally decrease AF prediction accuracies as also reported in a recent preprint (preprint:Bret et al, 2023). Furthermore, optimal cutoffs for various metrics such as the model confidence decreased when using longer protein fragments, making them less robust for interface prediction with AF. When evaluating performance differences for longer and shorter protein fragments we identified three DMI pairs involving the motif classes DEG_APCC_KENBOX_2, LIG_Pex14_3, and LIG_GYF, for which, during fragment extension, a second known motif occurrence was added to the fragment. This second motif was selected by AF during interface prediction, displacing the original motif and leading to a high RMSD score. We removed these instances from the dataset when evaluating AF's performance on fragment extension but they point to biologically correct variability in AF prediction outcomes due to existing multivalency of many DMIs in protein interactions. Other work suggested that AF is able to select the stronger binder among two motif occurrences (Chang and Perez, 2023), which might at least in some cases guide AF motif selections. However, in other cases this motif preference might also hinder discovery of multivalency in PPIs. For example, the use of smaller protein fragments for the protein pair SNRPB and GIGYF1 enabled the discovery of a proline-rich repeat motif in SNRPB.

In comparison to predictions made using full length proteins (Burke et al, 2023) we found that protein fragmentation increased the probability of obtaining a high confidence interface prediction, especially for cases involving proteins with long disordered regions such as GIGYF1. For smaller and more globular proteins like the PEX proteins studied above, full length predictions can identify the right binding sites but these can be further substantiated by running additional predictions with smaller fragments. The fragmentation approach increases the number of prediction runs per protein pair from one to a couple hundred, depending on the length and modularity of both proteins. The vast majority of these fragment pairs should not interact. With a FPR of 20%, this means that more actual non-interacting than truly interacting fragment pairs will result in a high confidence prediction. A big challenge is thus to identify likely correct interface predictions among the many false ones. This is also illustrated by the prediction results that we obtained for the seven protein pairs that we followed up experimentally. Clearly, AF's general limited specificity contributes to these false predictions. We observed that additional sources of error can arise from exposed intramolecular binding sites resulting from fragmentation, incorrectly designed boundaries of folded regions, and docking of protein fragments into enzymatic pockets of metabolic enzymes or sites for metal ion, DNA, or RNA binding. It seems that AF is overall well suited to find binding pockets on folded domains. However, our work also clearly demonstrates that AF is able to correctly dock the matching partner structure into these pockets without the need for a pre-existence of both partner structures in the bound conformation contrary to other state-of-the-art docking algorithms. AF's high sensitivity with respect to intramolecular binding sites and wrongly fragmented folded regions will make it particularly hard to fully automate the fragment design process. Despite these challenges we found that recurrent interface predictions from overlapping fragments can help gain confidence in predictions, as also highlighted in a recent study (Bronkhorst et al, 2023), since we rarely observed this recurrence for likely wrong predictions.

Given the reported uncertainties in AF predictions, even for high confidence cutoffs, experimental validation is essential. The BRET assay used here has been shown in previous studies to be sensitive enough to quantify weakening of binding introduced by point mutations and to detect motif-mediated PPIs (Ebersberger et al, 2023; Trepte et al, 2018; Mo et al, 2022). Using the BRET assay, we were able to detect 11 out of 28 PPIs from the HuRI dataset. This retest rate is actually higher compared to retest rates of gold standard PPI datasets used in the past to benchmark various binary PPI assays including this BRET assay, attesting the overall detectability of PPIs from HuRI (Braun et al, 2009; Trepte et al, 2018; Choi et al, 2019). The NL and mCit fusions used in the BRET assay allowed us to monitor the expression levels of wildtype and mutant constructs, which is important to rule out loss of binding because of a destabilization of the protein. However, we cannot exclude the possibility that some expressed mutants might still be partially unfolded or mislocalized and thus, some loss of binding detected in our study could be unspecific and not the result of a specific perturbation of the predicted interface. Furthermore, preservation of binding observed for some other mutants at the predicted interface might result from the mutations not being disruptive enough and thus, do not necessarily disprove the predicted interface.

Despite these limitations, we were able to assess the validity of seven interface predictions using experimentation. We discovered a likely novel DMI type that mediates binding between PEX3 and PEX16, and proposed a model for how PEX3, PEX16, and PEX19 form a trimeric complex at the peroxisomal membrane. We also validated a variation of the LIG_GYF motif class in SNRPB that mediates binding to GIGYF1 thereby potentially connecting mRNA splicing with posttranscriptional control mechanisms. These results confirm in principle that AF is able to predict novel interface types and that it can be used to extend existing interface type definitions. However, our experimental results also highlight clear limitations of AF predictions. Our data suggests that FBXO28 and STX1B as well as STX1B and VAMP2 interact via coiled-coil interfaces but likely at higher stoichiometries and different conformations than predicted. We confirmed the binding pocket in ESRRG but not the predicted interfaces in PSMC5 and we could not substantiate interface predictions for TRIM37 and PNKP. Highly confident interface predictions were obtained for seven additional PPIs that await experimental validation. In summary, we provided experimental evidence and structural information for PPIs whose disruption is likely associated with neurodevelopmental disorders. This information can be explored in future studies aimed at delineating potential molecular mechanisms causing disease. Our study furthermore laid out clear limitations, perspectives, and future needs in AI-based structure prediction to bring us closer to a fully structurally annotated human protein interactome.

Methods

Selection of structures for DMI benchmark dataset

To gather a list of ELM classes with structural evidence and annotate their minimal interacting fragments, we downloaded a dataset of solved structures of all ELM classes from ELM DB on 08.10.2021 (ELM class version 1.4) for instances that are annotated as true positives (Kumar et al, 2022). The structures were subject to a series of manual inspections to check their validity for further analysis. First, since AlphaFold can only model the 20 standard amino acids, we excluded any structures with post-translational modifications in the motif. Second, structures that do not resolve all of the residues in a motif as curated by ELM DB were excluded. Third, we restrict our studies to only binary interactions, so DMIs that require more than two proteins to form the binding interface were excluded. Likewise, DMIs with only intramolecular interaction evidence were excluded. We manually annotated the boundaries of the domains by visual inspection of the structures. After this filtering, we identified 136 structures from distinct ELM classes that formed our DMI benchmark dataset (Dataset EV2).

Sequence identity of the domains in the DMI benchmark dataset

We took all the binding domains in the DMI benchmark dataset and computed their pairwise sequence identity from a global alignment without gap penalties. Matching residues were given a score of 1, otherwise 0. The sum of these scores was divided by the length of the longer sequence to compute the sequence identity.

Selection of structures for the DDI benchmark dataset

We randomly selected 80 pairs of Pfam domain types that were described in the 3did resource (Mosca et al, 2014) to be in contact with each other in solved structures in the Protein Data Bank (PDB). We manually inspected all PDB entries listed to contain contacts between instances of a given Pfam domain pair until we found one that we considered a genuine domain-domain interaction. These decisions were primarily based on the number of atomic contacts observed and the validity that two folded domains were interacting with each other. Out of the 80 selected Pfam domain pairs, we identified 48 DDI types and 48 corresponding approved DDI structural instances that we selected for the DDI benchmark dataset. The sequences of the minimal interacting domain regions were manually annotated by visual inspection of the structures and used for prediction. A more detailed description of the curation procedure and information on the pairs will be soon published elsewhere (Geist et al, in preparation).

Generation of random reference sets with minimal interacting regions

Mutating motif sequences

Key conserved residues of the motifs in the DMI benchmark dataset were identified computationally using the regular expression of the corresponding ELM class in the ELM DB and SLiMSearch (Krystkowiak and Davey, 2017). The defined positions are any positions in the regular expression that are not wildcards. To mutate the key residues to the ones with opposite physico-chemical properties, we substituted one or two key residues with the ones that are of the largest Miyata distance (Miyata et al, 1979) (Dataset EV2).

Randomizing pairings of known domain-motif interfaces

To simulate non-binding domain-motif pairs, we randomized the pairings of known domain motif interfaces. As some domain types can bind to motifs from distinct ELM classes, we manually checked that the randomized pairings did not coincide with actual domain-motif interface types (Dataset EV2).

Randomizing pairings of known domain-domain interfaces

The pairings between known domain-domain interfaces were randomized to form the random reference set for DDIs.

Generation of positive DMI reference set with fragment extensions

Among the 136 solved structures that we selected previously, we further filtered for structures that consist of only human proteins. To test the potential effect of extension on DMIs that were predicted with different accuracies in their minimal forms, we selected 12 DMI types from the correct sidechain category, 8 DMI types from the correct backbone category and 11 DMI types from the correct pocket category as determined using the motif RMSD calculation. In total, 31 DMI types were selected for extension. Three additional DMI types were originally selected but later on discarded because they contained secondary motif occurrences complicating data analysis. The extensions were done on the canonical sequence of the proteins used to solve the structure. Motif extension 1 extended the motif sequence at both N and C termini by n residues where *n* is the length of the known motif. Motif extension 2 further extended the motif sequence by another *n* residues at both termini. Motif extension 3 and 4 each extended the motif sequence by 2*n* residues at both termini. Motif extension 5 extended the motif sequence by including neighboring domains and motif extension 6 used the full-length protein sequence. On the domain side, domain extension 1 extended the domain sequence to include the disordered regions N- and C-terminally of the binding domain until it reached neighboring domain(s) boundaries. Domain extension 2 included the sequence region of the neighboring domains and domain extension 3 used the full-length protein sequence. In cases where the known motif or binding domain is at the C terminus, we extended the motif or domain sequence on only the N terminus and vice versa. There were some cases where the last extension steps, motif extension 6 and domain extension 3, extended the protein minimally (<20 residues N or C terminal to the previous extension step). These cases were excluded from the analysis. The dataset of extended DMIs is in Dataset EV5. In total, 709 fragment pairs were submitted to AlphaFold. From these, 632 and 616 were successfully modeled by AF v2.2 and v2.3, respectively.

Generation of random DMI reference set with fragment extensions

To generate a random reference set using the extensions, we randomized the pairings of the 34 DMI types that we selected for extensions and paired their extensions for prediction. Motif extension 6 and domain extension 3 were excluded from the pairing. The dataset of DMIs with random pairings and their extensions can be found in Dataset EV6. In total, 612 predictions were generated, among which 566 and 522 predictions were successfully modeled by AF v2.2 and v2.3, respectively. Since motif extension 6 and domain extension 3 were excluded from the random reference set using the extensions, we also excluded them from the positive reference set extensions during ROC analysis. This resulted in 563 and 540 predictions from the positive reference set extensions for AF v2.2 and v2.3, respectively.

Selection of reference datasets for comparison of AF v2.2 with v2.3

All predictions for the minimal DMIs and the random DMIs involving minimal fragments were successfully modeled by both versions of AF. Some extensions from the positive reference set were not successfully modeled by AF v2.2 and v2.3 due to failure from HHblits. To compare AF v2.2 with v2.3, we used only predictions that were successfully modeled by both versions of AF. This resulted in 616 predictions from the extensions of the positive reference set.

Evaluation of AF sensitivity and specificity when using the fragmentation approach

Among the 34 DMIs selected for extension, we further selected 20 DMIs and retrieved the PPIs mediating these DMIs as the PRS and randomized their pairing to form random domain-motif protein pairs as the RRS. The 20 PPIs from the PRS and the 20 protein pairs from the RRS were subjected to the fragmentation approach, generating 8943 fragment pairs and 11,045 fragment pairs for the PRS and RRS, respectively. All fragment pairs from the PRS and all but one fragment pair from the RRS resulted in an AlphaFold model. Models were deemed highly confident, if the disordered fragment had a motif interface pLDDT of ≥70 or, in case of ordered-ordered models, the average interface pLDDT scored ≥70. To evaluate the sensitivity of the fragmentation approach, we considered all models that met the above mentioned cutoffs and which contained the motif and domain sequence. We superimposed the models onto the corresponding native structures using the minimal domain and computed the RMSD between the minimal motif residues in the native and modeled structure. A model was deemed accurate if the motif RMSD was ≤5 Å. At this cutoff the backbone of the native and modeled motif are well aligned but not necessarily their side chains (see also RMSD subsection below). We repeated the same procedure for each DMI protein pair using full length sequences as input into AF for modeling. In 18 cases AF did not return a model when using full length sequences. Here, we used the largest protein fragments instead for which AF returned a model. Information on the protein pairs, prediction results, and statistics is available in Dataset EV9.

AlphaFold versions and runs

We used local installations of AlphaFold Multimer version 2.2.0 and 2.3.0 (preprint:Evans et al, 2021) for all protein complex predictions with the following parameters:

```
--max_template_date=2020-05-14
--db_preset=full_dbs
--use_gpu_relax=False
```

For every AlphaFold run, five models were predicted with single seed per model by setting the following parameter:

```
--num_multimer_predictions_per_model=1
```

The databases queried during AlphaFold predictions were specified following the instructions from the github page of AlphaFold (https://github.com/deepmind/alphafold#running-alphafold):

For running AlphaFold Multimer v2.2, the following databases were queried:

--bfd_database_path=bfd_metaclust_clu_complete_id30_c90_-final_seq.sorted_opt

--mgnify_database_path=alphafold_v220_databases/mgy_clusters_2018_12.fa

--obsolete_pdbs_path=alphafold_v220_databases/pdb_mmcif/obsolete.dat

--pdb_seqres_database_path=alphafold_v220_databases/pdb_seqres/pdb_seqres.txt

--template_mmcif_dir=alphafold_v220_databases/pdb_mmcif/mmcif_files

--uniprot_database_path=alphafold_v220_databases/uniprot/uniprot.fasta

--uniclust30_database_path=alphafold_v220_databases/uniclust30/uniclust30_2018_08/uniclust30_2018_08

--uniref90_database_path=alphafold_v220_databases/uniref90/uniref90.fasta

For running AlphaFold Multimer v2.3, the following databases were queried:

--bfd_database_path=alphafold_v230_databases/bfd/bfd_metaclust_clu_complete_id30_c90_final_seq.sorted_opt

--mgnify_database_path=alphafold_v230_databases/mgnify/mgy_clusters_2022_05.fa

--obsolete_pdbs_path=alphafold_v230_databases/pdb_mmcif/obsolete.dat

--pdb_seqres_database_path=alphafold_v230_databases/pdb_seqres/pdb_seqres.txt

--template_mmcif_dir=alphafold_v230_databases/pdb_mmcif/mmcif_files

--uniprot_database_path=alphafold_v230_databases/uniprot/uniprot.fasta

--uniref30_database_path=alphafold_v230_databases/uniref30/UniRef30_2021_03

--uniref90_database_path=alphafold_v230_databases/uniref90/uniref90.fasta

To test the effect of template use on prediction accuracy, the following parameter setting was used to switch off the use of templates during the prediction:

--max_template_date=1950-01-01

For the fragmentation approach, the multiple sequence alignments (MSAs) of a given protein fragment can be reused in subsequent runs where the same fragment is involved. The MSAs were first moved to the prediction output folder and the following parameter was added to enable the reuse of MSAs.

--use_precomputed_msas=True

For efficient computing, we segregated the MSA generation part by using only the CPUs and the model fitting part using the GPUs.

Calculation of metrics for structural models

Motif RMSD

We used the software PyMOL (TM) Molecular Graphics System, Version 2.5.0. Copyright (c) Schrodinger, LLC., for the superimposition of AlphaFold models with corresponding solved structures. First, we used the align command to align the domain chain in AlphaFold models with the domain chain in the solved structure. Then, we used the rms_cur command to calculate the all-atom RMSD between the motif chain in AlphaFold models and the motif chain in the solved structure. To ensure that the RMSD calculation was done based on all atom identifiers and without any outlier rejection refinement, the arguments of the rms_cur command, matchmaker and cycles, were set to 0. Prediction accuracy categories were defined based on motif RMSD cutoffs: RMSD ≤ 2 Å for correct sidechain, between 2 Å and 5 Å for correct backbone, between 5 Å and 15 Å for correct pocket and >15 Å for wrong pocket.

DockQ

The calculation of DockQ scores of AlphaFold models was done in reference to their solved structures using the code available on the github repository of DockQ (https://github.com/bjornwallner/DockQ, (Basu and Wallner, 2016). DockQ classification was done using the cutoffs provided by DockQ (DockQ: <0.23 for incorrect, between 0.23 and 0.49 for acceptable, between 0.49 and 0.80 for medium and ≥0.80 for high).

pDockQ

The calculation of pDockQ of AlphaFold models was done by adapting the code available on the github repository from the Elofsson lab (https://gitlab.com/ElofssonLab/FoldDock/-/blob/main/src/pdockq.py, (Bryant et al, 2022)). The pDockQ score is created by fitting a sigmoidal curve to the DockQ scores of a series of AlphaFold predicted models. The score takes into account the number of interface contacts as well as their pLDDT scores. Of note, the calculation of pDockQ score takes Cβs (Cα for glycine) from different chains within 8 Å from each other as interface contacts which is different from our interface definition (see the subsection below *Domain chain and motif chain interface pLDDT and average interface pLDDT).*

iPAE

The calculation of iPAE of AlphaFold models was done by adapting code available on the github repository https://github.com/fteufel/alphafold-peptide-receptors/tree/main (Teufel et al, 2023). The iPAE is the median predicted aligned error at the interface. The authors consider residues in contact if their distance is below 0.35 nm (3.5 Å). The iPAE score could not be calculated for models generated by AlphaFold Multimer version 2.3.0 due to JAX dependency of the pickle files generated by AlphaFold Multimer version 2.3.0.

Model confidence

The model confidence of AlphaFold models was extracted from the ranking_debug json file. The model confidence is a weighted combination of pTM and ipTM to account for both intra- and interchain confidence:

$$model\ confidence = 0.8 \cdot ipTM + 0.2 \cdot pTM$$

Domain chain and motif chain interface pLDDT and average interface pLDDT

Since AlphaFold conveniently stores the pLDDT confidence measure for each residue in the B-factor field of the output PDB files, the pLDDT of residues at the interface was parsed from the output PDB files of AlphaFold. Residues at the interface are defined as those that have at least one heavy atom that is less than 5 Å away from any heavy atom of the other chain (calculated using the

PyMOL API). The pLDDT of the residues at the interface from the domain chain and motif chain was averaged to compute the domain chain and motif chain interface pLDDT, respectively. The pLDDT of all the residues from both chains was averaged to compute the average interface pLDDT.

Residue-residue and atom-atom contacts

Following the interface definition above, the number of unique residue-residue and atom-atom contacts were also quantified as measurements to assess AlphaFold models.

Mean DockQ between predicted models

The top five models generated by AF, determined based on their model confidence, were considered for computing this metric. To quantify the similarity among the models, we computed DockQ scores between all possible pairs of models by taking the higher ranked model as the “template” model and lower ranked model as the “predicted” model. The mean of these DockQ scores is taken as the similarity among the models in a given prediction. This calculation was done for AF models of minimal DMIs and their randomizations for ROC analysis. The data were stored in Dataset EV2.

Quantification of motif properties

Motif hydropathy score and symmetry score

By referring to the Kyte-Doolittle hydrophobicity scale, (Kyte & Doolittle, 1982) the hydropathy scores of the amino acids in a given motif were summed and averaged to compute the average hydropathy of the motif. The average motif symmetry score was computed by taking the sum of the absolute difference of hydropathy scores between motif position n and n - motif length + 1 and division of this sum by half of the motif length:

$$Peptide\ symmetry\ score = \frac{\sum_{n=1}^{a} |(H_n - H_{x-n+1})|}{a}$$

where x is the length of the motif and a is the floor division of x by 2.

Motif probability

The motif probability reflects the degeneracy of a given motif class as quantified by its regular expression that is annotated in the ELM DB. The motif probability was retrieved from the ELM DB version 1.4.

Secondary structure elements of motifs

We extracted the secondary structure elements of motifs using the PyMOL API. In cases where the motif adopts partial secondary structure, such as loop-helix-loop or loop-strand-loop, they are treated as helical or strand, respectively.

Selection of motif classes from ELM DB without annotated structural instances and prediction with AF

By querying the ELM DB for all ELM classes, we retrieved a list of ELM classes and the number of instances with a structure solved (column #instances_in_PDB). We filtered for ELM classes with 0 instances_in_PDB and selected 205 instances out of the filtered ELM classes for AF prediction. The ELM instances were extended at both N and C termini by n residues where *n* is the length of the ELM instance, according to the benchmarking results. The minimal binding domains of the ELM instances were detected in the interaction partner using Pfam HMMs (Mistry *et al*, 2021). As the domain boundaries detected by Pfam HMMs could be inaccurate, we also extended the domain sequence at the N and C terminus by 20 residues to ensure that the whole folded region was covered. The predictions were performed using AF version 2.3.0. To select a subset of these motif classes, where we can do experimental testing, we also used the InParanoid resource (Persson & Sonnhammer, 2023) to map ELM instances where both proteins are from mouse to their human orthologs. To verify that they indeed do not have structural homologues in the PDB, we both used the SIFTS mapping (Dana *et al*, 2019) between the Pfam domain in ELM and the PDB and also looked at the ELM classes that were listed as homologs on the ELM website.

Evaluation of effect of fragment extensions on AF prediction accuracies

We superimposed the AF models generated with DMI extensions onto the corresponding solved DMI structures to quantify AF prediction accuracy using motif RMSD calculations. To this end, we aligned the two structures on their minimal binding domains and calculated the all-atom RMSD between the minimal motif in the extension AF model and the minimal motif in the solved structure. To determine potential differences in DMI prediction accuracy when using minimal versus extended protein fragments, we computed the log2 fold change of the all-atom motif RMSD before and after extension.

$$Fold\ change\ in\ prediction\ accuracy = log_2\left(\frac{all\ atom\ RMSD\ motif_{minimal\ DMI}}{all\ atom\ RMSD\ motif_{extended\ DMI}}\right)$$

Fragment design and fragment pairing for fragmentation approach

We first inspected the monomeric structural models from the AlphaFold database (Varadi et al, 2022; Jumper et al, 2021) of both interacting proteins to determine the boundaries of their ordered and coiled-coil regions, which were also treated as “ordered”. All regions that were not annotated as ordered were annotated as disordered. In some cases, an extended loop with low pLDDT can be found within an ordered region. As they can also potentially carry a motif or mediate interactions in another way, these regions were also annotated as disordered in addition to their annotation as being part of a larger ordered region. The disordered regions of the proteins were fragmented into fragment sizes of 10, 20 and 30 residues. To allow AF to sample continuous sequences, we also generated another set of fragments of same sizes that overlap with the previous fragments by sliding the sequence by half the size of the fragment. The unfragmented disordered regions, as well as their fragments, from one protein were then paired with the ordered regions from its interacting partner and vice versa for prediction. The ordered regions from both proteins were also paired for prediction. We decided to manually define boundaries between ordered and disordered regions because testing available code developed for this purpose, like clustering using the PAE matrix, turned out to be too inaccurate. We observed that erroneous removal of residues close to the domain borders that are still contributing to the folding of a structured domain, can heavily mislead AF predictions.

Selection of NDD proteins

A list of NDD genes was assembled using whole exome and whole genome sequencing studies of cohorts of NDD patients from Gene4Denovo (Zhao et al, 2020) and Deciphering Developmental Disorders (DDD) study (Firth et al, 2011), respectively. From Gene4Denovo, we selected genes linked to autism-spectrum disorders (ASD), intellectual disability (ID), epilepsy (EE), undiagnosed developmental disorders (UDD) and NDDs in general. Genes with non-coding mutations as well as genes with a false discovery rate (FDR) >= 0.05 were excluded. Similarly, in the DDD study, genes associated with developmental disorders with a neurological component, as well as genes found to be mutated in at least three children with NDDs (labeled as confirmed genes) were retained. The final list included 984 NDD-risk genes. We filtered the HuRI network (Luck et al, 2020) for interactions mediated exclusively by proteins from this NDD gene list resulting in 67 PPIs excluding self-interactions. Since our fragmentation approach generates many fragments, we did not consider PPIs involving proteins that are more than 1500 amino acids in length, resulting in a final list of 62 PPIs that were subjected to AF modeling.

Manual inspection of interface predictions for NDD-NDD PPIs and selection for experimental validation

Paired fragments from NDD-NDD PPIs were predicted using AF version 2.2 and the prediction results are stored in Dataset EV10. Based on our benchmarking results, we started by manually inspecting all NDD-NDD PPIs that obtained at least one structural model with either a motif chain interface pLDDT of ≥70 for the disordered fragment or with an average interface pLDDT ≥ 70 for structural models with predicted ordered-ordered interfaces (DDIs). However, during the course of these manual inspections, we found that using in addition a model confidence of ≥0.7 for ordered-ordered fragment pairs helped discriminating good from bad structural models. We inspected the ranked_0 models for all fragment pairs that met the above cutoffs but also inspected models scoring somewhat below these cutoffs. For every NDD-NDD PPI we used Interactome3D (Mosca et al, 2013) and PDB database searches (https://www.rcsb.org/ (Berman et al, 2000)) to identify whether a structure already existed for this PPI. In our evaluation of the structural models we also considered if a certain interface was recurrently predicted for different overlapping fragments because this usually hints at increased confidences for the correctness of the interface prediction. We furthermore explored the number and kind of residue-residue contacts predicted by AF by visual inspection of the structural models using PyMol. We searched for functional annotations and existing structures for the monomers using the PDB, ProViz (Jehl et al, 2016), SMART (Letunic et al, 2021), and the scientific literature to identify enzymatic pockets or binding interfaces for DNA, RNA, or metal ions. Observations and justifications for the final evaluation of the predictions for every NDD-NDD PPI are provided in Appendix Supplementary Text S1.

Based on clone availability, we selected 49 of the 62 PPIs for experimental validation of the predicted interfaces using the BRET assay. For 30 of the 49 selected PPIs for experimental testing we obtained sequence-confirmed clones with luciferase and mCitrine fusions. For 28 of these PPIs both partners were expressed in our experimental system as determined by total luminescence and fluorescence measurements (Fig. 3D,F).

Softwares used

We used the software PyMOL (TM) Molecular Graphics System, Version 2.5.0. Copyright (c) Schrodinger, LLC., for the visualization and superimposition of AlphaFold models.

All codes were written in Python3 and analyses were done using Jupyter notebooks. We used the Python libraries, Biopython (Cock et al, 2009) for sequence similarity computation, pandas (McKinney, 2010) for data analysis, and Matplotlib (Hunter, 2007) and seaborn (Waskom, 2021) for data visualization. ROC and PR statistics were calculated using the Python package sci-kit learn (Pedregosa et al, 2012).

Cell line culture and maintenance

HEK293 cells were purchased from DSMZ (catalog number ACC305). These cells were grown and maintained in DMEM (Thermo Fisher), supplemented with 10% FBS (PAN-Biotech), 2 mM glutamine (Thermo Fisher) and 1% penicillin–streptomycin (Thermo Fisher). Cells were incubated at 37 °C with 5% CO_2. Subcultivation was performed with 1 ml of 0.05% trypsin every 2–3 days for up to 40 passages. For each passage 1–2 × 10^6 cells were seeded in T25 flasks (Sarstedt). Then, new cells were thawed from stocks containing 2 × 10^6 cells in 1 ml of growth medium, supplemented with 10% DMSO (Sigma). Every 3 months cells were checked for mycoplasma contamination using a PCR test (Dataset EV11). The cell line was purchased from DSMZ four years ago, expanded, aliquoted, and frozen. A new aliquot is thawed after every 40 passages. No further authentication of the cell line has been done.

Plasmid construction

Standard controls

The donor and acceptor vectors pcDNA3.1-cmyc-NL-GW (Addgene plasmid ID #113446), pcDNA3.1-GW-NL-cmyc (Addgene plasmid ID #113447), pcDNA3.1 GW-His3C-mCit, pcDNA3.1 mCit-His3C-GW as well as controls pcDNA3.1-NL-cmyc (Addgene plasmid ID #113442), pcDNA3.1-PA-mCit (Addgene plasmid ID #113443) were kindly provided by the Wanker Group (Max-Delbrück-Centrum für Molekulare Medizin, Germany) (Dataset EV12). By default we cloned all ORFs of interest into N-terminal NL and mCit fusion destination vectors and occasionally also transferred ORFs into C-terminal fusion vectors if N-terminal fusions did not result in sufficient BRET signals but the interaction was of high interest to this study and predicted interfaces were closer to the C-terminus. Trepte et al have shown that testing protein pairs in different configurations increases detection rates while maintaining low false detection rates and that BRET signals are higher if fusions are close to the actual interaction interface (Trepte et al, 2018; preprint:Trepte et al, 2021; preprint:Trepte et al, 2023).

GATEWAY cloning procedure

Full-length wild-type human open reading frames (ORFs) being cloned in GATEWAY entry vectors from the ORFeome collaboration are stored as bacterial glycerol stocks. (ORFeome Collaboration, 2016)

1. The ORFs were inoculated in 96-well plates (Corning), with each well containing 200 uL of LB medium and 100 μg/ml ampicillin. The plate was incubated at 37 °C and left to shake overnight at 190 rpm.
2. In a 96-well PCR plate (Brand) 10 ng of each selected ORF was used per 50 μl PCR reaction (denaturation at 98 °C for 10 s, annealing at 55 °C for 30 s and extension at 72 °C for 3 min, 30 cycles of amplification) using phusion high-fidelity polymerase (NEB) and primers annealing to the backbone of the plasmid (forward: 5′TTGTAAAACGACGGCCAGTC and reverse: 5′ GCCAGGAAACAGCTATGACC).
3. The PCR products (6 μl per well) were confirmed through 96-well E-gel with SYBR (Thermo Fisher, Catalog no G720801) using 25 μl of loading buffer (Thermo Fisher) and 20 μl of E-Gel 96 High range DNA marker (Thermo Fisher).
4. In a 96-well PCR plate 1 μl of each amplified PCR product together with 200 ng of above-mentioned destination vectors were directly used per 10 μl LR reaction using 4x LR clonase (Invitrogen), thereby generating expression vectors.
5. The full 10 μl of LR reaction was transformed into chemically competent DH5a cells (30 μl) in a 96-well PCR plate, then recovered in 80 μl of pre-warmed SOC medium at 37 ˚C for 1 h without shaking.
6. 70 μl of transformed bacteria was plated on 48-well square agar plates and incubated at 37 °C overnight.
7. Afterwards, colonies were selected and inoculated into a 96 deep-well plate containing 2 ml of LB medium and 100 μg/ml ampicillin. The plate was then incubated at 37 ˚C with continuous shaking at 700 rpm in the incumixer for 24 h.
8. The amplified vectors were extracted from the inoculated culture using Plasmid Plus 96-well Miniprep kit (Qiagen). The concentration of each vector was measured with a Nanophotometer and diluted to 100 ng/μl. Next, 600 ng of insert was used for full-length sequencing using the backbone primers (tag-specific NanoLuc forward: 5′GAACGGCAACAAAATTATCGAC, mCitrine forward: 5′AGCAGAATACGCCCATCG and reverse: 5′GGCAACTAGAAGGCACAGTC) and ORF-specific primers (Dataset EV11) to fully cover the ORFs where it was needed (Dataset EV12). All sequence-confirmed ORF sequences used in this study are available in Dataset EV13.

Site-directed mutagenesis

The primers were manually designed using the following criteria:

1. For point mutation the primers should overlap the site of mutation. The overlap should be 15–20 nucleotides (nt).
2. For the deletion the primers should be designed to exclude the deletion site, but still overlap and the overlap should be as mentioned in step 1.
3. Primer length should be in the range of 32–36 nt.
4. GC content should be between 40–60%.
5. Difference in melting temperature of primers should not exceed 5 °C.
6. The primer ideally should start and end with guanine or cytosine.
7. The designed oligos were grouped by annealing temperature for the next step.
8. In 96-well PCR plate 10 ng of DNA template together with oligos were used per 50 μL of PCR reaction (denaturation at at 98 °C for 2 min, annealing for 15 s and extension at 72 °C for 5 min, 25 cycles of amplification) using phusion high-fidelity polymerase (NEB).
9. 1 μL of DpnI (NEB) was added to the plate with PCR products and incubated at 37 °C for 1 h. The reaction was stopped at 65 °C for 20 min.
10. The PCR products (6 μl per well) were confirmed through 96-well E-gel with SYBR (Thermo Fisher, Catalog no G720801) using 25 μl of loading buffer (Thermo Fisher) and 20 μl of E-Gel 96 High range DNA marker (Thermo Fisher).
11. 3 μL of digested PCR product was transformed into chemically competent DH5a cells (30 μL) in a 96-well PCR plate, then recovered in 80 μL of pre-warmed SOC medium at 37 °C for 1 h without shaking.
12. 70 μL of transformed bacteria was plated on 48-well square agar plates and incubated at 37 °C overnight.
13. Afterwards, colonies were selected and inoculated into a 96 deep-well plate containing 2 ml of LB medium and 100 μg/ml ampicillin. The plate was then incubated at 37˚C with continuous shaking at 700 rpm in the incumixer for 24 h.
14. The amplified vectors were extracted from the inoculated culture with Plasmid Plus 96-well Miniprep kit (Qiagen). The concentration was measured with a Nanophotometer and diluted to 100 ng/μl. Next, 600 ng of insert was used for full-length sequencing using primers covering the mutation and ORF-specific primers (Dataset EV11) to fully cover the ORF length (Dataset EV12).

BRET assay

Transfection

HEK293 cells were grown and maintained in high-glucose (4.5 g/l) DMEM (Thermo Fisher) for BRET assays. Media was supplemented with 10% fetal bovine serum (PAN-Biotech) and 1% Penicillin/Streptomycin. Cells were grown at 37 °C, 5% $CO_{2,}$ and 85% RH. Cells were subcultured every 2–3 days and transfected with lipofectamine 2000 transfection reagent (Invitrogen) in Opti-MEM medium (Thermo Fisher) using the reverse transfection method according to the manufacturer's instructions. For transfections, cells were seeded at a density of 4.0×10^4 cells per well in a white 96-well microtiter plate (Greiner) in phenol-red-free, high-glucose DMEM media (Thermo Fisher) supplemented with 5% fetal bovine serum (Thermo Fisher). Transfections were performed with a total DNA amount of 200 ng per well. If the expression plasmid concentration amount was below 200 ng/well, pcDNA3.1 (+) was used as a carrier DNA to reach the total amount of DNA of 200 ng. All protein pairs were tested in both N-terminal fusion orientations (NL-A with mCit-B and NL-B with mCit-A). The following proteins were also tested as C-terminal fusions: CSNK2B-NL, ESRRG-NL, CUL3-NL, PEX3-NL, PEX19-NL, PSMC5-NL, PEX3-mCit, PEX19-mCit, PEX16-mCit, RORB-mCit, ESRRG-mCit, PAX6-mCit, CSNK2B-mCit, PSMC5-mCit, KCTD7-mCit (Dataset EV12).

Measurement

The plate was incubated 2 days at 37 °C, 5% CO_2, and 85% RH before measurements. All measurements were done with the Infinite M200 Pro microplate reader (Tecan). First, 100 µl of the medium was aspirated from each well. The mCitrine fluorescence (FL) was measured in intact cells (excitation/emission 513 nm/548 nm) using a gain of 100. On rare occasions, the plate reader recorded an overflow with these settings (i.e. for GIGYF1 constructs). In these cases, we repeated the measurement with optimal gain settings and used a fluorescein control to normalize fluorescence signals measured with different gain settings. For this purpose, Fluorescein was obtained from Sigma-Aldrich (Catalog No 46955-250MG-F) and used without further purification. A stock solution of Fluorescein (1 mg/ml in Ethanol) was prepared by dissolving 1.3 mg Fluorescein in 1.3 ml absolute ethanol. 100 µl of a 20 µg/ml solution of Fluorescein were added to an empty well immediately before starting the fluorescence measurements. The 20 µg/ml solution of Fluorescein was obtained by preparing a 1:50 dilution in water of the stock solution. After measuring the fluorescence, coelenterazine-h (PJK Biotech GmbH) was added to a final concentration of 5 µM. The cells were briefly shaken for 15 s and incubated for 15 min inside the plate reader at 37 °C. After incubation, total luminescence was measured first followed by short-wavelength (WL) and long-wavelength luminescence (LU) measurements using the BLUE1 (370–480 nm) and the GREEN1 (520–570 nm) filters at 1000 ms integration time. Corrected BRET ratios were calculated as described in (Trepte et al, 2018). Briefly, for every transfected protein pair NL-A and mCit-B, the following two control pairs were measured: NL-Stop with mCit-B and NL-A with mCit-Stop. The maximal BRET from both control pairs was subtracted from the actual test pair to correct for donor bleedthrough, unspecific binding to the tags, and background signal.

Determination of binding events in BRET assay

To determine whether a protein pair interacted in the BRET assay or not, we used donor:acceptor DNA transfection ratios of 2:50 ng in all cases except for PEX3-PEX16 where we used 8:25 and PEX3:PEX19 where we used 8:50 ng DNA ratios due to low expression levels of PEX3 and a degradation effect of higher PEX16 protein levels on PEX3 expression levels. We requested that cBRETs determined at these transfection ratios were ≥0.05, fluorescence measurements representing mCitrine fusion expression levels to be ≥500 units, and total luminescence measurements representing NL fusion expression levels to be ≥50,000.

Saturation assay

For donor saturation experiments various donor DNA amounts (1, 2, 4 and 8 ng) encoding NL-fused proteins were co-transfected with increasing amounts of acceptor DNA (12.5, 25, 50, 100, 200 ng) encoding mCitrine-fused proteins. Fluorescence, total luminescence, and BRET measurements were done as described before. BRET measurements were corrected for bleedthrough using NL-Stop transfections. Fluorescence and total luminescence measurements were corrected for background signal using transfections with pcDNA3.1(+) and subsequently used to estimate amounts of expressed proteins and to plot acceptor/donor ratios on the x-axis of titration plots.

Fitting of titration curves

Titration curves were fitted using the leastsq function from the scipy.optimize python package (Virtanen et al, 2020) using the model BRET = ((A/D) * BRETmax)/(BRET50 + (A/D)) described in (Drinovec et al, 2012), which assumes a 1:1 binding mode, to obtain estimates for the BRETmax and BRET50. Standard errors of the BRET50 estimates were obtained from the variance-covariance matrix, calculated by multiplying the fractional covariance matrix (output by leastsq function) by the residual variance. Measuring BRET signals in intact cells for increasing acceptor/donor protein expression ratios results in an eventual saturation of the signal. Fitting this curve allows extraction of the maximal BRET that can be reached and the BRET50, which is the acceptor/donor ratio at which half of the maximal BRET is obtained. The BRET50 is indicative of binding affinity, in analogy to the IC50, however, its accurate estimation requires saturation of the BRET to be observed in the experimental system, which cannot always be achieved because of limited amounts of DNA that cells can be transfected with. Alternatively, if mutations are unlikely to change the overall structure of the fusion constructs and do not alter expression levels compared to wildtype, single point BRET measurements at acceptor/donor ratios prior to BRET saturation are also indicative of changes in binding strength. The BRET titration curves that we obtained for the PNKP-TRIM37 interaction clearly deviated from the assumed 1:1 binding mode because at higher acceptor:donor ratios we observed a sudden increase in BRET again contrary to an expected saturation. The model could thus not be fitted to the titration data.

Antibodies

Purified anti-HA.11 Epitope Tag, Clone: [16B12], Mouse, Monoclonal (Biolegend, BLD-901502), 1:2000.

Purified anti-GIGYF1, Rabbit, Polyclonal (BETHYL laboratories, Cat. #A304-132A-1), 1:1000.

GAPDH Loading Control Monoclonal Antibody (GA1R), HRP-coupled (Thermo Fisher Cat. MA515738HRP), 1:3000.

Co-immunoprecipitation and western blot

Snrpb (full-length) and C-terminal truncation mutant (amino acids 1-190) was cloned from mouse cDNA and ligated into pFRT-TO destination plasmid using AscI and PacI restriction sites. The constructs additionally contain C-terminal 2xHA and mNeonGreen tags. Flp-In™ T-REx™ 293 Cell Lines (Thermo Fisher, catalog number: R78007) expressing Snrpb endogenously from a single locus were generated according to the manufacturer's instructions. In brief, pFRT-TO and pOG44 plasmids were co-transfected and hygromycin-resistant colonies were grown, picked and expanded. The Snrpb transgene expression was validated by western blot, RT-qPCR, and immunofluorescence, which showed that ectopic Snrpb-HA was expressed at levels highly similar to the endogenous Snrpb protein.

For the co-immunoprecipitation experiments, 8 × 106 cells were seeded in a 10 cm dish. The following day, expression of Snrpb-HA was induced by adding 0.1 µg/mL Doxycycline (D9891, Sigma Aldrich) to the culture medium. Parental cells not expressing any HA-tagged transgene were used as a negative control of immunoprecipitation. The next morning the cells were harvested by scraping in culture media, followed by centrifugation and a single wash in ice-cold PBS. The whole cell extract was prepared by 15 min incubation on ice with 0.3 mL of lysis buffer (200 mM NaCl, 50 mM HEPES, pH 7.6, 0.1% IGEPAL, 10 mM $MgCl_2$, 10% Glycerol, Protease Inhibitor Cocktail (P8340, Sigma Aldrich), Phosphatase Inhibitor (P5726, Sigma Aldrich) followed by 2 cycles of sonication in a Bioruptor Plus (30 s on, 30 s off) and centrifugation for 20 min at 16,000 × *g*. The extract was quantified by a Bradford assay and 1 mg was used for immunoprecipitation, for which the NaCl concentration was adjusted to 100 mM final concentration by diluting with an equal volume of Lysis Buffer containing 0 mM NaCl. 0.05 mg was set aside as input control (5%). 0.02 mL of Thermo Scientific™ Pierce™ Anti-HA Magnetic Beads (Thermo Fisher Cat. 13464229) were incubated with 1 mg protein extract for 1 h at 4 °C on a rotating wheel. The beads were washed three times before eluting the immunoprecipitated proteins with 0.02 mL of 1 x NuPAGE™ LDS Sample Buffer by incubating at 42 °C for 10 min while shaking at 800 rpm. Another 0.01 mL were used for elution, were then combined making a total of 30 µL, which were transferred to a fresh tube and to which 3 µL of 1 M DTT were added. Input and immunoprecipitated eluates were then separated on a 10% Tris-Glycine SDS PAGE using 1xMOPS buffer, immunoblotted on 0.45 µm PVDF membranes (Tris-Glycin Transfer Buffer, 10% Methanol, 300 mA, 1 hour), blocked with 5% milk in TBS-0.2% Tween for 30 min at RT. Primary antibodies were incubated overnight at 4 °C on a rocker followed by washes and incubation with secondary HRP-labeled antibodies (1 h at RT in 5% milk, TBS-0.2% Tween). Blots were developed using Pierce™ ECL Western Blotting Substrate (Thermo Fisher Cat. 32209) or SuperSignal West Femto Maximum Sensitivity Substrate Kit (Thermo Fisher Cat. 34095) and imaged on a ChemiDoc MP V3 (Bio-Rad). The cell line was authenticated via X-Gal staining, qPCR and Sanger Sequencing.

Data availability

The datasets and computer code produced in this study are available in the following databases:

- Interaction data: submitted to the IMEx (http://www.imexconsortium.org) consortium through IntAct (Del Toro et al, 2022) and assigned the identifier IM-29904.
- Computer scripts for data processing and analysis: available at GitHub under https://github.com/KatjaLuckLab/AlphaFold_manuscript.

Expanded view data, supplementary information, appendices are available for this paper at https://doi.org/10.1038/s44320-023-00005-6.

