# Structure-guided evolutionary analysis of protein-protein interactions and interactome network rewiring at single residue resolution in yeasts.

Léah Pollet

Biological and Biomedical Engineering Department, Faculty of Engineering

McGill University, Montreal

December 13th

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Ph.D. in Biological and Biomedical Engineering

## Table of contents

## Abstract

### English

Protein-protein interactions, or PPIs, are important phenomena, essential to proper protein function, and present in virtually all biological pathways of cells. Accordingly, in recent years, numerous experiments have been performed to survey all proteins that interact in a given species, as well as to uncover the molecular structure and 3D mechanisms of interactions between individual proteins. So far, this extensive work has generated large amounts of data, which now allows us to study the evolution of PPIs, a feat that was previously difficult due to a lack of high-quality experimental results. An investigation into the evolution of PPIs is essential to try and uncover the evolutionary design principles behind variations in PPIs, both within and between species. Here, we take advantage of PPI datasets made recently available for two yeast species, *Saccharomyces cerevisiae* (*S. cerevisiae*), and *Schizosaccharomyces pombe* (*S. pombe*), and perform their thorough analysis using bioinformatics tools. We first design a custom script pipeline to automate the curation of high-quality protein-protein interaction data from online databases and organize this data into structural models of PPIs for the two yeast species, *S. cerevisiae*, and *S. pombe*. These structural models are subsequently used to investigate the relationship between PPI structure and PPI evolution in yeast at the single residue level. This analysis yields significant insight into the design principles and structural mechanisms governing PPI evolution in yeast, uncovering several structural properties directly correlated with the evolutionary rates of PPIs. Finally, we use structural models of *S. cerevisiae* and *S. pombe* PPIs to construct structurally-resolved interactome networks for the two yeasts and compare PPIs that are preserved and PPI that are different between the two yeast species. This analysis yields further insight into the evolutionary design principles of PPIs and the mechanisms by which interactions are preserved or

rewired between species, improving our understanding of the molecular evolution of PPIs at the residue level. Overall, this work establishes a better picture of the evolution of PPIs, both (1) at the molecular level, by uncovering small-scale structural properties that influence the evolution of PPIs within a species; and (2) at the phylogenetic level, by identifying mechanisms leading to large-scale differences in PPIs between species. Our findings have wide-ranging applications to the study of mis-regulation and disruption of PPIs, two processes that are commonly associated with various diseases.

## Résumé

### Français

Titre de la thèse :

Analyse structurelle à haute résolution de l'évolution des interactions protéine-protéine et des changements de réseaux d'interactions chez les levures.

Les interactions protéine-protéine, ou IPPs, sont des processus importants qui jouent un rôle fondamental à tous les niveaux de la cellule, elles sont essentielles au bon fonctionnement des protéines. De nombreuses expériences ont donc été menées récemment afin de cataloguer les protéines qui interagissent dans une espèce donnée, et pour déterminer la structure moléculaire et les mécanismes d'interaction 3D entre ces protéines. Ces travaux approfondis ont généré de grandes quantités de données, qui permettent désormais d'étudier l'évolution des IPPs. Le manque de résultats expérimentaux de qualité rendait auparavant cette tâche difficile. Étudier l'évolution des IPP est nécessaire pour découvrir les principes de conception évolutive qui expliquent les variations qui peuvent être observées entre différentes IPPs au sein d'une même espèce, mais aussi entre différentes espèces. Nous utilisons donc les données d'IPP récemment rendues disponibles pour deux espèces de levures, *Saccharomyces cerevisiae* (*S. cerevisiae*) et *Schizosaccharomyces pombe* (*S. pombe*), et effectuons leur analyse approfondie à l'aide d'outils bio-informatiques. Nous concevons d'abord une succession de scripts informatiques afin d'automatiser la collecte de données d'interaction protéine-protéine de haute qualité à partir de bases de données. Cette tache permet alors l'organisation de ces données d'IPP en modèles structurels pour les IPPs des deux espèces de levure *S. cerevisiae* et *S. pombe*. Ces modèles structurels sont ensuite utilisés pour étudier la relation entre la structure d'une IPP et son évolution chez les levures, à haute résolution.

Cette analyse produit des informations pertinentes sur les principes de conception et les mécanismes structurels qui régulent l'évolution des IPP chez les levures, et révèle plusieurs propriétés structurelles directement corrélées avec les taux d'évolution des IPP. D'autre part, nous utilisons les modèles structurels des IPP chez *S. cerevisiae* et *S. pombe* pour construire des réseaux d'interactions structurels pour les deux levures, et ainsi comparer les IPPs qui sont préservées et les IPPs sont différentes entre les deux espèces de levures. Cette analyse produit des informations supplémentaires sur les principes de conception évolutive des IPPs, et quant aux mécanismes par lesquels certaines interactions sont préservées et d'autres sont différentes entre les espèces. Ceci améliore notre compréhension de l'évolution moléculaire des IPPs. En conclusion, ce travail établit une meilleure image de l'évolution des IPP à deux niveaux : (1) à l'échelle moléculaire, en découvrant des propriétés structurelles à petite échelle qui influencent l'évolution des IPP au sein d'une espèce. (2) à l'échelle phylogénétique, en identifiant les mécanismes conduisant à des différences à grande échelle d'IPP entre espèces. Nos résultats ont de nombreuses applications pour l'étude du dérèglement et de la perturbation des IPP, deux processus fréquemment associés à diverses maladies.

## Acknowledgements

First, I would like to thank my family for their unconditional love and support and for always cheering me on and encouraging me in my choices, even when those choices took me far across the world from them. I truly appreciate all the time and conversations about my work, your efforts to translate and make sense of everything I do mean the world to me.

I also want to express my gratitude to my supervisor Yu (Brandon) Xia for all the support, expertise and dedication throughout the years. Thank you for shaping me into the inquisitive scientist I am today and for many conversations, freely sharing and inspiring me with your fascinating outlooks on life and science.

I thank Luke Lambourne, for all his hard work and constructive input on the methodologies and results presented in Chapter 3. I would also like to thank the members of my advisory committee and defense committee Sebastian Wachsmann Hogiu, Paul Harrison, Allen Ehrlicher, Amin Emad, Codruta Ignea, Adrian Serohijos and Andrew Bateman for their invaluable feedback throughout this process. I appreciated your insights, suggestions, and willingness to guide my project.

Thank you to all my friends, colleagues and members of my lab for great discussions, insights and unwavering support through some of the rougher patches. This experience would not have been the same without you.

Finally, I want to thank all the funding sources that made my project possible. NSERC (Natural Sciences and Engineering Research Council of Canada), McGill University, the Canada Foundation for Innovation, and the Canada Research Chairs program.

## Contributions to Original knowledge

This work is divided into three separate aims, each with associated deliverables and original contributions:

The **first aim** focuses on the automated curation of protein-protein interaction (PPI) data from online databases, and its organization into molecular models of PPIs for two yeast species, *Saccharomyces cerevisiae* (*S. cerevisiae*), and *Schizosaccharomyces pombe* (*S. pombe*). This first step gathers and combines large amounts of PPI data from very different experimental fields, therefore allowing for novel analysis.

This aim utilizes a large amount of data obtained through extensive experimental work over the last decade. Records of all interactions between pairs of proteins in a given organism (also called interactomes) curated from numerous experimental projects are currently available on the BioGRID and IntAct databases. Additionally, over 25 000 molecular structures (detailed, atom-resolution, three-dimensional descriptions of individual PPIs) obtained from various experiments are currently available on the PDB database. For this aim, a custom script pipeline was, therefore, developed to automate the gathering and quality control of the PPI data described above for both *S. cerevisiae*, and *S. pombe*. Both *S. cerevisiae* and *S. pombe* PPI data have been successfully curated. Following data collection, additional processing and combining steps were automated to systematically store and organize the data into what we call molecular models of PPIs for both *S. cerevisiae* and *S. pombe*. The custom script pipeline designed to gather and combine large amounts of PPI data from very different experimental fields is novel and could be applied to future works in the two yeasts or in other species. Moreover, the detailed molecular models of PPIs in *S. cerevisiae* and *S. pombe* generated here, combine PPI data at two very different scales in a unique

manner, enabling novel future analysis. The pipeline and molecular models of PPIs generated in this aim are publicly available as a GitHub repository published in **Research Article No. 1: Structural Determinants of Yeast Protein-Protein interaction Interface Evolution at the Residue Level**, as well as further detailed in Chapter 3.

The **second aim** of the project focuses on utilizing the data gathered in Aim 1 to study the relationship between PPI structures and their evolution in yeast. Evolutionary rates are known to vary widely, even within the same PPI protein, with some residues being conserved during evolution, and others being much more variable. However, the design principles and structural mechanisms governing PPI evolution and responsible for those observed differences in evolutionary rates remain mostly unknown. In this aim we, therefore, uncover some of those principles and further our understanding of the evolution of PPIs. This feat broadens our knowledge of cellular mechanisms, and has practical applications to disease diagnosis and treatment, synthetic biology, and genome engineering.

Extensive work in identifying structural determinants (i.e., measurable quantities, characteristic of the structure of the microenvironment surrounding a residue) correlated with residue evolution in *S. cerevisiae* PPIs was performed. The final structural determinants selected in this aim are the change in relative solvent accessibility upon PPI binding (ΔRSA), the number of residue-residue contacts across the PPI interface (interRRC), and the distance from the center (dCenter) or the periphery (dEdges) of the PPI interface. Several significant correlations between these structural determinants and residue evolutionary rates in *S. cerevisiae* PPIs were uncovered. The relationships uncovered, while supporting results from previous work using single protein structures, also identified determinants uniquely important to the investigation of PPI structures.

Additional analysis aiming to estimate the relative importance of those determinants, as well as quantifying their overall contribution to our understanding of the relationship between structure and evolution of PPIs data was also performed. The following important conclusions were established: (i) interfacial residues in PPIs are subject to continuous, structure-based selective constraints proportional to their degree of interface involvement, (ii) interfacial burial (as measured by the structural determinant ΔRSA) is selectively equivalent to non-interfacial burial, (iii) in addition to ΔRSA, other measures of interface involvement (structural determinants interRRC, dCenter, and dEdges) independently constrain residue evolution, and (iv) in addition to these continuous structure-based selective constraints, interfacial residues are subject to a fixed function-based selective constraint independent of their degree of interface involvement. Those findings are published in **Research Article No. 1: Structural Determinants of Yeast Protein-Protein interaction Interface Evolution at the Residue Level**, as well as further detailed in Chapter 3.

The **third aim** of the project focuses on the comparison of PPIs between *S. cerevisiae* and *S. pombe* to uncover possible drivers for observed differences in interactomes between the two yeasts. We classify PPIs according to whether they are preserved or different between the two yeast species and compare site-specific evolutionary rates of interfacial versus non-interfacial residues for these different categories of PPIs. This last aim of comparative analysis between two species' interactomes uncovers some of the molecular mechanisms behind the phylogenetic loss or gain of an interaction between two species. Moreover, insights gained from studying the phylogenetic loss or gain of interactions could have wide-ranging applications to the study of misregulation and disruption of PPIs associated with various diseases.

Work towards this aim uncovered the following important trends: (i) residues in PPI interfaces evolve significantly more slowly than non-interfacial residues when using lineage-specific measures of evolutionary rate, but not when using non-lineage-specific measures, (ii) both lineage-specific and non-lineage-specific evolutionary rate measures can distinguish interfacial residues from non-interfacial residues for preserved PPIs between the two yeasts, but only the lineage-specific measure is appropriate for PPIs that are different between the two yeasts, (iii) both lineage-specific and non-lineage-specific evolutionary rate measures are appropriate for elucidating structural determinants of protein evolution for residues outside of PPI interfaces. Overall, our results demonstrate that unlike tertiary structures of single proteins, PPIs and PPI interfaces can be highly volatile in their evolution, thus requiring the use of lineage-specific measures when studying their evolution. These results yield insight into the evolutionary design principles of PPIs and the mechanisms by which interactions are preserved or different between species, improving our understanding of the molecular evolution of PPIs and PPI interfaces at the residue level. Those findings are published in **Research Article No. 2: Structure-guided evolutionary analysis of interactome network rewiring at single residue resolution in yeasts**, as well as further detailed in Chapter 4.

Overall, this project establishes a better picture of the evolution of PPIs, both at the molecular level, by uncovering small-scale structural properties that influence the evolution of protein interactions in a species; and at the phylogenetic level, by identifying mechanisms leading to large-scale differences in PPIs between species.

## Contributions of Authors

This thesis consists of two manuscripts (Chapters 3 and Chapter 4). I am the first author of both manuscripts. The contributions of authors to each manuscript are listed below.

| | |
|---|---|
| **Chapter 3:** | **Structural Determinants of Yeast Protein-Protein interaction Interface Evolution at the Residue Level** |
| Published in: | *Journal of Molecular Biology* 2022 |
| Authors: | **Léah Pollet,** Luke Lambourne, and Yu Xia |
| Contributions: | **LP:** Data curation, Formal analysis, Software, Writing – original draft.<br>LL: Conceptualization, Methodology, Investigation.<br>YX: Funding acquisition, Conceptualization, Supervision, Writing – review & editing |

| | |
|---|---|
| **Chapter 4:** | **Structure-guided evolutionary analysis of interactome network rewiring at single residue resolution in yeasts** |
| Published in: | *Journal of Molecular Biology* 2024 |
| Authors: | **Léah Pollet** and Yu Xia |
| Contributions: | **LP:** Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Software, Writing – original draft.<br>YX: Funding acquisition, Conceptualization, Supervision, Writing – review & editing |

## List of Figures and Tables

Figures and tables are grouped by chapter.

**Chapter 2. Literature review**

**List of figures**



**Chapter 3. Structural Determinants of Yeast Protein-Protein interaction Interface Evolution at the Residue Level**

**List of figures**

**Figure S1.** The difference in evolutionary rate between interfacial and non-interfacial residues.

**Figure S2.** The relationship between structural properties and evolutionary rate for residues in PPI interfaces.

**Figure S3.** The difference in evolutionary rate between interfacial and non-interfacial residues with additional species included in evolutionary rate calculations – Analysis S1, S2 and S3.

**Figure S4.** The relationship between solvent accessibility and evolutionary rate in PPI interfaces with additional species included in evolutionary rate calculations – Analysis S1, S2 and S3.

**Figure S5.** The relationship between interface involvement and evolutionary rate for residues in PPI interfaces with additional species included in evolutionary rate calculations – Analysis S1.

**Figure S6.** The relationship between interface involvement and evolutionary rate for residues in PPI interfaces with additional species included in evolutionary rate calculations -Analysis S2, S3.

**Figure S7.** Graphical representation of the homology-based structural annotation transfer and evolutionary sequence analysis portion of our data curation pipeline.

**Figure S8.** The distribution of evolutionary rate for interfacial and non-interfacial residues.

**Figure S9.** The difference in evolutionary rate between interfacial and non-interfacial residues (for high-sequence-identity PPIs– Analysis S4).

**Figure S10.** The relationship between solvent accessibility and evolutionary rate in PPI interfaces (for high-sequence-identity PPIs – Analysis S4).

**Figure S11.** The relationship between interface involvement and evolutionary rate for residues in PPI interfaces (for high-sequence-identity PPIs – Analysis S4).

**Figure S12.** Correlation between structural properties of a residue's microenvironment.

**Figure S13**. The difference in evolutionary rate between interfacial and non-interfacial residues.

**List of tables**



**Chapter 4. Structure-guided evolutionary analysis of interactome network rewiring at single residue resolution in yeasts**

**List of Figures**

**Figure 3.** The difference in evolutionary rate between interfacial and non-interfacial residues for preserved, missing ortholog, and rewired PPIs.

**Figure 4.** The difference in evolutionary rate between buried and exposed residues outside of PPI interfaces.

**Figure 5.** The difference in evolutionary rate between buried and exposed residues outside of PPI interfaces for preserved, missing ortholog and rewired PPIs.

**Figure S1.** The difference in evolutionary rate between interfacial and pseudo-interfacial residues within a species.

**Figure S2.** The difference in evolutionary rate between interfacial residues and pseudo-interfacial residues across species.

**Figure S3.** The difference in average evolutionary rate for non-interfacial surface and interior residues and interfacial rim support and core residues.

**Figure S4.** The difference in average evolutionary rate for non-interfacial surface and interior residues and interfacial rim support and core residues in preserved, missing ortholog, and rewired PPIs.

**Figure S5.** The difference in evolutionary rate between interfacial and non-interfacial residues. Repeat analysis using only experimentally determined protein complex structures, with no homology-based PPI structural models.

**Figure S6.** The difference in evolutionary rate between interfacial and non-interfacial residues for preserved, missing ortholog, and rewired PPIs. Repeat analysis using only experimentally determined protein complex structures, with no homology-based PPI.

**Figure S7.** The difference in evolutionary rate between buried and exposed residues outside of PPI interfaces. Repeat analysis using only experimentally determined protein complex structures, with no homology-based PPI structural models.

**Figure S8.** The difference in evolutionary rate between buried and exposed residues outside of PPI interfaces for preserved, missing ortholog and rewired PPIs. Repeat analysis using only experimentally determined protein complex structures, with no homolog.

**List of tables**

**Table S1.** PPI dataset summary.

**Table S2.** Interface size comparison between PPI types.

## List of Abbreviations

| | |
|---|---|
| PPI | Protein-protein interaction |
| *S. cerevisiae* | *Saccharomyces cerevisiae* |
| *S. pombe* | *Schizosaccharomyces pombe* |
| BioGRID | Biological General Repository for Interaction Datasets |
| IntAct | IntAct Molecular Interaction Database |
| PDB | Protein Data Bank |
| ΔRSA | Relative solvent accessibility upon PPI binding |
| interRRC | Number of residue-residue contacts across the PPI interface |
| dCenter | Distance from the center of the PPI interface |
| dEdges | Distance from the periphery of the PPI interface |
| 3D | Three-dimensional |
| GPCR | G-protein-coupled receptor |
| GTP | Guanosine triphosphate |
| GDP | Guanosine diphosphate |
| TCR | T-cell receptor |
| MHC | Major histocompatibility complex |
| NPC | Nuclear pore complex |
| ORC | Origin recognition complex |
| CML | Chronic myeloid leukemia |
| ALS | Amyotrophic lateral sclerosis |
| LDL | Low-density lipoproteins |
| ATP | Adenosine triphosphate |

| | |
|---|---|
| CRC | Colorectal cancer |
| FDA | Food and Drug Administration |
| CARs | Chimeric antigen receptors |
| CRISPR | Clustered regularly interspaced short palindromic repeats |
| Cas-9 | CRISPR-associated protein 9 |
| Y2H | Yeast Two-Hybrid |
| POI | Protein of interest |
| DBD | DNA-binding domain |
| AD | Activation domain |
| TAP-MS | Tandem affinity purification-mass spectroscopy |
| FRET | Fluorescence Resonance Energy Transfer |
| BiFC | Bimolecular Fluorescence Complementation |
| GFP | Green fluorescent protein |
| Co-IP | Co-Immunoprecipitation |
| SPR | Surface Plasmon Resonance |
| MD | Molecular Dynamics |
| STRING | Search Tool for the Retrieval of Interacting Genes/Proteins |
| DIP | Database of Interacting Proteins |
| MINT | Molecular INTeraction database |
| Å | Angstrom |
| NMR | Nuclear Magnetic Resonance |
| kDa | Kilodalton |
| Cryo-EM | Cryo-Electron Microscopy |

| | |
|---|---|
| HDX-MS | Hydrogen-Deuterium Exchange Mass Spectrometry |
| XL-MS | Cross-linking Mass Spectrometry |
| PINT | Protein–protein Interactions Thermodynamic Database |
| SKEMPI | Structural database of Kinetics and Energetics of Mutant Protein Interactions |
| DIPS | Database of Interacting Protein Structures |
| HOG | High-osmolarity glycerol |
| TOR, | Target of Rapamycin |
| PAML | Phylogenetic Analysis by Maximum Likelihood |
| ASA | Accessible surface area |
| SASA | Solvent accessible surface area |
| RSA | Relative solvent accessibility |
| ΔRSA | Change in relative solvent accessibility upon PPI binding |
| CN | Contact number |
| WCN | Weighted contact number |
| ORF | Open reading frame |

# 1. Introduction

Proteins are a vital component of all living organisms. In fact, next to water, they are the most plentiful substance in the human body [1]. These molecules can be found in all cells, where they perform many essential functions, such as allowing motion, distributing oxygen, clotting blood, fighting infections, transporting substances, controlling chemical reactions, and carrying messages from one part of the body to another. Proteins rarely act alone when accomplishing these complex tasks. Instead, they tend to cooperate with one another in a process termed protein-protein interaction (hereafter referred to as PPI) [2].

For instance, many major cellular processes, including DNA replication, transcription, translation, splicing, secretion, cell cycle control and signal transduction are carried out by stable protein-protein complexes, which behave as molecular machines, composed of protein components and organized by tightly regulated PPIs to ensure proper function [2–4]. Moreover, all manner of fundamental cellular processes, including cell growth, cell cycle, metabolic pathways, and signal transduction are controlled and regulated by more transient interactions such as the interactions of protein kinases, protein phosphatases, proteases and other enzymes with their substrate proteins [3,5]. Finally, transient protein-protein interactions are also crucial in the recruitment and assembly of the transcription complex to specific promoters, the transport of proteins across cellular membranes, the proper folding of proteins, and various individual steps of the translation cycle and cell cycle [3,6]. As such, mis-regulations and disruptions of the normal patterns of PPIs in human have been linked to various diseases including cancer, cardiomyopathies, diabetes, microbial infections, and genetic and neurodegenerative disorders [7–9].

Accordingly, in recent years, numerous experiments have been performed to try and survey all proteins that interact in a given species, as well as to uncover the molecular structure (detailed, atom-resolution, three-dimensional description of a PPI) and three-dimensional (3D) mechanisms of interactions between individual proteins. Records of all proteins that interact in a given species, also called interactomes, were obtained with high confidence for human [9], baker's yeast (*Saccharomyces cerevisiae*) [10], and as of recently, fission yeast (*Schizosaccharomyces pombe*) [11]. The BioGRID database [12] and IntAct database [13] are large databases aggregating such protein interactions curated from various high-throughput datasets and primary literature [14]. In addition, ongoing investigations into the molecular structures, or 3D shapes, of individual PPIs have so far yielded over 25 000 structures of complexes containing more than one protein. Those structures are currently accessible on the PDB database [15].

This wealth of available data now allows us to study the evolution of PPIs, a feat that was previously difficult due to a lack of high-quality experimental results. An investigation into the evolution of PPIs is essential to try and uncover the evolutionary design principles behind variations in PPIs, both within and between species. Moreover, such knowledge could in turn have practical applications to the identification of disease-specific patterns of PPIs which could serve as diagnostics biomarkers, help in the development of treatments and therapies targeting interactions that are functionally relevant to disease progression, as well as provide insights to the fields of synthetic biology, and genome engineering [7,8].

Here, we, therefore, take advantage of high-confidence PPI datasets made recently available for two yeast species, *Saccharomyces cerevisiae* (*S. cerevisiae*) [10], and

*Schizosaccharomyces pombe* (*S. pombe*) [11] and perform their thorough analysis using bioinformatics tools. More formally, we **hypothesize** that:

*Structural determinants influence the evolutionary rate of residues in protein-protein interactions* **and** *changes in those determinants for interfacial residues could be associated with the phylogenetic loss or gain of an interaction between two species.*

This study first focuses on creating an automated, custom pipeline to curate, preprocess, and build PPI molecular models from the above-mentioned data **(Aim 1)**. Those models are subsequently used to investigate the relationship between PPI structures and their evolution by studying the impact of various structural determinants on residue evolutionary rates in yeast **(Aim 2).** Finally, we compare PPIs that are preserved and PPIs that are different between *S. cerevisiae* and *S. pombe*, to identify possible drivers for differences in PPIs between the two species. The evolution of PPI interfaces is considered more specifically, as this region of contact between interacting proteins could be particularly important to PPI evolution **(Aim 3)**. Overall, this work yields great insight into the evolution of PPIs. Such knowledge could, in turn, be priceless to guide efforts in disease diagnosis and treatment, synthetic biology, or genome engineering.

# 2. Literature review

## 2.1 Importance of PPIs

### 2.1.1 Cellular function

Proteins are the main agents of biological function in cells. As such, the association of proteins with other proteins is one of the most common interactions in biology [1]. The term protein-protein interaction (PPI), therefore, refers to a variety of different types of interactions between proteins that are fundamental to virtually all biological processes within cells, playing critical roles in maintaining cellular structure, regulating metabolic pathways, and facilitating signal transduction. These interactions are also essential to the functionality of protein complexes that carry out various cellular activities [1–8]. Here we focus on physical PPIs, the molecular, physical contact between two or more proteins within a cell [3].

Physical PPIs can be transient. These interactions are typically short-lived, and are temporary and dynamic, allowing for rapid responses to changing cellular conditions, environmental cues and internal signals [6]. Transient PPIs play crucial roles in signal transduction pathways, where proteins must interact and dissociate quickly to propagate signals. For instance, ligand binding to a G-protein-coupled receptor (GPCR) triggers a conformational change, allowing a transient PPI between GPCR and a G-protein on the inner side of the plasma membrane. This transient interaction causes the G-protein to exchange GDP for GTP, activating it and enabling the transmission of signals to downstream effectors. The transient nature of this interaction ensures that signals are passed quickly and efficiently, enabling cells to respond promptly to external stimuli [16]. Transient PPIs also play a crucial role in the regulation of enzymes within cells. Transient PPIs between protein kinases, protein phosphatases, proteases and other enzymes with

their substrate proteins are crucial to the control and regulation of fundamental cellular pathways, including cell growth, cell cycle, metabolic pathways, and signal transduction [17]. For example, the phosphorylase kinase transiently interacts with glycogen phosphorylase, phosphorylating it and thus activating it to release glucose-1-phosphate from glycogen. This transient interaction ensures that glycogen breakdown is precisely regulated in response to cellular energy demands [18]. Transient PPIs also facilitate the rapid activation and deactivation of immune cells in response to pathogens. For instance, the transient interaction between T-cell receptor (TCR) proteins and major histocompatibility complex (MHC) proteins presenting antigenic peptides is crucial for T-cell activation. This interaction TCR and MHC proteins triggers a cascade of signaling events that lead to T-cell proliferation and differentiation. The transient nature of the interaction allows T-cells to quickly disengage from one antigen-presenting cell and interact with another, enhancing the immune response efficiency [19]. Overall, transient PPIs are fundamental to cellular responsiveness and adaptability, facilitating precise and dynamic control over a wide array of biological processes.

Physical PPIs can also be more stable. Stable PPIs are characterized by their long-lasting nature and play key roles in the formation and regulation of persistent multi-protein complexes that behave as molecular machines and carry out essential structural and functional roles within the cell [20]. Stable PPIs are crucial for providing structural integrity to cells and their organelles and form the basis of cellular architecture. For instance, the cytoskeleton is a network of protein filaments and tubules that provide structural support to the cell. Actin filaments, microtubules, and intermediate filaments are stabilized by PPIs, ensuring cellular shape, motility, and division [21]. Stable PPIs are also essential for maintaining the spatial organization of cellular components,

ensuring that biochemical processes occur at the right place and time, and facilitating the selective and efficient movement of molecules across cellular compartments. For example, the nuclear pore complex (NPC) is a stable assembly of nucleoporins proteins that regulates the transport of macromolecules between the nucleus and cytoplasm of cells. The stability of the NPC ensures selective and efficient transport, essential for cellular function [22]. Stable PPIs in regulatory complexes also ensure the precise control of critical cellular processes such as DNA replication, transcription, and cell cycle progression. For instance, the origin recognition complex (ORC) is a stable multi-protein assembly that binds to origins of DNA replication in eukaryotic cells. The ORC interacts stably with other replication initiation factors, to form the pre-replication complex, which is crucial for the initiation of DNA replication [23]. Overall, many major cellular processes are carried out by stable protein-protein complexes, which behave as molecular machines, composed of protein components and organized by tightly regulated PPIs to ensure proper function.

### 2.1.2 Mis-regulation and disruption in disease

Given the crucial and complex roles played by protein-protein interactions (PPIs) to ensure proper cellular function discussed above, the misregulation or disruption PPIs can profoundly impact cellular function. As such, misregulations and disruptions of PPIs have been associated with various diseases and disorders in human. Several disease-causing mutations are known to disrupt PPIs and single nucleotide polymorphisms associated with a number of diseases tend to occur in sites predicted to mediate interactions. Proper regulation of PPIs is, therefore, essential for maintaining normal cellular processes, and the study of PPIs is crucial to the study of various diseases.

Disruptions in PPIs have been associated with cancer initiation, progression, and treatment resistance. Many types of cancer exhibit dysregulated signaling pathways involved in cell growth, survival, and metastasis due to aberrant PPIs. For instance, in breast cancer, overexpression of HER2 leads to constitutive activation of downstream pathways by altering its interaction with other proteins like EGFR and Src kinases [24]. This disruption promotes uncontrolled cell division and aggressive cancer phenotypes. Mutations or overexpression of proteins involved in Ras-ERK signaling pathways are also known to disrupt normal PPIs, leading to uncontrolled proliferation and evasion of apoptosis in cancer cells [25]. Moreover, disruption of PPIs can lead to loss of tumor suppressor functions. For instance, mutations in the p53 protein, which normally interacts with MDM2 to regulate cell cycle and apoptosis, can disrupt these interactions and contribute to uncontrolled cell growth in various cancers [26]. Additionally, alterations in PPIs can confer resistance to cancer therapies. In chronic myeloid leukemia (CML), mutations in the BCR-ABL fusion protein alter its interactions with drug-binding sites, leading to resistance against tyrosine kinase inhibitors drugs [27]. This disruption reduces the effectiveness of targeted CML therapies designed to inhibit BCR-ABL signaling. Finally, disruptions in protein complexes critical for DNA repair, such as BRCA1 and BRCA2 interactions, can predispose individuals to breast and ovarian cancers [28]. Mutations in these proteins disrupt their interactions, compromising DNA repair mechanisms and increasing cancer susceptibility.

Disruptions in PPIs also play pivotal roles in the pathogenesis of neurodegenerative disorders, contributing to protein misfolding, aggregation, impaired cellular function, and ultimately neuronal degeneration and neuronal death. For instance, disrupted interactions between amyloid-beta and tau proteins in Alzheimer's disease leads to the formation of toxic aggregates,

such as neurofibrillary tangles and amyloid plaques, contributing to neuronal dysfunction and cell death [29]. Disrupted PPIs can also impair protein clearance mechanisms such as autophagy and the ubiquitin-proteasome system, exacerbating protein aggregation. In Parkinson's disease, mutations in α-synuclein proteins disrupt their interactions with chaperones, proteolytic systems and degradation pathways, leading to the accumulation of toxic protein aggregates [30]. Moreover, dysregulated PPIs can alter cellular signaling pathways critical for neuronal function and survival. For instance, disruptions in PPIs involving glutamate receptors and associated proteins contribute to excitotoxicity, which is neuronal death caused by overstimulation of glutamate transporters, in neurodegenerative disorders like amyotrophic lateral sclerosis (ALS) and Huntington's disease [31]. Finally, genetic mutations affecting PPIs can directly influence disease pathogenesis. In Huntington's disease, mutations in the huntingtin protein alter its interactions with cellular partners, disrupting processes such as vesicular transport and mitochondrial function, which contribute to neuronal degeneration [32].

Disruptions in PPIs can also affect the pathogenesis of infectious diseases, influencing pathogen virulence, host immune responses, and therapeutic resistance mechanisms. Pathogens can exploit disruptions in host cell PPIs to facilitate entry into host cells, manipulate host signaling pathways, evade immune surveillance, and promote disease progression. For example, viral proteins such as HIV-1 gp120 interact with host receptors CD4 and CCR5 to initiate viral entry by binding and altering receptor conformation [33]. Pathogens can also disrupt host immune responses by altering PPIs involved in immune signaling and evasion mechanisms. For instance, bacterial pathogens like *Yersinia spp.* inject effector proteins that disrupt PPIs in host immune signaling pathways, suppressing inflammatory responses and promoting pathogen survival [34]. Disruptions

in pathogen PPIs can also confer drug resistance and enhance virulence. In antibiotic-resistant bacteria, mutations in PPIs involved in drug-target interactions reduce antibiotic binding affinity, leading to treatment failure [35]. Finally, disruptions in host-pathogen PPIs can influence disease progression and severity. In malaria, interactions between plasmodium falciparum proteins and host erythrocyte receptors mediate parasite invasion and contribute to disease pathogenesis [36].

Dysregulations of PPIs that affect metabolic pathways, hormone regulation, and cellular signaling are also associated with several metabolic disorders. Disruptions in PPIs involved in insulin signaling pathways can lead to insulin resistance and impaired glucose metabolism. For example, in type 2 diabetes, alterations in PPIs between insulin receptor substrates proteins and downstream signaling molecules like PI3K/Akt disrupt insulin-mediated glucose uptake and metabolism [37]. Disruptions in PPIs can also affect lipid metabolism, contributing to dyslipidemia and cardiovascular risk. In familial hypercholesterolemia, mutations in LDL receptor PPIs impair receptor-mediated uptake of LDL cholesterol, leading to elevated blood cholesterol levels and increased cardiovascular disease risk [38]. Moreover, dysregulated PPIs can alter the mechanisms by which body energy status is sensed and have been linked to obesity. Leptin resistance, observed in obesity, involves disruptions in PPIs between leptin and its receptor, impairing signaling pathways that regulate appetite and energy expenditure [39]. Finally, disruptions in mitochondrial PPIs can impair oxidative phosphorylation and contribute to metabolic disorders. In mitochondrial diseases, mutations in proteins involved in electron transport chain complexes disrupt PPIs critical for ATP production, leading to energy deficiency and metabolic dysfunction [40].

Overall, PPIs are indispensable for the intricate network of biological processes that sustain cellular function and organismal health. From transient associations that enable rapid signaling, to stable complexes that maintain structural integrity, PPIs orchestrate essential functions within cells. Dysregulation or disruption of these interactions can play a pivotal role in the pathogenesis of numerous diseases. Understanding the mechanisms underlying PPI function and dysregulation, therefore, offers promising avenues for therapeutic interventions. Targeting specific PPIs involved in disease processes holds potential for developing novel treatments that restore normal interactions or inhibit aberrant ones. Advances in structural biology, proteomics, and computational modeling are enhancing our ability to identify and characterize critical PPIs, paving the way for precision medicine approaches tailored to intervene at the molecular level. Studying PPIs, thus, not only deepens our understanding of cellular biology but also offers hope for innovative strategies to combat a wide range of human diseases.

### 2.1.3 Applications to disease diagnosis and treatment, synthetic biology and genome engineering

Better understanding and studying the intricate networks of protein-protein interactions (PPIs) within cells offers new perspectives on how disruptions in these interactions contribute to disease pathogenesis and how they can be leveraged for therapeutic and technological innovations. This offers a wide range of current and future applications which extends beyond fundamental biology to various research fields including disease diagnosis, disease treatment, synthetic biology, and genome engineering.

PPIs can serve as potential biomarkers for disease detection and prognosis. Aberrant interactions or disrupted networks of interactions can indicate disease states or help predict treatment response, offering insights into disease mechanisms and guiding personalized medicine approaches. For instance, PPI based biomarkers have strong ability in distinguishing normal and disease samples in human cholangiocarcinoma dataset and diabetes dataset [41]. Studies on clinical samples have also shown success in diagnosing metastatic versus non-metastatic breast cancer tumors by overlaying a patient's expression profile onto the human protein-protein interaction map [42]. Investigating clusters in PPI networks for early onset colorectal cancer (CRC) patients uncovered five functional modules involved in the pathways of signal transduction, carcinogenesis and metastasis, that may serve as biomarkers of early onset CRC and have the potential to be targets for therapeutic intervention [43]. PPI network analysis also uncovered nine crucial proteins could form a candidate biomarker panel for esophageal adenocarcinoma, one of the most lethal cancers in the world with a very poor prognosis [44]. In Alzheimer's disease, altered PPIs involving amyloid-beta and tau proteins in cerebrospinal fluid are also being investigated as potential biomarkers for disease progression and response to treatment [45].

Moreover, understanding disease-associated PPIs facilitates the identification of novel therapeutic targets. Targeting specific PPIs implicated in disease pathology enables the development of more effective and precise therapies. Inhibitors targeting dysregulated PPIs in oncogenic signaling pathways have shown promise in preclinical and clinical studies, highlighting their potential as therapeutic interventions [46]. Several small molecule drugs targeting specific PPIs are approved by the FDA or in clinical studies for a wide range of diseases including chronic lymphocytic leukaemia [47], head and neck cancer [48], breast cancer [49], small-cell lung cancer [50] and

ovarian cancer [51]. In Alzheimer's disease, the interaction between beta-secretase and amyloid precursor protein is crucial for the production of amyloid-beta peptides. Inhibitors targeting this interaction aim to reduce amyloid-beta levels and slow disease progression and are currently being explored in clinical trials [52]. A class of drugs targeting the interaction between HIV-1 integrase and the host DNA, and therefore preventing the integration of viral DNA into the host genome and inhibiting viral replication has also been effective in the treatment of HIV [53]. Moreover, therapies targeting the interaction between SARS-CoV-2 spike protein and human ACE2 receptor are being developed to prevent viral entry and treat COVID-19 infections [54].

Insights from the study of PPIs also have crucial applications to the fields of synthetic biology and genome engineering. Synthetic biology harnesses knowledge of the principles of PPIs to engineer novel protein complexes or pathways with customized functions. These “designed” PPIs can be used to reprogram cellular behavior or create synthetic biomaterials with applications in medicine and biotechnology. For instance, synthetic PPIs have enabled the creation of inducible dimerization systems like the rapamycin-inducible FKBP12-FRB system, which can precisely control cellular functions and gene expression in response to specific stimuli, offering potential for targeted gene therapies [55]. Chimeric antigen receptors (CARs) are a new class of immunotherapy cancer drugs using synthetic PPIs to direct T-cells to recognize and kill cancer cells [56]. CARs have shown clinical benefit in patients by providing highly specific and effective treatment options. Synthetic transcription factors, such as zinc finger proteins, have also been successfully engineered to bind specific DNA sequences and recruit transcriptional regulators to allow precise control of gene expression, facilitating advances in gene therapy and functional genomics [57]. Moreover, PPIs have been utilized to create synthetic biomaterials. Engineered elastin-like polypeptides form

reversible hydrogels through specific PPIs and can be used in drug delivery systems and tissue engineering for medical applications [58]. Additionally, genome engineering tools such as CRISPR-Cas9 utilize PPIs to precisely edit DNA sequences. The efficiency and specificity of these tools depend on PPIs involving Cas proteins and guide RNAs, offering powerful capabilities for therapeutic genome editing and disease modeling [59]. Base editing, a technology which combine a catalytically impaired Cas9 with a deaminase enzyme, uses PPIs to introduce point mutations without double-strand breaks. This technology allows for precise nucleotide changes, offering potential for correcting genetic mutations in various diseases [60]. Prime editing involves engineered PPIs between a Cas9 nickase, a reverse transcriptase, and a pegRNA, enabling the introduction of targeted insertions, deletions, and base conversions with high precision. This tool expands the capabilities of genome editing for therapeutic applications [61].

Given the crucial roles that PPIs play in cellular function, the significant impact that their mis-regulation and disruption can have on disease, and the practical applications of PPI research in disease diagnosis, treatment, synthetic biology, and genome engineering, it is clear that research towards obtaining a more comprehensive understanding of PPIs is essential. Consequently, vast amounts of PPI data have been collected across various fields, driving advancements and applications in multiple scientific disciplines. In the following sections, we will summarize and describe the available data on PPIs and the experimental techniques utilized to collect PPI data in different experimental fields and at different experimental scales.

## 2.2 PPI data at the species scale

Understanding protein-protein interactions (PPIs) at the species scale involves the comprehensive mapping and analysis of a given species' interactome: the record of all PPIs which occur in an organism. This large-scale approach provides insights into the complex networks of interacting proteins that underpin cellular processes, disease mechanisms, and evolutionary biology. Here, we discuss various methods used to collect and analyze PPI interactome data at the species scale.

**Figure 1. Protein-protein interaction network (interactome network).** Graphical representation of a protein-protein interaction (PPI) network, or interactome network. Proteins in a species of interest are represented as circles. Lines illustrate PPIs that have been detected between the two connected proteins in the species.

### 2.2.1 Interactome data

Interactome data represents the entirety of protein-protein interactions (PPIs) in an organism and serves as a fundamental resource for understanding cellular functions and disease mechanisms. Accordingly, in recent years, numerous experimental techniques and computational predictions have been used to try and survey all proteins that interact in a given species, and thus

construct interactomes for various species. Interactomes, or records of all proteins that interact in a species, have been obtained with high confidence for human (*Homo sapiens*) [62], fruit fly (*Drosophila melanogaster*) [63], mouse (*Mus musculus*) [64], *Arabidopsis thaliana* [65], baker's yeast (*Saccharomyces cerevisiae*) [10], and as of recently, fission yeast (*Schizosaccharomyces pombe*) [11]. To construct these high quality interactomes, as well as to detect interacting proteins in many other species, several experimental systems and computational methods have been used and are further detailed below.

### 2.2.2 Yeast two-hybrid system (Y2H)

The Yeast Two-Hybrid (Y2H) system is a popular *in vivo* tool for detecting and studying protein-protein interactions (PPIs) in a high-throughput manner. The Y2H system utilizes the modular nature of transcription factors to detect interactions between proteins of interest (POIs) within the nucleus of yeast cells. It consists of two main components: the DNA-binding domain (DBD) and the activation domain (AD). The DBD is fused to a POI, acting as a bait. The AD is fused to another POI, serving as the prey [66].

Functionally, the Y2H system operates through the introduction of these bait and prey constructs into yeast cells. If the bait and prey proteins physically interact within the yeast nucleus, the DBD and AD domains come into proximity, reconstituting a functional transcription factor complex. This reconstituted complex then binds to specific DNA sequences upstream of reporter genes, such as lacZ for β-galactosidase or HIS3 for histidine biosynthesis, leading to the activation of these reporter genes and producing detectable phenotypic changes indicative of positive interactions [66].

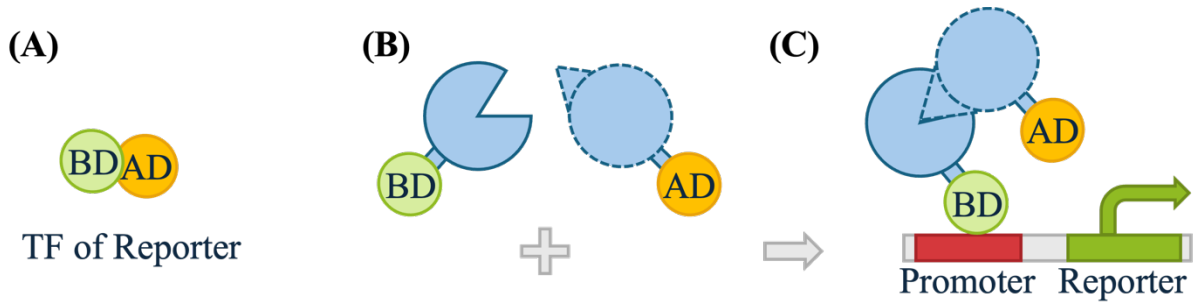



**Figure 2. Yeast two-hybrid system (Y2H).** Graphical representation of the yeast two hybrid system assay used to detect and study protein-protein interactions *in vivo*. **(A)** The transcription factor (TF) for a reporter gene. DNA-binding domain (BD) and activation domain (AD) are represented separately in green and yellow. **(B)** Cartoon diagram of a pair of proteins of interest shown in cross section, in blue. The bait protein is fused to the DNA-binding domain of the transcription factor (BD) and represented with a solid outline. The prey protein is fused to the activation domain of the transcription factor (AD) and represented with a dashed outline. **(C)** Cartoon diagram of the two proteins of interest physically interacting within the yeast nucleus, with AD and BD in close proximity forming a reconstituted functional transcription factor complex, activating transcription of the reporter gene.

The Y2H system's ability to screen a large number of potential interactions simultaneously makes it invaluable for generating comprehensive interactome maps. This high-throughput capability is particularly valuable in the initial stages of interactome mapping, where the goal is to identify as many interactions as possible to build a complete picture of cellular protein networks. Another advantage of the Y2H system is its relatively low cost and technical simplicity compared to other PPI detection methods, enabling many researchers to adopt and implement the system [67]. However, the method can falsely report proteins as interacting in the Y2H system, when they, in fact, do not interact in the species of interest (false positive interactions). Moreover, many true interactions may not be traced using Y2H assay or not have been investigated yet using the Y2H system (false negative interactions). In particular, only proteins localized to the nucleus of cells are appropriate to study using Y2H assays, since they are the only proteins able to activate reporter genes. Moreover, proteins that require post-translational modifications to carry out their functions are unlikely to behave or interact normally in a Y2H experiment where conditions for proper post-

translational modifications may not be met. Furthermore, if the proteins are not in their natural physiological environment, they may not fold properly to interact [68,69].

### 2.2.3 Tandem affinity purification combined with mass spectrometry (TAP-MS)

Tandem affinity purification-mass spectroscopy (TAP-MS) system is a widely used *in vitro* method for studying protein-protein interactions (PPIs) with high specificity and reliability under the intrinsic conditions of the cell. This technique leverages the affinity purification of protein complexes from cell lysates, followed by mass spectrometry analysis to identify the partner proteins interacting with a protein of interest (POI) in the complex [69].

The process begins by tagging a POI (bait) with a dual-affinity tag, typically consisting of a calmodulin-binding peptide and a streptavidin-binding peptide, then expressing the POI in the cell line of interest [70]. The cells are then lysed to release the protein complexes, and the lysate is passed through an affinity column that binds the first tag capturing the bait protein and its interacting partners while washing away other proteins. The captured complexes are then subjected to a second purification step using a column specific to the second tag. This secondary purification further refines the protein complex, significantly reducing background noise. Finally, the purified protein complexes are analyzed by mass spectrometry, which provides detailed information about the protein composition of the complex [71].

The TAP-MS system has several advantages, particularly its high specificity and responsivity as it can even detect weak protein interactions. The dual-step purification process also leads to high specificity of isolated protein complexes, thereby reducing non-specific binding and

background noise [70]. Additionally, mass spectrometry allows for the comprehensive identification of proteins within the complex, including low-abundance proteins and post-translational modifications, providing a detailed view of the protein composition of a PPI complex [71]. Another significant benefit of TAP-MS is that it can be performed under near-physiological conditions, preserving the native state of protein complexes and making the results more biologically relevant [71]. Despite these advantages, the TAP-MS system also has its drawbacks. False negatives can occur when some interactions are lost during the purification process due to the stringent washing steps. Moreover, false positives can arise, particularly if the bait protein is overexpressed, leading to non-specific interactions [72] .

### 2.2.4 Other experimental methods

Several other experimental methods have also been successfully used to detect protein-protein interactions (PPIs) both *in vivo* and *in vitro*. Each method having its own set of advantages and drawbacks, multiple techniques are typically used in order to crosscheck and verify results obtained.

Fluorescence Resonance Energy Transfer (FRET) is a *in vivo* method for studying PPIs in living cells. Proteins of interest (POIs) are tagged with donor and acceptor fluorophores, and interactions are detected based on the energy transfer between these fluorophores when they are in close proximity [73]. FRET, therefore, allows for real-time observation of PPIs in live cells, providing dynamic information about interaction kinetics and spatial localization. It is highly sensitive and can detect weak and transient interactions. FRET is also non-invasive, preserving the physiological conditions of the cellular environment. However, the efficiency of FRET depends

on the proper folding and orientation of the tagged POIs, which can affect the accuracy of interaction measurements. The method also requires sophisticated instrumentation and expertise in fluorescence microscopy. Moreover, the introduction of fluorescent tags may interfere with the native function and interaction of the POIs in cells [73].

Bimolecular Fluorescence Complementation (BiFC) is an *in vivo* method that visualizes PPIs by splitting a fluorescent protein (such as GFP) into two non-fluorescent fragments, each fused to a protein of interest (POI). When the POIs interact, the fragments come together to form a functional fluorescent protein, producing a detectable signal [74]. BiFC provides direct visualization of PPIs within their native cellular context, offering spatial information about where interactions occur. It is relatively simple and does not require complex instrumentation. BiFC is also highly specific, as fluorescence is only reconstituted upon protein interaction. However, false positives can still occur due to the high affinity of the fluorescent fragments. The irreversible nature of the fluorescent fusion to a POI can also prevent the study of dynamic interactions and lead to signal accumulation. Moreover, the large size of the fluorescent tags may interfere with the native function and interaction of the proteins in cells [75].

The split reporter assay is a similar, powerful *in vivo* technique used to study PPIs by employing a reporter system that consists of two inactive halves, each fused to a protein of interest (POI). Upon interaction of the POIs, the halves of the reporter protein (e.g., enzymes such as dihydrofolate reductase (DHFR), luciferase, or β-lactamase) reassemble to restore enzymatic activity, producing a detectable signal, such as fluorescence or chemiluminescence [76,77]. This method is versatile and can utilize different reporter systems, allowing for the detection of PPIs

with varying sensitivity and signal types. Unlike BiFC, which relies on fluorescence, split reporter assays can also measure enzyme activity, making them suitable for high-throughput applications. While split reporter assays provide valuable information about protein interactions in living cells, they also come with challenges. The size of the reporter fragments can affect the native function of the interacting proteins, and false positives may arise if the fragments are too small or have intrinsic affinity for each other. Additionally, the reassembly of the reporter can be irreversible, limiting the study of transient or dynamic interactions [76]. Nevertheless, split reporter assays offer a flexible and efficient tool for studying PPIs with high specificity and sensitivity in a variety of cellular environments [76,77].

Co-Immunoprecipitation (Co-IP) is a classical *in vitro* method for studying PPIs within their native cellular environment. This technique uses antibodies to capture a protein of interest (bait) along with its interacting partners (prey) from a cell lysate. The protein complexes are then analyzed using techniques such as Western blotting [78]. Co-IP is highly specific and preserves the native state of protein interactions. It can validate interactions identified by other methods and provides a physiologically relevant context for studying PPIs. However, the technique requires high-quality, specific antibodies, as non-specific binding can lead to false positives, and inefficient antibody binding or poor lysis conditions can result in false negatives. Therefore, Co-IP is not suitable for high-throughput screening [3].

Surface Plasmon Resonance (SPR) is an *in vitro*, label-free method for studying PPIs. This technique measures changes in the refractive index near a sensor surface to detect binding events between an immobilized protein of interest (bait) and an interacting partner (prey) in solution. This

technique offers high sensitivity and quantitative information on binding kinetics, affinities and interaction dynamics in real-time. Moreover, it is label-free, eliminating potential tag-induced artifacts [79]. However, immobilization of bait proteins on the sensor surface can alter their native conformation and affect binding properties. Moreover, SPR is typically limited to studying binary interactions and may not capture the complexity of multi-protein complexes [80].

### 2.2.5 Computational methods

Alongside experimental methods, computational approaches play a pivotal role in predicting, modeling, and analyzing protein-protein interactions (PPIs). These methods offer the advantage of high-throughput capabilities and can provide insights that complement experimental findings. Computational methods for studying PPIs can be broadly classified into *in silico* predictions based on sequence and structural data, as well as computational modeling and simulation techniques. Each approach has its unique strengths and challenges, contributing to a comprehensive understanding of PPIs.

Sequence-based prediction methods *in silico* utilize the amino acid sequences of proteins to predict potential interactions. Currently, the two main types of approaches used to predict PPIs from sequence data are similarity-based methods and machine learning-based methods [81]. Similarity-based methods such as PIPE4 [82]and SPRINT [83] score proteins based on the principle that if two query proteins resemble a pair of known interacting proteins, then evidence for an interaction between the query proteins can be inferred. Essentially, these methods quantify the strength of the interaction evidence under this assumption, using substitution matrices such as PAM120 or BLOSUM64 to measure the similarity between the query and interacting protein pairs

[81]. Machine learning-based methods "learn" to identify patterns or features that are commonly found in interacting proteins. To make predictions, these models examine query protein pairs for the presence or extent of these patterns. Depending on the approach, the predictors may learn from the physicochemical properties of protein sequences or simply from the amino acid composition. Recent models predominantly use the latter approach, leveraging advanced deep learning techniques to capture the "grammar" of protein interactions [84,85]. These sequence-based methods can handle large-scale datasets and are relatively fast. They do not require structural information, making them applicable to a wide range of proteins, including those with unknown structures [2]. However, the accuracy of predictions can be limited by the quality and size of the training datasets. False positives and false negatives are common, necessitating experimental validation. Moreover, these methods may not capture the full complexity of PPIs, such as those involving post-translational modifications [86].

Co-evolutionary analysis *in silico* leverages the concept that interacting proteins evolve in a coordinated manner. By analyzing correlated mutations in protein sequences across multiple species, co-evolutionary methods can predict potential PPIs and interaction interfaces. Co-evolutionary PPI inference techniques can be divided into two types: site-specific and full-sequence methods [87]. The site-specific method detects mutual sequence changes in binding interfaces of interacting partners to infer PPIs, but changes in such regions are hard to detect [88–90]. Full-sequence methods, such as the mirror-tree method compare a distance matrix between two proteins and use topological similarity of phylogenetic trees to predict PPIs [91]. These methods can predict PPIs without the need for structural information, making them applicable to a wide range of proteins. Moreover, they can identify evolutionary conserved interaction patterns, providing

insights into which PPIs are functionally important evolutionarily [92]. However, the accuracy of co-evolutionary predictions depends on the availability and quality of multiple sequence alignment data. Co-evolutionary signals can also be confounded by indirect interactions or phylogenetic relationships [93].

Structure-based prediction methods *in silico* use 3D protein structures to predict PPIs. These methods often involve docking simulations, where the physical and chemical properties of protein surfaces are analyzed to identify potential binding sites and interaction partners. Structure-based methods provide detailed insights into the molecular basis of interactions, including binding affinities and interaction interfaces. They can also identify specific residues critical for binding, facilitating the design of inhibitors or modulators [94]. However, high-resolution structures are required for accurate predictions, limiting the applicability to proteins with solved structures. Moreover, docking simulations can be computationally intensive and may not always accurately capture the dynamics of protein interactions. False positives can also occur due to the inherent flexibility of proteins [95].

Molecular Dynamics (MD) simulations can also be used to model the physical movements of atoms and molecules over time, providing dynamic insights into PPIs. By simulating the interactions between proteins in a virtual environment, MD simulations can reveal conformational changes, binding kinetics, and interaction stability. MD simulations offer a detailed, atom-level view of PPIs, capturing dynamic processes. They can provide insights into the energetics and mechanisms of interactions, helping to identify key residues and binding hotspots [96]. However, these simulations computationally demanding, requiring significant resources and time. The

accuracy of simulations also depends on the quality of the force fields used to model atomic interactions. Moreover, MD simulations are often not possible for very large protein complexes and not appropriate for exploring interactions over long timescales [97].

### 2.2.6 Databases for PPI data at the species scale

Comprehensive databases compiling protein-protein interaction (PPI) data at the species level, obtained via the various experimental and computation methods described above, are invaluable resources for PPI research, providing extensive information on experimentally validated and predicted interactions. These databases span various species, offering insights into the conserved and species-specific nature of PPIs. Here, we summarize some of the most prominent databases for PPI data at the species scale, highlighting their key features, strengths, and limitations.

The Biological General Repository for Interaction Datasets (BioGRID) is a comprehensive resource that catalogs PPIs across multiple species, including humans, yeast, and model organisms [12]. BioGRID compiles data from high-throughput experiments and individual studies. The database offers a wide coverage of species and interaction types, including physical, genetic and chemical interactions. It is frequently updated, ensuring access to the latest data. BioGRID is user-friendly, allowing easy search and download of interaction datasets. However, BioGRID relies on published data, which may introduce publication bias. The quality and reliability of the interactions can also vary, necessitating careful validation by users.

IntAct is a curated database of PPIs, primarily focusing on experimentally validated interactions [13]. It includes data from various species, with a strong emphasis on human interactions. IntAct provides detailed annotations for each interaction, including experimental conditions and interaction types. However, the focus on experimentally validated interactions may limit the number of interactions available compared to databases that include predicted data. IntAct's curation process can also lead to delays in data availability.

The Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) integrates known and predicted PPIs from various sources, including experimental data, computational prediction methods, and public text collections [98]. STRING covers a wide range of species, with an emphasis on integrating multiple lines of evidence. The database provides a confidence score for each interaction, helping users assess the reliability of the data. It also integrates interactive network visualization tools to facilitate the exploration of complex interaction networks. However, predicted interactions may include false positives, especially those derived from text mining and computational predictions. Users, therefore, need to interpret the confidence scores carefully and validate key interactions experimentally.

The Database of Interacting Proteins (DIP) focuses on experimentally determined PPIs, providing a curated dataset that includes interaction data for multiple species [99]. DIP emphasizes the quality and reliability of the interactions, offering a valuable resource for studying conserved and species-specific PPIs. DIP provides high-quality, experimentally validated interactions with rigorous curation standards. It offers tools for visualizing and analyzing PPI networks, facilitating hypothesis generation and experimental design. However, DIP's exclusive focus on

experimentally validated interactions may result in a smaller dataset compared to other databases that include predicted interactions. The update frequency may also be lower due to the curation process.

The Molecular INTeraction database (MINT) specializes in curated PPI data from experimental studies, with a focus on high-quality, manually annotated interactions [100]. MINT covers interactions from a variety of species, providing detailed information on interaction types and experimental methods. MINT offers detailed annotations and high-quality data, with a focus on experimentally validated interactions. It supports advanced search capabilities and integration with other PPI databases through the IMEx consortium. The focus on high-quality curation may limit the number of available interactions, and the update frequency may be lower compared to automated databases. Users may need to complement MINT data with other resources for comprehensive analyses.

In this study, PPI data at the species level were curated from the BioGRID [12] and IntAct [13] databases, which are currently the two most comprehensive resources for individual interactions. The BioGRID database contains nearly 1.49 million unique interactions derived from over 62,978 publications [12], while IntAct includes approximately 850,000 unique interactions from more than 23,462 publications [13]. Both databases offer extensive coverage of yeast interactions, further supporting their selection for this work. Additionally, the pairwise overlap of yeast PPIs between the databases used in this study and other publicly available PPI databases is relatively high [14], although incorporating data from additional databases could be a promising direction for future research. However, it is important to note that integrating data from multiple PPI databases is not

a straightforward task. While many databases provide interactions in a similar format, inconsistent or incorrect use of controlled vocabulary is common. Moreover, different gene and protein identifiers are used across databases, and sometimes even within a single database [14]. Finally, for this study, the primary limitation in PPI data was not at the species level, but rather at the molecular scale, as discussed in the following section.

### 2.3 PPI data at the molecular scale

Understanding protein-protein interactions (PPIs) at the molecular scale involves studying the detailed, specific molecular interactions between individual protein molecules within a cell. This encompasses studies of the 3D structure of interacting proteins and PPI complexes, the precise binding sites between individual proteins, PPI interfaces, and dynamic conformational changes that occur when proteins interact. This detailed, small-scale approach provides insight into the fundamental mechanisms of cellular functions, the structural basis of protein functions, the effects of mutations on interactions, and the mechanisms underlying disease states. Here, we discuss various methods used to collect and analyze PPI data at the molecular scale.

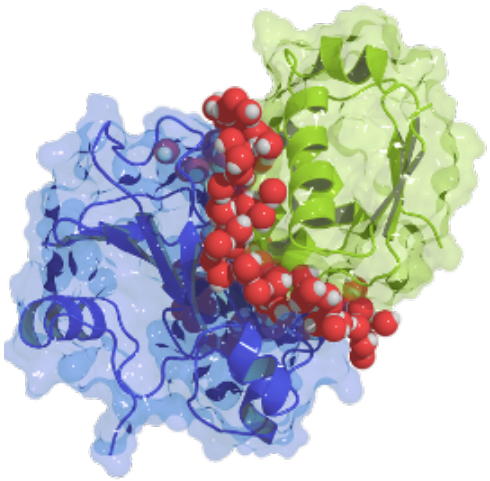


**Figure 3. Protein-protein interaction (PPI) structure.** Graphical representation of the molecular structure of a protein-protein interaction (PPI). Two interacting protein partners are illustrated in blue and green respectively. The interface of contact between the two interacting protein partners is illustrated in red.

### 2.3.1 PPI structure data

Protein-protein interaction (PPI) structure data (detailed, atom-resolution, three-dimensional descriptions of individual PPIs) provides detailed insights into the three-dimensional arrangements and atomic-level features of the interaction between individual proteins. At its core, PPI structure data elucidates how proteins physically interact with each other to perform biological functions essential for cellular processes. For instance, one of the first solved structure of a protein-protein complex, the barnase-barstar complex, revealed the precise binding interface between the polypeptide inhibitor barstar and its target the bacterial ribonuclease barnase [101]. The analysis of this solved structure uncovered small structural changes that can dramatically affect interaction specificity and affinity between the two proteins. Since then, numerous experimental techniques and computational predictions have been used to obtain structural details for a wide range of PPIs in numerous species and are further detailed below.

### 2.3.2 X-ray crystallography

X-ray crystallography is a popular method used to obtain high-resolution structural data for PPIs, revealing details of molecular arrangements within complex assemblies. This technique starts with the crystallization of purified protein complexes, where proteins are induced to form ordered arrays in a crystalline lattice. The crystals are then exposed to X-rays, which scatter off the electron clouds of atoms within the crystal. The resulting diffraction pattern are subsequently deciphered using mathematical algorithms to provide information about the spatial arrangement of atoms within the PPI and reconstruct the three-dimensional structure of the protein complex [102,103].

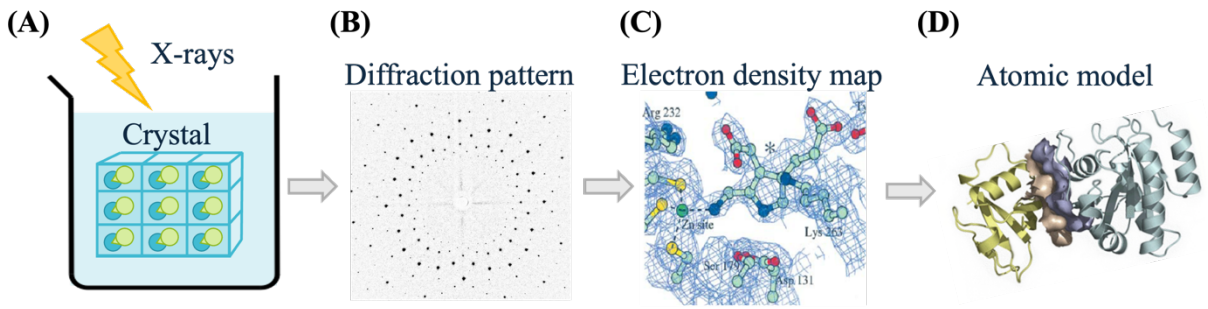



**Figure 4. X-ray crystallography.** Graphical representation of the x-ray crystallography process used to detect obtain high-resolution structural data for protein-protein interactions. **(A)** Interacting proteins are induced to form ordered arrays in a crystalline lattice exposed to X-rays. **(B)** X-rays scatter off the electron clouds of atoms within the crystal resulting in a diffraction pattern. **(C)** An electron density map is generated from the diffraction pattern. **(D)** The three-dimensional structure of the protein complex is reconstructed.

One of the major advantages of X-ray crystallography is its ability to resolve atomic-level details, typically achieving resolutions in the range of 1 to 3 angstroms (Å). This high level of resolution allows researchers to discern individual atoms, identify key binding interfaces between proteins, and understand the specific interactions that stabilize the complex [103]. However, the technique requires the production of large, homogeneous protein crystals, which can be challenging for some protein complexes and may limit its applicability. Additionally, the crystallization process itself may induce artifacts or alter protein conformations, potentially affecting the accuracy of the structural data obtained. Moreover, X-ray crystallography is generally unable to capture transient or flexible interactions that do not form stable crystals, limiting its utility for studying dynamic PPIs [102].

### 2.3.3 Nuclear Magnetic Resonance (NMR) spectroscopy

Nuclear Magnetic Resonance (NMR) spectroscopy is a method used to study protein-protein interactions (PPIs) in solution, providing valuable structural insights into their dynamic behavior. In NMR spectroscopy, proteins are studied in solution phase, allowing interactions to be

observed under near-physiological conditions. NMR detects the nuclear magnetic resonance of atomic nuclei in proteins, particularly hydrogen and carbon atoms, which emit signals that are influenced by their local chemical environment and interactions with neighboring atoms. By analyzing these signals, NMR can elucidate the spatial arrangement of atoms and infer the three-dimensional structure of protein complexes [104,105].

One of the major advantages of NMR spectroscopy is its ability to study protein dynamics and flexibility. Unlike X-ray crystallography, NMR does not require the formation of large protein crystals and can analyze proteins in solution, capturing transient interactions and conformational changes. This capability makes NMR particularly valuable for studying flexible regions within proteins or disordered proteins, which are challenging for other structural techniques [105]. However, NMR spectroscopy also has limitations. It is generally less sensitive than X-ray crystallography, requiring higher concentrations of purified proteins and longer data acquisition times. The interpretation of NMR data can also be complex, necessitating sophisticated computational methods for structure determination and validation. Additionally, NMR is limited in the size of proteins that can be studied effectively, typically up to 30-40 kDa, although advancements in technology have extended this limit [104,105].

### 2.3.4 Other experimental methods

Other experimental methods have also been successfully used to elucidate molecular details and three-dimensional (3D) structures of protein-protein interactions (PPIs). Each method has its own set of advantages and drawbacks, making them most appropriate for the study of different types of PPIs.

Cryo-Electron Microscopy (Cryo-EM) enables the visualization of large protein complexes at near-atomic resolution [106,107]. In Cryo-EM, proteins are flash-frozen in vitreous ice to preserve their native structure and imaged using electron microscopy. Advanced computational algorithms then reconstruct 3D maps of protein complexes and can reveal detailed architectures and conformational changes. Cryo-EM is particularly useful for studying flexible or transient interactions that are challenging for X-ray crystallography and NMR. However, Cryo-EM requires significant expertise, specialized equipment, and computational resources for data processing and analysis. Moreover, PPI complexes within samples may exhibit structural variability, making it challenging to obtain a uniform 3D structure [107].

Hydrogen-Deuterium Exchange Mass Spectrometry (HDX-MS) is a method for probing protein-protein interactions and dynamics at the atomic level [108]. In HDX-MS, proteins are exposed to deuterium oxide, and the exchange of hydrogen atoms with deuterium atoms in solvent-accessible regions is monitored. By comparing the exchange rates between free and complexed proteins, HDX-MS can map protein interaction interfaces and identify conformational changes upon binding. This technique provides valuable structural information on PPIs in solution without the need for crystallization, making it suitable for studying transient or weak interactions as well as for membrane proteins and large complexes. However, HDX-MS requires complex data analysis and bioinformatics tools for accurate data interpretation. Moreover, while the method provides valuable information on PPI dynamics, HDX-MS typically offers lower spatial resolution [109].

Cross-linking Mass Spectrometry (XL-MS) uses cross-linking reagents to covalently link amino acid residues that are in close proximity within protein or protein complexes. Cross-linked peptides are then analyzed by mass spectrometry to identify and characterize the cross-linked residues. XL-MS data provides distance constraints between cross-linked residues, allowing reconstruction of protein structures and mapping of interaction interfaces in protein complexes [110]. XL-MS is advantageous for studying large and dynamic complexes that are challenging for traditional structural methods. Moreover, the method identifies residues involved in interactions, aiding in the characterization of binding interfaces. However, interpretation of XL-MS data is complex and requires specialized software and expertise in bioinformatics for accurate interpretation. Non-specific cross-linking or background noise can also lead to false positives, requiring careful validation and control experiments [111].

### 2.3.5 Computational methods

Alongside experimental methods, computational approaches play a pivotal role in predicting, modeling, and analyzing protein-protein interactions (PPIs). These methods offer the advantage of high-throughput capabilities and can provide insights that complement experimental findings. Each approach has its unique strengths and challenges, contributing to a comprehensive understanding of PPI structural data.

Molecular docking is a computational technique that predicts the preferred orientation of one molecule when bound to another, thereby modeling the structure of their complex. This method is particularly useful for identifying potential binding sites and understanding the specificity and affinity of interactions. It operates by simulating the physical interactions between

the molecules and scoring them based on predicted binding affinity. Despite being high throughput and cost-effective, molecular docking has limitations in accuracy due to the simplifications made in modeling protein flexibility and the scoring functions used. Nevertheless, it remains a powerful tool for preliminary screening of large libraries of molecules against target proteins [94].

Molecular dynamics (MD) simulations offer another layer of detail by providing insights into the temporal evolution of protein structures and interactions. Unlike docking, MD simulations account for the dynamic nature of biomolecules, capturing their movements and conformational changes over time. This method involves solving Newton's equations of motion for the system, providing a trajectory that shows how the atoms in a protein or PPI move. MD simulations can reveal detailed information about the stability of protein complexes, the pathways of molecular interactions, and the effects of mutations on protein function. However, these simulations are computationally intensive and often limited by the timescales they can realistically cover [96].

Bioinformatics approaches leverage computational algorithms to predict and analyze PPIs based on existing sequence and structural data. These methods include co-evolution analysis, a technique which examines correlated mutations across protein sequences to infer interaction interfaces. The underlying principle is that interacting proteins evolve together, with mutations in one protein often compensated by mutations in its interacting partner to maintain the interaction. This analysis can reveal which residues are likely to be in contact in a PPI, providing clues about the structure of the protein complex [92]. However, the accuracy of co-evolutionary predictions depends on the availability and quality of multiple sequence alignment data. Co-evolutionary signals can also be confounded by indirect interactions or phylogenetic relationships [93].

Another approach involves using machine learning algorithms to predict the structural details of PPIs. These algorithms are trained on datasets of known protein complexes, learning to recognize features indicative of interaction interfaces. Features can include sequence motifs, physicochemical properties, and evolutionary conservation. Once trained, these models can predict the interaction surfaces of novel protein pairs. For instance, Wang et al. (2017) developed a deep learning framework that combines sequence and structural data to accurately predict protein interaction interfaces, demonstrating the potential of machine learning in structural PPI prediction [112]. However, these methods are still relatively new and require extensive testing and experimental validation.

Finally, homology modeling is a key bioinformatics tool for predicting the structure of protein-protein interactions. This method uses known structures of homologous proteins and PPIs as templates to model the structure of a protein complex without a solved structure. By aligning the sequences of the target proteins with those of known complexes, structural models for the PPI without a solved structure can be built [113]. Using homology modeling, researchers have been able to generate accurate models of protein complexes, providing insights into their functional mechanisms [114]. However, the accuracy of homology models is highly dependent on the availability of suitable templates. For proteins with low sequence similarity to known structures, the models may be unreliable.

### 2.3.6 Databases for PPI data at the molecular scale

Databases compiling large amounts of protein-protein interaction (PPI) data at the molecular level, obtained using the various experimental and computation methods described

above are valuable resources for PPI research, providing detailed, high-resolution data on the structure of PPIs, PPI interfaces, and structural dynamics at play in various interactions. Several key databases offer comprehensive structural information for a large number of protein complexes in a wide range of species, each with unique strengths and limitations.

The Protein Data Bank (PDB) is the most established repository for three-dimensional structural data of biomolecules, including protein complexes [15]. The PDB provides high-resolution structures obtained through methods like X-ray crystallography, NMR spectroscopy, and Cryo-EM. Over 25,000 structures of complexes containing more than one protein are currently available on the PDB. This extensive database offers detailed atomic coordinates, allowing researchers to visualize and analyze the interaction interfaces at a very high resolution. The major advantage of PDB is its comprehensive collection of high-quality, experimentally determined structures. However, its limitation lies in the static nature of the structures, which do not capture the dynamic aspects of protein interactions. Additionally, the PDB's coverage is limited to proteins and complexes that have been successfully crystallized or otherwise structurally resolved, leaving a gap for many transient and flexible interactions.

Databases like Interactome3D offer a more specialized resource by integrating structural details into interaction networks [115]. Interactome3D combines species-level interactome data with three-dimensional structural information, derived from PDB and other sources. It annotates interaction interfaces and provides models for complexes where experimental structures are not available. This combination of data allows researchers to not only confirm whether proteins interact but also to understand the structural context of these interactions. A significant advantage

of Interactome3D is its ability to provide structural models for interactions, thus filling gaps where direct experimental data may be lacking. However, its reliance on existing structural data can be a limitation, as not all protein interactions have corresponding structural information available.

Databases of experimentally determined thermodynamic data for PPI complexes are also valuable resources to the study of PPI data at the molecular scale [116]. These include the Protein–protein Interactions Thermodynamic Database (PINT), which records thermodynamic data on PPIs along with experimental conditions, sequence, structure and literature information [117]. The Protein–Protein Interaction Affinity Database also contains information on the binding affinity of complexes along with the structures of free proteins and complex [118]. The PDBbind database [119] and Structural database of Kinetics and Energetics of Mutant Protein Interactions (SKEMPI) [120] databases also record experimental binding affinity measures and thermodynamic data for protein complexes with known structure. These databases are particularly useful for researchers interested in the quantitative aspects of protein interactions. However, their scopes are limited by the availability of detailed binding affinity data, which can be challenging to obtain for all interactions.

The Database of Interacting Protein Structures (DIPS) provides detailed structural information about protein-protein interfaces [121]. It includes not only experimentally determined structures but also homology-modeled interactions, offering a broader coverage of the interactome. DIPS allows researchers to explore the geometric and physicochemical properties of interaction interfaces, facilitating studies on interface evolution, binding specificity, and the effects of mutations. A significant advantage of DIPS is its inclusion of modeled structures, which extends its applicability beyond experimentally resolved complexes. However, the accuracy of these

modeled structures can vary, depending on the quality of the template and the modeling techniques used.

In this study, PPI data at the molecular level were curated from the Protein Data Bank (PDB) [15], which is the most established and widely recognized repository for three-dimensional structural data of protein complexes.

## 2.4 Species of interest

With the large amount of protein-protein interaction (PPI) data available, both at the species scale and the molecular scale, we elected to focus our analysis on yeasts, the only group of species with two high-quality interactomes currently available [10,11]. More specifically we elected to study these two yeast species with available interactomes: baker's yeast, *Saccharomyces cerevisiae* (*S. cerevisiae*) and fission yeast, *Schizosaccharomyces pombe* (*S. pombe*). The large amounts of high-quality data in both species allows us to study the evolution of PPIs, and compare PPIs between both yeast species, a feat that is essential to try and uncover the evolutionary design principles behind variations in PPIs, both within and between species.

### 2.4.1 Baker's yeast

*Saccharomyces cerevisiae* (*S. cerevisiae*), commonly known as baker's yeast, is one of the most extensively studied eukaryotic model organisms. Its significance in biological research stems from its relatively simple eukaryotic structure, ease of genetic manipulation, and rapid growth rate. These attributes make *S. cerevisiae* an invaluable tool for investigating fundamental biological processes such as DNA replication, transcription, translation, and cell cycle regulation [122]. The

fully sequenced genome of *S. cerevisiae* and the availability of comprehensive genetic and genomic resources further enhance its utility in research [123]. *S. cerevisiae* can serve as a model for understanding various eukaryotic processes, including human diseases. Many human genes and their corresponding pathways have functional homologs in baker's yeast, enabling the study of proteins and their interactions in a simpler context [124]. For instance, *S. cerevisiae* has been crucial in elucidating the regulation mechanisms of the eukaryotic cell cycle, including the identification of cyclins and cyclin-dependent kinases which are conserved across eukaryotes [125,126]. Research in baker's yeast also uncovered key components of various signaling pathways, such as the MAPK/ERK pathway, which is critical for cell growth and differentiation [127]. Moreover, studies in *S. cerevisiae* have led to a deeper understanding of homologous recombination and DNA repair processes, with significant implications for cancer research [128,129]. Additionally, *S. cerevisiae* has provided insights into chromatin remodeling, histone modifications, and gene silencing mechanisms [130]. Baker's yeast has also been pivotal in studying protein folding and degradation including the ubiquitin-proteasome system and chaperone-mediated protein folding, processes essential for cellular homeostasis [131,132].

In the context of the study of protein-protein interactions (PPIs), *S. cerevisiae* is commonly used as a model organism for a wide range of applications. Signal transduction pathways are crucial for cells to respond to their environment, and the study of these pathways in *S. cerevisiae* has greatly furthered our understanding of how signals are transmitted and regulated by PPIs. For instance, the high-osmolarity glycerol (HOG) pathway in baker's yeast, which responds to osmotic stress, involves a series of PPIs that activate the Hog1 MAPK, which in turn regulates gene expression to adapt to high osmolarity conditions [133]. Studies of the HOG pathway have provided

insights into osmoregulation mechanisms that are applicable to higher eukaryotes. The regulation of the cell cycle in *S. cerevisiae* has also been pivotal in understanding cell division processes in eukaryotes. For example, studies of the binding of the cyclin Cln2 to the CDK Cdc28, an interaction that is essential for the transition from the G1 to the S phase in yeast, helped elucidate how similar interactions regulate the cell cycle in higher eukaryotes [134]. Baker's yeast research has also highlighted the role of checkpoint proteins that monitor DNA integrity and ensure that damaged DNA is repaired before cell cycle progression continues. The *S. cerevisiae* Rad9 protein, for example, interacts with other proteins to halt the cell cycle in response to DNA damage, a mechanism that is conserved in humans [135]. Moreover, high-throughput PPI studies in baker's yeast have greatly enhanced the functional annotation of a wide range of proteins. Comprehensive interactome mapping projects, such as the yeast two-hybrid screens conducted by Gavin et al. (2002) or Yu et al. (2008), have identified thousands of PPIs in yeast [10,71]. These large-scale studies have not only provided a detailed map of protein interactions but also facilitated the prediction of protein functions based on their interaction partners. Moreover, functional genomics in baker's yeast has been applied to study human disease genes. For instance, the yeast ortholog of the human PARK9 gene, implicated in Parkinson's disease, interacts with several proteins involved in metal ion homeostasis. Studying these interactions in yeast has helped elucidate the molecular mechanisms underlying Parkinson's disease [136].

### 2.4.2 Fission yeast

*Schizosaccharomyces pombe* (*S. pombe*), or fission yeast, has emerged as a powerful model organism in biological research more recently. Similarly to *S. cerevisiae*, *S. pombe* offers advantages such as a well-characterized genome, ease of genetic manipulation, and conserved

cellular machinery, making it invaluable for investigating cellular processes like DNA replication, transcription, translation, and cell cycle regulation [137,138]. Research into these processed uncovered some of their key components, providing insights that are applicable to higher eukaryotes, including humans. For instance, research in *S. pombe* helped elucidate crucial aspects of cell cycle control in eukaryotes. The identification of core regulators like cyclins and cyclin-dependent kinases (CDKs), which govern progression through cell cycle phases in the species, has been pivotal [139,140]. Studies on DNA damage response pathways, mediated by proteins like Rad3 and Chk1, have highlighted mechanisms of genome stability maintenance conserved across eukaryotes [141,142]. Additionally, *S. pombe* has contributed significantly to our understanding of chromosome structure and dynamics, including telomere maintenance and centromere function, processes critical for genome stability [143,144]. Studies in *S. pombe* have also elucidated the components and mechanisms of the RNA interference (RNAi) pathway, which regulates gene expression through small RNA molecules. This pathway is conserved across eukaryotes and plays essential roles in gene silencing and genome stability, contributing to our understanding of epigenetic regulation and RNAi-based therapeutic approaches [145].

In the context of the study of PPIs, *S. pombe* serves as an excellent model organism for studying signal transduction pathways and regulatory networks. For instance, studies of the Target of Rapamycin pathway, a pathway that regulates growth and metabolism in response to nutrient availability, in fission yeast have advanced our knowledge of how nutrient signaling pathways integrate with PPI networks to regulate cellular responses in higher eukaryotes [146]. Research on the components of the spindle pole body in *S. pombe* also helped further our understanding of how protein interactions govern mitotic processes including spindle formation and chromosome

segregation, crucial events for proper cell division [147]. Moreover, studies in *S. pombe* have provided foundational knowledge about mechanisms that regulate telomere length and function across species, informing our understanding of telomere maintenance mechanisms relevant to human aging and cancer [148,149].

### 2.4.3 Comparison of the two model organisms

*Saccharomyces cerevisiae* (*S. cerevisiae*) and *Schizosaccharomyces pombe* (*S. pombe*) are prominent model organisms in biological research, each offering distinct advantages rooted in their evolutionary history and genetic characteristics. *S. cerevisiae* and *S. pombe* belong to different fungal clades and diverged from a common ancestor approximately 500 million years ago [142,150]. Their genomes have since undergone significant changes, including gene duplications, deletions, and rearrangements [151]. In particular, *S. cerevisiae* has undergone a whole genome duplication during its evolutionary history, and only a small fraction of the genes was subsequently retained in duplicates while most were deleted. Gene order was then rearranged by many reciprocal translocations between chromosomes [152]. In contrast, comparisons of chromosomal sequences and searches for conserved gene did not reveal evidence for large-scale genome duplications in *S. pombe* [137].

*S. cerevisiae* is known for its ease of genetic manipulation, the species exhibits haploid and diploid phases, facilitating genetic screens, knockout studies, and high-throughput assays. Its rapid growth and ability to ferment sugars make it valuable for industrial applications and metabolic engineering [153]. While also genetically tractable, *S. pombe* offers a different set of experimental advantages. The two yeasts show major differences in cell cycles, therefore, the study and

identification of similarities in underlying control mechanisms have major implications for other eukaryotes such as humans [154,155]. While *S. cerevisiae* and *S. pombe* share conserved pathways, there are notable differences in response mechanisms to DNA damage and oxidative stress between the two species, reflecting divergent evolutionary adaptations [11]. Moreover, *S. pombe* is an important model organism for studying fundamental biological processes such as RNA splicing, cell-cycle regulation, RNAi, and centromeric maintenance, which are conserved in metazoans but divergent in budding yeast [11,137]. For instance, research on RNAi and epigenetic in *S. pombe* has not only contributed to our understanding of its mechanisms and consequences but also helps to explain biological differences between fission and budding yeasts [155]. The regulation of centromeric silencing also is a well-conserved process in *S. pombe* and metazoans but is divergent in *S. cerevisiae* [156]. Research on centromeric silencing in fission yeast, therefore, helped uncover some of the key components of the RNA-induced transcriptional silencing complex [11].

In the context of the study of PPIs, it is estimated that only ~40% of *S. pombe* interactions are conserved in *S. cerevisiae*, while ~65% of *S. pombe* interactions are conserved in human [11]. This is despite the two yeasts species sharing a greater fraction of protein-coding genes than either yeast does with human. This suggests that a large fraction of interactions is conserved between human and *S. pombe* but have been lost specifically in the *S. cerevisiae* lineage. In particular, PPIs involved in biological processes such as chromosomal organization, chromosome segregation, and cell cycle are far better conserved between *S. pombe* and human than in *S. cerevisiae*, and accordingly *S. pombe* has been used as a model organism for studying these processes [137]. This is highly relevant to the use of both *S. cerevisiae* and *S. pombe* as model organisms, as they appear to be complementary, with some biological functions that can be better studied using fission yeast

and reciprocally. Initial comparisons of PPIs between *S. cerevisiae* and *S. pombe* also found evidence of co-evolution in a large fractions of interactions that are preserved between the two yeasts, with major implications for studies reliant on the expression of human proteins in model organisms to identify functional mechanisms [11].