Peer review information

A peer review file is available at https://doi.org/10.1038/s44320-023-00005-6

References

Ajuh P, Chusainow J, Ryder U, Lamond AI (2002) A novel function for human factor C1 (HCF-1), a host protein required for herpes simplex virus infection, in pre-mRNA splicing. EMBO J 21:6590–6602

Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL, Laskowski RA, Pozzati G et al (2022) A structural biology community assessment of AlphaFold2 applications. Nat Struct Mol Biol 29:1056–1067

Basu S, Wallner B (2016) DockQ: a quality measure for protein-protein docking models. PLoS ONE 11:e0161879

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Braun P, Tasan M, Dreze M, Barrios-Rodiles M, Lemmens I, Yu H, Sahalie JM, Murray RR, Roncari L, de Smet AS, Venkatesan K, Rual JF, Vandenhaute J, Cusick ME, Pawson T, Hill DE, Tavernier J, Wrana JL, Roth FP, Vidal M (2009) An experimentally derived confidence score for binary protein-protein interactions. Nature Methods 6:91–97

Bret H, Andreani J, Guerois R (2023) From interaction networks to interfaces: Scanning intrinsically disordered regions using AlphaFold2. Preprint at BioRxiv https://doi.org/10.1101/2023.05.25.542287

Bronkhorst AW, Lee CY, Möckel MM, Ruegenberg S, de Jesus Domingues AM, Sadouki S, Piccinno R, Sumiyoshi T, Siomi MC, Stelzl L, Luck K, Ketting RF (2023) An extended Tudor domain within Vreteno interconnects Gtsf1L and Ago3 for piRNA biogenesis in Bombyx mori. EMBO J 42(24):e114072 https://doi.org/10.15252/embj.2023114072

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

Buel GR, Walters KJ (2022) Can AlphaFold2 predict the impact of missense mutations on structure? Nat Struct Mol Biol 29:1–2

Bugge K, Brakti I, Fernandes CB, Dreier JE, Lundsgaard JE, Olsen JG, Skriver K, Kragelund BB (2020) Interactions by disorder - a matter of context. Front Mol Biosci 7:110

Burke DF, Bryant P, Barrio-Hernandez I, Memon D, Pozzati G, Shenoy A, Zhu W, Dunham AS, Albanese P, Keller A et al (2023) Towards a structurally resolved human protein interaction network. Nat Struct Mol Biol 30:216–225

Chang L, Perez A (2023) Ranking peptide binders by affinity with AlphaFold. Angew Chem Int Ed 62:e202213362

Choi SG, Olivet J, Cassonnet P, Vidalain PO, Luck K, Lambourne L, Spirohn K, Lemmens I, Dos Santos M, Demeret C, Jones L, Rangarajan S, Bian W, Coutant EP, Janin YL, van der Werf S, Trepte P, Wanker EE, De Las Rivas J, Tavernier J, Twizere JC, Hao T, Hill DE, Vidal M, Calderwood MA, Jacob Y (2019) Maximizing binary interactome mapping with a minimal number of assays. Nature Communications 10:3907

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B et al (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423

Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, Velankar S (2019) SIFTS: updated structure integration with function, taxonomy and sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Res 47:D482–D489

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ (2012) Attributes of short linear motifs. Mol Biosyst 8:268–281

Del Toro N, Shrivastava A, Ragueneau E, Meldal B, Combe C, Barrera E et al (2022) The IntAct database: efficient access to fine-grained molecular interaction data. Nucleic Acids Res 50(D1):D648–53

Drew K, Lee C, Huizar RL, Tu F, Borgeson B, McWhite CD, Ma Y, Wallingford JB, Marcotte EM (2017) Integration of over 9000 mass spectrometry experiments builds a global map of human protein complexes. Molecular Systems Biology 13:932

Drinovec L, Kubale V, Nøhr Larsen J, Vrecl M (2012) Mathematical models for quantitative assessment of bioluminescence resonance energy transfer:

application to seven transmembrane receptors oligomerization. Front Endocrinol 3:104

Durocher D, Taylor IA, Sarbassova D, Haire LF, Westcott SL, Jackson SP, Smerdon SJ, Yaffe MB (2000) The Molecular Basis of FHA Domain:Phosphopeptide Binding Specificity and Implications for Phospho-Dependent Signaling Mechanisms. Molecular Cell 6:1169–1182

Ebersberger S, Hipp C, Mulorz MM, Buchbender A, Hubrich D, Kang HS, Martínez-Lumbreras S, Kristofori P, Sutandy FXR, Llacsahuanga Allcca L, Schönfeld J, Bakisoglu C, Busch A, Hänel H, Tretow K, Welzel M, Di Liddo A, Möckel MM, Zarnack K, Ebersberger I, Legewie S, Luck K, Sattler M, König J (2023) FUBP1 is a general splicing factor facilitating 3′ splice site recognition and splicing of long introns. Molecular Cell 83:2653–2672

Ernst JA, Brunger AT (2003) High Resolution Structure Stability and Synaptotagmin Binding of a Truncated Neuronal SNARE Complex. Journal of Biological Chemistry 278:8630–8636

Evans R, O'Neill M, Pritzel A, Antropova N, Senior AW, Green T, Žídek A, Bates R, Blackwell S, Yim J et al (2021) Protein complex prediction with AlphaFold-Multimer. Preprint at BioRxiv https://doi.org/10.1101/2021.10.04.463034

Firth HV, Wright CF, DDD Study (2011) The deciphering developmental disorders (DDD) study. Dev Med Child Neurol 53:702–703

Freiman RN, Herr W (1997) Viral mimicry: common mode of association with HCF by VP16 and the cellular protein LZIP. Genes Dev 11:3122–3127

Freund C, Kühne R, Yang H, Park S, Reinherz EL, Wagner G (2002) Dynamic interaction of CD2 with the GYF and the SH3 domain of compartmentalized effector molecules. EMBO J 21:5985–5995

Fujiki Y, Matsuzono Y, Matsuzaki T, Fransen M (2006) Import of peroxisomal membrane proteins: the interplay of Pex3p- and Pex19p-mediated interactions. Biochim Biophys Acta 1763:1639–1646

Fujiki Y, Okumoto K, Honsho M, Abe Y (2022) Molecular insights into peroxisome homeostasis and peroxisome biogenesis disorders. Biochim Biophys Acta Mol Cell Res 1869:119330

Henrie A, Hemphill SE, Ruiz-Schultz N, Cushman B, DiStefano MT, Azzariti D, Harrison SM, Rehm HL, Eilbeck K (2018) ClinVar Miner: demonstrating utility of a Web-based tool for viewing and filtering ClinVar data. Hum Mutat 39:1051–1060

Hunter JD (2007) Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95

Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, Gygi MP, Thornock A, Zarraga G, Tam S et al (2021) Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell 184:3022–3040.e28

Jehl P, Manguy J, Shields DC, Higgins DG, Davey NE (2016) ProViz-a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. Nucleic Acids Res 44:W11–5

Johansson-Åkhe I, Mirabello C, Wallner B (2021) Interpeprank: assessment of docked peptide conformations by a deep graph network. Front Bioinform 1:763102

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

Krystkowiak I, Davey NE (2017) SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. Nucleic Acids Res 45:W464–W469

Kumar M, Michael S, Alvarado-Valverde J, Mészáros B, Sámano-Sánchez H, Zeke A, Dobson L, Lazar T, Örd M, Nagpal A et al (2022) The Eukaryotic Linear Motif resource: 2022 release. Nucleic Acids Res 50:D497–D508

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Letunic I, Khedkar S, Bork P (2021) SMART: recent updates, new developments and status in 2020. Nucleic Acids Res 49:D458–D460

Leung AKW, Nagai K, Li J (2011) Structure of the spliceosomal U4 snRNP core domain and its implication for snRNP biogenesis. Nature 473:536–539

Lu R, Yang P, O'Hare P, Misra V (1997) Luman, a new member of the CREB/ATF family, binds to herpes simplex virus VP16-associated host cellular factor. Mol Cell Biol 17:5117–5126

Luck K, Charbonnier S, Travé G (2012) The emerging contribution of sequence context to the specificity of protein interactions mediated by PDZ domains. FEBS Lett 586:2648–2661

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

Machida YJ, Machida Y, Vashisht AA, Wohlschlegel JA, Dutta A (2009) The deubiquitinating enzyme BAP1 regulates cell growth via interaction with HCF-1. J Biol Chem 284:34179–34188

Matsuzaki T, Fujiki Y (2008) The peroxisomal membrane protein import receptor Pex3p is directly transported to peroxisomes by a novel Pex19p- and Pex16p-dependent pathway. J Cell Biol 183:1275–1286

McKinney W (2010) Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference pp 56–61. SciPy

Mishra M, Jiang H, Wei Q (2023) New insights on the differential interaction of sulfiredoxin with members of the peroxiredoxin family revealed by protein-protein docking and experimental studies. Eur J Pharmacol 954:175873

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ et al (2021) Pfam: the protein families database in 2021. Nucleic Acids Res 49:D412–D419

Miyata T, Miyazawa S, Yasunaga T (1979) Two types of amino acid substitutions in protein evolution. J Mol Evol 12:219–236

Mo X, Niu Q, Ivanov AA, Tsang YH, Tang C, Shu C, Li Q, Qian K, Wahafu A, Doyle SP, Cicka D, Yang X, Fan D, Reyna MA, Cooper LAD, Moreno CS, Zhou W, Owonikoko TK, Lonial S, Khuri FR, Du Y, Ramalingam SS, Mills GB, Fu H (2022) Systematic discovery of mutation-directed neo-protein-protein interactions in cancer. Cell 185:1974–1985

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

Mosca R, Céol A, Stein A, Olivella R, Aloy P (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res 42:D374–9

O'Reilly FJ, Graziadei A, Forbrig C, Bremenkamp R, Charles K, Lenz S, Elfmann C, Fischer L, Stülke J, Rappsilber J (2023) Protein complexes in cells by AI-assisted structural proteomics. Mol Syst Biol 19:e11544

ORFeome Collaboration (2016) The ORFeome Collaboration: a genome-scale human ORF-clone resource. Nat Methods 13:191–192

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Müller A, Nothman J, Louppe G et al (2012) Scikit-learn: Machine Learning in Python. arXiv

Persson E, Sonnhammer ELL (2023) InParanoiDB 9: ortholog groups for protein domains and full-length proteins. J Mol Biol 435:168001

Pozzati G, Zhu W, Bassot C, Lamb J, Kundrotas P, Elofsson A (2022) Limits and potential of combined folding and docking. Bioinformatics 38:954–961

Schmidt F, Treiber N, Zocher G, Bjelic S, Steinmetz MO, Kalbacher H, Stehle T, Dodt G (2010) Insights into peroxisome function from the structure of PEX3 in complex with a soluble fragment of PEX19. J Biol Chem 285:25410–25417

Sobti M, Mead BJ, Stewart AG, Igreja C, Christie M (2023) Molecular basis for GIGYF–TNRC6 complex assembly. RNA 29:724–734

Teufel F, Refsgaard JC, Kasimova MA, Deibler K, Madsen CT, Stahlhut C, Grønborg M, Winther O, Madsen D (2023) Deorphanizing peptides using structure prediction. J Chem Inf Model 63:2651–2655

Tompa P, Davey NE, Gibson TJ, Babu MM (2014) A million peptide motifs for the molecular biologist. Mol Cell 55:161–169

Trepte P, Kruse S, Kostova S, Hoffmann S, Buntru A, Tempelmeier A, Secker C, Diez L, Schulz A, Klockmeier K et al (2018) LuTHy: a double-readout bioluminescence-based two-hybrid technology for quantitative mapping of protein-protein interactions in mammalian cells. Mol Syst Biol 14:e8071

Trepte P, Secker C, Choi SG, Olivet J, Ramos ES, Cassonnet P, Golusik S, Zenkner M, Beetz S, Sperling M et al (2021) A quantitative mapping approach to identify direct interactions within complexomes. Preprint at BioRxiv https://doi.org/10.1101/2021.08.25.457734

Trepte P, Secker C, Kostova S, Maseko SB, Choi SG, Blavier J, Minia I, Ramos ES, Cassonnet P, Golusik S et al (2023) AI-guided pipeline for protein-protein interaction drug discovery identifies a SARS-CoV-2 inhibitor. Preprint at BioRxiv https://doi.org/10.1101/2023.06.14.544560

Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O (2022) Harnessing protein folding neural networks for peptide-protein docking. Nat Commun 13:176

Van Roey K, Gibson TJ, Davey NE (2012) Motif switches: decision-making in cell regulation. Curr Opin Struct Biol 22:378–385

Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A et al (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50:D439–D444

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17:261–272

Waskom M (2021) seaborn: statistical data visualization. JOSS 6:3021

Weatheritt RJ, Jehl P, Dinkel H, Gibson TJ (2012) iELM-a web server to explore short linear motif-mediated interactions. Nucleic Acids Res 40:W364–W369

Zhao G, Li K, Li B, Wang Z, Fang Z, Wang X, Zhang Y, Luo T, Zhou Q, Wang L et al (2020) Gene4Denovo: an integrated database and analytic platform for de novo mutations in humans. Nucleic Acids Res 48:D913–D926

Acknowledgements

We thank all members of the Luck, Gibson, and Schueler-Furman labs as well as Julian König and Anton Khmelinskii for helpful discussions and input. We thank Izabella Krystkowiak and Norman Davey for helping us access the SLiMSearch resource with an API. We thank Fridolin Kielisch for advice on statistical analysis as well as the media lab and protein production core facilities of IMB. Support from IMB's IT department and especially help from Christian Dietrich for local installations of AlphaFold is gratefully acknowledged. The GPU cluster on which part of the AlphaFold predictions were performed was funded by the Ministry of Science and Health (MWG), Rhineland Palatinate (funding ID: TB-Nr.:3658/19). We are very thankful for support from EMBL IT Services and the HPC resources for running AlphaFold predictions for this project. This work is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-IDs LU 2568/1-1 and SFB1551 Project No 464588647 awarded to KL. JS acknowledges a PhD stipend from IMB's collaborative research initiative. JKV was supported by the European Union's Horizon 2020 UBIMOTIF programme (860517). This work was supported, in whole or in part, by the Israel Science Foundation, founded by the Israel Academy of Science and Humanities (grant number 301/2021 to OS-F).

Author contributions

Chop Yan Lee: Data curation; Formal analysis; Investigation; Visualization; Methodology; Writing—original draft; Project administration; Writing—review and editing. **Dalmira Hubrich**: Data curation; Formal analysis; Investigation; Visualization; Methodology; Writing—original draft; Writing—review and editing. **Julia K Varga**: Data curation; Formal analysis; Investigation; Visualization; Writing—original draft; Writing—review and editing. **Christian Schäfer**: Data curation; Investigation; Methodology. **Mareen Welzel**: Investigation. **Eric Schumbera**: Methodology. **Milena Djokic**: Data curation. **Joelle M Strom**: Formal analysis; Investigation; Visualization. **Jonas Schönfeld**: Investigation. **Johanna L Geist**: Investigation. **Feyza Polat**: Investigation. **Toby J Gibson**: Resources; Supervision; Writing—review and editing. **Claudia Isabelle Keller Valsecchi**: Supervision; Funding acquisition; Investigation; Writing—review and editing. **Manjeet Kumar**: Resources; Formal analysis; Methodology; Writing—review and editing. **Ora Schueler-Furman**: Conceptualization; Supervision; Funding acquisition; Writing—original draft; Writing—review and editing. **Katja Luck**: Conceptualization; Data curation; Formal analysis; Supervision; Funding acquisition; Investigation; Visualization; Methodology; Writing—original draft; Project administration; Writing—review and editing.

Disclosure and competing interest statement

The authors declare no competing interests.

4.1.1 Supplementary material

Appendix

Systematic discovery of protein interaction interfaces using AlphaFold and experimental validation

Chop Yan Lee[1,†], Dalmira Hubrich[1,†], Julia K. Varga[2,†], Christian Schäfer[1], Mareen Welzel[1], Eric Schumbera[3], Milena Đokić[1], Joelle M. Strom[1], Jonas Schönfeld[1], Johanna L. Geist[1], Feyza Polat[1], Toby J. Gibson[4], Claudia Isabelle Keller Valsecchi[1], Manjeet Kumar[4], Ora Schueler-Furman[2,*], Katja Luck[1,**]

Affiliations
1 Institute of Molecular Biology (IMB) gGmbH, 55128 Mainz, Germany.
2 Department of Microbiology and Molecular Genetics,Institute for Biomedical Research Israel-Canada, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 9112001, Israel.
3 Institute of Molecular Biology (IMB) gGmbH, 55128 Mainz, Germany. Current address: Computational Biology and Data Mining Group Biozentrum I 55128 Mainz, Germany.
4 Structural and Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, 69117, Germany.

*Corresponding author. Tel: +972-2-675-7094, E-mail: ora.furman-schueler@mail.huji.ac.il
**Corresponding author. Tel: +49-(0)6131-3921440, E-mail: k.luck@imb-mainz.de
†These authors contributed equally to this work.

Table of content

Appendix Figure S5. Comparison of AF v2.2 and v2.3 prediction performance.

Appendix Figure S6. Performance of different metrics derived from structural models when benchmarking AF v2.3 for DMI predictions.

Appendix Figure S7. Expression and BRET50 plots for TRIM37-PNKP and ESRRG-PSMC5.

Appendix Figure S8. Structural models, expression, and BRET50 plots for STX1B-FBXO28 and STX1B-VAMP2.

Appendix Figure S9. Structural models, expression, and BRET50 plots for PEX3-PEX19 and PEX3-PEX16.

Appendix Figure S10. Expression and BRET50 plots for SNRPB-GIGYF1.

Appendix Text S1. Summary of observations from the manual inspection of AlphaFold models generated from fragmentation approach on PPIs connecting NDD proteins.

run14: PLP1-MFF

Top prediction involves an ordered region from PLP1 and a disordered fragment from MFF, with a model confidence of 0.75. Looking at the predicted model, the peptide is tilted at an angle to the bundle of helices of PLP1, not like the usual coiled-coil interaction. No trend in increasing confidence with shorter fragments too. The interface does not look very convincing. While the disordered region in MFF is likely to be a functional motif, the 4-helix bundle domain in PLP1 that AF models it to bind to is known to be a transmembrane domain, so the binding site is actually buried inside the membrane. AF is also not very confident about the domain structure, especially for the parts that are at the membrane surface or outside of it. The prediction is likely wrong.

run17: PAX6-CSNK2A1

CSNK2A is a widely active kinase, involved in many processes. Overlapping fragments from PAX6 show trend of increasing confidence the shorter the fragment. CSNK2A1 is predicted to bind with its kinase domain (it doesn't really has anything else than the kinase domain) to a peptide in PAX6 which seems to be a good looking linear motif, i.e. conserved, not part of a folded domain as predicted by AF and predicted by AF to form an alpha helix. The motif though overlaps with a putative NLS. The PAX6 motif is predicted to bind clearly to a pocket that exists in N-lobe of the kinase domain at the bottom of it, away from the catalytic side. Digging deeper, I found a structure, 1JWH, that shows that this is the pocket that is bound by CSNK2B, the regulatory subunit, that interacts with the catalytic subunit to form an active holoenzyme. This, however, does not eliminate the possibility that the AF prediction is right since the peptide looks like a functional motif.

run18: PAX6-SET

Top prediction is ordered-ordered, PAX6 Homeodomain and SET NAP domain. The structure 6PAX shows the PAX domain consisting of two similar folds like the homeodomain bound to DNA but the three-helix bundles are not oriented in exactly the same way like in the homeodomain so I am having a hard time to see where the homeodomain would bind DNA; AF models the homeodomain interface with the NAP domain of SAP via a charged interface with a lot of positively charged residues on the homeodomain contacting a patch of negatively charged residues on the SAP domain. It could be that this patch of positively charged residues on the homeodomain would usually interact with the negatively charged backbone of DNA, but the predicted structure from AF looks interesting since the interface likely does not interfere with SET homodimerization (2E50).

run19: PAX6-TLK2

All predictions with >0.7 model confidence are paired with the Pkinase domain of TLK2 and they are all predicted to bind at the bottom of the beta barrel fold (N-lobe) of the kinase domain. However, almost all peptides come from very different regions in PAX6, no recurrent predictions here.

When looking at the motif pLDDT metric then top predictions also involve two distinct motifs predicted to bind to the long helices in TLK2. However, AF predicts the two helices to form intramolecular contacts. By taking them apart into separate fragments it could be that intramolecular contact sites are now used for interface prediction.

The pair of interactions has a DMI predicted, MOD_GSK3_1 (PAX6 395-402). The peptide PAX6 394-404 was paired with the Pkinase domain but similar to the previous point, it is also put at the beta barrel fold in the N lobe and not the substrate binding site.

run20: PAX6-NGLY1

The PUB domain from Q96IV0, NGLY1, gives good model confidence, >0.8, in binding overlapping disordered fragments of P26367, PAX6. The PUB domain has been solved before alone (2CCQ), the catalytic domain has also been solved bound to RAD23 (2F4M); in the paper that published the PUB domain structure (Allen et al JBC 2006, 10.1074/jbc.M601173200) they also did some mutational analysis to show that there is an interface on the PUB domain that binds the AAA ATPase domain of p97 but the experimental evidence looks not very convincing. Indeed, AF modelled the peptide from PAX6 to bind to an interface adjacent to the one found by Allen et al. There is indeed some hydrophobic pocket and the best 4 predictions comprise that peptide binding to this pocket, however, which hydrophobic residue of the peptide is docked into the pocket varies depending on the length of the peptide; I think that this region in PAX6 could indeed be a linear motif, it is adjacent to the homeobox domain but I don’t think that it is part of the homeobox domain.

run21: PAX6-ESRRG

Many short fragments with high model confidence that are scattered over the disordered region. The binding pocket on ESRRG is in the hormone receptor domain and is a known binding pocket for binding to L..LL motifs (ELMDB: LIG_NRBOX).

According to ELMDB, the first and last L go into a hydrophobic pocket and all fragments of PAX6 with high model confidence have more or less two hydrophobic amino acids with three residues in between: PAX6 319-329: DTA**L**TNT**Y**SA, PAX6 203-213: RLQ**L**KRK**L**QR, PAX6 374-384: PPH**M**QTH**M**NS, PAX6 198-208: DEAQ**M**RLQ**L**K, PAX6 128-148: GADG**M**YDK**L**R.

Looking at structures with ESRRG and two different bound peptides: 1KV6 and 1TFC: NCOA1 686-700: RHKI**L**HRL**L**QEGSPS, 2GPO and 2GPP: NRIP1 378-387: SL**L**LHL**L**KSQ, it furthermore became apparent that the hydrophobic residues right before both Leucines are also important for binding since they contact a hydrophobic patch on the other side of the pocket. However, none of the AlphaFold predicted motifs really fit, it is thus questionable whether they can actually bind the pocket.

Structurally speaking, the peptide does not fit that nicely in the hydrophobic pocket. In 2GPO and 2GPV, there is a triad of hydrophobic residues (V/L/I) making contact with the hydrophobic pocket on the domain but here only 2 residues are making contact. Therefore, it seems doubtful to me that this is a motif that can bind to the domain.

run22: PAX6-QRICH1

Difficult to dig deeper because QRICH1 has only one domain (DUF) that binds to C terminus peptide from PAX6. The high confidence peptide is 20 aa long and seems nice with 0.88 model confidence.

The same DUF is also modelled with 0.76 confidence with a very long disordered region (85 aa) that is at the N terminus of PAX6. However, the predicted complex of this disordered region is quite odd, as it has many twists and turns that seem weird to me.

Overall, these predictions look good but it’s hard to be very certain about it because nothing is known about the domain in QRICH1 and PAX6 has a long disordered C-terminal region full of S, T, but also some Ps and hydrophobics.

run23: PAX6-KCTD7

The top prediction involves the disordered region of PAX6 (198-208) and BTB_2 domain of KCTD7, with 0.74 model confidence. No trend of increasing confidence when fragments shorten. InterPro describes this domain as one that multimerises for its protein function, e.g. KCTD1 as a transcriptional repressor (3DRX, solves KCTD5 that has a similar fold but shorter in length). Since BTB domain mediates the multimerisation of KCTD, it could be that it requires a certain stoichiometry for binding to its partner. In the HuRI database, KCTD7 was indeed detected to interact with itself. The two highest predicted models put both peptides into the same pocket and both peptides have some sequence similarity albeit from different regions in PAX6. These peptides were also predicted with high model confidence in other runs. Based on the structure 5FTA, BTB domains in their homodimerized form do expose the surface predicted in the top prediction. Therefore, the surface predicted to bind to the peptide would be available. Taken together, the prediction looks plausible.

run24: TTC19-FH

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run25: PEX3-PEX16

PEX3 and PEX16 are two proteins that seem to cooperate to help inserting new peroxisome membrane proteins (PMPs) into the peroxisome membrane. They do so via interaction between PEX3 and PEX19. PEX19 brings the PMPs to the peroxisome where PEX3 and PEX16 sit and mediate then further insertion of the cargo (this is described in review Smith and Aitchison 2013 Nature Rev Mol Cell Biol in Fig 2). However, there is also a study that describes how PEX16 localizes to ER and from there traffics to peroxisomes (Kim et al JCB 2006). The structure 6AJB that has been solved for the interaction between PEX3 and PEX19 was published by Sato et al EMBO J 2010 and describes how an N-terminal SLiM in PEX19 binds to the domain of PEX3. They tried to crystalize the whole protein of PEX3 but only observed residues 52-368. The domain has the exact same fold as predicted by AF. The predicted cytosolic and peroxisomal localization of protein regions and the two TM helices that are shown in Uniprot seem to be wrong for PEX3 according to work cited in Sato et al. They summarize that the N-terminal region of PEX3 contains a targeting signal or anchor for PEX3 to the peroxisomal membrane followed by the domain that is located in the cytosol. No structure has been solved yet for PEX16 but it seems likely that the prediction of two TM helices that are shown in UniProt in this protein is also wrong. AF predicts a globular domain containing the two TM helices and has a nicely exposed loop that carries the putative SLiM that AF predicted to bind to PEX3. It binds onto PEX3 on the opposite side to PEX19 binding, so PEX19 and PEX16 could bind simultaneously to PEX3. Further work on these interactions can be done by submitting the three protein sequences to AF to see what it does. Some other study observed interaction between PEX3 and PEX16 according to Uniprot but the interface really does not seem to have been looked at before nor the interaction studied in detail. All the fragments that contain the putative SLiM in PEX16 are predicted in the exact same way to bind to PEX3; always anchored via a conserved region sitting in PEX16 between residue 160 and 190. Interestingly, the most conserved residues are also those that seem most important for binding. This smells really good.

run26: PEX3-PEX19

This is a positive control interaction since the structure has been solved for this PPI (3AJB) and it is a well known and well studied PPI with an entry for it in the ELM DB: LIG_Pex3_1 (L..LL...L..F). This ELM instance is indeed predicted by AF to be the highest model confidence. Another peptide from PEX19 121-141, FTSCLKETLSGL, scored equally high model confidence. It could be that this other predicted binding site is also true but I believe that it is rather an artefact from AF's insensitivity to mutations.

run27: GABARAPL2-UBA5

The structure 6H8C shows binding of GABARAPL2 domain to LIR motif in UBA5. This motif is not listed on the ELM website for LIG_LIR_Gen_1 because it does not quite fit the regular expression which seems to be defined too narrowly. AlphaFold correctly predicts this interface but only as third highest based on model confidence just hitting the cutoff of 0.7 while using chainB_inf_avg_plddt it scores as fourth best prediction far below the cutoff (67). However, AF recurrently finds peptides including the motif following each other when ranked by model confidence or pLDDT. The top three motifs predicted to bind to GABARAPL2 are not finding the hydrophobic pocket that is filled by a key big hydrophobic residue in the motif and these peptides are also not recurrently predicted. So, I think these are wrong predictions.

run28: GABARAPL2-LZTR1

GABARAPL2 (P60520) has Atg8 domain that is known to bind motifs (LIG_LIR_Gen_1). The domain is modelled with high confidence to bind to different disordered fragments of interacting partner LZTR1 (Q8N653). The second top confident model (when ranked by model confidence) has an aromatic residue tucked into a deep pocket and a branched alipathic residue tucked into another shallow pocket. The top confident model has some kind of increasing trend in model confidence as fragments get shorter, with the shortest one getting the highest confidence. The highest confidence model has a nice increasing model confidence trend but it does not have an aromatic residue fitting into the deep pocket as it is known for LIG_LIR motifs.

Looking at the structure 2LUE, the second top model LZTR1 46-52 GP**F**ET**V**H looks more similar in sidechain positioning compared to 2LUE. Residues highlighted in bold get tucked into the mentioned pockets. This model seems more likely to be true than the best model. However, it also is predicted to bind in reverse order compared to structure 3WIM.

run29: CUL3-KCTD7

Has an ordered-ordered prediction with quite high confidence (0.66) but the contact interface is a tetramerization domain from KCTD7. Therefore it seems unlikely that it is a functional interface.

Two N terminus disordered fragments from KCTD7 with > 0.7 model confidence when paired with the Cullin domain of CUL3. These two fragments are modelled to be binding at the same site of Cullin domain (the site where RING proteins bind to, 1LDJ). In the case of 1LDJ, the RING protein has a long disordered region inserted into the Cullin domain of CUL1, burying a series of hydrophobic residues in the long disordered region. However, the same binding site of the Cullin domain of CUL3 is a bit different, with more surface exposed than CUL1. In this case, the contacts modelled in KCTD7 16-26, with a triple Serine making contact with the Cullin domain, look plausible. The other high confidence peptide KCTD7 1-11, with triple Valine making contact with the Cullin domain, also looks plausible to me.

In the structure of 1LDJ it is really amazing how the partner protein interacts with CUL1 via beta-sheet augmentation but how this extra beta strand becomes part of the integral fold, it is kind of in the middle of the domain. I think AlphaFold feels that there is something missing and is trying to put a peptide there but the overall conformation of the domain is also different at places so that the predicted peptide does not sit at the same position like the one shown in 1LDJ. AlphaFold predicts two different motifs of very different sequence from the N-terminus of KCTD7 to bind there. Given how different the sequences are, this adds another negative point towards questioning the specificity of these predictions.

run30: PNKP-SYP

Top prediction is a disordered fragment from SYP (7-19) paired with the kinase domain of PNKP. The binding surface is different from the nucleotide binding surface (1RC8). This binding interface looks plausible. It was later found that the kinase and phosphatase domain form a structural unit based on published structures. The run is modified to use the kinase and phosphatase domain as an ordered region for prediction with disordered fragments of SYP.

The rerun with a fragment comprising the phosphatase and kinase domain now resulted in one prediction that makes the cutoff. This prediction puts a motif from SYP into the DNA binding pocket of the kinase domain (according to Bernstein et al Mol Cell 2005, 1RC8).

There is another predicting docking a peptide from SYP into the FHA domain of PNKP. It puts it where FHA domains bind their phosphorylated peptides but the SYP peptide has no Ser or Thr.

run31: PNKP-TRIM37

The first prediction involving the combined kinase-phosphatase structure puts a peptide of TRIM37 into the binding pocket where the phosphatase domain would bind single stranded DNA.

Following up is a prediction that involves a disordered region in PNKP binding to the surface of MATH domain of TRIM37 where MATH domain-binding peptides generally bind to. The PNKP peptide differs slightly in sequence from regular expression patterns described for MATH domains in the ELM database. This peptide in PNKP has a known phosphorylation site that stabilizes PNKP protein levels, making the peptide very interesting since this suggests a regulatory role of phosphorylation on the peptide.

There is a second peptide of PNKP predicted to bind to the MATH domain also with high confidence but the sequence is quite different from the first one and very close to the phosphatase domain. There is also a prediction where the FHA domain of PNKP is predicted to bind to a peptide of TRIM37 but the peptide looks very different from known FHA-binding motifs (peptide with phosphorylated threonines), which is of course difficult to predict for AF.

run32: PNKP-XRCC4

XRCC4 and PNKP prediction, there is a peptide from XRCC4 that binds to the phosphatase domain with high confidence. But then I am not sure if this is right because it could be a false prediction of a small peptide easily fitting into the catalytic site of the phosphatase domain. There is a Serine in the peptide, so it is possible that this is where the phosphate group gets cleaved off by the phosphatase domain. After checking more, it is found that XRCC4 is known to bind to PNKP via a phosphorylated motif that binds to the FHA domain in PNKP.

In principle, it would be better to make a rerun where the kinase and phosphatase domain are taken as one fragment since they form 1 structural unit but I think in this case it would not have changed anything. The best prediction put a peptide from XRCC4 into the pocket of the phosphatase domain where it would bind the single-stranded DNA as seen in 3U7G. Among the first 9 predictions AF put 7 different peptides from XRCC4 into the phosphatase, the others go to the kinase domain. The first prediction that involves the FHA domain of PNKP and contains the FHA-binding motif in the sequence fragment of XRCC4 has a confidence score of 60 and does not put the FHA-binding motif in the pocket but another negatively charged peptide in the sequence (the FHA pocket is very positively charged). The correct prediction where AF puts the FHA-binding motif in the right pocket has a confidence score of 0.58.

run33: TNPO3-GCH1

Top prediction involves the disordered region of GCH1 (16-26) and the superhelical structure of TNPO3, with model confidence 0.71. Since TNPO3 (transportin) is known to transport cargo into the nucleus by releasing the cargo via the competitive binding of GTP-bound Ran (2X19), the peptides from GCH1 are modelled to be at a binding site near where Ran binds in 2X19. It is therefore biologically sound where the peptides are modelled at. The binding site of the peptides from GCH1 is also lined with many arginines, making it very positively charged. The contact modelled by AF in the top prediction looks good, with many charge-charge interactions at the interface. The N terminus of GCH1 has many prolines that are conserved, with three repeats of PAEK or PEAK and two repeats of PPRP.

run34: TNPO3-CAMK2G

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run35: GNAI3-GPSM2

This interaction has been structurally solved (4G5S) and AlphaFold predicted the interface 100% accurately. GPSM2 has multiple GoLoco motifs that AlphaFold predicts individually with high confidence to bind in the pocket on GNAI3.

run36: SYT1-MIP

Both are transmembrane proteins. The top prediction involves the linker between two C2 domains of SYT1 and the MIP domain of MIP. MIP domain is also known as aquaporin domain (transmembrane). However, when the linker is fragmented, it receives lower confidence. I think this is unlikely to be the interaction interface. The linker could be a motif for some other interaction because of its moderately high plddt. There is a structure of a homodimer of SYT1, 2R83, that shows that both C2 domains of one chain are actually interacting with each other and that the linker between both domains interacts with one domain. It is this linker where AF predicts that a peptide would bind to the porin domain of MIP; interestingly, AF predicts the two C2 domains to be independent from each other in the monomeric structure of SYT1, so either AF is wrong or crystallization introduced the packing of both domains against each other but I would rather believe the Xray structure and in this case the peptide would not be accessible to bind to the MIP domain.

run37: FTSJ1-CERT1

FTSJ domain of FTSJ1 is known to bind S-ADENOSYLMETHIONINE (see structure 1EJ0). The top predictions all look very different in that different regions or partially overlapping regions of CERT1 are docked into different sites onto the FTSJ domain. Sometimes the peptide is docked into the catalytic pocket where the protein methylates adenosines on tRNAs but the peptide is also docked elsewhere. Because of these ambiguities, I believe that the predictions are questionable since they seem to lack specificity but I don't think we can call them definitely wrong.

Another interface was found involving CERT1 368-388, with model confidence 0.70. However, the contacts modelled are mostly backbone-to-backbone. I have previously noticed that AF tends to give higher confidence to complex modelled with secondary structure. So I think this is also a likely false interface.

run38: CAMK2A-SOX5

Kinase domain of CAMK2A with a disordered fragment predicted showed high confidence. The structure predicted by the highest confidence model is weird, with both beta sheet and helix structure.

Kinase domain of CAMK2A is likely serine threonine kinase and in kinase domain prediction, one has to be careful with the two lobes that bind substrate and ATP. It might be interesting to check other high scoring peptide to see if they have S/T that can be phosphorylated and check the crystal structure to find substrate binding pocket. The first two highest scoring peptides do not look convincing because the first one has no S/T in the peptide but it is fit into the catalytic cleft while the second one has positioned the sidechain of a T out of the cleft. The third highest scoring peptide (P35711 131-141) looks nice because it positions the sidechain of an S into the catalytic cleft.

The highest ranking peptides are essentially all over the place from SOX5 and I don't think that AF can predict very well kinase-substrate interactions. Overall, the high-scoring predictions all do not look very convincing.

run39: CAMK2A-CAMK2G

Many high confidence predictions involving different regions in the protein pair. Among them, one ordered-ordered interface gives a really high confidence. The interface is a known DDI in 3did with high zscore. The structure 3SOA only shows one CAMK2A monomer but the publication talks about a dodecamer for which one can download a model from the PDB as well. Looking at this dodecamer and the paper, it becomes clear that downstream of the kinase domain there is another domain referred to as hub domain in the paper which mediates oligomerization, together with the linker between the kinase and hub domain. The best AF prediction for the interaction between CAMK2A and CAMK2G involves both hub domains and is an accurate prediction of the interface seen in the dodecamer.

The second best prediction made by AF involves the hub domain and a bit of the linker sequence from the other partner. Looking at the dodecamer, one can see that where the peptide is predicted to bind on the hub domain is part of the linker sequence bound from the same monomer, so an intra-molecular interaction. So, there is indeed some binding site but not for inter-molecular interaction. Because the linker sequences are different in the structure and canonical uniprot sequence it is very difficult to know which part of the linker is binding on the hub domain and whether this corresponds to the bit of the linker sequence predicted by AF to bind there. In the paper accompanying the 3SOA structure they also investigate how different linker sequences from different isoforms influence Ca-binding site accessibility and thus activation of the complex. There is evidence from 3 other studies that CAMK2G and CAMK2A interact with each other from co-IP experiments but these were large-scale studies. It is likely that no one has studied the interface between CAMK2G and CAMK2A and thus would be something new.

run40: ACTB-ACTG1

Two actin proteins are predicted to have high confidence DDI. The interface itself that is predicted by AlphaFold looks very interesting, it indeed looks like a polymerization interface because both domains interact with opposite sites. interactome3D would model this interaction with the structure 4JHD as a template but this one looks quite different, it's not the same interface and needs according to the authors a third protein for polymerization. Digging deeper in PDB for structures of ACTB, I found structure 6ANU which shows the same interface that AF predicted between ACTB and ACTG1, so the interface is probably right.

This is also a very interesting case. Based on the review by Vedula and Kashina (J of Cell Signal 2018, 10.1242/jcs.215509), it is still an open question whether the different actin forms that exist in human can form heteropolymers or not. Some studies find this in vitro, other find intermingled homopolymers of beta and gamma actin. Both actins co-occur in many cell types while alpha-actin is more specifically expressed in muscle. It seems really tricky to solve this since actins are highly studied and actins are also super similar in their sequence, so it could be that in a somewhat artificial system, beta and gamma actin can interact because the interface residues are identical but in vivo they would rather not interact and rather form homopolymers. In the end, whether ACTB and ACTG1 indeed interact in vivo is the only open question.

run41: RARB-PSMC5

PSMC5 has been repeatedly modelled by AF to have a high confidence peptide that binds to partners with Hormone_recep domain. The peptide is 132-141 DP**L**VS**LM**MVE. Residues highlighted in bold are the ones tucked into the hydrophobic pocket. However, this peptide does not match with the consensus of LIG_NRBox (^PL..LL^P), especially in this peptide P precedes the first L. I am not sure why P is disallowed at first position as ELM has not described much about the sequence composition of the motif. I think it might be too early to reject this peptide because the highlighted residues are indeed hydrophobic and can serve similar functions as those in the regex.

I looked at the HuRI network of PSMC5 too, and found that the interactors seem to be enriched with the Hormone_recep domain, making this interface even more plausible.

run42: DCX-BICD2

DCX has two DCX domains and all good predictions involve the N terminus DCX domain. The N terminus DCX domain is known to bind Tubulin. AF modelled a different interface on the N terminus DCX domain to bind to disordered fragments from BICD2.

The DCX domains have a C-terminal part that is not confidently predicted by AF to be part of the fold. When excluding this part from the first DCX domain, AF models peptides to bind to the area where this last part is predicted to be located in the monomeric structure from AF. When we use a DCX domain that contains this last bit, then AF predicts other peptides from BICD2 to bind on the opposite side of DCX. There is no consistency in these predictions.

There are no other predictions between ordered-ordered or disordered fragments binding to ordered domains in BICD2 that make the cutoffs. BICD2 however, also only consists of large helices. Nonetheless, it could be that both DCX domains together bind to one of these coiled coil helices in BICD2.

run43: DCX-ZBTB10

A possible prediction involves the first DCX domain of DCX and a peptide of ZBTB10 261-271. This prediction is not influenced by the actual domain boundaries because the peptide is not docked into the pocket where a region a little C-terminal of the domain might bind to. This is the case for the second best prediction involving the first DCX domain and peptide 604-614. According to chainB inf avg plddt these are the only two prediction that make the cutoff when looking at chainB as a disordered region. ZBTB10 has a lot of disorder and probably many motifs. DCX has two DCX domains and a bit of disorder. Looking into available PDB structures then the DCX domains are known to bind to microtubules. There is one structure with the first DCX domain bound to microtubules (6RFD). It seems though that the pocket where ZBTB10 261-271 is predicted to bind is not occupied in this complex. AF does not predict slightly extended versions of this peptide with reasonable confidence to bind to this pocket.

A peptide was also predicted to bind in beta-sheet augmentation to the last beta strand of the BTB domain with reasonable model confidence and chainA_intf_avg_plddt scores but the ZBTB10 model might have its own beta strand C-terminal of the current domain boundaries that AF predicted to complement the last beta strand of the domain as predicted in the full length model of ZBTB10.

AF also predicts a contact between the ZnF domain of ZBTB10 and the first DCX domain but it does not look very likely and I think the ZnF fold is perturbed.

run44: PSMC5-ESRRG

The interaction has quite some high confidence predictions. The highest scoring peptide is P62195, PSMC5, 132-141, DP**L**VS**LM**MVE. The three hydrophobic residues make nice contacts with the hydrophobic pocket and surface of the domain. Another disordered fragment from PSMC5 binding to the same domain, IKKLWK, also looks promising. However, there is some possibility that these are artefacts because AF is not very specific when it comes to detecting single mutation in known motifs. The sequence alignments are not helpful unfortunately because the whole PSMC5 is super conserved.

Nonetheless, interaction between PSMC5 and ESRRG looks promising because the alternative name is thyroid hormone receptor-interacting protein 1, TRIP1.

run45: PSMC5-RORB

The highest confidence prediction involves a disordered fragment from PSMC5 and it is the same as run44. The ordered region from RORB is the same domain, hormone receptor domain, as run44.

It is interesting to see AF predicting similar DMI with high confidence from two different proteins. Same observation as run44.

run46: WAC-NFE2L2

WAC and NFE2L2 are largely disordered. WAC has a WW domain. AF predicts recurrently a sequence close to the N-terminus of NFE2L2 to bind to the WW domain that are known to bind proline-rich motifs. The putative motif in NFE2L2 does not contain prolines and is not docked onto the WW domain in any way like other WW domains, e.g. 1EG4. These are likely wrong predictions. While the motif interface pLDDT is reasonably high for these predictions, the model confidence does not reach the 0.6. There are no other predictions that make the cutoff.

run47: WAC-MOBP

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run48: STX1B-FBXO28

The top model has 0.76 model confidence that utilizes the disordered region 1-22 of STX1B and Fbox + helix bundle domain (63-221) of FBXO28. The interface involves the disordered region of STX1B forming a 310 helix structure with the helices from the Fbox domain. Note that the Fbox domain annotated by InterPro is from 61-109, while the ordered region that I used for prediction is 63-221. The Fbox domain is known to mediate PPI but it is not used by AF to model the interaction in this prediction. Region 1-22 of STX1B is conserved only in recent homologs. The plddt of the disordered region is low, <60 for all residues.

The second top model has 0.75 model confidence that involves the syntaxin domain (23-237) of STX1B and disordered region 354-368 of FBXO28. The disordered region of FBXO28 is at the C terminus and conserved. However, the plddt of the peptide is low and adopts a 310 helix kind of structure. A slightly different prediction involving fragments of the proteins (27-219 STX1B and 345-363 FBXO28) returned 0.73 model confidence. The peptide adopts a helical structure but is placed on a different surface of the syntaxin domain. Although the peptide 345-363 has good plddt (mosty >60), I am not sure if this is the right interface. One prediction pairs the full length of STX1B with the disordered region 354-368 of FBXO28 and returned 0.71 model confidence. The interface is similar to that of the syntaxin domain (23-237) of STX1B and disordered region 354-368 of FBXO28 with low plddt. This region 354-368 in FBXO28 could be an nuclear localization signal (NLS), where ELMDB also predicts quite a few NLS, and therefore unlikely to be the interface for the interaction.

Next top prediction has 0.749 model confidence that involves the C terminus of the syntaxin domain (220-232) of STX1B as disordered region and the Fbox + helices domain of FBXO28 (63-221). The interface is formed by the peptide adopting a helical structure with the Fbox + helices domain. The plddt of the peptide is good, with all residues above 60 plddt. Nonetheless, another prediction involving slightly longer peptide from the same region of STX1B has a much lower model confidence (0.55). The interface modelled is not exactly the same as it is a little bit shifted. Unsure if this is a good interface.

I tried to find more molecular studies on the two proteins but I can’t find much. STX1B is known to function in docking of synaptic vesicles at presynaptic active zones while FBXO28 probably recognizes and binds to some phosphorylated proteins and promotes their ubiquitination and degradation. Weirdly, STX1B is known to localize to membrane while FBXO28 has not much information on subcellular localization but studies have shown that it interacts with topoisomerase using its Fbox domain (the bundled helices are not needed for interaction). Out of all the predictions, I think STX1B 27-219 + FBXO28 345-363 and STX1B 220-232 + FBXO28 63-221 are most likely to be the interface, as their peptides are modelled with good plddt and both achieved model confidence higher than 0.7.

run49: STX1B-MMGT1

Top prediction involves the Syntaxin domain of STX1B and the disordered region of MMGT1 (23-31) with confidence 0.73. A slightly longer fragment has a slightly lower confidence but looking at the structure, the two peptides have different angles to the Syntaxin helical bundle. Since the interfaces modelled by AF differ a lot despite using the same peptide and its extended counterpart, the modelled interfaces do not look genuine.

run50: STX1B-VAMP2

Interactome3D models an interface between both proteins based on the structure 3HD7/3IDP where STX1A interacts with VAMP2. STX1A and STX1B are very similar in structure.

STX1B is predicted in closed conformation, which we know because structures exist of STX1A bound to Munc18 where it is in this closed conformation with the long C-terminal helix comprising the SNARE domain folding back onto the syntaxin domain. However, when bound to VAMP2 we can see the open conformation where the long helix is made available to bind in coiled-coil like manner to VAMP2 and SNAP25 helices.

Based on this available structural information we designed different fragments of the extended SNARE domain of variable length. VAMP2 is a short protein of 116 residues consisting of a long helix and about 30 disordered residues at the N-terminus. The most confident predictions obtained for these fragments is the one modeling a coiled-coil interaction between the extended SNARE domain and the helix of VAMP2 but the model confidence is slightly below the cutoff. Predictions with the disordered N-terminal region of VAMP2 remain far below cutoffs.

run51: CSNK2A1-CSNK2B

Nice prediction with overlapping fragments showing increasing model confidence. This interface has been solved before in two structures: 4DGL and 6Q38; prediction is highly accurate, and is probably a DMI that is not in ELM yet.

run52: EBF3-EBF2

Dimerization of the EBF family already known and solved (3MUJ). AF predicts the middle domain of both proteins called TIG as the dimerization interface as top prediction but in head to tail orientation while the structure 3MUJ shows head to head orientation. Followed closely up in terms of score (avg_intf_plddt) is the fragment comprising the TIG domain and the helix loop helix domain which are predicted accurately as seen in the structure.

The third best prediction involves the N-terminal DNA binding domain as the dimerization interface. Does not look so convincing to me but still got a very high score. The fourth best prediction is the helix-loop-helix domain alone as dimerization interface, still with a score of 90. There are more predictions that make the cutoff that involve various disordered regions of either protein and ordered fragments from the other involving interfaces used for dimerization but I guess that these predictions are likely wrong.

run53: PEX12-TREX1

The disordered region of PEX12 215-312 (98 residues long) is predicted with high confidence. One fragment of it achieved even higher confidence but when this fragment is further fragmentate, their confidence is not as high anymore. After checking the protein on InterPro, this domain is the exonuclease domain of TREX1 that binds to ssDNA (2OA8). In this crystal structure, it shows the pocket modelled by AF to bind PEX12 215-312 is bound to a ssDNA, with the phosphodiester bond of ssDNA making interactions with the backbone of the domain chain and some hydrophobic side chain (leucine) making hydrophobic interaction with the base of the nucleotide. Interestingly, AF seems to have memorized this crystal structure because the bound ssDNA has a curved structure and AF also models the long disordered region to have an odd curve. I think this interface is unlikely to be true because the bound magnesium ions coordinate with the oxygen in the phosphodiester bond of ssDNA and the modelled helix places hydrophobic sidechains to the cavity where magnesium ions bind.

A very short fragment of PEX12 12-16 at the N terminus is modelled with high confidence with a very negatively charged pocket in the domain of TREX1. It is unusual to have a peptide binding pocket with such a high negative charge. Further checking revealed that this domain binds magnesium ion and nucleotides. The short fragment fits into the magnesium binding pocket and thus this is unlikely to be true.

run54: PRKAR1A-PRKAR1B
Best model is an ordered-ordered prediction with 0.83 confidence. It is a homo-DDI (RIIa domain) dimerization and has been solved in 2EZW.

An additional disordered fragment (PRKAR1A 360-372) predicted with high model confidence but low pLDDT with the cyclic nucleotide binding domain of PRKAR1B. Referencing available structure of cNMP binding domain (1NE4), there are two beta barrel folds in the domain that bind to cyclic nucleotides. AF fits the disordered fragment on a hydrophobic surface near the beta barrel but not in the cNMP binding pocket. Although this could be another binding site, the binding makes little sense to me because the disordered fragment is at the C terminus of cNMP binding domain of PRKAR1A, meaning that the sequence would have to loop back to make this contact. In the previous bullet point, it seems very likely that the dimerization of the two proteins are mediated by the RIIa domain (N terminus), so it seems not so plausible to me that at the C terminus they make contact again. This is likely a false positive interface.

run55: ASF1A-H4C8
The interaction between both proteins has been solved (5C3I). However, this structure shows that the motif in H4 sits at the very C-terminus and binds in beta sheet augmentation to ASF1A in the same pocket like AF predicted but using an N-terminal peptide of H4. I think the problem is that the C-terminal region of H4 was made part of the domain of H4, which I agree was hard to see from looking at the monomeric AF structure for full length H4; I checked further down in the predicted structures but the first ordered-ordered prediction has a model confidence of 0.25 and does not find this mode of binding either. One could rerun this by taking the C-terminal peptide of H4 as disordered region just to see whether AF would then get it right but in principle this is a false positive prediction; the N-terminal peptide also shares no sequence similarity with the C-terminal motif.

run56: RARS1-CCDC115
There is only one prediction that makes the cutoff for model confidence or/and motif pLDDT. This prediction involves RARS1 1-21 as a disordered fragment that is modelled to bind as a helix to the two helix coiled-coil domain of CCDC115. A shorter fragment of the motif is placed elsewhere. The helix of CCDC115 to which the peptide is predicted to bind has more hydrophobic residues along the helix on that side so I would think that a longer partner chain would be able to bind there. Thus, this interface does not seem likely to be true.

run57: UBE3A-TAT
Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run58: VAMP4-MFF

Top prediction is two ordered regions that are both helical. Both proteins have only helical regions and the rest are disordered. Interestingly, despite the top predicted interface having only 0.71 model confidence, both chains have very high plddt for their residues at the interface (95 for VAMP4 and 90 for MFF). Because of their high plddt, it could be a genuine interface. The helix in VAMP4 definitely has an interface there because one side is rather hydrophobic while the other side is rather hydrophilic. MFF could bind there with its helix or via another helix that it has. The binding does not show that many nice contacts, i.e. some hydrophobic residues on the VAMP4 helix still remain exposed.

run59: PEX16-MMGT1

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run60: PLP1-SLC16A2

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run62: SNRPB-GIGYF1

GIGYF1 is a very long protein with many disordered regions. It has a GYF domain that is known to bind proline-rich sequences. SNRPB has many proline-rich sequences in its C terminus. Some proline-rich motifs are predicted with high pLDDT to bind the GYF domain (these are the top predictions).