Overall, *S. cerevisiae* and *S. pombe* complement each other as model organisms, each offering unique strengths rooted in their evolutionary history and genetic characteristics. Their distinct genomic structures, evolutionary divergence, and experimental advantages make them indispensable tools for studying a wide range of biological processes and disease mechanisms. Moreover, the large amounts of PPI data available at the species scale and the molecular scale, in both species allows us to compare PPIs, both within and between the two yeasts and study the evolution of PPIs, a feat that is essential to try and uncover the evolutionary design principles behind variations in PPIs, both within and between species.

## 2.5 Structure-evolution relationship within a species

Protein evolution refers to the process by which protein sequences change over time through genetic mutations, leading to the development of new protein functions. Understanding the nature of constraints on protein evolution has long been the focus of much scientific interest. In particular, the three-dimensional structure of a protein is known to play an important role in constraining protein evolution, with different sites, even within the same protein often having very different evolutionary rates [157–159]. Studying the relationship between protein structure and protein evolution has, therefore, been a fundamental aspect of molecular biology, with profound implications for understanding cellular functions and the mechanisms underlying various

biological processes [160]. However, as previously discussed, proteins do not act in isolation in cells, but rather act via protein-protein interactions (PPIs). Yet very little is known about how PPI structure can shape evolutionary dynamics and influence the evolution of single proteins, PPIs, and PPI interfaces. With the available data on PPIs described above for yeast, we can now study structural constraints on PPI evolution, and thus better understand the evolutionary design principles behind variations in PPIs within a species. Moreover, such knowledge on natural, evolutionary, variations in PPIs within a species is crucial when investigating dysregulation or disruption of PPIs associated with disease, as well as provides insights to the fields of synthetic biology, and genome engineering [7,8]. Here, we, therefore, first describe current methods to estimate evolutionary rates for individual residue sites within a protein. We then discuss molecular traits and structural properties known to influence residue evolutionary rates in the single protein literature. Finally, we review previous works investigating the structure-evolution relationship of PPIs more specifically.

### 2.5.1 Site-specific evolutionary rates

Evolutionary rates (the speed at which genetic or protein changes accumulate in a species over time) are known to vary widely between proteins. Genes encoding highly expressed proteins, proteins that carry out crucial functions, or proteins that interact with many partner proteins tend to be more conserved (i.e. evolve more slowly) [161–163]. For instance, in the genome of the model organism *Saccharomyces cerevisiae* (baker's yeast), evolutionary rates among the roughly 6,000 genes are spread out over three orders of magnitude [164]. However, in addition to this gene-wide variation, and perhaps more interestingly, evolutionary rates can vary significantly among different residue sites, even within the same protein. For example, sites in the core of most proteins

or regions involved in enzymatic activity typically evolve more slowly than other sites [159]. Moreover, sites associated with ligand binding activity are also known to be more evolutionarily conserved [165].

Estimating site-specific evolutionary rates is a non-trivial task and, thus, various methods have been proposed for this inference in the literature [163,166,167]. However, broadly speaking, methods to estimate site-specific evolutionary rates compute rates of substitution at each individual site in a protein using two pieces of data: a multiple sequence alignment and a corresponding phylogeny. The two primary approaches for this estimation differ in the multiple sequence alignment data used: codon data, or amino acid data.

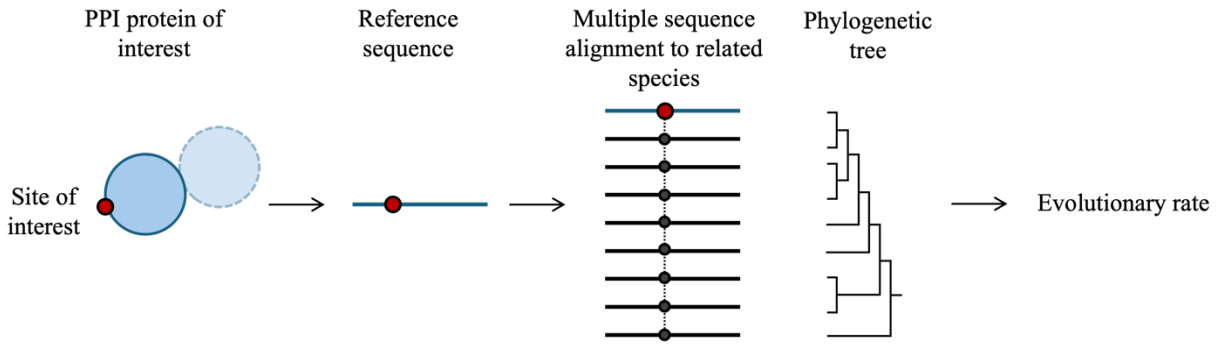



**Figure 5. Site specific evolutionary rate calculation.** Graphical representation of the general principle behind site specific evolutionary rate calculations. The protein of interest to the calculation is illustrated in blue along with its interaction partner differentiated with a dashed outline. The protein site of interest to the calculation is represented in red.

In the context of protein-coding sequences, evolutionary rates are typically estimated from codon data by calculating the ratio $\omega = dN/dS$, where $dN$ is the evolutionary rate of non-synonymous substitutions (the rate at which non-synonymous substitution mutations are fixed per unit of evolutionary time) and $dS$, is the evolutionary rate of synonymous substitutions (the rate at

which synonymous substitution mutations are fixed per unit of evolutionary time). To make *dN* and *dS* directly comparable, they are typically normalized to account for the higher likelihood that a random mutation is non-synonymous rather than synonymous [166]. Older *dN/dS* inference methods calculate *dN/dS* simply by counting the observed changes either between pairs of sequences or along a phylogenetic tree [168,169]. Most current inference approaches expand on this idea by using a Markov model of sequence evolution to infer evolutionary rate parameters, typically in a maximum-likelihood framework [170]. For instance, tools like the PAML (Phylogenetic Analysis by Maximum Likelihood) software use codon substitution models that incorporate factors such as transition/transversion rate bias and codon frequency to estimate *dN* and *dS* rates [171]. PAML then computes the likelihood of the observed codon sequence data given a phylogenetic tree and estimates the site-specific *dN/dS* ratios by maximizing this likelihood. The *dN/dS* ratios obtained provide insights into the selective pressures acting on individual codons. A *dN/dS* ratio less than 1 indicates purifying selection, where non-synonymous mutations are deleterious and thus selected against. A ratio greater than 1 suggests positive selection, where non-synonymous mutations are advantageous and spread throughout the population.

The primary approach for inferring rates from amino acid data is implemented in the Rate4Site algorithm [172]. The rate4site algorithm works by first constructing a multiple sequence alignment of homologous proteins and a corresponding phylogenetic tree. It then calculates the likelihood of the data under different models of rate variation among sites, assigning a per-site rate-scaling factor that indicates how rapidly each residue evolves relative to the average evolutionary rate for the full protein. Rate4site is implemented using a Bayesian framework with a random-effects approach, assuming that evolutionary rate at each site follows a gamma

distribution while allowing for variability in rates across sites [173]. The relative rates are then normalized such that the average rate across all sites is 1, with lower rates indicating higher evolutionary conservation and higher rates indicating more variable positions [172]. Tools like ConSurf further extend this approach by mapping these rates onto the protein's three-dimensional structure, facilitating the identification of functional sites, such as active sites or binding interfaces, that are evolutionarily conserved [174]. Finally, the ConSurf-DB database includes pre-calculated estimates of the evolutionary rates for a large number of proteins of known structure obtained using the Rate4Site algorithm [175].

Overall, the estimation of site-specific evolutionary rates is a powerful tool for understanding the selective pressures and evolutionary dynamics that shape protein sequences. By analyzing either codon or amino acid data, researchers can identify conserved and variable regions, shedding light on the structural and functional importance of different residues. Moreover, conclusions obtained using the two different techniques discussed here are typically correlated [176]. These insights are crucial for advancing our knowledge of protein evolution, with implications for drug design, synthetic biology, and understanding the molecular basis of diseases. As computational methods continue to evolve, they will help provide more precise and detailed pictures of how proteins adapt and function over evolutionary time.

### 2.5.2 Structural constraints on evolutionary rates

Variations in evolutionary rates among different protein sites measured and discussed above, are driven, to a large extent, by the requirement that proteins fold properly and stably into their required, active conformation, enabling them to interact with protein partners and perform

important cellular roles. Purifying selection, therefore, ensures that sites at which mutations would disrupt folding, stability or interaction are the most conserved. Understanding the precise nature of these constraints and selective forces on sequence evolution has, therefore, long been the focus of much scientific interest [164,167,177].

Amongst all possible constraints on protein evolution, constraints imposed by three-dimensional (3D) structure are of particular interest, as they connect protein evolution with fundamental biophysical principles [157,160]. Moreover, features directly connected to protein structure have previously been estimated to explain roughly 10% of the variation in protein evolutionary rate [178]. The highest possible resolution for investigating this structure-evolution relationship is to correlate structural properties of the micro-environment surrounding individual residues with site-specific residue evolutionary rates in order to identify factors influencing evolution at this most basic level of fixation and elimination of single amino acid residue mutations. These forces are interesting in their own right, and their summed effects on protein evolutionary rates contribute to our understanding of system-level evolutionary phenomena.

One of the oldest observations linking protein structure to evolution involved the influence of solvent exposure or burial on residue mutations. The homologous proteins hemoglobin and myoglobin were observed to differ far more dramatically on their solvent-exposed surfaces, than in their cores that are largely buried from solvent [179]. Since this initial observation solvent accessibility (accessible surface area (ASA), or solvent accessible surface area (SASA)), the surface area of a given residue that is accessible to the external environment, has become a well-known structural correlate of molecular evolution. ASA and SASA values are commonly

normalized by the largest possible solvent accessibility for a given amino acid type, resulting in relative solvent accessibility (RSA), a relative measure ranging from 0 for completely buried residues to 1 for completely exposed residues [180]. Numerous works have studied the relationship between RSA and residue evolutionary rates using large sequence and structure datasets, and universally support the notion that buried residues (for instance residues buried in a protein's core) are under tighter constraint, and therefore evolve more slowly [160,181–186]. Moreover, several studies established RSA as the dominant structural constraint on residue evolutionary rate. For example, when investigating the evolutionary rates of residues in several different types of secondary structures including helices, sheets and coils, exposed sites were found to evolve more rapidly than buried ones, regardless of secondary structure [182]. Other works modeling selective constraints for site-specific evolutionary rate predictions found that properties such as secondary structure and H-bonding information [187], or amino acid hydrophobicity and size [188] had little influence on site rates beyond the strong effect of RSA. Finally, more recent works established that the relationship between solvent exposure and selective constraint is strong, positive, and linear across its parameter range [160,189,190].

Although solvent exposure gets the most attention as a structural driving force in protein evolution, it is not the only property to be studied in this context. In recent years packing density, or the degree to which residues are surrounded by other atoms, has been investigated as a structural correlate for residue evolutionary rate. Similarly to how the extent to which a given residue comes into contact with solvent can be quantified using RSA, we can quantify the extent to which a given residue comes into contact with other residues in a protein. This idea, known as packing density or contact density, represents how densely packed a residue is within a protein structure. Two

measures are most commonly used to estimate packing density for a given residue: the contact number (CN) and the weighted contact number (WCN). For a given amino acid residue, CN simply counts the number of other residues within a local, structural neighbourhood. WCN instead considers all residues in a proteins and weighs them by the square of their inverse distance to the amino acid of interest to the calculation [191,192]. Previous works found a small but significant correlation between CN and site-specific evolutionary rates after controlling for RSA, establishing that CN influences evolutionary rates independently of RSA [160]. Additional studies using WCN instead of CN to estimate packing density found much stronger correlations with evolutionary rates [191,193]. Whether solvent accessibility or packing density is a more dominant determinant of site-specific evolutionary rates remains highly debated in the literature, as both their relative performance, and the overall performance of either predictor can vary widely between different protein structures [166].

The type of secondary structure in which a residue is located can also significantly influence its evolutionary rate. For instance, work on mammalian proteins showed that residues in helices and strands evolve more slowly than those in the less ordered loops and turns [184]. Increased conservation is also observed in immunoglobulin domains, where maintaining the hydrogen-bonding network is crucial for the structural integrity and proper function of the protein [194]. Other works investigating the occurrence of different structure elements in conserved sequence regions found that stands are often located in highly conserved regions of proteins, highlighting that these structural elements could be crucial to a protein's overall architecture and function, and resulting in lower evolutionary rates for residues within these regions [195]. In contrast, residues in loop regions tend to evolve more rapidly. Loops often connect secondary structural elements and are

not directly involved in maintaining the core structure of the protein. This flexibility allows for greater sequence variability, as seen in surface loops of enzymes that can tolerate changes without significantly affecting the overall protein function. The variability of loops can be crucial for adapting to new functions or interactions, illustrating a balance between structural conservation and evolutionary adaptability [196]. Moreover, disordered regions of proteins are generally known to evolve more rapidly than their ordered counterparts [197].

Finally, properties related to the dynamics and flexibility of a protein are also thought to influence site-specific evolutionary rates. Proteins are not static structures, but instead undergo constant conformational changes that are often critical to protein function. For instance, enzymes undergo conformational shifts to expose active sites for substrate binding. As such sites in highly flexible regions of proteins can be more tolerant to mutations than sites in less flexible regions [198,199]. The overall flexibility of a site can be estimated using measures such as mean square fluctuations or B-factors, both measures of the extent to which a given residue changes its position over time. Several works have found correlations between these measures of local flexibility and site-specific evolutionary rates, where rigid sites are more conserved and evolve more slowly than flexible sites [198,200]. However, whether flexibility is a structural determinant of residue evolutionary rates in its own right or correlates with evolutionary rates simply because it is also correlated with other structural determinants such as RSA or packing density remains debated.

Overall, understanding the relationship between protein structure and evolution provides crucial insights into the fundamental mechanisms of biological function and adaptation. Current literature highlights that structural properties play a significant role in determining the evolutionary

rates of residues in single proteins. However, proteins rarely act alone in cells, and instead tend to be involved in protein-protein interactions (PPIs). Yet very little is known about how PPI structure can shape evolutionary dynamics and influence the evolution of single proteins, PPIs, and PPI interfaces. Studying structural constraints on PPI evolution, could, therefore, help us better understand the evolutionary design principles behind variations in PPIs within a species, or enhances our ability to predict and manipulate PPIs, with implications for drug design, synthetic biology, and understanding the molecular basis of diseases. Continued research in this field promises to uncover further intricacies of the structure-evolution relationship, advancing our knowledge of protein and PPI and evolution.

### 2.5.3 Review of work in PPIs

In more recent years, studies of the relationship between structure and evolution have begun to shift from simply considering single proteins in isolation, to taking protein-protein interactions (PPIs) into account. Some proteins in cells are never found as free-floating stable monomers in vivo, and instead form obligate PPI complexes [20]. Considering the structure of the entire PPI, rather than single protein structures when studying evolution for proteins involved in such obligate PPIs could, therefore, be crucial. Moreover, even if involved in stable (although not obligate) complexes, or more transient complexes, most proteins in cells perform crucial functions via interactions with partner proteins. As new tools become available and become more sensitive, a shift in protein studies, from the single structure level to the PPI level is, therefore, underway, bringing the science of proteins from the primary, secondary and tertiary structural level to the quaternary level [201].

One of the key structural properties studied in the context of PPIs, is the region of contact between interacting partners, also referred to as the interface. The interface is a key PPI feature, both structurally and functionally, as mutations to interfacial residues can lead to altered protein-protein binding affinities and improper PPI function, which can have implications for organismal fitness, and health [176,202]. Interfacial residues are considered exposed surface residues in most structure-based studies that treat the two protein partners of a PPI as free-floating. However, the structure of the micro-environment surrounding interfacial residues can be drastically altered upon formation of a PPI complex, when they become buried in the interface of contact between protein partners. Previous works uncovered that interfacial residues tend to be more evolutionarily conserved than non-interfacial residues, establishing the critical role that they play in mediating the interactions between protein partners and ensuring proper PPI formation and function [203–206]. Additional works studying evolutionary rates for interfacial residues in transient and obligate PPIs found that residues at the interfaces of obligate complexes evolve more slowly than those in transient complexes, highlighting the importance of interface conservation for stable protein complexes [207].

Solvent accessibility, a known correlate of evolutionary rates for single proteins, is another crucial factor influencing evolutionary rates of residues in PPIs. Generally, buried residues, which are less exposed to solvent, evolve more slowly than exposed residues in single proteins. This trend is also observed in protein interfaces, where buried interfacial residues show higher conservation than solvent-exposed interfacial residues. For instance, studies of six homodimers families uncovered that interface residues, particularly those completely buried in the interface, were more conserved than other surface-exposed residues, suggesting that the structural and

functional integrity of the interface is maintained by conserving buried residues [203]. Moreover, sub-divining interfaces into different regions, including a core, a rim, and support regions based on solvent accessibility, helped uncover that core residues are typically more evolutionarily conserved than other interfacial residues [208,209]. Finally, interface involvement, the change in solvent accessibility of a residue upon complex formation (as its parent protein transitions from a free-floating to a co-complexed state), was also found to independently constrain residue evolutionary rates in yeast: if two residues are similarly buried from a solvent accessibility standpoint, but one resides in a protein–protein interface, then it will evolve much more slowly [160].

Binding energy, which quantifies the strength of interaction between proteins, can also plays a significant role in the evolution of PPIs. Residues that contribute significantly to binding energy, often referred to as "hot spots" are typically highly conserved. This conservation is likely due to the crucial role these residues play in maintaining high-affinity interactions necessary for the stability and functionality of protein complexes [210]. Hot spot residues often form hydrogen bonds, salt bridges, and hydrophobic contacts that are essential for binding. The evolutionary pressure to maintain these interactions results in lower substitution rates for these residues, as even minor changes can drastically affect the binding affinity and, consequently, the biological function of the complex [211].

Overall, insight from the study of PPIs have been valuable in improving our understanding of the structure-evolution relationship, uncovering various structural properties, including interface location, solvent accessibility, and binding energy that influence residue evolutionary rates. Understanding these relationships is crucial for unraveling the molecular mechanisms of

protein interactions and provides insights into the evolutionary pressures that shape the PPI landscape. However, the works discussed here typically only consider a single structural constraint on the evolution of interfacial residue at a time. Moreover, these previous studies typically rely upon coarse distinctions between "interfacial" and "non-interfacial", or "core" and "rim" residues. Consequently, our knowledge of any continuous structure-based evolutionary constraints on PPI and interfacial residues remains limited at the proteome scale. A systematic comparison, integrating a comprehensive list of different structural features in terms of their impact on residue conservation therefore remains needed in order to settle long-standing debates over which structural features are most important at constraining residue evolution in PPIs and PPI interfaces. Addressing this blind spot in current literature by systematically studying and comparing structural determinants that influence the evolutionary rates of residues in protein-protein interactions in yeasts is, therefore, the focus of Chapter 3. This investigation into the evolution of PPIs helps uncover some of the evolutionary design principles behind variations in PPIs within a species.

## 2.6 Evolution of PPI network rewiring between species

As discussed in previous sections, the wealth of newly available data on protein-protein interaction (PPI), including species-level interactome maps of all interacting proteins in a given species, and molecular-level detailed structural data on individual PPIs allows us to better study variations in PPIs within a species. Findings from these types of analyses have given us critical insights into the structural properties and evolutionary constraints that shape interactions between proteins within a species. However, this large amount of PPI data can also be used to examine variations in PPIs (or PPI rewiring) between species, an investigation that is essential to fully understand PPIs and PPI evolution. Indeed, PPIs can be drastically different even between closely

related species, reflecting significant lineage-specific and species-specific changes in molecular processes during evolution. Comparative analyses of PPIs across different species are, therefore, crucial to uncover how PPI networks evolve and rewire over time, to better understand the molecular mechanisms driving these changes, and to study implications of PPI rewiring for cellular function and adaptation. Moreover, such knowledge on natural, evolutionary, variations in PPIs between species is key when investigating dysregulation or disruption of PPIs associated with disease, as well as provides insights to the fields of synthetic biology, and genome engineering [7,8]. Here, we, therefore, first describe the general principles governing the evolution of PPI networks (interactomes). We then discuss previous comparative analyses of interactomes across different species, summarizing the key findings and conclusions drawn from these studies. Finally, we review previous works investigating the molecular evolutionary mechanisms underlying PPI rewiring across species.

### 2.6.1 Evolution of PPI networks

Once a large number of protein-protein interactions (PPIs) have been studied in a given species, a protein-protein interaction network (or interactome) can be constructed. These interactomes are comprehensive maps of all physical binary interactions between proteins detected thus far in a given organism, representing the intricate web of molecular interactions between various proteins crucial for cellular processes. Over time, interactomes can evolve and change, either via protein sequence evolution while interactions are maintained, via gain or loss of genes and proteins, or via gain or loss of interactions while proteins are maintained. These different mechanisms can lead to extensive rewiring in interactomes between different species.

One mechanism by which interactomes evolve is by accruing changes and mutations in protein sequences over time while maintaining PPIs. In order to maintain interactions despite sequence changes, protein sites that are particularly important to PPIs are often under increased evolutionary pressures. For instance, PPI interfaces tend to be highly conserved and under strong purifying selection, while mutations occurring outside of the interaction interface, can often be neutral with respect to the interaction [160,162,205]. In some cases, a mutation in one protein can also be compensated by a nearby mutation in the binding partner, maintaining the interaction through a co-evolutionary process [212,213]. For instance, ~33% of PPIs that are preserved between *S. cerevisiae* and *S. pombe* (i.e. both species have a corresponding, homologous PPI) show evidence of co-evolution [11]. This fraction could, however, be much smaller for more closely related species [214].

Interactome evolution can also be driven the gain or loss of protein-coding genes. In eukaryotes, new genes are most often introduced by either small-scale duplication or whole-genome duplication events which produces a pair of paralogous genes. The resulting paralogous protein pairs initially share the same PPIs, but can subsequently undergo sub-functionalization, a process by which each daughter protein maintains only a subset of the interactions of the parent protein [215]. Duplicate proteins can also gain new interactions and acquire new functions in a process called neo-functionalization [216]. In addition to gene duplication, new genes can also be introduced by horizontal gene transfer [217] and *de novo* gene birth [218]. New proteins introduced in such a manner are initially subject to tight regulation and gradually integrated into PPI networks by forming new interactions [219].

Finally, interactions can be lost, gained, or rewired while the proteins themselves are maintained. For instance, it is estimated that only ~50% of interactions are preserved between proteins that have homologs in both *S. cerevisiae* and *S. pombe* [11]. This means that even if a pair of proteins in one species has a corresponding orthologous pair of proteins in the other species, the interactions between each respective pair of proteins could be different or rewired between the two species. Rates of PPI evolution and PPI rewiring also vary greatly among different parts of interactomes. For instance, some stable protein complexes in animals that were inherited from unicellular ancestors are only modified slightly during evolution [220]. On the other hand, some domain-motif interactions rewire at much higher rates and can rapidly adapt to lineage-specific conditions [221,222]. Moreover, interactions connecting different functional modules of the interactome tend to rewire faster than interactions within specific functional modules [11].

Overall, the evolution of PPI networks is a complex and multifaceted process and differences in interactome networks between species can be varied, with some interaction being preserved despite changes in protein sequences, and some being different or rewired either with or without gene gain or loss. Comparing interactomes between species can help uncover the principles and molecular mechanisms governing PPI network evolution. Understanding these mechanisms provides insights into how organisms can adapt their molecular networks to meet the demands of their environments and evolutionary histories. This knowledge could also be applied to broader biological contexts, including disease mechanisms and synthetic biology applications.

### 2.6.2 Comparative analysis of interactomes between species

High confidence protein-protein interactions (PPI) networks (or interactomes) have been constructed for a variety of species, including human (*Homo sapiens*) [62], fruit fly (*Drosophila melanogaster*) [63], mouse (*Mus musculus*) [64], *Arabidopsis thaliana* [65], baker's yeast (*Saccharomyces cerevisiae*) [10], and fission yeast (*Schizosaccharomyces pombe*) [11]. These interactomes provide valuable insights into the protein networks that underpin biological functions and offer a foundation for comparative analyses. With the availability of these detailed interactome maps, comparative analysis has become a powerful tool for understanding the evolutionary dynamics of PPI networks. By comparing interactomes across different species, researchers can identify conserved and divergent interactions, revealing how PPI networks evolve and adapt over time. This comparative approach is crucial for uncovering the molecular evolutionary mechanisms underlying network rewiring and for understanding the functional implications of these changes. Such analyses also inform our knowledge of evolutionary biology, disease mechanisms, and have potential applications in synthetic biology and genome engineering [7,8].

For instance, Cesareni et al. (2005) conducted a comparative analysis of interactome networks between baker's yeast and fruit fly, estimating that only approximately 24% of yeast PPIs are present in the fly interactome [223]. This substantial difference underscores the extent of PPI network rewiring that occurs between species, highlighting the evolutionary adaptability of PPI networks. Gandhi et al. (2006) compared interactomes between human, baker's yeast, worm, and fruit fly and found the overlap in protein interactions between the four species to be very small, indicating a high level of network rewiring [224]. This finding suggests that, while some interactions may be conserved, the majority of PPIs undergo significant evolutionary changes, reflecting

species-specific functional adaptations. Vo et al. (2016) performed a comparative analysis between fission yeast, baker's yeast and human, estimating that only about 40% of fission yeast PPIs were conserved in baker's yeast, and approximately 65% of fission yeast PPIs were conserved in humans [11]. This study highlights varying degrees of conservation across species and the evolutionary pressures driving the divergence of PPI networks, emphasizing that while essential interactions are often conserved, many are rewired to reflect evolutionary adaptations. These and other previous works highlight that interactome networks undergo significant rewiring during evolution, via either gain or loss of proteins or interactions, leading to lineage-specific and species-specific changes in molecular processes and cellular functions [225,226].

### 2.6.3 Molecular mechanisms underlying interactome rewiring

While interactome networks have been compared between species, uncovering both PPIs that are preserved even between evolutionarily distant lineages, and PPIs that are very different and extensively rewired between species, the detailed molecular evolutionary mechanisms underlying interactome network rewiring, and the site-specific selective pressures acting on rewired PPIs are not well-understood, especially on the genomic scale [158]. Understanding these mechanisms is crucial for comprehending how cellular networks adapt and evolve over time. As such, several previous studies have attempted to elucidate molecular evolutionary mechanisms underlying biological network rewiring as a whole, and PPI rewiring more specifically for some types of interactions.

Studies in the field of graph theory have given us interesting insights on factors that could influence evolutionary rewiring in biological networks, including PPI networks. For instance,

Shou et al. (2011) used computational models to measure the evolutionary rewiring of biological networks, including PPIs, across different species [226]. They identified key factors influencing network rewiring, such as gene duplication, mutation, and changes in regulatory elements. Their findings provided a framework for understanding how genetic and environmental factors drive the evolution of biological networks, highlighting the importance of functional innovation and adaptation in network dynamics. Yamada and Bork (2009) reviewed the evolution of biomolecular networks, including metabolic and protein interactions [225]. They discussed the conservation and divergence of network motifs and modules, showing that while some network components are highly conserved due to their essential roles, significant rewiring occurs to adapt to new environmental conditions and functional demands. Their review underscored the importance of studying network evolution to understand the underlying molecular mechanisms of adaptation and innovation.

Other works have aimed to uncover molecular evolutionary mechanisms underlying PPI network rewiring for specific types of PPIs. For example, Xin et al. (2013), surveyed and compared interactions mediated by 79 SH3 domains in worm, baker's yeast, and human [227]. They observed drastic rewiring between worm and yeast. This extensive rewiring was attributed to variations in the sequence of the motifs recognized by the SH3 domains, as well as changes in binding specificities between orthologous SH3 domains. This study highlighted the role of protein domain evolution in PPI network rewiring, showing that even small changes in domain sequences can lead to significant shifts in interaction patterns across species. Reinke et al. (2013), experimentally compared interactions between 53 human bZIP proteins to their homologs in four other species including fly and worm [228]. They found significant rewiring of the bZIP interactome, driven by

changes to just one or two amino acids. This study demonstrated how minor genetic variations can lead to substantial changes in interaction networks, emphasizing the plasticity and adaptability of PPIs over evolutionary timescales. While those previous studies are crucial in furthering our understanding of the mechanisms of PPI rewiring for these two groups of PPIs, those are specific examples of PPIs and a large-scale analysis of the detailed molecular evolutionary mechanisms underlying interactome network rewiring on the genomic scale remains needed. We address this blind spot in the current literature by studying the detailed molecular evolutionary mechanisms underlying PPI network rewiring, and the site-specific selective pressures acting on rewired PPIs between the baker's yeast (*S. cerevisiae*) and fission yeast (*S. pombe*) interactomes in Chapter 4. This investigation into the evolution of PPI rewiring helps uncover some of the evolutionary design principles behind variations in PPIs between species.

## Preface to Chapter 3

With this context now in place, the upcoming chapter covers details on Aim 1 and Aim 2 of this thesis.

**Aim 1** focuses on the creation of an automated, custom pipeline to curate, process and organize protein-protein interaction (PPI) data from online databases into molecular models of PPIs for two yeast species, *Saccharomyces cerevisiae* (*S. cerevisiae*), and *Schizosaccharomyces pombe* (*S. pombe*). This pipeline gathers and combines large amounts of PPI data from very different experimental fields, therefore allowing for novel analysis.

**Aim 2** focuses on utilizing molecular models of PPIs in *S. cerevisiae* to study the relationship between PPI structure and PPI evolution, by studying the impact of various structural determinants on residue evolutionary rates in yeast. No systematic comparison, integrating a comprehensive list of different structural features in terms of their impact on residue conservation at the proteome scale has previously been performed in the literature. As such, this aim uncovers design principles and structural mechanisms that influence the evolution of PPIs within a species.

The article included here covers the development of a custom script pipeline to automate the gathering, quality control and organization of PPI data into molecular models of PPIs for a species. Interactome data (records of all interactions between pairs of proteins in a species), and molecular structure data (detailed, atom-resolution, three-dimensional descriptions of individual PPIs) were successfully curated for both *S. cerevisiae* and *S. pombe*, and molecular models of PPIs in both species have been constructed. The custom script pipeline designed to gather and combine large amounts of PPI data from very different experimental fields is further described in the article and could be applied to future works in the two yeasts or in other species. Moreover, the detailed

molecular models of PPIs in *S. cerevisiae* and *S. pombe* combine PPI data at two very different scales in a unique manner, enabling novel future analysis.

The article also covers the use of molecular models of PPIs in *S. cerevisiae* to investigate the relationship between PPI structure and PPI evolution. Extensive work in identifying structural determinants (measurable characteristics of the structure of the microenvironment surrounding a residue) correlated with residue evolution in *S. cerevisiae* PPIs was performed. The final structural determinants selected are the change in relative solvent accessibility upon PPI binding (ΔRSA), the number of residue-residue contacts across the PPI interface (interRRC), and the distance from the center (dCenter) or the periphery (dEdges) of the PPI interface. These four determinants are further described in the article. Moreover, several significant correlations between these structural determinants and residue evolutionary rates in *S. cerevisiae* PPIs were uncovered and are discussed in depth. Estimations of the overall, and relative importance of these determinants to our understanding of the relationship between structure and evolution of PPIs were also performed. Overall, the following important conclusions were reached: (i) interfacial residues in PPIs are subject to continuous, structure-based selective constraints proportional to their degree of interface involvement, (ii) interfacial burial (as measured by the structural determinant ΔRSA) is selectively equivalent to non-interfacial burial, (iii) in addition to ΔRSA, other measures of interface involvement (structural determinants interRRC, dCenter, and dEdges) independently constrain residue evolution, and (iv) in addition to these continuous structure-based selective constraints, interfacial residues are subject to a fixed function-based selective constraint independent of their degree of interface involvement. Those findings help establish some of the design principles and structural mechanisms that influence the evolution of PPIs within a species.

# 3. Research Article No. 1: Structural Determinants of Yeast Protein-Protein interaction Interface Evolution at the Residue Level

Léah Pollet [1], Luke Lambourne [2,3,4*] and Yu Xia [1*]

1 - Department of Bioengineering, Faculty of Engineering, McGill University, Montreal, QC, Canada

2 - Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA, USA

3 - Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA, USA

4 - Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA

Correspondence to Luke Lambourne and Yu Xia: Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA, USA (L. Lambourne).

luke_lambourne@dfci.harvard.edu (L. Lambourne), brandon.xia@mcgill.ca (Y. Xia)

https://doi.org/10.1016/j.jmb.2022.167750

Edited by Michael Sternberg

## 3.1 Abstract

Interfaces of contact between proteins play important roles in determining the proper structure and function of protein–protein interactions (PPIs). Therefore, to fully understand PPIs, we need to better understand the evolutionary design principles of PPI interfaces. Previous studies have uncovered that interfacial sites are more evolutionarily conserved than other surface protein sites. Yet, little is known about the nature and relative importance of evolutionary constraints in PPI interfaces. Here, we explore constraints imposed by the structure of the microenvironment surrounding interfacial residues on residue evolutionary rate using a large dataset of over 700 structural models of baker's yeast PPIs. We find that interfacial residues are, on average, systematically more conserved than all other residues with a similar degree of total burial as measured by relative solvent accessibility (RSA). Besides, we find that RSA of the residue when the PPI is formed is a better predictor of interfacial residue evolutionary rate than RSA in the monomer state. Furthermore, we investigate four structure-based measures of residue interfacial involvement, including change in RSA upon binding (ΔRSA), number of residue-residue contacts across the interface, and distance from the center or the periphery of the interface. Integrated modeling for evolutionary rate prediction in interfaces shows that ΔRSA plays a dominant role among the four measures of interfacial involvement, with minor, but independent contributions from other measures. These results yield insight into the evolutionary design of interfaces, improving our understanding of the role that structure plays in the molecular evolution of PPIs at the residue level.

## 3.2 Introduction

Understanding the nature of constraints on protein evolution has long been the focus of much scientific interest. Evolutionary rates are known to vary widely between proteins. For instance, genes encoding highly expressed proteins, proteins that carry out crucial functions, or proteins that interact with many partner proteins tend to be more conserved [1–3]. In addition to molecular function, three-dimensional structure is known to play an important role in constraining protein evolution [4–5]. Moreover, evolutionary rates can vary significantly among different sites, even within a given protein. For example, sites in the core of most proteins or catalytic residues in enzymes typically evolve more slowly than other sites [6–7]. In order to develop a complete picture of protein evolution, work in this field has focussed on uncovering determinants of site-specific evolution, highlighting that site- specific evolutionary rates are under both structural and functional constraints in single proteins [8].

However, proteins seldom work in isolation in cells, but rather act via protein–protein interactions (hereafter referred to as PPIs). Indeed, protein function tends to be regulated via transient interactions with protein kinases and other enzymes [9]. Moreover, many cellular processes are carried out by stable protein complexes, which behave as molecular machines, composed of protein components, and organized by tightly regulated PPIs to ensure proper function [10–11]. Additionally, changes and mis-regulations in PPIs can have important consequences for organismal fitness: several disease-causing mutations are known to disrupt PPIs and single nucleotide polymorphisms associated with a number of diseases tend to occur in sites predicted to mediate interactions [12–14]. Yet, PPIs are typically not considered when investigating proteome-wide, quantitative agents of selective constraint on site-specific evolutionary rates in the literature.

Therefore, little is known about factors influencing the evolution of PPIs at this most basic level of fixation and elimination of single amino acid residue mutations.

To contribute to our understanding of PPI evolution, we investigate the region of contact between partner proteins in a PPI, also referred to as the interface. The interface is a key PPI feature, both structurally and functionally. Interfacial residues are considered surface residues in most structure-based studies that treat the two protein partners of a PPI as free-floating monomers [15]. However, upon formation of a PPI complex, the structure of the micro-environment surrounding those residues can be drastically altered. Moreover, interfacial residues are functionally unique: mutations to interfacial residues can alter protein–protein binding affinities and proper PPI function with a wide range of implications for human health [16–17]. Interfacial residues also tend to be more evolutionarily conserved than non-interfacial surface residues, although debate still remains in the literature as this conclusion varies in significance depending on the dataset considered [18–21].

While residues in PPI interfaces have both a unique structural context and a specific function in mediating crucial interactions between protein partners, structure-based studies of protein evolution typically rely upon coarse distinctions between "interfacial" and "non-interfacial" residues, if considering interfaces at all. The few key studies that go beyond this binary distinction and investigate structure-based evolutionary constraints within PPI interfaces, uncovered that considering whether an interfacial residue is accessible to solvent in a co-complexed PPI structure can help further subdivide interfaces into different regions. These regions include a core, a rim, and support regions, each with predictable and distinct amino acid

compositions, and sequence entropy [22–26]. However, these studies only consider a single possible structural constraint on the evolution of interfacial residue: the residue's exposure to solvent. Moreover, this structural constraint is discretized to categorize residues into each region: for instance, all interfacial residues that are inaccessible to solvent in a PPI structure are assigned to the core region of the interface, while residues accessible to solvent in a PPI structure are labeled as belonging to the rim or support regions of the interface. Consequently, our knowledge of any continuous structure-based evolutionary constraints on interfacial residues remains limited at the proteome scale. In the main previous study on such continuous structural constraints, Franzosa and Xia developed a quantitative measure of PPI interface involvement based on solvent accessibility and investigated its relationship with residue evolutionary rate [27], but no other structural measures of interfacial involvement were investigated, and the number of interfaces studied by the authors was relatively small.

Overall, interest in uncovering structural features that can influence, or help predict evolutionary rates in interfaces is high, as evidenced by several studies on the topic in recent years. These previous works uncovered various structural features that correlate with residue evolutionary rates ranging from solvent accessibility and interface involvement [27–28], to location of a residue within an interface [22–24] and local packing surrounding the residue [8,29–30]. However, these previous studies have only investigated the relationship between residue evolution and a single structural feature for interfacial residues. In contrast, our current work systematically compares and integrates a comprehensive list of different structural features of residue interfacial involvement, in terms of their impact on interfacial residue conservation. As such, our current work settles the long-standing debate over which structural features are most important at

constraining residue evolution in PPI interfaces. Here, we, therefore, study interfacial residues in a large dataset of *Saccharomyces cerevisiae* PPI structures and a comprehensive list of structural features of PPI interfacial involvement that could influence their evolutionary rates. We elected to use *S. cerevisiae* in this analysis, as the large amount of experimentally derived physical interactions reported in the species in recent years, combined to the relatively small size of the yeast genome, make it one of the most complete interactomes currently available as well as one of the most complete interactome with regards to structural experimental data [31–32].

We find that, on average, interfacial residues in *S. cerevisiae* PPIs are more conserved than non-interfacial residues. Moreover, interfacial residues are systematically more conserved than non-interfacial residues with an equivalent degree of total burial, including residues buried at the core of a protein. The result is surprising because core residues are thought to evolve under very strong constraints to maintain protein structure and stability [33–34]. This confirms the existence of strong evolutionary constraints associated with the function of the interfacial residue, independent of its degree of burial [27–28]. In addition, we find a strong relationship between the relative solvent accessibility (RSA), or burial, of an interfacial residue in a PPI and its evolutionary rate. This result confirms previous work establishing solvent exposure as a major structural determinant of residue evolutionary rate in single proteins [27,35]. Moreover, we find that residue RSA in complexed state is a better predictor of interfacial residue evolutionary rate than RSA in monomeric state. Hence, from an evolutionary standpoint, interfacial residues are mainly constrained by solvent exposure when the PPI proteins are co-complexed.

To further probe the quantitative structural basis of residue-level evolutionary constraints in PPI interfaces, we investigate several continuous structure-based measures of residue interfacial involvement and their relationship with residue evolutionary rate. These structural features include local properties of an interfacial residue's microenvironment such as the change in its RSA upon binding (ΔRSA), and the number of residue- residue contacts it makes across the interface, as well as more global properties such as its distance from the center or the periphery of the interface. We find significant correlations between the four structure-based measures of interface involvement and residue evolutionary rate in PPI interfaces, establishing that, even within the already highly constrained context of the interface, different residues may be under different evolutionary pressures [23–24,27,36–37]. Integrated modeling shows that these measures of interface involvement are predictive of interfacial residue evolutionary rate, with ΔRSA playing a dominant role, and minor but independent contributions from other measures.

This work yields insight into the evolutionary design principles of interfaces and the identification of some of their key features, improving our understanding of the molecular evolution of PPI interfaces at the residue level.

## 3.3 Results

**Residues in PPI interfaces are, on average, more conserved than non-interfacial residues**

We gathered a dataset of three-dimensional structures for a list of high confidence PPIs in *S. cerevisiae*. For PPIs with no structure currently available, we used the structure of a closely related homologous PPI. This homology-based structural annotation transfer process was successfully used in previous work [27,38]. The final dataset comprises structures for 718 PPIs between 611 S. cerevisiae proteins containing more than 400,000 residues (data curation pipeline summarized in **Figure 1** and further detailed in **Materials and Methods**). Interfacial residues in our data were defined as amino acids exhibiting a change in solvent accessibility upon formation of a PPI complex.