Another highly ranked prediction involves the LSM domain of SNRPB with various disordered fragments from GIGYF1. However, checking InterPro entry as well as structures showing LSM domain, it seems like LSM domain is predominantly involved in multimerization with other SNRP proteins to form the SMN complex involved in splicing (1H64). Therefore, the models involving this domain with disordered fragments look unlikely to be true to me.

Digging deeper into the top predictions, comparing the binding modelled by AF between SNRPB 231-240 and GYF domain of GIGYF1 with 1L2Z, the peptide is oriented differently. However, from 3FMA, one can see different ways a peptide binds to the same surface of GYF. In 3FMA chain E and P show a similar way of binding to that modelled by AF. The peptide sequence in 3FMA is also different from 1L2Z, but importantly, there are three prolines in the peptide that always orient the same to the hydrophobic surface formed by the GYF motif on the GYF domain. This orientation of the 3 prolines is captured by AF.

AlphaFold repeatedly predicts the PPGM motif in the same pocket. This motif occurs multiple times in the C-ter tail of SNRPB. On the ELM website, the LIG_GYF motif is described to bind proline-rich sequences and they also cite the structure 1L2Z but they say that flanking positively charged residues seem to be important for binding to the GYF domain. Indeed, in the crystal structure there are some negatively charged residues on the GYF domain. Interestingly, the GYF domain from GIGYF1 does not or only partially has those. It also differs in that it has a deeper hydrophobic pocket which is filled with a Trp in the crystal structure. So, it could well be that the GYF domain from GIGYF1 binds somewhat different proline-rich peptides. The interaction between GIGYF1 and SNRPB has not been described before other than in HuRI. Functionally, it would be probably a new connection because GIGYF1 is not known to function in splicing as far as I can see and thought to be localized to the cytoplasm. GIGYF1 however, has also interacted with SNRPA and SNRPC in HuRI. They also have 1 or some more occurrences of the PPGM motif. If this mode of binding is true then it would be somewhat of a new mode of binding or in the most conservative case an extension of the known binding mode of LIG_GYF.

Alignment of 1L2Z chain A (GYF domain) with the GYF domain from GIGYF1 (476-535) shows that the sequences are not very conserved. Structural superimposition of the two GYF domains reveal that the overall fold is conserved, including the majority of the binding pocket except for the hydrophobic pocket filled with a W. The peptides of the two structures have their PPPG in similar orientation. Following this sequence is a M from SNRPB that is tucked into the hydrophobic pocket and H for 1L2Z that is exposed to the environment. The sequence that follows is R for both, with the one in SNRPB exposed to the environment and possibly forming a hydrogen bond with the Q on the domain, and that in CD2 (1L2Z) forming salt bridge with an E from the domain.

Later a structure of the GYF domain of GIGYF1 was published binding to a similar motif found in TNRC6 further supporting the correctness of these predictions.

run63: ARHGEF9-VEZF1

Top prediction has 0.74 model confidence with the fragment from VEZF1 (375-385) making contact with the RhoGEF domain of ARHGEF9. The top predictions all put the peptide at the same binding site of the RhoGEF domain. In terms of conservation, all the peptides from VEZF1 are well conserved. Nonetheless, the prediction looks like a very questionable one, at least it seems like the predictions do not make use of the GTP/GDP binding pocket for which I did not find a structure that shows where it precisely is located but based on an abstract of an article and InterPro entries it seems to be between both structural entities that form one larger domain, the GEF domain and the PH domain (IPR000219). There is absolutely no consistency in the two peptides from VEZF1 selected to bind to the same surface on the GEF domain of ARHGEF9; VEZF1 also seems to be of very weird type, AF has a hard time to make sense out of this protein.

run64: MIP-MFF

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run65: VEZF1-PRKAR1B

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run66: VEZF1-KCTD7

Top prediction involves the disordered region of VEZF1 (360-380) and the BTB domain of KCTD7. The disordered region overlaps with the top prediction in run63 that models the interface between VEZF1 (375-385) and the RhoGEF domain of ARHGEF9. Despite AF modelling a 310 helix structure in the disordered region of VEZF1 (360-380), the contacts modelled at the interface do not look very convincing. It could be that the disordered region (360-385) is a functional motif for other interactions and AF detects that and tries to fit it into the domain. It could also be that, to form the binding interface, it needs multiple copies of BTB domain, which is not used in this prediction. The VEZF1 peptide is put in the same pocket like the PAX6 peptides from run23 but the sequences look different, it is however the same peptide in VEZF1 like in the prediction with ARHGEF9.

run67: APTX-FLAD1

Has overlapping fragments with increasing confidence: APTX, N terminus disordered region 5-12 and 6-13, paired with MoCF_biosynth or a domain of unknown type (not matched to a Pfam or SMART domain) that is between the MoCF and PAPS_reduct domain of FLAD1. It also predicts the same N-terminal region of APTX into the PAPS_reduct domain. The disordered fragments from the region 8-15 of APTX showed high confidence model confidence but below the cutoff pLDDT score when modelled with the PAPS_reduct domain of FLAD1. Checking the structure of PAPS_reduct domain in complex with adenosine phosphosulfate shows that the peptide is modelled by AF to be in the binding pocket of adenosine phosphosulfate. This is likely a false prediction.

For the N-terminal part of APTX AF is quite confident when it models it into the MoCF domain or the other unknown domain of FLAD1. There are multiple predictions with different overlapping fragments that make the cutoff. However, AF is more confident with both metrics when the peptide is modelled into the MoCF domain. This domain has a pretty substantial pocket that is actually in the monomeric structure of FLAD1 occupied by another region of FLAD1 with low pLDDT. However, when APTX 10-15 is used for modelling, the orientation of the peptide is reversed. MoCF_biosynth domain is known to trimerize for its activity and is known to bind molybdopterin. MoCF_biosynth binds molybdopterin on a site close to where AF models the peptide to be (refer to 1DI6, https://doi.org/10.1074/jbc.275.3.1814 that solves the structure of a bacterial protein with the same domain. They mentioned 49D and 82D to be important for catalytic activity)

APTX with the unknown domain of FLAD1 does not reach the model confidence cutoff, only the motif pLDDT cutoff. It puts the same peptide as beta-sheet augmentation to the domain while in the predictions for the MoCF domain, the peptide is put in helical conformation.

The only predictions where disordered regions in FLAD1 are predicted to bind to folded regions in APTX involve the FHA domain of APTX and correspond to two completely different disordered regions in FLAD1.

run68: FBXO28-PSMC3

Top prediction is coiled-coil interaction between regions from the two proteins that are modelled by AF monomer as long helices. The plddt of all residues are very high. This interaction looks convincing. The only problem is that one helix is shorter than the other, while for a common coiled coil interaction, both helices are usually equally long.

The second best prediction based on model confidence involves a disordered region from FBXO28 (51-61). The modelled complex does not look convincing because the peptide is quite hydrophobic and the residues do not make much contact with the domain. The peptide is predicted to bind to the first domain of PSMC3 which as far as I was able to find, does not have catalytic activity.

There are only these two predictions that make the cutoff for model confidence, none make the cutoff when looking for disordered regions in PSMC3 predicted to bind to FBXO28. The other way round there is the peptide mentioned above and a C-terminal disordered region of FBXO28 predicted to bind to the same first domain in PSMC3 but predicted to bind to a different side. The C-terminus of FBXO28 is very charged, maybe a localization signal. Both motifs in FBXO28 are somewhat recurrently predicted to bind to the domain in PSMC3.

run69: CAMK2G-ESRRG

Many high confidence predictions in a disordered region of CAMK2G. The whole disordered region used as a fragment for prediction also returned high confidence (0.78). In this long disordered region, AF puts the third highest model confidence peptide in the domain pocket. The top three highest confidences are very similar in terms of confidence. The motif detected by AF resembles LIG_NRBOX with the motif L..LL. CAMK2G 300-310: LKGAI**L**TT**ML**V -> looks plausible to me because the M is hydrophobic and it is possible to substitute for the role of L in the regex. CAMK2G 315-325: SA**A**KS**LL**NKKS -> Also possible but the A is fitted into a quite deep hydrophobic pocket where known structure (refer to run21) shows that it is L that gets fit into the pocket. A might have too short of a hydrophobic side chain to make a good contact with the deep pocket. CAMK2G 355-365: QEPAP**L**QTAME -> not so good IMO because the hydrophobic contact is less extensive as the peptide found above. Another interesting observation: CAMK2G 285-423 (139 aa) prediction resulted in 0.78 model confidence, which is very high for a disordered region that long. In this case, **CAMK2G 300-310** is fitted into the hydrophobic pocket, adding weight to the fact that this could be the correct peptide. This reminds me of the extension analysis with DMI where extension of motif can improve prediction results.

A pairing of ordered-ordered region prediction returned high confidence (0.83). This involves Zn finger from ESRRG and CaMKII association domain at the C terminus of CAMK2G. The binding is close to but not in the Zn binding pocket, which is good. CaMKII association domain of CAMK2 has been shown to oligomerize with other CAMK2 in 1HKX.

Looking at the monomeric structure of ESRRG and CAMK2G, it looks possible that the C terminus association domain of CAMK2G to bind to ESRRG via Zn finger domain of ESRRG and the hormone receptor domain of ESRRG binds to the long and disordered region separating the two domains found in CAMK2G. This makes a multi-site binding between two proteins and a very interesting case.

run70: XRCC4-LIG4

The structure for this interaction has been solved: 3II6 and 1IK9. Looking at the structure of 3II6, the two proteins interact with each other via XRCC4 first forming a homodimer with its coiled-coil domain, then around the homodimer binds the tandem BRCT domains of LIG4. The BRCT domains are separated by a structurally less defined region that most likely forms two helices upon binding to XRCC4. Not sure if this can be seen as domain-motif or domain-domain interaction, probably something in between. It is not so clear from the monomeric AF model of full length LIG4 that both BRCT domains form a functional unit but I guess one could have also made a fragment comprising both domains and the linker sequence. Runs so far were made with both BRCT domains individually and the linker sequence individually and further rerun has to be done by using the BRCT domain tandem as one structural unit.

The top prediction involves a motif at the C-terminus of XRCC4 that is predicted to bind to the last BRCT domain of LIG4. I think the prediction is wrong because of the solved structure. The prediction also does not look like how other motifs bind to BRCT, i.e. the protein FANCJ (LIG_BRCT_BRCA_1). However, the C-terminus of XRCC4 certainly carries one or two motifs. One is annotated in Proviz as WD40 domain binding. The very C-terminus is a class 3 PDZ-binding motif. The whole region is very conserved. Maybe this is why AF tries to put peptides from this C-terminus in various domains, including the DNA ligase domain of LIG4 (fourth top prediction). So, the top two predictions involve this C-terminus and reach high confidences in both metrics (model confidence and intf_avg_plddt).

The third highest prediction involves the XRCC4 N-terminal domain plus one long helix (taken as one ordered region) and the 2nd BRCT domain. This interface is exactly the same interface that is seen in the structure 3II6 where part of the BRCT domain also contacts the XRCC4 helix.

The 6th best prediction involves the linker between both BRCT domains and the XRCC4 helix. Despite the fact that XRCC4 is in monomeric form in our prediction and that the BRCT domains are missing, AF correctly models the contacts between the linker and the single XRCC4 domain as they can be seen in the structure 3II6. This model meets both cutoffs, for model confidence and pLDDT.

Rerun using the BRCT domain tandem as one structural unit completed. The tandem BRCT fragment ranks 7th with the coiled coil XRCC4 fragment based on model confidence and second for ordered-ordered fragment pairs when ranked by avg interface plddt. The prediction that is still ranked first is the single BRCT domain binding to the coiled coil fragment (92 vs 89 avg intf plddt score).

run71: TMEM237-MFF

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run72: HNRNPK-TH

In the full length structural model of HNRNPK the first 2 KH domains are predicted to pack against each other using an interface that is also predicted to bind to the TH peptide 61-71. This region indeed overlaps with a Pfam HMM that seems to find some pattern in this disordered region but nothing is known about this “structural”(?) motif. It predicts 3 occurrences of it in the N-terminal region of TH but the third one is the most conserved and this is the one predicted to bind to the second KH domain. Two other motifs overlapping with 61-71 are also predicted to bind to this KH domain. The residues that are part of all three motifs are predicted to bind to the KH domain in the same way. One prediction below the model confidence cutoff predicts the motif to bind to the third KH domain but in a different way.

run73: OTX2-RPS26

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run74: MFF-MMGT1

Not inspected because none of the predictions returns model confidence or average interface pLDDT above cutoff (model confidence ≥ 0.7, ordered-disordered prediction with disordered fragment interface plddt ≥ 70; ordered-ordered prediction with intf_avg_plddt ≥ 75).

run75: PUF60-TH

The top prediction involves using both RRM domains of PUF60 as one ordered region and a disordered polyA peptide from TH. The peptide is put at the same position where the Nbox would bind as shown in the NMR structure 2KXH. However, the predicted peptide has some different sequence: solved structure: LxxAxxI, model: VxxAxxV, and there are no recurrent predictions. Another prediction involves the third RRM domain of PUF60 and another peptide in TH which tugs a Trp in a pocket but it does not look very convincing.

Prediction involving disordered fragments from PUF60 and ordered region (Biopterin_H domain) from TH returned a maximum of 0.78 model confidence. This is likely false interface because the short peptide is fit into the biopterin and iron binding pocket of the enzymatic domain (refer to run72 for example). The second best prediction is also fitted at the same site, therefore also likely a false interface.

Interestingly, the disordered region of PUF60 302-461 is modelled with 0.69 model confidence with the Biopterin_H domain of TH. The long disordered region makes contacts with two regions of the domain, one at the iron binding site (likely false) and another coiled-coil interaction at the C terminal helix of Biopterin_H domain. This coiled-coil interaction is repeated in a shorter disordered fragment of PUF60 (317-347, third best prediction (0.77), the same C terminal helix in the long disordered region). This coiled-coil interaction looks like a plausible interface.

I tried finding more information about this ACT-like domain but to no avail. InterPro says that it homo-dimerizes using the beta strands like in 1Q5V, but the fold is not exactly the same. The ACT-like domain in TH is special in the way that the last beta strand is formed by its N and C termini by looping back to meet each other. I cannot find much information about this domain.

run76: PUF60-QRICH1

One long disordered region of PUF60 (1-128) is modelled with high model confidence with DUF of QRICH1. In this region, 111-121 is modelled at the interface. This region when fragmented from the long disordered region also showed high confidence (0.86). This fragment tucks a R into a very deep negatively charged pocket but the rest of the peptide seems to make questionable contact with the DUF domain.

Top prediction with ordered region in QRICH1 and peptides in PUF60 either put the linker helix between the first two RRM domains or the N-terminal long helix in PUF60 or another helical peptide at 442-461 at two different places on the DUF domain. I think that the helical linker between both RRM domains is not accessible for this mode of binding because the key residues are making intramolecular contacts to the RRM domains in the AF monomer PUF60 model.

3 different peptides are predicted to bind to the tandem PUF60 domain. In principle, the long disordered N-terminal region of QRICH1 is full of potential helical peptides of pattern hydrophobic-x-x-Ala-x-x-hydrophobic, which is the kind of peptide that is like the Nbox motif that can bind to PUF60 and the three different peptides are also predicted to bind to the same pocket.

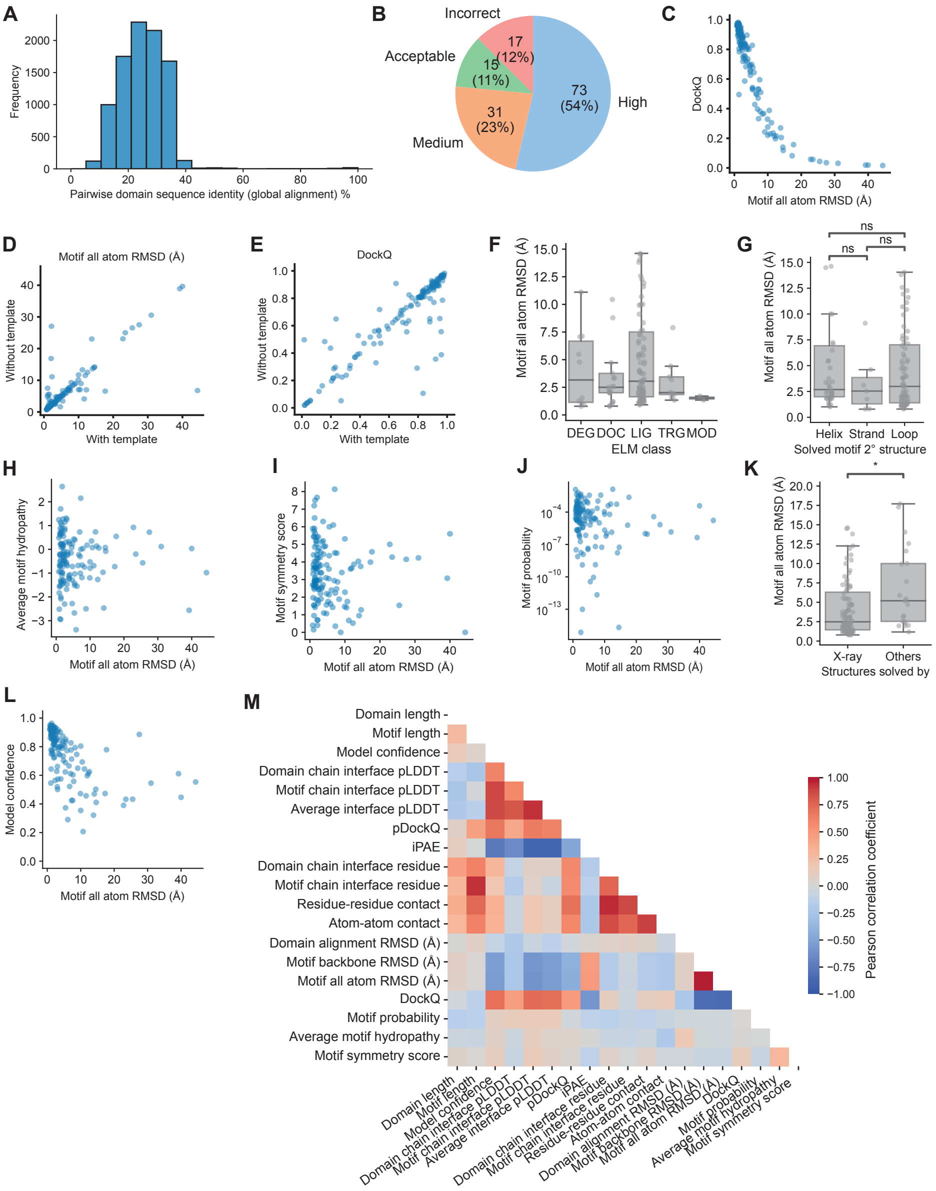
There are also 4 different peptides in QRICH1 predicted to bind to the third RRM domain.

run77: MAB21L2-AP1S2

The top prediction involves Clat_adaptor_s domain of AP1S2 with the disordered fragment (215-220) of MAB21L2 (78 motif pLDDT, 0.77 model confidence). The motif is predicted recurrently with variable length but the disordered region is generally very short because it is a loop within the domain of MAB21L2. AF also made a disulfide bridge between motif and domain. Not sure this is correct. Looking at the structure 1W63 that shows the large Ap1 clathrin adaptor core complex where there is a fold similar to the one in AP1S2, one can see that the region where the peptide is predicted to bind would in principle be accessible for binding. This domain Clat_adaptor_s is known to bind motifs from ELMDB but no structure has been solved in terms of this domain and its bound peptide. The disordered fragments from the previous point also do not match with any ELM class that binds to Clat_adaptor_s. Other good predictions use the Mab-21 domain of MAB21L2. Two overlapping disordered fragments (146-154, 0.68 and 153-157, 0.75) had good confidence with the domain but they are modelled to be at different binding sites, so it does not look likely to me that this is the binding region.

run78: PRKAR1B-QRICH1
The motif in PRKAR1B is at the very C-terminus of the protein and also matches a PDZ-binding motif. There is only one prediction that makes the model confidence cutoff but it does not meet the pLDDT cutoff. The C-terminal peptide of PRKAR1B binds to the only domain of QRICH1 but extended or smaller versions of the motif are only predicted with very low score then to bind to the domain so no recurrence here. The prediction therefore looks unlikely to be functional. No other predictions make the pLDDT cutoff.

Figure S1

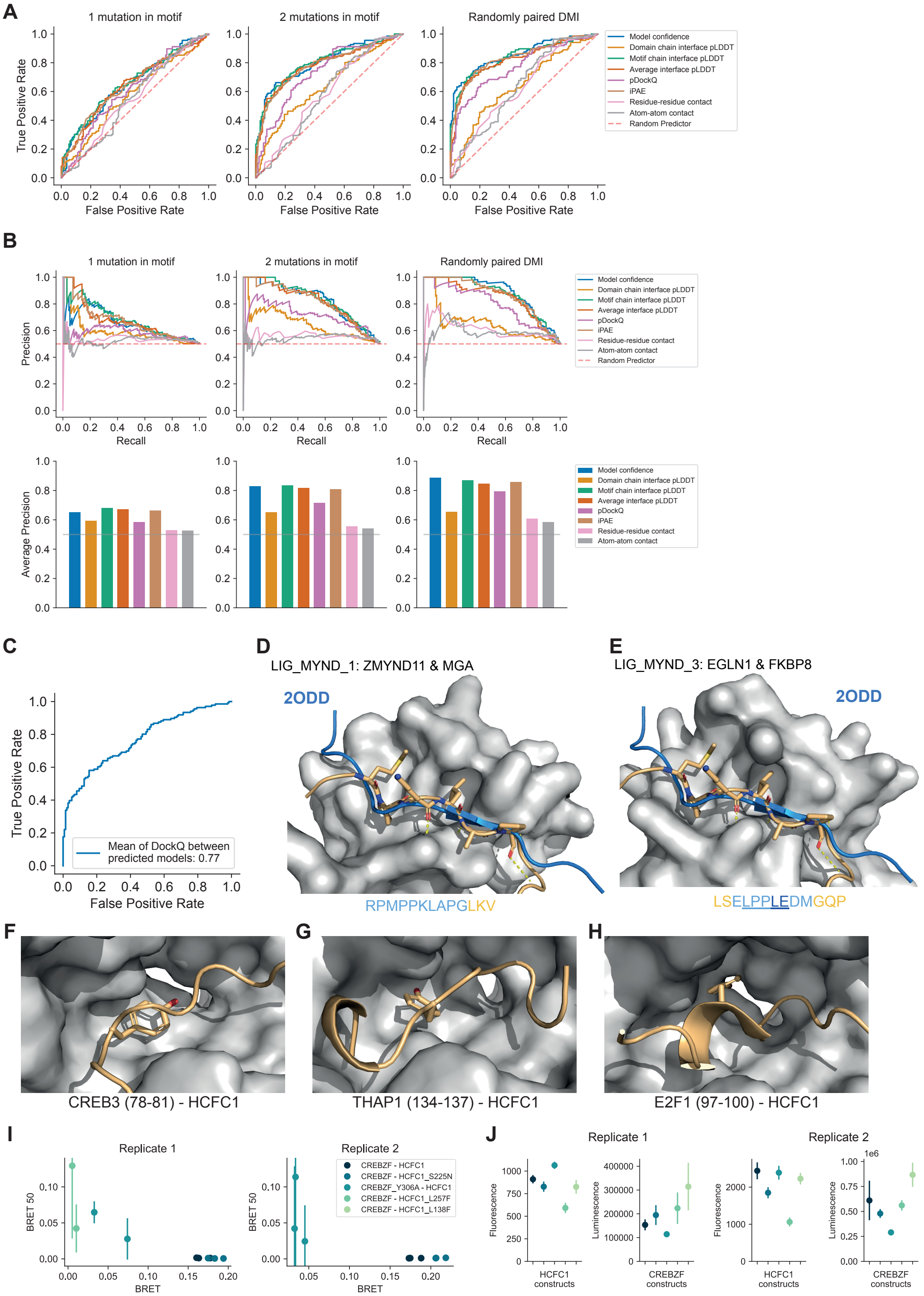


Appendix Figure S1. Benchmarking of AF on DMI interfaces using minimal interacting regions.

A Pairwise sequence identity of domains in the DMI positive reference dataset. **B** Proportion of high, medium, acceptable and incorrect models predicted by AF from the positive reference dataset as classified by the DockQ score. **C** Scatterplot of DockQ vs motif RMSD for DMIs from positive benchmark dataset. Pearson r = -0.85, p-value < 0.0001. **D-E** Motif RMSD and DockQ scores of structures for DMIs from positive benchmark dataset predicted by AF with and without the use of templates. Motif RMSD: Pearson r = 0.81, p-value < 0.0001. DockQ: Pearson r = 0.88, p-value < 0.0001. **F** Accuracy of AF DMI predictions stratified according to the annotated functional categories of DMIs in the ELM DB. DEG=degron, DOC=docking, LIG=ligand, TRG=targeting, MOD=modification. **G** Accuracy of AF DMI predictions stratified according to the secondary structure element formed by the motif in the solved structure. **H-J** Scatterplot of various motif features vs motif RMSD determined for models and structures of DMIs from positive benchmark dataset: H motif hydropathy, Pearson r = -0.03, p-value = 0.72, I motif symmetry, Pearson r = -0.08, p-value

= 0.38, J motif regular expression degeneracy, Pearson r = -0.04, p-value = 0.66. **K** Accuracy of AF DMI predictions stratified according to the method used to solve the structures in the benchmark dataset, Mann-Whitney-Wilcoxon test two-sided p-value = 0.017 test statistics = 811 **L** Scatterplot of model confidence of predicted models vs motif RMSD determined from superimposing the predicted models with structures of DMIs from the positive benchmark dataset. Pearson r = -0.55, p-value < 0.0001. **M** Correlation matrix of different prediction variables and prediction outcomes.

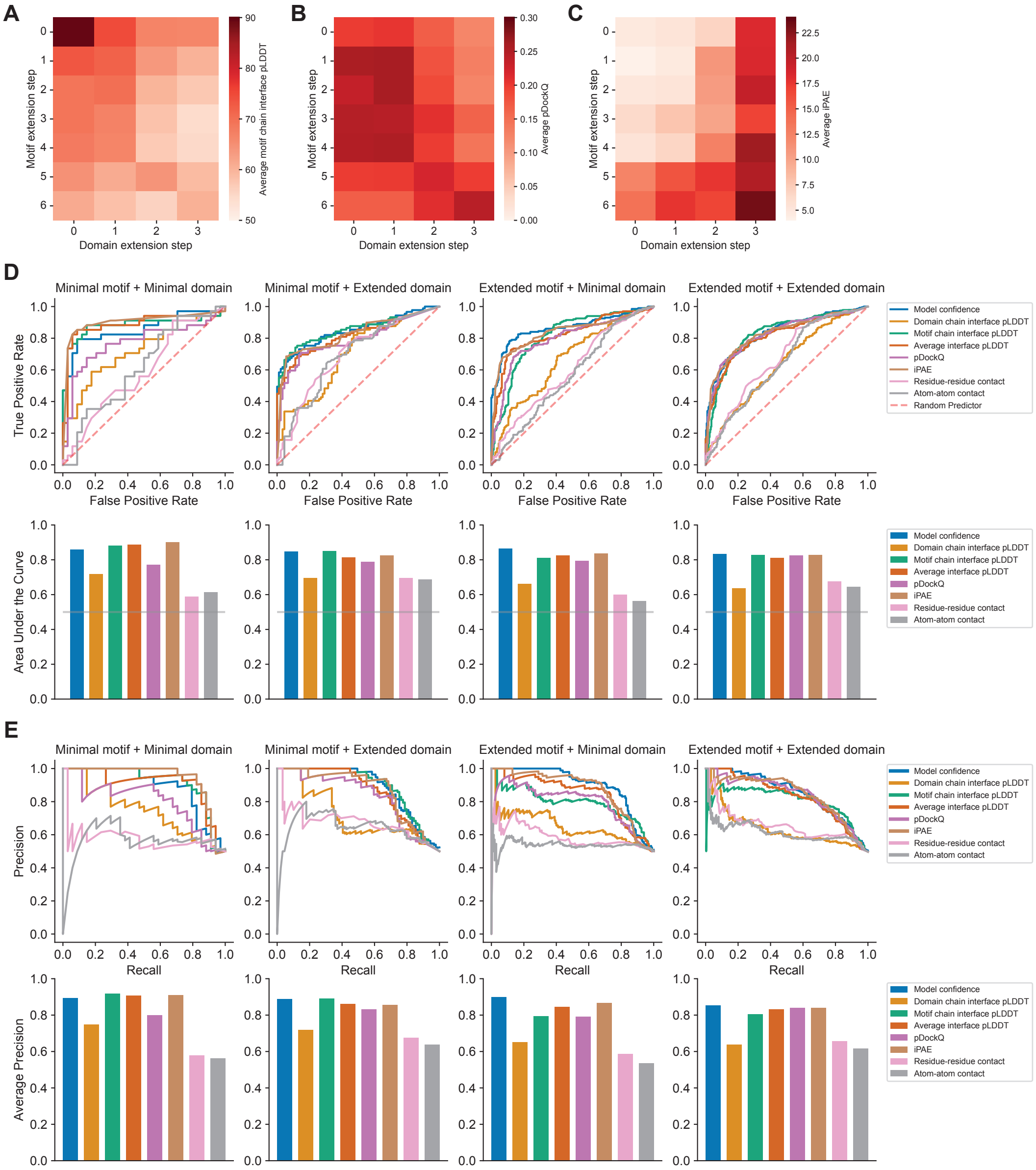
Figure S2



Appendix Figure S2. Benchmarking and application of AF for DMI interface prediction using minimal interacting fragments.

A Receiver operating characteristic (ROC) curve of various metrics extracted from AF models when using the DMI benchmark dataset as the positive reference and the following sets as random reference: Left, 1 mutation introduced in conserved motif position; middle, 2 mutations introduced in conserved motif positions, right, randomly shuffled domain-motif pairs. **B** Precision recall curve of various metrics determined for benchmark datasets as in A. **C** ROC curve of mean DockQ between the top five AF structural models returned for a given input, assessed using the DMI positive reference set and random pairings of domains and motifs as in A. The AUROC of the metric is indicated in the legend of the ROC curve. **D-E** Superimposition of AF structural model for motif class LIG_MYND_1 (D) and LIG_MYND_3 (E) (orange) with homologous solved structures (PDB:2ODD) from motif class LIG_MYND_2 (blue). The motif sequence used for prediction is indicated at the bottom, colored by pLDDT (dark blue=highest pLDDT). **F-H** AF models for three motif instances (orange) of LIG_HCF-1_HBM_1 predicted to bind into a pocket on the Kelch domain of HCFC1 (gray). Motif positions are indicated below the figures. The key tyrosines of the motif sequences are drawn as sticks. **I** BRET50 estimates from fitting titration curves shown in Fig 1G are plotted vs. BRET values that were corrected for bleedthrough and measured at a 2:50 ng DNA transfection ratio for wildtype and mutant CREBZF-HCFC1 pairs. Error bars indicate the standard error. Data is shown for two technical replicates for the first biological replicate and three technical replicates for the second biological replicate. **J** Fluorescence and total luminescence are shown for wildtype and mutant CREBZF-HCFC1 pairs measured at a 2:50 ng DNA transfection ratio. Error bars indicate STD of two technical replicates for the first biological replicate and three technical replicates for the second biological replicate. Coloring as in I.

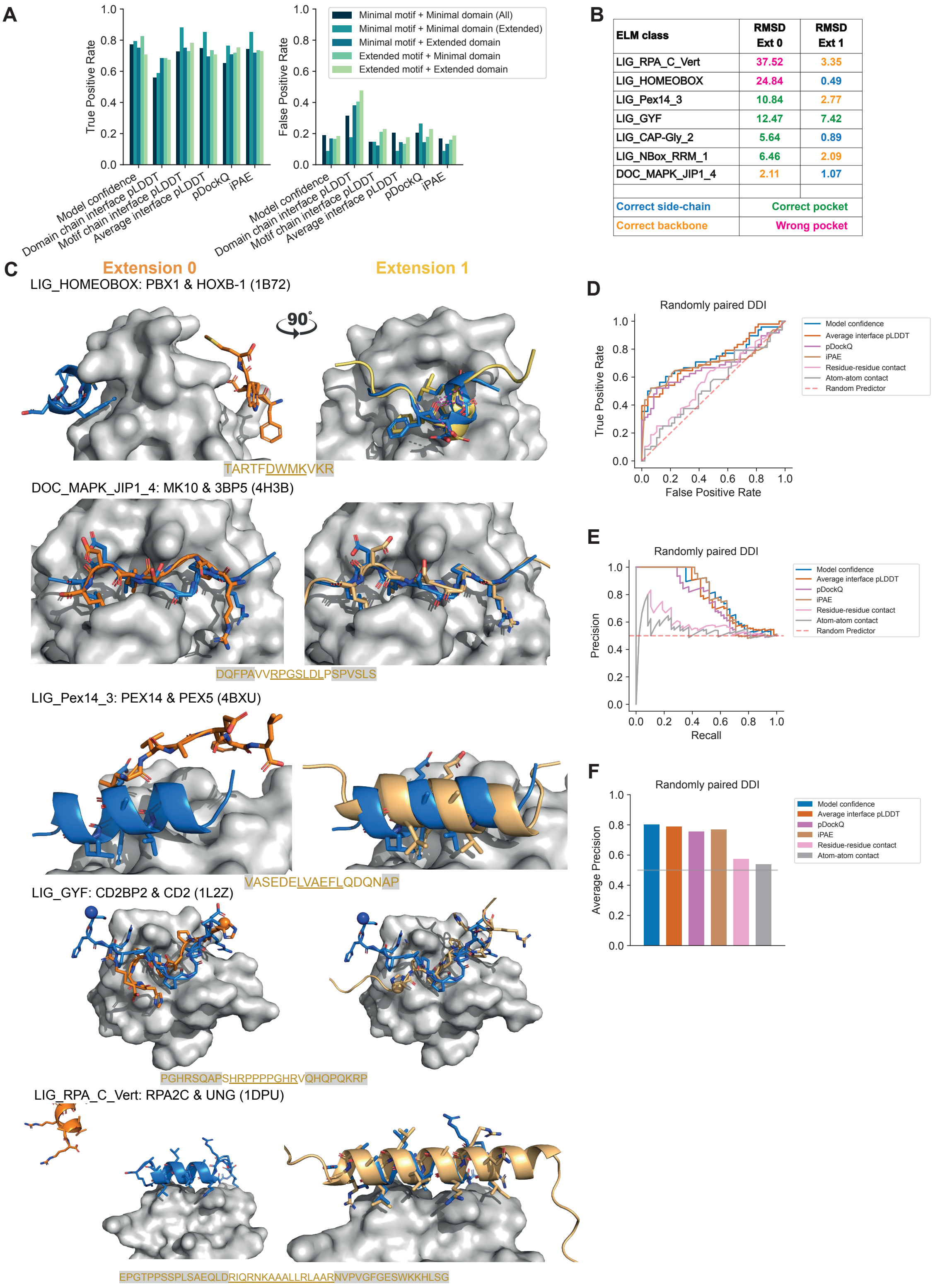
Figure S3



Appendix Figure S3. Effect of protein fragment extensions on the accuracy of AF predictions.

A-C Heatmap of the average motif interface pLDDT (A), pDockQ (B), and iPAE (C) for combinations of different motif and domain sequence extensions using a positive reference set consisting of 31 DMI structures. Extensions like in Fig 2A. **D** ROC curves (top) and corresponding AUROC values (bottom) of various metrics extracted from AF models when using the DMI extension dataset split by different combinations of motif and domain extensions as indicated on the top of each graph. Gray horizontal line indicates the AUROC of a random predictor. **E** Precision recall curves (top) and area under the precision recall curve as quantified by average precision (bottom) for various metrics extracted from AF models determined for benchmark datasets as in D.

Figure S4

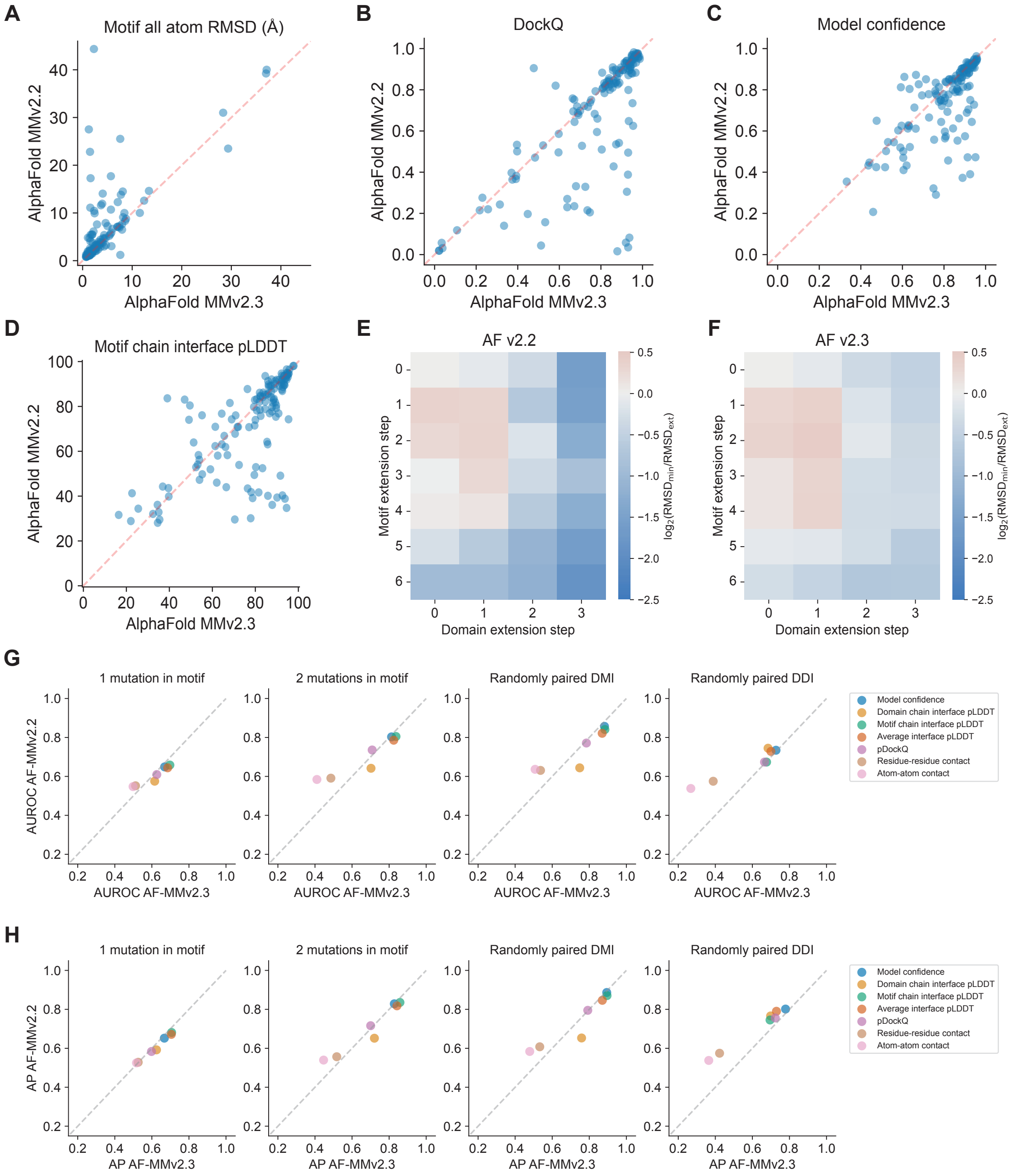


ELM class	RMSD Ext 0	RMSD Ext 1
LIG_RPA_C_Vert	37.52	3.35
LIG_HOMEOBOX	24.84	0.49
LIG_Pex14_3	10.84	2.77
LIG_GYF	12.47	7.42
LIG_CAP-Gly_2	5.64	0.89
LIG_NBox_RRM_1	6.46	2.09
DOC_MAPK_JIP1_4	2.11	1.07
Correct side-chain	Correct pocket	
Correct backbone	Wrong pocket	

Appendix Figure S4. Effect of protein fragment extensions on the accuracy of AF predictions.

A True and false positive rate (left and right, respectively) based on optimal cutoffs from Fig 2D derived for different metrics from ROC analysis for benchmarking AF with different motif and domain extensions from the reference dataset illustrated in Fig 2A and random pairings of domain and motif sequences. **B** Table indicating the motif RMSD achieved when using minimal (extension 0) or extended motif sequences for structure prediction for all inspected motif extension cases. Extension 1 refers to extension of the minimal motif sequence by the length of the motif to the left and right. Color coding indicates the accuracy classes of the respective structural models as shown in Fig 1A. **C** Superimposition of the structural model of the minimal (left, orange) or extended (right, yellow) motif sequence with the solved structure (motif in blue) for five different motif classes as indicated on the top of each panel. The motif sequence from the solved structure is indicated at the bottom of each panel. Motif residues are underlined, motif residues not resolved in the structure have a gray background. Sticks indicate the motif residues, domain surfaces are shown in gray based on experimental structures. **D** ROC curves of different metrics using the DDI benchmark dataset as positive reference and random shuffling of domain-domain pairs as negative reference. **E** Precision recall curves of different metrics extracted from AF models determined for benchmark datasets as in D. **F** Area under the precision recall curve as quantified by average precision for metrics extracted from AF models determined for benchmark datasets as in D. Gray horizontal line indicates the average precision of a random predictor.

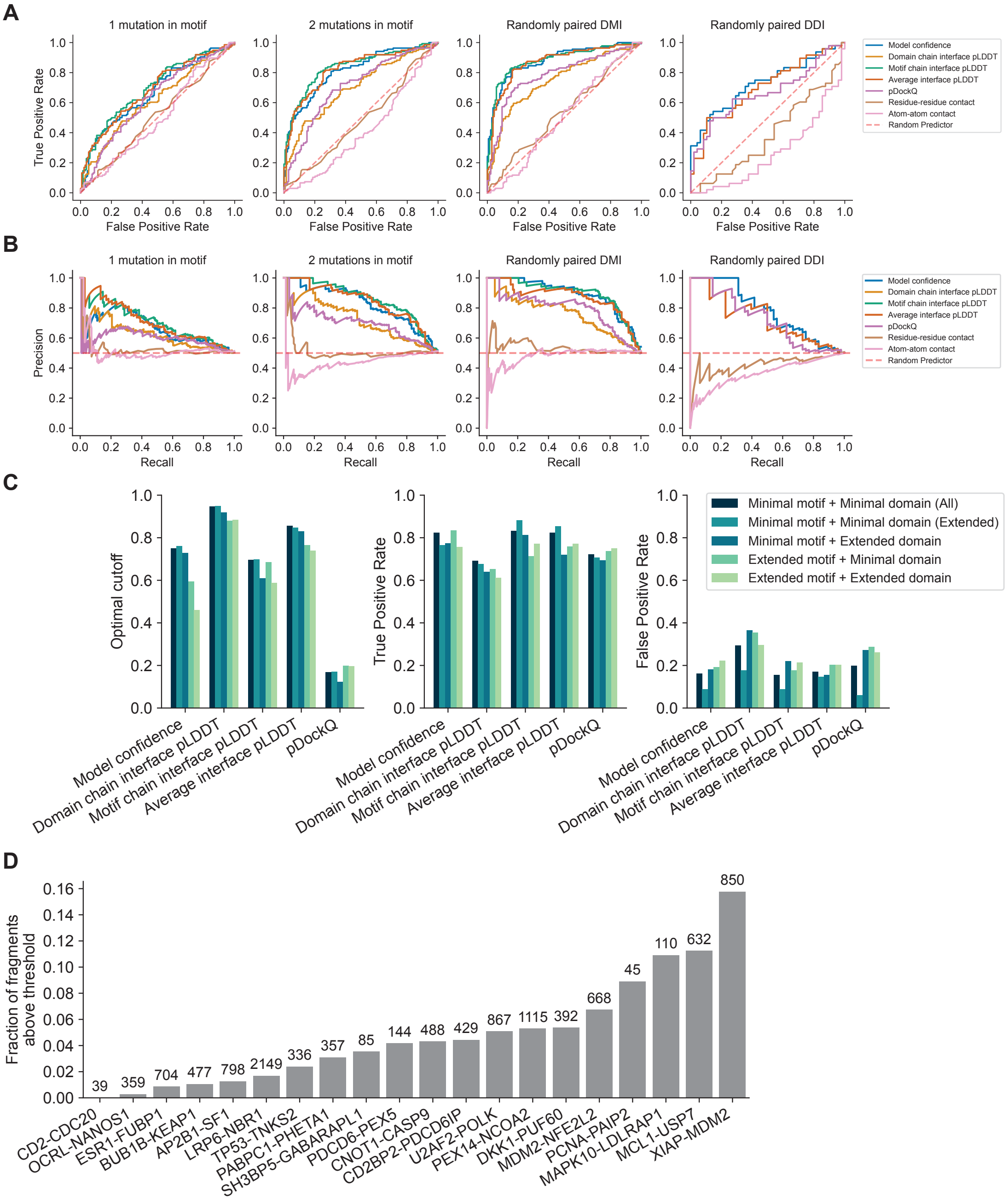
Figure S5



Appendix Figure S5. Comparison of AF v2.2 and v2.3 prediction performance.
A Scatterplot showing the motif RMSD obtained from structural models computed either with AF v2.2 or AF v2.3 using the minimal interacting regions of all annotated DMIs. **B-D** Scatterplots computed as in A showing the DockQ (B), model confidence (C), and motif chain interface pLDDT (D) for both AF versions. **E-F** Heatmaps showing the fold change in motif RMSD obtained for structural models from AF v2.2 (E) and AF v2.3 (F) upon domain or/and motif sequence extension compared to when using minimal interacting regions. Positive values indicate improved predictions from extension and negative values indicate worse prediction outcomes. **G** Scatterplots showing the AUROC obtained for different metrics derived from structural models from benchmarking AF v2.2 and AF v2.3 using the minimal interacting regions of all annotated DMIs or DDIs as the positive reference dataset and different random reference datasets: Left (DMI), 1 mutation introduced in conserved

motif position; middle-left (DMI), 2 mutations introduced in conserved motif positions, middle-right (DMI), randomly shuffled domain-motif pairs; right (DDI), randomly shuffled domain-domain pairs. Corresponding ROC curves for AF v2.2 and AF v2.3 are shown in Fig. S2A, S4D, and S6A. **H** Scatterplots as in G plotting the average precision (AP) obtained from PR curves from the same analysis as in G. Corresponding PR curves for AF v2.2 and AF v2.3 are shown in Fig S2B, S4E and S6B.