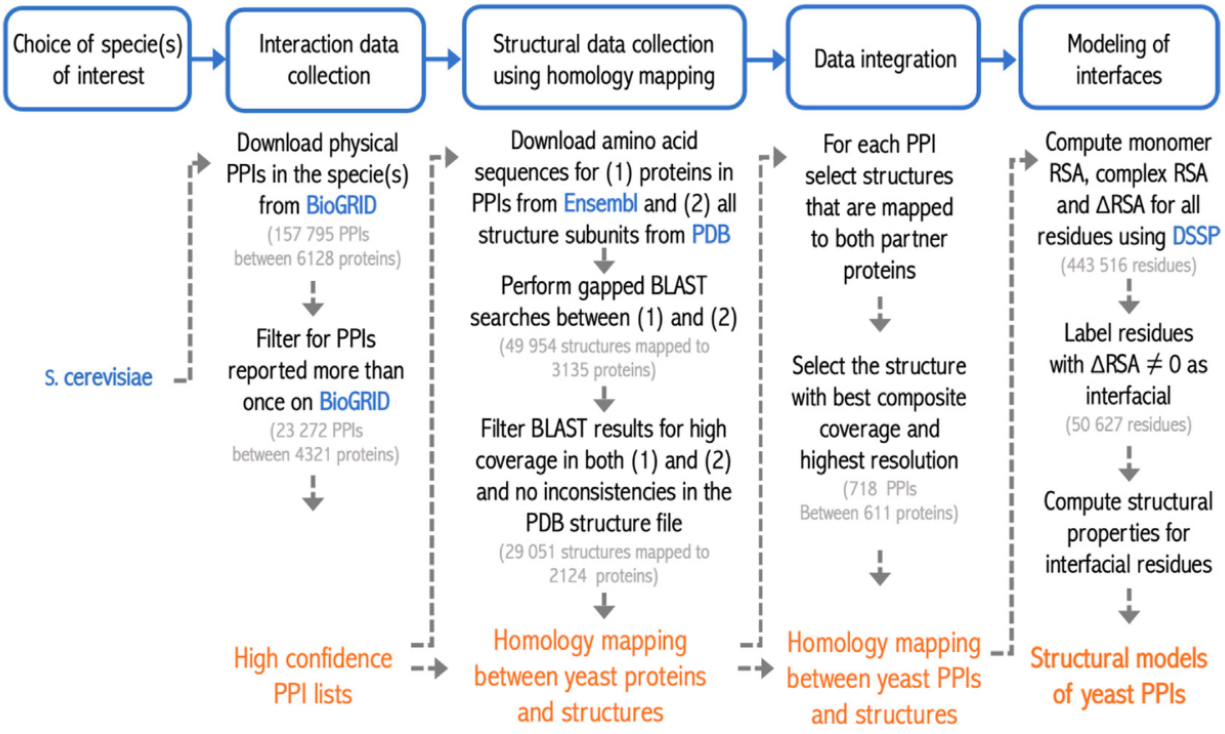



**Figure 1. Computational pipeline.** Graphical representation of the pipeline used for automated curation and homology-based structural annotation transfer of PPIs in *S. cerevisiae*. The pipeline is available in a GitHub repository.

**(A) Definitions**

| Symbol | Residue property | Description |
|---|---|---|
| **RSA** | Relative solvent accessibility | Computed for all residues. The area traced by the center of a solvent molecule (1.4-Å sphere) while in contact with the residue's molecular surface, divided by the maximum such area observed across all residues of that type (hydrogen atoms and outlier residues excluded). Property computed both for free floating proteins (monomer RSA) and co-complexed proteins (complex RSA) . Illustrations in **(B)**. |
| **ΔRSA** | Change in burial upon complex formation | Computed for all residues. The change in burial of a residue as its parent protein transitions from free-floating to co-complexed state (ΔRSA = monomer RSA - complex RSA). Any residue with ΔRSA ≠ 0 is defined as an interfacial residue. Illustration in **(B)**. |
| **interRRC** | Inter-protein residue-residue contacts | Computed for interfacial residues. The number of nonpolypeptide adjacent residues belonging to the partner protein with at least one non-hydrogen atom within a 4.5-Å radius of some nonhydrogen atom in the residue of interest. Illustration in **(C)**. |
| **dCenter** | Distance from interface center | Computed for interfacial residues. The Euclidian distance between the Cα atom of the residue and the geometric center of the interface that it belongs to in the parent protein. Illustration in **(D)**. |
| **dEdges** | Distance from interface edges | Computed for interfacial residues. The minimum Euclidian distance between the Cα atom of the residue and the Cα atom of any residue in the 90$^{th}$ percentile of largest values of complex RSA (edges of the interface) in the parent protein. Illustration in **(E)**. |

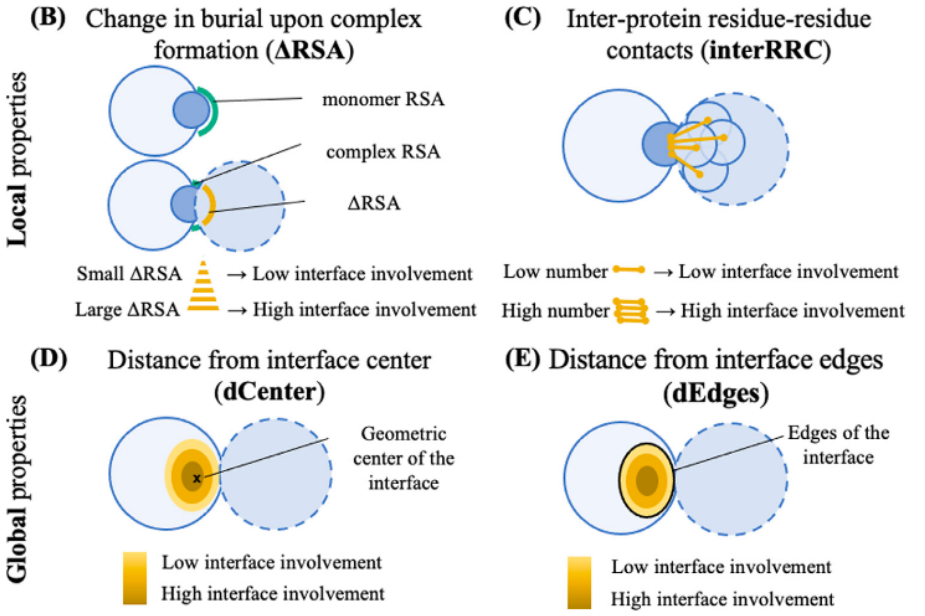



**Figure 2. Structural properties of the residue microenvironment. (A)** Symbols, names, and descriptions of the structural properties of a residue's microenvironment. **(B)-(D)** Graphical representation of the structural properties of a residue's microenvironment. For each property, a cartoon diagram of a pair of interacting proteins is shown in cross section, in blue, with one of the protein partners differentiated with a dashed outline. For local properties (**(B), (C)**), the residue of interest to the calculation is highlighted in darker blue and structural properties are illustrated in green (monomer RSA, complex RSA) and yellow (ΔRSA, interRRC). The yellow properties have an associated degree of involvement of the residue in the interface. Global properties (**(D), (E)**) are determined by relative positions, so their value and corresponding degree of interface involvement are represented by a shade of yellow.

We, therefore, computed two measures of relative solvent accessibility (**Figure 2**, monomer RSA, complex RSA) for all residues in our data. For a residue in a PPI protein, monomer RSA measures its accessibility to solvent when the protein is in a monomeric state, while complex RSA measures its accessibility or burial when the protein is co-complexed with a partner protein. ΔRSA, the change in a residue's burial upon complex formation, was computed as the difference between monomeric and co-structured RSA values for each residue (ΔRSA = monomer RSA – complex RSA), and any residue with ΔRSA ≠ 0 was labeled as interfacial. This yielded more than 50,000 interfacial residues. We then assessed the degree of evolutionary constraint on interfacial and non-interfacial residues in *S. cerevisiae* PPIs separately using two techniques: ω = *dN/dS* ratio (hereafter referred to as dN/dS) and ConSurf score (see **Materials and methods**). These represent two popular ways to estimate residue evolutionary rate in the literature.

**Figure 3(A)** shows the average evolutionary rates for interfacial residues (ΔRSA ≠ 0) and non- interfacial residues (ΔRSA = 0) in our data. Interfacial residues are, on average, significantly more conserved than their non-interfacial counterparts using both estimates of evolutionary rates (*dN/dS* and *ConSurf* score). The difference in average *dN/dS* value between the two groups (average difference: 0.0199) signifies that 1 fewer amino acid substitution every 50 silent mutations is expected for interfacial residues compared to non-interfacial ones. T-tests for the difference in average evolutionary rates between the two groups (**Figure 3(B)**) indicate a statistically significant difference using both evolutionary rate estimates (P-value < 0.01). Previous work showed that interfacial residues are under unique evolutionary constraints and tend to be more conserved than non-interfacial surface residues, although evaluations of those constraints vary depending on the dataset considered [18–20]. Here we find that, for proteins involved in *S.*

*cerevisiae* PPIs, interfacial residues are, on average, more conserved than non-interfacial residues, even when including residues buried at the core of a protein in the comparison. This is surprising as core residues are under very strong evolutionary pressure to maintain protein structure and stability [6,33–34] and confirms the existence of strong evolutionary constraints associated with the involvement of a residue in an interface [27–28].

**Interfacial residues in PPIs are systematically more conserved than all other residues of the same total burial**

As previously noted, residues buried in the core of a protein structure, are typically highly conserved. However, this observation is only a single example of a continuous, quantitative trend identified in single proteins: the exposure or burial of a residue to solvent is usually correlated with its evolutionary rate [27,35,39].

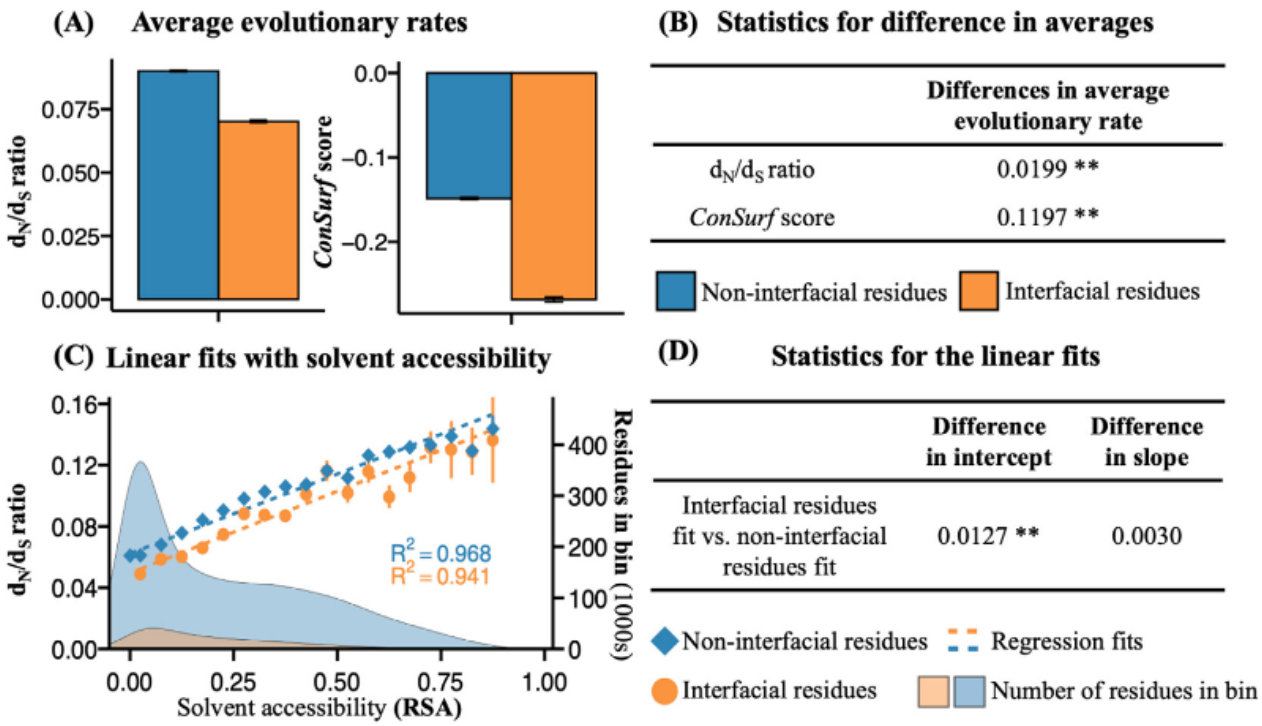


**Figure 3. The difference in evolutionary rate between interfacial and non-interfacial residues. (A)** Average evolutionary rates (estimated using both *dN/dS* ratio and *ConSurf* score), plotted for interfacial and non-interfacial residues from all PPIs in our data. Standard errors for the average values of each group of residues are also shown. **(B)** Results of t-tests for differences in average evolutionary rates between interfacial and non-interfacial residues using both evolutionary rate estimates. Differences significant at the P-value < 0.01 level are denoted with a double asterisk (**). **(C)** Linear fits between binned measures of solvent accessibility (complex RSA) and evolutionary rate (*dN/dS*) for interfacial and non-interfacial residues. Distributions of the number of residues per bin, weighted linear regression lines, and $R^2$ values of the fits are also shown. **(D)** Results of t-tests for differences in slope and intercept between the two fits in **(C)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**).

Indeed, solvent exposure or degree of burial (quantified by measures such as relative solvent accessibility – RSA) has been established as a significant structural predictor of residue evolutionary rate in single proteins [27]. The difference in average evolutionary rate between interfacial and non-interfacial residues in **Figure 3 (A)** could, therefore, be due to differences in the average total burial of residues in the two groups. To test this possibility, we binned interfacial and non-interfacial residues in our data separately, in 5% intervals over the range of possible complex RSA values, from fully buried residues (complex RSA = 0) to fully exposed residues (complex RSA = 1). Evolutionary rate (*dN/dS*) was then calculated for the residues in each bin by concatenating the aligned codons of *S. cerevisiae* and eight other closely related yeast species.

**Figure 3(C)** shows the relationship between evolutionary rate and complex RSA for interfacial residues and non-interfacial residues binned as described above. Taking into account the *dN/dS* estimation error associated with each complex RSA bin, we generated least-squares regression lines with coefficient $R^2 = 0.968$ and $R^2 = 0.941$ for non-interfacial and interfacial residues respectively. Both regression lines show a strong, linear, and positive relationship over the full range of complex RSA values. The slopes of the two regression lines are not statistically different (P- value > 0.05), however, the intercepts are (difference in intercept = 0.0127, P-value

< 0.01). These results indicate that interfacial residues are more conserved than non-interfacial residues of similar total burial for the full range of complex RSA values considered. More specifically, for a pair of interfacial and non-interfacial residues equivalently buried in the PPI complex, the interfacial residue will evolve more slowly than the non-interfacial residue (expected difference in *dN/ dS*: 0.0127, i.e., one fewer amino acid substitution for every 79 silent mutations for the interfacial residue compared to the non-interfacial one). Previous work showed that interfacial residues are more conserved than non-interfacial surface residues [18,20,23,25–26]. Here, however, we find that, when correcting for total burial, interfacial residues are systematically more conserved than all non- interfacial residues, including residues at the core of proteins. This surprising result confirms the existence of a strong functional constraint associated with the involvement of a residue in an interface, beyond what can be explained by total burial [27]. We verified these observations using another proxy for residue evolutionary rate: *ConSurf* score, and the results remain consistent. Details of this second regression analysis are available **in Supplementary material** and the resulting plots can be found in **Figure S1**.

**Solvent exposure in co-complexed state better predicts PPI interfacial residue evolutionary rate than solvent exposure in monomer state**

Previous work showed that residues at the core of proteins are typically highly conserved, while residues on the surface tend to evolve faster [8,27,33–34]. Here, we study a third category of residues: interfacial residues. These residues are unique as they are buried in the PPI complex when two protein members of a PPI come together but remain surface residues if considering the two proteins separately, as monomers. Therefore, either of those two structural contexts (monomeric or complexed) could be used when investigating structure-based constraints on their

evolution. Extensive work has already been performed in the monomeric context, uncovering structural features such as RSA correlated with residue evolutionary rate in monomeric proteins [4–5,8,27]. RSA can also be computed from PPI complex structures, measuring the accessibility of a residue to solvent when co-complexed with a partner protein. Hence, to uncover whether co-complexed PPI structures should also be considered when studying structure-based constraints on interface evolution, we examined and compared the relationship between RSA and site-specific evolutionary rates in the monomeric and the complexed contexts.

We computed two measures of RSA for 50,627 interfacial residues in our data: monomer RSA using monomeric structures, and complex RSA using co-complexed PPI structures (**Figure 2**, monomer RSA, complex RSA). *ConSurf* score was calculated for each interfacial residue. To compute dN/dS, interfacial residues were binned in 5% intervals over the range of possible RSA values for monomer RSA and complex RSA separately. *dN/dS* was calculated for each bin by concatenating the aligned codons of *S. cerevisiae* and eight other closely related yeast species. We then correlated RSA values with both estimates of evolutionary rate (*dN/dS* and *ConSurf* score). The results in **Figure 4(A)** show significant, positive correlations with *dN/dS* for both monomer RSA ($r = 0.957$) and complex RSA ($r = 0.717$), indicating that sites more accessible to solvent (either in the complex or monomer state) evolve progressively more quickly. Results remain significant when weighting correlations by standard errors on *dN/dS* calculations for both complex RSA ($r = 0.952$) and monomer RSA ($r = 0.964$). These trends are confirmed by the *ConSurf* score estimate of evolutionary rate: the correlations remain positive and statistically significant for both complex RSA ($r = 0.131$) and monomer RSA ($r = 0.091$) when weighted by the standard error on

*ConSurf* score calculations. It is worth noting that correlations with *ConSurf* scores are smaller than the ones with *dN/dS*.

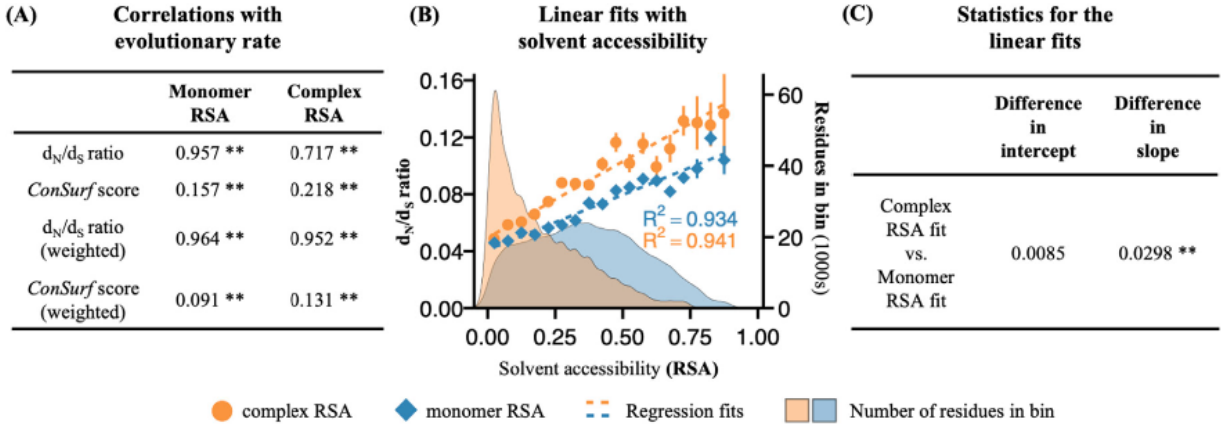



**Figure 4. The relationship between solvent accessibility and evolutionary rate in PPI interfaces. (A)** Results of a Pearson product-moment correlation test between values of solvent accessibility and measure of evolutionary rates for interfacial residues in our models. The first two rows in the table show standard Pearson correlations, whereas the two last rows, show weighted Pearson correlations, taking the standard error on evolutionary rate estimates into consideration. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**). **(B)** Linear fits between binned measures of solvent accessibility and evolutionary rate (*dN/dS*), for monomer solvent accessibility (monomer RSA) and complex solvent accessibility (complex RSA). Distributions of the number of residues per bin, weighted linear regression lines, and $R^2$ values of the fits are also shown. **(C)** Results of t-tests for differences in slope and intercept between the two fits in **(B)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**).

This difference in correlation value is expected as *ConSurf* data are not binned whereas *dN/dS* data are. The difference also indicates that values of correlations with *dN/dS* may be inflated by the binning process. Little import should therefore be placed on the numerical value of correlations with *dN/dS* besides their sign and whether they are significant, and correlation values should not be directly compared to each other. Instead, this correlation analysis can be used to establish significant linear relationships between both measures of RSA and evolutionary rate. While previous work uncovered a direct relationship between residue burial and protein stability

[40], and a close connection between protein stability and organismal fitness [41], our findings suggest that burial in a PPI and PPI stability is also selected for and linked to organismal fitness.

We further investigated the high correlations between solvent accessibility and residue evolutionary rate in PPI interfaces using weighted least-square regression. **Figure 4(B)** shows the relationship between *dN/dS* and both measures of solvent accessibility for interfacial residues binned as described above. Taking into account the *dN/dS* estimation error associated with each RSA bin we generated least-squares regression lines with coefficient $R^2 = 0.934$ and $R^2 = 0.941$ for monomer and complex RSA respectively. Both regression lines share the same intercept (difference in intercept is not statistically significant, P-value > 0.05), indicating that fully buried residues, either in a monomeric protein or in a PPI, will have the same, low, evolutionary rate. Moreover, the slopes of the monomeric and complex regression lines, show a statistically significant difference (difference in slope = 0.0298, P-value < 0.01). For monomer RSA, the slope of 0.08 indicates that a 1% increase in RSA is associated with a *dN/dS* increase of approximately 0.0008 (i.e., one extra amino acid substitution for every 1,250 silent mutations). This result is in agreement with previous work in yeast [27]. For complex RSA, the slope of 0.1 indicates that a 1% increase in RSA can be related to a *dN/dS* increase of approximately 0.001 (i.e., one additional amino acid substitution for every 1000 silent mutations). This large difference demonstrates that, during evolution, interfacial residues are mainly constrained by solvent exposure when in complexed state rather than in monomeric state. We confirmed these observations using another proxy for residue evolutionary rate: *ConSurf* score. Details of this second regression analysis are available in **Supplementary material** and the resulting plots can be found in **Figure S2**. Overall,

using PPI complex structures to compute RSA values yields a better correlation with interfacial residue evolutionary rates than simply considering single protein structures.

**The degree of structural involvement of a residue in a PPI interface influences its evolutionary rate**

The above observation that complex RSA better correlates with interfacial residue evolutionary rate than monomer RSA suggests that the change in RSA upon PPI complex formation (ΔRSA) plays an important role in constraining interfacial residue evolution. Moreover, ΔRSA is one of several possible structure-based measures of residue interfacial involvement [27]. To further probe the quantitative structural basis of residue-level evolutionary constraints in PPI interfaces, we, therefore, investigate several continuous, structure-based measures of residue interfacial involvement, and their relation- ship with residue evolutionary rate.

The first measure of residue interfacial involvement that we considered is ΔRSA. ΔRSA quantifies changes in the solvent accessibility of a residue upon formation of a PPI complex. Residues with large ΔRSA values could, thus, be particularly important to an interface as their micro-environment is significantly affected by complex formation, while residues with small ΔRSA values may be less involved in the interaction [27]. In addition to ΔRSA, we computed three other structural properties designed to estimate the degree of involvement of a residue in an interface. These properties, summarized in **Figure 2**, include the number of contacts that a residue makes with residues in the partner protein (interRRC), the distance between a residue and the geometric center of the interface (dCenter), and the distance between a residue and the edges of the interface (dEdges). The structural properties include local features of interfacial residues'

microenvironment (ΔRSA and interRRC) as well as more global measures estimating overall position within the interface structure (dCenter and dEdges). The four properties are only weakly correlated to each other (**Figure 5**), and we, therefore, reason that each has the potential to convey some amount of independent information.

We binned the 50,627 interfacial residues in our data in 10% intervals over the range of each of the structure-based measures of interface involvement (ΔRSA, interRRC, dCenter, dEdges). We correlated the four measures of interface involvement with estimates of evolutionary rate (*dN/dS* for each bin, and *ConSurf* score for each residue). The first two rows of **Table 1** show standard Pearson correlation tests between each structural measure of interface involvement and evolutionary rate. Results in the last two rows of **Table 1** are weighted Pearson correlations, using the standard error on evolutionary rate estimates to weigh the correlation analysis. Here, again, correlations with dN/dS are performed on binned data, and thus, direct conclusions should not be drawn from the numerical values of correlations in **Table 1**. Furthermore, correlation values should not be directly compared to each other. Instead, the significance of the correlations, tested using a Pearson product-moment correlation test, can be used to establish linear relationships between measures of interface involvement and evolutionary rate.

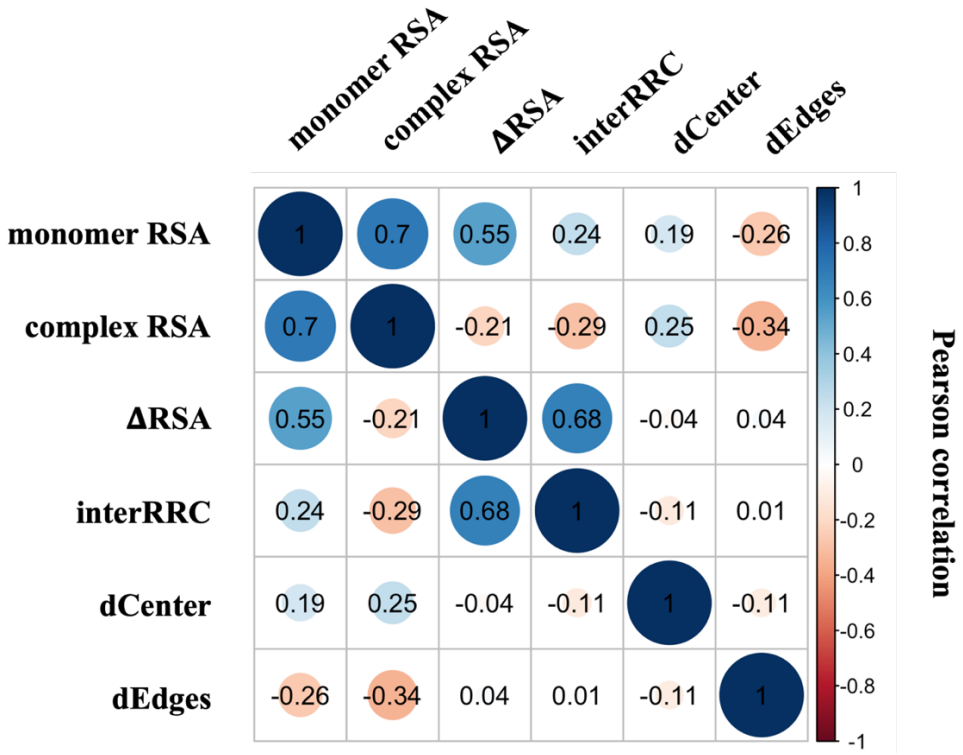



**Figure 5. Correlation between structural properties of a residue's microenvironment**. Pairwise Pearson correlation matrix for the structural properties of interest in this study computed for all interfacial residues.

The four structural properties considered here showed statistically significant correlations (P-value < 0.01) with at least one measure of evolutionary rate. The correlations with ΔRSA and dCenter were found significant regardless of the estimate of evolutionary rate used. These significant correlations indicate that interfacial residues are subject to continuous structure-based selective constraints which are proportional to their degree of interface involvement: residues with a large change in burial upon complex formation, residues which make numerous contacts with the partner protein, and residues closer to the geometric center of interfaces evolve progressively more slowly.

We further investigated the correlations between structural measures of interface involvement and evolutionary rate using weighted least-square regression. The blue diamonds and blue lines in **Figure 6(A)-(D)** show the observed *dN/dS* values of interfacial residues for different bins of ΔRSA, interRRC, dCenter, and dEdges (binned as described above), as well as the corresponding least-squares regression lines, taking into account the *dN/dS* estimation error associated with each interface involvement bin. The slope of the regression line with ΔRSA (slope = -0.02) indicates that, within interfaces, a 10% increase in ΔRSA is associated with a *dN/d*S decrease of approximately 0.002 (i.e., one fewer amino acid substitution for every 500 silent mutations). The regression line with dCenter has a slope of 0.001, showing that a 10% decrease in the distance between a residue and the geometric center of an interface is associated with a *dN/dS* decrease of around 0.0001 (i.e., one fewer amino acid substitution for every 10,000 silent mutations). These are small, incremental changes, but over the whole range of dCenter values, residues closest to the geometric center of an interface are, on average, two times more conserved than the ones furthest away. The regressions with interRRC and dEdges also indicate a continuous, linear relationship with *dN/ dS* over the full range of interRRC and dEdges values. These results confirm the high correlations observed between the four structural measures of interface involvement and evolutionary rate (**Table 1**).

**Table 1 Results of a Pearson product-moment correlation test between structural measures of interface involvement and evolutionary rate estimates for interfacial residues**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**), values significant at the P-value < 0.05 level are denoted with a single asterisk (*).

| | ΔRSA | InterRRC | dCenter | dEdges |
|---|---|---|---|---|
| Correlation with *dN/dS* ratio | −0.619 * | −0.462 | 0.745 * | −0.689 ** |
| Correlation with *ConSurf* score | −0.043 ** | −0.066 ** | 0.07 ** | −0.101 ** |
| Correlation with *dN/dS* ratio (weighted) | −0.846 ** | −0.327 | 0.913 ** | −0.03 |
| Correlation with *ConSurf* score (weighted) | −0.022 ** | −0.039 ** | 0.047 ** | −0.047 ** |

In addition to plotting the observed *dN/dS* values of interfacial residues for different bins of ΔRSA, interRRC, dCenter, and dEdges (as blue diamonds and blue regression lines in **Figure 6**), we also computed expected *dN/dS* values for each bin of interfacial residues (as yellow circles and yellow regression lines in **Figure 6**), operating under the assumption that interfacial burial (as measured by ΔRSA) is selectively equivalent to non-interfacial burial, and that interfacial residues are subject to the same evolutionary constraints as non-interfacial residues with the same total burial in PPIs. In other words, we computed the expected *dN/dS* values for each bin of interfacial residues based on the evolutionary behavior of non-interfacial residues in the following way: (i) we determined the average total burial (as measured by complex RSA) for each bin, and (ii) using the average complex RSA value for each bin, we predicted the expected dN/dS value for each bin, based on the *dN/dS* versus RSA trend for non- interfacial residues in **Figure 3(C)**. We plotted these expected *dN/dS* values as yellow circles and regression lines in **Figure 6**. Notably, the slope of the observed regression line is not significantly different from expected for the trend between *dN/dS* and ΔRSA (P-value > 0.05), suggesting that interfacial burial (as measured by ΔRSA) is indeed selectively equivalent to non-interfacial burial. In contrast, the slopes of the observed regression lines are significantly different from expected for the trends between *dN/dS* and interRRC (P-value < 0.05), dCenter (P- value < 0.01), and dEdges (P-value < 0.05) respectively, suggesting that these three measures of interface involvement make contributions to *dN/dS* that are independent of ΔRSA.

The yellow "X"s in **Figure 6** denote the average *dN/dS* for non-interfacial residues of yeast PPIs (average *dN/dS* for non-interfacial residues = 0.0901). Notably, the yellow "X" for non-interfacial residues in **Figure 6** does not lie on the blue observed regression line for interfacial residues for any of the four measures of interface involvement (P-value < 0.01), but rather lies

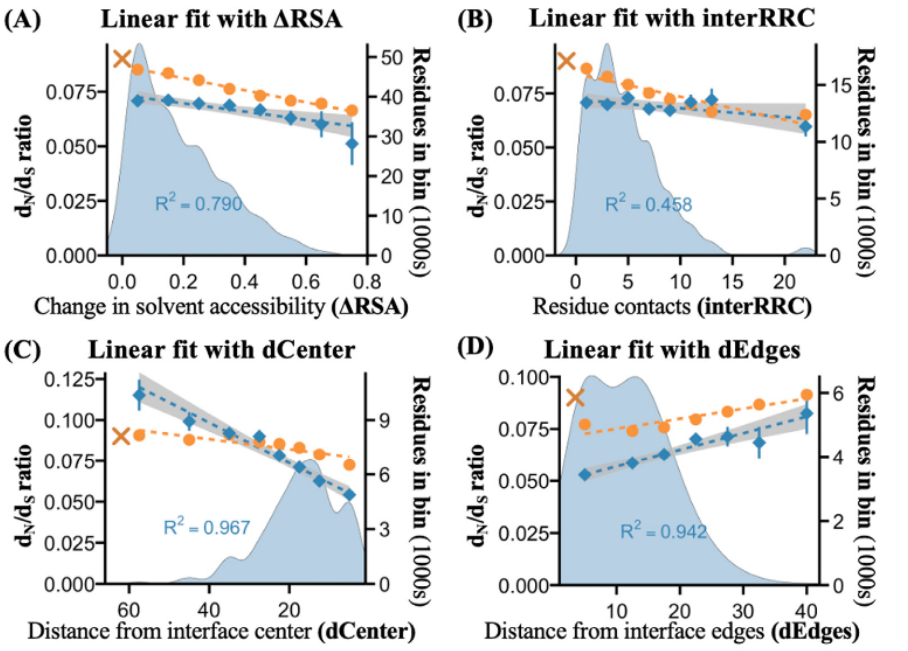
significantly closer to the yellow expected regression line for interfacial residues for all measures of interface involvement (P-value < 0.05). In other words, there is a significant difference in average evolutionary rate between non-interfacial residues and interfacial residues, even for interfacial residues with minimal interface involvement (i.e., residues in the first ΔRSA, interRRC, dEdges, or dCenter bin). These observations suggest that, in addition to the continuous structure-based selective constraints imposed by their degree of interface involvement, interfacial residues are subject to a fixed function-based selective constraint which is independent of their degree of interface involvement.

In addition to *dN/dS*, we confirmed all of these observations using another estimate of residue evolutionary rate: *ConSurf* score. Details of this *ConSurf*-based regression analysis are available in **Supplementary material** and the resulting plots can be found in **Figure S2**. Overall, our results suggest that: (i) interfacial residues are subject to continuous, structure-based selective constraints proportional to their degree of interface involvement; (ii) interfacial burial (as measured by ΔRSA) is selectively equivalent to non-interfacial burial; (iii) in addition to ΔRSA, other measures of interface involvement (interRRC, dCenter, and dEdges) independently constrain residue evolution; and (iv) in addition to these continuous structure-based selective constraints, interfacial residues are subject to a fixed function-based selective constraint independent of their degree of interface involvement.

**Among the four structure-based measures of interface involvement, ΔRSA is the major determinant of interfacial residue evolution**

To further investigate the quantitative nature of the selective constraints imposed by the four structure-based measures of interface involvement (ΔRSA, interRRC, dCenter, and dEdges),

we tested several statistical models to combine their respective power in predicting interfacial residue evolutionary rates.




**(D) Statistics for linear fits**

| | Difference in intercept between observed and expected fit | Difference in slope between observed and expected fit |
|---|---|---|
| ΔRSA | 0.0133 ** | 0.0088 |
| interRRC | 0.0120 ** | 0.0007 * |
| dCenter | 0.0269 ** | 0.0009 ** |
| dEdges | 0.0214 ** | 0.0003 * |

**Figure 6. The relationship between interface involvement and evolutionary rate for residues in PPI interfaces. (A)-(D)** Linear fits in blue between binned measures of interface involvement and evolutionary rate (*dN/ dS*), for change in burial upon complex formation (ΔRSA), inter-protein residue-residue contacts (interRRC), distance from interface center (dCenter), and distance from interface edges (dEdges) respectively. Distributions of the number of residues per bin, weighted linear regression lines, and $R^2$ values of the fits are also shown. In addition to the observed fit in blue, we also show the expected fit in yellow, assuming that interfacial burial and non-interfacial burial are selectively equivalent and that interfacial residues are subject to the same evolutionary constraints as non-interfacial residues with the same total burial in PPIs. For each panel, the average *dN/dS* value for non-interfacial residues (average *dN/dS* = 0.0901) is marked by a yellow "X". **(E)** Results of t-tests for differences in slope and intercept between observed and expected fits in **(A)-(D)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**), values significant at the P-value < 0.05 level are denoted with a single asterisk (*).

We used *ConSurf* scores as proxies for evolutionary rate, as *ConSurf* scores can be calculated for individual interfacial residues, which is preferable to binned *dN/dS* for training multiple regression models. Multiple linear regression is the natural choice here as we want to predict a continuous dependent variable (*ConSurf* score of residues) by integrating a set of heterogeneous features (structural measures of interface involvement described in **Figure 2**). We used all interfacial residues with an available *ConSurf* score and structural measures of interface involvement (17,443 residues) and 10-fold cross-validation to train a multiple linear regression model using various residue structural properties, aiming to compare the relative strengths of the different residue structural properties for predicting residue evolutionary rate in PPI interfaces.

**Table 2 Regression results for different models aiming to predict residue evolutionary rate (ConSurf score) from structural properties in PPI interfaces.** All models were trained using 10-fold cross-validation, and the results shown here are average adjusted $R^2$ values across all cross-validation trials.

| Structural properties included in the model | Linear regression with *ConSurf* score $R^2$ |
|---|---|
| Monomer RSA | 8.93% |
| Monomer RSA + ΔRSA | 15.50% |
| Monomer RSA + interRRC | 11.64% |
| Monomer RSA + dCenter | 10.92% |
| Monomer RSA + dEdges | 9.04% |
| Monomer RSA + ΔRSA + interRRC + dCenter + dEdges | 16.27% |

The first row in **Table 2** shows our baseline model, where we predict interfacial residue evolutionary rates using monomer RSA values only, operating under the assumption that interfacial involvement plays no role in constraining interfacial residue evolution. This baseline model gives us an $R^2$ value of 8.9%. All subsequent rows in **Table 2** show adjusted $R^2$ values for

models integrating the baseline model with measures of interface involvement for evolutionary rate prediction. Adjusted $R^2$s take the number of predictors in a model into consideration before estimating the goodness-of-fit of the regression analysis, to avoid overfitting. Therefore, adjusted $R^2$ values will only increase from the baseline model if the addition of one or more structure-based measures of interface involvement to the initial model increases prediction fit significantly more than expected by chance (i.e., if each additional structural property contributes independent information to evolutionary rate prediction).

Among models integrating the baseline model (monomer RSA) with a single structural measure of interface involvement (**Table 2**, rows 2–5), the model integrating monomer RSA with ΔRSA has the highest adjusted $R^2$ value (adjusted $R^2$ = 15.50%). The prediction performance of the model increases from 8.9% to 15.50% in adjusted $R^2$ following the addition of ΔRSA to our baseline, indicating that interface involvement – as measured by ΔRSA – makes a significant contribution to evolutionary rate prediction in PPI interfaces. Moreover, the increase in $R^2$ achieved by including ΔRSA in the model is similar in magnitude to the contribution imposed by residue burial in the monomer state (monomer RSA). In contrast, models integrating our baseline with each of the other three measures of interface involvement show smaller increases in adjusted $R^2$ (2.71%, 1.99%, and 0.11% increase in adjusted $R^2$ value for the addition of interRRC, dCenter, dEdges respectively). Therefore, compared to ΔRSA, each of these three other measures of interface involvement make a minor contribution to evolutionary rate prediction in PPI interfaces. Moreover, since adjusted $R^2$ values increase when integrating our baseline model with each of the four measures of interface involvement, their individual contributions are all distinct from the constraints imposed by residue burial in the monomer state.

The highest adjusted $R^2$ value (adjusted $R^2$ = 16.27%) is achieved when all four measures of interface involvement are added to our baseline model (**Table 2**, row 6). Overall, to predict evolutionary rates in PPI interfaces, simply considering non-interfacial structural constraints (monomer RSA) yields an $R^2$ value of 8.9%, including the ΔRSA estimate of the degree of interface involvement, increases the adjusted $R^2$ value to 15.50%, and the addition of the other three measures of interface involvement (interRRC, dCenter, and dEdges) contributes a further increase in adjusted $R^2$ value smaller than 1%. We conclude that interface involvement is a significant constraint on residue evolutionary rate in PPI interfaces, and ΔRSA is the dominant structural measure of interface involvement in PPIs for constraining interface evolution, with minor contributions from other measures.

### 3.4 Discussion

In this work, we carried out a systematic and quantitative analysis of protein–protein interactions (PPIs) structures in *S. cerevisiae*, aiming to uncover structural determinants of PPI interface evolution at the residue level. We found that residue burial in a co-complexed PPI (as measured by complex RSA) better predicts interfacial residue evolutionary rate than residue burial in the monomer state (as measured by monomer RSA) and that interfacial burial (as measured by ΔRSA) is selectively equivalent to non-interfacial burial. These results are surprising because, while stable, permanent PPI complexes may be over-represented in our data as they are easier to study experimentally [42], many proteins investigated in our study are involved in transient PPIs and can mostly be found as free-floating monomers in cells. For all molecular interactions (and especially transient interactions), interfacial residues are buried in a PPI complex only when the two protein members of a PPI come together as a complex, but remain surface residues, exposed

to solvent, when the two proteins are separate monomers. One could, therefore, reasonably assume that the structural constraints on interfacial residues' evolution are a mixture between constraints imposed in the monomer state and constraints imposed in the complexed state. In this work, however, we discover that the evolutionary behavior of interfacial buried residues mainly resembles the behavior of non-interfacial buried residues, and not the behavior of non-interfacial surface residues, indicating that interfacial residues are mainly constrained by structure when in complexed state. The dominant role of the complexed state (rather than the monomeric state) in constraining interface evolution likely reflects the importance of maintaining proper PPI function and stability, as disruption and mis-regulations of PPIs are known to have dire consequences for organismal fitness [12–14].

Our second surprising finding is that, while interfacial burial is selectively equivalent to non- interfacial burial, considering interfacial buried residues to behave similarly to non-interfacial buried residues does not fully explain the low evolutionary rate of interfacial residues. Indeed, interfacial residues evolve significantly more slowly than non-interfacial residues, even after controlling for total residue burial (as measured by complex RSA). Moreover, we investigated several structure-based measures of the degree of involvement of a residue in an interface, and consistently observed a significant difference in evolutionary rate across the interface boundary, going from non-interfacial residues to interfacial residues that are marginally involved in the interface, even though there are no large changes in other structural properties (such as residue total burial) across the interface boundary. Hence, this jump in evolutionary conservation from non-interfacial residues to interfacial residues cannot be fully explained by differences in structural properties (such as residue total burial) between the interface and the rest of the protein. These

observations are evidence of a fixed evolutionary constraint, associated with the function of the interface, and independent of a residue's degree of interfacial involvement [27].