Figure S6

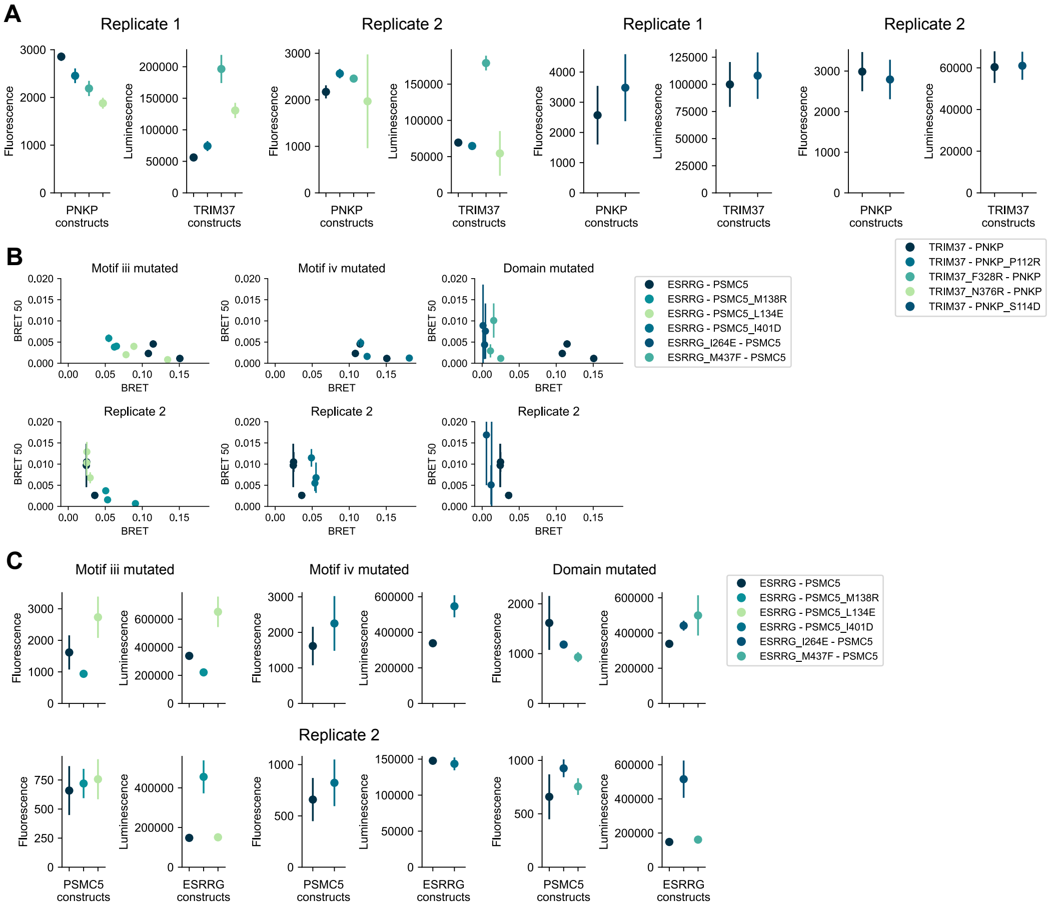


Appendix Figure S6. Performance of different metrics derived from structural models when benchmarking AF v2.3 for DMI predictions.

A ROC curves obtained for different metrics derived from structural models from benchmarking AF v2.3 using the minimal interacting regions of all annotated DMIs or DDIs as the positive reference dataset and different random reference datasets: Left (DMI), 1 mutation introduced in conserved motif position; middle-left (DMI), 2 mutations introduced in conserved motif positions, middle-right (DMI), randomly shuffled domain-motif pairs; right (DDI), randomly shuffled domain-domain pairs. **B** PR curves computed for the same datasets and AF version as in A. **C** Optimal cutoff, true, and false positive rate derived for different metrics from ROC analysis for benchmarking AF v2.3 with different motif and domain extensions from the reference dataset used in Fig 2A and randomly shuffled domain

-motif pairs. **D** Fraction of fragment pairs with structural models scoring above thresholds for 20 randomly shuffled domain-motif pairs. Numbers on top of the bars indicate the total number of fragment pairs submitted for interface prediction to AF for each random protein pair.

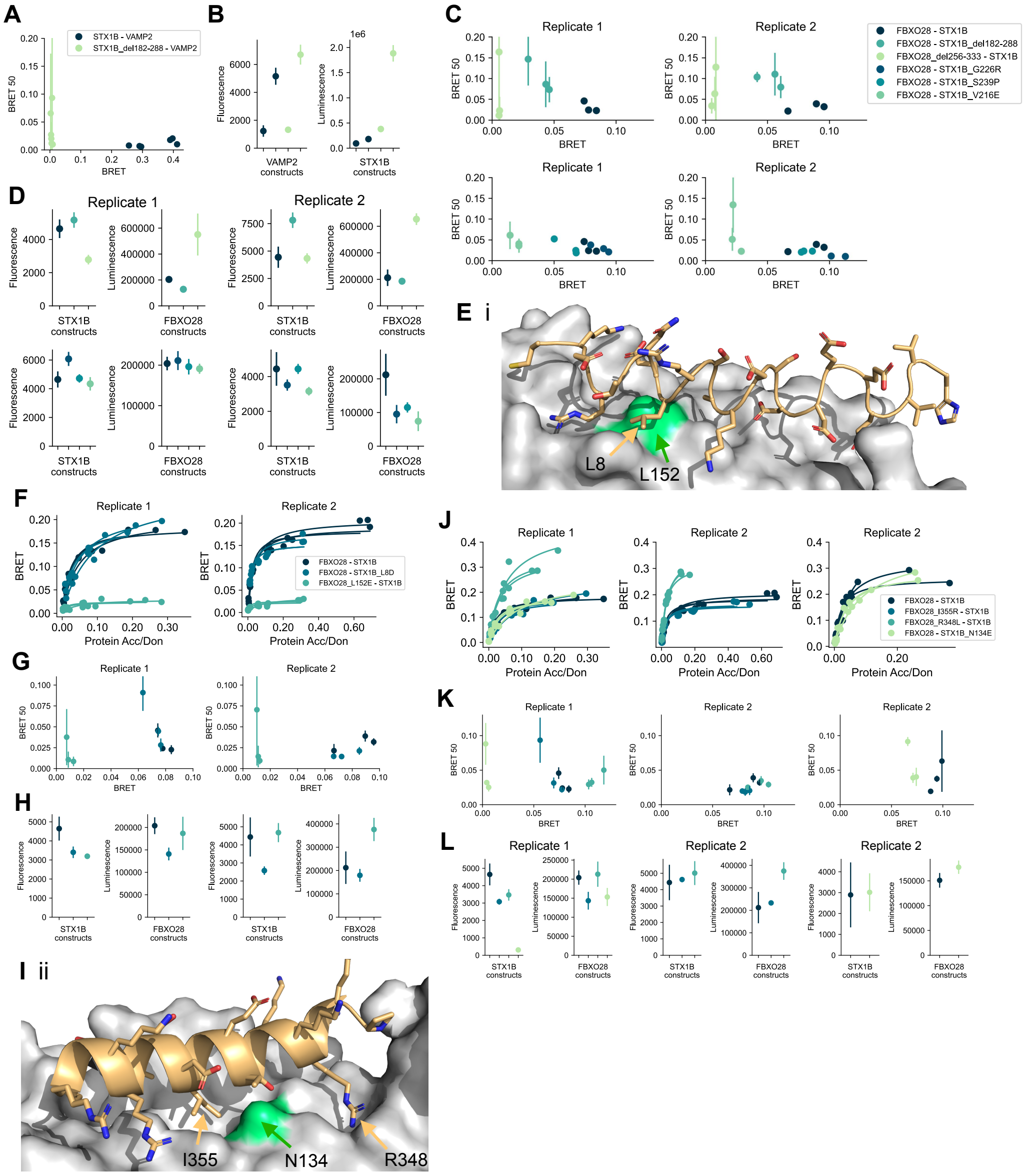
Figure S7



Appendix Figure S7. Expression and BRET50 plots for TRIM37-PNKP and ESRRG-PSMC5.

A Fluorescence and total luminescence are shown for wildtype and mutant TRIM37-PNKP pairs measured at a 2:50 ng DNA transfection ratio. Error bars indicate STD of three technical replicates. Data is shown for two biological replicates. **B** BRET50 estimates from fitting titration curves shown in Fig 4H are plotted vs. BRET values that were corrected for bleedthrough and measured at a 2:50 ng DNA transfection ratio for wildtype and mutant ESRRG-PSMC5 pairs. Error bars indicate the standard error. Data is shown for three technical replicates for two biological replicates each. BRET50 estimates for the second biological replicate for the ESRRG_M437F-PSMC5 pair were omitted from the graph because they exceeded the upper y-axis limit. Roman labels refer to interfaces shown in Fig 4E. **C** Fluorescence and total luminescence are shown for wildtype and mutant ESRRG-PSMC5 pairs measured at a 2:50 ng DNA transfection ratio. Error bars indicate STD of three technical replicates. Data is shown for two biological replicates.

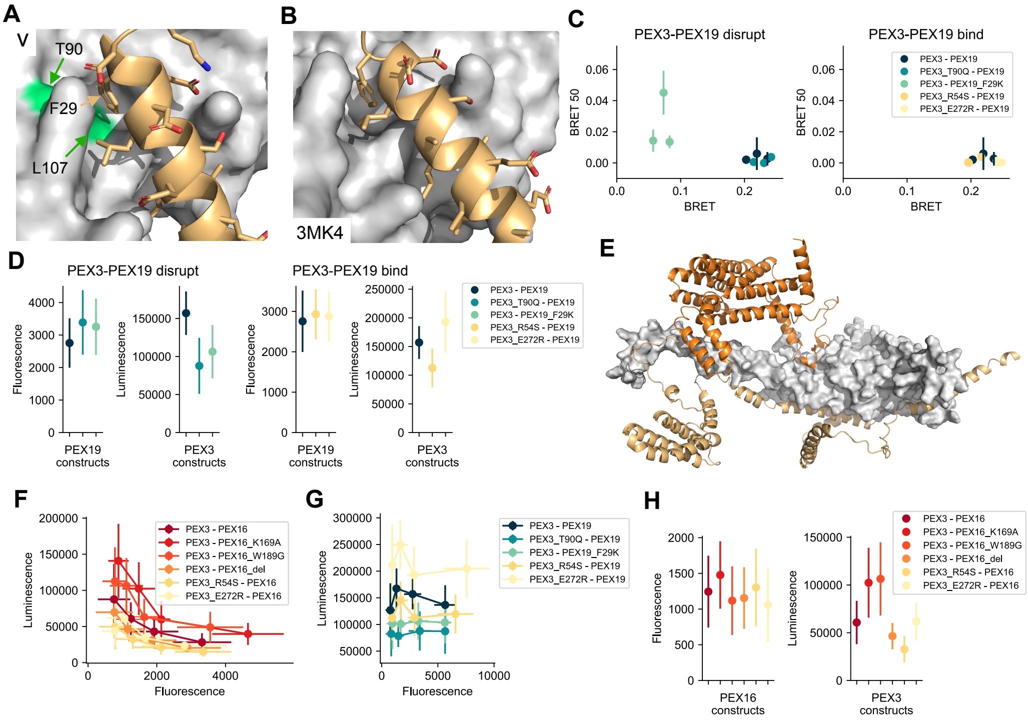
Figure S8



Appendix Figure S8. Structural models, expression, and BRET50 plots for STX1B-FBXO28 and STX1B-VAMP2.

A BRET50 estimates from fitting titration curves shown in Fig 5C are plotted vs. BRET values that were corrected for bleedthrough and measured at a 2:50 ng DNA transfection ratio for wildtype and mutant STX1B-VAMP2 pairs. Error bars indicate the standard error. Data is shown for three technical replicates for two biological replicates each. **B** Fluorescence and total luminescence are shown for wildtype and mutant STX1B-VAMP2 pairs measured at a 2:50 ng DNA transfection ratio. Error bars indicate STD of three technical replicates. Data is shown for two biological replicates. **C** Data shown as in A for wildtype and mutant FBXO28-STX1B pairs relating to interface iii (Fig 5A,D). **D** Data shown as in B for wildtype and mutant FBXO28-STX1B pairs shown in C. **E** Structural model corresponding to interface i shown in Fig 5A. Mutated residues on the domain (green) and motif side are labeled. **F** BRET titration curves are shown for wildtype and mutant FBXO28-STX1B pairs relating to interface i shown in E with two biological replicates, each with three technical replicates. Protein acceptor over protein donor expression levels are plotted on the x-axis determined from fluorescence and luminescence measurements, respectively. **G** Data shown as in A for wildtype and mutant FBXO28-STX1B pairs relating to interface i. **H** Data shown as in B for wildtype and mutant FBXO28-STX1B pairs relating to interface i. **I** Structural model corresponding to interface ii shown in Fig 5A. Mutated residues on the domain (green) and motif side are labeled. **J** Data shown as in F for wildtype and mutant FBXO28-STX1B pairs relating to interface ii. **K** Data shown as in A for wildtype and mutant FBXO28-STX1B pairs relating to interface i. **L** Data shown as in B for wildtype and mutant FBXO28-STX1B pairs relating to interface i.

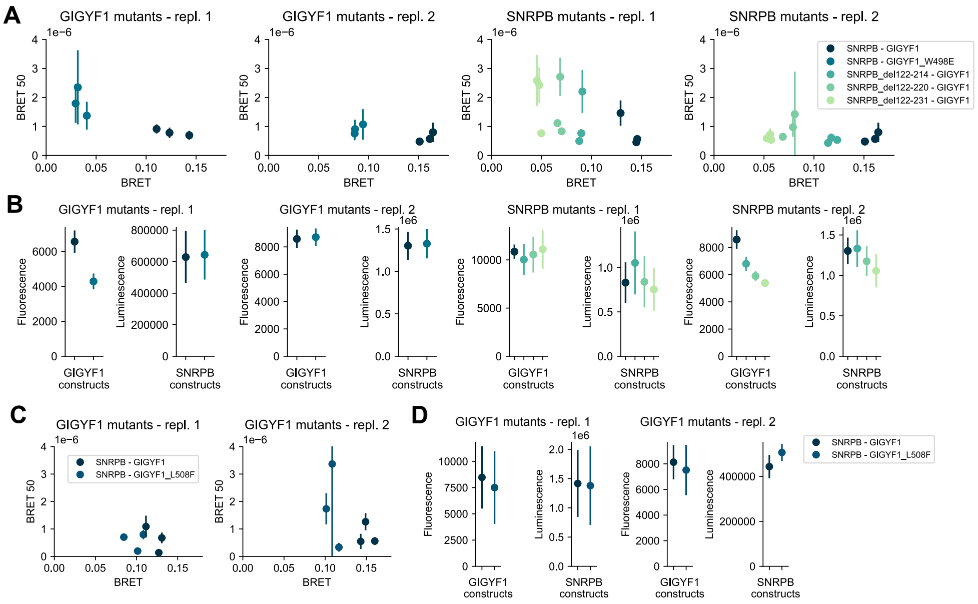
Figure S9



Appendix Figure S9. Structural models, expression, and BRET50 plots for PEX3-PEX19 and PEX3-PEX16.

A Structural model of PEX3-PEX19 corresponding to interface v as shown in Fig 5G. Mutated residues on the domain (green) and motif side are labeled. **B** Structure from PDB:3MK4 showing the PEX19 N-terminal motif bound to the PEX3 domain. **C** BRET50 estimates from fitting titration curves shown in Fig 5H are plotted vs. BRET values that were corrected for bleedthrough and measured at a 2:50 ng (for PEX3 and PEX3_T90Q) or 8:50 ng (for PEX3, PEX3_R54S, PEX3_E272R) DNA transfection ratio for wildtype and mutant PEX3-PEX19 pairs. Error bars indicate the standard error. Data is shown for three technical replicates. The left panel corresponds to mutant constructs that should disrupt binding while mutants shown in the right panel were aimed to disrupt binding to PEX16 and thus should not disrupt binding to PEX19. **D** Fluorescence and total luminescence are shown for wildtype and mutant PEX3-PEX19 pairs measured at a 2:50 or 8:50 ng DNA transfection ratio (see panel C). Error bars indicate STD of three technical replicates. **E** Structural model obtained with AF for the trimeric complex of PEX3 (gray), PEX19 (yellow), and PEX16 (orange) using full length sequences as input. **F** PEX3 expression levels measured in luminescence units plotted for co-transfections with increasing PEX16 protein amounts measured in fluorescence units. Error bars indicate STD of three technical replicates. **G** PEX3 expression levels measured in luminescence units plotted for co-transfections with increasing PEX19 protein amounts measured in fluorescence units. Error bars indicate STD of three technical replicates. **H** Data shown as in D for wildtype and mutant constructs of PEX3-PEX16 pairs. Measures are taken for 2:25 ng DNA transfection ratios.

Figure S10



Appendix Figure S10. Expression and BRET50 plots for SNRPB-GIGYF1. **A** BRET50 estimates from fitting titration curves shown in Fig 6D are plotted vs. BRET values that were corrected for bleedthrough and measured at a 2:50 ng DNA transfection ratio for wildtype and mutant SNRPB-GIGYF1 pairs. Error bars indicate the standard error. Data is shown for three technical replicates for two biological replicates each. **B** Fluorescence and total luminescence are shown for wildtype and mutant SNRPB-GIGYF1 pairs measured at a 2:50 ng DNA transfection ratio. Error bars indicate STD of three technical replicates. Data is shown for two biological replicates. Coloring as in A. **C** Data shown as in A for wildtype and mutant SNRPB-GIGYF1 pairs fitted from titration curves shown in Fig 6E. **D** Data shown as in B for wildtype and mutant SNRPB-GIGYF1 pairs shown in C.

4.2 Article III: An extended Tudor domain within Vreteno interconnects Gtsf1L and Ago3 for piRNA biogenesis in *Bombyx mori*

Summary

This project focused on characterizing the molecular mechanisms of PPIs involved in the piRNA biogenesis.

Piwi-interacting RNAs (piRNAs) play a crucial role in preserving genome integrity by silencing transposons. The piRNA population is known to expand through the ping-pong amplification loop. Many proteins are involved in the amplification process, but the molecular mechanism of the process is not fully understood. In silkworm cells, we showed that the protein GTSF1L interacts with piRNA-loaded Ago3 and co-localizes with unloaded Ago3 and Vreteno. Further biochemical analyses revealed that GTSF1L was able to directly interact with Vreteno through its disordered C terminal tail. By applying the fragmentation strategy presented in Chapter 4, Article II, we generated a high-confidence structural model for the detected interface. Interestingly, the structural model revealed a hydrophobic pocket on the Tudor domain that was previously not known. Atomistic molecular dynamics simulations were also performed using the structural model predicted by AF-MM, and the result further corroborated the interface predicted by AF-MM. The predicted interface was experimentally validated.

In summary, this project demonstrated the use of AF-MM in deciphering the molecular mechanisms of PPIs and the application of AF-MM models for more fine-grained structural studies.

Statement of contribution

This is a collaborative project where I conducted the AF-MM modelling aspect of the study, as well as related data visualization and analysis. I also took part in the writing of the Method section.

Supervisor confirmation

Article

SOURCE DATA TRANSPARENT PROCESS OPEN ACCESS

THE EMBO JOURNAL

An extended Tudor domain within Vreteno interconnects Gtsf1L and Ago3 for piRNA biogenesis in *Bombyx mori*

Alfred W Bronkhorst[1,*], Chop Y Lee[2,3], Martin M Möckel[4], Sabine Ruegenberg[4], Antonio M de Jesus Domingues[1,†], Shéraz Sadouki[1], Rossana Piccinno[5], Tetsutaro Sumiyoshi[6,‡], Mikiko C Siomi[6], Lukas Stelzl[7,8], Katja Luck[3,**] & René F Ketting[1,9,***]

Abstract

Piwi-interacting RNAs (piRNAs) direct PIWI proteins to transposons to silence them, thereby preserving genome integrity and fertility. The piRNA population can be expanded in the ping-pong amplification loop. Within this process, piRNA-associated PIWI proteins (piRISC) enter a membraneless organelle called nuage to cleave their target RNA, which is stimulated by Gtsf proteins. The resulting cleavage product gets loaded into an empty PIWI protein to form a new piRISC complex. However, for piRNA amplification to occur, the new RNA substrates, Gtsf-piRISC, and empty PIWI proteins have to be in physical proximity. In this study, we show that in silkworm cells, the Gtsf1 homolog BmGtsf1L binds to piRNA-loaded BmAgo3 and localizes to granules positive for BmAgo3 and BmVreteno. Biochemical assays further revealed that conserved residues within the unstructured tail of BmGtsf1L directly interact with BmVreteno. Using a combination of AlphaFold modeling, atomistic molecular dynamics simulations, and *in vitro* assays, we identified a novel binding interface on the BmVreteno-eTudor domain, which is required for BmGtsf1L binding. Our study reveals that a single eTudor domain within BmVreteno provides two binding interfaces and thereby interconnects piRNA-loaded BmAgo3 and BmGtsf1L.

Keywords Ago3; eTudor; Gtsf; piRNA; Vreteno

Subject Categories RNA Biology; Structural Biology

DOI 10.15252/embj.2023114072 | Received 22 March 2023 | Revised 25 October 2023 | Accepted 30 October 2023 | Published online 20 November 2023

The EMBO Journal (2023) 42: e114072

Introduction

Animal germ cells utilize the Piwi-interacting (pi)RNA pathway as a mechanism to silence transposons, thereby maintaining genome stability and fertility (Czech *et al*, 2018; Ozata *et al*, 2019). A defective piRNA pathway leads to transposon derepression, DNA damage, gametogenesis defects, and sterility. The piRNA pathway can also have non-transposon targets, such as in the silk moth *Bombyx mori*, where piRNAs regulate sex determination (Kiuchi *et al*, 2014). piRNAs are about 24–31 nucleotides in size and associate with PIWI-clade Argonaute proteins to guide them to complementary targets (Ghildiyal & Zamore, 2009). Therefore, piRNAs are a key, specificity-determining component of the Piwi pathway.

In *Drosophila* germ cells, precursor piRNAs (pre-pre-piRNAs) are transcribed within large dual-strand clusters and exported from the nucleus for subsequent processing within the cytoplasm (Brennecke *et al*, 2007; Czech *et al*, 2018; Ozata *et al*, 2019). piRNAs can be loaded into different PIWI proteins. *Drosophila* expresses three PIWI proteins: Piwi, Aubergine (Aub), and Ago3. Piwi and Aub are predominantly loaded with antisense piRNAs, whereas Ago3 mostly incorporates sense piRNAs (Brennecke *et al*, 2007; Gunawardane *et al*, 2007). In the silkworm, only two cytoplasmic PIWI proteins are expressed: Siwi and BmAgo3 (Kawaoka *et al*, 2009). Siwi-associated piRNAs are mostly antisense and are responsible for the cleavage of sense transposon mRNA, whereas BmAgo3 binds sense piRNAs and triggers antisense piRNA precursor cleavage.

The current model for piRNA biogenesis suggests that cytoplasmic piRNA processing occurs through two interconnected mechanisms.

1 Biology of Non-coding RNA Group, Institute of Molecular Biology, Mainz, Germany
2 International PhD Programme on Gene Regulation, Epigenetics & Genome Stability, Mainz, Germany
3 Integrative Systems Biology Group, Institute of Molecular Biology, Mainz, Germany
4 Protein Production Core Facility, Institute of Molecular Biology, Mainz, Germany
5 Microscopy Core Facility, Institute of Molecular Biology, Mainz, Germany
6 Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo, Japan
7 Faculty of Biology, Johannes Gutenberg University Mainz, Mainz, Germany
8 KOMET 1, Institute of Physics, Johannes Gutenberg University Mainz, Mainz, Germany
9 Institute of Developmental Biology and Neurobiology, Johannes Gutenberg University, Mainz, Germany
*Corresponding author. Tel: +49 6131 39 21474; E-mail: w.bronkhorst@imb-mainz.de
**Corresponding author. Tel: +49 6131 39 21440; E-mail: k.luck@imb-mainz.de
***Corresponding author. Tel: +49 6131 39 21470; E-mail: r.ketting@imb-mainz.de
†Present address: Dewpoint Therapeutics GmbH, Dresden, Germany
‡Present address: Department of Medical Innovations, Osaka Research Center for Drug Discovery, Otsuka Pharmaceutical Co., Ltd., Osaka, Japan

One step of piRNA processing takes place within the nuage, a germline-specific phase-separated structure that surrounds the nuclear membrane. Here, piRNA-guided (called trigger piRNA) endonuclease activity of one PIWI protein (e.g., Ago3) generates the 5′ monophosphate end of a complementary piRNA precursor transcript (Han *et al*, 2015; Homolka *et al*, 2015; Mohn *et al*, 2015; Gainetdinov *et al*, 2018). This so-called responder pre-piRNA is subsequently incorporated into an unloaded PIWI protein (*Drosophila* Aub or silkworm Siwi), which is often still too long at its 3′-end. For further pre-piRNA processing, the PIWI protein then migrates to the mitochondrial outer membrane. Here, the second step of piRNA processing is mediated by the endonuclease Zucchini (Zuc), which mediates responder pre-piRNA 3′-end formation by cleaving 5′ to an available uridine (Han *et al*, 2015; Mohn *et al*, 2015). Pre-piRNA 3′-end resection is further completed by trimming and methylation to generate a mature piRISC complex (Horwich *et al*, 2007; Kawaoka *et al*, 2011; Hayashi *et al*, 2016; Izumi *et al*, 2016). The mature piRISC complex is liberated from the mitochondria into the cytosol to cleave complementary RNA, resulting in target RNA degradation. Alternatively, the mature piRISC complex (Aub or Siwi) can transit back to the nuage (now serving as a trigger piRNA) to bind complementary target RNA. This initiates a new round of PIWI-catalyzed responder piRNA biogenesis, leading to new piRISC formation. This consecutive and continuous process of responder and trigger piRNA production, which requires reciprocal cleavages by two paired PIWI proteins, is called the piRNA amplification cycle (or ping-pong loop). Thus, mature piRNAs within the ping-pong cycle are generated by the combined action of PIWI-slicing and Zuc-cleavage.

Notably, Zuc-mediated processing of the responder pre-piRNA 3′-end simultaneously generates the 5′-end of a new pre-piRNA substrate for phased piRNA biogenesis, which results in the production of trailer piRNAs (Han *et al*, 2015; Mohn *et al*, 2015). Initiator and responder piRNAs that are generated via the ping-pong cycle increase the abundance of an existing pool of piRNAs, whereas the Zuc-dependent trailer piRNAs expand the repertoire of piRNA sequences. In *Drosophila*, phased piRNAs predominantly associate with Piwi and translocate to the nucleus to induce transcriptional gene silencing through the deposition of repressive chromatin marks (Czech *et al*, 2018). Even though trailer piRNAs are produced, the silkworm does not possess a nuclear piRNA-based silencing pathway (Gainetdinov *et al*, 2018; Izumi *et al*, 2020).

Efficient piRNA amplification within the ping-pong cycle requires that PIWI-mediated target cleavage is confined to molecular surroundings that are compatible with an empty PIWI protein receiving one of the cleavage products. It has been suggested that Tudor-domain-containing proteins that reside in the nuage can provide such an environment by acting as a molecular scaffold (Chen *et al*, 2011; Siomi *et al*, 2011). Tudor domains that harbor an aromatic cage can bind to symmetrically dimethylated arginine residues (sDMAs) on client proteins but can also establish sDMA-independent protein interactions (Siomi *et al*, 2010; Chen *et al*, 2011). For example, the *Drosophila* Krimper protein makes sure that cleavage products resulting from Aub-slicing are efficiently loaded into empty Ago3 (Sato *et al*, 2015; Webster *et al*, 2015). Krimper binds sDMA-methylated piRISC-Aub via its aromatic cage-containing Tudor domain, whereas an upstream Tudor domain within Krimper establishes the sDMA-independent interaction with empty Ago3 (Sato *et al*, 2015; Webster *et al*, 2015; Huang *et al*, 2021). Likewise, the multi-Tudor domain-containing protein Qin also promotes heterotypic ping-pong between piRISC-Aub and empty, unmethylated Ago3 (Zhang *et al*, 2011, 2014). In silkworms, the handover of piRISC-Siwi-cleaved target RNA to empty BmAgo3 is mediated by the RNA-helicase Vasa (Xiol *et al*, 2014; Nishida *et al*, 2015). Notably, Vasa contains intrinsically disordered regions that are involved in the formation of phase-separated structures and seems to be the scaffold for nuage formation (Nott *et al*, 2015). Moreover, the Vasa N-terminus is strongly methylated, indicating that multivalent interactions with Tudor domain-containing proteins also contribute to nuage assembly (Kirino *et al*, 2010). Additional studies in silkworm revealed that BmVreteno brings piRNA-loaded BmAgo3 and empty Siwi together via their Tudor domains to allow new piRISC-Siwi formation (Nishida *et al*, 2020). This may involve the dimerization of two BmVreteno isoforms (BmVreteno-Long and -Short), where the BmVreteno-Long isoform anchors the RNA target through its unique RRM domain. Thus, BmVreteno also acts in the ping-pong cycle but has an opposite role compared to Krimper and Qin, as it enforces Siwi loading instead of BmAgo3 loading.

Recently, Arif *et al* (2022) reported that gametocyte-specific factor (Gtsf) proteins stimulate the catalytic activity of PIWI proteins. Gtsf proteins act in the piRNA pathway in different species, including flies, silkworms, and mice (Ipsaro & Joshua-Tor, 2022). In *Drosophila*, Gtsf1 is required for piRNA-mediated transcriptional gene silencing, but Gtsf1 is not essential for piRNA biogenesis (Dönertas *et al*, 2013; Ohtani *et al*, 2013). In contrast, mouse Gtsf1 is involved in piRNA amplification and associates with the mouse PIWI proteins Miwi2 and Mili (Yoshimura *et al*, 2018). Moreover, mouse Gtsf1 can enhance the piRNA-directed target cleavage of both Mili and Miwi *in vitro*. Likewise, silkworm Gtsf1 associates with Siwi and was found to enhance slicing activity (Chen *et al*, 2020; Arif *et al*, 2022; Izumi *et al*, 2022). We note that the conditional cleavage by PIWI proteins, which is dependent on Gtsf, provides an interesting possibility to restrict target cleavage to conditions in which an empty PIWI protein may be available to accept a cleavage product and to prevent RNA cleavage in the absence of such empty PIWI proteins. However, it is not known how the Gtsf-piRISC complex is brought in physical proximity with empty PIWI and target RNA.

In this study, we show that silkworm Gtsf1-like (BmGtsf1L), a Gtsf1 paralog, binds piRNA-loaded BmAgo3. The BmGtsf1L-BmAgo3 piRISC interaction is stimulated by BmVreteno, a protein known to aid Siwi loading following BmAgo3-mediated target cleavage. Surprisingly, we find that BmGtsf1L and BmAgo3 bind to the same eTudor domain of BmVreteno. Using AlphaFold predictions, we uncover a novel binding interface on this eTudor domain that additionally accommodates BmGtsf1L binding. Thus, a single eTudor domain within BmVreteno can serve as a molecular scaffold and interconnect BmGtsf1L and piRISC-BmAgo3 to allow efficient target cleavage only within a context that enables Siwi loading.

Results

BmGtsf1L associates with piRNA-loaded BmAgo3

A previous study showed that Gtsf1 is involved in piRNA-regulated sex determination and transposon silencing in the silkworm (Chen *et al*, 2020). The role of its paralog, BmGtsf1L, however, remained elusive. Alignment of Gtsf proteins from different species shows that BmGtsf1L possesses two conserved CHHC-type zinc (Zn) fingers followed by a short C-terminal tail (Fig EV1A). To find potential binding partners of BmGtsf1L, we transiently expressed HA-tagged BmGtsf1L in BmN4 cells and performed immunoprecipitation followed by quantitative mass spectrometry (IP/qMS). Interestingly, many of the enriched proteins, such as BmAgo3, BmVreteno, and Siwi, are known to play a role in piRNA biogenesis (Fig 1A and Dataset EV2). Next, we transiently expressed HA-BmGtsf1L together with FLAG-tagged BmAgo3, BmVreteno, or Siwi and confirmed that these candidates interact with BmGtsf1L, both in the presence or absence of RNA (Fig 1B). Endogenous BmAgo3 is also associated with transiently expressed BmGtsf1L (Fig 1C).

Next, we generated an anti-BmGtsf1L monoclonal antibody, which detected both endogenous as well as epitope-tagged BmGtsf1L (Fig EV1B). Despite the low expression levels of BmGtsf1L, we could detect endogenous BmGtsf1L specifically in BmAgo3 precipitates (Fig 1D). Unfortunately, the BmGtsf1L antibody was not suitable for immunoprecipitation assays and did not function in immunostainings. To be able to study BmGtsf1L function in further detail, we generated a BmN4 cell line stably expressing HA-BmGtsf1L-eGFP. Using this stable cell line, we confirmed that BmGtsf1L is mostly enriched in BmAgo3 IPs and hardly in Siwi IPs (Fig 1E). *Vice versa*, BmAgo3 and Siwi were both co-precipitated with BmGtsf1L, and again, we observed a stronger enrichment for BmAgo3 (Fig 1F). Using stable cell lines expressing epitope-tagged PIWI proteins (Fig EV1C), we could show that the increased affinity of BmAgo3 for BmGtsf1L was not due to differences in PIWI antibody specificities (Fig EV1D). Moreover, BmSpn-E and BmQin were also co-purified by BmGtsf1L, confirming our initial IP/qMS hits (Fig 1A and F).

To reveal the loading status of endogenous BmAgo3 that associates with BmGtsf1L, we performed BmGtsf1L immunoprecipitation followed by small RNA sequencing. BmGtsf1L-associated small RNA profiles resembled those of BmAgo3-bound small RNAs, showing a clear, defined peak of 27 nucleotides in size, a strong sense bias, and enrichment for adenine at the 10th position (Figs 1G and H, and EV1E). Together, these results indicate that BmGtsf1L associates with piRNA-loaded BmAgo3.

Interestingly, BmVreteno has been shown to also interact with piRNA-loaded BmAgo3, and to do so as a dimer (Nishida *et al*, 2020). Using an independently generated anti-BmVreteno antibody that also detects the BmVreteno-Long (L) and BmVreteno-Short (S) isoforms (Fig EV1F–H), we confirm that BmVreteno retrieves BmAgo3 (Fig EV1I). Consistently, both BmVreteno isoforms were also found in BmAgo3 precipitates (Fig EV1J). Next, we assessed the interaction between BmGtsf1L and endogenous BmVreteno. This revealed that BmGtsf1L also brings down both isoforms of endogenous BmVreteno (Fig 1I). Likewise, BmGtsf1L is also co-precipitated by BmVreteno (Fig 1J). Together, these data suggest that BmVreteno, BmGtsf1L, and piRNA-loaded BmAgo3 form a complex.

BmGtsf1L resides in BmAgo3 bodies

Nishida *et al* (2020) recently described that the formation of BmAgo3 bodies is dependent on BmVreteno. To further analyze whether BmGtsf1L also resides in BmAgo3 bodies we exploited fluorescent lifetime imaging microscopy (FLIM). Using this technique, we observed the co-localization of BmGtsf1L to BmAgo3-BmVreteno

Figure 1. BmGtsf1L associates with piRNA-loaded BmAgo3.

A Anti-HA immunoprecipitation on BmN4 cell lysates where either HA-BmGtsf1L or HA-eGFP was ectopically expressed. The experiment was performed using technical duplicates to perform quantitative mass-spectrometry-based detection of peptides using stable dimethyl isotope labeling. A scatterplot showing the $\log_2$ converted normalized ratio data for the individual label pairs. The threshold was set to 2-fold enrichment; known piRNA factors are indicated (red dots) as well as the bait protein (BmGtsf1L, green dot).

B Anti-HA immunoprecipitation on BmN4 lysates made from cells that were transfected with the indicated constructs either in presence or absence of RNase A/T1. BmGtsf1L was immunoprecipitated followed by Western blot detection using the indicated antibodies. Expression of 3xFLAG-eGFP served as a negative control.

C Anti-HA immunoprecipitation of HA-BmGtsf1L or HA-eGFP from BmN4 cell lysates followed by immunodetection of endogenous BmAgo3. Anti-tubulin probing served as a loading control.

D Immunoprecipitation using the indicated endogenous antibodies or using non-immune serum (n.i.) as a control in the presence or absence of naïve BmN4 cell lysates. Oriole stain was used to detect the immunopurified BmAgo3 and Siwi complexes, whereas retrieval of endogenous BmGtsf1L was verified by Western blot.

E Immunoprecipitation of endogenous BmAgo3 and Siwi complexes on BmN4 cell lysates from the HA-BmGtsf1L-eGFP stable cell line, followed by Western blot detection using the indicated antibodies.

F GFP (BmGtsf1L) or control (Ctrl) immunoprecipitation on BmN4 cell extracts stably expressing HA-BmGtsf1L-eGFP was followed by Western blot detection using the indicated antibodies.

G Small RNA size profiles of input samples and from anti-HA immunopurified complexes. Immunoprecipitations were performed in technical duplicates on BmN4 cell lysates from cells that were transiently transfected, denoted in the two lines on the right-hand panels.

H Violin plot showing the $\log_2$ transformed strand bias of sense to antisense small RNAs from input and IP samples of technical duplicates. The mean strand bias is indicated with the color code, where negative and positive values represent antisense and sense bias, respectively. The boxplot inside the data distribution shows the summary of the data as follows: The top and bottom of the embedded box represent the 75th and 25th percentile of the distribution, respectively, and the line inside the box represents the median; the lines extend to the smallest or largest observation that falls within a distance of the inter-quartile range (IQR), Q1–1.5 × IQR or Q3 + 1.5 × IQR respectively.

I GFP (BmGtsf1L) and control immunoprecipitation, followed by Western blot detection of endogenous BmVreteno using an antibody that detects the two BmVreteno isoforms.

J Reciprocal immunoprecipitation using endogenous anti-BmVreteno antibody or rabbit IgG as an isotype control, followed by immunodetection using the indicated antibodies.

K Fluorescence lifetime imagining of BmN4 cells co-transfected with eGFP-BmVreteno, BmGtsf1L- mOrange2, and mCardinal-BmAgo3. Inset shows the zoom-in of the boxed area. Two representative images of two biological experiments are shown. Scale bars:10 µm. Plots on the right show the intensity gray values for each channel of the line that has been drawn in the inset frame. Quantification data can be found in Dataset EV4.

Source data are available online for this figure.

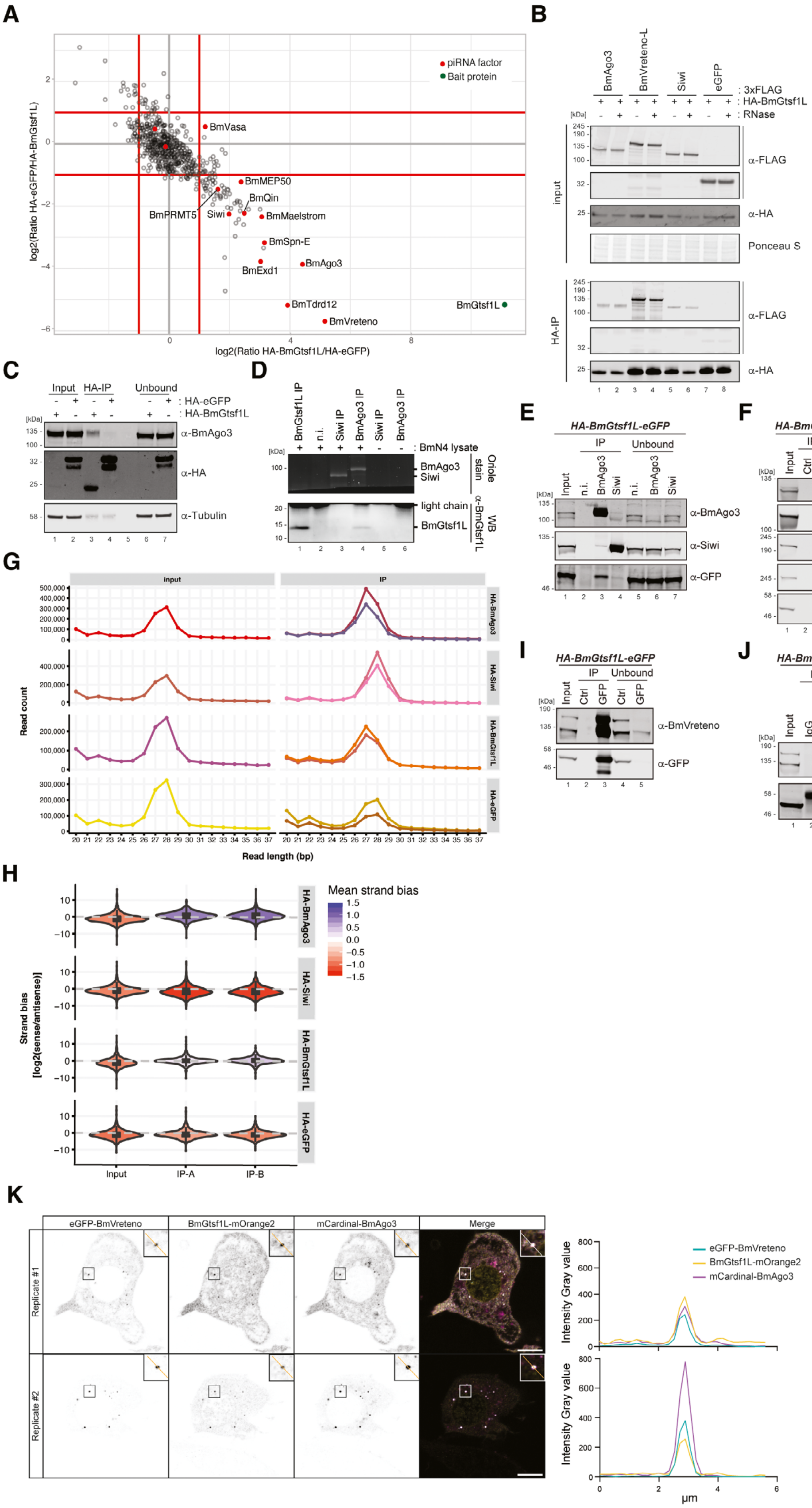

A
piRNA factor
Bait protein
BmVasa
BmMEP50
BmQin
BmPRMT5
Siwi
BmMaelstrom
BmSpn-E
BmAgo3
BmExd1
BmTdrd12
BmGtsf1L
BmVreteno
log2(Ratio HA-eGFP/HA-BmGtsf1L)
log2(Ratio HA-BmGtsf1L/HA-eGFP)
B
BmAgo3
BmVreteno-L
Siwi
eGFP
: 3xFLAG
: HA-BmGtsf1L
: RNase
[kDa]
input
HA-IP
α-FLAG
α-HA
Ponceau S
C
Input
HA-IP
Unbound
: HA-eGFP
: HA-BmGtsf1L
α-BmAgo3
α-HA
α-Tubulin
D
BmGtsf1L IP
n.i.
Siwi IP
BmAgo3 IP
Siwi IP
BmAgo3 IP
: BmN4 lysate
BmAgo3
Siwi
Oriole stain
light chain
BmGtsf1L
WB α-BmGtsf1L
E
HA-BmGtsf1L-eGFP
IP
Unbound
Input
n.i.
BmAgo3
Siwi
α-BmAgo3
α-Siwi
α-GFP
F
HA-BmGtsf1L-eGFP
IP
Unbound
Input
Ctrl
GFP
α-BmAgo3
α-Siwi
α-BmSpn-E
α-BmQin
α-GFP
G
input
IP
HA-BmAgo3
HA-Siwi
HA-BmGtsf1L
HA-eGFP
Read count
Read length (bp)
H
Mean strand bias
Strand bias [log2(sense/antisense)]
HA-BmAgo3
HA-Siwi
HA-BmGtsf1L
HA-eGFP
Input
IP-A
IP-B
I
HA-BmGtsf1L-eGFP
IP
Unbound
Input
Ctrl
GFP
α-BmVreteno
α-GFP
J
HA-BmGtsf1L-eGFP
IP
Unbound
Input
IgG
BmVreteno
α-BmVreteno
α-GFP
K
eGFP-BmVreteno
BmGtsf1L-mOrange2
mCardinal-BmAgo3
Merge
Replicate #1
Replicate #2
Intensity Gray value
µm
eGFP-BmVreteno
BmGtsf1L-mOrange2
mCardinal-BmAgo3

Figure 1.

marked granules (Fig 1K). Single transfection of the individual proteins served as controls and revealed the specificity of each fluorescently tagged protein (Fig EV2A).

Interdependence of BmVreteno-BmAgo3-BmGtsf1L interaction

Next, we assessed the consequences of BmGtsf1L knockdown on the interactions we described above. Knockdown of BmGtsf1L followed by BmVreteno immunoprecipitation revealed that the interaction between BmVreteno and BmAgo3 does not require BmGtsf1L (Fig 2A). *Vice versa*, both isoforms of BmVreteno were also still retrieved by BmAgo3 following BmGtsf1L depletion (Fig 2B). These results indicate that the interaction between BmVreteno and BmAgo3 does not require BmGtsf1L.