In addition to this fixed, function-based, evolutionary constraint on any interfacial residue, we found evidence of structure-based constraints within PPI interfaces, scaling continuously with a residue's degree of interfacial involvement. In particular, we found significant, monotonic, and continuous relationships between interfacial residue evolutionary rate and four structure-based measures of residue interfacial involvement (ΔRSA, interRRC, dCenter, and dEdges), with residues more involved in the interface evolving progressively more slowly on average. Among the four structure-based measures of residue interfacial involvement, ΔRSA plays a dominant role in predicting interfacial residue evolutionary rate, with independent yet minor contributions from other measures. Surprisingly, despite its simplicity and local nature, interfacial burial (ΔRSA) significantly outperforms other local and non-local measures of interface involvement in evolutionary rate prediction within the interface. It is known that the choices of sequences included in the alignment can have a large effect on evolutionary rate values [25]. We have thus repeated our analyses using different criteria for including more species and more sequences. While these additional analyses yield results that are, in general, consistent with our main conclusions, inclusion of more species and more sequences does not necessarily lead to better results due to the highly lineage- specific nature of PPI structure and evolution (**Supplementary Materials Analyses S1-S3, Figures S3- S6, Tables S1-S2**). Furthermore, we know that solvent accessibility calculations are specific to side chains [43]. RSA calculations in our analysis could, therefore, be affected when computed and transferred from a PDB structure with low sequence identity to a yeast protein sequence. We have thus repeated our analysis excluding all PPIs for which either of

the two partner proteins has sequence identity lower than 50% between their yeast protein sequence and the PDB protein sequence used to compute structural properties. All conclusions remain unchanged (**Supplementary Materials Analyses S4, Figures S9-S11, Tables S3-S4**).

The data curation process used to construct this large dataset of PPIs with both sequence and structure information comes with several caveats. Estimates of the proportion of known protein– protein interactions in *S. cerevisiae* suggest that 50% of yeast PPIs have been identified [31]. Current yeast PPI networks are, therefore, a sample of the complete network, and the PPI data used in this analysis comes with the set of biases typically associated with PPI interaction measurements. Our data and results may be biased towards proteins from particular cellular environments, more ancient and conserved proteins, commonly studied proteins, and highly expressed proteins [44]. Furthermore, this analysis relied on homology-based structural annotation transfer to gather structural data for a large dataset of PPIs. Due to the relatively small number of solved protein structures compared to the set of known protein sequences, homology-based structural annotation transfer was necessary, but could further bias our dataset toward easily structured and often well-ordered proteins. We further assume that differences at the sequence level among close homologs do not produce measurable structural differences and align yeast homologs to the same three-dimensional structure. But even when the sequence-structure alignment is perfect, we cannot be fully confident that a given homology-mapped structure accurately reports on in-vivo properties of its residues [27]. However, we believe that that the data curation process used here is still the best existing and the most reliable method for integrating structural details with molecular evolutionary properties of PPIs on a proteomic scale. Moreover, this method will only improve as the spaces of known PPIs and known structures grow.

In summary, we have presented several strong, proteome-wide relationships between residue- level structural properties of PPI interfaces and evolutionary rate in yeast. Moreover, the results found here have broader significance as they yield insight into the evolutionary design principles of interfaces and the identification of some of their key features, improving our understanding of the role that structure plays in the molecular evolution of PPIs at the residue level. Our study also has implications for interfacial residue prediction. While residue evolutionary rate by itself is a weak predictor for interfacial versus non-interfacial residues, residue evolutionary rate information can be used in combination with other weak predictors (based on sequence, structure, and co- evolutionary information, among others) to boost accuracy in predicting whether a residue is interfacial or not. As more PPI and structural data become available, future work in additional species could further help in our understanding and identification of key interfacial residues mediating molecular interactions. Such residues, under unique evolutionary pressures, even within the already constrained context of PPI interfaces, could be useful in the development of drugs aiming to target specific PPI interfaces [12,36,45], the prediction of existing PPIs [46–47] or the design of novel protein–protein interactions [48–49].

## 3.5 Materials and Methods

### Homology-based structural annotation transfer

First, we curated a high confidence set of physical interactions between *Saccharomyces cerevisiae* (*S. cerevisiae*) proteins: we filtered the most recent release of the BioGRID database (April 2020) for physical PPIs reported in *S. cerevisiae* by two or more independent experiments (determined by different PubMed IDs), yielding 23,272 high confidence PPIs between 4,321 *S. cerevisiae* proteins [50–51]. We then individually mapped the proteins involved in the aforementioned

PPIs to three-dimensional structures by performing gapped BLAST [52] searches under default settings between (i) a database built from the proteins' translated open reading frame sequences (ORFs) obtained on Ensembl [53] and (ii) 510,817 biological unit structure subunit sequences from the Protein Data Bank (PDB) [54]. For each ORF in the database, we constructed a list of potential structural matches by selecting biological unit structures which (i) produced E-values below a 1 105 cut- off in the alignment, (ii) had high coverage (>50%) in the alignment for both the ORF and the subunit sequence, and (iii) showed no inconsistencies (e.g., insufficient atomic detail, unreasonable distances between alpha-carbons, non-sensible heavy atom counts). We found 2,212 *S. cerevisiae* proteins involved in our high confidence set of PPIs with at least one biological unit structure mapped to their ORF meeting those initial conditions, and an average of 11 structures mapped to each protein. We further excluded all biological unit structures annotated as "low" or "very low" confidence on the QSbio database [55] as those structures could be doubtful biological assembly assignments. Moreover, only 25% of the structures mapped in this process were annotated as yeast protein structures; all other mappings are, therefore, obtained from between-species homology. Finally, to select the best structural match to each interacting pair of *S. cerevisiae* proteins, we looked at the list of potential structural matches for each partner protein in a high confidence PPI, and retained only the one which (i) met our initial alignment conditions above for both protein partners, (ii) showed the two protein partners in physical contact (i.e., mapped to spatially adjacent chains in the structure), (iii) had the highest composite coverage (sum of the coverage for each partner protein) in the BLAST alignment, and (iv) had a resolution better than 3 Å. If more than one potential structure remained for the PPI following the above process, the structure with the best resolution was kept. We further note that no homology modeling was performed in this analysis: the structures curated as best structural matches for *S. cerevisiae* PPIs

were all obtained directly from the PDB, using the process referred to as homology-based structural annotation transfer above. Structures that are not annotated as yeast structures on the PDB but are, nonetheless, the best structural match for a known *S. cerevisiae* PPI were taken as is, assuming that with high sequence conservation between two known PPIs, structural conservation must also be high. This homology-based structural mapping pipeline yielded structural models for 718 PPIs between 611 *S. cerevisiae* proteins and is illustrated in **Figure 1** and **Figure S7**. The code pipeline, as well as curated data from this analysis, are available in a GitHub repository.

**Calculation of structural properties at the residue level**

**Figure 2** specifies basic definitions of the structural properties used here. Solvent Accessible Surface Area (SASA) was calculated using the DSSP program [56–57] with hydrogen atoms excluded. SASA values were normalized using reliable normalization values from Tien et al [58]. to produce Relative Solvent Accessibility (RSA). For each residue in our structural models, two values of RSA were computed: monomer RSA, which was calculated using the structure of monomeric proteins (discarding the chain mapped to the partner protein in a structure), and complex RSA, which was obtained from the co-complexed structure of both protein partners (PPI structure). ΔRSA, the change in residue burial upon complex formation, was computed as the difference between monomeric and co-structured RSA values for each residue in the structural models (ΔRSA = monomer RSA – complex RSA). ΔRSA was subsequently used in the definition of interfaces: any residue with a change in burial upon complex formation (ΔRSA ≠ 0) was defined as an interfacial residue. For interfacial residues in our structural models, inter-protein Residue-Residue Contacts (interRRC) number was calculated by adapting the Residue–Residue Contacts (RRC) definition from the single protein literature [28–29] to the context of PPIs: interRRCs were

taken as the subset of all RRCs for a residue which occurs between the chains mapped to two PPI partner proteins, ignoring contacts within the same protein. Distance from interface center (dCenter) was computed for all interfacial residues as the Euclidian distance between the residue and the geometric center of the interface it belongs to. To calculate distance from interface edges (dEdges), residues in the 90th percentile of largest values of complex RSA for each interface (most exposed interfacial residues upon complex formation) were assigned a distance of zero and defined as belonging to the edges of the interface. Distances for all other interfacial residues were calculated as the Euclidian distance to the closest edge residue.

**Evolutionary sequence analysis**

Estimating residue-level evolutionary rates is a non-trivial task and, thus, various methods have been proposed for this inference in the literature [3,8,34]. Here, we used an established technique that infers rates from codon data: ω = dN/dS ratio [59–60]. One additional measure of site-specific evolutionary rates, *ConSurf* score, which uses amino acid data for rate inference was also investigated to confirm results obtained from codon data [61]. Conclusions obtained using these two techniques are typically correlated [62].

For each protein involved in a high confidence *S. cerevisiae* PPI, we generated alignments using ClustalW [63] between (i) its translated ORF, (ii) the sequence of its mapped protein structure subunit, and (iii) orthologous translated ORFs in *Saccharomyces paradoxus* (*S. paradoxus*), *Saccharomyces mikatae* (*S. mikatae*), *Saccharomyces bayanus* (*S. bayanus*), *Naumovozyma castellii* (*N. castellii*), *Candida glabrata* (*C. glabrata*), *Eremothecium gossypii* (*E. gossypii*), *Kluyveromyces lactis* (*K. lactis*) and *Candida albicans* (*C. albicans*) obtained from Ensembl [53].

We reconstructed the codon alignments between the nine genomic yeast sequences using the protein alignments as guides. This alignment process is illustrated in **Figure S7**.

The ω = *dN/dS* ratio compares the rate of amino acid changing substitutions (*dN*) with the rate of silent substitutions (*dS*) at the codon level. The former is presumed to be more selectable than the latter (thus *dS* acts as a normalization factor). To compute ω = *dN/dS*, we concatenated codons within each yeast species into groups binned according to whether they are interfacial or non-interfacial, or in uniform intervals of a given structural property (summarized in **Figure 2**). We then calculated a single ω = *dN/dS* value for a group using the codeml program within the PAML software package [60]. We considered a single *dN/dS* value for the entire tree, which we specified as [[[[[*S. cerevisiae*, *S. paradoxus*], [*S. mikatae*, *S. bayanus*]], *C. glabrata*], *N. castellii*], [[*E. gossypii*, *K. lactis*], *C. albicans*]] following previous work [27,64–65]. The final value of *dN/dS* and associated error for a group of codons was taken as the average and standard deviation of *dN/dS* values obtained for 100 bootstrap resamples (with replacement) from the original codons in the group. Other parameters in codeml were left to their default values. *ConSurf* score estimates of residue evolutionary rates were computed from amino acid data using the Rate4Site program [61,66]. This method computes relative conservation scores for each site in a protein using empirical Bayesian methods and a multiple sequence alignment of homologous sequences to essentially rank residues from most to least conserved within a protein. The *ConSurf* score obtained from running the program is lower for more conserved residues, and higher for less conserved ones. We ran *ConSurf* score calculations for all amino acids in our models using the Rate4Site program, and the same closely related species, and phylogenetic tree mentioned above for *dN/dS* calculations. Other parameters in Rate4site were left to their default values. When binning *ConSurf* scores for plotting

in this analysis, the final *ConSurf* scores value and associated error for a group of residues were taken as the average and standard error of *ConSurf* score values for each residue in the bin.

## Statistical analysis

### Correlations between residue structural properties.

We calculated pairwise correlations between each residue structural property in R using the Visualization of a Correlation Matrix ('corrplot') package [67]. The results in **Figure 5** show pairwise Pearson correlation coefficients between different residue structural properties. Correlations between measures of RSA (monomer RSA, complex RSA, and ΔRSA) are computed for all residues in our models. Correlations with the other structural properties (interRRC, dCenter, dEdges) are computed for all interfacial residues in our models.

### Correlations with evolutionary rate

We computed Pearson correlation coefficients between structural properties and measures of evolutionary rate for interfacial residues in our models (**Figure 4(A)**, **Table 1**). Correlations in the first 2 rows of both **Figure 4(A)** and **Table 1** are standard Pearson correlation obtained in R using the Feature Selection (Including Multiple Solutions) and Bayesian Networks ('MXM') package [68]. As *dN/dS* values are computed for binned residues, all correlations with *dN/dS* were made with the center of each bin. Correlations labeled as "(weighted)" are weighted Pearson correlation, using the standard error on each *dN/dS* and *ConSurf* score value, respectively, to weigh the correlation analysis and were obtained from the Weighting and Weighted Statistics ('weights') package [69]. The structural and evolutionary data used in this analysis does not follow a normal

distribution, therefore, significance for each correlation coefficient (as shown in **Figure 4 (A)** and **Table 1**) was determined from 1,000 rounds of randomizing permutations.

**Modeling the relationship between individual structural properties and dN/dS**

We studied the relationship between individual structural properties and *dN/dS* in PPI interfaces using a weighted least-square regression technique that takes the error associated with calculating *dN/dS* for each residue bin into account [27,70–71]. One advantage of this approach is that residue bins with small *dN/dS* estimation errors receive greater weight in the line fitting process. This technique has been used in previous literature and was adapted in R. The regression model takes the following form:

$$y(x) = w_0 + w_1 x_1 + e_x$$

where $y(x)$ is the *dN/dS* score of residues in bin $x$ (binned according to similar structural property as previously described), $x_1$ is the center value of bin $x$ for the structural property investigated, $w_0$, $w_1$ are the intercept, and the weight associated with the structural feature in the regression model, and $e_x$ is a random variable ("noise term") following a Gaussian distribution with zero mean and standard deviation equal to the standard error associated with the *dN/dS* score for bin $x$. This method also reports a standard error for the slope and intercept of the resulting linear fit which we used in t-tests to compare slope and intercept across different fits (**Figure 3(D)**, **Figure 4(C)**, **Figure 6(E)**). One model was trained for each structural property, and the resulting linear fits can be seen in **Figures 3, 4, and 6**.

**Integrated modeling of the relationship between structural properties and evolutionary rate**

We investigated the combined influences of residue structural properties on evolutionary rate (*ConSurf* score) in PPI interfaces using a weighted multiple linear regression technique, again aiming to take the error associated with calculating *ConSurf* scores into account [72]. The regression model was implemented in R with the Classification and Regression Training ('caret') package [73] and takes the following form:

$$y(x) = w_0 + w_1 x_1 + \cdots + w_n x_n + e_x$$

where $y(x)$ is the *ConSurf* score of residue $x$ (value normalized within each protein so that the average score for all residues is zero, and the standard deviation is one, with low scores associated with the most conserved positions in a protein), $x_1, \ldots, x_n$ are the values of each structural property investigated for residue $x$, ), $w_0, \ldots, w_n$ are the intercept and weights associated with the structural features, and $e_x$is a random variable ("error term") following a Gaussian distribution with zero and standard deviation equal to the standard error associated with the *ConSurf* score for residue $x$. The model is trained using a 10-fold cross-validation process: the set of residues is randomly partitioned into 10 subsamples of equal size. A single subsample is retained as validation data for testing the model and the remaining 9 subsamples are used as training data. This cross-validation process is repeated 10 times so that each of the 10 subsamples is used exactly once as validation data. The advantage of this method is that all observations are used for both training and validation and each observation is used for validation exactly once [74]. The overall performance of a model is taken as the average performance across all cross-validation trials and compared across models including different subsets of structural properties (**Table 2**).

## Code Availability

The code pipeline used to construct structural models of *S. cerevisiae* PPIs is available as a GitHub repository: https://github.com/LeahPollet/ interface_structure_evolution. Curated data underlying this article, including a list of high confidence PPIs in *S. cerevisiae*, and homology mapped PDB structures for *S. cerevisiae* PPIs are also available in the GitHub repository and can be accessed using the following https://doi.org/10. 5281/zenodo.4737637.

## CRediT authorship contribution statement

**Léah Pollet**: Data curation, Formal analysis, Software, Writing – original draft. **Luke Lambourne**: Conceptualization, Methodology, Investigation. **Yu Xia**: Funding acquisition, Conceptualization, Supervision, Writing – review & editing.

## DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.


## Acknowledgement and Funding Information

This work was supported by Natural Sciences and Engineering Research Council of Canada grants RGPIN-2019-05952 and RGPAS-2019-00012, Canada Foundation for Innovation grants JELF-33732 and IF- 33122, and Canada Research Chairs program.

## 3.6 References


1. Kimura, M., Ohta, T., (1974). On Some Principles Governing Molecular Evolution. Proc. Natl. Acad. Sci. U. S.A. 71 (7), 2848–2852.
2. Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., Feldman, M.W., (2002). Evolutionary Rate in the Protein Interaction Network. Science 296 (5568), 750–752.
3. Zhang, J., Yang, J.-R., (2015). Determinants of the rate of protein sequence evolution. Nat. Rev. Genet. 16 (7), 409– 420.
4. Goldstein, R.A., (2008). The structure of protein evolution and the evolution of protein structure. Curr. Opin. Struct. Biol. 18 (2), 170–177.
5. Liberles, D.A., Teichmann, S.A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L.J., De Koning, A. J., et al., (2012). The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 21 (6), 769–785.
6. Hubbard, T.J.P., Blundell, T.L., (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. Protein Eng. Des. Sel. 1 (3), 159–171.
7. Bartlett, G.J., Porter, C.T., Borkakoti, N., Thornton, J.M., (2002). Analysis of Catalytic Residues in Enzyme Active Sites. J. Mol. Biol. 324 (1), 105–121.
8. Echave, J., Spielman, S.J., Wilke, C.O., (2016). Causes of evolutionary rate variation among protein sites. Nat. Rev. Genet. 17 (2), 109–121.
9. Cohen, P., (2000). The regulation of protein function by multisite phosphorylation – a 25 year update. Trends Biochem. Sci. 25 (12), 596–601.
10. Jones, S., Thornton, J.M., (1996). Principles of protein- protein interactions. Proc. Natl. Acad. Sci. U.S.A. 93 (1), 13–20.
11. De Las Rivas, J., Fontanillo, C., (2010). Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks. PLoS Comput. Biol. 6 (6), e1000807.
12. Ryan, D., Matthews, J., (2005). Protein-protein interactions in human disease. Curr. Opin. Struct. Biol. 15 (4), 441–446.
13. Zhong, Q., Simonis, N., Li, Q.R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., et al., (2009). Edgetic perturbation models of human inherited disorders. Mol. Syst. Biol. 5 (1), 321.
14. Schoenrock, A., Burnside, D., Moteshareie, H., Pitre, S., Hooshyar, M., Green, J.R., Golshani, A., Dehne, F., et al., (2017). Evolution of protein-protein interaction networks in yeast. PLoS ONE 12 (3), e0171920.
15. Tonddast-Navaei, S., Skolnick, J., (2015). Are protein- protein interfaces special regions on a protein's surface?. J. Chem. Phys. 143 (24), 12B631_1.
16. Engin, H.B., Kreisberg, J.F., Carter, H., (2016). Structure- Based Analysis Reveals Cancer Missense Mutations Target Protein Interaction Interfaces. PLoS ONE 11 (4), e0152929.
17. Jubb, H.C., Pandurangan, A.P., Turner, M.A., Ochoa- Montan˜ o, B., Blundell, T.L., Ascher, D.B., (2017). Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. Prog. Biophys. Mol. Biol. 128, 3–13.
18. Valdar, W.S., Thornton, J.M., (2000). Protein-protein interfaces: Analysis of amino acid conservation in homodimers. Proteins 42 (1), 108–124.
19. Ma, B., Elkayam, T., Wolfson, H., Nussinov, R., (2003). Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc. Natl. Acad. Sci. U.S.A. 100 (10), 5772–5777.

20. Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., Huang, E.S., (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci. 13 (1), 190–202.
21. Eames, M., Kortemme, T., (2007). Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. Structure 15 (11), 1442–1451.
22. Chakrabarti, P., Janin, J., (2002). Dissecting protein- protein recognition sites. Proteins 47 (3), 334–343.
23. Guharoy, M., Chakrabarti, P., (2005). Conservation and relative importance of residues across protein-protein interfaces. Proc. Natl. Acad. Sci. U.S.A. 102 (43), 15447– 15452.
24. Levy, E.D., (2010). A simple definition of structural regions in proteins and its use in analysing interface evolution. J. Mol. Biol. 403 (4), 660–670.
25. Duarte, J.M., Srebniak, A., Scha¨rer, M.A., Capitani, G., (2012). Protein interface classification by evolutionary analysis. BMC Bioinf. 13 (1), 1–16.
26. Schärer, M.A., Grütter, M.G., Capitani, G., (2010). CRK: an evolutionary approach for distinguishing biologically relevant interfaces from crystal contacts. Proteins: Struct. Funct. Genet. 78 (12), 2707–2713.
27. Franzosa, E.A., Xia, Y., (2009). Structural Determinants of Protein Evolution Are Context-Sensitive at the Residue Level. Mol. Biol. Evol. 26 (10), 2387–2395.
28. Franzosa, E.A., Xia, Y., (2008). Structural Perspectives on Protein Evolution. Annu Rep Comput Chem. 4 (1), 3–21.
29. Bloom, J.D., Drummond, D.A., Arnold, F.H., Wilke, C.O., (2006). Structural Determinants of the Rate of Protein Evolution in Yeast. Mol. Biol. Evol. 23 (9), 1751–1761.
30. Zhou, T., Drummond, D.A., Wilke, C.O., (2008). Contact density affects protein evolutionary rate from bacteria to animals. J. Mol. Evol. 66, 395–404.
31. Hakes, L., Pinney, J.W., Robertson, D.L., Lovell, S.C., (2008). Protein-protein interaction networks and biology— what's the connection? Nat. Biotechnol. 26 (1), 69–72.
32. Stumpf, M.P.H., Thorne, T., de Silvia, E., Stewart, R., Jun An, H., Lappe, M., Wiuf, C., (2008). Estimating the size of the human interactome. Proc. Natl. Acad. Sci. U.S.A. 105 (19), 6959–6964.
33. Choi, S.S., Vallender, E.J., Lahn, B.T., (2006). Systematically Assessing the Influence of 3-Dimensional Structural Context on the Molecular Evolution of Mammalian Proteomes. Mol. Biol. Evol. 23 (11), 2131– 2133.
34. Pa´l, C., Papp, B., Lercher, M.J., (2006). An integrated view of protein evolution. Nat. Rev. Genet. 7 (5), 337–348.
35. Shahmoradi, A., Sydykova, D.K., Spielman, S.J., Jackson,
E.L., Dawson, E.T., Meyer, A.G., Wilke, C.O., (2014). Predicting Evolutionary Site Variability from Structure in Viral Proteins: Buriedness, Packing, Flexibility, and Design. J. Mol. Evol. 79 (3–4), 130–142.
36. Moreira, I.S., Fernandes, P.A., Ramos, M.J., (2007). Hot spots-A review of the protein-protein interface determinant amino-acid residues. Proteins 68 (4), 803–812.
37. David, A., Sternberg, M.J.E., (2015). The Contribution of Missense Mutations in Core and Rim Residues of Protein- Protein Interfaces to Human Disease. J. Mol. Biol. 427 (17), 2886–2898.
38. Kim, P.M., Lu, L.J., Xia, Y., Gerstein, M.B., (2006). Relating three-dimensional structures to protein networks provides evolutionary insights. Science 314 (5807), 1938–1941.

39. Ramsey, D.C., Scherrer, M.P., Zhou, T., Wilke, C.O., (2011). The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution. Genetics 188 (2), 479–488.
40. Zhou, H., Zhou, Y., (2003). Quantifying the effect of burial of amino acid residues on protein stability. Proteins 54 (2), 315–322.
41. DePristo, M.A., Weinreich, D.M., Hartl, D.L., (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet. 6 (9), 678–687.
42. Perkins, J.R., Diboun, I., Dessailly, B.H., Lees, J.G., Orengo, C., (2010). Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. Structure 18 (10), 1233–1243.
43. Lesk, A.M., Chothia, C., (1980). Solvent accessibility, protein surfaces, and protein folding. Biophys. J. 32 (1), 35–47.
44. Hart, G.T., Ramani, A.K., Marcotte, E.M., (2006). How complete are current yeast and human protein-interaction networks? Genome Biol. 7 (11), 1–9.
45. Tuncbag, N., Gursoy, A., Keskin, O., (2011). Prediction of protein-protein interactions: unifying evolution and structure at protein interfaces. Phys. Biol. 8 (3), 035006.
46. Singh, R., Park, D., Xu, J., Hosur, R., Berger, B., (2010). Struct2Net: a web service to predict protein-protein interactions using a structure-based approach. Nucleic Acids Res. 38 (suppl_2), W508–W515.
47. Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., et al., (2012). Structure- based prediction of protein-protein interactions on a genome-wide scale. Nature 490 (7421), 556–560.
48. Kortemme, T., Baker, D., (2004). Computational design of protein-protein interactions. Curr Opin Chem Biol. 8 (1), 91–97.
49. Mandell, D.J., Kortemme, T., (2009). Computer-aided design of functional protein interactions. Nat. Chem. Biol. 5 (11), 797–807.
50. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M., (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34 (suppl_1), D535–D539.
51. Chatr-aryamontri, A., Breitkreutz, B.-J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., et al., (2014). The BioGRID interaction database: 2015 update. Nucleic Acids Res. 43 (D1), D470– D478.
52. Altschul, S.F., Madden, T.L., Scha¨ffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402.
53. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., et al., (2019). Ensembl 2020. Nucleic Acids Res. 48 (D1), D682– D688.
54. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., (2000). The Protein Data Bank. Nucleic Acids Res. 28 (1), 235– 242.
55. Dey, S., Ritchie, D.W., Levy, E.D., (2018). PDB-wide identification of biological assemblies from conserved quaternary structure geometry. Nat. Methods 15 (1), 67– 72.
56. Kabsch, W., Sander, C., (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen- bonded and geometrical features. Biopolymers 22 (12), 2577–2637.
57. Joosten, R.P., Te Beek, T.A., Krieger, E., Hekkelman, M. L., Hooft, R.W., Schneider, R., Sander, C., Vriend, G., (2010). A series of PDB related databases for everyday needs. Nucleic Acids Res. 39 (suppl_1), D411–D419.

58. Tien, M.Z., Meyer, A.G., Sydykova, D.K., Spielman, S.J., Wilke, C.O., (2013). Maximum allowed solvent accessibilites of residues in proteins. PLoS ONE 8 (11), e80635.
59. Goldman, N., Yang, Z., (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11 (5), 725–736.
60. Yang, Z., (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. Bioinformatics 13 (5), 555–556.
61. Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., Ben-Tal, N., (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18 (suppl_1), S71–S77.
62. Sydykova, D.K., Wilke, C.O., (2017). Calculating site- specific evolutionary rates at the amino-acid or codon level yields similar rate estimates. PeerJ. 5
63. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R., et al., (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47 (W1), W636–W641.
64. Wall, D.P., Hirsh, A.E., Fraser, H.B., Kumm, J., Giaever, G., Eisen, M.B., Feldman, M.W., (2005). Functional genomic analysis of the rates of protein evolution. Proc. Natl. Acad. Sci. U.S.A. 102 (15), 5483–5488.
65. Marcet-Houben, M., Gabaldo´n, T., (2015). Beyond the Whole-Genome Duplication: Phylogenetic Evidence for an Ancient Interspecies Hybridization in the Baker's Yeast Lineage. PLoS Biol. 13 (8), e1002220.
66. Mayrose, I., Graur, D., Ben-Tal, N., Pupko, T., (2004). Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior. Mol. Biol. Evol. 21 (9), 1781–1791.
67. Wei, T. & Simko, V. (2021). R package 'corrplot': Visualization of a Correlation Matrix. (Version 0.90), https://github.com/taiyun/corrplot.
68. Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., Tsamardinos, I., (2017). Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets. J. Stat. Softw. 80 (7)
69. Pasek, J., Tahk, A., Culter, G. & Schwemmle, M. (2020). weights: Weighting and Weighted Statistics. R package (Version 1.0.1), https://CRAN.R-project.org/package= weights.
70. Meer, P., Mintz, D., Rosenfeld, A., Kim, D.Y., (1991). Robust regression methods for computer vision: A review. Int. J. Comput. Vis. 6 (1), 59–70.
71. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., (2007). Numerical recipes: the art of scientific computing. Cambridge University Press, Cambridge.
72. Schaffrin, B., Wieser, A., (2007). On weighted total least- squares adjustment for linear regression. J. Geod. 82 (7), 415–421.
73. Kuhn, M., (2008). Caret package. J. Stat. Softw. 28 (5)
74. Arlot, S., Celisse, A., (2010). A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79.

## Preface to Chapter 4

Chapter 3, and the article discussed therein, focuses on utilizing the wealth of newly available data on protein-protein interaction (PPI), including species-level interactome maps of all interacting proteins in a given species, and molecular-level detailed structural data on individual PPIs, to study variations in PPIs within a species. The findings and conclusions from this work give us critical insights into the structural properties and evolutionary constraints that shape interactions between proteins within a species. However, this large amount of PPI data can also be used to examine variations in PPIs (or PPI rewiring) between species, an investigation that is essential to fully understand PPIs and PPI evolution. This is the focus of Aim 3 in this thesis.

**Aim 3** focuses on the comparison of PPIs between *Saccharomyces cerevisiae* (*S. cerevisiae*) and *Schizosaccharomyces pombe* (*S. pombe*) to uncover possible drivers for observed differences in interactomes between the two yeasts. No large-scale analysis of the detailed molecular evolutionary mechanisms underlying interactome network rewiring on the genomic scale has previously been performed in the literature. As such, this aim uncovers some of the molecular mechanisms behind the phylogenetic loss or gain of an interaction between two species.

The article included here covers the use of molecular models of PPIs in *S. cerevisiae* and *S. pombe* to classify PPIs according to whether they are preserved or different between the two yeast species. Site-specific evolutionary rates for residues in these different categories of PPIs are then compared. The evolution of PPI interfaces is considered more specifically, as this region of contact between interacting proteins could be particularly important to PPI evolution and PPI rewiring between species. Overall, the following important conclusions were reached: (i) residues

in PPI interfaces evolve significantly more slowly than non-interfacial residues when using lineage-specific measures of evolutionary rate, but not when using non-lineage-specific measures, (ii) both lineage-specific and non-lineage-specific evolutionary rate measures can distinguish interfacial residues from non-interfacial residues for preserved PPIs between the two yeasts, but only the lineage-specific measure is appropriate for PPIs that are different between the two yeasts, (iii) both lineage-specific and non-lineage-specific evolutionary rate measures are appropriate for elucidating structural determinants of protein evolution for residues outside of PPI interfaces. These findings demonstrate that, PPIs and PPI interfaces can be highly volatile in their evolution, thus requiring the use of lineage-specific measures when studying their evolution. The article also helps establish some of the evolutionary design principles and mechanisms that influence the evolution of PPIs between species.

# 4. Research Article No. 2: Structure-guided evolutionary analysis of interactome network rewiring at single residue resolution in yeasts

Léah Pollet and Yu Xia *

Department of Bioengineering, Faculty of Engineering, McGill University, Montreal, QC, Canada

Correspondence to Yu Xia: brandon.xia@mcgill.ca (Y. Xia)

https://doi.org/10.1016/j.jmb.2024.168641

Edited by Michael Sternberg

## 4.1 Abstract

Protein-protein interactions (PPIs) are known to rewire extensively during evolution leading to lineage- specific and species-specific changes in molecular processes. However, the detailed molecular evolutionary mechanisms underlying interactome network rewiring are not well-understood. Here, we combine high-confidence PPI data, high-resolution three-dimensional structures of protein complexes, and homology-based structural annotation transfer to construct structurally-resolved interactome networks for the two yeasts *S. cerevisiae* and *S. pombe*. We then classify PPIs according to whether they are preserved or different between the two yeast species and compare site-specific evolutionary rates of interfacial versus non-interfacial residues for these different categories of PPIs. We find that residues in PPI interfaces evolve significantly more slowly than non-interfacial residues when using lineage-specific measures of evolutionary rate, but not when using non-lineage-specific measures. Furthermore, both lineage-specific and non-lineage-specific evolutionary rate measures can distinguish interfacial residues from non-interfacial residues for preserved PPIs between the two yeasts, but only the lineage-specific measure is appropriate for rewired PPIs. Finally, both lineage-specific and non-lineage-specific evolutionary rate measures are appropriate for elucidating structural determinants of protein evolution for residues outside of PPI interfaces. Overall, our results demonstrate that unlike tertiary structures of single proteins, PPIs and PPI interfaces can be highly volatile in their evolution, thus requiring the use of lineage-specific measures when studying their evolution. These results yield insight into the evolutionary design principles of PPIs and the mechanisms by which interactions are preserved or rewired between species, improving our understanding of the molecular evolution of PPIs and PPI interfaces at the residue level.

## 4.2 Introduction

Proteins rarely work alone within cells, and instead often act through protein–protein interactions (hereafter referred to as PPIs) that are essential to proper cellular function [1]. For instance, PPIs are crucial to the regulation of protein function, as many proteins are regulated by transient interactions with protein kinases and other enzymes [2,3]. Moreover, stable protein complexes that act as molecular machines and involve numerous PPIs are necessary for most cellular processes [4–6]. Disruptions in PPIs can also lead to changes in organismal fitness, as numerous disease-causing mutations have been shown to disrupt protein interactions [7–11]. Accordingly, works in recent years have focused on large-scale mapping of PPIs in different species, yielding high-confidence interaction networks (also known as "interactome networks" or "interactomes") for various species [12–14], as well as detailed studies of many individual PPIs, including the determination of their three-dimensional (3D) molecular structures [15].

This wealth of data allows for better comparative analysis of PPIs between species. In previous work, Cesareni et al. (2005) compared interactome networks between yeast and drosophila, estimating that only 24% of yeast PPIs are present in fly [16]. Gandhi et al. (2006) compared interactomes between human, baker's yeast, worm, and fly, and found the overlap in protein interactions between the four species to be very small [17]. However, the small overlap between interactome networks from different species in these early studies may not be an accurate measure of the extent of network rewiring during evolution, as comparative interactomics analysis can be significantly affected by inaccuracies in early interactome datasets (both false positive and false negative errors). Recently, through careful consideration and controlling of both false positive and false negative errors in interactome datasets, and by focusing on comparing

interactions where both interacting protein partners have orthologs in two species, Vo et al. (2016) [13] estimated that only 40% of fission yeast PPIs were conserved in baker's yeast, and only 65% of fission yeast PPIs were conserved in human. These and other previous works [18,19] highlight that interactome networks undergo significant rewiring during evolution, via either gain or loss of proteins or interactions, leading to lineage-specific and species-specific changes in molecular processes and cellular functions. However, the detailed molecular evolutionary mechanisms underlying interactome network rewiring, and the site-specific selective pressures acting on rewired PPIs are not well-understood, especially on the genomic scale [20].

Several previous studies have attempted to elucidate molecular evolutionary mechanisms underlying PPI rewiring for specific types of molecular interactions. For example, Xin et al. (2013) [21], surveyed and compared interactions mediated by 79 SH3 domains in worm, baker's yeast, and human, and observed drastic rewiring between worm and yeast, attributing this rewiring to variations in the sequence of the motifs recognized by the SH3 domains, as well as changes in binding specificities between orthologous SH3 domains. Reinke et al. (2013) [22], experimentally compared interactions between 53 human bZIP proteins to their homologs in four other species including fly and worm and found significant rewiring, highlighting the plasticity of the bZIP interactome, which can be dramatically rewired with changes to just one or two amino acids. While those previous studies are crucial in furthering our understanding of the mechanisms of PPI rewiring for these two groups of PPIs, those are specific examples of PPIs and a large-scale analysis of the detailed molecular evolutionary mechanisms underlying interactome network rewiring on the genomic scale remains needed.

We, therefore, focus on two yeast species with large, high-quality protein interactome networks available in the literature: baker's yeast (*S. cerevisiae*) and fission yeast (*S. pombe*). The two species diverged from a common ancestor approximately 500 million years ago [23], and their genomes have since undergone significant changes, including gene duplications, deletions, and rearrangements [24]. Estimates of the extent of conservation of PPIs between the two yeasts in the literature range from 36.3%25 to 40% [13].Here, we combine high-confidence PPI data, high-resolution 3D structures of protein complexes, and homology-based structural annotation transfer to construct structurally-resolved interactome networks for the two yeasts. This allows us to compute and study the evolutionary rates of specific residue sites in the PPIs, rather than evolutionary rates for full proteins [20,25,26]. We focus on the region of contact between protein partners in a PPI, known as the interface. The interface is a critical feature of PPIs both structurally and functionally, as mutations to interfacial residues can lead to altered protein–protein binding affinities and improper PPI function, which can have implications for organismal fitness, and health [27,28]. Interfacial residues are also typically more evolutionarily conserved than non-interfacial residues, further establishing the critical role that they play in mediating the interactions between protein partners and ensuring proper PPI formation and function [25,29–32]. Furthermore, we classify PPIs according to whether they are preserved or different between the two yeast species, then compute and compare site-specific evolutionary rates of interfacial versus non-interfacial residues for these different categories of PPIs. Finally, we investigate the use of lineage-specific (*ConSurf-rate4site* scores computed from closely related species [33,34] and non-lineage-specific (*ConSurf-DB* scores from the *ConSurf* database [35] evolutionary rates in this work, as they may be particularly interesting to the study of PPI rewiring, a highly lineage-specific event.

We find that, in both *S. cerevisiae* and *S. pombe*, residues in PPI interfaces evolve significantly more slowly than non-interfacial residues when using lineage-specific measures of evolutionary rate (*ConSurf-rate4site*), but not when using non-lineage-specific measures of evolutionary rate (*ConSurf-DB*). Furthermore, only the lineage-specific evolutionary rate measure, but not the non-lineage-specific evolutionary rate measure, is able to distinguish interfacial residues from non- interfacial residues for PPIs that are rewired or different between the two yeasts. In contrast, both lineage-specific and non-lineage-specific evolutionary rate measures can distinguish interfacial residues from non-interfacial residues for PPIs that are preserved between the two yeasts. It is expected that non-lineage-specific evolutionary rates may not be appropriate to study certain types of PPIs. For instance, if a PPI only occurs in one specific lineage, using species outside of that lineage in evolutionary rate calculations may incorrectly estimate the selective pressures acting on some residues, especially those at the interface of contact between the two partner proteins. However, the magnitude of the difference in results obtained using lineage-specific versus non-lineage specific evolutionary rates in this work is truly surprising and shows that PPI rewiring, and the corresponding selective pressure on interfacial residues as measured by site-specific evolutionary rates, are highly lineage specific. Finally, both lineage-specific and non-lineage-specific evolutionary rate measures are appropriate for elucidating structural determinants of protein evolution for residues outside of PPI interfaces which are expected to be much less lineage-specific, such as residue burial or exposure to solvent.

Overall, this work is the first large-scale evolutionary rate analysis of PPI network rewiring between *S. cerevisiae* and *S. pombe* at the level of single residues. Our results demonstrate that PPIs, PPI interfaces, as well as the selective pressures acting on interfacial residues, can be highly

volatile in their evolution and can vary greatly between different species, thus requiring the use of lineage-specific measures when studying their evolution. These results yield insight into the evolutionary design principles of PPIs and the molecular evolutionary mechanisms by which interactomes rewire between species, improving our understanding of the molecular evolution of PPIs and PPI interfaces at the residue level.

## 4.3 Results

### Interfacial residues tend to evolve more slowly than non-interfacial residues, but only when using lineage-specific evolutionary rate measures

We assembled a dataset of high-quality three- dimensional (3D) structural models for a list of high-confidence protein–protein interactions (PPIs) in baker's yeast (*S. cerevisiae*) and fission yeast (*S. pombe*) (see **Materials and Methods**). For PPIs with experimentally-determined structure in the Protein Data Bank (PDB), we used the experimental structure for all subsequent analyses. For PPIs with no experimental structure in the PDB, if they have a homologous PPI with experimental structure in the PDB, we transferred the structural annotations from the homologous PPI (with available experimental structure) to the PPI of interest via sequence alignment. This homology-based structural annotation transfer process was successfully used in previous work and is further described in these works [26,44]. The final dataset comprises high-quality structural models for 717 PPIs between 611 *S. cerevisiae* proteins containing more than 400,000 residues, and 191 PPIs between 191 *S. pombe* proteins containing more than 140,000 residues.

We then used manually curated ortholog mappings [38], and reciprocal BLAST alignments [39] between proteins in the two yeast species to further classify *S. cerevisiae* and *S. pombe* PPIs in

our dataset as either preserved, having a missing ortholog, or rewired between the two yeasts. Preserved PPIs are PPIs that can be found in both *S. cerevisiae* and *S. pombe*. Rewired PPIs are PPIs that can be found in only one of *S. cerevisiae* and *S. pombe*, despite both protein partners having orthologs in the two species. We note that this set of rewired PPIs includes both PPIs that were absent in the most recent common ancestor of *S. cerevisiae* and *S. pombe* and subsequently gained in one species, as well as PPIs that were present in the most recent common ancestor of *S. cerevisiae* and *S. pombe* and subsequently lost in the other species and does not distinguish between those two cases. Missing ortholog PPIs are PPIs where at least one of the interacting protein partners has no ortholog or even homolog in either *S. cerevisiae* or *S. pombe*, and thus, we know that the interaction that is present in one of the yeast species is truly absent in the other species. Here we also note that this set of missing ortholog PPIs includes both PPIs where an interaction partner was absent in the most recent common ancestor of *S. cerevisiae* and *S. pombe* and subsequently gained in one species, as well as PPIs where an interaction partner was present in the most recent common ancestor of *S. cerevisiae* and *S. pombe* and subsequently lost in the other species and does not distinguish between those two cases. This process yielded 102 preserved PPIs, 447 missing ortholog PPIs, and 202 rewired PPIs between the two yeast species (additional details on the number of PPIs curated from each species listed in **Table S1**).