To test whether BmAgo3 affects the interaction between BmGtsf1L and BmVreteno, we analyzed their association following efficient BmAgo3 knockdown. This revealed that the interaction between BmGtsf1L and BmVreteno was reduced but not fully abrogated (Fig 2C). Given that no BmAgo3 was detected in this BmGtsf1L IP, these results suggest that BmAgo3 enhances, but is

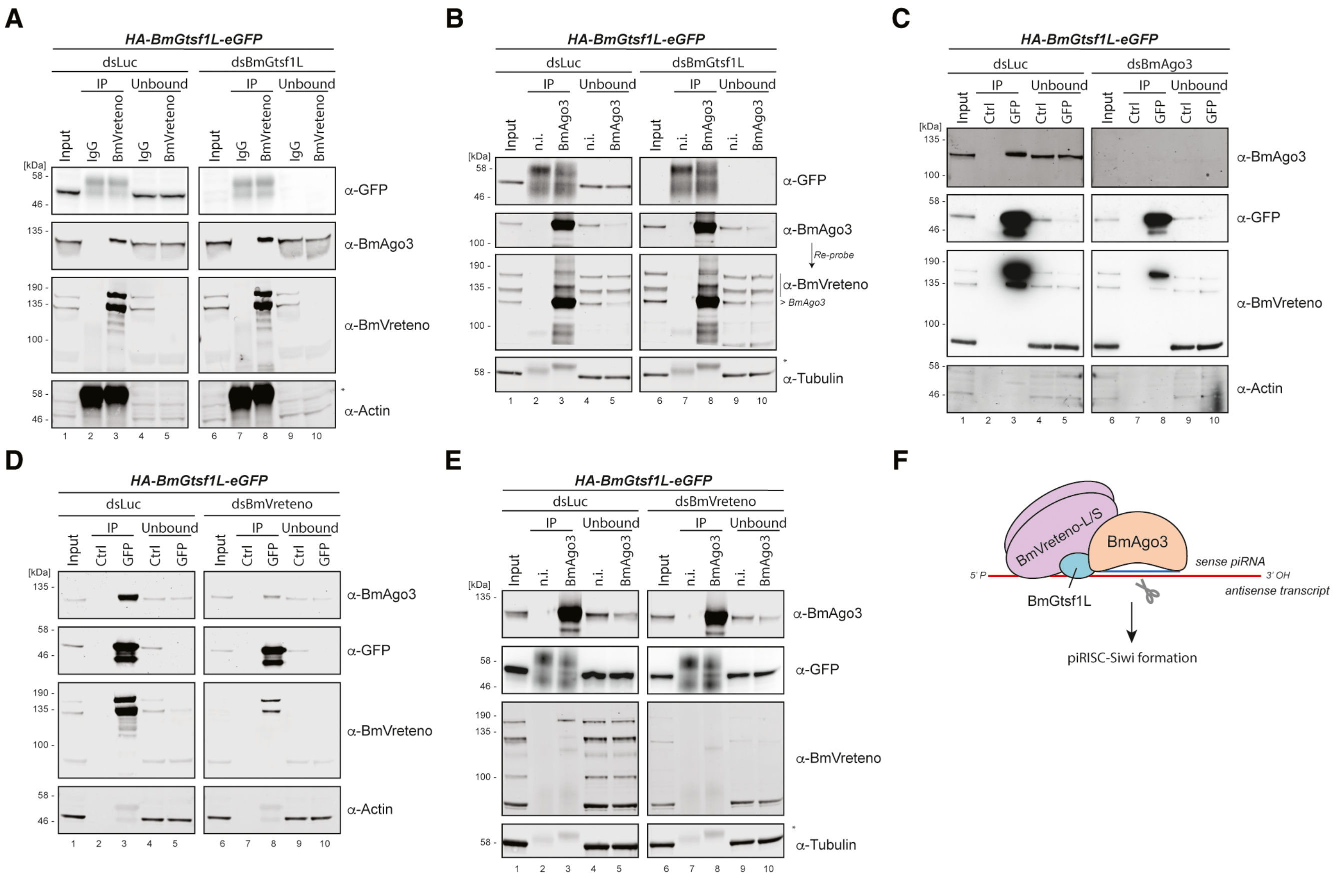


Figure 2. Interdependence of BmVreteno-BmAgo3-BmGtsf1L interaction.

dsRNA-mediated gene depletion on BmN4 cells, which stably express HA-BmGtsf1L-eGFP, followed by immunoprecipitation and Western blot detection.

A Luciferase (dsLuc) control knockdown and BmGtsf1L (dsBmGtsf1L) depletion, followed by IgG control or anti-BmVreteno immunoprecipitations and detection of retrieved proteins by Western blot using the indicated antibodies. Anti-actin detection served as the loading control. An asterisk indicates detection of the antibody light chain.

B Knockdown in BmN4 cells as described in panel (A), followed by non-immune (n.i.) serum control IP or anti-BmAgo3 IP and detection of retrieved proteins by Western blot. Anti-tubulin probing was used as a loading control. An asterisk indicates detection of the antibody light chain.

C Knockdown of luciferase (dsLuc) or BmAgo3 (dsBmAgo3) in BmN4 cells. Immunoprecipitation using GFP (BmGtsf1L) or control (Ctrl) magnetic beads, followed by Western blot detection using the indicated antibodies. Anti-actin probing was performed as a loading control.

D Immunoprecipitation and Western blot detection were performed as described in panel (C), but on BmN4 cell extracts from which endogenous BmVreteno was depleted by dsRNA transfection.

E Knockdown of endogenous BmVreteno, followed by immunoprecipitation using anti-BmAgo3 antibodies or non-immune serum as a control. Western blot detection of precipitated proteins was performed using the indicated antibodies and detection of anti-tubulin served as a loading control. Asterisk indicates detection of the antibody light chain.

F Model on the interconnection between BmGtsf1L, BmAgo3, and BmVreteno. The majority of BmGtsf1L is found in complex with BmVreteno. BmVreteno, which can stimulate the BmGtsf1L-BmAgo3 interaction, whereas BmAgo3 also fosters the BmGtsf1L-BmVreteno interaction. BmVreteno exists as a long (L) and short (S) isoform, which can interact with each other (Nishida *et al*, 2020) and is therefore schematically depicted as a heterodimer.

Source data are available online for this figure.

not essential for the interaction between BmGtsf1L and BmVreteno. Finally, we tested the effects of BmVreteno depletion on the BmAgo3-BmGtsf1L interaction. While the dsRNA treatment strongly affected BmVreteno levels, we were not able to fully eliminate it, as evidenced by its presence in the BmGtsf1L IPs. Nevertheless, RNAi against BmVreteno strongly diminished the amount of BmAgo3 that was brought down by BmGtsf1L (Fig 2D), indicating that BmVreteno stimulates the interaction between BmGtsf1L and BmAgo3. We note, however, that in reciprocal BmAgo3 immunoprecipitations, small amounts of BmGtsf1L could be retrieved, independent of BmVreteno knockdown status (Fig 2E). In this experiment no residual BmVreteno was detected in the IPs. We conclude that a small fraction of BmGtsf1L binds BmAgo3 independent of BmVreteno, but that the majority of BmGtsf1L is found in complex with BmVreteno and that this stimulates the BmGtsf1L-BmAgo3 interaction (Fig 2F).

BmAgo3 interacts with BmVreteno eTD1 via methylated arginine residues

BmVreteno contains two Tudor domains (TDs) that are confidently predicted by Pfam and SMART Hidden Markov Models (HMMs) (Letunic *et al*, 2021; Mistry *et al*, 2021). A third match of the Pfam Tudor HMM around residue 500 exceeded the e-value threshold and thus was not significant. Alignment of the two confidently predicted Tudor domains to those of *Drosophila* Tudor-SN and Tudor-eTD11, for which crystal structures have been resolved (Friberg *et al*, 2009; Liu *et al*, 2010), indicates that BmVreteno TD1 contains an aromatic cage (Fig EV2B). It is well known that aromatic cages within TD domains can bind to methylated arginines of client proteins and thereby mediate protein–protein interactions (Siomi *et al*, 2010; Chen *et al*, 2011). When co-expressing BmVreteno with a BmAgo3 variant that cannot be methylated at its N-terminus (5RK), we lost interaction between both (Fig EV2C), suggesting that arginine methylation is a prerequisite for its association with BmVreteno. This is in line with previous work, which showed that the aromatic cage of TD1 is involved in BmAgo3 interaction (Nishida *et al*, 2020).

In addition, a BmAgo3 piRNA-loading defective mutant (YK > LE) also revealed a strong loss of interaction with BmVreteno (Fig EV2C), which is in line with the observation that unloaded BmAgo3 does not co-localize with BmVreteno and fails to form BmAgo3 bodies (Nishida *et al*, 2020). This could indicate that BmAgo3 becomes methylated only following piRNA-binding, which has been observed for *Drosophila* Aubergine (Webster *et al*, 2015; Huang *et al*, 2021). Notably, a BmAgo3 slicing mutant (DADH) does not show loss of interaction with endogenous BmVreteno (Fig EV2C). Taken together, our data, combined with the findings of Nishida *et al* (2020), suggests that the aromatic cage of the BmVreteno TD1 domain mediates the interaction with the methylated N-terminus of BmAgo3.

The BmGtsf1L C-terminus establishes an interaction with BmVreteno

To understand how BmGtsf1L binds to BmVreteno, we tested which region of BmGtsf1L interacted with endogenous BmVreteno. At the same time, we also probed for BmAgo3. A BmGtsf1L fragment missing the N-terminal part, including the two CHHC Zn fingers, could still retrieve BmVreteno as well as BmAgo3 (Fig 3A and B). By contrast, deletion of the likely disordered BmGtsf1L C-terminus completely abolished the interaction with BmVreteno, while it allowed some interaction with BmAgo3 (Fig 3A and B). Studies in *Drosophila* showed that aromatic residues within the C-terminus of Gtsf1 regulate Piwi binding (Dönertas *et al*, 2013; Ohtani *et al*, 2013). Therefore, we checked for the presence of aromatic residues within BmGtsf1L and studied how the mutagenesis of these residues would affect its interaction with either BmAgo3 or BmVreteno. The BmGtsf1L tyrosine residue mutant (Y88A) retrieved BmAgo3 and BmVreteno to a similar extent as wildtype BmGtsf1L, whereas mutation of the conserved tryptophan residue (W99A) affected the BmAgo3 interaction and completely abrogated the interaction with BmVreteno (Figs 3A and C, and EV1A). The BmGtsf1L double point mutant (YW > AA) displayed similar effects when compared to the W99A single mutant. We could also show that BmGtsf1L (W99A) remained uniformly distributed in the nucleus and cytoplasm, even though BmAgo3 granules were still present (Fig 3D and E). We thus identified the conserved tryptophan residue (W99) within the C-terminal tail of BmGtsf1L to be essential for interaction with BmVreteno and to enhance binding to BmAgo3.

BmVreteno directly interacts with BmGtsf1L

The above results prompted us to test the hypothesis that BmVreteno and BmGtsf1L interact directly. Using an *E. coli* expression system, we succeeded in the expression and purification of full-length BmVreteno-L, BmVreteno-S, and BmGtsf1L. Notably, GST-BmVreteno was eluted as multimeric proteins from gel filtration columns and associated with nucleic acids (Fig EV3A–C). Using these proteins in GSH pull-down assays, we could show that BmGtsf1L interacts directly with both isoforms of BmVreteno (Fig 3F). Furthermore, we could recapitulate the effects of the mutations described above on the BmVreteno-BmGtsf1L interaction *in vitro*: recombinant BmGtsf1L-W99A failed to interact with BmVreteno (Fig 3G). We conclude that the C-terminal end of BmGtsf1L is sufficient to bind directly to BmVreteno and that BmGtsf1L-W99 plays a crucial role in this interaction.

BmGtsf1L binds to BmVreteno eTD1

We next analyzed which region of BmVreteno is involved in its interaction with BmGtsf1L. Using truncation analysis, we found that the C-terminal region of BmVreteno, containing the two PFAM/SMART predicted TDs, was required (Fig EV3D and E). However, purification of fragments containing individual predicted TD domains to probe binding with BmGtsf1L failed. To improve fragment design, we turned to AlphaFold as a novel artificial intelligence-based tool for protein structure prediction. To our surprise, AlphaFold confidently predicted three extended Tudor domain folds within full length BmVreteno, which are also referred to as TSN folds (Fig 4A and Dataset EV1) (Liu *et al*, 2010). Hereafter, we refer to these three domains as AF-eTD0, 1, and 2, where AF-eTD0 corresponds to the newly identified eTD domain. The prediction of this additional Tudor domain is in line with IUPred predictions suggesting that this region is structured (Fig 4A). It also overlaps with the nonsignificant Tudor HMM match from SMART/Pfam mentioned earlier. In addition, AlphaFold predicted with high confidence the structures and boundaries of the RRM and MYND domains

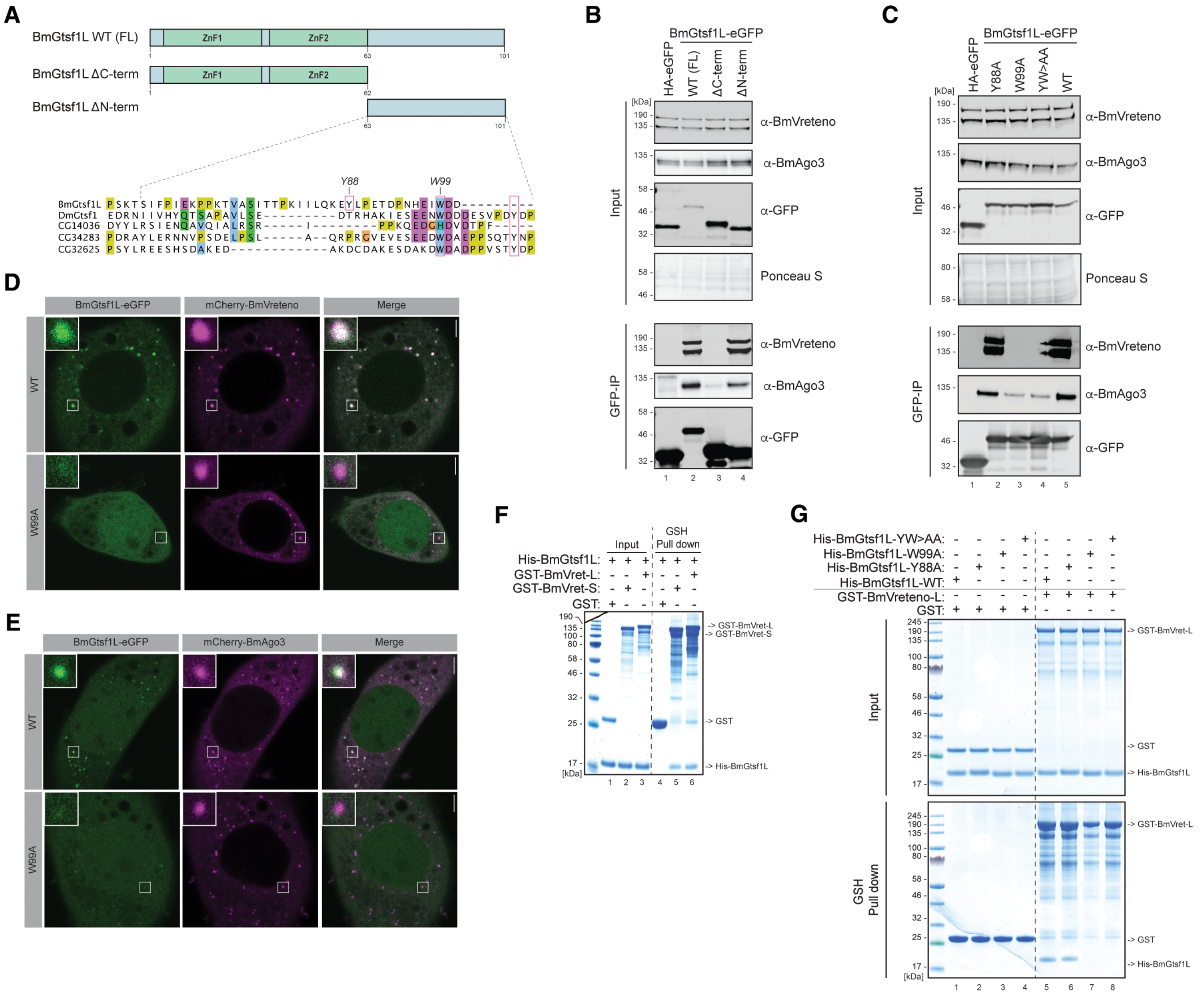


Figure 3. The BmGtsf1L C-terminus establishes a direct interaction with BmVreteno.

A Overview of BmGtsf1L domain architecture with two N-terminally located zinc fingers (ZnF1 and ZnF2, respectively). The two deletion variants used in panel (B) to address binding to BmAgo3 and BmVreteno are also depicted (top). Alignment of BmGtsf1L to GTSF proteins from *Drosophila*. Conserved tryptophan (W) and tyrosine (T) residues of DmGtsf1 that were shown to be involved in Ago3 interaction are boxed in magenta (Dönertas *et al*, 2013). BmGtsf1L contains another aromatic residue (Y88) in its C-terminus (small magenta box), which is not conserved. Clustal Omega alignment was processed with Jalview software (bottom).

B GFP-IP (BmGtsf1L) on BmN4 cell extracts from cells that were transfected with full length (FL) BmGtsf1L-eGFP and with their deletion variants. Transfection of HA-eGFP served as a control. Western blot was performed with indicated antibodies and Ponceau S staining served as a loading control.

C Same as in panel (B) but now BmGtsf1L-eGFP mutants, in which aromatic residues were substituted with alanine, were transfected.

D Single-plane confocal micrographs of BmN4 cells transfected with BmGtsf1L-eGFP wildtype (WT) or the W99A mutant together with mCherry-BmVreteno. Inset shows the zoom-in of the boxed area. Scale bars – 4 µm.

E Single-plane confocal micrographs of BmN4 cells transfected with BmGtsf1L-eGFP wildtype (WT) or the W99A mutant together with mCherry-BmAgo3. Inset shows the zoom-in of the boxed area. Scale bars – 4 µm.

F Analysis of the interaction between BmVreteno and BmGtsf1L by GSH pull-down assays. GST alone or GST-BmVreteno L/S was incubated with His-BmGtsf1L. Input and elution fractions were analyzed by SDS-PAGE followed by Coomassie staining.

G *In vitro* GSH pull-down assay for GST alone or for GST-BmVreteno-L incubated with His-tagged BmGtsf1L variants. Proteins from the SDS-PAGE gel were detected by Coomassie staining.

Source data are available online for this figure.

further upstream of the three eTD domains (Fig 4B). As indicated by the predicted aligned error matrix, AlphaFold is very uncertain about the relative orientation of the RRM, MYND, and AF-eTD0 domain to the rest of the protein (Fig 4B). This suggests that they are unlikely to establish intramolecular contacts with other regions in BmVreteno, while AlphaFold is somewhat more certain about the predicted relative orientation of AF-eTD1 and AF-eTD2 to each other. Based on the predicted AF-eTD boundaries, we designed novel fragments carrying individual eTD domains to test if any of these AF-eTDs mediate the binding to BmGtsf1L. BmN4 cells were transfected with individual HA-eGFP-tagged AF-eTDs of BmVreteno together with mCherry-3xFLAG-BmGtsf1L. Only BmVreteno AF-eTD1 was retrieved in BmGtsf1L immunoprecipitations (Fig 4C). Likewise, recombinant BmGtsf1L was only co-precipitated with now purifiable GST-tagged BmVreteno-AF-eTD1 *in vitro*, while AF-eTD0 and AF-eTD2 could not bind BmGtsf1L (Fig 4D).

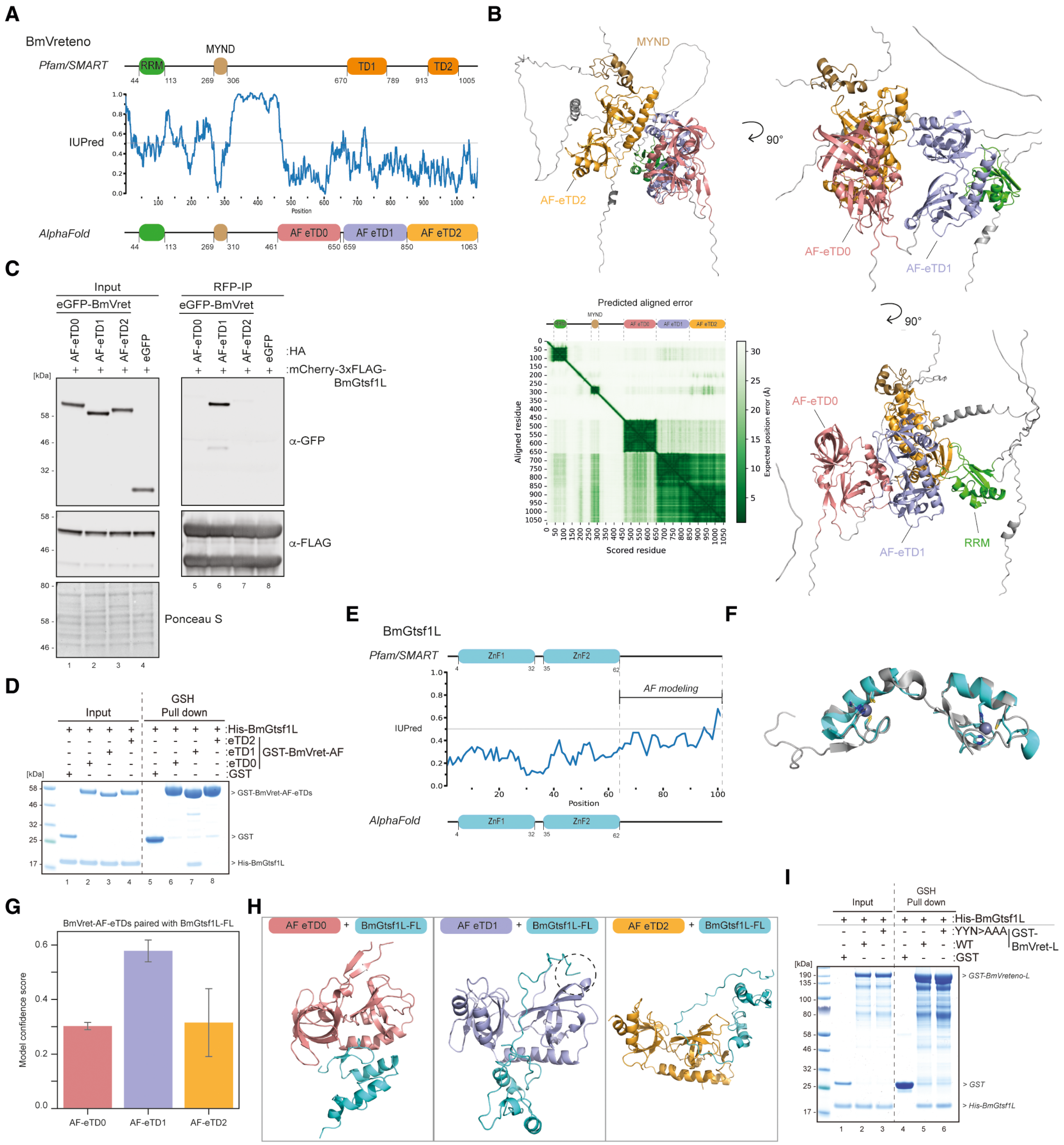


Figure 4.

◀

Figure 4. BmVreteno AF-eTD1 interacts with a C-terminal motif in BmGtsf1L.

A Domain organization of BmVreteno is based on Pfam/SMART annotations (top) or domain annotations from AlphaFold predictions (bottom). IUPred predictions (center) indicate structural disorder propensities for BmVreteno (values > 0.5 indicate disorder). Disorder scores and amino acid positions are shown on the *X*-axis and *Y*-axis, respectively. Abbreviations: RRM = RNA recognition motif; MYND = Myeloid translocation protein 8, Nervy and DEAF-1; TD = Tudor domain; AF-eTD = AlphaFold predicted extended Tudor domain.

B AlphaFold predicted the structure of full-length BmVreteno shown from different angles. Individual domains within the displayed structure are color coded as in panel (A). The BmVreteno domain organization is displayed on top of the PAE matrix. The PAE plot displays the scored residues and aligned residues on the *X*-axis and *Y*-axis, respectively. The expected position error in angstroms (Å) is color coded, where a dark green color indicates low PAE (high confidence) and a white color indicates high PAE (low confidence).

C Transfection of BmN4 cells with mCherry-3xFLAG-BmGtsf1L together with individual eTD domains of BmVreteno, carrying an HA-eGFP tag. The transfection of HA-eGFP served as a control. An RFP immunoprecipitation was performed on BmN4 lysates, and input as well as elution samples were resolved by SDS-PAGE. Proteins were detected by Western blot using the indicated antibodies, and Ponceau S staining served as a loading control.

D Analysis of the interaction between individual eTDs of BmVreteno and BmGtsf1L by GSH pull-down assays. GST alone or GST-BmVreteno-AF-eTDs were incubated with His-BmGtsf1L. Input and elution fractions were analyzed by SDS-PAGE, followed by Coomassie staining.

E Domain organization of BmGtsf1L is based on Pfam/SMART annotations (top) or domain annotations from AlphaFold predictions (bottom). IUPred predictions (center) indicate structural disorder propensities for BmGtsf1L. Disorder scores and amino acid positions are shown on the *X*-axis and *Y*-axis, respectively. The more disordered C-terminal tail was used for AlphaFold predictions (related to Fig 5). Abbreviation: ZnF1 = zinc finger 1; ZnF2 = zinc finger 2.

F Superimposition of the structure of the Zn fingers from MmGtsf1 (in gray, PDB: 6X46) with the predicted structure of the Zn fingers of BmGtsf1L by AlphaFold (in cyan). Zinc-binding residues within MmGtsf1 and BmGtsf1L that coordinate zinc ion-binding are displayed as sticks.

G Bar chart showing the mean of the different model confidence scores that were obtained from AlphaFold predictions (*Y*-axis) using individual AF-eTDs of BmVreteno (*X*-axis) that were paired with full-length BmGtsf1L (error bars indicate the standard deviation of the five predicted models).

H AlphaFold-predicted structures for each individual eTD of BmVreteno with full-length BmGtsf1L. The inset in the middle panel shows that the C-terminal tail of BmGtsf1L establishes contacts with the ordered structure of BmVreteno AF-eTD1.

I Analysis of the interaction between BmVreteno-L full-length wildtype (WT) and the aromatic cage mutant (YYN > AAA) with BmGtsf1L by GSH pull-down assays. GST alone, or GST-BmVreteno-L, was incubated with His-BmGtsf1L. Input and elution fractions were analyzed by SDS-PAGE followed by Coomassie staining.

Source data are available online for this figure.

BmVreteno AF-eTD1 interacts with a C-terminal motif in BmGtsf1L

Results from the previous sections indicate that a region around the W99 residue in the disordered C-terminal tail of BmGtsf1L can bind to AF-eTD1, pointing to the possibility that this interaction is mediated by a so-called short linear motif-folded domain interaction (Van Roey *et al*, 2014). Various reports suggest that AlphaFold has some ability to predict domain-motif interfaces between two submitted protein sequences (Akdel *et al*, 2022; Tsaban *et al*, 2022). However, prior to probing AlphaFold for interface prediction between BmVreteno and BmGtsf1L, we first tested whether AlphaFold could predict the structure of full-length BmGtsf1L with high confidence. AlphaFold confidently predicted the two N-terminal Zn-finger domains and a disordered C-terminal tail in line with Pfam/SMART domain annotations and IUPred disorder propensity predictions (Fig 4E and Dataset EV1). The superimposition of the BmGtsf1L Zn fingers with the resolved structure of mouse Gtsf1 (PDB: 6X46) showed a very similar overall structure (Fig 4F). Each MmGtsf1 Zn finger coordinates the binding of an individual zinc ion (Ipsaro *et al*, 2021). Displaying the zinc-coordinating residues of BmGtsf1L revealed that AlphaFold accurately modeled these residues, despite the fact that AlphaFold cannot model the zinc ions themselves (Fig 4F).

Encouraged by these observations, we submitted full-length BmGtsf1L and full-length BmVreteno for interface prediction by AlphaFold. Unfortunately, predicted structural models were of very low model confidence (at most 0.27) and docked the BmGtsf1L Zn finger domains between AF-eTD0 and AF-eTD2 of BmVreteno in an unlikely mode of binding that also contradicts our experimental results (Dataset EV1).

Next, we submitted sequences of individual AF-eTDs from BmVreteno with the full-length sequence of BmGtsf1L for interface prediction. Interestingly, while predictions involving AF-eTD0 and AF-eTD2 resulted again in low confidence predictions, structural models involving AF-eTD1 resulted in substantially higher model confidences (Fig 4G and Dataset EV1). Inspection of the structural models revealed that AlphaFold predicted binding of a region involving W99 in BmGtsf1L to AF-eTD1 exclusively (Fig 4H), in line with our experimental data. Interestingly, AlphaFold docked W99 of BmGtsf1L into a small hydrophobic pocket of AF-eTD1 that was different from the aromatic cage (see below). Indeed, BmGtsf1L can still interact with BmVreteno *in vitro* when the aromatic cage was disrupted (Fig 4I). This data would suggest that BmAgo3 and BmGtsf1L both bind to BmVreteno AF-eTD1 while using different interaction interfaces on AF-eTD1.

AF predicts BmGtsf1L motif binding to a novel hydrophobic pocket on the BmVreteno AF-eTD1 domain

Despite these encouraging agreements between our experimental data and AlphaFold predictions, the structural models of the interface between AF-eTD1 and full-length BmGtsf1L were still only of moderate model confidence (max. 0.64). To further gain in prediction accuracies, we *in silico* fragmented the unstructured C-terminal tail of BmGtsf1L (39 AA in length), starting off with a fragment of five residues in length at the start, middle, and end of the C-terminal tail and gradually extending these fragments by five residues in each step (Fig 5A). We submitted each fragment individually for interface prediction by AlphaFold with each of the three eTD domains of BmVreteno. This resulted in 72 prediction runs in total (Table EV1). Since the C-terminal fragments were overlapping among each other, we were able to compute the fraction of prediction runs involving a specific pair of residues from BmVreteno and BmGtsf1L where this pair of residues was predicted to be in contact with each other. This computed fraction was visualized as a heatmap between all residues from BmVreteno AF-eTD domains that were observed to be at least once in contact with a residue from the C-terminal tail of BmGtsf1L and *vice versa* (Fig 5A). This residue-residue contact heat map revealed a clear hotspot of residues within AF-eTD1 and residues in BmGtsf1L, including W99, and residues close by that were consistently predicted to be in contact with each other (Fig 5A). No such hotspot was observed for AF-eTD0 and AF-eTD2 suggesting that AlphaFold specifically predicted an interface between AF-eTD1 and

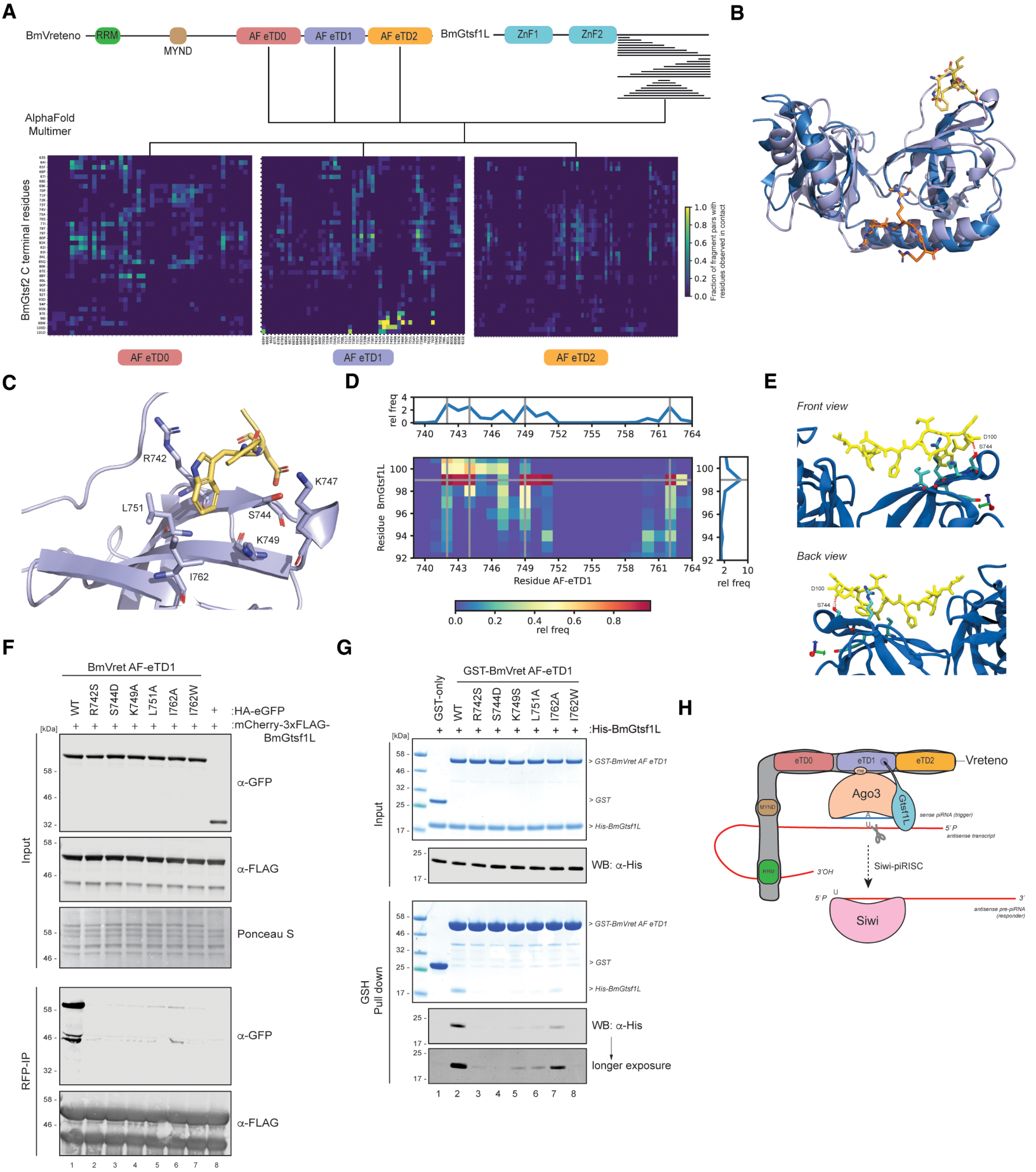


Figure 5.

Figure 5. The Gtsf1L motif binds to a novel hydrophobic pocket on the BmVreteno AF-eTD1 domain.

A AlphaFold-based domain organization of BmVreteno and BmGtsf1L and a schematic overview of the fragmentation approach of the C-terminus of BmGtsf1L. BmGtsf1L fragments were paired for interface predictions with AlphaFold with each of the eTDs of BmVreteno (top). The frequency by which a pair of residues, one from BmVreteno and one from BmGtsf1L, was predicted to be in contact with each other among all fragment pairs submitted to AlphaFold that contained this residue pair is visualized as a heatmap for each individual eTD domain of BmVreteno. Only residues of BmVreteno and BmGtsf1L that were at least observed once to be in contact with a partner residue are displayed on the *X* and *Y*-axis, respectively.

B Superimposition of eTudor11 from *Drosophila* Tudor (in dark blue, PDB: 3NTH) crystallized with a peptide containing a methylated arginine residue of Aubergine (orange) with the structural model of AF-eTD1 (light blue) and the C-terminal five residue-long peptide of BmGtsf1L (in yellow). Peptide residues are represented as sticks.

C Zoom-in on the novel hydrophobic binding pocket of BmVreteno AF-eTD1 (light blue) and contacts between the hydrophobic residues (shown as sticks) with BmGtsf1L W99 and D100 residues (shown as yellow sticks).

D Contact map of BmVreteno AF-eTD1 with the BmGtsf1L 10-AA residue peptide, predicted by atomistic molecular dynamics simulations. The plot summarizes 10 runs of one microsecond each. The blue color in the heatmap indicates a low relative frequency of contacts between the BmGtsf1L-BmVreteno residues, and red indicates a high frequency of contacts. Marginal plots that display the relative frequency (rel freq) show the relative probability of a residue interacting with residues from the binding partner, which is the sum of the probability for each column (for the sum of the contacts along the AF-eTD1 sequence) or row (for the sum of the contacts along the BmGtsf1L 10-AA residue peptide sequence).

E Snapshot on the novel hydrophobic binding pocket of BmVreteno AF-eTD1 (blue) and contacts between residues R742, S744, K749, and I762 (shown as sticks) with BmGtsf1L C-terminal 10-AA residues (shown as yellow sticks). The snapshot additionally displays the BmVreteno-S744 residue that can form a hydrogen bond with the backbone carbonyl of BmGtsf1L-D100.

F Anti-RFP (BmGtsf1L) immunoprecipitation from BmN4 lysates made from cells that were transfected with HA-eGFP-tagged BmVreteno AF-eTD1 variants. Cells were co-transfected with mCherry-BmGtsf1L. The transfection of HA-eGFP served as a control. Proteins from input and elution samples were resolved by SDS-PAGE, followed by Western blot detection using the indicated antibodies. Ponceau-S staining served as a loading control.

G *In vitro* GSH pull-down assay for GST alone or for GST-BmVreteno-AF eTD1 variants incubated with His-tagged BmGtsf1L. Proteins from the input and elution fractions are separated by SDS-PAGE and detected by Coomassie staining. For more sensitive detection, a fraction of the same input and elution samples were subjected to gel electrophoresis in parallel followed by Western blot detection using anti-His antibodies.

H Model showing that a novel binding interface on BmVreteno AF-eTD1 facilitates the binding of BmGtsf1L (via the hydrophobic pocket) and BmAgo3 (aromatic cage).

Source data are available online for this figure.

a motif in BmGtsf1L involving W99. Importantly, model confidences reached 0.87 for these fragment pairs (Table EV1).

We superimposed the structural model involving AF-eTD1 and the last 5 residues of BmGtsf1L (Table EV1) with the solved structure of eTD11 of the *Drosophila* Tudor protein in complex with a synthetic peptide representing the methylated arginine residues of Aubergine (PDB: 3NTH, Fig 5B) (Liu *et al*, 2010). This clearly shows that the predicted interface between BmVreteno and BmGtsf1L does not involve the aromatic cage and lies on the opposite site of AF-eTD1. Closer inspection of the interface revealed that W99 of BmGtsf1L is predicted to bind in a hydrophobic pocket formed by the side chains of R742, K747, K749, L751, and I762 of AF-eTD1 (Fig 5C). Furthermore, the conserved D100 of BmGtsf1L is predicted to be in contact with K722 and R742, suggesting charge–charge contacts. This is also true for BmGtsf1L-D101 and BmVreteno-K747 (Fig 5C). To understand why AlphaFold predictions and experimental results suggest that BmGtsf1L motif-binding is specific to the AF-eTD1 domain of BmVreteno, we superimposed the structural models of all three AF-eTD domains. We observed that the described hydrophobic pocket as well as the aromatic cage are specific to AF-eTD1 (Fig EV4A).

To gain further confidence in the stability of the predicted interface between BmGtsf1L and BmVreteno, we performed atomistic molecular dynamics simulations using the AlphaFold structural model involving AF-eTD1 of BmVreteno and the ten last residues of BmGtsf1L as starting point. In nine out of ten 1 μs simulations, we observed that W99 anchors the BmGtsf1L motif into the predicted hydrophobic pocket of AF-eTD1 (Fig 5D). However, in one of the ten simulations, W99 moves away from the shallow hydrophobic pocket suggesting that additional contacts between both proteins are required to further stabilize the interaction. Our contact analysis of the ten simulation runs demonstrate that the flanking residues I98, D100, and D101 also contribute to anchoring the BmGtsf1L motif, whereas the remaining part of the BmGtsf1L peptide forms fewer contacts with the AF-eTD1 domain and is highly dynamic (Fig 5D and E). On average W99 is interacting with BmVreteno residues R742 and K749 as well as I762. Movie EV1, which visualizes one of the ten trajectories, shows how W99 can be "sandwiched" between the side chains of R742 and K749. The simulations also suggested an important contribution of S744 in AF-eTD1 to BmGtsf1L motif-binding by mostly interacting with W99 but also with D100 (Fig 5E). S744 forms transient hydrogen bonds with I98 and D100 of BmGtsf1L (Figs 5E and EV4B and C), with the proton of the S744 OH group interacting both with the carboxyl group of the D100 side chain and the D100 backbone carbonyl (Fig EV4C). Interestingly, W99 and D100 also display high conservation scores across orthologous Gtsf sequences (Fig EV1A), suggesting that this may be a conserved mode of binding.

Experimental verification of the AF-predicted BmGtsf1L-BmVreteno interface

We set out to further probe this predicted mode of binding using mutagenesis. To this end, we selected residues within AF-eTD1 that contribute to forming the hydrophobic pocket or/and mediate interaction with the aspartate (D100, D101) residues in the BmGtsf1L motif. Individual point mutations (R742S, S744D, K749S, L751A, and I762A) were generated for which we hypothesized that these would perturb the formation of the hydrophobic pocket. We also generated a point mutant (I762W) in which the hydrophobic pocket would be filled and as such would sterically hinder the binding of BmGtsf1L. These mutations were also designed such that an overall impact on the stability and folding of AF-eTD1 should be minimal. Immunoprecipitations on BmN4 cell extracts derived from cells that were co-transfected with BmGtsf1L and the panel of AF-eTD1 mutants, showed that the interaction between BmGtsf1L and AF-eTD1 was abolished in all the variants that we tested (Fig 5F). Using recombinant GST-tagged AF-eTD1 variants in GSH pull-down assays, we also observed that the direct interaction between AF-eTD1 and BmGtsf1L was strongly impaired by these mutations (Fig 5G).

Finally, we studied the effect of these mutations on the BmGtsf1L-BmVreteno interaction in the context of full-length proteins. We assessed these interactions through co-IP experiments, as well as via subcellular localization. Mutations within the hydrophobic pocket on AF-eTD1 were introduced into full -length BmVreteno and were co-transfected with BmGtsf1L. By microscopy, it was apparent that all the hydrophobic pocket mutants resulted in nuclear BmGtsf1L localization (Fig EV4D). This indicates a loss of binding, since BmVreteno wildtype overexpression results in BmGtsf1L exclusion from the nucleus. In addition, we find that BmGtsf1L is still present in granules of BmVreteno mutants, which could be explained by the presence of endogenous BmVreteno that can form a complex with the transiently expressed BmVreteno mutant and suffices to recruit a small fraction of BmGtsf1L. Mutation of the aromatic cage did not result in nuclear BmGtsf1L (Fig EV4D), consistent with it not playing a role in BmGtsf1L binding.

Immunoprecipitations of BmGtsf1L revealed that single point mutations within the hydrophobic pocket had little effect on BmVreteno retrieval (Fig EV4E). More significant reduction in binding could be observed for the serine (S744D) and isoleucine (I762W) mutations that also showed the strongest effects in our *in vitro* assay (Figs EV4E and 5E). A stronger loss of interaction was observed when generating double mutants for the hydrophobic pocket residues that initially showed no or marginal effects when compared to wildtype BmVreteno. We note that the full-length BmVreteno that we transiently expressed likely still dimerizes or oligomerizes with endogenous BmVreteno, which in turn would still be able to interact with endogenous BmAgo3 (via the aromatic cage). This might lead to the observed residual BmGtsf1L binding. Taken together, our data reveal a novel binding interface on BmVreteno AF-eTD1 that facilitates the simultaneous binding of BmGtsf1L and BmAgo3 (Fig 5H).

Discussion

In this study, we show that one of the eTudor domains of BmVreteno acts on its own as a molecular scaffold to bring piRNA-loaded BmAgo3 and BmGtsf1L in close proximity. BmGtsf1L slightly stimulates BmAgo3-directed cleavage of an RNA target (Arif *et al*, 2022; Izumi *et al*, 2022). Concurrently, BmVreteno provides an environment that promotes the handover of the cleaved target to empty Siwi (Murakami *et al*, 2021). Therefore, we propose that the interactions that we identify may help to restrict BmAgo3 cleavage to a molecular surrounding in which its cleavage products can fuel piRNA biogenesis, and to prevent futile BmAgo3 cleavage events. As the targets of BmAgo3 are antisense transcripts, its cleavage activity will not directly contribute to transposon silencing. Only cleavage in the presence of empty Siwi protein will be beneficial to transposon silencing. Therefore, making BmAgo3 cleavage dependent on BmGtsf1L and confining this to the BmVreteno environment would represent an effective way of optimizing BmAgo3 cleavage effectivity.

While it has been revealed that BmVreteno can establish an environment where piRNA-loaded BmAgo3 and empty Siwi are brought together, it is not fully understood how empty Siwi is provided (Nishida *et al*, 2020). Using our BmGtsf1L stable cell line, we did detect an interaction between BmGtsf1L and Siwi, but clearly much weaker than the BmAgo3-BmGtsf1 interaction. Our small RNA sequencing results further suggests that BmGtsf1L interacts with unloaded Siwi. Taken together, we hypothesize that Siwi is retrieved as a tertiary interaction via BmVreteno. It is possible that, in analogy to *Drosophila* Krimper, other eTudor domains of BmVreteno may bind unloaded Siwi (Sato *et al*, 2015; Webster *et al*, 2015). Using AlphaFold modeling, we uncovered in total three eTudor domains in BmVreteno. Two of these do not have an intact aromatic cage, suggesting they may bind empty, unmethylated Siwi.

We and others have shown that BmVreteno is spliced into two isoforms: Long (L) and Short (S). In addition, we know that these two isoforms form heterodimers. *In vitro* RNA cross-linking experiments have revealed that the RRM domain contained within BmVreteno-L binds RNA (Nishida *et al*, 2020). Therefore, it is tempting to speculate that the BmVreteno heterodimer can bring in one target RNA molecule, which is then processed by BmAgo3 and whose 3′-end cleavage product is subsequently loaded onto empty Siwi. Nonetheless, a single BmVreteno-L molecule could in principle also recruit both BmAgo3, Siwi and at least one more protein, so the question of why BmVreteno heterodimerizes remains unanswered.

BmVreteno homologs that are expressed within the germ cells of flies (Vreteno), fish, and mice (Tdrd1) are all required for piRNA biogenesis but have different domain organizations (Reuter *et al*, 2009; Vagin *et al*, 2009; Handler *et al*, 2011; Huang *et al*, 2011; Zamparini *et al*, 2011). *Drosophila* and silkworm Vreteno contain an RRM domain and a MYND domain followed by two or three eTudor domains, respectively. Notably, the expression of two Vreteno isoforms seems to be restricted to silkworm. However, in addition to Vreteno, flies also express a Vreteno-like protein in their ovaries, which is called Veneno (Brosh *et al*, 2022). DmVeneno has a very similar domain architecture compared to DmVreteno but is lacking an N-terminal RRM and, as such, mimics the domain organization of the BmVreteno-S isoform (Nishida *et al*, 2020; Brosh *et al*, 2022). Vreteno and Veneno orthologs can also be found in the mosquito species *Aedes aegypti*. Here, Veneno acts as an adaptor protein that brings the ping-pong partners Piwi5 and Ago3 in close proximity for viral piRNA biogenesis (Joosten *et al*, 2019). It would be interesting to study whether Veneno and Vreteno can dimerize and if they co-localize within the nuage of flies and mosquitoes. The mouse and fish Vreteno homologs (Tdrd1) also lack an RRM domain but contain four eTudor domains instead, raising the question how target RNA is provided within these complexes to facilitate *de novo* piRISC assembly. Multivalent interactions within the nuage that are (in part) established by Tudor domains may play an important role here.

The above examples illustrate that many nuage-residing proteins contain multiple eTudor domains, which contribute to the assembly of this phase-separated structure through the formation of multivalent interactions (Chen *et al*, 2011). Importantly, the depletion of a single nuage component can affect nuage integrity and concurs with a significant reduction in piRNA levels. Interestingly, however, in

C. elegans most eTudor-domain-containing proteins that reside in germ granules only harbor one eTudor domain. So how can a single eTudor domain establish a binding platform to recruit multiple proteins? In this study we reveal that a single eTudor domain (AF-eTD1) of BmVreteno can do so by establishing dual binding interfaces. The aromatic cage facilitates the binding of piRNA-loaded, methylated BmAgo3, whereas the hydrophobic pocket allows for binding of BmGtsf1L. The BmGtsf1L C-terminal residues (W99, D100) that are indispensable for binding to the BmVreteno hydrophobic pocket are broadly conserved. This indicates that Gtsf proteins might have a preserved mode of binding, which corresponds to a novel type of domain-linear motif interaction. However, more structural studies are needed to understand to which extent the hydrophobic pocket is conserved among other eTudor domains. Interestingly, we recently uncovered another, novel binding interface on an eTudor domain of the *C. elegans* protein TOFU-6 (Podvalnaya *et al*, 2023). This implicates that eTudor domains are much more versatile in establishing multivalent interactions than previously anticipated.