Next, we identified interfacial and non-interfacial residues in all *S. cerevisiae and S. pombe* PPIs with high-quality structural models. Interfacial residues are defined as amino acid residues exhibiting a change in solvent accessibility upon formation of a PPI complex. We, therefore, computed ΔRSA, the change in a residue's relative solvent accessibility (RSA) upon complex formation, for all residues in our PPI structural models, and residues with $\Delta RSA \neq 0$ were labeled

as interfacial. This yielded more than 50,000 *S. cerevisiae* interfacial residues and over 8,000 *S. pombe* interfacial residues (data curation pipeline summarized in **Figure 1** and further detailed in **Materials and Methods**).

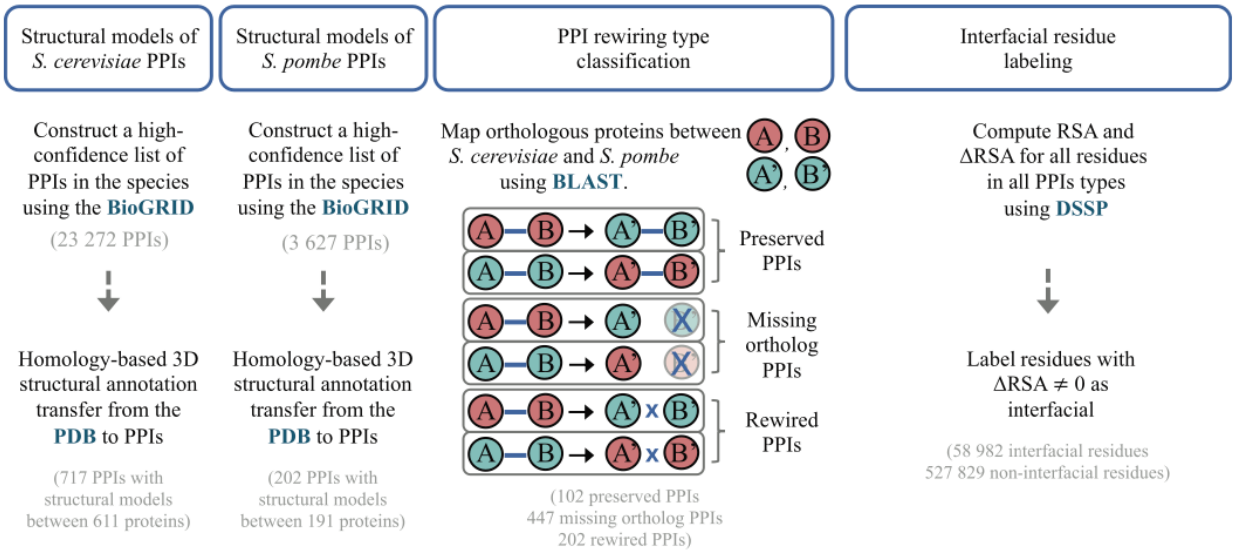



**Figure 1. Computational pipeline.** Graphical representation of the pipeline used for curation, homology-based structural annotation transfer, and rewiring type classification of PPIs in *S. cerevisiae* and *S. pombe*. Proteins are illustrated in cross-section as circles. Pairs of interacting proteins (PPIs) are illustrated in cross-section as circles connected by a blue line. Proteins that do not interact or are missing in a species are highlighted with a blue cross. Proteins and PPIs in *S. cerevisiae* and *S. pombe* are illustrated with different colors. Detailed description of the pipeline can be found in **Materials and Methods**.

We then assessed the degree of evolutionary constraint on interfacial and non-interfacial residues in all yeast PPIs with structural models using two measures: *ConSurf-rate4site* score and *ConSurf-DB* score (see **Materials and Methods**). These represent two popular ways to estimate residue evolutionary rate in the literature.

*ConSurf-rate4site* score is a measure of residue conservation in protein structures. It is calculated by comparing the amino acid at each site in a protein sequence to the corresponding amino acids in a chosen set of aligned closely related species. This measure is lineage-specific as

it considers only the evolution of the protein in the species of interest and in a chosen set of closely related species [34]. In contrast, *ConSurf-DB* scores available on the *ConSurf* database are not lineage-specific, and instead their calculations use sequence alignments to as many species as possible, including species outside the close lineage of the species of interest [35]. *ConSurf-DB* scores are calculated by comparing the amino acid at each sites in a protein sequence to the corresponding amino acids in all available sequences. While this measure provides a comprehensive view of residue conservation in all species containing proteins homologous to the protein of interest, it may not be ideal when studying PPIs, as PPIs are highly lineage-specific. Indeed, PPIs are often characterized by small changes in amino acid residues that enable or disable specific interactions between two or more proteins. These changes are frequently lineage-specific, meaning that they have evolved in the species of interest and its closely related species. As a result, using a sequence alignment to as many species as possible, including those that are not in the close lineage of the species of interest, may fail to capture lineage-specific signals that are important to the study of PPIs.

**Figure 2(A)** shows the average evolutionary rates for interfacial residues ($\Delta RSA \neq 0$) and non-interfacial residues ($\Delta RSA = 0$) in *S. cerevisiae* PPIs. Interfacial residues are, on average, significantly more conserved than non-interfacial residues using the lineage-specific estimate of evolutionary rates (*ConSurf-rate4site* score), but not significantly more conserved using the non-lineage-specific estimate of evolutionary rates (*ConSurf-DB* score). T-tests for the difference in average evolutionary rates between interfacial and non-interfacial residues in *S. cerevisiae* PPIs (**Figure 2(B)**) indicate a statistically significant difference for the lineage-specific measure of evolutionary rates (P-value < 0.01), but not for the non-lineage-specific measure of evolutionary

rates (P-value > 0.05). The results in *S. pombe* are similar, with interfacial residues, on average, significantly more conserved than non-interfacial residues using the lineage-specific measure of evolutionary rates (*ConSurf-rate4site* score), but the difference in evolutionary rates is much smaller when using the non-lineage-specific estimate of evolutionary rates (*ConSurf-DB* score) (**Figure 2(C)**). The t-tests for the difference in average evolutionary rates between interfacial and non-interfacial residues in *S. pombe* PPIs (**Figure 2(D)**) indicate a statistically significant difference for the lineage-specific measure of evolutionary rates (P-value < 0.01) with a larger effect size, and a less statistically significant difference for the non-lineage-specific measure of evolutionary rates (P-value < 0.05) with a smaller effect size.

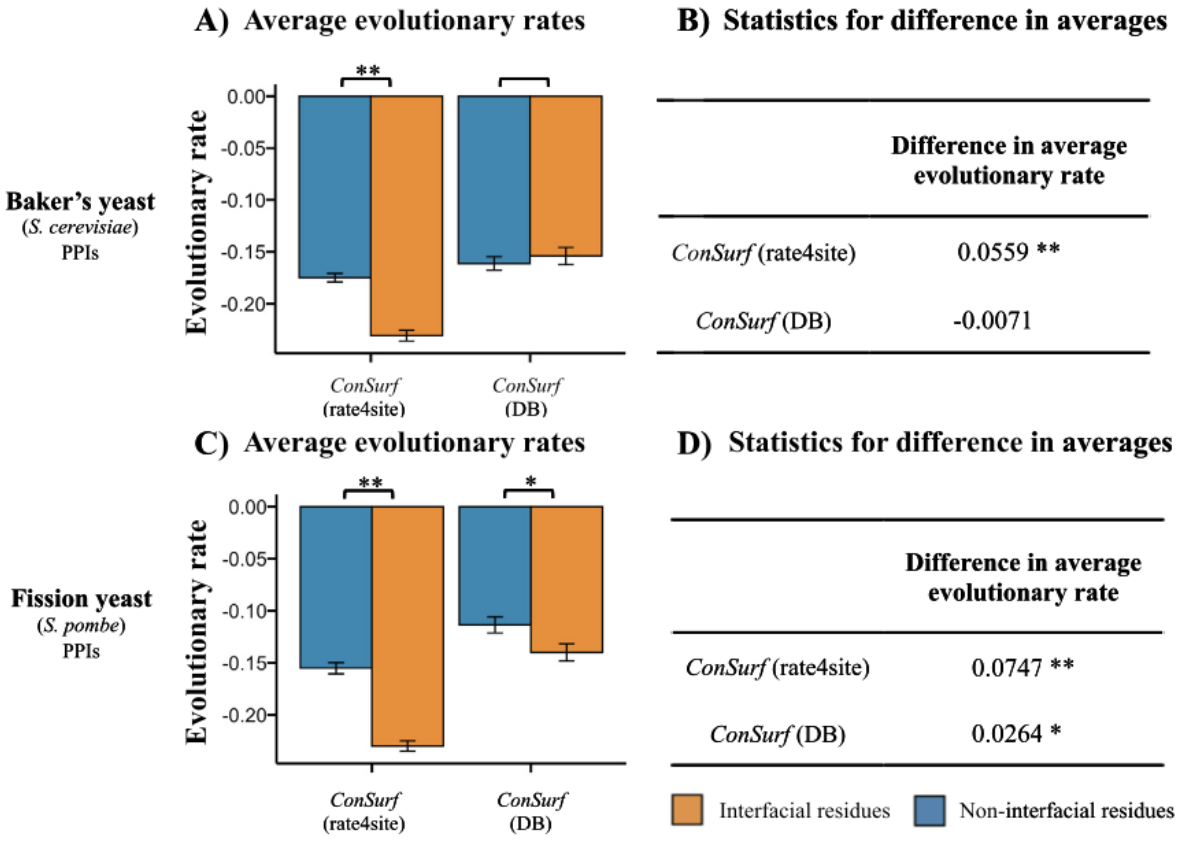



**Figure 2. The difference in evolutionary rate between interfacial and non-interfacial residues. (A)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for interfacial and non-interfacial residues from all *S. cerevisiae* PPIs in our datasets. Standard errors for the average values of each group of residues are also shown. **(B)**

Results of t-tests for differences in average evolutionary rates between interfacial and non-interfacial residues in *S. cerevisiae* PPIs using both evolutionary rate measures. **(C)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for interfacial and non-interfacial residues from all *S. pombe* PPIs in our datasets. Standard errors for the average values of each group of residues are also shown. **(D)** Results of t-tests for differences in average evolutionary rates between interfacial and non-interfacial residues in *S. pombe* PPIs using both evolutionary rate measures. Comparisons significant at the P-value < 0.05 level are denoted with a single asterisk (*) and comparisons significant at the P-value < 0.01 level are denoted with a double asterisk (**).

Previous work showed that interfacial residues are under strong evolutionary constraints and tend to be more conserved than non-interfacial residues, although estimates of those constraints vary depending on the dataset considered [30–33,37,44]. Here we find that, for proteins involved in *S. cerevisiae* and *S. pombe* PPIs, interfacial residues are indeed, on average, more conserved than non-interfacial residues, when using lineage-specific measures of evolutionary rates, with this difference in conservation being much smaller when using non-lineage-specific measures of evolutionary rates. This is surprising as the non-lineage-specific measure of evolutionary rate used here (*ConSurf-DB* score) is a popular and widely used evolutionary rate measure in the literature. The discrepancy of results observed may, therefore, be evidence that to study PPIs, which are highly lineage-specific, using lineage-specific evolutionary rate measures is important.

**Lineage-specific evolutionary rate measures are required to distinguish interfacial from non-interfacial residues in rewired PPIs, as well as PPIs with a missing ortholog between species**

To further probe the importance of lineage specificity in studying PPI structure and evolution, we studied preserved PPIs, PPIs missing an ortholog, and rewired PPIs in baker's yeast and fission yeast separately, investigating differences in evolutionary conservation between the three PPI types. Preserved PPIs are PPIs that are present in both baker's yeast and fission yeast.

As such, preserved PPIs are expected to be more universal and less specific to a given lineage. In contrast, rewired PPIs, as well as PPIs where an ortholog is missing in one of the two yeast species, are expected to be unique to a particular species and its closely related lineage.

**Figure 3(A)** and **(E)** shows the average evolutionary rates for interfacial residues (ΔRSA ≠ 0) and non-interfacial residues (ΔRSA = 0) in preserved PPIs in baker's yeast, and preserved PPIs in fission yeast respectively.

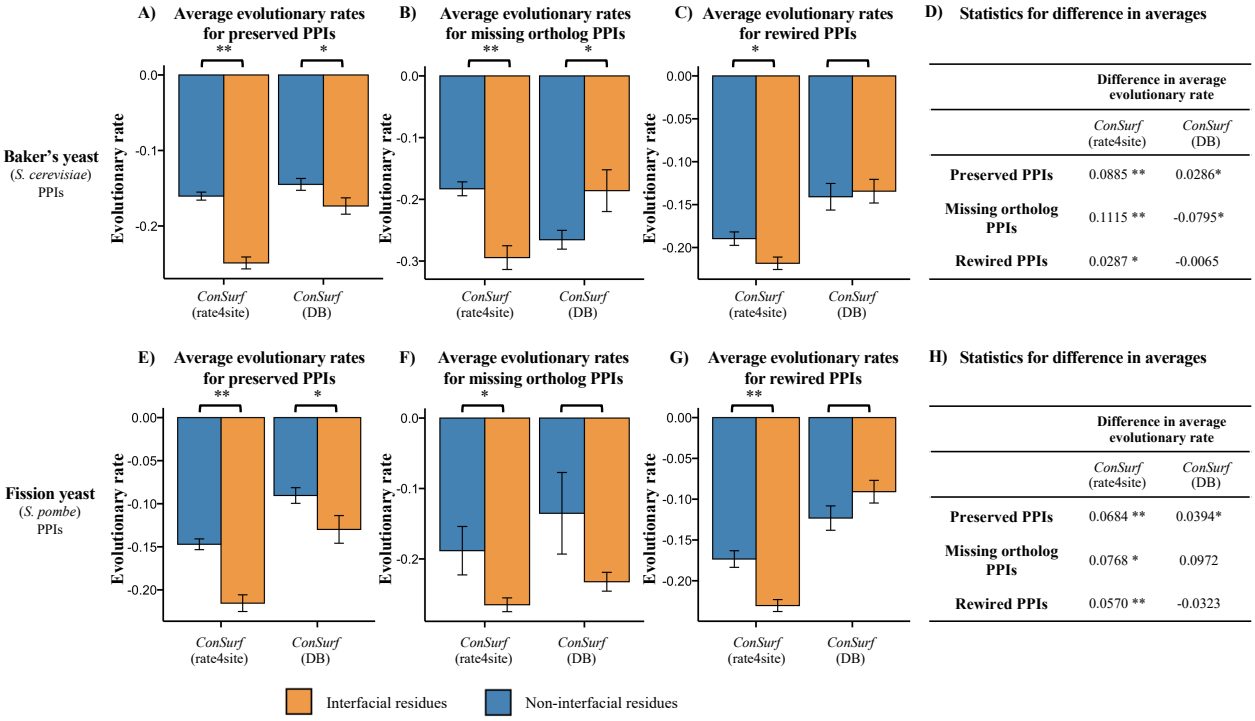



**Figure 3. The difference in evolutionary rate between interfacial and non-interfacial residues for preserved, missing ortholog, and rewired PPIs**. **(A-C)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for interfacial and non-interfacial residues from all preserved PPIs, missing ortholog PPIs and rewired PPIs in *S. cerevisiae* respectively. Standard errors for the average values of each group of residues are also shown. **(D)** Results of t-tests for differences in average evolutionary rates between interfacial and non- interfacial residues in preserved PPIs, missing ortholog PPIs and rewired PPIs in *S. cerevisiae* using both measures of evolutionary rate. **(E-G)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for interfacial and non-interfacial residues from all preserved PPIs, missing ortholog PPIs and rewired PPIs in *S. pombe* respectively. Standard errors for the average values of each group of residues are also shown. **(H)** Results of t-

tests for differences in average evolutionary rates between interfacial and non-interfacial residues in preserved PPIs, missing ortholog PPIs and rewired PPIs in *S. pombe* using both measures of evolutionary rate. Comparisons significant at the P-value < 0.05 level are denoted with a single asterisk (*) and comparisons significant at the P-value < 0.01 level are denoted with a double asterisk (**).

Preserved PPIs are PPIs where two protein partners have been found to interact in one of the two yeast species (*S. cerevisiae* or *S. pombe*), and orthologs of the two protein partners also interact in the other yeast species. For preserved PPIs in both yeast species, interfacial residues are, on average, significantly more conserved than non-interfacial residues using both the lineage-specific measure of evolutionary rates (*ConSurf-rate4site* score) and the non-lineage-specific measure of evolutionary rates (*ConSurf-DB* score). T-tests for the difference in average evolutionary rates between interfacial and non- interfacial residues in preserved PPIs in baker's yeast (**Figure 3(D)**) and preserved PPIs in fission yeast (**Figure 3(H)**) indicate a statistically significant difference for the lineage-specific estimate of evolutionary rates (P-value < 0.01), as well as for the non-lineage-specific estimate of evolutionary rates (P-value < 0.05). Thus, we conclude that lineage specificity is less important for the investigation of PPIs that are preserved between the two yeast species. Indeed, as baker's yeast and fission yeast are very distantly related evolutionarily, PPIs that are present in both species tend to be more universally conserved and found across many branches of the tree of life. Using lineage-specific measures of evolutionary rates to study those types of PPIs may, therefore, be less important.

**Figure 3(B)** and **(F)** shows the average evolutionary rates for interfacial residues ($\Delta RSA \neq 0$) and non-interfacial residues ($\Delta RSA = 0$) in missing ortholog PPIs in baker's yeast, and missing ortholog PPIs in fission yeast respectively. Missing ortholog PPIs are PPIs where the two protein partners interact in one of the two yeast species (*S. cerevisiae* or *S. pombe*), and at least

one of the protein partners has no ortholog or even homolog in the other yeast species. We can, therefore, state with high confidence, that the PPI found in one of the yeast species has no counterpart in the other yeast species, as at least one of the interacting partners is truly absent. For missing ortholog PPIs in baker's yeast (**Figure 3(B)**), interfacial residues are, on average, significantly more conserved than non-interfacial residues when using the lineage-specific *ConSurf-rate4site* score (P- value < 0.01, **Figure 3(D)**), but are significantly less conserved than non-interfacial residues when using the non-lineage-specific *ConSurf-DB* score (P-value < 0.05, **Figure 3(D)**), which is completely opposite to the trend expected. In addition, we observed similar results for missing ortholog PPIs in fission yeast (**Figure 3(F)**). Here, interfacial residues are, on average, significantly more conserved than non-interfacial residues when using the lineage-specific *ConSurf-rate4site* score (P-value < 0.05, **Figure 3(H)**), but not when using the non-lineage specific *ConSurf-DB* score (P- value > 0.05, **Figure 3(H)**). Therefore, we conclude that lineage specificity is crucial to the study of PPIs with missing orthologs between species.

Finally, we investigated rewired PPIs in both yeast species. **Figure 3(C)** and **(G)** shows the average evolutionary rates for interfacial residues ($\Delta RSA \neq 0$) and non-interfacial residues ($\Delta RSA = 0$) in rewired PPIs in baker's yeast, and rewired PPIs in fission yeast respectively. Rewired PPIs are PPIs where two protein partners have been found to interact in one of the two yeast species (*S. cerevisiae* or *S. pombe*), but orthologs of the protein partners do not interact in the other yeast species. For rewired PPIs in baker's yeast (**Figure 3(C)**), interfacial residues are, on average, significantly more conserved than non-interfacial residues when using *ConSurf-rate4site* score, the lineage-specific measure of evolutionary rates (P- value < 0.05, Figure 3(D)), but not significantly more conserved when *ConSurf-DB* score, the non-lineage-specific measure

of evolutionary rates is used (P-value > 0.05, **Figure 3(D)**). In addition, we observed similar results for rewired PPIs in fission yeast (**Figure 3(G)**) where interfacial residues are, on average, significantly more conserved than non-interfacial residues when using *ConSurf-rate4site* score, the lineage-specific measure of evolutionary rates (P-value < 0.01, **Figure 3(H)**), but not significantly more conserved when *ConSurf-DB* score, the non-lineage-specific measure of evolutionary rates is used (P-value > 0.05, **Figure 3(H)**). Thus, we conclude that lineage specificity is important to the study of PPI rewiring between species.

Overall, these results are surprising and show that interfacial residues in preserved PPIs are significantly more conserved than non-interfacial residues in both *S. cerevisiae* and *S. pombe*, no matter which measure of evolutionary rate is used. However, for missing ortholog PPIs and rewired PPIs in both yeast species, the significantly increased conservation of interfacial residues is only observed when using lineage-specific evolutionary rate measures, and not observed at all when using non-lineage-specific evolutionary rate measures.

**Buried residues tend to evolve more slowly than exposed residues for non-interfacial residues, when using both lineage-specific and non-lineage-specific evolutionary rate measures**

While PPIs and PPI interfaces are highly lineage-specific and can vary significantly between different species, residue solvent exposure and burial for non-interfacial residues are expected to be more universal and less variable across different lineages. Hence, we hypothesized that non-lineage-specific measures of evolutionary rate are more appropriate for the study of buried versus exposed residues outside of PPI interfaces. To test this hypothesis, we compared

evolutionary rates obtained using both the lineage-specific measure and the non-lineage-specific measure for buried and exposed residues outside of PPI interfaces in *S. cerevisiae* and *S. pombe* PPIs. We computed residue burial as measured by relative solvent accessibility (RSA) for all non-interfacial residues in our PPI structural models and investigated evolutionary rates for buried residues (RSA < 0.25) and exposed residues (RSA ≥ 0.25) separately.

**Figure 4. (A)** shows the average evolutionary rates for buried residues (RSA < 0.25) and exposed residues (RSA ≥ 0.25) outside of PPI interfaces in baker's yeast PPIs. Buried residues are, on average, significantly more conserved than exposed residues using both *ConSurf-rate4site* score, the lineage-specific measure of evolutionary rates, and *ConSurf-DB* score, the non-lineage-specific measure of evolutionary rates (P-value < 0.01, **Figure 4(B)**). In addition, we observed similar results for fission yeast PPIs (**Figure 4(C)**), where buried residues are, on average, significantly more conserved than exposed residues using both the *ConSurf-rate4site* score and the *ConSurf-DB* score measures of evolutionary rates (P-value < 0.01, **Figure 4(D)**). These results are in agreement with previous works in single proteins showing buried residues are under very strong evolutionary pressure to maintain protein structure and stability [41–43]. Moreover, solvent exposure or degree of burial has been established as a significant structural predictor of residue evolutionary rate in single proteins [44].

To further confirm these results, we compared evolutionary rates for buried and exposed residues outside of PPI interfaces in all three types of PPIs (preserved PPIs, missing ortholog PPIs and rewired PPIs) in both yeast species. We computed residue burial as measured by relative solvent accessibility (RSA) for all non-interfacial residues in our PPI structural models and

investigated evolutionary rates for buried residues (RSA < 0.25) and exposed residues (RSA  0.25) separately in the three PPI types.

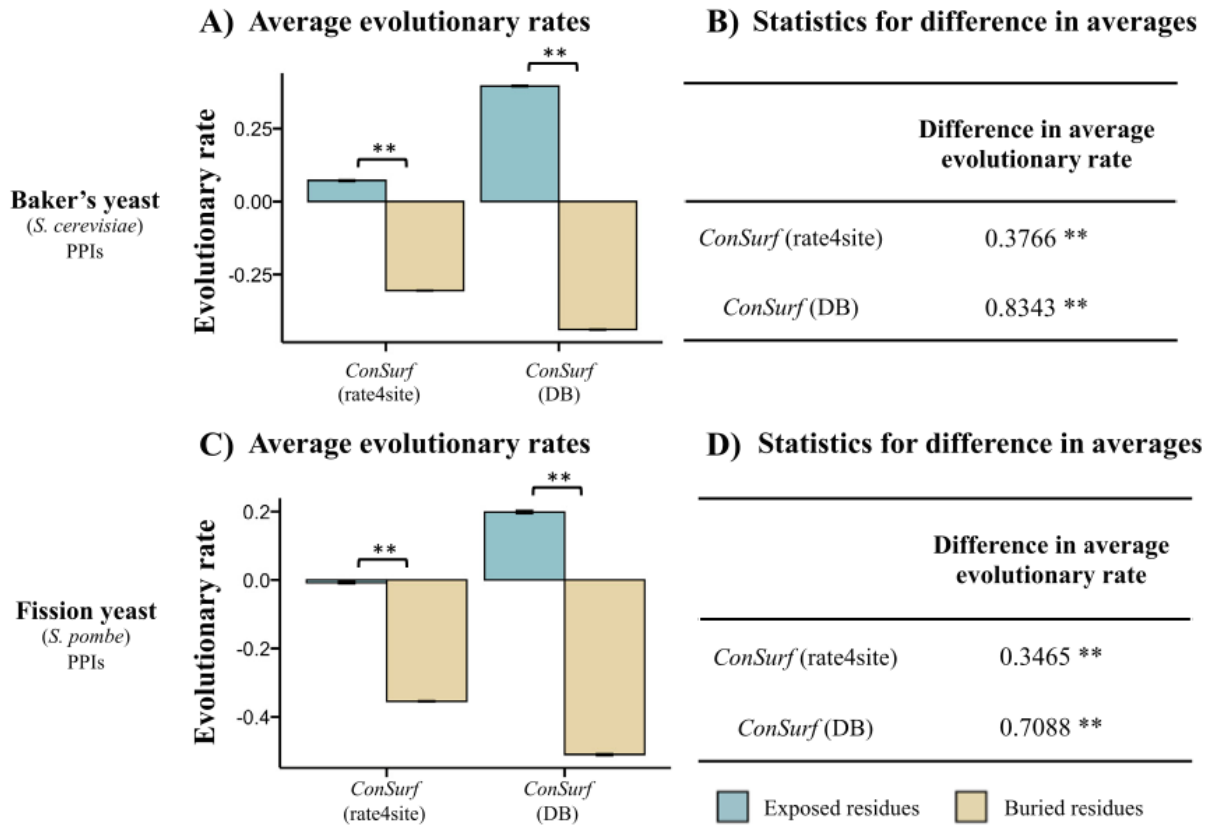



**Figure 4. The difference in evolutionary rate between buried and exposed residues outside of PPI interfaces. (A)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for buried and exposed residues outside of PPI interfaces from all *S. cerevisiae* PPIs in our data. Standard errors for the average values of each group of residues are also shown. **(B)** Results of t-tests for differences in average evolutionary rates between buried and exposed residues outside of PPI interfaces in *S. cerevisiae* PPIs using both evolutionary rate measures. **(C)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for buried and exposed residues outside of PPI interfaces from all *S. pombe* PPIs in our data. Standard errors for the average values of each group of residues are also shown. **(D)** Results of t-tests for differences in average evolutionary rates between buried and exposed residues outside of PPI interfaces in *S. pombe* PPIs using both evolutionary rate measures. Comparisons significant at the P-value < 0.01 level are denoted with a double asterisk (**).

**Figure 5(A–C)** shows the average evolutionary rates for buried residues (RSA < 0.25) and exposed residues (RSA ≥ 0.25) outside of PPI interfaces in preserved PPIs, missing ortholog PPIs,

and rewired PPIs in *S. cerevisiae* respectively. Buried residues are, on average, significantly more conserved than exposed residues for all three PPI types and using both the *ConSurf-rate4site* score and the *ConSurf-DB* score measures of evolutionary rates (P-value < 0.01, **Figure 5. (D)**). Moreover, we observed similar results for fission yeast PPIs (**Figure 3(E–G)**), where buried residues are also, on average, significantly more conserved than exposed residues for all three PPI types and using both the *ConSurf-rate4sit*e score and the *ConSurf-DB* score measures of evolutionary rates (P-value < 0.01, **Figure 5(H)**).

Overall, these results confirm that non-lineage-specific measures of evolutionary rates such as the *ConSurf-DB* scores are appropriate to study more universally-conserved processes such as residue exposure or burial for non-interfacial residues, as well as PPIs that are conserved across species. In contrast, when studying lineage-specific processes that can be highly variable between different species or lineages (e.g., interfacial residues within rewired PPIs or PPIs missing an ortholog between two species), using lineage-specific measures of evolutionary rates is necessary.

## 4.4 Discussion

In this work, we study the detailed molecular evolutionary mechanisms underlying interactome network rewiring, and the site-specific selective pressures acting on rewired protein–protein interactions (PPIs) between the interactomes of *S. cerevisiae* (baker's yeast) and *S. pombe* (fission yeast).

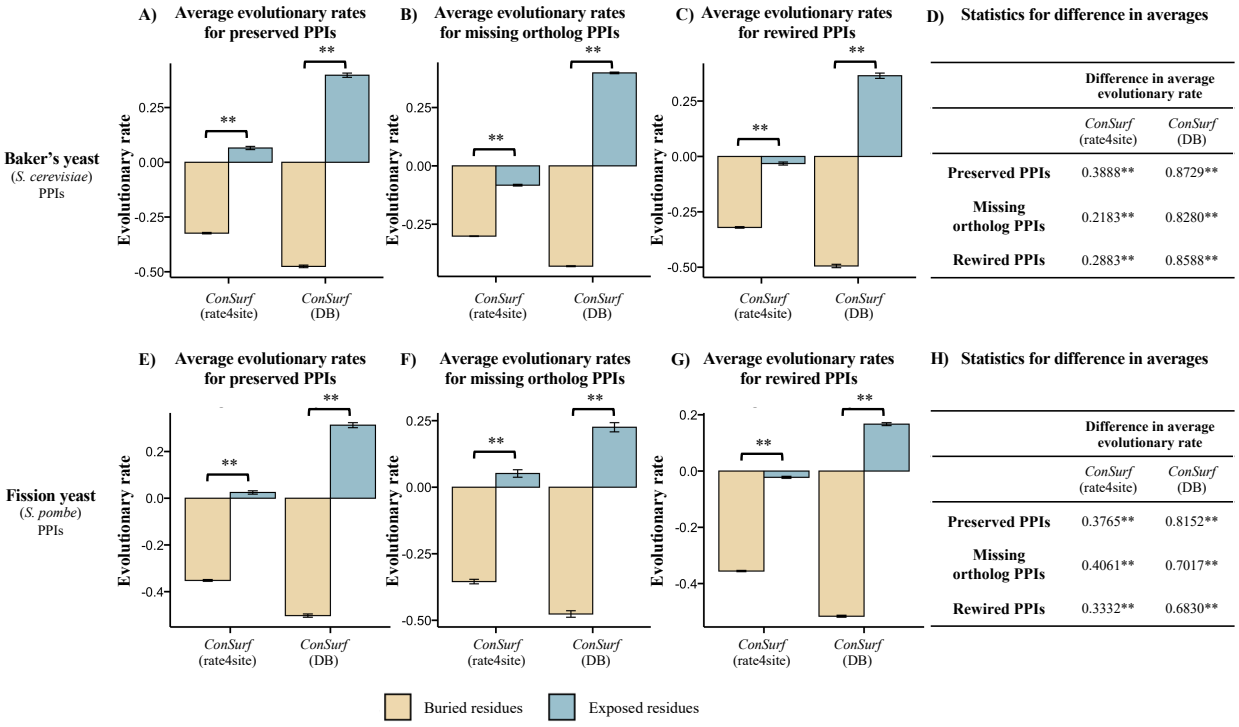



**Figure 5. The difference in evolutionary rate between buried and exposed residues outside of PPI interfaces for preserved, missing ortholog and rewired PPIs. (A–C)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for buried and exposed residues outside of PPI interfaces from all preserved PPIs, missing ortholog PPIs, and rewired PPIs in *S. cerevisiae* respectively. Standard errors for the average values of each group of residues are also shown. **(D)** Results of t-tests for differences in average evolutionary rates between buried and exposed residues outside of PPI interfaces in preserved PPIs, missing ortholog PPIs, and rewired PPIs in *S. cerevisiae* using both measures of evolutionary rate. **(E–G)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for buried and exposed residues outside of PPI interfaces from all preserved PPIs, missing ortholog PPIs, and rewired PPIs in *S. pombe* respectively. Standard errors for the average values of each group of residues are also shown. **(H)** Results of t-tests for differences in average evolutionary rates between buried and exposed residues outside of PPI interfaces in preserved PPIs, missing ortholog PPIs, and rewired PPIs in *S. pombe* using both measures of evolutionary rate. Comparisons significant at the P-value < 0.05 level are denoted with a single asterisk (*), and comparisons significant at the P-value < 0.01 level are denoted with a double asterisk (**).

We construct structurally-resolved interactome networks for the two yeasts, and compute and compare the site-specific evolutionary rates of interfacial versus non-interfacial PPI residues in both species. We find that interfacial residues in both baker's and fission yeast PPIs are significantly more conserved than non-interfacial residues when using a lineage-specific measure

of residue evolutionary rates. Those results are in agreement with previous works showing the critical role that interfacial residues play in mediating the interactions between protein partners and ensuring proper PPI formation and function [26,29–32,37,44]. However, when using a non-lineage-specific measure of evolutionary rate, the difference in evolutionary rate between interfacial residues and non-interfacial residues is drastically reduced for both species. Hence, the extent to which interfacial residues can be distinguished from other residues depends sensitively on the details of the site-specific evolutionary rate measures used, specifically whether the evolutionary rate measures are lineage-specific or not. This surprising sensitivity suggests that considering the close lineage of a species is important when studying evolutionary rates in PPIs. To explore possible drivers for this interesting result, we subdivided all interfacial residues in our data into three distinct categories commonly used in the literature: interfacial rim residues, interfacial support residues and interfacial core residues [59]. We then compared evolutionary rates for these three types of interfacial residues to non-interfacial residue evolutionary rates (**Supplementary material Analysis S2**). The results (**Figure S3, Figure S4**) indicate that interfacial core residues appear to be ones most sensitive to the choice of closely related species used in evolutionary rate calculation. This interesting observation supports previous works on the evolution of PPIs stating that interfacial surface and interfacial rim may be pre-existing in monomeric proteins, and that evolving a new PPI could require mutations to form an interface core only [59]. Under the above-described model of evolution of a PPI, interfacial core residues would indeed be highly lineage-specific, and, therefore, particularly sensitive to the choice of closely related species considered when calculating their evolutionary rates, which is what we observe here.

To further probe the basis of this conclusion, we investigated preserved PPIs, PPIs missing an ortholog and rewired PPIs in *S. cerevisiae* and *S. pombe* separately. Preserved PPIs are PPIs that are present in both baker's yeast and fission yeast, and thus are more universal and less specific to a given lineage. We, therefore, expect that non-lineage-specific measures of evolutionary rates may be appropriate to the study of preserved PPIs. In contrast, PPIs that are different between the two yeast species, including cases where an interaction is gained or lost but both binding partners are retained (rewired PPIs), and cases where a binding partner is gained or lost between species (missing ortholog PPIs) may be unique to a species and its closely related lineage, and thus much more sensitive to the choice of species and lineages used in evolutionary rate calculations. Indeed, we find that interfacial residues in preserved PPIs are significantly more conserved than non-interfacial residues in both *S. cerevisiae* and *S. pombe* when using both lineage-specific and non-lineage-specific measures of evolutionary rates. In contrast, for residues in PPIs that are different between the two yeast species (including rewired PPIs, as well as PPIs where an ortholog is missing in one of the yeast species), the increased conservation of interfacial residues is only observed when using a lineage-specific evolutionary rate measure, and not observed at all when using a non-lineage- specific evolutionary rate measure. These remarkable differences show that the loss or gain of an ortholog in a PPI, as well as PPI rewiring are highly lineage-specific events, and therefore the choice of evolutionary rate measures used to study these types of PPIs is very important.

Furthermore, each rewired PPI in our data is associated with a set of "interfacial" residues (in the species with the interaction), and a set of "pseudo-interfacial" residues (a set of non-interfacial residues in the species without the interaction that align to the functional interface in

the species with the interaction). We therefore compared evolutionary rates for interfacial residues and pseudo-interfacial residues in all *S. cerevisiae* and *S. pombe* rewired PPIs and their corresponding single, non-interacting proteins, and the results show that interfacial residues are, on average, more conserved than their pseudo- interfacial **Supplementary material Analysis S1**, **Figure S1, S2**).

Finally, both lineage-specific and non-lineage-specific measures of evolutionary rates are able to distinguish buried residues from exposed residues outside of PPI interfaces in all three types of PPIs (preserved PPIs, missing ortholog PPIs, and rewired PPIs) in both yeast species. Overall, these results suggest that including more distantly related species in evolutionary rate calculations may be appropriate when elucidating structural determinants of protein evolution for non- interfacial residues in PPIs, such as residue burial or exposure to solvent, and when studying more universally-conserved PPIs such as PPIs that are preserved between *S. cerevisiae* and *S. pombe*. In contrast, when investigating the loss or gain of an ortholog, or the rewiring of PPIs between different species, using lineage-specific evolutionary rates is crucial.

The data curation process used to construct our dataset of PPIs with both sequence and structure information comes with several caveats. PPI assays used to determine whether proteins interact in a species have associated false positive rates. As such, two proteins can be falsely labeled as interacting due to experimental errors. To address this caveat, only high-confidence PPIs, detected in at least two independent experiments, were used for subsequent analyses in this work. The false positive rate of our PPI dataset is further minimized by removing all PPIs that do not map (via sequence homology) to any physically interacting subunits in experimental 3D

structures of protein complexes. These multiple validation steps ensure that the false positive rate in our PPI dataset is minimal, and that PPIs found to be present in both yeast species (*S. cerevisiae* and *S. pombe*) can be labeled as "preserved PPIs" with high confidence.

In addition to false positive errors, PPI assays also have associated false negative errors. For example, it is estimated that 50% of PPIs in *S. cerevisiae* have thus far been experimentally identified [45]. As a result of the incomplete nature of the current yeast PPI networks, our dataset of PPIs labeled as rewired in this work not only includes PPIs that are truly present in one of the two yeast species (*S. cerevisiae* or *S. pombe*) and absent in the other, but also includes some PPIs that are truly preserved between the two yeasts but have simply not been detected in one of the two yeast species thus far. Indeed, our rewired PPI dataset appears to be enriched in weaker interactions, as PPIs in our rewired PPI sets have smaller interfaces and a larger proportion of rim to core interfacial residues in the interface than other PPIs in our data [59,60] (**Supplementary Material Table S2**). As weak interactions are significantly harder to detect and study, some interactions classified as rewired here could in fact be preserved, and simply not have been detected in one of the two yeast species thus far [60]. Despite this possible inclusion of some preserved PPIs in our rewired PPI dataset, we still observe a significantly better performance of lineage-specific evolutionary rate measures over non-lineage-specific evolutionary rate measures in discriminating interfacial from non-interfacial residues. Given the similar performance of both evolutionary rate measures on the preserved PPI dataset, we expect that the performance of lineage-specific evolutionary rate measures for the pure rewired PPI dataset will be even better than observed in this work.

To further investigate this issue, the set of missing ortholog PPIs was constructed as a gold-standard dataset of PPIs that are truly missing in one of the two yeast species. Indeed, if a PPI occurs in *S. cerevisiae*, but one of the interacting protein partners has no ortholog or even homolog in *S. pombe*, we can be highly confident that the counterpart PPI does not exist in *S. pombe*, and vice versa. Indeed, we find that the difference in performance between the lineage-specific evolutionary rate measures and non-lineage-specific evolutionary rate measures, when discriminating interfacial from non-interfacial residues, is the largest and most significant for this gold standard dataset. This result further confirms that the choice of lineage used in evolutionary rate calculations is highly important to the study of PPI differences and PPI rewiring between interactomes.

Finally, the PPI structural data used in this analysis comes with biases typically associated with experimental 3D structural measurements. Our PPI structural data may be biased towards proteins from particular cellular environments, more ancient and conserved proteins, commonly studied proteins, and highly expressed proteins [44,45]. To address these caveats, we include in our PPI structural datasets not only high- resolution experimental 3D structures of protein complexes, but also high-quality homology-based structural models for PPIs with no known 3D structures. Including such high-quality homology-based PPI structural models not only increases the cover-age but also reduces the biases in our PPI structural datasets. In addition, our results and conclusions are based on comparisons of different subsets within our PPI structural data where the same biases exist. Hence, these biases are likely cancelled out during the comparisons and unlikely to affect our results and conclusions. When performing homology-based structural annotation transfer, we assume that differences at the sequence level among close homologs do not produce

measurable structural differences and align yeast homologs to the same 3D structure. But even when the sequence-structure alignment is perfect, we cannot be fully confident that a given homology-mapped structure accurately reports on in-vivo properties of its residues [44]. However, we believe that the data curation process used here is still the best existing and the most reliable method for integrating structural details with molecular evolutionary properties of PPIs on a proteomic scale. Moreover, this method will only improve as the spaces of known PPIs and known structures grow. To further validate that the use of homology-based structural models does not unduly bias our results, we repeat the analysis using only experimentally determined protein complex structures, with no homology-based PPI structural models. Despite this additional analysis having a significantly reduced amount of data it yields results that are consistent with our main conclusions (**Supplementary material Analysis S3**, **Figure S5–S8**).

In summary, this work yields insight into the evolutionary design principles of PPIs and the mechanisms by which interactions are preserved or rewired between species, improving our understanding of the molecular evolution of PPI and PPI interfaces at the residue level.