In this study, we developed a successful strategy based on AlphaFold for the prediction of protein interaction interfaces involving linear motifs. We note that interface predictions by AlphaFold always return the two protein fragments in contact with each other, making it most of the time very difficult to distinguish good from bad structural models simply by visual inspection. Confidence in reported structural models can be gained from metrics such as the model confidence that is computed by AlphaFold and, as we showed, the recurrent observation of residues predicted in contact with each other when alternating the length of protein fragments submitted for interface prediction. Our work also suggests that interface predictions with AlphaFold using full-length proteins might be unsuccessful but more systematic studies are needed to confirm this. Our results further indicate that AlphaFold is able to extrapolate from the training set of protein structures within the PDB to accurately predict protein interaction interfaces it has never seen before. Physics-based models such as molecular dynamics simulations as we employed here also offer a route to investigate and critically assess the importance of binding interfaces predicted by AlphaFold (preprint: Zhang *et al*, 2023).

A recent study from Arif *et al* (2022) revealed that Gtsf proteins contribute to the piRNA-guided endonuclease activity of PIWI proteins *in vitro*. The authors proposed a model in which the binding of Gtsf would induce a conformational change in the piRISC-PIWI complex upon pairing with its RNA target. In addition, while our manuscript was in preparation, another paper reported that BmGtsf1L associates with BmAgo3 and modestly enhances its slicing activity, whereas BmGtsf1 specifically increases Siwi endonuclease activity (Izumi *et al*, 2022). Thus, the stimulation of catalytic activity of PIWI proteins seems to be a conserved function of Gtsf proteins. We note, however, that a molar excess of Gtsf was required to stimulate PIWI slicing activity (Arif *et al*, 2022) and the stimulatory effect of BmGtsf1L on BmAgo3 was relatively small (Izumi *et al*, 2022). We hypothesize that the catalytic effects of Gtsf proteins on PIWI proteins are generally enhanced by eTudor domains that interact with both the loaded PIWI protein and a Gtsf protein, thereby effectively restricting PIWI cleavage to those environments that contain the required eTudor domain. Indeed, in flies and mouse, conserved aromatic residues within the C-terminus of Gtsf1, that we show to interact with an eTudor domain, contribute to PIWI binding (Dönertas *et al*, 2013; Ohtani *et al*, 2013; Yoshimura *et al*, 2018) and PIWI target cleavage kinetics (Arif *et al*, 2022). It is possible that these reported interactions are in fact mediated via, or at least stimulated by eTudor domains. To firmly test our hypothesis, piRISC kinetics will need to be studied in settings that recapitulate the interactions between PIWI, Gtsf and eTudor domains, using recombinant proteins (Dönertas *et al*, 2013; Ohtani *et al*, 2013; Yoshimura *et al*, 2018; Arif *et al*, 2022).

To conclude, our studies start to address the question of why PIWI proteins evolved the requirement of a co-factor for target cleavage. Given that other Argonaute proteins can efficiently cleave target RNA without such co-factors, it seems reasonable to pose that the Gtsf dependence of PIWI proteins serves a purpose. We propose that Gtsf proteins are required to dictate where and possibly when target RNAs are cleaved by PIWI proteins to allow for piRNA amplification. Interestingly, in flies and mouse Gtsf proteins also contribute to PIWI-induced transcriptional gene silencing. However, the exact role of Gtsf1 in this process, which does not involve target RNA cleavage, still remains elusive (De Fazio *et al*, 2011; Dönertas *et al*, 2013; Ohtani *et al*, 2013; Yoshimura *et al*, 2018). Perhaps Gtsf proteins restrict conformational changes of PIWI proteins upon target recognition to loci of strong homology, preventing the establishment of transcriptional silencing at erroneous loci. Further studies will be required to test these ideas.

Materials and Methods

BmN4 cell culture and transfection

BmN4 cells were cultured at 27°C in IPL-41 insect medium (Gibco) supplemented with 10% FBS (Gibco) and 0.5% Pen-Strep (Gibco). Twenty-four hours prior to transfection, ~4 × 10^6 cells were seeded in a 10-cm dish (using one 10-cm dish for each condition). Cells were transfected with plasmid DNA using X-tremeGene HP (Roche) transfection reagent, according to the manufacturer's instructions. Seventy-two hours post transfection cells were harvested, washed once in 5 ml ice-cold PBS and once more in 1 ml ice-cold PBS. Subsequently, cells were pelleted by centrifugation for 5 min at 500 *g* at 4°C and frozen at −80°C. The BmN4 cell line was obtained from R. Pillai. BmN4 cells were obtained from T. Kusakabe. Further details are available online (https://www.cellosaurus.org/CVCL_Z634). It was not authenticated and was not tested for mycoplasma.

RNAi in BmN4 cells

For preparation of dsRNA, template DNAs were prepared by PCR using primers that contained flanking T7 promoter sequences. Primers for preparation of dsRNA can be found in Supplementary Materials (Dataset EV3). dsRNA was generated by *in vitro* transcription using the HiScribe T7 kit (NEB), according to the manufacturer's instructions. Transcribed RNA was purified by phenol/chloroform extraction, precipitated with ethanol, and annealed in water. For dsRNA-mediated gene knockdown, ~2 × 10^6 BmN4 cells were transfected with 10 µg of dsRNA using X-tremeGene HP. Seventy-two hours after transfection, cells were again transfected with dsRNA and the dsRNA-treatment was repeatedly performed every 3 days for at least three times for BmGtsf1L depletion and four times for BmVreteno or BmAgo3 knockdown.

Generation of stable cell lines

For the generation of 3xFLAG-eGFP, 3xFLAG-BmAgo3, and 3xFLAG-Siwi stable cell lines, ~4 × 10^6 BmN4 cells were seeded in a 10-cm dish. Cells were transfected with 10 μg of plasmid DNA (Dataset EV3) and cultured under Puromycin (Gibco) selection (5 μg/ml) for at least 4 additional weeks. Stable integration of plasmid DNA was verified by Western blot. The HA-BmGtsf1L-eGFP stable cell line was generated in a similar manner. All stable cell lines are polyclonal.

Plasmid construction

For expression of plasmids in BmN4 cells, all genes were PCR amplified using BmN4 cDNA and then cloned into the pBEMBL vector (kind git of Ramesh Pillai), which harbors an OpIE2 promoter and an OpIE2 polyA tail (Xiol *et al*, 2012). The plasmids that were used to generate stable cell lines additionally contain a puromycin cassette, where the BmA3 promoter drives the expression of the puromycin-N-acetyltransferase (*pac*) gene, followed by the OpIE2 polyA sequence.

For recombinant protein expression in *E. coli*, coding sequences were cloned into the pET28a(plus) vector that contains an N-terminal $(HIS)_6$-tag or into the pGEX-6p vector for GST-tagged protein expression (kind gift from H. Ullrich lab). All primers, vector backbones, and detailed cloning strategies can be found in Dataset EV3.

Immunoprecipitations

Directly before use, BmN4 cell pellets were thawed on ice and lysed in 1 ml IP-150 Lysis Buffer (30 mM Hepes [pH7.4], 150 mM KOAc, 2 mM $Mg(OAc)_{2,}$ and 0.1% Igepal freshly supplemented with EDTA-free protease inhibitor cocktail and 5 mM DTT) for 1 h by end-over-end rotation at 4°C. Cells were further lysed by passing the lysate 10 times through a 20-gauge syringe needle followed by five passes through a 30-gauge needle. Cell debris was pelleted by centrifugation at 17,000 *g* for 20 min at 4°C. Supernatant fractions were collected and subjected to immunoprecipitations. In case of RNase treatment, 20 μl of RNaseA/T1 (Thermo Scientific, #EN0551) were added to 1 ml of IP-150 lysis buffer prior to lysis (according to the manufacturer's instructions).

Immunoprecipitations were performed using Pierce™ Anti-HA Magnetic Beads (30 μl bead suspension per reaction, Thermo Fisher, #88836), Anti-FLAG M2 Magnetic Beads (20 μl bead suspension per reaction, Sigma, #M8823) or GFP/RFP-Trap Magnetic Agarose beads (15 μl bead suspension per reaction, Chromotek). When using endogenous antibodies, 3 μg of affinity-purified antibodies were coupled to 15 μl of pre-equilibrated Protein G Dynabeads (30 μl bead suspension, Invitrogen) for one reaction. Normal rabbit IgG (Cell Signaling, #2729) or mouse non-immune serum (n.i.) served as controls. Beads and antibodies were incubated for 1 h at 4°C by end-over-end rotation in 500 μl of IP-150 Lysis Buffer. Bead-conjugated antibodies were then washed three times in 1 ml of IP-150 Lysis Buffer. Equilibrated beads were subsequently incubated with the BmN4 cell lysate and incubated overnight by end-over-end rotation at 4°C. The next day, immunoprecipitated complexes were washed five times using 1 ml of IP-150 Lysis Buffer and were subsequently used for immunodetection using Western Blot analysis.

Western blot

Samples were prepared in 1× Novex NuPage LDS sample buffer (Invitrogen) supplemented with 100 mM DTT and were heated at 95°C for 10 min prior to resolving on a 4–12% Bis-Tris NuPage NOVEX gradient gel (Invitrogen) in 1× Novex NuPAGE MOPS SDS Running Buffer (Invitrogen) at 140 V. For the detection of endogenous BmGtsf1L, proteins were resolved on a 15% Bis-Tris polyacrylamide gel. Separated proteins were transferred to a nitrocellulose membrane (Amersham) overnight at 20 V using 1× NuPAGE Transfer Buffer (Invitrogen) supplemented with 10% methanol. The next day, the membrane was blocked for 1 h in 1× PBS-Tween (0.05%) supplemented with 5% skim milk and incubated for 1 h with primary antibodies diluted in blocking buffer (1:1,000 anti-Flag; 1:1,000 anti-GFP; 1:1,000 anti-HA; 1:1,000 anti-actin, 1:2,500 anti-tubulin; 1:1,000 for all endogenous antibodies). Subsequently, the membrane was washed three times for 5 min in PBS-Tween, prior to 1 h incubation with the secondary antibody, using 1:10,000 IRDye 800CW Goat anti-mouse and IRDye 680LT Donkey anti-rabbit IgG (LI-COR) and imaged on an Odyssey CLx imaging system (LI-COR). Secondary antibodies used for chemiluminescence-based detection were 1:1,000 rat monoclonal anti-mouse Ig HRP (Clone eB144, Mouse TrueBlot ULTRA, Rockland #18-8817-30), 1:10,000 goat anti-rabbit IgG HRP-linked antibody (Cell Signaling Technology, #7074), and 1:10,000 horse anti-mouse IgG, HRP (Cell Signaling Technology, #7076). Chemiluminescence signals were detected using ECL select Western Blotting detection reagent (Cytvia, #GERPN2235) and imaged on a Fusion FX imaging system (Vilber).

Recombinant protein purification

GST alone as well as GST-3C-BmVreteno variants and fragments were expressed from pGEX6p vectors. His6-thrombin-BmGtsf1L variants were expressed in pET vectors. Transformed plasmids were expressed in *E. coli* (BL21 DE3 codon+, Agilent) overnight at 18°C using 0.5 mM IPTG in LB media. Cells were lysed in ice-cold lysis buffer (30 mM Tris-Cl pH 8.0, 500 mM NaCl, 0.5 mM TCEP, 5% glycerol, an EDTA-free complete protease inhibitor cocktail, and an additional 10 mM imidazole pH 8.0 for BmGtsf1L purifications), using a CF1 continuous flow cell disruptor from constant systems at 29 kpsi and cleared by centrifugation at 40,000 *g* for 30 min at 4°C. Recombinant proteins were affinity-purified from cleared lysates using a NGC Quest Plus FPLC system (Biorad) and GSTrap HP (GST-tagged BmVreteno and GST), or HisTrap HP (His6-tagged BmGtsf1L) 5 ml columns (Cytiva), according to the manufacturers protocols. Eluted proteins were concentrated using Amicon spin concentrators (Merck Millipore) and subjected to gel filtration (Superdex 75 and 200 16/60 pg, Cytiva, in 25 mM Na-Hepes, 300 mM NaCl, 10% Glycerol, pH 7.4).

To obtain an untagged BmVreteno (181-386) antigen fragment for immunization, the GST-tagged fragment from the affinity step was digested with 3C protease (1:100 w/w) overnight at 4°C during dialysis (30 mM Tris-Cl pH 8.0, 500 mM NaCl, 1 mM DTT, 5%

glycerol). The digested protein was re-applied to a GSTrap HP 5 ml column to absorb the free GST. The flow through from this step, containing the untagged BmVreteno (181-386) antigen was concentrated using Amicon spin concentrators and subjected to gel filtration (Superdex 75 16/60 pg in PBS). Another round of free GST absorption via GSTrap, followed by gel filtration (Superdex 75 16/60 pg in PBS) was performed to remove residual free GST from the untagged BmVreteno (181-386) antigen.

For all recombinant proteins, peak fractions after the final gel filtration were pooled and protein concentration was determined by using absorbance spectroscopy and the respective extinction coefficient at 280 nm, before aliquots were flash frozen in liquid nitrogen and stored at −80°C.

GSH pull-downs

Glutathione Sepharose 4B beads (Cityva) were equilibrated (20 µl beads suspension for each reaction) by three washes in PBS containing 0.1% Triton-X100 (PBS-T), and the resin was pelleted by mild centrifugation at 1,000 *g* for 2 min at 4°C. Next, 5 µM of GST-BmVreteno was added to the beads together with 10 µM of His-BmGtsf1L, and samples were incubated for 2 h by end-over-end rotation at 4°C. Beads were pelleted by centrifugation at 1,000 *g* for 2 min at 4°C and washed three times with PBS-T. Finally, pelleted beads were resuspended in 25 µl 1× Novex NuPage LDS sample buffer (Invitrogen) supplemented with 100 mM DTT and were heated at 95°C for 5 min prior to resolving (40% of the sample) on a 4–12% Bis-Tris NuPage NOVEX gradient gel (Invitrogen) in 1× Novex NuPAGE MES SDS Running Buffer (Invitrogen) at 180 V. Proteins on the gel were visualized by staining with InstantBlue Coomassie protein stain (Abcam).

Antibodies

Monoclonal antibodies for detection and/or immunoprecipitation of endogenous Siwi, BmAgo3, BmSpn-E, and BmQin were a kind gift from Mikiko Siomi (Nishida *et al*, 2015). Rabbit polyclonal antibodies for BmAgo3 detection were provided by Ramesh Pillai (Xiol *et al*, 2012).

The monoclonal anti-BmGtsf1L antibody was generated in the Siomi lab by immunizing mice with purified GST-tagged full-length BmGtsf1L. Fusing myeloma generated hybridomas as described previously (Nishida *et al*, 2015).

The rabbit polyclonal anti-BmVreteno antibody was generated by immunizing rabbits with the affinity-purified BmVreteno (186–381) antigen (Eurogentec). Two milliliters of sulfolink resin (Thermo Fisher Scientific) were covalently conjugated with 3 mg of GST-tagged BmVreteno (181-386) according to the manufacturer's protocol. Ten milliliters of final bleed of each rabbit serum were incubated with 1 ml of GST-BmVreteno (181-386)-conjugated sulfolink resin at 4°C overnight while rotating. After incubation, the resin was washed with PBS containing 0.1% Triton X-100, followed by a wash with PBS in a gravity-flow poly-prep column (Biorad). Elution of polyclonal antibody species was performed using low pH (100 mM Glycine-Cl, 150 mM NaCl, pH 2.3), followed by immediate neutralization of elution fractions with Tris-Cl pH 8.0. The eluted antibodies were re-buffered using a PD-10 column (PBS, 10% glycerol, 0.05% NaN_3) and concentrated to 1 mg/ml using Amicon spin concentrators before flash freezing in liquid nitrogen and storage at −80°C.

Monoclonal anti-HA was produced in house (clone 12CA5, Core Facility Protein Production). Rabbit polyclonal Anti-HA (Sigma-Aldrich, #SAB4300603), mouse monoclonal anti-Flag M2 (Sigma-Aldrich, #F3165), rabbit polyclonal anti-FLAG (Milipore, #F7425), rabbit polyclonal anti-actin (Sigma-Aldrich, #A5060), mouse monoclonal anti-alpha Tubulin (clone B-5-1-2, Sigma-Aldrich, #T6074), rabbit polyclonal anti-GFP (Origene, #TP401), and mouse monoclonal anti-GFP (clone B-2, Santa Cruz, #sc-9996) are all commercially available.

Sequence alignment

Clustal W (Larkin *et al*, 2007) and Jalview software (Waterhouse *et al*, 2009) was used for protein alignment and visualization.

Microscopy

For co-localization studies, approximately 2 × 10^4 cells were seeded per well in an 8-well µ-slide (Ibidi, #80826). The next day, cells were transfected with 100 ng of each corresponding plasmid using X-tremeGene HP. Twenty-four hours post transfection, live cells were fixed and imaged. Confocal imaging (Figs 1K and EV2A) was performed using a STELLARIS 8 FALCON microscope (Leica Microsystems, Mannheim, Germany) equipped with a White Light Laser (WLL). Images (512 × 512 pixel format, pixel size 180 nm) were acquired with a 63×/1.40NA oil immersion objective, using: Channel 0: 488 nm excitation line and the emission band ranging from 500 to 540 nm using a detector HyD X2; Channel 1: 548 nm excitation line and the emission band ranging from 560 to 590 nm using a detector HyD X4; Channel 2: 600 nm excitation line and the emission band ranging from 620 to 750 nm using a detector HyD R5. A sequential scan was performed line by line, and accumulation mode was set to 300 to have enough photon counts for fluorescence lifetime imagining.

The Leica TCS SP5 with a 60× oil immersion objective lens was used for images in Figs 3D and E, and EV4D.

All images were processed using FIJI (Schindelin *et al*, 2012) and Adobe Illustrator software.

RNA isolation and small RNA sequencing

Per condition, one well of a 6-well plate was seeded with 6 × 10^5 BmN4 cells 24 h prior to transfection with either HA-eGFP, HA-BmAgo3, HA-Siwi, or HA-BmGtsf1L using X-tremeGene HP transfection reagent. Seventy-two hours post transfection, cells were harvested, and an anti-HA immunoprecipitation was performed as described above. The experiment was performed in duplicate. Anti-BmAgo3 and anti-Siwi IPs were performed in IP-500 (30 mM Hepes [pH 7.4], 500 mM KOAc, 2 mM $Mg(OAc)_{2,}$ and 0.1% Igepal, freshly supplemented with an EDTA-free protease inhibitor cocktail and 5 mM DTT). Immunopurified RNAs were extracted from beads by adding 1 ml of Trizol LS (Invitrogen #10296028), according to the manufacturer's instructions. The lysate was incubated at RT for 5 min to allow complete dissociation of the nucleoprotein complex. Next, 200 µl of chloroform was added to 1 ml of lysate, followed by harsh mixing and centrifugation at 12,000 *g* for 15 min at 4°C.

Another round of chloroform extraction was performed, and the aqueous phase was transferred to a fresh tube to which 1 volume (500 µl) of ice-cold isopropanol was added for RNA precipitation. RNA pellets were washed twice in 1 ml of 70% ice-cold ethanol and centrifuged at 7,500 *g* for 10 min at 4°C. The RNA pellet was air-dried and dissolved in nuclease-free water.

NGS library prep was performed with NEXTflex Small RNA-Seq Kit V3 following Step A to Step G of Bioo Scientific's standard protocol (V16.06) using the NEXTFlex 3′ SR Adaptor and 5′ SR Adaptor (5′ rApp/NNNNTGGAATTCTCGGGTGCCAAGG/3ddC/ and 5′ GUUCAGAGUUCUACAGUCCGACGAUCNNNN, respectively). Libraries were prepared with a starting amount of 7 ng and amplified in 25 PCR cycles.

Amplified libraries were purified by running an 8% TBE gel and size-selected for 15–35 nt.

Libraries were profiled on a high-sensitivity DNA chip on a 2100 Bioanalyzer (Agilent Technologies) and quantified using the Qubit dsDNA HS Assay Kit, in a Qubit 2.0 Fluorometer (Life Technologies).

All samples were pooled in an equimolar ratio and sequenced on 1 Highoutput NextSeq 500/550 Flowcell, SR for 1× 84 cycles plus 7 cycles for the index read.

Bioinformatic analyses

The quality of raw sequenced reads was accessed with FastQC, Illumina adapters were then removed with cutadapt (-O 5 -m 28 -M 45), reads with low-quality calls were filtered out with fastq quality_filter (-q 20 -p 100 -Q 33). Using information from unique molecule identifiers (UMIs) added during library preparation, reads with the same sequence (including UMIs) were collapsed to remove putative PCR duplicates using a custom script. Prior to mapping, UMIs were trimmed (seqtk trimfq -b 4 -e 4) and library quality re-assessed with FastQC. Reads were aligned against the silkworm (*Bombyx mori*) genome assembly obtained from lepbase GCA_000151625.1 with bowtie v1.1.1 (-l 40 -n 2 -e 70 -m 1 –tryhard –best –strata –chunkmbs 256 –phred33-quals). The locations of repeat elements were also downloaded from lepbase, repeat masker scaffolds (ASM15162v1), converted to genomic location with rmsk2bed. These locations were used to select reads mapping to repeats by intersecting with bedtools intersect (-wa -wb -bed -f 1.0 -nonamecheck) with either the flags -s or -S to determine which small RNAs map sense or antisense, respectively, to the annotated repeats. After filtering, length profiles were obtained by summarizing the length of these reads. Sense/antisense bias was determined by calculating the ratio of reads mapping in the same or the opposite strand for each annotated repeat - repeats with 10 or fewer mapped reads were excluded. Nucleotide bias of piRNAs was determined by summarizing the number of times a base is present in any given piRNA (read sequence) position.

Mass-spectrometry

About 4 × 10^6 BmN4 cells were transfected with HA-tagged BmGtsf1L or with HA-eGFP, which served as a control to detect nonspecific binders. Cells were harvested 72 h post transfection and an anti-HA immunoprecipitation was performed (as described above) on 4 mg total protein lysate. The experiment was performed using two technical duplicates to perform quantitative mass-spectrometry based detection of unique peptides using stable dimethyl isotope labeling (Hsu *et al*, 2003).

Protein in-gel digestion

Proteins were separated briefly in a 10% NuPAGE Bis-Tris gel, stained with Coomassie blue, and cut into small gel cubes, followed by destaining in 50% ethanol/25 mM ammonium bicarbonate. Afterwards, proteins were reduced in 10 mM DTT at 56°C and alkylated by 50 mM iodoacetamide in the dark at room temperature. Enzymatic digestion of proteins was performed using trypsin (1 µg per sample) in 50 mM TEAB (triethylammonium bicarbonate) overnight at 37°C. Following peptide extraction sequentially using 30 and 100% acetonitrile, the sample volume was reduced in a centrifugal evaporator to remove residual acetonitrile. The sample volume was filled up to 100 µl by adding 100 mM TEAB.

Dimethyl-labeling

Dimethyl-labeling was performed as previously reported (Boersema *et al*, 2009). Briefly, the digested samples were labeled as "Light" or "Heavy" by adding formaldehyde or formaldehyde-d_2, respectively. This was followed by addition of $NaBH_3CN$. Thereafter, the samples were incubated at room temperature with orbital shaking for 1 h. The labeling reaction was quenched by adding ammonia solution. Next, peptides were acidified with formic acid to reach pH ~3. The paired-labeled samples were then combined. The resultant peptide solution was purified by solid-phase extraction in C_{18} StageTips (Rappsilber *et al*, 2003).

Liquid chromatography tandem mass spectrometry

Peptides were separated in an in-house packed 30-cm analytical column (inner diameter: 75 µm; ReproSil-Pur 120 C_{18}-AQ 1.9-µm beads, Dr. Maisch GmbH; heated at 40°C) by online reverse phase chromatography through a 105-min nonlinear gradient of 1.6–32% acetonitrile with 0.1% formic acid at a nanoflow rate of 225 nl/min. The eluted peptides were sprayed directly by electrospray ionization into a Q Exactive Plus Orbitrap mass spectrometer (Thermo Scientific). Mass spectrometry measurement was conducted in data-dependent acquisition mode using a top10 method with one full scan (mass range: 300–1,650 *m/z*; resolution: 70,000, target value: 3 × 10^6, maximum injection time: 20 ms) followed by 10 fragmentation scans via higher energy collision dissociation (HCD; normalized collision energy: 25%, resolution: 17,500, target value: 1 × 10^5, maximum injection time: 120 ms, isolation window: 1.8 *m/z*). Precursor ions of the unassigned or +1 charge state were rejected. Additionally, precursor ions already isolated for fragmentation were dynamically excluded for 20 s.

Mass spectrometry data processing and statistical analysis

Raw data files were processed by the MaxQuant software package (version 1.5.2.8) (Cox & Mann, 2008) using its built-in Andromeda search engine (Cox *et al*, 2011) and default settings. Spectral data were searched against a target-decoy database consisting of the forward and reverse sequences of the bait proteins (HA-eGFP and HA-BmGtsf1L), *Bombyx mori* proteomes (UniProt 18,382 entries; NCBI 29,282 entries) downloaded on 8th January 2018, a collection of self-cloned *Bombyx mori* genes (28 entries), and a list of 245 common contaminants. Corresponding labels were selected for "Light" (DimethLys0 and DimethNter0) and "Heavy" (DimethLys4 and

DimethNter4) labels. A maximum of 3 labeled amino acids per peptide were considered. Trypsin/P specificity was assigned. Carbamidomethylation of cysteine was set as a fixed modification. Oxidation of methionine and acetylation of the protein N-terminus were chosen as variable modifications. A maximum of 2 missed cleavages were tolerated. The minimum peptide length was set to be 7 amino acids. The false discovery rate (FDR) was set to 1% for both peptide and protein identifications.

For protein quantification, minimum ratio count of two was required. Both the unique and razor peptides were used for quantification. The "re-quantify" function was switched on. The "advanced ratio estimation" option was also chosen. Downstream data analysis was performed in R statistical environment. Reverse hits, potential contaminants and protein groups "only identified by site" were filtered out. Protein groups with at least two peptides including at least one unique peptide were retained.

AlphaFold predictions

We used the following sequences for AlphaFold predictions:

BmGtsf1L:

MDDPFVSCPYNPIHRVPRSRLQRHIVKCEWINPTMIACPYNATHRYTQED

MKFHVLNCPSKTSIFPIEKPPKTVASITTPKIILQKEYLPETDPNHEIWDD

BmVreteno:

MSNHSRPQRRREWDPMRDDFNEHTYDVQYADDNAGEQVQLDHTKLYIINI

PRGLSEDGIRAAFSKHGKVLSARLSKNPNKRFAIVQFETASEAKLAMMKM

NGSEPLNLKISIAHKTIRKTQHDNKDRNYSTSRNGHCSRDEASSISSKGW

NMRNLDDVMNNDEIDEIDDMIHEDHDDNLDLELDMLTLKQLKIKEEQLMC

KRRLLLRHAEKRQVAPHSSAGRSVLPDGRIVVRNNANETDSAEVEPSFAG

AGSESLKTPGLERNASRQCVKCGAPADWYCSRCAITPYCSQTCQTRDWTE

RHKSVCHYLAPLKTAGGFEAEATSSKSVSNSTPMRSSHSPPTKQQRGEAD

ETDNKAKNIQEPRQNYHRPSNSGPNKNIPGKNQDPRRPATSREAIEEETE

ERGARNPKPAEATKDKHHPMNPVTFQRRQLKSNPVVDAQPAPREQQQPAA

TRAPEASPTEQRESTRRTLVPDRCLIDSLSEGDVVLVSVELKASECCTKQ

GGYVCLSMHEKYESDYQKLCEDYVLDCEADSDEYKIITGDTFSYLSPEDG

GWYRARALNTTMAALLDGSKVVYLRMNDKVKKLPAKYSGIPEFCCVLNAD

VEVGLNLKCSLLSKTPNGFKVTLENVETEANVGEGEITRWIPEVDYPPPV

KNVPVQRSVEIPEVPRPEIKNKSRVILVDATDVQRVFVRPADTRSQKAFD

NILQDVLLYGTTAEPLKEPPSKGQTVVSKYTDNLHYRALCKRTSVNKNKY

LLEYIEYGNIEITQLNRLYPCPEHLSVTSLASLTSHVQLDTTVGELTPRA

LEYIETIKEEEMILTLSSGGDTAQSGAALVNLTLVKNNDNVNKRIEELCT

PEWKKLELKGVDVIETERLMYGTALDYIELPAAPFDLQVLDEVGLDSGNI

SGCPTNSDYVRYVMTKLPARMREYCESEFGRQPYLPAAEELCIAQLPPSS

EWHRAVVLEQILGPGGGTARVLFVDHGNVAEVPVSSLRKMLAEFVTDLPA

VACQIVIEDFPKQATAEMLAKARRFMSGPDKARAAQLPVRGCDKQDVGIY

AIRVPELLEAMTE

We ran AlphaFold v2.2 (Jumper *et al*, 2021) for all monomeric protein predictions and AlphaFold-Multimer v2.2 (preprint: Evans *et al*, 2022) for all protein complex predictions with the following parameters:

--max_template_date=2020-05-14
--db_preset=full_dbs
--use_gpu_relax=False

For every AlphaFold run, 5 models were predicted with one seed per model by setting the following parameter:

--num_multimer_predictions_per_model=1

Out of the five models generated, we used only the model ranked_0 for further processing and interpretation. We defined two residues, one from each protein fragment, to be in contact with each other in predicted AlphaFold models if at least one heavy atom from one residue is less than 5 Å away from any heavy atom from the other residue. Distance measurements between heavy atoms were obtained using the function cmd.distance from PyMOL. The model confidence was extracted from the ranking_debug json file. The PAE matrix was extracted from the pickle file of the model. We used the software PyMOL (TM) Molecular Graphics System, Version 2.5.0. Copyright (c) Schrodinger, LLC., for the visualization and superimposition of AlphaFold models. The superimposition of the structural model involving AF-eTD1 and the last 5 residues of BmGtsf1L with the solved structure 3NTH was done using the cealign command, where AF-eTD1 was set as the mobile entity and the chain A of 3NTH as the target entity for superimposition. For the superimposition of AlphaFold-predicted BmGtsf1L with the solved structure 6X46, we extracted the two Zn fingers from AlphaFold-predicted BmGtsf1L (residue 8–34 and 35–64 for the two Zn fingers, respectively) and aligned them to the two Zn fingers from chain A of 6X46 (residue 14–41 and 48–75 for the two Zn fingers, respectively). The align command was used for the superimposition where AlphaFold-predicted BmGtsf1L Zn fingers were set as the mobile entities and the Zn fingers from the first ensemble state of 6X46 as the target entities.

IUPred predictions were obtained by submitting full-length sequences to the webserver of IUPred2A (Mészáros *et al*, 2018) and selecting the option IUPred2 long disorder (default) for disorder propensity predictions.

We used the Python libraries, pandas (McKinney, 2010) for data analysis, and Matplotlib (Hunter, 2007) and seaborn (Waskom, 2021) for data visualization.

Molecular dynamics simulations

We ran atomistic molecular dynamics simulations using the AlphaFold structural model involving AF-eTD1 of BmVreteno and the 10 last residues of BmGtsf1L. We used the Amber99SB*-ILDN-q protein force field (Hornak *et al*, 2006; Best & Hummer, 2009; Lindorff-Larsen *et al*, 2010; Best *et al*, 2012) and the TIP4P-D water model (Piana *et al*, 2015). Molecular dynamics simulations were run in GROMACS 2021 (www.gromacs.org) (Abraham *et al*, 2015).

The protein-peptide complex was simulated in a rhombic dodecahedron, with a minimum distance of 12 Å between protein atoms and box edges. One hundred and fifty millimolar NaCl were added to the solvated simulation system. The system was energy minimized and equilibrated for 1 ns in using the Berendsen thermostat and barostat at 300 K and 1 bar (Berendsen *et al*, 1984).

We run 10 independent simulations starting from the equilibrated starting structure, each with a different set of initial velocities. Each of the 10 simulations was run for 1 µs. The Bussi-Donadio-Parinello thermostat was used to maintain a simulation temperature of 300 K (Bussi *et al*, 2007). Parrinello-Rahman barostat was employed to keep pressure at 1 bar (Parrinello & Rahman, 1981). Electrostatics were described by the particle mesh Ewald method (PME). The cut-off for van der Waals interactions was 12 Å.

Simulations were analyzed with the MDAnalysis Python library (Michaud-Agrawal *et al*, 2011; Gowers *et al*, 2016). Two residues were deemed to be in contact if one pair of atoms was within 4.5 Å. The contacts maps from the 10 simulation runs, which were started from the same starting structured were averaged to produce a single contact map. Hydrogen bonds were quantified as described by Smith *et al* (2019).

Data availability

The datasets produced in this study are available in the following databases: (i) The accession number for the smRNA-seq data generated in this study is PRJNA940809 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA940809), (ii) Mass spectrometry proteomics data is available at https://massive.ucsd.edu under accession ID MSV000091404, (iii) All plasmids and reagents are available upon request.

Expanded View for this article is available online.

Acknowledgements

We thank all members of the Ketting Lab for fruitful discussions. We would like to thank Ramesh Pillai for providing plasmids, antibodies, and the BmN4 cell line. We are very grateful to the Siomi Lab for their generous supply of antibodies. The IMB Core Facility Microscopy (Sandra Ritz, Márton Gelléri, and Petri Turunen) is acknowledged for support with confocal microscopy. STELLARIS 8 Falcon Confocal was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 497669232. Support by the IMB Proteomics Core Facility is gratefully acknowledged (instrument funded by DFG INST 247/766-1 FUGG). We wish to thank Mario Dejung in particular, Proteomics Core Facility, for his continuous support and for extensive bioinformatic analysis. Support by the IMB Genomics Core Facility and the use of its NextSeq500 (funded by the DFG – INST 247/870-1 FUGG) are gratefully acknowledged. Support from IMB's IT department and especially help from Christian Diedrich for a local installation of AlphaFold is gratefully acknowledged. The GPU cluster on which all AlphaFold predictions were performed was funded by the Ministry of Science and Health (MWG), Rhineland Palatinate (funding id: TB-Nr.:3658/19). AWB was supported by the Peter und Traudl Engelhorn Foundation, by the Klaus Tschira Boost Fund and the German Scholars Organization (GSO/KT 14), as well as by an SPP-1935 start-up fund from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation). MC Siomi is supported by MEXT KAKENHI Grant Number JP19H05466. This project was funded by SFB1551 Project No. 464588647 of the DFG (Deutsche Forschungsgemeinschaft) for RFK, KL, and LS. This work was further supported by the DFG-funded project LU 2568/1-1 granted to KL. LSS acknowledges support from ReALity (Resilience, Adaptation, and Longevity), M^{3}ODEL, and the Forschungsinitiative des Landes Rheinland-Pfalz. Further, we gratefully acknowledge the advisory services offered and the computing time granted on the supercomputer Mogon II at Johannes Gutenberg University Mainz, which is a member of the AHRP (Alliance for High Performance Computing in Rhineland Palatinate) and the Gauss Alliance e.V. Open Access funding enabled and organized by Projekt DEAL.

Author contributions

Alfred W Bronkhorst: Conceptualization; data curation; supervision; funding acquisition; validation; investigation; visualization; writing – original draft; project administration; writing – review and editing. **Katja Luck:** Conceptualization; resources; software; supervision; funding acquisition; validation; visualization; writing – original draft. **Lukas Stelzl:** Data curation; software; formal analysis; validation; investigation; visualization; writing – original draft. **René F Ketting:** Conceptualization; resources; supervision; funding acquisition; validation; writing – original draft; writing – review and editing. **Shéraz Sadouki:** Investigation. **Martin M Möckel:** Investigation. **Sabine Ruegenberg:** Investigation. **Chop Y Lee:** Data curation; formal analysis; validation; investigation; visualization. **Antonio M de Jesus Domingues:** Data curation; formal analysis; visualization. **Tetsutaro Sumiyoshi:** Investigation; visualization. **Mikiko C Siomi:** Resources; supervision; validation. **Rossana Piccinno:** Data curation; validation; investigation; visualization.

Disclosure and competing interests statement

The authors declare that they have no conflict of interest.

References

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2: 19 – 25

Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, Bryant P, Good LL, Laskowski RA, Pozzati G *et al* (2022) A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* 29: 1056 – 1067

Arif A, Bailey S, Izumi N, Anzelon TA, Ozata DM, Andersson C, Gainetdinov I, MacRae IJ, Tomari Y, Zamore PD (2022) GTSF1 accelerates target RNA cleavage by PIWI-clade Argonaute proteins. *Nature* 608: 618 – 625

Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81: 3684 – 3690

Best RB, Hummer G (2009) Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. *J Phys Chem B* 113: 9004 – 9015

Best RB, de Sancho D, Mittal J (2012) Residue-specific α-helix propensities from molecular simulation. *Biophys J* 102: 1462 – 1467

Boersema PJ, Raijmakers R, Lemeer S, Mohammed S, Heck AJR (2009) Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat Protoc* 4: 484 – 494

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, Hannon GJ (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128: 1089 – 1103

Brosh O, Fabian DK, Cogni R, Tolosana I, Day JP, Olivieri F, Merckx M, Akilli N, Szkuta P, Jiggins FM (2022) A novel transposable element-mediated mechanism causes antiviral resistance in *Drosophila* through truncating the Veneno protein. *Proc Natl Acad Sci USA* 119: e2122026119

Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. *J Chem Phys* 126: 014101

Chen C, Nott TJ, Jin J, Pawson T (2011) Deciphering arginine methylation: Tudor tells the tale. *Nat Rev Mol Cell Biol* 12: 629 – 642

Chen K, Yu Y, Yang D, Yang X, Tang L, Liu Y, Luo X, Walters JR, Liu Z, Xu J *et al* (2020) Gtsf1 is essential for proper female sex determination and transposon silencing in the silkworm, *Bombyx mori*. *PLoS Genet* 16: e1009194

Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26: 1367 – 1372

Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* 10: 1794 – 1805

Czech B, Munafò M, Ciabrelli F, Eastwood EL, Fabry MH, Kneuss E, Hannon GJ (2018) piRNA-guided genome defense: from biogenesis to silencing. *Annu Rev Genet* 52: 131 – 157

De Fazio S, Bartonicek N, Di Giacomo M, Abreu-Goodger C, Sankar A, Funaya C, Antony C, Moreira PN, Enright AJ, O'Carroll D (2011) The endonuclease activity of Mili fuels piRNA amplification that silences LINE1 elements. *Nature* 480: 259 – 263

Dönertas D, Sienski G, Brennecke J (2013) *Drosophila* Gtsf1 is an essential component of the Piwi-mediated transcriptional silencing complex. *Genes Dev* 27: 1693 – 1705

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* https://doi.org/10.1101/2021.10.04.463034 [PREPRINT]

Friberg A, Corsini L, Mourão A, Sattler M (2009) Structure and ligand binding of the extended Tudor domain of D. melanogaster Tudor-SN. *J Mol Biol* 387: 921 – 934

Gainetdinov I, Colpan C, Arif A, Cecchini K, Zamore PD (2018) A single mechanism of biogenesis, initiated and directed by PIWI proteins, explains piRNA production in most animals. *Mol Cell* 71: 775 – 790

Ghildiyal M, Zamore PD (2009) Small silencing RNAs: an expanding universe. *Nat Rev Genet* 10: 94 – 108

Gowers R, Linke M, Barnoud J, Reddy T, Melo M, Seyler S, Domański J, Dotson D, Buchoux S, Kenney I *et al* (2016) MDAnalysis: a Python package for the rapid analysis of molecular dynamics simulations. pp 98–105

Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, Siomi H, Siomi MC (2007) A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* 315: 1587 – 1590

Han BW, Wang W, Li C, Weng Z, Zamore PD (2015) Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. *Science* 348: 817 – 821

Handler D, Olivieri D, Novatchkova M, Gruber FS, Meixner K, Mechtler K, Stark A, Sachidanandam R, Brennecke J (2011) A systematic analysis of *Drosophila* TUDOR domain-containing proteins identifies Vreteno and the Tdrd12 family as essential primary piRNA pathway factors. *EMBO J* 30: 3977 – 3993

Hayashi R, Schnabl J, Handler D, Mohn F, Ameres SL, Brennecke J (2016) Genetic and mechanistic diversity of piRNA 3′-end formation. *Nature* 539: 588 – 592

Homolka D, Pandey RR, Goriaux C, Brasset E, Vaury C, Sachidanandam R, Fauvarque M-O, Pillai RS (2015) PIWI slicing and RNA elements in precursors instruct directional primary piRNA biogenesis. *Cell Rep* 12: 418 – 428

Hornak V, Abel R, Okur A, Strockbine B, Roitberg A, Simmerling C (2006) Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65: 712 – 725

Horwich MD, Li C, Matranga C, Vagin V, Farley G, Wang P, Zamore PD (2007) The *Drosophila* RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr Biol* 17: 1265 – 1272

Hsu J-L, Huang S-Y, Chow N-H, Chen S-H (2003) Stable-isotope dimethyl labeling for quantitative proteomics. *Anal Chem* 75: 6843 – 6852

Huang H-Y, Houwing S, Kaaij LJT, Meppelink A, Redl S, Gauci S, Vos H, Draper BW, Moens CB, Burgering BM *et al* (2011) Tdrd1 acts as a molecular scaffold for Piwi proteins and piRNA targets in zebrafish. *EMBO J* 30: 3298 – 3308

Huang X, Hu H, Webster A, Zou F, Du J, Patel DJ, Sachidanandam R, Toth KF, Aravin AA, Li S (2021) Binding of guide piRNA triggers methylation of the unstructured N-terminal region of Aub leading to assembly of the piRNA amplification complex. *Nat Commun* 12: 4061

Hunter JD (2007) Matplotlib: a 2D graphics environment. *Comput Sci Eng* 9: 90 – 95

Ipsaro JJ, Joshua-Tor L (2022) Developmental roles and molecular mechanisms of Asterix/GTSF1. *Wiley Interdiscip Rev RNA* 13: e1716

Ipsaro JJ, O'Brien PA, Bhattacharya S, Palmer AG, Joshua-Tor L (2021) Asterix/Gtsf1 links tRNAs and piRNA silencing of retrotransposons. *Cell Rep* 34: 108914

Izumi N, Shoji K, Sakaguchi Y, Honda S, Kirino Y, Suzuki T, Katsuma S, Tomari Y (2016) Identification and functional analysis of the pre-piRNA 3′ trimmer in silkworms. *Cell* 164: 962 – 973

Izumi N, Shoji K, Suzuki Y, Katsuma S, Tomari Y (2020) Zucchini consensus motifs determine the mechanism of pre-piRNA production. *Nature* 578: 311 – 316

Izumi N, Shoji K, Kiuchi T, Katsuma S, Tomari Y (2022) The two Gtsf paralogs in silkworms orthogonally activate their partner PIWI proteins for target cleavage. *RNA* 29: 18 – 29

Joosten J, Miesen P, Taşköprü E, Pennings B, Jansen PWTC, Huynen MA, Vermeulen M, Van Rij RP (2019) The Tudor protein Veneno assembles the ping-pong amplification complex that produces viral piRNAs in Aedes mosquitoes. *Nucleic Acids Res* 47: 2546 – 2559

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

Kawaoka S, Hayashi N, Suzuki Y, Abe H, Sugano S, Tomari Y, Shimada T, Katsuma S (2009) The Bombyx ovary-derived cell line endogenously expresses PIWI/PIWI-interacting RNA complexes. *RNA* 15: 1258 – 1264

Kawaoka S, Izumi N, Katsuma S, Tomari Y (2011) 3′ end formation of PIWI-interacting RNAs *in vitro*. *Mol Cell* 43: 1015 – 1022

Kirino Y, Vourekas A, Kim N, de Lima AF, Rappsilber J, Klein PS, Jongens TA, Mourelatos Z (2010) Arginine methylation of vasa protein is conserved across phyla. *J Biol Chem* 285: 8148 – 8154

Kiuchi T, Koga H, Kawamoto M, Shoji K, Sakai H, Arai Y, Ishihara G, Kawaoka S, Sugano S, Shimada T *et al* (2014) A single female-specific piRNA is the primary determiner of sex in the silkworm. *Nature* 509: 633 – 636

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R *et al* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947 – 2948

Letunic I, Khedkar S, Bork P (2021) SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res* 49: D458 – D460

Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis JL, Dror RO, Shaw DE (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78: 1950 – 1958

Liu H, Wang J-YS, Huang Y, Li Z, Gong W, Lehmann R, Xu R-M (2010) Structural basis for methylarginine-dependent recognition of Aubergine by Tudor. *Genes Dev* 24: 1876 – 1881

McKinney W (2010) Data structures for statistical computing in Python. In *Proceedings of the 9th* Python in Science Conference, van der Walt S, Millman J (eds), pp 56 – 61. Austin, TX: SciPy

Mészáros B, Erdős G, Dosztányi Z (2018) IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res* 46: W329 – W337

Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O (2011) MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* 32: 2319 – 2327

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ *et al* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res* 49: D412 – D419

Mohn F, Handler D, Brennecke J (2015) Noncoding RNA. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. *Science* 348: 812 – 817

Murakami R, Sumiyoshi T, Negishi L, Siomi MC (2021) DEAD-box polypeptide 43 facilitates piRNA amplification by actively liberating RNA from Ago3-piRISC. *EMBO Rep* 22: e51313

Nishida KM, Iwasaki YW, Murota Y, Nagao A, Mannen T, Kato Y, Siomi H, Siomi MC (2015) Respective functions of two distinct Siwi complexes assembled during PIWI-interacting RNA biogenesis in Bombyx germ cells. *Cell Rep* 10: 193 – 203

Nishida KM, Sakakibara K, Sumiyoshi T, Yamazaki H, Mannen T, Kawamura T, Kodama T, Siomi MC (2020) Siwi levels reversibly regulate secondary piRISC biogenesis by affecting Ago3 body morphology in *Bombyx mori*. *EMBO J* 39: e105130

Nott TJ, Petsalaki E, Farber P, Jervis D, Fussner E, Plochowietz A, Craggs TD, Bazett-Jones DP, Pawson T, Forman-Kay JD *et al* (2015) Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. *Mol Cell* 57: 936 – 947

Ohtani H, Iwasaki YW, Shibuya A, Siomi H, Siomi MC, Saito K (2013) DmGTSF1 is necessary for Piwi-piRISC-mediated transcriptional transposon silencing in the *Drosophila* ovary. *Genes Dev* 27: 1656 – 1661

Ozata DM, Gainetdinov I, Zoch A, O'Carroll D, Zamore PD (2019) PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet* 20: 89 – 108

Parrinello M, Rahman A (1981) Polymorphic transitions in single crystals: a new molecular dynamics method. *J Appl Phys* 52: 7182 – 7190

Piana S, Donchev AG, Robustelli P, Shaw DE (2015) Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J Phys Chem B* 119: 5113 – 5123

Podvalnaya N, Bronkhorst AW, Lichtenberger R, Hellmann S, Nischwitz E, Falk T, Karaulanov E, Butter F, Falk S, Ketting RF (2023) piRNA processing by a trimeric Schlafen-domain nuclease. *Nature* 622: 402 – 409

Rappsilber J, Ishihama Y, Mann M (2003) Stop and go extraction tips for matrix-assisted laser desorption/ionization, nanoelectrospray, and LC/MS sample pretreatment in proteomics. *Anal Chem* 75: 663 – 670

Reuter M, Chuma S, Tanaka T, Franz T, Stark A, Pillai RS (2009) Loss of the Mili-interacting Tudor domain-containing protein-1 activates transposons and alters the Mili-associated small RNA profile. *Nat Struct Mol Biol* 16: 639 – 646

Sato K, Iwasaki YW, Shibuya A, Carninci P, Tsuchizawa Y, Ishizu H, Siomi MC, Siomi H (2015) Krimper enforces an antisense bias on piRNA pools by binding AGO3 in the *Drosophila* germline. *Mol Cell* 59: 553 – 563

Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S, Rueden C, Saalfeld S, Schmid B *et al* (2012) Fiji: an open-source platform for biological-image analysis. *Nat Methods* 9: 676 – 682

Siomi MC, Mannen T, Siomi H (2010) How does the royal family of Tudor rule the PIWI-interacting RNA pathway? *Genes Dev* 24: 636 – 646

Siomi MC, Sato K, Pezic D, Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nat Rev Mol Cell Biol* 12: 246 – 258

Smith P, Ziolek RM, Gazzarrini E, Owen DM, Lorenz CD (2019) On the interaction of hyaluronic acid with synovial fluid lipid membranes. *Phys Chem Chem Phys* 21: 9845 – 9857

Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O (2022) Harnessing protein folding neural networks for peptide-protein docking. *Nat Commun* 13: 176

Vagin VV, Wohlschlegel J, Qu J, Jonsson Z, Huang X, Chuma S, Girard A, Sachidanandam R, Hannon GJ, Aravin AA (2009) Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. *Genes Dev* 23: 1749 – 1762

Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, Gibson TJ, Davey NE (2014) Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev* 114: 6733 – 6778

Waskom ML (2021) seaborn: statistical data visualization. *J Open Source Softw* 6: 3021

Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25: 1189 – 1191

Webster A, Li S, Hur JK, Wachsmuth M, Bois JS, Perkins EM, Patel DJ, Aravin AA (2015) Aub and Ago3 are recruited to nuage through two mechanisms to form a Ping-Pong complex assembled by Krimper. *Mol Cell* 59: 564 – 575

Xiol J, Cora E, Koglgruber R, Chuma S, Subramanian S, Hosokawa M, Reuter M, Yang Z, Berninger P, Palencia A *et al* (2012) A role for Fkbp6 and the chaperone machinery in piRNA amplification and transposon silencing. *Mol Cell* 47: 970 – 979

Xiol J, Spinelli P, Laussmann MA, Homolka D, Yang Z, Cora E, Couté Y, Conn S, Kadlec J, Sachidanandam R *et al* (2014) RNA clamping by Vasa assembles a piRNA amplifier complex on transposon transcripts. *Cell* 157: 1698 – 1711

Yoshimura T, Watanabe T, Kuramochi-Miyagawa S, Takemoto N, Shiromoto Y, Kudo A, Kanai-Azuma M, Tashiro F, Miyazaki S, Katanaya A *et al* (2018) Mouse GTSF1 is an essential factor for secondary piRNA biogenesis. *EMBO Rep* 19: e42054

Zamparini AL, Davis MY, Malone CD, Vieira E, Zavadil J, Sachidanandam R, Hannon GJ, Lehmann R (2011) Vreteno, a gonad-specific protein, is essential for germline development and primary piRNA biogenesis in *Drosophila*. *Development* 138: 4039 – 4050

Zhang Z, Xu J, Koppetsch BS, Wang J, Tipping C, Ma S, Weng Z, Theurkauf WE, Zamore PD (2011) Heterotypic piRNA Ping-Pong requires qin, a protein with both E3 ligase and Tudor domains. *Mol Cell* 44: 572 – 584

Zhang Z, Koppetsch BS, Wang J, Tipping C, Weng Z, Theurkauf WE, Zamore PD (2014) Antisense piRNA amplification, but not piRNA production or nuage assembly, requires the Tudor-domain protein Qin. *EMBO J* 33: 536 – 539

Zhang I, Rufa DA, Pulido I, Henry MM, Rosen LE, Hauser K, Singh S, Chodera JD (2023) Identifying and overcoming the sampling challenges in relative binding free energy calculations of a model protein:protein complex. *bioRxiv* https://doi.org/10.1101/2023.03.07.530278 [PREPRINT]

4.2.1 Supplementary material

Expanded View Figures

Figure EV1. BmGtsf1L and BmVreteno both interact with BmAgo3.

A Domain organization (top) and ClustalW alignment of GTSF proteins from different species. The alignment and conservation scores are depicted using the Jalview software. Residues that are highlighted in blue reveal a 20% identity threshold

B Western blot detection using the mouse monoclonal anti-BmGtsf1L antibody on naïve BmN4 cell extracts or on BmN4 cells that were transfected with FLAG-BmGtsf1L.

C Validation of stable integration of FLAG-Siwi, FLAG-BmAgo3, or FLAG-eGFP expression cassettes into BmN4 cells after extensive puromycin selection by Western blot using the indicated antibodies. Anti-actin probing served as a loading control.

D Control or GFP (BmGtsf1L) immunoprecipitation on BmN4 cell extracts from FLAG-PIWI stable cells that were co-transfected with BmGtsf1L-eGFP. Western blot was performed using the indicated antibodies, and anti-actin probing as well as Ponceau S staining served as loading controls.

E Nucleotide composition of small RNAs that were sequenced from input samples or from anti-HA immunoprecipitated samples.

F Pfam/SMART-based domain organization of the BmVreteno-Long and BmVreteno-Short isoforms, showing the RNA-recognition motif (RRM), Myeloid translocation protein 8, Nervy and DEAF-1 (MYND) domain, and two C-terminal Tudor domains (TD).

G Western blot using the rabbit polyclonal anti-BmVreteno antibody on cell extracts from BmN4 cells that were either untransfected or transfected with HA-BmVreteno (L)-FL. Anti-tubulin probing served as a loading control.

H Validation of anti-BmVreteno antibody specificity on cell extracts from BmN4 cells that were transfected four consecutive times with dsRNA against Luciferase (Luc) or against BmVreteno.

I IgG or anti-BmVreteno immunoprecipitation on naïve BmN4 cells, followed by Western blot detection of endogenous BmVreteno and BmAgo3.

J Reciprocal IP on BmN4 cell extracts using non-immune (n.i.) serum or anti-BmAgo3 antibodies as well as endogenous BmVreteno antibodies for Western blot detection of retrieved proteins.

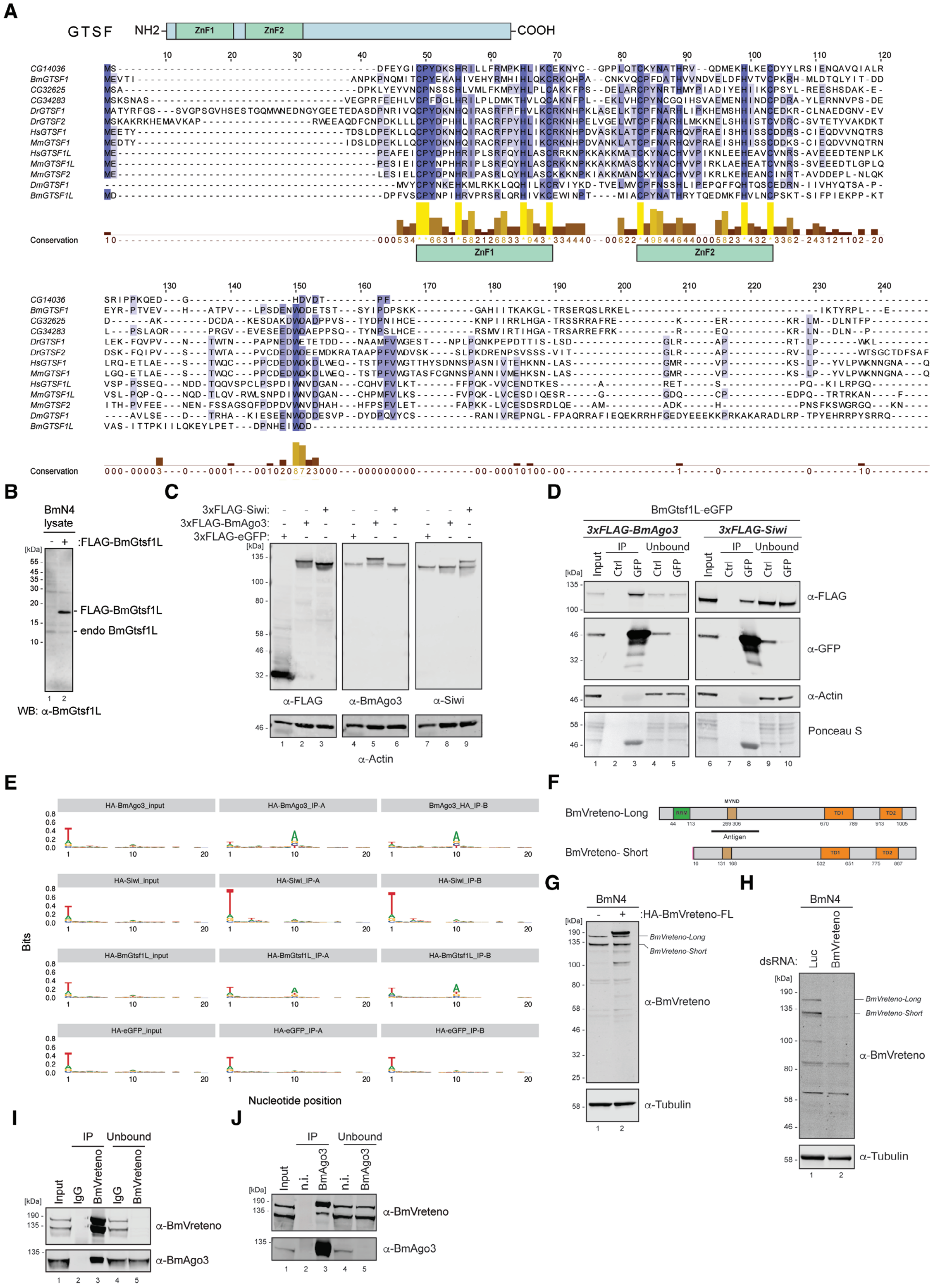

A
GTSF
NH2
ZnF1
ZnF2
COOH
CG14036
BmGTSF1
CG32625
CG34283
DrGTSF1
DrGTSF2
HsGTSF1
MmGTSF1
HsGTSF1L
MmGTSF1L
MmGTSF2
DmGTSF1
BmGTSF1L
Conservation
B
BmN4 lysate
:FLAG-BmGtsf1L
FLAG-BmGtsf1L
endo BmGtsf1L
WB: α-BmGtsf1L
C
3xFLAG-Siwi:
3xFLAG-BmAgo3:
3xFLAG-eGFP:
α-FLAG
α-BmAgo3
α-Siwi
α-Actin
D
BmGtsf1L-eGFP
3xFLAG-BmAgo3
3xFLAG-Siwi
IP
Unbound
Input
Ctrl
GFP
α-FLAG
α-GFP
α-Actin
Ponceau S
E
HA-BmAgo3_input
HA-BmAgo3_IP-A
BmAgo3_HA_IP-B
HA-Siwi_input
HA-Siwi_IP-A
HA-Siwi_IP-B
HA-BmGtsf1L_input
HA-BmGtsf1L_IP-A
HA-BmGtsf1L_IP-B
HA-eGFP_input
HA-eGFP_IP-A
HA-eGFP_IP-B
Bits
Nucleotide position
F
BmVreteno-Long
BmVreteno- Short
RRV
MYND
TD1
TD2
Antigen
G
BmN4
:HA-BmVreteno-FL
BmVreteno-Long
BmVreteno-Short
α-BmVreteno
α-Tubulin
H
BmN4
dsRNA:
Luc
BmVreteno
BmVreteno-Long
BmVreteno-Short
α-BmVreteno
α-Tubulin
I
IP
Unbound
Input
IgG
BmVreteno
α-BmVreteno
α-BmAgo3
J
IP
Unbound
Input
n.i.
BmAgo3
α-BmVreteno
α-BmAgo3
[kDa]

Figure EV1.

Figure EV2. BmAgo3 interacts with BmVreteno eTD1 via methylated arginine residues.

A Fluorescence lifetime imaging of BmN4 cells (using all channels) transfected with either eGFP-BmVreteno, BmGtsf1L- mOrange2, or mCardinal-BmAgo3. Inset shows the zoom-in of the boxed area. Two representative images of two biological experiments are shown. Scale bars: 10 μm. Plots on the right show the normalized intensity (gray values) for each channel of the line that has been drawn in the inset frame. Quantification data can be found in Dataset EV4.

B Multiple sequence alignments of Tudor domains expressed in BmVreteno and its orthologue in *Drosophila* (DmVret). In addition, eTudor11 from the *Drosophila* Tudor protein as well as the eTudor domain of *Drosophila* Tudor-SN (p100), for which crystal structures have been resolved (PDB: 3NTH and 2WAC, respectively), were included. Alignments were performed using Clustal Omega and processed using Jalview software. Aromatic cage residues are depicted as green boxes, and the asparagine residue that is involved in directly binding to the methylated arginine residue (sDMA) is highlighted in yellow. Identical residues (*), conserved substitutions (:) or substitutions by weakly similar residues (.) are indicated below the alignment.

C Anti-FLAG immunoprecipitation of FLAG-BmAgo3 variants that were transiently expressed in the HA-BmGtsf1L-eGFP stable BmN4 cell line. The transfection of 3xFLAG-mCherry served as a control. Immunoprecipitated proteins were analyzed by Western blot using the indicated antibodies, whereas Ponceau S staining served as a loading control.

▶

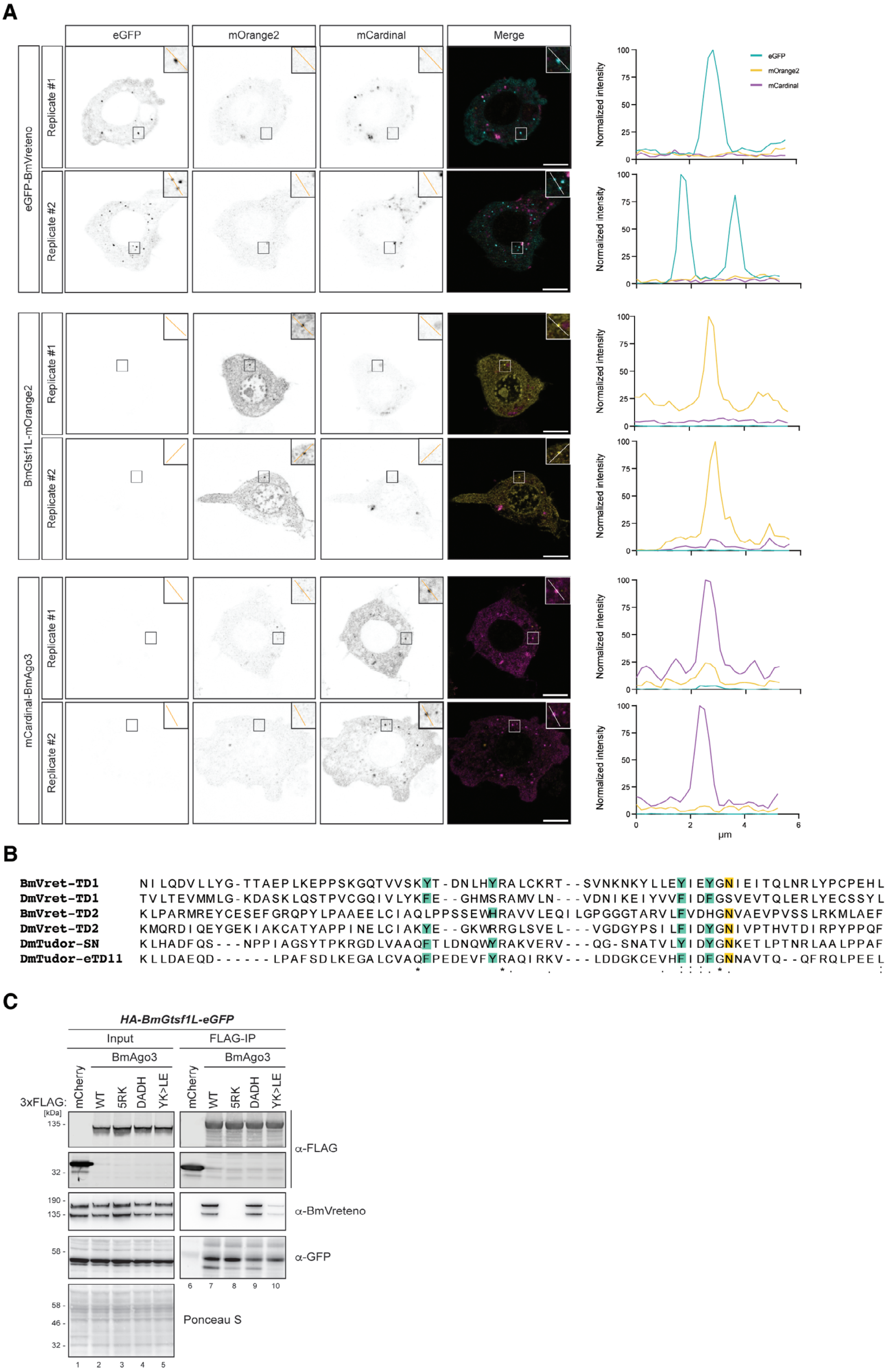

A
eGFP
mOrange2
mCardinal
Merge
eGFP-BmVreteno
Replicate #1
Replicate #2
BmGtsf1L-mOrange2
mCardinal-BmAgo3
Normalized intensity
eGFP
mOrange2
mCardinal
0
25
50
75
100
2
4
6
µm
B
BmVret-TD1 NILQDVLLYG-TTAEPLKEPPSKGQTVVSKYT-DNLHYRALCKRT--SVNKNKYLLEYIEYGNIEITQLNRLYPCPEHL
DmVret-TD1 TVLTEVMMLG-KDASKLQSTPVCGQIVLYKFE--GHMSRAMVLN---VDNIKEIYVVFIDFGSVEVTQLERLYECSSYL
BmVret-TD2 KLPARMREYCESEFGRQPYLPAAEELCIAQLPPSSEWHRAVVLEQILGPGGGTARVLFVDHGNVAEVPVSSLRKMLAEF
DmVret-TD2 KMQRDIQEYGEKIAKCATYAPPINELCIAKYE--GKWRRGLSVEL---VGDGYPSILFIDYGNIVPTHVTDIRPYPPQF
DmTudor-SN KLHADFQS---NPPIAGSYTPKRGDLVAAQFTLDNQWYRAKVERV---QG-SNATVLYIDYGNKETLPTNRLAALPPAF
DmTudor-eTD11 KLLDAEQD------LPAFSDLKEGALCVAQFPEDEVFYRAQIRKV---LDDGKCEVHFIDFGNNAVTQ--QFRQLPEEL
C
HA-BmGtsf1L-eGFP
Input
FLAG-IP
BmAgo3
3xFLAG:
mCherry
WT
5RK
DADH
YK>LE
[kDa]
135 -
32 -
190 -
135 -
58 -
58 -
46 -
32 -
α-FLAG
α-BmVreteno
α-GFP
Ponceau S
1
2
3
4
5
6
7
8
9
10

Figure EV2.

Figure EV3. The BmGtsf1L C-terminus establishes an interaction with BmVreteno.

A Outlined strategy for the purification of recombinant GST-tagged BmVreteno-S, showing the profiles from the size-exclusion column (left) and the peak fractions that were analyzed by SDS-PAGE followed by Coomassie staining (right).

B Similar to panel (A), but for GST-BmVreteno-L.

C Size-exclusion profiles of BmVreteno-S (left) and BmVreteno-L (right) to compare the shift in molecular weight before (green line) and after (blue line) 3C-mediated cleavage of the GST tag.

D Pfam/SMART-based domain organization of the BmVreteno-Long isoform, showing the RNA-recognition motif (RRM), Myeloid translocation protein 8, Nervy and DEAF-1 (MYND) domain, and two C-terminal Tudor domains (TD). In addition to the full-length (FL) BmVreteno, the C-and N-terminal truncation variants are depicted and are used in panel (E) to analyze the interaction between BmGtsf1L and BmVreteno variants.

E Co-transfection of BmGtsf1L-eGFP with HA-BmVreteno truncation variants in which different domains were omitted. The transfection of HA-LacZ served as a control. BmGtsf1L was retrieved by GFP-IP, and input and elution fractions were analyzed by SDS-PAGE, followed by Western blot using the indicated antibodies. Anti-actin immunodetection served as a loading control.

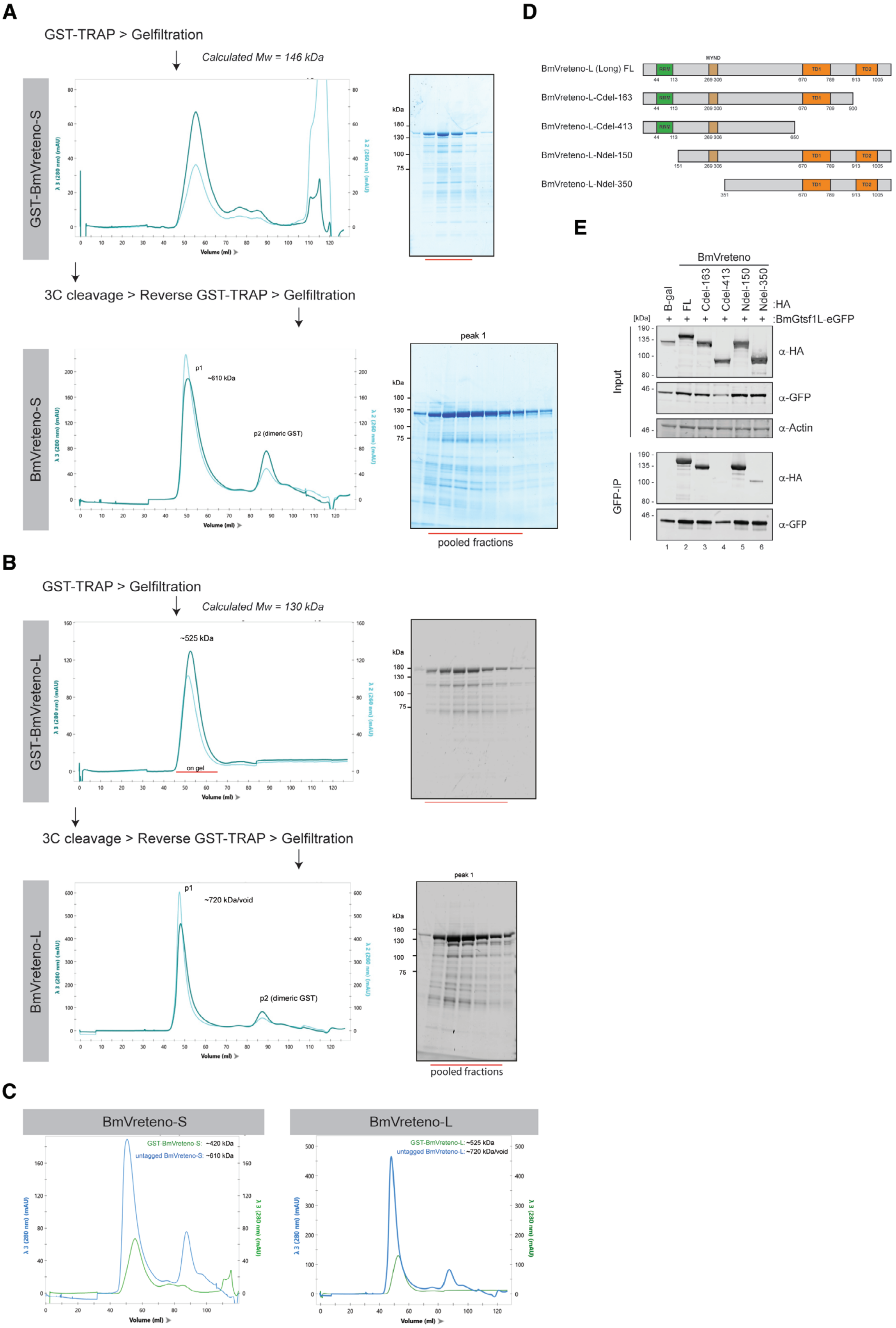

A
GST-TRAP > Gelfiltration
Calculated Mw = 146 kDa
GST-BmVreteno-S
λ 3 (280 nm) (mAU)
λ 2 (260 nm) (mAU)
Volume (ml)
kDa
180
130
100
75
3C cleavage > Reverse GST-TRAP > Gelfiltration
BmVreteno-S
p1
~610 kDa
p2 (dimeric GST)
peak 1
pooled fractions
B
GST-TRAP > Gelfiltration
Calculated Mw = 130 kDa
GST-BmVreteno-L
~525 kDa
on gel
3C cleavage > Reverse GST-TRAP > Gelfiltration
BmVreteno-L
p1
~720 kDa/void
p2 (dimeric GST)
peak 1
pooled fractions
C
BmVreteno-S
GST-BmVreteno-S: ~420 kDa
untagged BmVreteno-S: ~610 kDa
BmVreteno-L
GST-BmVreteno-L: ~525 kDa
untagged BmVreteno-L: ~720 kDa/void
Volume (ml)
D
BmVreteno-L (Long) FL
BmVreteno-L-Cdel-163
BmVreteno-L-Cdel-413
BmVreteno-L-Ndel-150
BmVreteno-L-Ndel-350
RRM
MYND
TD1
TD2
44 113 269 306 670 789 913 1005
900
650
151
351
E
BmVreteno
B-gal
FL
Cdel-163
Cdel-413
Ndel-150
Ndel-350
:HA
:BmGtsf1L-eGFP
[kDa]
Input
GFP-IP
α-HA
α-GFP
α-Actin
190
135
100
80
46
1 2 3 4 5 6

Figure EV3.

Figure EV4. The hydrophobic binding pocket is unique to BmVreteno AF-eTD1 and facilitates BmGtsf1L binding.

A Superimposition of the AlphaFold structural models of all three BmVreteno AF-eTD domains. The central inset (closed circle) shows the side view of the aromatic cage, which is only present in AF eTD1 and indicated with a dashed circle. The top inset (closed circle) shows a top view of the novel interface, which is unique to AF eTD1. The hydrophobic pocket is indicated with a dashed circle, and the enlarged view additionally shows the docked C-terminal motif of BmGtsf1L.

B Snapshot on the novel hydrophobic binding pocket of BmVreteno AF-eTD1 (blue) and contacts between the residues R742, S744, K749, and I762 (shown as sticks) with BmGtsf1L C-terminal 10-AA residues (shown as yellow sticks). This snapshot displays the BmVreteno S744 residue that forms a hydrogen bond with the backbone carbonyl of BmGtsf1L I98.

C Plots showing the distances between the atoms forming the four most important inter-chain hydrogen bonds of the side chain of BmVreteno S744 in two out of the ten 1 μs simulation runs. The run presented in the upper panel displays the hydrogen bond between the side chain of S744 and the backbone carbonyl of BmGtsf1L I98. The simulation presented in the bottom panel reveals that S744 is engaged in different interactions with BmGtsf1L residues I98 and D100. Overall, the simulations reveal that some of the hydrogen bonds are transiently formed and broken.

D Single-plane confocal micrographs of BmN4 cells co-transfected with different eGFP-BmVreteno constructs (upper panel) and BmGtsf1L-mCherry (middle panel). Yellow triangles indicate a formed granule. Scale bars: 4 μm.

E Transfection of BmN4 cells with BmGtsf1L-eGFP together with HA-BmVreteno-FL. The transfection of HA-LacZ served as a control. A GFP (BmGtsf1L) immunoprecipitation was performed on BmN4 lysates, and input as well as elution samples were resolved by SDS-PAGE. Proteins were detected by Western blot using the indicated antibodies, and anti-actin probing as well as Ponceau S staining served as a loading control.

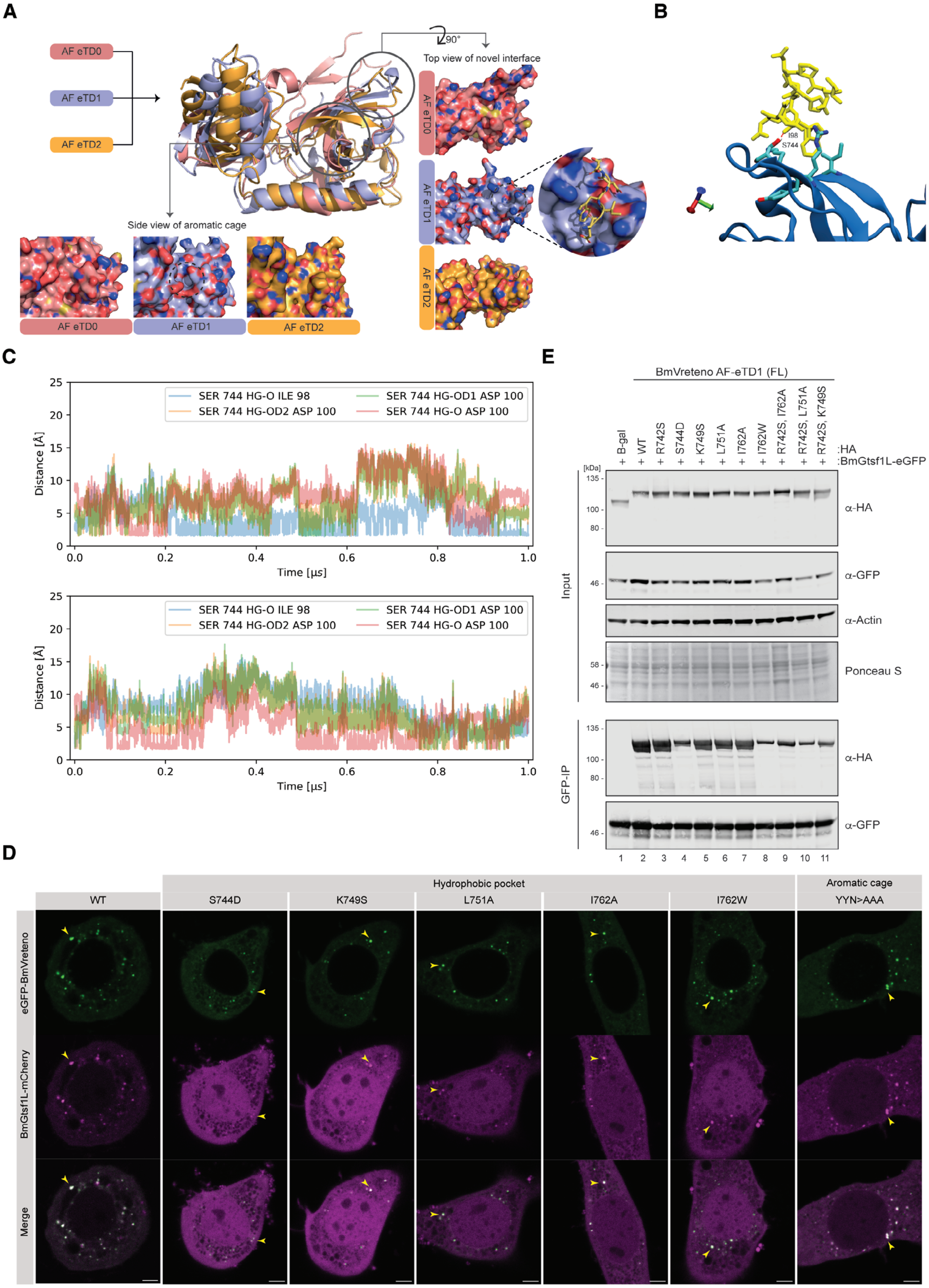

A
AF eTD0
AF eTD1
AF eTD2
90°
Top view of novel interface
Side view of aromatic cage
AF eTD0
AF eTD1
AF eTD2
B
I98
S744
C
Distance [Å]
Time [μs]
SER 744 HG-O ILE 98
SER 744 HG-OD2 ASP 100
SER 744 HG-OD1 ASP 100
SER 744 HG-O ASP 100
0.0
0.2
0.4
0.6
0.8
1.0
0
5
10
15
20
25
D
WT
Hydrophobic pocket
S744D
K749S
L751A
I762A
I762W
Aromatic cage
YYN>AAA
eGFP-BmVreteno
BmGtsf1L-mCherry
Merge
E
BmVreteno AF-eTD1 (FL)
B-gal
WT
R742S
S744D
K749S
L751A
I762A
I762W
R742S, I762A
R742S, L751A
R742S, K749S
:HA
:BmGtsf1L-eGFP
[kDa]
135
100
80
46
58
Input
GFP-IP
α-HA
α-GFP
α-Actin
Ponceau S
1
2
3
4
5
6
7
8
9
10
11

Figure EV4.

Chapter 5

Conclusions and future perspectives

5.1 Deciphering protein interaction interfaces: detecting the known and charting the unknown

As many proteins interact with others to mediate their functions, the interactions between proteins have been mapped extensively as a means to study their biological functions. Nonetheless, without further mechanistic information on the detected PPIs, it remains difficult to further probe their molecular functions. While significant effort has been dedicated to developing efficient methods to detect functional motifs within protein sequences, these methods only offer partial information on the molecular function of a PPI because potential binding domains in the interaction partner are not considered. For an interaction-specific detection of motifs and their potential binding domains, this thesis delved into the development of a DMI predictor that detects and scores potential DMIs using the sequences of PPIs. The DMI predictor achieved similar performance as iELM, another DMI detection program that is currently obsolete. Notably, the similar performance was achieved with a considerably more simplified workflow, therefore posing a lower risk of overfitting compared to iELM.

Owing to the folded nature of domains and the stable interactions that DDIs mediate, structural information on DDIs is significantly more abundant compared to DMIs. 3did is one of the databases that makes use of this type of information to extensively catalog DDIs. Article I investigated the quality of DDIs cataloged in 3did by manually curating a randomly sampled subset of DDIs from 3did. The manual curation process, albeit labor-intensive, provided us with important insights regarding features that can aid in scoring predicted DDIs for their abilities to mediate PPIs. For instance, a comparison between the AF-MM models of approved and non-approved DDIs revealed that the AF-MM models of approved DDIs had a stronger tendency than non-approved DDIs to resemble their respective solved structures. While the incorporation of this feature to score DDIs is not feasible due to the time-consuming modelling process of

AF-MM, the agreement between AF-MM and our evaluation was striking and further substantiated our evaluation on the DDIs. On the other hand, perhaps not so surprisingly, the publications accompanying the structures used by 3did to annotate DDIs provided the most informative evidence to evaluate DDIs. While our attempt on text mining yielded poor results, more sophisticated natural language processing (NLP) methods are now available for further exploration. For example, zero-shot learning approaches can be applied using large language models like Generative Pretrained Transformer (GPT) from OpenAI or Bidirectional Encoder Representations from Transformers (BERT) from Google to classify publications based on whether they discuss DDIs or not ("What's the next Word in Large Language Models?", 2023). In our experience, the publications for a given DDI type either discuss the DDI type or none at all, so classifying whether a publication talks about DDIs or not would be sufficient to aid in scoring DDI types. Alternatively, the STRING database has an elaborate text-mining pipeline designed to detect protein-protein associations by scanning for the co-occurrence of protein names within scientific publications (Franceschini et al., 2013; Szklarczyk et al., 2019). This text-mining pipeline can also be explored for its potential application in detecting the co-occurrence of domain names. Regardless, this set of manually curated DDIs will definitely serve as a useful dataset for the development and benchmarking of subsequent tools.

By applying the developed DMI predictor and a filtered list of DDI types to HuRI, approximately 8% of the PPIs could be mapped with at least one interface of high-confidence. Although Y2H-based screens primarily detect regulatory and signalling-related PPIs that are more likely mediated by DMIs, the overwhelmingly large number of PPIs that are not mapped with any known interfaces points to the fact that many interface types remain uncharted, especially those involving motifs (Lambourne et al., 2022; Rolland et al., 2014; Tompa et al., 2014). To detect these interfaces that are not previously known, Article II and III investigated the use of AF-MM to model the interaction interfaces between known interacting proteins. With the finding that sequence length is detrimental to AF-MM prediction performance, we devised the so-called fragmentation approach to boost AF-MM sensitivity. Albeit fruitful in some cases, using the fragmentation approach to scan for potential interaction interfaces without prior binding region information is too slow to be scaled up to higher throughput. One way to circumvent this limitation is to incorporate tools that are capable of identifying potential protein-binding regions into the fragmentation pipeline. An example of such tools is PeSTo that uses a deep learning model to predict protein-binding surfaces on monomeric protein structures (Krapp et al., 2023). Alternatively, the DMI detection pipeline developed in this thesis can also be modified to scan for potential protein-binding regions. The degree of conservation of a protein region as well as its propensity to be disordered and to undergo secondary structure transition are all useful indicators of the region's capability to bind proteins. All these features can be computed for all the amino acids in a given protein sequence, and they can be subsequently used to score and identify putative protein-binding regions. To detect the interfaces that are potentially formed by these putative protein-binding regions, domains that are significantly enriched in the interaction partners can be paired with the putative protein-binding regions for AF-MM modelling. While incorporating these steps can narrow down likely protein-binding regions in a protein sequence to make the fragmentation approach more efficient, another shortcoming of the approach is the lack of specificity. In our experience, the primary source of false positives arises from non-protein binding surfaces being modelled to bind to small peptides. Since PeSTo also scores protein surfaces for their propensity to bind to different types of organic or inorganic molecules, including amino acids, nucleic acids, and metal ions, integrating these scores into the model evaluation step can help to filter out false positives. All in all, using AF-MM to discover novel interfaces holds great potentials as it bypasses the need for a reference list of interface types for interface searching. Nevertheless, for its application to be scaled up to higher throughput, more work is warranted.

5.2 For a more user-friendly bioinformatic tool

Article II extensively benchmarked AF-MM across various aspects of structural modelling. Among the benchmarked aspects, the length of input sequence was emphasized as it carried important implications on the application of AF-MM on characterizing PPI interfaces. One key result from this benchmark was that longer input sequence led to worse prediction performance of AF-MM. In retrospect, given that many protein complexes have been solved using minimal interacting regions, such as a domain complexed with a peptide or another domain, it is reasonable that AF-MM struggles with interface prediction when sequences beyond minimal interacting regions are used, especially the full-length sequences of multi-domain proteins. Accordingly, the publication accompanying AF-MM also tested the accuracy of AF-MM using only structures solved with minimal interacting regions. Nevertheless, since this information is not explicitly stated in the publication, its clarity to non-expert users is not always guaranteed. Similarly, in spite of AF-MM outputting various metrics to evaluate the models that it generates, it is not always clear what are the thresholds that need to applied on the metrics to deem a model to be of high-confidence. To this end, the benchmark also enabled the establishment of useful thresholds on different metrics to discriminate between good and bad models. These thresholds are particularly valuable for non-expert users, as they facilitate a standard and unambiguous assessment of the models.

An important aspect of assessing the performance of machine learning or AI models is to use unseen data for model evaluation. While this was not addressed in our investigation, an independent study conducted a similar benchmarking on AF-MM using a smaller dataset of domain-motif complex structures that are previously unseen by AF-MM during its training. Reassuringly, their findings were similar to those in our investigation (Bret et al., 2024). Without a doubt, with their proven capability to transform a massive amount of data into useful insights and testable hypotheses, machine learning and AI models have been and will continue to be incorporated into the workflow of bioinformatic tools whose purposes extend beyond PPI interface prediction (Baek et al., 2024; Cheng et al., 2023; H. Zhao et al., 2024). For non-expert users to derive the most benefit from these tools, it is imperative to consider potential applications of the tools and benchmark them with use cases in mind.

5.3 Validating protein interaction interface using interaction assay

The BRET assay has been used extensively in Article II to validate the interaction interfaces predicted by AF-MM. To use BRET for the validation of predicted interfaces, the proteins were first tested for their interactions in a mammalian cell line. Once the interactions between the proteins were detected, we proceeded to design mutations based on AF-MM models to perturb the predicted interfaces between the proteins. The mutations were then introduced into the proteins, and their interactions were tested again to probe the effect of the mutations on the binding of the proteins. A reduction in binding strength or the absence of binding between the mutated protein and its wild-type partner was thus indicative of the predicted interface being responsible, or at least partially responsible, for the detected interaction.

Though designed with the intention of perturbing predicted interfaces, introduced mutations can alter the localization of the proteins or affect the integrity and subsequently the expression of the proteins by misfolding the proteins. These possibilities can also lead to a reduction or loss in binding, thereby confounding the results of the study. While these possibilities are not fully addressed in Article II, future work that incorporates microscopy imaging into the interface validation workflow can help to elucidate the localization of mutant proteins. While the expression levels of mutant proteins were always monitored in the study, it is important to note that misfolded proteins are not always degraded and may remain expressed. To probe the integrity of the mutated proteins, an elegant approach was demonstrated in the validation of the predicted interface between PEX3 and PEX16 in Article II. In brief, PEX3 is known to interact with PEX19 through a surface that is distal from the predicted interface between PEX3 and PEX16. This information was leveraged to design mutations on the surface of PEX3 that are aimed at perturbing its interaction with PEX16 and not PEX19. Since this surface is not involved in the interaction between PEX3 and PEX19, mutations on this surface should not affect the interaction between PEX3 and PEX19, unless the mutations lead to the misfolding of PEX3. Indeed, when these mutations on PEX3 were tested against PEX16 and PEX19, it was found that they perturbed the interaction with PEX16 and had no effect on the interaction with PEX19, indicating that the integrity of PEX3 was maintained despite the presence of the mutations. Alternatively, the interactions between mutated proteins with protein chaperones like HSP70 and HSP90 can also be checked as an indicator of protein misfolding induced by mutations. Several studies have applied this approach to investigate the impact of mutations on the conformational stability of proteins (Gracia et al., 2023; Sahni et al., 2015).

While the BRET assay is, without a doubt, a potent quantitative interaction assay, its inherent sensitivity to the orientation and distance of the tags poses significant technical hurdles for scaling up the method. Moreover, the location of the tags, either at the N terminus or at the C terminus, can also block the interaction between the proteins due to steric hindrance. The detection of interactions between transmembrane proteins is especially sensitive to these factors. To minimize these limitations, we tested the PPIs using different configurations by fusing the tags at different termini and selected the configuration with the highest BRET measurement for subsequent interface validations. Besides, estimating the binding strength of a PPI using BRET also requires saturating the donors with an increasing amount of acceptors. As the number of BRET experiments scales with the number of mutations designed for interface validation, such saturation experiments can drastically inflate the number of BRET experiments to be performed. Taken together, the use of BRET assay to characterize protein interaction interfaces can be efficient as it allows the simultaneous monitoring of protein expression levels. Nonetheless, for it to be scaled up to a higher throughput, the use of robotics may be necessary to handle the labor-intensive cloning procedures and saturation experiments.

5.4 General outlook

This thesis has delved into different ways of predicting known and novel PPI interfaces between interacting proteins. In doing so, we have gained profound insights into the intricate interplay between different functional modules in proteins that facilitate their interactions. Given the useful mechanistic insights that prediction tools like AF-MM can provide, I anticipate greater inclusion of these tools in experimental workflows, where they can generate hypotheses to guide experiments.

Another avenue worth exploring is the use of predicted protein structures to chart the structural space of PPI interfaces. Previous studies have suggested that PPI interfaces are degenerate and have put forth an estimate of 10,000 unique PPI interfaces based on available PDB structures (Aloy & Russell, 2004; Gao & Skolnick, 2010). It would be interesting to revisit this estimate, given that highly accurate structural models of proteins are now available thanks to AF.

Binary interaction assays like BRET have proven to be useful in validating PPI interfaces. Nonetheless, many proteins are involved in multi-subunit protein complexes, and the disruption of one interaction can have an effect on other interactions in the complex. To obtain a holistic understanding of PPIs and their roles in biological systems, functional studies focusing on certain protein complexes or cellular pathways will be instrumental.