## 4.5 Materials and Methods

### Homology-based structural annotation transfer

First, we curated a high-confidence set of physical interactions between *Saccharomyces cerevisiae* (*S. cerevisiae*) proteins, and *Schizosaccharomyces pombe* (*S. pombe*) proteins separately: we filtered the most recent release of the BioGRID database (May 2023) for physical PPIs reported in *S. cerevisiae* or *S. pombe* by two or more independent experiments (determined by different PubMed IDs), yielding 23,272 high-confidence PPIs between 4,321 *S. cerevisiae*

proteins, and 3,627 high-confidence PPIs between 2,059 *S pombe* proteins [46,47]. We then individually mapped the proteins involved in the aforementioned PPIs to 3D structures by performing gapped BLAST [39] searches under default settings between (i) a database built from the proteins' translated open reading frame sequences (ORFs) obtained on Ensembl [48] and (ii) 510,817 biological unit structure subunit sequences from the Protein Data Bank (PDB) [15]. For each ORF in the database, we constructed a list of potential structural matches by selecting biological unit structures which (i) produced E-values below a cut-off of $1 \times 10^{-5}$ in the alignment, (ii) had high coverage (>50%) in the alignment for both the ORF and the subunit sequence, and (iii) showed no inconsistencies (e.g., insufficient atomic detail, unreasonable distances between alpha-carbons, non-sensible heavy atom counts). We found 2,212 *S. cerevisiae* proteins and 710 *S. pombe* proteins involved in our high-confidence set of PPIs with at least one biological unit structure mapped to their ORF meeting those initial conditions. We further excluded all biological unit structures annotated as "low" or "very low" confidence on the QSbio database [49] as those structures could be doubtful biological assembly assignments. Finally, to select the best structural match to each interacting pair of *S. cerevisiae* or *S. pombe* proteins, we looked at the list of potential structural matches for each partner protein in a high-confidence PPI, and retained only the one which (i) met our initial alignment conditions above for both protein partners, (ii) showed the two protein partners in physical contact (i.e., mapped to spatially adjacent chains in the structure), (iii) had the highest composite coverage (sum of the coverage for each partner protein) in the BLAST alignment, and (iv) had a resolution better than 3 Å. If more than one potential structure remained for the PPI following the above process, the structure with the best resolution was kept. We further note that no explicit structural model building, and refinement were performed in this analysis: the structures curated as best structural matches for *S. cerevisiae* or *S. pombe* PPIs were all obtained

directly from the PDB, using the process referred to as homology-based structural annotation transfer above. Structures that are not annotated as yeast structures on the PDB but are, nonetheless, the best structural match for a known *S. cerevisiae* or *S. pombe* PPI were taken as is, assuming that with high sequence conservation between two known PPIs, structural conservation must also be high. This homology-based structural mapping pipeline yielded structural models for 717 PPIs between 611 *S. cerevisiae* proteins containing more than 400,000 residues, and 191 PPIs between 191 *S. pombe* proteins containing more than 140,000 residues and is illustrated in **Figure 1**. This homology-based structural annotation transfer process was successfully used in previous work and is further described in those works [26,44].

**Ortholog mapping between the two yeast species**

For each protein involved in *S. cerevisiae* or *S. pombe* PPIs, we established whether they have an ortholog in the other yeast species. First, using a manually curated lists of orthologs between the two yeasts, available on the Pombase server [38], we constructed an initial list of orthologs between PPI proteins in the two species. As this manually curated list is a *S. pombe* resource, it includes all *S. pombe* proteins in our data but only 20% of *S. cerevisiae* proteins in our data. No similar manually curated list of orthologs is currently available for the mapping from *S. cerevisiae* to *S. pombe*. Therefore, to improve our coverage of ortholog mapping from *S. cerevisiae* to *S. pombe*, we performed Reciprocal Best Hits BLAST (RBHB) using an E-value cut-off of $1 \times 10^{-5}$ between databases built from proteins' translated open reading frame sequences (ORFs) obtained on Ensembl[48] for each yeast species. If two proteins, each encoded in a different genome, find each other as the highest-scoring matches among the proteome of the opposite genome, they are reciprocal best hits (RBH) and thus inferred to be orthologs [50]. This yielded an ortholog

mapping for 108 additional proteins involved in structurally modeled *S. pombe* and *S. cerevisiae* PPIs. Finally, proteins with no homologs in the other yeast species (via BLAST alignment using an E-value cut-off of $1 \times 10^{-5}$) were labeled as proteins that definitely do not have any ortholog in the other species. Overall, this ortholog mapping process between the two yeasts yielded an ortholog mapping for 191 proteins involved in structurally modeled *S. pombe* PPIs, and 517 proteins involved in structurally modeled *S. cerevisiae* PPIs.

**PPI type classification**

Using the ortholog mappings described above, we further classified *S. cerevisiae* and *S. pombe* PPIs in our dataset as either preserved, missing an ortholog, or rewired between the two yeast species. PPIs between two *S. cerevisiae* or *S. pombe* proteins that have a corresponding PPI between orthologs of the two proteins in the other species were labeled as preserved PPIs. We found 102 preserved PPIs in our structurally modeled PPI dataset. Rewired PPIs are PPIs where two protein partners interact in *S. cerevisiae* or *S. pombe* but orthologs of the protein partners do not interact in the other yeast species. Our dataset of structurally modeled PPIs contains 202 rewired PPIs. Finally, missing ortholog PPIs are PPIs between two *S. cerevisiae* or *S. pombe* proteins, where at least one of the two interacting protein partners has no ortholog or even homolog in the other species. Thus, we know that the PPI observed in one of the yeast species is truly absent in the other species. We found 447 missing ortholog PPIs in our structurally modeled PPI dataset. The data curation pipeline used to classify PPIs as preserved, rewired, or missing an ortholog is illustrated in **Figure 1**.

**Calculation of structural properties at the residue level**

Solvent Accessible Surface Area (SASA) was calculated using the DSSP program [51,52] with hydrogen atoms excluded. SASA values were normalized using reliable normalization values from Tien et al. [53] to produce Relative Solvent Accessibility (RSA). For each residue in our structural models, two values of RSA were computed: monomer RSA, which was calculated using the structure of monomeric proteins (discarding the chain mapped to the partner protein in a structure), and complex RSA, which was obtained from the co-complexed structure of both protein partners (PPI structure). ΔRSA, the change in residue burial upon complex formation, was computed as the difference between monomeric and co-structured RSA values for each residue in the structural models (ΔRSA = monomer RSA − complex RSA). ΔRSA was subsequently used in the definition of interfaces: any residue with a change in burial upon complex formation ($\Delta RSA \neq 0$) was defined as an interfacial residue. This yielded more than 50,000 *S. cerevisiae* interfacial residues and over 8,000 *S. pombe* interfacial residues.

**Evolutionary sequence analysis**

Estimating residue-level evolutionary rates is a non-trivial task and, thus, various methods have been proposed for this inference in the literature [43,54,55]. Here, we used an established technique to measure site-specific evolutionary rates in proteins, *ConSurf* score, which uses protein multiple sequence alignment data for rate inference [40]. Lineage-specific *ConSurf* scores, termed *ConSurf-rate4site* scores in this analysis, were computed using the *Rate4Site* program [35]. Non-lineage-specific *ConSurf* scores, termed *ConSurf-DB* scores in this analysis, were downloaded from the *ConSurf* Database [35,36].

For each protein involved in a high-confidence *S. cerevisiae* PPI, we generated multiple sequence alignments using *ClustalW* [56] between (i) its translated ORF, (ii) the sequence of its mapped protein structure subunit, and (iii) orthologous translated ORFs in *Saccharomyces paradoxus* (*S. paradoxus*), *Saccharomyces mikatae* (*S. mikatae*), *Saccharomyces bayanus* (*S. bayanus*), *Naumovozyma castellii* (*N. castellii*), *Candida glabrata* (*C. glabrata*), *Eremothecium gossypii* (*E. gossypii*), *Kluyveromyces lactis* (*K. lactis*) and *Candida albicans* (*C. albicans*) obtained from Ensembl [48,57].

For each protein involved in a high-confidence *S. pombe* PPI, we generated multiple sequence alignments using ClustalW [56] between (i) its translated ORF, (ii) the sequence of its mapped protein structure subunit, and (iii) orthologous translated ORFs in *Schizosaccharomyces japonicus* (*S. japonicus*), *Schizosaccharomyces octosporus* (*S. octosporus*), *Schizosaccharomyces cryophilus* (*S. cryophilus*), *Neolecta irregularis* (*N. irregularis*), *Pneumocystis jirovecii* (*P. jirovecii*), *Pneumocystis murina* (*P. murina*), *Saitoella complicata* (*S. complicata*) and *Protomyces lactucaedebilis* (*P. lactucaedebilis*) obtained from Ensembl [48,58].

*ConSurf-rate4site* scores were computed from protein multiple sequence alignment data using the *Rate4Site* program [34,35]. This method computes relative conservation scores for each site in a protein using empirical Bayesian methods and a multiple sequence alignment of homologous sequences to essentially rank residues from most to least conserved within a protein. The *ConSurf* score obtained from running the program is lower for more conserved residues, and higher for less conserved ones. We ran *ConSurf* score calculations for all residues in our models using the *Rate4Site* program, with the closely related species and phylogenetic tree mentioned above as

inputs. Other parameters in *Rate4Site* were left to their default values. When binning *ConSurf* scores for plotting in this analysis, the final *ConSurf* scores value and associated error for a group of residues were taken as the average and standard error of *ConSurf* score values for each residue in the bin.

*ConSurf-DB* scores were downloaded from the *ConSurf* database [36]. The *ConSurf* database provides pre-computed evolutionary rates for structures on the PDB. *ConSurf* database evolutionary rate calculations are performed using the *Rate4Site* program on multiple sequence alignments constructed using PSI-BLAST with an E-value cut-off of $10^{-3}$to find all potential homologs in the Uni-ProtKB/SwissProt database for proteins on the PDB.

## Code availability

The code pipeline used to construct structural models of *S. cerevisiae* and *S. pombe* PPIs is available as a GitHub repository: https://github.com/LeahPollet/interactome_network_rewiring.git. Curated data underlying this article, including a list of high confidence PPIs in *S. cerevisiae* and *S. pombe*, and homology mapped PDB structures for *S. cerevisiae* and *S. pombe* PPIs are also available in the GitHub repository and can be accessed using the following https://doi.org/10. 5281/zenodo.10222227.

## CRediT authorship contribution statement

**Léah Pollet**: Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yu Xia**: Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

## DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.


## Acknowledgments and funding information

This work was supported by Natural Sciences and Engineering Research Council of Canada grants RGPIN-2019-05952 and RGPAS-2019-00012, Canada Foundation for Innovation grants JELF-33732 and IF-33122, and Canada Research Chairs program (CRC-2022-00424).

## 4.6 References


1. Jones, S., Thornton, J.M., (1996). Principles of protein- protein interactions. Proc. Natl. Acad. Sci. U.S.A. 93 (1), 13–20. https://doi.org/10.1073/pnas.93.1.13.
2. Pawson, T., Nash, P., (2000). Protein-protein interactions define specificity in signal transduction. Genes Dev. 14 (9), 1027–1047. https://doi.org/10.1101/gad.14.9.1027.
3. Cohen, P., (2000). The regulation of protein function by multisite phosphorylation – a 25 year update. Trends Biochem. Sci. 25 (12), 596–601. https://doi.org/10.1016/ S0968-0004(00)01712-6.
4. Alberts, B., (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92 (3), 291–294. https://doi.org/10.1016/ S0092-8674(00)80922-8.
5. Gavin, A.C., Superti-Furga, G., (2003). Protein complexes and proteome organization from yeast to man. Curr. Opin. Chem. Biol. 7 (1), 21–27. https://doi.org/10.1016/S1367-5931(02)00007-8.
6. De Las Rivas, J., Fontanillo, C., (2010). Protein-protein interactions essentials: Key concepts to building and analyzing interactome networks. PLoS Comput. Biol. 6 (6), e1000807.
7. Landry, C.R., Levy, E.D., Michnick, S.W., (2009). Weak functional constraints on phosphoproteomes. Trends Genet. 25 (5), 193–197. https://doi.org/10.1016/j. tig.2009.03.003.
8. Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T., et al., (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490 (7421), 556–560. https://doi.org/10.1038/nature11503.
9. Ryan, D.P., Matthews, J.M., (2005). Protein-protein interactions in human disease. Curr. Opin. Struct. Biol. 15 (4), 441–446. https://doi.org/10.1016/j.sbi.2005.06.001.
10. Zhong, Q., Simonis, N., Li, Q.R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D., et al., (2009). Edgetic perturbation models of human inherited disorders. Mol. Syst. Biol. 5 (1), 321. https://doi. org/10.1038/msb.2009.80.
11. Schoenrock, A., Burnside, D., Moteshareie, H., Pitre, S., Hooshyar, M., Green, J.R., Golshani, A., Dehne, F., Wong, A., (2017). Evolution of protein-protein interaction networks in yeast. PloS One 12 (3), e0171920.
12. Yu, H., Braun, P., Yıldırım, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., et al., (2008). High- quality binary protein interaction map of the yeast interactome network. Science 322 (5898), 104–110. https://doi.org/10.1126/science.1158684.
13. Vo, T.V., Das, J., Meyer, M.J., Cordero, N.A., Akturk, N., Wei, X., Fair, B.J., Degatano, A.G., Fragoza, R., Liu, L.G., et al., (2016). A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. Cell 164 (1), 310–323. https://doi.org/10.1016/ j.cell.2015.11.037.
14. Rolland, T., Tasßan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al., (2014). A proteome-scale map of the human interactome network. Cell 159 (5), 1212–1226. https://doi.org/10.1016/j.cell.2014.10.050.
15. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., (2000). The protein data bank. Nucleic Acids Res. 28 (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

16. Cesareni, G., Ceol, A., Gavrila, C., Palazzi, L.M., Persico, M., Schneider, M.V., (2005). Comparative interactomics. FEBS Lett. 579 (8), 1828–1833. https://doi.org/10.1016/j.febslet.2005.01.064.
17. Gandhi, T.K.B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K.N., Mohan, S.S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B., et al., (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nature Genet. 38 (3), 285–293. https://doi.org/10.1038/ng1747.
18. Shou, C., Bhardwaj, N., Lam, H.Y., Yan, K.K., Kim, P.M., Snyder, M., Gerstein, M.B., (2011). Measuring the evolutionary rewiring of biological networks. PLoS Comput. Biol. 7 (1), e1001050.
19. Yamada, T., Bork, P., (2009). Evolution of biomolecular networks—lessons from metabolic and protein interactions. Nature Rev. Mol. Cell Biol. 10 (11), 791–803. https://doi.org/10.1038/nrm2787.
20. Liberles, D.A., Teichmann, S.A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L.J., De Koning, A. J., Dokholyan, N.V., Echave, J., et al., (2012). The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 21 (6), 769–785. https://doi.org/ 10.1002/pro.2071.
21. Xin, X., Gfeller, D., Cheng, J., Tonikian, R., Sun, L., Guo, A., Lopez, L., Pavlenco, A., Akintobi, A., Zhang, Y., et al., (2013). SH3 interactome conserves general function over specific form. Mol. Syst. Biol. 9 (1), 652. https://doi.org/ 10.1038/msb.2013.9.
22. Reinke, A.W., Baek, J., Ashenberg, O., Keating, A.E., (2013). Networks of bZIP protein-protein interactions diversified over a billion years of evolution. Science 340 (6133), 730–734. https://doi.org/10.1126/science.1233465.
23. Jeffares, D.C., Rallis, C., Rieux, A., Speed, D., Pˇrevorovsky´, M., Mourier, T., Marsellach, F.X., Iqbal, Z., Lau, W., Cheng, T.M., et al., (2015). The genomic and phenotypic diversity of Schizosaccharomyces pombe. Nature Genet. 47 (3), 235–241. https://doi.org/10.1038/ng.3215.
24. Seoighe, C., Wolfe, K.H., (1998). Extent of genomic rearrangement after genome duplication in yeast. Proc. Natl. Acad. Sci. U.S.A. 95 (8), 4447–4452. https://doi.org/10.1073/pnas.95.8.4447.
25. Das, J., Vo, T.V., Wei, X., Mellor, J.C., Tong, V., Degatano, A.G., Wang, X., Wang, L., Cordero, N.A., Kruer-Zerhusen, N., et al., (2013). Cross-species protein interactome mapping reveals species-specific wiring of stress response pathways. Sci. Signal. 6 (276), ra38-ra38. https://doi.org/10.1126/scisignal.2003350.
26. Pollet, L., Lambourne, L., Xia, Y., (2022). Structural determinants of yeast protein-protein interaction interface evolution at the residue level. J. Mol. Biol. 434, (19) https://doi.org/10.1016/j.jmb.2022.167750 167750.
27. Sydykova, D.K., Claus, O.W., (2017). Calculating site- specific evolutionary rates at the amino-acid or codon level yields similar rate estimates. PeerJ 5, e3391. https://doi.org/10.7717/peerj.3391.
28. Engin, H.B., Kreisberg, J.F., Carter, H., (2016). Structure- based analysis reveals cancer missense mutations target protein interaction interfaces. PLoS One 11 (4), e0152929.
29. Jubb, H.C., Pandurangan, A.P., Turner, M.A., Ochoa- Montano, B., Blundell, T.L., Ascher, D.B., (2017). Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health. Prog. Biophys. Mol. Biol. 128, 3–13. https://doi.org/10.1016/j.pbiomolbio.2016.10.002.

30. Valdar, W.S., Thornton, J.M., (2000). Protein-protein interfaces: Analysis of amino acid conservation in homodimers. Proteins 42 (1), 108–124. https://doi.org/ 10.1002/1097-0134(20010101)42:1<108::AID- PROT110>3.0.CO;2-O.
31. Ma, B., Elkayam, T., Wolfson, H., Nussinov, R., (2003). Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces. Proc. Natl. Acad. Sci. U.S.A. 100 (10), 5772–5777. https://doi.org/10.1073/pnas.1030237100.
32. Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., Huang, E.S., (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci. 13 (1), 190–202. https://doi.org/10.1110/ ps.03323604.
33. Eames, M., Kortemme, T., (2007). Structural mapping of protein interactions reveals differences in evolutionary pressures correlated to mRNA level and protein abundance. Structure 15 (11), 1442–1451. https://doi.org/ 10.1016/j.str.2007.09.010.
34. Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., Ben-Tal, N., (2002). Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18 (suppl_1), S71–S77. https://doi.org/ 10.1093/bioinformatics/18.suppl_1.s71.
35. Mayrose, I., Graur, D., Ben-Tal, N., Pupko, T., (2004). Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol. Biol. Evol. 21 (9), 1781–1791. https://doi. org/10.1093/molbev/msh194.
36. Ben Chorin, A., Masrati, G., Kessel, A., Narunsky, A., Sprinzak, J., Lahav, S., Ashkenazy, H., Ben-Tal, N., (2020). ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. Protein Sci. 29 (1), 258–267. https://doi.org/ 10.1002/pro.3779.
37. Andreani, J., Guerois, R., (2014). Evolution of protein interactions: from interactomes to interfaces. Arch. Biochem. Biophys. 554, 65–75. https://doi.org/10.1016/j. abb.2014.05.010.
38. Wood, V., Harris, M.A., McDowall, M.D., Rutherford, K., Vaughan, B.W., Staines, D.M., Aslett, M., Lock, A., Ba¨hler, J., Kersey, P.J., et al., (2012). PomBase: a comprehensive online resource for fission yeast. Nucleic Acids Res. 40 (D1), D695–D699. https://doi.org/10.1093/nar/gkr853.
39. Altschul, S.F., Madden, T.L., Scha¨ffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25 (17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389.
40. Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., Ben-Tal, N., (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res. 44 (W1), W344–W350. https://doi.org/10.1093/nar/gkw408.
41. Hubbard, T.J.P., Blundell, T.L., (1987). Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. Protein Eng. Des. Sel. 1 (3), 159–171. https://doi.org/10.1093/protein/ 1.3.159.
42. Choi, S.S., Vallender, E.J., Lahn, B.T., (2006). Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. Mol. Biol. Evol. 23 (11), 2131–2133. https://doi. org/10.1093/molbev/msl086.
43. Pa´l, C., Papp, B., Lercher, M.J., (2006). An integrated view of protein evolution. Nature Rev. Genet. 7 (5), 337–348. https://doi.org/10.1038/nrg1838.

44. Franzosa, E.A., Xia, Y., (2009). Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol. 26 (10), 2387–2395. https://doi.org/ 10.1093/molbev/msp146.
45. Hart, G.T., Ramani, A.K., Marcotte, E.M., (2006). How complete are current yeast and human protein-interaction networks? Genome Biol. 7 (11), 1–9. https://doi.org/ 10.1186/gb-2006-7-11-120.
46. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., Tyers, M., (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34 (suppl_1), D535–D539. https://doi.org/10.1093/nar/gkj109.
47. Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., et al., (2014). The BioGRID interaction database: 2015 update. Nucleic Acids Res. 43 (D1), D470–D478. https://doi.org/ 10.1093/nar/gku1204.
48. Martin, F.J., Amode, M.R., Aneja, A., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., et al., (2023). Ensembl 2023. Nucleic Acids Res. 51 (D1), D933–D941. https://doi.org/10.1093/nar/ gkac958.
49. Dey, S., Ritchie, D.W., Levy, E.D., (2018). PDB-wide identification of biological assemblies from conserved quaternary structure geometry. Nature Methods 15 (1), 67–72. https://doi.org/10.1038/nmeth.4510.
50. Hernández-Salmerón, J.E., Moreno-Hagelsieb, G., (2020). Progress in quickly finding orthologs as reciprocal best hits: comparing blast, last, diamond and MMseqs2. BMC Genom. 21 (741), 1–9. https://doi.org/10.1186/s12864- 020-07132-6.
51. Kabsch, W., Sander, C., (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen- bonded and geometrical features. Biopolymers 22 (12), 2577–2637. https://doi.org/10.1002/bip.360221211.
52. Touw, W.G., Baakman, C., Black, J., Te Beek, T.A., Krieger, E., Joosten, R.P., Vriend, G., (2015). A series of PDB related databases for everyday needs. Nucleic Acids Res. 43 (D1), D364–D368. https://doi.org/10.1093/nar/ gku1028.
53. Tien, M.Z., Meyer, A.G., Sydykova, D.K., Spielman, S.J., Wilke, C.O., (2013). Maximum allowed solvent accessibilites of residues in proteins. PloS One 8 (11), e80635.
54. Zhang, J., Yang, J.R., (2015). Determinants of the rate of protein sequence evolution. Nature Rev. Genet. 16 (7), 409–420. https://doi.org/10.1038/nrg3950.
55. Echave, J., Spielman, S.J., Wilke, C.O., (2016). Causes of evolutionary rate variation among protein sites. Nature Rev. Genet. 17 (2), 109–121. https://doi.org/10.1038/ nrg.2015.18.
56. Madeira, F., Park, Y.M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A.R., Potter, S. C., Finn, R.D., et al., (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic Acids Res. 47 (W1), W636–W641. https://doi.org/10.1093/nar/gkz268.
58. Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R., et al., (2012). Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res. 40 (D1), D700–D705. https://doi. org/10.1093/nar/gkr1029.
59. Harris, M.A., Rutherford, K.M., Hayles, J., Lock, A., Ba¨hler, J., Oliver, S.G., Mata, J., Wood, V., (2022). Fission stories: using PomBase to understand Schizosaccharomyces pombe biology. Genetics 220, (4) https://doi.org/ 10.1093/genetics/iyab222 p.iyab222.

60. Levy, E.D., (2010). A simple definition of structural regions in proteins and its use in analyzing interface evolution. J. Mol. Biol. 403 (4), 660–670. https://doi.org/10.1016/j. jmb.2010.09.028.
61. Qin, J., Gronenborn, A.M., (2014). Weak protein complexes: challenging to study but essential for life. FEBS J. 281 (8), 1948–1949. https://doi.org/10.1111/ febs.12744.

# 5. Discussion

Protein-protein interactions (PPIs) are crucial for proper protein function and are involved in virtually all biological pathways within cells [1–6]. As such, numerous recent experiments have aimed to catalog the interactions of all proteins in a given species [9–14], and to elucidate the molecular structure and three-dimensional (3D) mechanisms of these interactions [15,115–121]. This extensive research has generated vast amounts of high-quality PPI data, which now enables us to study the evolution of PPIs. Investigating the evolution of PPIs is essential to further our understanding of a wide range of biological processes within cells [6,16,17,20], to better understand misregulation or disruption of PPIs associated with various diseases and disorders [24–40], as well as to inform practical applications in various research fields including disease diagnosis [41–45], disease treatment [46–54], synthetic biology [55–58], and genome engineering [59–61].

The work presented in this thesis, therefore, makes use of this newly available PPI data to perform large-scale systematic analyses of the detailed molecular evolutionary design principles that drive variations in PPIs, first within a species, and then between different species. More formally, we hypothesize that *structural determinants influence the evolutionary rate of residues in protein-protein interactions*; and that *changes in those determinants for interfacial residues could be associated with the phylogenetic loss or gain of an interaction between two species.* To demonstrate this hypothesis, we first create an automated, custom pipeline to combine high-confidence PPI data and 3D structures of protein complexes in order to build molecular models for all PPIs in a species (Chapter 3). We then show how these molecular models of PPIs can be used to investigate the relationship between PPI structures and their evolution, uncovering some of the molecular evolutionary design principles driving variations in PPIs in baker's yeast (Chapter

3). Finally, we show how these molecular models of PPIs can be used to compare PPIs across species, uncovering some of the molecular drivers leading to difference, or rewiring, of PPIs between baker's yeast and fission yeast (Chapter 4).

The first aim and significant contribution of this thesis is the creation of a custom automated pipeline for curating, processing, and organizing PPI data. This pipeline integrates data from diverse experimental fields enabling the construction of molecular models for PPIs in *S. cerevisiae* and *S. pombe*. The pipeline includes curation of a high-confidence set of physical interactions between proteins in a species from the BioGRID database [12] and IntAct database [13]. PPI datasets, such as the two databases used in this work are known to contain experimental false positives (erroneously reported PPIs) [229,230]. These inaccuracies arise primarily due to the limitations and inherent biases of the experimental techniques used to detect PPIs detailed in Chapter 2 of this thesis. For instance, high-throughput experimental methods such as yeast two-hybrid screens and tandem affinity purification-mass spectroscopy are prone to false positives due to non-specific interactions or cross-reactivity [231]. Additionally, the stringent conditions of *in vitro* experiments often do not accurately reflect *in vivo* cellular environments, leading to interactions that may not occur under physiological conditions [232]. Due to these biases, PPIs reported on online PPI databases could be erroneous, including, among others, non-reproducible experimental artifacts, *in vitro* physical interactions that do not occur *in vivo*, or pairs of proteins from the same complex that do not directly interact with each other. We, therefore, used several methods to minimize such false positive errors in our PPI data. First, we only considered experimentally-determined physical PPIs, excluding any computational predictions that can be less accurate [86,93,95,97]. Second, we only use PPIs reported by two or more independent experiments (determined

by different PubMed IDs) to ensure reproducibility and mitigate possible experimental artifacts [233,234]. Third, the false positive rate of our PPI datasets is further minimized by removing all PPIs that do not map (via sequence homology) to physically interacting subunits in experimental 3D structures of protein complexes on the PDB [15]. These multiple validation steps ensure that the false positive rate in our PPI dataset is minimal. In addition to false positive errors, PPI datasets are also known to contain false negatives [230]. Some PPIs that do occur in the species could be missing on PPI databases due to the incompleteness of interactome networks [235]. For instance, estimates of the proportion of known PPIs in *S. cerevisiae* suggest that only ~50% of yeast PPIs have been identified thus far [236]. True physical interactions that occur *in vivo* could also not be detected *in vitro* due to experimental biases [230]. Current PPI networks are, therefore, a sample of the complete networks. To address these biases, we use two of the most comprehensive PPI databases, the BioGRID database [12] and the IntAct database [13]. We also include PPIs detected using diverse experimental methods including, among others, yeast-two-hybrid screens, affinity capture experiments, and co-crystal structure experiments. Although incorporating data from additional databases could be a promising direction for future research, we note that integrating data from multiple PPI databases is not a straightforward task: while many databases provide interactions in a similar format, inconsistent or incorrect use of controlled vocabulary is common. Different gene and protein identifiers are also used across databases, and sometimes even within a single database [14]. Additionally, in this work, the primary limitation in PPI data was not at the species level, but rather due to the availability of high-resolution 3D structure for PPIs, as discussed below. Finally, as the space of known PPIs grows this curation method will only improve. More than 20,000 high-confidence *S. cerevisiae* PPIs, and more than 3,000 high-confidence *S. pombe* PPIs were curated in this work.

Once a high-confidence set of PPIs is curated in a species, we individually map each PPI to a high-resolution 3D structure. We use gapped BLAST searches between the PPI proteins' translated open reading frame sequences (ORFs) obtained on Ensembl [237] and biological unit structure subunit sequences from the Protein Data Bank (PDB) [15]. Several possible caveats and biases with this mapping process were identified and addressed. First, while rare, artifacts in PDB structures are possible. For instance, X-ray crystallography experiments can introduce crystal contacts artifacts between different chains that usually do not have a binding interface [102]. In order to distinguish "true" biological interfaces from fortuitous crystal-packing contacts in our data we only use structures annotated as biological assemblies on the PDB, and discard asymmetric units. We further exclude any biological assembly with "low" or "very low" confidence on the QSbio database [238]. Finally, we exclude biological assembly structures showing inconsistencies such as insufficient atomic detail, unreasonable distances between alpha-carbons, or non-sensible heavy atom counts. PPI structural data used in this work also comes with the set of biases typically associated with experimental 3D structural measurements. As the space of known structures in a species is not complete, PPI structural data may be biased towards proteins from particular cellular environments, more ancient and conserved proteins, commonly studied proteins, and highly expressed proteins [160,239]. To address these biases, for PPIs without a known structure in our species of interest, we use the structure of a closely related homologous PPI solved in another species if available, assuming that with high sequence conservation between two known PPIs, structural conservation must also be high [233,234]. We ensure that only structures from PPIs that are closely related to our PPI of interest are used, by applying both a stringent E-value cut-off and a high coverage cut-off to the alignment. In addition to reducing the above-mentioned biases in our structural data, the inclusion of such high-quality homology-based PPI structural models for PPIs

without a known structure also increases coverage. More than 700 *S. cerevisiae* PPIs, and more than 190 *S. pombe* PPIs were mapped to high-resolution 3D structures in this work.

Overall, we believe that the data curation process described in this thesis is the best and the most reliable existing method for integrating structural details with molecular evolutionary properties of PPIs on a whole-proteome scale. Moreover, this method will only improve as the spaces of known PPIs and known structures grow. The custom automated pipeline for curating, processing, and organizing PPI data into molecular models of PPIs for a species described here, as well as the molecular models of PPIs generated for *S. cerevisiae* and *S. pombe* in this work are publicly available and could be applied to future works in yeast or in other species.

The second aim of this thesis makes use of molecular models of PPIs in *S. cerevisiae* to study the relationship between PPI structure and PPI evolution. This analysis uncovers several strong, proteome-wide relationships between residue-level structural properties of PPI interfaces and residue evolutionary rate in baker's yeast. First, we find that interfacial residues are primarily constrained by their structural role in the complexed state (i.e. when the two partner proteins come together to form a PPI) rather than in the monomeric state (i.e. when the two partner proteins are free-floating monomers), underscoring their importance in maintaining PPI function and stability. Additionally, we observe evidence of a fixed evolutionary constraint associated with the function of the interface: if two residues have similar structural micro-environments but one is interfacial and the other is non-interfacial, the interfacial residue will typically be more evolutionarily conserved. Finally, we investigate structural constraints on residue evolution within PPI interfaces. We uncover significant, monotonic, and continuous relationships between interfacial residue

evolutionary rate and four structure-based measures of the overall involvement of a residue in an interface, with residues more involved in the interface evolving progressively more slowly. These results are surprising as interfacial residues experience two very different structural micro-environments depending on whether the two protein members of a PPI are in complexed state or in monomer state. In complexed state, interfacial residues are buried away from solvent, while in monomer state interfacial residues are surface residues, exposed to solvent. As such, one could expect the structural constraints on interfacial residue evolution to be a mixture between constraints imposed in the monomer state and constraints imposed in the complexed state. Contrary to this expectation, here, we find that the evolutionary behavior of interfacial buried residues mainly resembles the behavior of non-interfacial buried residues, and not the behavior of non-interfacial surface residues, indicating that interfacial residues are mainly constrained by structure when in complexed state. Moreover, we find structural constraints that are unique to the interface, both a fixed function-based, evolutionary constraint on any interfacial residue, and structure-based constraints within PPI interfaces, scaling continuously with a residue's degree of interfacial involvement. One possible explanation is that stable, permanent PPI complexes are over-represented in our data, and thus interfacial residues in these PPIs are rarely exposed constraints imposed in the monomer state [240]. However, we find that many proteins investigated in our study are instead involved in transient PPIs and can mostly be found as free-floating monomers in cells. We, therefore, conclude that the dominant role of the complexed state (rather than the monomeric state) in constraining interface evolution, and the unique structural constraints within PPI interfaces, likely reflect the importance of maintaining proper PPI function and stability, as disruption and mis-regulations of PPIs are known to have dire consequences for organismal fitness [12–14]. Future works investigating structural constraints on interfacial residue

evolution for transient and stable PPIs separately, using a framework similar to the one used in this study could be particularly interesting in order to further validate the conclusions reached here.

Work towards this aim also includes results which suggests that considering the close lineage of a species is important when studying evolutionary rates in PPIs. To compute evolutionary rate for individual residues in a PPI protein, the sequence of the protein is compared to aligned sequences of homologous proteins in related species. It is known that the choices of species and sequences included in this alignment can have a large effect on evolutionary rate values [209]. Indeed, both the structure of PPIs (e.g., presence or absence of interface) and the evolution of PPIs are highly species-specific and lineage-specific. Interactome network rewiring is known to be a widespread phenomenon, where PPIs existing in one species may be lost or rewired in another species [13]. Consequently, interfacial residues that are highly constrained and evolve slowly in one species due to the existence of the PPI may be completely free of such selective pressure and evolve much faster in a different species where the PPI is lost. Thus, a trade-off exists between accuracy and precision when more species are included in evolutionary rate calculations. Evolutionary rate estimates are likely more precise when more species are included due to the inclusion of additional data, but can be less accurate due to the possibility of PPI rewiring in one or more of the additional species included. Therefore, inclusion of more species does not necessarily lead to better estimation of evolutionary rates for highly lineage-specific processes such as PPI evolution. This point is further investigated in aim 3.

The third aim of this thesis makes use of molecular models of PPIs in *S. cerevisiae* and *S. pombe* to examine variations in PPIs (or PPI rewiring) between species. This analysis uncovers

some of the detailed molecular evolutionary mechanisms, and site-specific selective pressures underlying interactome network rewiring between two yeasts. First, we find that interfacial residues in both *S. cerevisiae* and *S. pombe* PPIs are significantly more conserved than non-interfacial residues when using a lineage-specific measure of residue evolutionary rates. This confirms that the results from our previous work in *S. cerevisiae* are also true in *S. pombe*. Furthermore, as *S. cerevisiae* and *S. pombe* are very distantly related evolutionarily the reproducibility discussed here could be evidence that our findings are fundamental and universal. Future works investigating structural constraints on interfacial residue evolution for additional species, using a framework similar to the one used in this study may be particularly interesting, and further establish the conclusions reached in this thesis as universal principles of PPI evolution. Additionally, we find that the difference in evolutionary rate between interfacial residues and non-interfacial residues is significantly less pronounced when using a non-lineage-specific measure of evolutionary rate. This surprising sensitivity further highlights that considering the close lineage of a species is important when studying evolutionary rates in PPIs. Finally, we establish that including more distantly related species in evolutionary rate calculations may be appropriate when studying more universally-conserved PPIs such as PPIs that are preserved between *S. cerevisiae* and *S. pombe*. In contrast, when investigating the loss or gain of an ortholog, or the rewiring of PPIs between the two species, using lineage-specific evolutionary rates is crucial. One possible explanation is that PPIs that are preserved between the two yeasts are more universal and less specific to a given lineage. Indeed, *S. cerevisiae* and *S. pombe* diverged from a common ancestor approximately 500 million years ago, and their genomes have since undergone significant changes [13, 23]. Non-lineage-specific measures of evolutionary rates may, therefore, be appropriate to the study of preserved PPIs. In contrast, PPIs that are different between the two yeast species, may be

unique to a species and its closely related lineage, and thus much more sensitive to the choice of species and lineages used in evolutionary rate calculations. Another possible biological mechanism for PPI rewiring between *S. pombe* and *S. cerevisiae* interactomes is that weaker PPIs or transient PPIs are more easily rewired between species with fewer amino acid changes [208]. Indeed, we find evidence of enrichment in weaker interactions for PPIs that are different or rewired between the two species' interactomes. We note that some experimental bias may also exist regarding our rewired PPI dataset, where weaker interactions are more likely to be detected in one species and not the other, even if they occur in both species. To fully address this possible experimental bias regarding our rewired PPI dataset, we constructed a separate dataset of missing ortholog PPIs as a gold-standard dataset of PPIs that are truly missing in only one of the two yeast species. Indeed, if a PPI occurs in *S. cerevisiae*, but one of the interacting protein partners has no ortholog or even homolog in *S. pombe*, we can be highly confident that the counterpart PPI does not exist in *S. pombe*, and vice versa. All our conclusions remain unchanged when the analyses are repeated on the missing ortholog PPI dataset (for which the above-mentioned experimental bias does not apply), demonstrating the robustness of our results and conclusions. Moreover, as highlighted in our previous article (Aim 2) and in other works [208–211] there is a high degree of heterogeneity within interfacial residues, as well as within non-interfacial residues. As more experimentally-determined protein-protein interactions and 3D structures become available, future works further sub-dividing both non-interfacial residues and interfacial residues and using a similar framework to the one used in this study could be particularly interesting to fully investigate the causes and consequences of structural and evolutionary heterogeneity within interfacial and non-interfacial residues in PPIs that are preserved and rewired between different species. Overall, this work yields insight into the evolutionary design principles of PPIs and the mechanisms by which interactions are preserved or

rewired between species, improving our understanding of the molecular evolution of PPI and PPI interfaces at the residue level.

This thesis advances our understanding of the evolutionary design principles of PPIs and the structural determinants influencing residue-level evolution both within a species, and between species. Our findings have wide-ranging applications which extends beyond furthering our understanding of fundamental biology and cellular processes, to various research fields including disease diagnosis, disease treatment, synthetic biology, and genome engineering. Moreover, the methodologies and findings from this thesis can be applied to future works in yeast or in other species, providing a framework for studying the evolutionary design principles of PPIs and the mechanisms by which interactions are preserved or rewired between species.

## Conclusion and summary

In conclusion, this thesis provides a comprehensive analysis of the structural and evolutionary principles underlying variations in PPIs, both within and between species. We designed a custom script pipeline to automate the curation of high-quality protein-protein interaction (PPI) data from online databases and organize this data into structural models of PPIs for the two yeast species, *S. cerevisiae*, and *S. pombe*. These structural models were subsequently used to investigate the relationship between PPI structure and PPI evolution in yeast at the single residue level. This analysis yielded significant insight into the design principles and structural mechanisms governing PPI evolution. Finally, we used structural models of *S. cerevisiae* and *S. pombe* PPIs to compare PPIs that are preserved and PPIs that are different between the two yeast species. This analysis yielded further insight into the evolutionary design principles of PPIs and the mechanisms by which interactions are preserved or rewired between species. Overall, this work establishes a better picture of the evolution of PPIs, both (1) at the molecular level, by uncovering small-scale structural properties that influence the evolution of PPIs within a species; and (2) at the phylogenetic level, by identifying mechanisms leading to large-scale differences in PPIs between species. These findings broaden our knowledge of PPI evolution as a whole. Moreover, the insights on natural, evolutionary, variations in PPIs both within and between species established in this work are crucial to the study of mis-regulation and disruption of PPIs associated with disease, as well as have wide-ranging applications to fields such as synthetic biology, and genome engineering.

## Master reference list


1. Omotayo, A.R., El-Ishaq, A., Tijjani, L.M. and Segun, D.I., 2016. Comparative analysis of protein content in selected meat samples (cow, rabbit, and chicken) obtained within damaturu metropolis. *American Journal of Food Science and Health*, *2*(6), pp.151-155.
2. De Las Rivas, J. and Fontanillo, C., 2010. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS computational biology*, *6*(6), p.e1000807.
3. Phizicky, E.M. and Fields, S., 1995. Protein-protein interactions: methods for detection and analysis. *Microbiological reviews*, *59*(1), pp.94-123.
4. Jones, S. and Thornton, J.M., 1996. Principles of protein-protein interactions. *Proceedings of the National Academy of Sciences*, *93*(1), pp.13-20.
5. Cohen, P., 2000. The regulation of protein function by multisite phosphorylation–a 25 year update. *Trends in biochemical sciences*, *25*(12), pp.596-601.
6. Acuner Ozbabacan, S.E., Engin, H.B., Gursoy, A. and Keskin, O., 2011. Transient protein–protein interactions. *Protein Engineering, Design & Selection*, *24*(9), pp.635-648.
7. Kuzmanov, U. and Emili, A., 2013. Protein-protein interaction networks: probing disease mechanisms using model systems. *Genome medicine*, *5*, pp.1-12.
8. Yohannes, D., 2003. Disruption of protein-protein interactions. pp.295-303.
9. Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S. and Timm, J., 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, *122*(6), pp.957-968.
10. Yu, H., Braun, P., Yıldırım, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N. and Hao, T., 2008. High-quality binary protein interaction map of the yeast interactome network. *Science*, *322*(5898), pp.104-110.
11. Vo, T.V., Das, J., Meyer, M.J., Cordero, N.A., Akturk, N., Wei, X., Fair, B.J., Degatano, A.G., Fragoza, R., Liu, L.G. and Matsuyama, A., 2016. A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. *Cell*, *164*(1), pp.310-323.
12. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M., 2006. BioGRID: a general repository for interaction datasets. *Nucleic acids research*, *34*(suppl_1), pp.D535-D539.
13. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., Del-Toro, N. and Duesbury, M., 2014. The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, *42*(D1), pp.D358-D363.
14. Lehne, B. and Schlitt, T., 2009. Protein-protein interaction databases: keeping up with growing interactomes. *Human genomics*, *3*, pp.1-7.
15. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E., 2000. The protein data bank. *Nucleic acids research*, *28*(1), pp.235-242.
16. Rosenbaum, D.M., Rasmussen, S.G. and Kobilka, B.K., 2009. The structure and function of G-protein-coupled receptors. *Nature*, *459*(7245), pp.356-363.
17. Hunter, T., 2000. Signaling—2000 and beyond. *Cell*, *100*(1), pp.113-127.
18. Johnson, L.N., 1992. Glycogen phosphorylase: control by phosphorylation and allosteric effectors. *The FASEB journal*, *6*(6), pp.2274-2282.