Bibliography

Akdel, M., Pires, D. E. V., Pardo, E. P., Jänes, J., Zalevsky, A. O., Mészáros, B., Bryant, P., Good, L. L., Laskowski, R. A., Pozzati, G., Shenoy, A., Zhu, W., Kundrotas, P., Serra, V. R., Rodrigues, C. H. M., Dunham, A. S., Burke, D., Borkakoti, N., Velankar, S., . . . Beltrao, P. (2022). A structural biology community assessment of AlphaFold2 applications. *Nature Structural & Molecular Biology*, *29*(11), 1056–1067. https://doi.org/10.1038/s41594-022-00849-w

Alexander, J., Lim, D., Joughin, B. A., Hegemann, B., Hutchins, J. R. A., Ehrenberger, T., Ivins, F., Sessa, F., Hudecz, O., Nigg, E. A., Fry, A. M., Musacchio, A., Stukenberg, P. T., Mechtler, K., Peters, J.-M., Smerdon, S. J., & Yaffe, M. B. (2011). Spatial Exclusivity Combined with Positive and Negative Selection of Phosphorylation Motifs Is the Basis for Context-Dependent Mitotic Signaling. *Science Signaling*, *4*(179), ra42–ra42. https://doi.org/10.1126/scisignal.2001796

Aloy, P., Ceulemans, H., Stark, A., & Russell, R. B. (2003). The relationship between sequence and interaction divergence in proteins. *Journal of Molecular Biology*, *332*(5), 989–998. https://doi.org/10.1016/j.jmb.2003.07.006

Aloy, P., & Russell, R. B. (2003). InterPreTS: Protein Interaction Prediction through Tertiary Structure. *Bioinformatics*, *19*(1), 161–162. https://doi.org/10.1093/bioinformatics/19.1.161

Aloy, P., & Russell, R. B. (2004). Ten thousand interactions for the molecular biologist. *Nature Biotechnology*, *22*(10), 1317–1321. https://doi.org/10.1038/nbt1018

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., & Murzin, A. G. (2008). Data growth and its impact on the SCOP database: New developments. *Nucleic Acids Research*, *36*, D419–425. https://doi.org/10.1093/nar/gkm993

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., & Murzin, A. G. (2014). SCOP2 prototype: A new approach to protein structure mining. *Nucleic Acids Research*, *42*(D1), D310–D314. https://doi.org/10.1093/nar/gkt1242

Andreeva, A., Kulesha, E., Gough, J., & Murzin, A. G. (2020). The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Research*, *48*(D1), D376–D382. https://doi.org/10.1093/nar/gkz1064

Araya, C. L., & Fowler, D. M. (2011). Deep mutational scanning: Assessing protein function on a massive scale. *Trends in Biotechnology*, *29*(9), 435–442. https://doi.org/10.1016/j.tibtech.2011.04.003

Babu, M. M. (2016). The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochemical Society Transactions*, *44*(5), 1185–1200. https://doi.org/10.1042/BST20160172

Baek, M., McHugh, R., Anishchenko, I., Jiang, H., Baker, D., & DiMaio, F. (2024). Accurate prediction of protein–nucleic acid complexes using RoseTTAFoldNA. *Nature Methods*, *21*(1), 117–121. https://doi.org/10.1038/s41592-023-02086-5

Bagowski, C. P., Bruins, W., & te Velthuis, A. J. (2010). The Nature of Protein Domain Evolution: Shaping the Interaction Network. *Current Genomics*, *11*(5), 368–376. https://doi.org/10.2174/138920210791616725

Basile, W., Salvatore, M., Bassot, C., & Elofsson, A. (2019). Why do eukaryotic proteins contain more intrinsically disordered regions? *PLOS Computational Biology*, *15*(7), e1007186. https://doi.org/10.1371/journal.pcbi.1007186

Benz, C., Ali, M., Krystkowiak, I., Simonetti, L., Sayadi, A., Mihalic, F., Kliche, J., Andersson, E., Jemth, P., Davey, N. E., & Ivarsson, Y. (2022). Proteome-scale mapping of binding sites in the unstructured regions of the human proteome. *Molecular Systems Biology*, *18*(1), e10584. https://doi.org/10.15252/msb.202110584

Borgia, A., Borgia, M. B., Bugge, K., Kissling, V. M., Heidarsson, P. O., Fernandes, C. B., Sottini, A., Soranno, A., Buholzer, K. J., Nettels, D., Kragelund, B. B., Best, R. B., & Schuler, B. (2018). Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, *555*(7694), 61–66. https://doi.org/10.1038/nature25762

Bret, H., Gao, J., Zea, D. J., Andreani, J., & Guerois, R. (2024). From interaction networks to interfaces, scanning intrinsically disordered regions using AlphaFold2. *Nature Communications*, *15*(1), 597. https://doi.org/10.1038/s41467-023-44288-7

Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J., & Dunker, A. K. (2002). Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of Molecular Evolution*, *55*(1), 104–110. https://doi.org/10.1007/s00239-001-2309-6

Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. *Nature Communications*, *13*(1), 1265. https://doi.org/10.1038/s41467-022-28865-w

Bystroff, C., & Krogh, A. (2008). Hidden Markov Models for Prediction of Protein Features. In M. J. Zaki & C. Bystroff (Eds.), *Protein Structure Prediction* (pp. 173–198). Humana Press. https://doi.org/10.1007/978-1-59745-574-9_7

Canman, C. E., Lim, D. S., Cimprich, K. A., Taya, Y., Tamai, K., Sakaguchi, K., Appella, E., Kastan, M. B., & Siliciano, J. D. (1998). Activation of the ATM kinase by ionizing radiation and phosphorylation of p53. *Science (New York, N.Y.)*, *281*(5383), 1677–1679. https://doi.org/10.1126/science.281.5383.1677

Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G., & Raschellà, G. (2017). Zinc-finger proteins in health and disease. *Cell Death Discovery*, *3*(1), 1–12. https://doi.org/10.1038/cddiscovery.2017.71

Chen, T. S., Petrey, D., Garzon, J. I., & Honig, B. (2015). Predicting Peptide-Mediated Interactions on a Genome-Wide Scale. *PLOS Computational Biology*, *11*(5), e1004248. https://doi.org/10.1371/journal.pcbi.1004248

Chen, X.-W., & Liu, M. (2005). Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics*, *21*(24), 4394–4400. https://doi.org/10.1093/bioinformatics/bti721

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*, *381*(6664), eadg7492. https://doi.org/10.1126/science.adg7492

Chien, C. T., Bartel, P. L., Sternglanz, R., & Fields, S. (1991). The two-hybrid system: A method to identify and clone genes for proteins that interact with a protein of interest. *Proceedings of the National Academy of Sciences*, *88*(21), 9578–9582. https://doi.org/10.1073/pnas.88.21.9578

Chothia, C., & Janin, J. (1975). Principles of protein–protein recognition. *Nature*, *256*(5520), 705–708. https://doi.org/10.1038/256705a0

Christian, F., Smith, E. L., & Carmody, R. J. (2016). The Regulation of NF-ϰB Subunits by Phosphorylation. *Cells*, *5*(1), 12. https://doi.org/10.3390/cells5010012

Clabbers, M. T. B., Holmes, S., Muusse, T. W., Vajjhala, P. R., Thygesen, S. J., Malde, A. K., Hunter, D. J. B., Croll, T. I., Flueckiger, L., Nanson, J. D., Rahaman, M. H., Aquila, A., Hunter, M. S., Liang, M., Yoon, C. H., Zhao, J., Zatsepin, N. A., Abbey, B., Sierecki, E., … Ve, T. (2021). MyD88 TIR domain higher-order assembly interactions revealed by microcrystal electron diffraction and serial femtosecond crystallography. *Nature Communications*, *12*(1), 2578. https://doi.org/10.1038/s41467-021-22590-6

Cortese, M. S., Uversky, V. N., & Keith Dunker, A. (2008). Intrinsic disorder in scaffold proteins: Getting more from less. *Progress in Biophysics and Molecular Biology*, *98*(1), 85–106. https://doi.org/10.1016/j.pbiomolbio.2008.05.007

Coulon, V., Audet, M., Homburger, V., Bockaert, J., Fagni, L., Bouvier, M., & Perroy, J. (2008). Subcellular Imaging of Dynamic Protein Interactions by Bioluminescence Resonance Energy Transfer. *Biophysical Journal*, *94*(3), 1001. https://doi.org/10.1529/biophysj.107.117275

Davey, N. E., Cowan, J. L., Shields, D. C., Gibson, T. J., Coldwell, M. J., & Edwards, R. J. (2012). SLiMPrints: Conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Research*, *40*(21), 10628–10641. https://doi.org/10.1093/nar/gks854

Davey, N. E., Cyert, M. S., & Moses, A. M. (2015). Short linear motifs – ex nihilo evolution of protein regulation. *Cell Communication and Signaling*, *13*(1), 43. https://doi.org/10.1186/s12964-015-0120-z

Del Pozo, M. A., Kiosses, W. B., Alderson, N. B., Meller, N., Hahn, K. M., & Schwartz, M. A. (2002). Integrins regulate GTP-Rac localized effector interactions through dissociation of Rho-GDI. *Nature Cell Biology*, *4*(3), 232–239. https://doi.org/10.1038/ncb759

Deng, M., Mehta, S., Sun, F., & Chen, T. (2002). Inferring Domain–Domain Interactions From Protein–Protein Interactions. *Genome Research*, *12*(10), 1540–1548. https://doi.org/10.1101/gr.153002

Di Fiore, B., Davey, N. E., Hagting, A., Izawa, D., Mansfeld, J., Gibson, T. J., & Pines, J. (2015). The ABBA Motif Binds APC/C Activators and Is Shared by APC/C Substrates and Regulators. *Developmental Cell*, *32*(3), 358–372. https://doi.org/10.1016/j.devcel.2015.01.003

Dias, R. V. R., Pedro, R. P., Sanches, M. N., Moreira, G. C., Leite, V. B. P., Caruso, I. P., de Melo, F. A., & de Oliveira, L. C. (2023). Unveiling Metastable Ensembles of GRB2 and the Relevance of Interdomain Communication during Folding. *Journal of Chemical Information and Modeling*, *63*(20), 6344–6353. https://doi.org/10.1021/acs.jcim.3c00955

Disfani, F. M., Hsu, W.-L., Mizianty, M. J., Oldfield, C. J., Xue, B., Dunker, A. K., Uversky, V. N., & Kurgan, L. (2012). MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics*, *28*(12), i75–i83. https://doi.org/10.1093/bioinformatics/bts209

Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics (Oxford, England)*, *21*(16), 3433–3434. https://doi.org/10.1093/bioinformatics/bti541

Dosztányi, Z., Mészáros, B., & Simon, I. (2009). ANCHOR: Web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, *25*(20), 2745–2746. https://doi.org/10.1093/bioinformatics/btp518

Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., & Obradovic, Z. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, *19*(1), 26–59. https://doi.org/10.1016/S1093-3263(00)00138-8

Dyla, M., & Kjaergaard, M. (2020). Intrinsically disordered linkers control tethered kinases via effective concentration. *Proceedings of the National Academy of Sciences*, *117*(35), 21413–21419. https://doi.org/10.1073/pnas.2006382117

Ekman, D., Björklund, Å. K., Frey-Skött, J., & Elofsson, A. (2005). Multi-domain Proteins in the Three Kingdoms of Life: Orphan Domains and Other Unassigned Regions. *Journal of Molecular Biology*, *348*(1), 231–243. https://doi.org/10.1016/j.jmb.2005.02.007

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., . . . Hassabis, D. (2022, March 10). *Protein complex prediction with AlphaFold-Multimer*. https://doi.org/10.1101/2021.10.04.463034

Eyckerman, S., Verhee, A., der Heyden, J. V., Lemmens, I., Ostade, X. V., Vandekerckhove, J., & Tavernier, J. (2001). Design and application of a cytokine-receptor-based interaction trap. *Nature Cell Biology*, *3*(12), 1114–1119. https://doi.org/10.1038/ncb1201-1114

Fields, S., & Song, O.-k. (1989). A novel genetic system to detect protein–protein interactions. *Nature*, *340*(6230), 245–246. https://doi.org/10.1038/340245a0

Finn, R. D., Miller, B. L., Clements, J., & Bateman, A. (2014). iPfam: A database of protein family and domain interactions found in the Protein Data Bank. *Nucleic Acids Research*, *42*(D1), D364–D373. https://doi.org/10.1093/nar/gkt1210

Flaugh, S. L., Kosinski-Collins, M. S., & King, J. (2005). Contributions of hydrophobic domain interface interactions to the folding and stability of human γD-crystallin. *Protein Science*, *14*(3), 569–581. https://doi.org/10.1110/ps.041111405

Flock, T., Weatheritt, R. J., Latysheva, N. S., & Babu, M. M. (2014). Controlling entropy to tune the functions of intrinsically disordered regions. *Current Opinion in Structural Biology*, *26*, 62–72. https://doi.org/10.1016/j.sbi.2014.05.007

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., & Jensen, L. J. (2013). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, *41*(D1), D808–D815. https://doi.org/10.1093/nar/gks1094

Freund, C., Kühne, R., Yang, H., Park, S., Reinherz, E. L., & Wagner, G. (2002). Dynamic interaction of CD2 with the GYF and the SH3 domain of compartmentalized effector molecules. *The EMBO Journal*, *21*(22), 5985–5995. https://doi.org/10.1093/emboj/cdf602

Fuller, J. C., Burgoyne, N. J., & Jackson, R. M. (2009). Predicting druggable binding sites at the protein–protein interface. *Drug Discovery Today*, *14*(3), 155–161. https://doi.org/10.1016/j.drudis.2008.10.009

Gao, M., & Skolnick, J. (2010). Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences*, *107*(52), 22517–22522. https://doi.org/10.1073/pnas.1012820107

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., . . . Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, *415*(6868), 141–147. https://doi.org/10.1038/415141a

Gödde, N. J., D'Abaco, G. M., Paradiso, L., & Novak, U. (2006). Efficient ADAM22 surface expression is mediated by phosphorylation-dependent interaction with 14-3-3 protein family members. *Journal of Cell Science*, *119*, 3296–3305. https://doi.org/10.1242/jcs.03065

Gostissa, M., Hengstermann, A., Fogal, V., Sandy, P., Schwarz, S. E., Scheffner, M., & Del Sal, G. (1999). Activation of p53 by conjugation to the ubiquitin-like protein SUMO-1. *The EMBO journal*, *18*(22), 6462–6471. https://doi.org/10.1093/emboj/18.22.6462

Gough, J., Karplus, K., Hughey, R., & Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure11Edited by G. Von Heijne. *Journal of Molecular Biology*, *313*(4), 903–919. https://doi.org/10.1006/jmbi.2001.5080

Gracia, B., Montes, P., Gutierrez, A. M., Arun, B., & Karras, G. I. (2023, September 15). *Protein-Folding Chaperones Predict Structure-Function Relationships and Cancer Risk in BRCA1 Mutation Carriers*. https://doi.org/10.1101/2023.09.14.557795

Gratten, J., & Visscher, P. M. (2016). Genetic pleiotropy in complex traits and diseases: Implications for genomic medicine. *Genome Medicine*, *8*(1), 78. https://doi.org/10.1186/s13073-016-0332-x

Green, D. W., Ingram, V. M., & Perutz, M. F. (1954). The Structure of Haemoglobin. IV. Sign Determination by the Isomorphous Replacement Method. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *225*(1162), 287–307. Retrieved October 13, 2023, from https://www.jstor.org/stable/99481

Gunasekaran, K., Tsai, C.-J., Kumar, S., Zanuy, D., & Nussinov, R. (2003). Extended disordered proteins: Targeting function with less scaffold. *Trends in Biochemical Sciences*, *28*(2), 81–85. https://doi.org/10.1016/S0968-0004(03)00003-3

Hadarovich, A., Chakravarty, D., Tuzikov, A. V., Ben-Tal, N., Kundrotas, P. J., & Vakser, I. A. (2021). Structural motifs in protein cores and at protein–protein interfaces are different. *Protein Science : A Publication of the Protein Society*, *30*(2), 381–390. https://doi.org/10.1002/pro.3996

Hames, R. S., Wattam, S. L., Yamano, H., Bacchieri, R., & Fry, A. M. (2001). APC/C-mediated destruction of the centrosomal kinase Nek2A occurs in early mitosis and depends upon a cyclin A-type D-box. *The EMBO journal*, *20*(24), 7117–7127. https://doi.org/10.1093/emboj/20.24.7117

Han, J.-H., Batey, S., Nickson, A. A., Teichmann, S. A., & Clarke, J. (2007). The folding and evolution of multidomain proteins. *Nature Reviews Molecular Cell Biology*, *8*(4), 319–330. https://doi.org/10.1038/nrm2144

He, J., Chao, W. C. H., Zhang, Z., Yang, J., Cronin, N., & Barford, D. (2013). Insights into Degron Recognition by APC/C Coactivators from the Structure of an Acm1-Cdh1 Complex. *Molecular Cell*, *50*(5), 649–660. https://doi.org/10.1016/j.molcel.2013.04.024

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., . . . Tyers, M. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, *415*(6868), 180–183. https://doi.org/10.1038/415180a

Holehouse, A. S., & Kragelund, B. B. (2023). The molecular basis for cellular function of intrinsically disordered protein regions. *Nature Reviews Molecular Cell Biology*, 1–25. https://doi.org/10.1038/s41580-023-00673-0

Howard, C. J., Hanson-Smith, V., Kennedy, K. J., Miller, C. J., Lou, H. J., Johnson, A. D., Turk, B. E., & Holt, L. J. (2014). Ancestral resurrection reveals evolutionary mechanisms of kinase plasticity (J. Ferrell, Ed.). *eLife*, *3*, e04126. https://doi.org/10.7554/eLife.04126

Huang, H., Li, L., Wu, C., Schibli, D., Colwill, K., Ma, S., Li, C., Roy, P., Ho, K., Songyang, Z., Pawson, T., Gao, Y., & Li, S. S.-C. (2008). Defining the Specificity Space of the Human Src Homology 2 Domain *. *Molecular & Cellular Proteomics*, *7*(4), 768–784. https://doi.org/10.1074/mcp.M700312-MCP200

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Huttlin, E. L., Bruckner, R. J., Navarrete-Perea, J., Cannon, J. R., Baltier, K., Gebreab, F., Gygi, M. P., Thornock, A., Zarraga, G., Tam, S., Szpyt, J., Gassaway, B. M., Panov, A., Parzen, H., Fu, S., Golbazi, A., Maenpaa, E., Stricker, K., Thakurta, S. G., … Gygi, S. P. (2021). Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. *Cell*, *184*(11), 3022–3040.e28. https://doi.org/10.1016/j.cell.2021.04.011

Hwang, H., Petrey, D., & Honig, B. (2016). A hybrid method for protein–protein interface prediction. *Protein Science : A Publication of the Protein Society*, *25*(1), 159–165. https://doi.org/10.1002/pro.2744

Iakoucheva, L. M., Radivojac, P., Brown, C. J., O'Connor, T. R., Sikes, J. G., Obradovic, Z., & Dunker, A. K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research*, *32*(3), 1037–1049. https://doi.org/10.1093/nar/gkh253

Igraph – Network analysis software. (n.d.). Retrieved January 25, 2024, from https://igraph.org/

Iqbal, S., Brünger, T., Pérez-Palma, E., Macnee, M., Brunklaus, A., Daly, M. J., Campbell, A. J., Hoksza, D., May, P., & Lal, D. (2023). Delineation of functionally essential protein regions for 242 neurodevelopmental genes. *Brain*, *146*(2), 519–533. https://doi.org/10.1093/brain/awac381

Jones, S., Marin, A., & M.Thornton, J. (2000). Protein domain interfaces: Characterization and comparison with oligomeric protein interfaces. *Protein Engineering, Design and Selection*, *13*(2), 77–82. https://doi.org/10.1093/protein/13.2.77

Joo, H. K., Park, Y. W., Jang, Y. Y., & Lee, J. Y. (2018). Structural Analysis of Glutamine Synthetase from Helicobacter pylori. *Scientific Reports*, *8*, 11657. https://doi.org/10.1038/s41598-018-30191-5

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaczor, A. A., Bartuzi, D., Stępniewski, T. M., Matosiuk, D., & Selent, J. (2018). Protein–Protein Docking in Drug Design and Discovery. In M. Gore & U. B. Jagtap (Eds.), *Computational Drug Discovery and Design* (pp. 285–305). Springer. https://doi.org/10.1007/978-1-4939-7756-7_15

Kann, M. G., Jothi, R., Cherukuri, P. F., & Przytycka, T. M. (2007). Predicting protein domain interactions from coevolution of conserved regions. *Proteins: Structure, Function, and Bioinformatics*, *67*(4), 811–820. https://doi.org/10.1002/prot.21347

Karan, B., Mahapatra, S., Sahu, S. S., Pandey, D. M., & Chakravarty, S. (2023). Computational models for prediction of protein–protein interaction in rice and Magnaporthe grisea. *Frontiers in Plant Science*, *13*. Retrieved January 27, 2024, from https://www.frontiersin.org/articles/10.3389/fpls.2022.1046209

Kim, A. S., Kakalis, L. T., Abdul-Manan, N., Liu, G. A., & Rosen, M. K. (2000). Autoinhibition and activation mechanisms of the Wiskott-Aldrich syndrome protein. *Nature*, *404*(6774), 151–158. https://doi.org/10.1038/35004513

Kim, W. K., Henschel, A., Winter, C., & Schroeder, M. (2006). The Many Faces of Protein–Protein Interactions: A Compendium of Interface Geometry. *PLOS Computational Biology*, *2*(9), e124. https://doi.org/10.1371/journal.pcbi.0020124

Kozakov, D., Hall, D. R., Xia, B., Porter, K. A., Padhorny, D., Yueh, C., Beglov, D., & Vajda, S. (2017). The ClusPro web server for protein–protein docking. *Nature Protocols*, *12*(2), 255–278. https://doi.org/10.1038/nprot.2016.169

Krapp, L. F., Abriata, L. A., Cortés Rodriguez, F., & Dal Peraro, M. (2023). PeSTo: Parameter-free geometric deep learning for accurate prediction of protein binding interfaces. *Nature Communications*, *14*(1), 2175. https://doi.org/10.1038/s41467-023-37701-8

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A. P., Punna, T., Peregrín-Alvarez, J. M., Shales, M., Zhang, X., Davey, M., Robinson, M. D., Paccanaro, A., Bray, J. E., Sheung, A., . . . Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, *440*(7084), 637–643. https://doi.org/10.1038/nature04670

Krystkowiak, I., & Davey, N. E. (2017). SLiMSearch: A framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. *Nucleic Acids Research*, *45*, W464–W469. https://doi.org/10.1093/nar/gkx238

Krystkowiak, I., Manguy, J., & Davey, N. E. (2018). PSSMSearch: A server for modeling, visualization, proteome-wide discovery and annotation of protein motif specificity determinants. *Nucleic Acids Research*, *46*(W1), W235–W241. https://doi.org/10.1093/nar/gky426

Kumar, M., Michael, S., Alvarado-Valverde, J., Zeke, A., Lazar, T., Glavina, J., Nagy-Kanta, E., Donagh, J. M., Kalman, Z. E., Pascarelli, S., Palopoli, N., Dobson, L., Suarez, C. F., Van Roey, K., Krystkowiak, I., Griffin, J. E., Nagpal, A., Bhardwaj, R., Diella, F., . . . Gibson, T. J. (2024). ELM—the Eukaryotic Linear Motif resource—2024 update. *Nucleic Acids Research*, *52*(D1), D442–D455. https://doi.org/10.1093/nar/gkad1058

Lambourne, L., Yadav, A., Wang, Y., Desbuleux, A., Kim, D.-K., Cafarelli, T., Pons, C., Kovács, I. A., Jailkhani, N., Schlabach, S., Ridder, D. D., Luck, K., Bian, W., Shen, Y., Yang, Z., Mee, M. W., Helmy, M., Jacob, Y., Lemmens, I., ... Vidal, M. (2022, July 22). *Binary interactome models of inner- versus outer-complexome organisation*. https://doi.org/10.1101/2021.03.16.435663

Letunic, I., Khedkar, S., & Bork, P. (2021). SMART: Recent updates, new developments and status in 2020. *Nucleic Acids Research*, *49*(D1), D458–D460. https://doi.org/10.1093/nar/gkaa937

Levy, Y. (2017). Protein Assembly and Building Blocks: Beyond the Limits of the LEGO Brick Metaphor. *Biochemistry*, *56*(38), 5040–5048. https://doi.org/10.1021/acs.biochem.7b00666

Liang, S. H., & Clarke, M. F. (1999). A bipartite nuclear localization signal is required for p53 nuclear import regulated by a carboxyl-terminal domain. *The Journal of Biological Chemistry*, *274*(46), 32699–32703. https://doi.org/10.1074/jbc.274.46.32699

Lievens, S., Gerlo, S., Lemmens, I., De Clercq, D. J. H., Risseeuw, M. D. P., Vanderroost, N., De Smet, A.-S., Ruyssinck, E., Chevet, E., Van Calenbergh, S., & Tavernier, J. (2014). Kinase Substrate Sensor (KISS), a Mammalian In Situ Protein Interaction Sensor. *Molecular & Cellular Proteomics : MCP*, *13*(12), 3332–3342. https://doi.org/10.1074/mcp.M114.041087

Lievens, S., Peelman, F., De Bosscher, K., Lemmens, I., & Tavernier, J. (2011). MAPPIT: A protein interaction toolbox built on insights in cytokine receptor signaling. *Cytokine & Growth Factor Reviews*, *22*(5), 321–329. https://doi.org/10.1016/j.cytogfr.2011.11.001

Liu, M., Chen, X.-w., & Jothi, R. (2009). Knowledge-guided inference of domain–domain interactions from incomplete protein–protein interaction networks. *Bioinformatics*, *25*(19), 2492–2499. https://doi.org/10.1093/bioinformatics/btp480

Lu, D., Hsiao, J. Y., Davey, N. E., Van Voorhis, V. A., Foster, S. A., Tang, C., & Morgan, D. O. (2014). Multiple mechanisms determine the order of APC/C substrate degradation in mitosis. *Journal of Cell Biology*, *207*(1), 23–39. https://doi.org/10.1083/jcb.201402041

Lu, H., Zhou, Q., He, J., Jiang, Z., Peng, C., Tong, R., & Shi, J. (2020). Recent advances in the development of protein–protein interactions modulators: Mechanisms and clinical trials. *Signal Transduction and Targeted Therapy*, *5*(1), 1–23. https://doi.org/10.1038/s41392-020-00315-3

Luciani, M. G., Hutchins, J. R., Zheleva, D., & Hupp, T. R. (2000). The C-terminal regulatory domain of p53 contains a functional docking site for cyclin A. *Journal of Molecular Biology*, *300*(3), 503–518. https://doi.org/10.1006/jmbi.2000.3830

Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B. E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F. J., Charloteaux, B., Choi, D., Coté, A. G., Daley, M., Deimling, S., Desbuleux, A., Dricot, A., Gebbia, M., Hardy, M. F., Kishore, N., … Calderwood, M. A. (2020). A reference map of the human binary protein interactome. *Nature*, *580*(7803), 402–408. https://doi.org/10.1038/s41586-020-2188-x

Luther, C. H., Brandt, P., Vylkova, S., Dandekar, T., Müller, T., & Dittrich, M. (2023). Integrated analysis of SR-like protein kinases Sky1 and Sky2 links signaling networks with transcriptional regulation in Candida albicans. *Frontiers in Cellular and Infection Microbiology*, *13*. Retrieved January 27, 2024, from https://www.frontiersin.org/articles/10.3389/fcimb.2023.1108235

Maier, R. H., Maier, C. J., Hintner, H., Bauer, J. W., & Önder, K. (2012). Quantitative real-time PCR as a sensitive protein–protein interaction quantification method and a partial solution for non-accessible autoactivator and false-negative molecule analysis in the yeast two-hybrid system. *Methods*, *58*(4), 376–384. https://doi.org/10.1016/j.ymeth.2012.09.001

Mantovani, F., Collavin, L., & Del Sal, G. (2019). Mutant p53 as a guardian of the cancer cell. *Cell Death & Differentiation*, *26*(2), 199–212. https://doi.org/10.1038/s41418-018-0246-9

Marasco, M., & Carlomagno, T. (2020). Specificity and regulation of phosphotyrosine signaling through SH2 domains. *Journal of Structural Biology: X*, *4*, 100026. https://doi.org/10.1016/j.yjsbx.2020.100026

Maris, A. E., Sawaya, M. R., Kaczor-Grzeskowiak, M., Jarvis, M. R., Bearson, S. M. D., Kopka, M. L., Schröder, I., Gunsalus, R. P., & Dickerson, R. E. (2002). Dimerization allows DNA target site recognition by the NarL response regulator. *Nature Structural Biology*, *9*(10), 771–778. https://doi.org/10.1038/nsb845

McCraith, S., Holtzman, T., Moss, B., & Fields, S. (2000). Genome-wide analysis of vaccinia virus protein–protein interactions. *Proceedings of the National Academy of Sciences*, *97*(9), 4879–4884. https://doi.org/10.1073/pnas.080078197

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference*, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

McPherson, A., & Gavira, J. A. (2014). Introduction to protein crystallization. *Acta Crystallographica Section F: Structural Biology Communications*, *70*(1), 2–20. https://doi.org/10.1107/S2053230X13033141

Mészáros, B., Erdős, G., & Dosztányi, Z. (2018). IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research*, *46*(W1), W329–W337. https://doi.org/10.1093/nar/gky384

Mészáros, B., Simon, I., & Dosztányi, Z. (2009). Prediction of Protein Binding Regions in Disordered Proteins. *PLoS Computational Biology*, *5*(5), e1000376. https://doi.org/10.1371/journal.pcbi.1000376

Michaelis, A. C., Brunner, A.-D., Zwiebel, M., Meier, F., Strauss, M. T., Bludau, I., & Mann, M. (2023). The social and structural architecture of the yeast protein interactome. *Nature*, *624*(7990), 192–200. https://doi.org/10.1038/s41586-023-06739-5

Milne, J. L. S., Borgnia, M. J., Bartesaghi, A., Tran, E. E. H., Earl, L. A., Schauder, D. M., Lengyel, J., Pierson, J., Patwardhan, A., & Subramaniam, S. (2013). Cryo-electron microscopy – a primer for the non-microscopist. *The FEBS Journal*, *280*(1), 28–45. https://doi.org/10.1111/febs.12078

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., & Steinegger, M. (2022). ColabFold: Making protein folding accessible to all. *Nature Methods*, *19*(6), 679–682. https://doi.org/10.1038/s41592-022-01488-1

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Mitrea, D. M., & Kriwacki, R. W. (2016). Phase separation in biology; functional organization of a higher order. *Cell Communication and Signaling*, *14*(1), 1. https://doi.org/10.1186/s12964-015-0125-7

Mohan, A., Uversky, V. N., & Radivojac, P. (2009). Influence of Sequence Changes and Environment on Intrinsically Disordered Proteins. *PLOS Computational Biology*, *5*(9), e1000497. https://doi.org/10.1371/journal.pcbi.1000497

Mok, J., Kim, P. M., Lam, H. Y. K., Piccirillo, S., Zhou, X., Jeschke, G. R., Sheridan, D. L., Parker, S. A., Desai, V., Jwa, M., Cameroni, E., Niu, H., Good, M., Remenyi, A., Ma, J.-L. N., Sheu, Y.-J., Sassi, H. E., Sopko, R., Chan, C. S. M., ... Turk, B. E. (2010). Deciphering Protein Kinase Specificity Through Large-Scale Analysis of Yeast Phosphorylation Site Motifs. *Science Signaling*, *3*(109), ra12–ra12. https://doi.org/10.1126/scisignal.2000482

Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2007). Hot spots—A review of the protein–protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics*, *68*(4), 803–812. https://doi.org/10.1002/prot.21396

Mosca, R., Céol, A., & Aloy, P. (2013). Interactome3D: Adding structural details to protein networks. *Nature Methods*, *10*(1), 47–53. https://doi.org/10.1038/nmeth.2289

Mosca, R., Céol, A., Stein, A., Olivella, R., & Aloy, P. (2014). 3did: A catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Research*, *42*(D1), D374–D379. https://doi.org/10.1093/nar/gkt887

Nim, S., Jeon, J., Corbi-Verge, C., Seo, M.-H., Ivarsson, Y., Moffat, J., Tarasova, N., & Kim, P. M. (2016). Pooled screening for antiproliferative inhibitors of protein-protein interactions. *Nature Chemical Biology*, *12*(4), 275–281. https://doi.org/10.1038/nchembio.2026

Northrop, J. H. (1930). CRYSTALLINE PEPSIN : I. ISOLATION AND TESTS OF PURITY. *Journal of General Physiology*, *13*(6), 739–766. https://doi.org/10.1085/jgp.13.6.739

O'Keefe, K., Li, H., & Zhang, Y. (2003). Nucleocytoplasmic shuttling of p53 is essential for MDM2-mediated cytoplasmic degradation but not ubiquitination. *Molecular and Cellular Biology*, *23*(18), 6396–6405. https://doi.org/10.1128/MCB.23.18.6396-6405.2003

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., . . . Hermjakob, H. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research*, *42*(D1), D358–D363. https://doi.org/10.1093/nar/gkt1115

Pagel, P., Wong, P., & Frishman, D. (2004). A Domain Interaction Map Based on Phylogenetic Profiling. *Journal of Molecular Biology*, *344*(5), 1331–1346. https://doi.org/10.1016/j.jmb.2004.10.019

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunić, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Orengo, C. A., Pandurangan, A. P., Rivoire, C., . . . Bateman, A. (2023). InterPro in 2022. *Nucleic Acids Research*, *51*(D1), D418–D427. https://doi.org/10.1093/nar/gkac993

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. Retrieved January 25, 2024, from http://jmlr.org/papers/v12/pedregosa11a.html

Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M., & Lupas, A. N. (2021). High-accuracy protein structure prediction in CASP14. *Proteins: Structure, Function, and Bioinformatics*, *89*(12), 1687–1699. https://doi.org/10.1002/prot.26171

Petrey, D., Zhao, H., Trudeau, S. J., Murray, D., & Honig, B. (2023). PrePPI: A Structure Informed Proteome-wide Database of Protein–Protein Interactions. *Journal of Molecular Biology*, *435*(14), 168052. https://doi.org/10.1016/j.jmb.2023.168052

Petsalaki, E., Stark, A., García-Urdiales, E., & Russell, R. B. (2009). Accurate Prediction of Peptide Binding Sites on Protein Surfaces. *PLoS Computational Biology*, *5*(3), e1000335. https://doi.org/10.1371/journal.pcbi.1000335

Pfleger, K. D. G., & Eidne, K. A. (2006). Illuminating insights into protein-protein interactions using bioluminescence resonance energy transfer (BRET). *Nature Methods*, *3*(3), 165–174. https://doi.org/10.1038/nmeth841

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., … Vidal, M. (2014). A Proteome-Scale Map of the Human Interactome Network. *Cell*, *159*(5), 1212–1226. https://doi.org/10.1016/j.cell.2014.10.050

Romero, P., Obradovic, Z., Li, X., Garner, E. C., Brown, C. J., & Dunker, A. K. (2001). Sequence complexity of disordered protein. *Proteins: Structure, Function, and Bioinformatics*, *42*(1), 38–48. https://doi.org/10.1002/1097-0134(20010101)42:1<38::AID-PROT50>3.0.CO;2-3

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G. I., Wang, Y., Kovács, I. A., Kamburov, A., Krykbaeva, I., Lam, M. H., Tucker, G., Khurana, V., Sharma, A., Liu, Y.-Y., Yachie, N., … Vidal, M. (2015). Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell*, *161*(3), 647–660. https://doi.org/10.1016/j.cell.2015.04.013

Sakaguchi, K., Saito, S., Higashimoto, Y., Roy, S., Anderson, C. W., & Appella, E. (2000). Damage-mediated phosphorylation of human p53 threonine 18 through a cascade mediated by a casein 1-like kinase. Effect on Mdm2 binding. *The Journal of Biological Chemistry*, *275*(13), 9278–9283. https://doi.org/10.1074/jbc.275.13.9278

Sali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, *234*(3), 779–815. https://doi.org/10.1006/jmbi.1993.1626

Santelli, E., Leone, M., Li, C., Fukushima, T., Preece, N. E., Olson, A. J., Ely, K. R., Reed, J. C., Pellecchia, M., Liddington, R. C., & Matsuzawa, S.-i. (2005). Structural analysis of Siah1-Siah-interacting protein interactions and insights into the assembly of an E3 ligase multiprotein complex. *The Journal of Biological Chemistry*, *280*(40), 34278–34287. https://doi.org/10.1074/jbc.M506707200

Schlessinger, A., Liu, J., & Rost, B. (2007). Natively Unstructured Loops Differ from Other Loops. *PLOS Computational Biology*, *3*(7), e140. https://doi.org/10.1371/journal.pcbi.0030140

Sekar, R. B., & Periasamy, A. (2003). Fluorescence resonance energy transfer (FRET) microscopy imaging of live cell protein localizations. *The Journal of Cell Biology*, *160*(5), 629–633. https://doi.org/10.1083/jcb.200210140

Sheng, Y., Saridakis, V., Sarkari, F., Duan, S., Wu, T., Arrowsmith, C. H., & Frappier, L. (2006). Molecular recognition of p53 and MDM2 by USP7/HAUSP. *Nature Structural & Molecular Biology*, *13*(3), 285–291. https://doi.org/10.1038/nsmb1067

Shin, J.-S., Ha, J.-H., Lee, D.-H., Ryu, K.-S., Bae, K.-H., Park, B. C., Park, S. G., Yi, G.-S., & Chi, S.-W. (2015). Structural convergence of unstructured p53 family transactivation domains in MDM2 recognition. *Cell Cycle (Georgetown, Tex.)*, *14*(4), 533–543. https://doi.org/10.1080/15384101.2014.998056

Shivhare, D., Musialak-Lange, M., Julca, I., Gluza, P., & Mutwil, M. (2021). Removing auto-activators from yeast-two-hybrid assays by conditional negative selection. *Scientific Reports*, *11*, 5477. https://doi.org/10.1038/s41598-021-84608-9

Sillitoe, I., Bordin, N., Dawson, N., Waman, V. P., Ashford, P., Scholes, H. M., Pang, C. S. M., Woodridge, L., Rauer, C., Sen, N., Abbasian, M., Le Cornu, S., Lam, S. D., Berka, K., Varekova, I. H., Svobodova, R., Lees, J., & Orengo, C. A. (2020). CATH: Increased structural coverage of functional space. *Nucleic Acids Research*, *49*(D1), D266–D273. https://doi.org/10.1093/nar/gkaa1079

Sobti, M., Mead, B. J., Stewart, A. G., Igreja, C., & Christie, M. (2023). Molecular basis for GIGYF-TNRC6 complex assembly. *RNA (New York, N.Y.)*, *29*(6), 724–734. https://doi.org/10.1261/rna.079596.123

Sørensen, C. S., & Kjaergaard, M. (2019). Effective concentrations enforced by intrinsically disordered linkers are governed by polymer physics. *Proceedings of the National Academy of Sciences*, *116*(46), 23124–23131. https://doi.org/10.1073/pnas.1904813116

Sprinzak, E., & Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction1 1Edited by G. von Heijne. *Journal of Molecular Biology*, *311*(4), 681–692. https://doi.org/10.1006/jmbi.2001.4920

Stein, A., & Aloy, P. (2010). Novel Peptide-Mediated Interactions Derived from High-Resolution 3-Dimensional Structures. *PLOS Computational Biology*, *6*(5), e1000789. https://doi.org/10.1371/journal.pcbi.1000789

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., . . . Wanker, E. E. (2005). A Human Protein-Protein Interaction Network: A Resource for Annotating the Proteome. *Cell*, *122*(6), 957–968. https://doi.org/10.1016/j.cell.2005.08.029

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., & Mering, C. v. (2019). STRING v11: Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, *47*, D607–D613. https://doi.org/10.1093/nar/gky1131

Titeca, K., Lemmens, I., Tavernier, J., & Eyckerman, S. (2019). Discovering cellular protein-protein interactions: Technological strategies and opportunities. *Mass Spectrometry Reviews*, *38*(1), 79–111. https://doi.org/10.1002/mas.21574

To, P., Bhagwat, A. M., Tarbox, H. E., Ecer, A., Wendorff, H., Jamieson, Z., Trcek, T., & Fried, S. D. (2023, June 25). *Intrinsically Disordered Regions Promote Protein Refoldability and Facilitate Retrieval from Biomolecular Condensates*. https://doi.org/10.1101/2023.06.25.546465

Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Letters*, *579*(15), 3346–3354. https://doi.org/10.1016/j.febslet.2005.03.072

Tompa, P., Davey, N. E., Gibson, T. J., & Babu, M. M. (2014). A Million Peptide Motifs for the Molecular Biologist. *Molecular Cell*, *55*(2), 161–169. https://doi.org/10.1016/j.molcel.2014.05.032

Truebestein, L., & Leonard, T. A. (2016). Coiled-coils: The long and short of it. *Bioessays*, *38*(9), 903–916. https://doi.org/10.1002/bies.201600062

Tsaban, T., Varga, J. K., Avraham, O., Ben-Aharon, Z., Khramushin, A., & Schueler-Furman, O. (2022). Harnessing protein folding neural networks for peptide–protein docking. *Nature Communications*, *13*(1), 176. https://doi.org/10.1038/s41467-021-27838-9

Turenne, G. A., & Price, B. D. (2001). Glycogen synthase kinase3 beta phosphorylates serine 33 of p53 and activates p53's transcriptional activity. *BMC cell biology*, *2*, 12. https://doi.org/10.1186/1471-2121-2-12

Ubersax, J. A., & Ferrell Jr, J. E. (2007). Mechanisms of specificity in protein phosphorylation. *Nature Reviews Molecular Cell Biology*, *8*(7), 530–541. https://doi.org/10.1038/nrm2203

Uetz, P., Dong, Y.-A., Zeretzke, C., Atzler, C., Baiker, A., Berger, B., Rajagopala, S. V., Roupelieva, M., Rose, D., Fossum, E., & Haas, J. (2006). Herpesviral Protein Networks and Their Interaction with the Human Proteome. *Science*, *311*(5758), 239–242. https://doi.org/10.1126/science.1116804

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., & Rothberg, J. M. (2000). A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature*, *403*(6770), 623–627. https://doi.org/10.1038/35001009

Van Roey, K., Dinkel, H., Weatheritt, R. J., Gibson, T. J., & Davey, N. E. (2013). The switches.ELM Resource: A Compendium of Conditional Regulatory Interaction Interfaces. *Science Signaling*, *6*(269), rs7–rs7. https://doi.org/10.1126/scisignal.2003345

Van Roey, K., Uyar, B., Weatheritt, R. J., Dinkel, H., Seiler, M., Budd, A., Gibson, T. J., & Davey, N. E. (2014). Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. *Chemical Reviews*, *114*(13), 6733–6778. https://doi.org/10.1021/cr400585q

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., . . . Velankar, S. (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. https://doi.org/10.1093/nar/gkab1061

Verma, R., & Pandit, S. B. (2019). Unraveling the structural landscape of intra-chain domain interfaces: Implication in the evolution of domain-domain interactions. *PLOS ONE*, *14*(8), e0220336. https://doi.org/10.1371/journal.pone.0220336

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Vyncke, L., Masschaele, D., Tavernier, J., & Peelman, F. (2019). Straightforward Protein-Protein Interaction Interface Mapping via Random Mutagenesis and Mammalian Protein Protein Interaction Trap (MAPPIT). *International Journal of Molecular Sciences*, *20*(9), 2058. https://doi.org/10.3390/ijms20092058

Wallace, L. A., Burke, J., & Dirr, H. W. (2000). Domain–domain interface packing at conserved Trp-20 in class α glutathione transferase impacts on protein stability. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, *1478*(2), 325–332. https://doi.org/10.1016/S0167-4838(00)00023-6

Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., & Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology*, *30*(2), 159–164. https://doi.org/10.1038/nbt.2106

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Weatheritt, R. J., Luck, K., Petsalaki, E., Davey, N. E., & Gibson, T. J. (2012). The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics*, *28*(7), 976–982. https://doi.org/10.1093/bioinformatics/bts072

Weatheritt, R. J., Jehl, P., Dinkel, H., & Gibson, T. J. (2012). iELM—a web server to explore short linear motif-mediated interactions. *Nucleic Acids Research*, *40*(W1), W364–W369. https://doi.org/10.1093/nar/gks444

Weihs, F., Wang, J., Pfleger, K. D. G., & Dacres, H. (2020). Experimental determination of the bioluminescence resonance energy transfer (BRET) Förster distances of NanoBRET and red-shifted BRET pairs. *Analytica Chimica Acta: X*, *6*, 100059. https://doi.org/10.1016/j.acax.2020.100059

Weinfeld, M., Mani, R. S., Abdou, I., Acetuno, R. D., & Glover, J. M. (2011). Tidying up loose ends: The role of polynucleotide kinase/phosphatase in DNA strand break repair. *Trends in biochemical sciences*, *36*(5), 262–271. https://doi.org/10.1016/j.tibs.2011.01.006

Wells, J. A., & McClendon, C. L. (2007). Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature*, *450*(7172), 1001–1009. https://doi.org/10.1038/nature06526

Wetlaufer, D. B., & Ristow, S. (1973). Acquisition of Three-Dimensional Structure of Proteins. *Annual Review of Biochemistry*, *42*(1), 135–158. https://doi.org/10.1146/annurev.bi.42.070173.001031

What's the next word in large language models? (2023). *Nature Machine Intelligence*, *5*(4), 331–332. https://doi.org/10.1038/s42256-023-00655-z

Wilson, C. J., Choy, W.-Y., & Karttunen, M. (2022). AlphaFold2: A Role for Disordered Protein/Region Prediction? *International Journal of Molecular Sciences*, *23*(9), 4591. https://doi.org/10.3390/ijms23094591

Winston, J. T., Strack, P., Beer-Romero, P., Chu, C. Y., Elledge, S. J., & Harper, J. W. (1999). The SCFbeta-TRCP-ubiquitin ligase complex associates specifically with phosphorylated destruction motifs in IkappaBalpha and beta-catenin and stimulates IkappaBalpha ubiquitination in vitro. *Genes & Development*, *13*(3), 270–283. https://doi.org/10.1101/gad.13.3.270

Wodak, S. J., & Janin, J. (1978). Computer analysis of protein-protein interaction. *Journal of Molecular Biology*, *124*(2), 323–342. https://doi.org/10.1016/0022-2836(78)90302-9 wwPDB consortium. (2019). Protein Data Bank: The single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, *47*(D1), D520–D528. https://doi.org/10.1093/nar/gky949

Xu, Q., & Dunbrack, R. L. (2019). Principles and characteristics of biological assemblies in experimentally determined protein structures. *Current opinion in structural biology*, *55*, 34–49. https://doi.org/10.1016/j.sbi.2019.03.006

Yang, A. S., & Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *Journal of Molecular Biology*, *301*(3), 665–678. https://doi.org/10.1006/jmbi.2000.3973

Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., & Jothi, R. (2011). DOMINE: A comprehensive collection of known and predicted domain-domain interactions. *Nucleic Acids Research*, *39*, D730–D735. https://doi.org/10.1093/nar/gkq1229

Zacchi, P., Gostissa, M., Uchida, T., Salvagno, C., Avolio, F., Volinia, S., Ronai, Z., Blandino, G., Schneider, C., & Del Sal, G. (2002). The prolyl isomerase Pin1 reveals a mechanism to control p53 functions after genotoxic insults. *Nature*, *419*(6909), 853–857. https://doi.org/10.1038/nature01120

Zámocký, M., & Koller, F. (1999). Understanding the structure and function of catalases: Clues from molecular evolution and in vitro mutagenesis. *Progress in Biophysics and Molecular Biology*, *72*(1), 19–66. https://doi.org/10.1016/S0079-6107(98)00058-3

Zhang, Q. C., Deng, L., Fisher, M., Guan, J., Honig, B., & Petrey, D. (2011). PredUs: A web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Research*, *39*, W283–287. https://doi.org/10.1093/nar/gkr311

Zhang, Q. C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C. A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., Maniatis, T., Califano, A., & Honig, B. (2012). Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, *490*(7421), 556–560. https://doi.org/10.1038/nature11503

Zhang, Q. C., Petrey, D., Norel, R., & Honig, B. H. (2010). Protein interface conservation across structure space. *Proceedings of the National Academy of Sciences*, *107*(24), 10896–10901. https://doi.org/10.1073/pnas.1005894107

Zhao, H., Du, H., Zhao, S., Chen, Z., Li, Y., Xu, K., Liu, B., Cheng, X., Wen, W., Li, G., Chen, G., Zhao, Z., Qiu, G., Liu, P., Zhang, T. J., Wu, Z., & Wu, N. (2024). SIGMA leverages protein structural information to predict the pathogenicity of missense variants. *Cell Reports Methods*, *4*(1), 100687. https://doi.org/10.1016/j.crmeth.2023.100687

Zhao, X.-M., Chen, L., & Aihara, K. (2010). A discriminative approach for identifying domain–domain interactions from protein–protein interactions. *Proteins:*

Structure, Function, and Bioinformatics, *78*(5), 1243–1253. https://doi.org/10.1002/prot.22643

CHOP YAN LEE

lcy9977@hotmail.com | +49 152 13858677 | Mainz, Germany
LinkedIn | GitHub

Profile

An an **aspiring data scientist** with a profound interest in **artificial intelligence** and a solid background in **bioinformatics**, my professional journey includes the successful design of an **automation program** for search tasks, coupled with a **random forest** model to enhance search accuracy. Proficient in **Python** and **SQL**, I have a proven track record in the full lifecycle of **machine learning** models, from **training** to **deployment**, as well as the storage and analysis of complex data. I am eager to apply my skills to tackle real-world challenges in data science.

Professional Experience

Researcher **April 2020 - present**

Institute of Molecular Biology, Germany

- Built a program to automate search tasks and **trained a machine learning model** to score the search result.
- **Benchmarked deep learning tools** and developed approach to boost its sensitivity.
- Developed robust pipelines to **process, store, analyse,** and **visualise** complex biological data.
- **Curated high quality dataset** for the development of machine learning models.
- Demonstrated excellent leadership in supervising master's students for their theses.
- Provided data processing and analysis support to PhD student network and colleagues.

Education

- **Doctor of Philosophy in Life Science** | Johannes Gutenberg University, Germany | April 2020 - present
- **Master in Genetics and Evolution** | University of Granada, Spain | September 2018 - September 2019
- **Bachelor of Science** | University of Melbourne, Australia | February 2015 - November 2017

Skills

- Proficient in **Python OOP** and **packages**, including **pandas, scipy** and **numpy** for data wrangling and analysis; **seaborn** and **matplotlib** for data visualization; **scikit-learn** for machine learning models; and the API of bioinformatic tools for automation.
- Trained in **MySQL** for efficient **data management** and proficient in query formulation for data manipulation in relational databases.
- Experienced in **machine learning statistics** and **techniques**, such as **confusion matrix**, **ROC** and **PR curve analysis**, feature importance analysis, **hyperparameter tuning** with **GridSearch,** and **k-fold cross-validation**, among others.
- Competent in distilling complex concepts and presenting them to diverse audiences.
- Excellent communicator and team player in cross-disciplinary collaborations.
- Organized, adaptable, and down-to-earth with a steadfast commitment to learning.

Volunteering Experience

Cultural Exchange in Trujillo, Peru **November 2016 - January 2017**

AIESEC Australia

- Lived with a local host family and worked with international teams to promote environmental awareness and enhance local infrastructure.

Languages

Chinese | Native   English | Fluent   Spanish | Fluent   German | Basic