19. Merwe, P.A.V.D. and Davis, S.J., 2003. Molecular interactions mediating T cell antigen recognition. *Annual review of immunology*, *21*(1), pp.659-684.
20. Nooren, I.M. and Thornton, J.M., 2003. Diversity of protein–protein interactions. *The EMBO journal*, *22*(14), pp.3486-3492.
21. Fletcher, D.A. and Mullins, R.D., 2010. Cell mechanics and the cytoskeleton. *Nature*, *463*(7280), pp.485-492.
22. Beck, M. and Hurt, E., 2017. The nuclear pore complex: understanding its function through structural insight. *Nature reviews Molecular cell biology*, *18*(2), pp.73-89.
23. Bell, S.P. and Dutta, A., 2002. DNA replication in eukaryotic cells. *Annual review of biochemistry*, *71*(1), pp.333-374.
24. Yarden, Y. and Sliwkowski, M.X., 2001. Untangling the ErbB signalling network. *Nature reviews Molecular cell biology*, *2*(2), pp.127-137.
25. Stephen, A.G., Esposito, D., Bagni, R.K. and McCormick, F., 2014. Dragging ras back in the ring. *Cancer cell*, *25*(3), pp.272-281.
26. Vousden, K.H. and Prives, C., 2009. Blinded by the light: the growing complexity of p53. *Cell*, *137*(3), pp.413-431.
27. Gorre, M.E., Mohammed, M., Ellwood, K., Hsu, N., Paquette, R., Rao, P.N. and Sawyers, C.L., 2001. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science*, *293*(5531), pp.876-880.
28. Venkitaraman, A.R., 2002. Cancer susceptibility and the functions of BRCA1 and BRCA2. *Cell*, *108*(2), pp.171-182.
29. Karran, E. and De Strooper, B., 2016. The amyloid cascade hypothesis: are we poised for success or failure?. *Journal of neurochemistry*, *139*, pp.237-252.
30. Lim, K.L. and Zhang, C.W., 2013. Molecular events underlying Parkinson's disease–an interwoven tapestry. *Frontiers in neurology*, *4*, p.33.
31. Albin, R.L. and Greenamyre, J.T., 1992. Alternative excitotoxic hypotheses. *Neurology*, *42*(4), pp.733-733.
32. Li, S.H. and Li, X.J., 2004. Huntingtin–protein interactions and the pathogenesis of Huntington's disease. *TRENDS in Genetics*, *20*(3), pp.146-154.
33. Wyatt, R. and Sodroski, J., 1998. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science*, *280*(5371), pp.1884-1888.
34. Black, D.S. and Bliska, J.B., 2000. The RhoGAP activity of the Yersinia pseudotuberculosis cytotoxin YopE is required for antiphagocytic function and virulence. *Molecular microbiology*, *37*(3), pp.515-527.
35. Blair, J.M., Webber, M.A., Baylay, A.J., Ogbolu, D.O. and Piddock, L.J., 2015. Molecular mechanisms of antibiotic resistance. *Nature reviews microbiology*, *13*(1), pp.42-51.
36. Hadley, T.J., 1986. Invasion of erythrocytes by malaria parasites: a cellular and molecular overview. *Annual review of microbiology*, *40*, pp.451-477.
37. White, M.F., 2002. IRS proteins and the common path to diabetes. *American Journal of Physiology-Endocrinology and Metabolism*, *283*(3), pp.E413-E422.
38. Ishibashi, S., Brown, M.S., Goldstein, J.L., Gerard, R.D., Hammer, R.E. and Herz, J., 1993. Hypercholesterolemia in low density lipoprotein receptor knockout mice and its reversal by adenovirus-mediated gene delivery. *The Journal of clinical investigation*, *92*(2), pp.883-893.

39. Bjørbæk, C., Lavery, H.J., Bates, S.H., Olson, R.K., Davis, S.M., Flier, J.S. and Myers, M.G., 2000. SOCS3 mediates feedback inhibition of the leptin receptor via Tyr985. *Journal of Biological Chemistry*, *275*(51), pp.40649-40657.
40. Wallace, D.C., 1999. Mitochondrial diseases in man and mouse. *Science*, *283*(5407), pp.1482-1488.
41. Zhang, W., Zeng, T. and Chen, L., 2014. EdgeMarker: identifying differentially correlated molecule pairs as edge-biomarkers. *Journal of theoretical biology*, *362*, pp.35-43.
42. Chuang, H.Y., Lee, E., Liu, Y.T., Lee, D. and Ideker, T., 2007. Network-based classification of breast cancer metastasis. *Molecular systems biology*, *3*(1), p.140.
43. Luo, T., Wu, S., Shen, X. and Li, L., 2013. Network cluster analysis of protein–protein interaction network identified biomarker for early onset colorectal cancer. *Molecular biology reports*, *40*, pp.6561-6568.
44. Rezaei-Tavirani, M., Rezaei-Tavirani, S., Mansouri, V., Rostami-Nejad, M. and Rezaei-Tavirani, M., 2017. Protein-protein interaction network analysis for a biomarker panel related to human esophageal adenocarcinoma. *Asian Pacific journal of cancer prevention: APJCP*, *18*(12), p.3357.
45. Andreasson, U., Lautner, R., Schott, J.M., Mattsson, N., Hansson, O., Herukka, S.K., Helisalmi, S., Ewers, M., Hampel, H., Wallin, A. and Minthon, L., 2014. CSF biomarkers for Alzheimer's pathology and the effect size of APOE ε4. *Molecular psychiatry*, *19*(2), pp.148-149.
46. Wu, D., Li, Y., Zheng, L., Xiao, H., Ouyang, L., Wang, G. and Sun, Q., 2023. Small molecules targeting protein–protein interactions for cancer therapy. *Acta Pharmaceutica Sinica B*, *13*(10), pp.4060-4088.
47. Souers, A.J., Leverson, J.D., Boghaert, E.R., Ackler, S.L., Catron, N.D., Chen, J., Dayton, B.D., Ding, H., Enschede, S.H., Fairbrother, W.J. and Huang, D.C., 2013. ABT-199, a potent and selective BCL-2 inhibitor, achieves antitumor activity while sparing platelets. *Nature medicine*, *19*(2), pp.202-208.
48. Bourhis, J., Burtness, B., Licitra, L.F., Nutting, C., Schoenfeld, J.D., Ait Sarkouh, R., Bouisset, F., Nauwelaerts, H., Urfer, Y., Zanna, C. and Cohen, E.E., 2021. TrilynX: A phase 3 trial of xevinapant and concurrent chemoradiation for locally advanced head and neck cancer.
49. Granqvist, V., Holmgren, C. and Larsson, C., 2022. The combination of TRAIL and the Smac mimetic LCL-161 induces an irreversible phenotypic change of MCF-7 breast cancer cells. *Experimental and Molecular Pathology*, *125*, p.104739.
50. Rew, Y. and Sun, D., 2014. Discovery of a small molecule MDM2 inhibitor (AMG 232) for treating cancer. *Journal of medicinal chemistry*, *57*(15), pp.6332-6341.
51. Wang, T.H., Wang, H.S. and Soong, Y.K., 2000. Paclitaxel-induced cell death: where the cell cycle and apoptosis come together. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, *88*(11), pp.2619-2628.
52. Vassar, R., Kovacs, D.M., Yan, R. and Wong, P.C., 2009. The β-secretase enzyme BACE in health and Alzheimer's disease: regulation, cell biology, function, and therapeutic potential. *Journal of neuroscience*, *29*(41), pp.12787-12794.
53. Hazuda, D.J., Felock, P., Witmer, M., Wolfe, A., Stillmock, K., Grobler, J.A., Espeseth, A., Gabryelski, L., Schleif, W., Blau, C. and Miller, M.D., 2000. Inhibitors of strand transfer that prevent integration and inhibit HIV-1 replication in cells. *Science*, *287*(5453), pp.646-650.

54. Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.H., Nitsche, A. and Müller, M.A., 2020. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *cell*, *181*(2), pp.271-280.
55. Rivera, V.M., Clackson, T., Natesan, S., Pollock, R., Amara, J.F., Keenan, T., Magari, S.R., Phillips, T., Courage, N.L., Cerasoli Jr, F. and Holt, D.A., 1996. A humanized system for pharmacologic control of gene expression. *Nature medicine*, *2*(9), pp.1028-1032.
56. Sadelain, M., Brentjens, R. and Rivière, I., 2013. The basic principles of chimeric antigen receptor design. *Cancer discovery*, *3*(4), pp.388-398.
57. Beerli, R.R., Dreier, B. and Barbas III, C.F., 2000. Positive and negative regulation of endogenous genes by designed transcription factors. *Proceedings of the National Academy of Sciences*, *97*(4), pp.1495-1500.
58. Urry, D.W., 1997. Physical chemistry of biological free energy transduction as demonstrated by elastic protein-based polymers. *The Journal of Physical Chemistry B*, *101*(51), pp.11007-11028.
59. Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A. and Charpentier, E., 2012. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *science*, *337*(6096), pp.816-821.
60. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A. and Liu, D.R., 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, *533*(7603), pp.420-424.
61. Anzalone, A.V., Randolph, P.B., Davis, J.R., Sousa, A.A., Koblan, L.W., Levy, J.M., Chen, P.J., Wilson, C., Newby, G.A., Raguram, A. and Liu, D.R., 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, *576*(7785), pp.149-157.
62. Luck, K., Kim, D.K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B. and Choi, D., 2020. A reference map of the human binary protein interactome. *Nature*, *580*(7803), pp.402-408.
63. Guruharsha, K.G., Rual, J.F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D.Y., Cenaj, O. and McKillip, E., 2011. A protein complex network of Drosophila melanogaster. *Cell*, *147*(3), pp.690-703.
64. Alanis-Lobato, G., Möllmann, J.S., Schaefer, M.H. and Andrade-Navarro, M.A., 2020. MIPPIE: the mouse integrated protein–protein interaction reference. *Database*, *2020*, p.baaa035.
65. Arabidopsis Interactome Mapping Consortium, Dreze, M., Carvunis, A.R., Charloteaux, B., Galli, M., Pevzner, S.J., Tasan, M., Ahn, Y.Y., Balumuri, P., Barabási, A.L. and Bautista, V., 2011. Evidence for network evolution in an Arabidopsis interactome map. *Science*, *333*(6042), pp.601-607.
66. Fields, S. and Song, O.K., 1989. A novel genetic system to detect protein–protein interactions. *Nature*, *340*(6230), pp.245-246.
67. Vidal, M. and Legrain, P., 1999. Yeast forward and reverse 'n'-hybrid systems. *Nucleic acids research*, *27*(4), pp.919-929.
68. Semple, J.I., Sanderson, C.M. and Campbell, R.D., 2002. The jury is out on 'guilt by association'trials. *Briefings in Functional Genomics*, *1*(1), pp.40-52.
69. Rao, V.S., Srinivas, K., Sujini, G.N. and Kumar, G.S., 2014. Protein-protein interaction detection: methods and analysis. *International journal of proteomics*, *2014*(1), p.147648.

70. Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M. and Séraphin, B., 1999. A generic protein purification method for protein complex characterization and proteome exploration. *Nature biotechnology*, *17*(10), pp.1030-1032.
71. Gavin, A.C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M. and Remor, M., 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, *415*(6868), pp.141-147.
72. Aebersold, R. and Mann, M., 2003. Mass spectrometry-based proteomics. *Nature*, *422*(6928), pp.198-207.
73. Piston, D.W. and Kremers, G.J., 2007. Fluorescent protein FRET: the good, the bad and the ugly. *Trends in biochemical sciences*, *32*(9), pp.407-414.
74. Hu, C.D., Chinenov, Y. and Kerppola, T.K., 2002. Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation. *Molecular cell*, *9*(4), pp.789-798.
75. Kerppola, T.K., 2009. Visualization of molecular interactions using bimolecular fluorescence complementation analysis: characteristics of protein fragment complementation. *Chemical Society Reviews*, *38*(10), pp.2876-2886.
76. Moustaqil, M., Bhumkar, A., Gonzalez, L., Raoul, L., Hunter, D. J., Carrive, P., Sierecki, E. and Gambin, Y. 2017. A split-luciferase reporter recognizing GFP and mCherry tags to facilitate studies of protein–protein interactions. *International Journal of Molecular Sciences*, *18*(12), p.2681.
77. Azad, T., Tashakor, A. and Hosseinkhani, S., 2014. Split-luciferase complementary assay: applications, recent developments, and future perspectives. *Analytical and bioanalytical chemistry*, *406*, pp.5541-5560.
78. Harlow, E.D. and Lane, D., 1988. A laboratory manual. *New York: Cold Spring Harbor Laboratory*, *579*, p.44.
79. Homola, J., 2003. Present and future of surface plasmon resonance biosensors. *Analytical and bioanalytical chemistry*, *377*, pp.528-539.
80. Schasfoort, R.B. ed., 2017. *Handbook of surface plasmon resonance*. Royal Society of Chemistry.
81. Charih, F., Biggar, K.K. and Green, J.R., 2022. Assessing sequence-based protein–protein interaction predictors for use in therapeutic peptide engineering. *Scientific Reports*, *12*(1), p.9610.
82. Dick, K., Samanfar, B., Barnes, B., Cober, E.R., Mimee, B., Tan, L.H., Molnar, S.J., Biggar, K.K., Golshani, A., Dehne, F. and Green, J.R., 2020. Pipe4: Fast ppi predictor for comprehensive inter-and cross-species interactomes. *Scientific reports*, *10*(1), p.1390.
83. Li, Y. and Ilie, L., 2017. SPRINT: ultrafast protein-protein interaction prediction of the entire human interactome. *BMC bioinformatics*, *18*, pp.1-11.
84. Yao, Y., Du, X., Diao, Y. and Zhu, H., 2019. An integration of deep learning with feature embedding for protein–protein interaction prediction. *PeerJ*, *7*, p.e7126.
85. Chen, M., Ju, C.J.T., Zhou, G., Chen, X., Zhang, T., Chang, K.W., Zaniolo, C. and Wang, W., 2019. Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. *Bioinformatics*, *35*(14), pp.i305-i314.
86. Eisenberg, D., Marcotte, E.M., Xenarios, I. and Yeates, T.O., 2000. Protein function in the post-genomic era. *Nature*, *405*(6788), pp.823-826.

87. Clark, G.W., Dar, V.U.N., Bezginov, A., Yang, J.M., Charlebois, R.L. and Tillier, E.R., 2011. Using coevolution to predict protein–protein interactions. *Network Biology: Methods and Applications*, pp.237-256.
88. Chen, X.W. and Jeong, J.C., 2009. Sequence-based prediction of protein interaction sites with an integrative method. *Bioinformatics*, *25*(5), pp.585-591.
89. Wass, M.N., Fuentes, G., Pons, C., Pazos, F. and Valencia, A., 2011. Towards the prediction of protein interaction partners using physical docking. *Molecular systems biology*, *7*(1), p.469.
90. Lovell, S.C. and Robertson, D.L., 2010. An integrated view of molecular coevolution in protein–protein interactions. *Molecular biology and evolution*, *27*(11), pp.2567-2575.
91. Pazos, F. and Valencia, A., 2001. Similarity of phylogenetic trees as indicator of protein–protein interaction. *Protein engineering*, *14*(9), pp.609-614.
92. Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R. and Sander, C., 2011. Protein 3D structure computed from evolutionary sequence variation. *PloS one*, *6*(12), p.e28766.
93. Weigt, M., White, R.A., Szurmant, H., Hoch, J.A. and Hwa, T., 2009. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences*, *106*(1), pp.67-72.
94. Vakser, I.A., 2014. Protein-protein docking: From interaction to interactome. *Biophysical journal*, *107*(8), pp.1785-1793.
95. Janin, J., 2005. Assessing predictions of protein–protein interaction: the CAPRI experiment. *Protein science*, *14*(2), pp.278-283.
96. Karplus, M. and McCammon, J.A., 2002. Molecular dynamics simulations of biomolecules. *Nature structural biology*, *9*(9), pp.646-652.
97. Shaw, D.E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R.O., Eastwood, M.P., Bank, J.A., Jumper, J.M., Salmon, J.K., Shan, Y. and Wriggers, W., 2010. Atomic-level characterization of the structural dynamics of proteins. *Science*, *330*(6002), pp.341-346.
98. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P. and Jensen, L.J., 2019. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*, *47*(D1), pp.D607-D613.
99. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D., 2004. The database of interacting proteins: 2004 update. *Nucleic acids research*, *32*(suppl_1), pp.D449-D451.
100. Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E. and Castagnoli, L., 2012. MINT, the molecular interaction database: 2012 update. *Nucleic acids research*, *40*(D1), pp.D857-D861.
101. Buckle, A.M., Schreiber, G. and Fersht, A.R., 1994. Protein-protein recognition: Crystal structural analysis of a barnase-barstar complex at 2.0-. ANG. resolution. *Biochemistry*, *33*(30), pp.8878-8889.
102. Drenth, J., 2007. *Principles of protein X-ray crystallography*. Springer Science & Business Media.
103. Chapman, H.N., Fromme, P., Barty, A., White, T.A., Kirian, R.A., Aquila, A., Hunter, M.S., Schulz, J., DePonte, D.P., Weierstall, U. and Doak, R.B., 2011. Femtosecond X-ray protein nanocrystallography. *Nature*, *470*(7332), pp.73-77.

104. Gardner, K.H. and Kay, L.E., 1998. The use of 2h, 13c, 15n multidimensional nmr gto study the structure and dynamics of proteins. *Annual review of biophysics and biomolecular structure*, *27*(1), pp.357-406.
105. Clore, G.M. and Gronenborn, A.M., 1998. NMR structure determination of proteins and protein complexes larger than 20 kDa. *Current opinion in chemical biology*, *2*(5), pp.564-570.
106. Cheng, Y., 2015. Single-particle cryo-EM at crystallographic resolution. *Cell*, *161*(3), pp.450-457.
107. Nogales, E. and Scheres, S.H., 2015. Cryo-EM: a unique tool for the visualization of macromolecular complexity. *Molecular cell*, *58*(4), pp.677-689.
108. Wales, T.E. and Engen, J.R., 2006. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass spectrometry reviews*, *25*(1), pp.158-170.
109. Konermann, L., Pan, J. and Liu, Y.H., 2011. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chemical Society Reviews*, *40*(3), pp.1224-1234.
110. Sinz, A., 2006. Chemical cross-linking and mass spectrometry to map three-dimensional protein structures and protein–protein interactions. *Mass spectrometry reviews*, *25*(4), pp.663-682.
111. Leitner, A., Faini, M., Stengel, F. and Aebersold, R., 2016. Crosslinking and mass spectrometry: an integrated technology to understand the structure and function of molecular machines. *Trends in biochemical sciences*, *41*(1), pp.20-32.
112. Wang, S., Sun, S., Li, Z., Zhang, R. and Xu, J., 2017. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS computational biology*, *13*(1), p.e1005324.
113. Baker, D. and Sali, A., 2001. Protein structure prediction and structural genomics. *Science*, *294*(5540), pp.93-96.
114. Martí-Renom, M.A., Stuart, A.C., Fiser, A., Sánchez, R., Melo, F. and Šali, A., 2000. Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure*, *29*(1), pp.291-325.
115. Mosca, Roberto, Arnaud Céol, and Patrick Aloy. "Interactome3D: adding structural details to protein networks." *Nature methods* 10.1 (2013): 47-53.
116. Gromiha, M.M., Yugandhar, K. and Jemimah, S., 2017. Protein–protein interactions: scoring schemes and binding affinity. *Current opinion in structural biology*, *44*, pp.31-38.
117. Kumar, M.S. and Gromiha, M.M., 2006. PINT: protein–protein interactions thermodynamic database. *Nucleic acids research*, *34*(suppl_1), pp.D195-D198.
118. Vreven, T., Moal, I.H., Vangone, A., Pierce, B.G., Kastritis, P.L., Torchala, M., Chaleil, R., Jiménez-García, B., Bates, P.A., Fernandez-Recio, J. and Bonvin, A.M., 2015. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *Journal of molecular biology*, *427*(19), pp.3031-3041.
119. Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y. and Wang, R., 2015. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, *31*(3), pp.405-412.
120. Moal, I.H. and Fernández-Recio, J., 2012. SKEMPI: a structural kinetic and energetic database of mutant protein interactions and its use in empirical models. *Bioinformatics*, *28*(20), pp.2600-2607.
121. Morehead, A., Chen, C., Sedova, A. and Cheng, J., 2023. Dips-plus: The enhanced database of interacting protein structures for interface prediction. *Scientific data*, *10*(1), p.509.

122. Botstein, D., Chervitz, S.A. and Cherry, M., 1997. Yeast as a model organism. *Science*, *277*(5330), pp.1259-1260.
123. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. and Louis, E.J., 1996. Life with 6000 genes. *Science*, *274*(5287), pp.546-567.
124. Foury, F., 1997. Human genetic diseases: a cross-talk between man and yeast. *Gene*, *195*(1), pp.1-10.
125. Hartwell, L.H., 1974. Saccharomyces cerevisiae cell cycle. *Bacteriological reviews*, *38*(2), pp.164-198.
126. Nurse, P., 1990. Universal control mechanism regulating onset of M-phase. *Nature*, *344*(6266), pp.503-508.
127. Gustin, M.C., Albertyn, J., Alexander, M. and Davenport, K., 1998. MAP kinase pathways in the yeast Saccharomyces cerevisiae. *Microbiology and Molecular biology reviews*, *62*(4), pp.1264-1300.
128. Game, J.C., 1993, April. DNA double-strand breaks and the RAD50-RAD57 genes in Saccharomyces. In *Seminars in cancer biology* (Vol. 4, No. 2, pp. 73-83).
129. Symington, L.S., 2002. Role of RAD52 epistasis group genes in homologous recombination and double-strand break repair. *Microbiology and molecular biology reviews*, *66*(4), pp.630-670.
130. Laurenson, P. and Rine, J., 1992. Silencers, silencing, and heritable transcriptional states. *Microbiological reviews*, *56*(4), pp.543-560.
131. Ciechanover, A., 1998. The ubiquitin–proteasome pathway: on protein death and cell life. *The EMBO journal*.
132. Hartl, F.U., 1996. Molecular chaperones in cellular protein folding. *Nature*, *381*(6583), pp.571-580.
133. Hohmann, S., 2002. Osmotic stress signaling and osmoadaptation in yeasts. *Microbiology and molecular biology reviews*, *66*(2), pp.300-372.
134. Mendenhall, M.D. and Hodge, A.E., 1998. Regulation of Cdc28 cyclin-dependent protein kinase activity during the cell cycle of the yeast Saccharomyces cerevisiae. *Microbiology and Molecular Biology Reviews*, *62*(4), pp.1191-1243.
135. Weinert, T.A. and Hartwell, L.H., 1988. The RAD9 gene controls the cell cycle response to DNA damage in Saccharomyces cerevisiae. *Science*, *241*(4863), pp.317-322.
136. Gitler, A.D., Chesi, A., Geddie, M.L., Strathearn, K.E., Hamamichi, S., Hill, K.J., Caldwell, K.A., Caldwell, G.A., Cooper, A.A., Rochet, J.C. and Lindquist, S., 2009. α-Synuclein is part of a diverse and highly conserved interaction network that includes PARK9 and manganese toxicity. *Nature genetics*, *41*(3), pp.308-315.
137. Wood, V., Gwilliam, R., Rajandream, M.A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S. and Basham, D., 2002. The genome sequence of Schizosaccharomyces pombe. *Nature*, *415*(6874), pp.871-880.
138. Rhind, N., Chen, Z., Yassour, M., Thompson, D.A., Haas, B.J., Habib, N., Wapinski, I., Roy, S., Lin, M.F., Heiman, D.I. and Young, S.K., 2011. Comparative functional genomics of the fission yeasts. *Science*, *332*(6032), pp.930-936.
139. Russell, P. and Nurse, P., 1986. cdc25+ functions as an inducer in the mitotic control of fission yeast. *Cell*, *45*(1), pp.145-153.
140. Gould, K.L. and Nurse, P., 1989. Tyrosine phosphorylation of the fission yeast cdc2+ protein kinase regulates entry into mitosis. *Nature*, *342*(6245), pp.39-45.

141. Walworth, N.C. and Bernards, R., 1996. rad-dependent response of the chk1-encoded protein kinase at the DNA damage checkpoint. *Science*, *271*(5247), pp.353-356.
142. Rhind, N., Furnari, B. and Russell, P., 1997. Cdc2 tyrosine phosphorylation is required for the DNA damage checkpoint in fission yeast. *Genes & development*, *11*(4), pp.504-511.
143. Allshire, R.C. and Karpen, G.H., 2008. Epigenetic regulation of centromeric chromatin: old dogs, new tricks?. *Nature Reviews Genetics*, *9*(12), pp.923-937.
144. Moser, B.A. and Nakamura, T.M., 2009. Protection and replication of telomeres in fission yeast. *Biochemistry and cell biology*, *87*(5), pp.747-758.
145. Volpe, T.A., Kidner, C., Hall, I.M., Teng, G., Grewal, S.I. and Martienssen, R.A., 2002. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science*, *297*(5588), pp.1833-1837.
146. Weisman, R. and Choder, M., 2001. The fission yeast TOR homolog, tor1+, is required for the response to starvation and other stresses via a conserved serine. *Journal of Biological Chemistry*, *276*(10), pp.7027-7032.
147. Hagan, I.M. and Hyams, J.S., 1988. The use of cell division cycle mutants to investigate the control of microtubule distribution in the fission yeast Schizosaccharomyces pombe. *Journal of cell science*, *89*(3), pp.343-357.
148. Baumann, P. and Cech, T.R., 2000. Protection of telomeres by the Ku protein in fission yeast. *Molecular biology of the cell*, *11*(10), pp.3265-3275.
149. Hayashi, M.T., 2017. Telomere biology in aging and cancer: early history and perspectives. *Genes & genetic systems*, *92*(3), pp.107-118.
150. Jeffares, D.C., Rallis, C., Rieux, A., Speed, D., Převorovský, M., Mourier, T., Marsellach, F.X., Iqbal, Z., Lau, W., Cheng, T.M. and Pracana, R., 2015. The genomic and phenotypic diversity of Schizosaccharomyces pombe. *Nature genetics*, *47*(3), pp.235-241.
151. Seoighe, C. and Wolfe, K.H., 1998. Extent of genomic rearrangement after genome duplication in yeast. *Proceedings of the National Academy of Sciences*, *95*(8), pp.4447-4452.
152. Wolfe, K.H. and Shields, D.C., 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, *387*(6634), pp.708-713.
153. Duina, A.A., Miller, M.E. and Keeney, J.B., 2014. Budding yeast for budding geneticists: a primer on the Saccharomyces cerevisiae model system. *Genetics*, *197*(1), pp.33-48.
154. Nurse, P. and Bissett, Y., 1981. Gene required in G1 for commitment to cell cycle and in G2 for control of mitosis in fission yeast. *Nature*, *292*(5823), pp.558-560.
155. Hoffman, C.S., Wood, V. and Fantes, P.A., 2015. An ancient yeast for young geneticists: a primer on the Schizosaccharomyces pombe model system. *Genetics*, *201*(2), pp.403-423.
156. Holoch, D. and Moazed, D., 2015. RNA-mediated epigenetic regulation of gene expression. *Nature Reviews Genetics*, *16*(2), pp.71-84.
157. Goldstein, R.A., 2008. The structure of protein evolution and the evolution of protein structure. *Current opinion in structural biology*, *18*(2), pp.170-177.
158. Liberles, D.A., Teichmann, S.A., Bahar, I., Bastolla, U., Bloom, J., Bornberg-Bauer, E., Colwell, L.J., De Koning, A.J., Dokholyan, N.V., Echave, J. and Elofsson, A., 2012. The interface of protein structure, protein biophysics, and molecular evolution. *Protein Science*, *21*(6), pp.769-785.
159. Hubbard, T.J.P. and Blundell, T.L., 1987. Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Engineering, Design and Selection*, *1*(3), pp.159-171.

160. Franzosa, E.A. and Xia, Y., 2009. Structural determinants of protein evolution are context-sensitive at the residue level. *Molecular biology and evolution*, *26*(10), pp.2387-2395.
161. Kimura, M. and Ohta, T., 1974. On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences*, *71*(7), pp.2848-2852.
162. Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C. and Feldman, M.W., 2002. Evolutionary rate in the protein interaction network. *Science*, *296*(5568), pp.750-752.
163. Zhang, J. and Yang, J.R., 2015. Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics*, *16*(7), pp.409-420.
164. McInerney, J.O., 2006. The causes of protein evolutionary rate variation. *Trends in ecology & evolution*, *21*(5), pp.230-232.
165. Bartlett, G.J., Porter, C.T., Borkakoti, N. and Thornton, J.M., 2002. Analysis of catalytic residues in enzyme active sites. *Journal of molecular biology*, *324*(1), pp.105-121.
166. Echave, J., Spielman, S.J. and Wilke, C.O., 2016. Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, *17*(2), pp.109-121.
167. Pál, C., Papp, B. and Lercher, M.J., 2006. An integrated view of protein evolution. *Nature reviews genetics*, *7*(5), pp.337-348.
168. Li, W.H., Wu, C.I. and Luo, C.C., 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular biology and evolution*, *2*(2), pp.150-174.
169. Nei, M. and Gojobori, T., 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular biology and evolution*, *3*(5), pp.418-426.
170. Goldman, N. and Yang, Z., 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular biology and evolution*, *11*(5), pp.725-736.
171. Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences*, *13*(5), pp.555-556.
172. Pupko, T., Bell, R.E., Mayrose, I., Glaser, F. and Ben-Tal, N., 2002. Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, *18*(suppl_1), pp.S71-S77.
173. Mayrose, I., Graur, D., Ben-Tal, N. and Pupko, T., 2004. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Molecular biology and evolution*, *21*(9), pp.1781-1791.
174. Armon, A., Graur, D. and Ben-Tal, N., 2001. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *Journal of molecular biology*, *307*(1), pp.447-463.
175. Ben Chorin, A., Masrati, G., Kessel, A., Narunsky, A., Sprinzak, J., Lahav, S., Ashkenazy, H. and Ben-Tal, N., 2020. ConSurf-DB: An accessible repository for the evolutionary conservation patterns of the majority of PDB proteins. *Protein Science*, *29*(1), pp.258-267.
176. Sydykova, D.K. and Wilke, C.O., 2017. Calculating site-specific evolutionary rates at the amino-acid or codon level yields similar rate estimates. *PeerJ*, *5*, p.e3391.
177. Rocha, E.P., 2006. The quest for the universals of protein evolution. *Trends in genetics*, *22*(8), pp.412-416.
178. Bloom, J.D., Drummond, D.A., Arnold, F.H. and Wilke, C.O., 2006. Structural determinants of the rate of protein evolution in yeast. *Molecular biology and evolution*, *23*(9), pp.1751-1761.

179. Perutz, M.F., Kendrew, J.C. and Watson, H.C., 1965. Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. *Journal of Molecular Biology*, *13*(3), pp.669-678.
180. Tien, M.Z., Meyer, A.G., Sydykova, D.K., Spielman, S.J. and Wilke, C.O., 2013. Maximum allowed solvent accessibilites of residues in proteins. *PloS one*, *8*(11), p.e80635.
181. Overington, J., Donnelly, D., Johnson, M.S., Šali, A. and Blundell, T.L., 1992. Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Science*, *1*(2), pp.216-226.
182. Goldman, N., Thorne, J.L. and Jones, D.T., 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, *149*(1), pp.445-458.
183. Bustamante, C.D., Townsend, J.P. and Hartl, D.L., 2000. Solvent accessibility and purifying selection within proteins of Escherichia coli and Salmonella enterica. *Molecular biology and evolution*, *17*(2), pp.301-308.
184. Choi, S.S., Vallender, E.J. and Lahn, B.T., 2006. Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. *Molecular biology and evolution*, *23*(11), pp.2131-2133.
185. Conant, G.C. and Stadler, P.F., 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Molecular biology and evolution*, *26*(5), pp.1155-1161.
186. Franzosa, E. and Xia, Y., 2008. Structural perspectives on protein evolution. *Annu Rep Comput Chem*, *4*(1), pp.3-21.
187. Dean, A.M., Neuhauser, C., Grenier, E. and Golding, G.B., 2002. The pattern of amino acid replacements in α/β-barrels. *Molecular biology and evolution*, *19*(11), pp.1846-1864.
188. Conant, G.C. and Stadler, P.F., 2009. Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Molecular biology and evolution*, *26*(5), pp.1155-1161.
189. Ramsey, D.C., Scherrer, M.P., Zhou, T. and Wilke, C.O., 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics*, *188*(2), pp.479-488.
190. Scherrer, M.P., Meyer, A.G. and Wilke, C.O., 2012. Modeling coding-sequence evolution within the context of residue solvent accessibility. *BMC Evolutionary Biology*, *12*, pp.1-11.
191. Shih, C.H., Chang, C.M., Lin, Y.S., Lo, W.C. and Hwang, J.K., 2012. Evolutionary information hidden in a single protein structure. *Proteins: Structure, Function, and Bioinformatics*, *80*(6), pp.1647-1657.
192. Lin, C.P., Huang, S.W., Lai, Y.L., Yen, S.C., Shih, C.H., Lu, C.H., Huang, C.C. and Hwang, J.K., 2008. Deriving protein dynamical properties from weighted protein contact number. *Proteins: Structure, Function, and Bioinformatics*, *72*(3), pp.929-935.
193. Yeh, S.W., Huang, T.T., Liu, J.W., Yu, S.H., Shih, C.H., Hwang, J.K. and Echave, J., 2014. Local packing density is the main structural determinant of the rate of protein sequence evolution at site level. *BioMed research international*, *2014*(1), p.572409.
194. Chothia, C. and Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, *5*(4), pp.823-826.
195. Sitbon, E. and Pietrokovski, S., 2007. Occurrence of protein structure elements in conserved sequence regions. *BMC structural biology*, *7*, pp.1-15.
196. Fersht, A., 1999. Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding. In *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding* (pp. 631-p).

197. Brown, C.J., Takayama, S., Campen, A.M., Vise, P., Marshall, T.W., Oldfield, C.J., Williams, C.J. and Keith Dunker, A., 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *Journal of molecular evolution*, *55*(1).
198. Gerek, Z.N., Kumar, S. and Ozkan, S.B., 2013. Structural dynamics flexibility informs function and evolution at a proteome scale. *Biophysical Journal*, *104*(2), p.228a.
199. Marsh, J.A. and Teichmann, S.A., 2014. Parallel dynamics and evolution: Protein conformational fluctuations and assembly reflect evolutionary changes in sequence and structure. *BioEssays*, *36*(2), pp.209-218.
200. Shahmoradi, A., Sydykova, D.K., Spielman, S.J., Jackson, E.L., Dawson, E.T., Meyer, A.G. and Wilke, C.O., 2014. Predicting evolutionary site variability from structure in viral proteins: buriedness, packing, flexibility, and design. *Journal of molecular evolution*, *79*, pp.130-142.
201. Ngounou Wetie, A.G., Sokolowska, I., Woods, A.G., Roy, U., Loo, J.A. and Darie, C.C., 2013. Investigation of stable and transient protein–protein interactions: past, present, and future. *Proteomics*, *13*(3-4), pp.538-557.
202. Engin, H.B., Kreisberg, J.F. and Carter, H., 2016. Structure-based analysis reveals cancer missense mutations target protein interaction interfaces. *PloS one*, *11*(4), p.e0152929.
203. Valdar, W.S. and Thornton, J.M., 2001. Protein–protein interfaces: analysis of amino acid conservation in homodimers. *Proteins: Structure, Function, and Bioinformatics*, *42*(1), pp.108-124.
204. Bloom, J.D., Drummond, D.A., Arnold, F.H. and Wilke, C.O., 2006. Structural determinants of the rate of protein evolution in yeast. *Molecular biology and evolution*, *23*(9), pp.1751-1761.
205. Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J. and Huang, E.S., 2004. Are protein–protein interfaces more conserved in sequence than the rest of the protein surface?. *Protein Science*, *13*(1), pp.190-202.
206. Ma, B., Elkayam, T., Wolfson, H. and Nussinov, R., 2003. Protein–protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proceedings of the National Academy of Sciences*, *100*(10), pp.5772-5777.
207. Mintseris, J. and Weng, Z., 2005. Structure, function, and evolution of transient and obligate protein–protein interactions. *Proceedings of the National Academy of Sciences*, *102*(31), pp.10930-10935.
208. Levy, E.D., 2010. A simple definition of structural regions in proteins and its use in analyzing interface evolution. *Journal of molecular biology*, *403*(4), pp.660-670.
209. Duarte, J.M., Srebniak, A., Schärer, M.A. and Capitani, G., 2012. Protein interface classification by evolutionary analysis. *BMC bioinformatics*, *13*, pp.1-16.
210. Moreira, I.S., Fernandes, P.A. and Ramos, M.J., 2007. Hot spots—A review of the protein–protein interface determinant amino-acid residues. *Proteins: Structure, Function, and Bioinformatics*, *68*(4), pp.803-812.
211. Keskin, O., Ma, B. and Nussinov, R., 2005. Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. *Journal of molecular biology*, *345*(5), pp.1281-1294.
212. Ivankov, D.N., Finkelstein, A.V. and Kondrashov, F.A., 2014. A structural perspective of compensatory evolution. *Current opinion in structural biology*, *26*, pp.104-112.
213. Lovell, S.C. and Robertson, D.L., 2010. An integrated view of molecular coevolution in protein–protein interactions. *Molecular biology and evolution*, *27*(11), pp.2567-2575.

214. Leducq, J.B., Charron, G., Diss, G., Gagnon-Arsenault, I., Dubé, A.K. and Landry, C.R., 2012. Evidence for the robustness of protein complexes to inter-species hybridization. *PLoS genetics*, *8*(12), p.e1003161.
215. Wagner, A., 2001. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Molecular biology and evolution*, *18*(7), pp.1283-1292.
216. He, X. and Zhang, J., 2005. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics*, *169*(2), pp.1157-1164.
217. Soucy, S.M., Huang, J. and Gogarten, J.P., 2015. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, *16*(8), pp.472-482.
218. Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B. and Brar, G.A., 2012. Proto-genes and de novo gene birth. *Nature*, *487*(7407), pp.370-374.
219. Abrusán, G., 2013. Integration of new genes into cellular networks, and their structural maturation. *Genetics*, *195*(4), pp.1407-1417.
220. Wan, C., Borgeson, B., Phanse, S., Tu, F., Drew, K., Clark, G., Xiong, X., Kagan, O., Kwan, J., Bezginov, A. and Chessman, K., 2015. Panorama of ancient metazoan macromolecular complexes. *Nature*, *525*(7569), pp.339-344.
221. Beltrao, P. and Serrano, L., 2007. Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Computational Biology*, *3*(2), p.e25.
222. Sun, M.G., Sikora, M., Costanzo, M., Boone, C. and Kim, P.M., 2012. Network evolution: rewiring and signatures of conservation in signaling. *PLoS computational biology*, *8*(3), p.e1002411.
223. Cesareni, G., Ceol, A., Gavrila, C., Palazzi, L.M., Persico, M. and Schneider, M.V., 2005. Comparative interactomics. *FEBS letters*, *579*(8), pp.1828-1833.
224. Gandhi, T.K.B., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K.N., Mohan, S.S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B. and Mishra, G., 2006. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nature genetics*, *38*(3), pp.285-293.
225. Yamada, T. and Bork, P., 2009. Evolution of biomolecular networks—lessons from metabolic and protein interactions. *Nature Reviews Molecular Cell Biology*, *10*(11), pp.791-803.
226. Shou, C., Bhardwaj, N., Lam, H.Y., Yan, K.K., Kim, P.M., Snyder, M. and Gerstein, M.B., 2011. Measuring the evolutionary rewiring of biological networks. *PLoS computational biology*, *7*(1), p.e1001050.
227. Xin, X., Gfeller, D., Cheng, J., Tonikian, R., Sun, L., Guo, A., Lopez, L., Pavlenco, A., Akintobi, A., Zhang, Y. and Rual, J.F., 2013. SH3 interactome conserves general function over specific form. *Molecular systems biology*, *9*(1), p.652.
228. Reinke, A.W., Baek, J., Ashenberg, O. and Keating, A.E., 2013. Networks of bZIP protein-protein interactions diversified over a billion years of evolution. *Science*, *340*(6133), pp.730-734.
229. Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S. and Bork, P., 2002. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, *417*(6887), pp.399-403.
230. Wodak, S.J., Vlasblom, J., Turinsky, A.L. and Pu, S., 2013. Protein–protein interaction networks: the puzzling riches. *Current opinion in structural biology*, *23*(6), pp.941-953.

231. Cusick, Michael E., et al. "Interactome: gateway into systems biology." *Human molecular genetics* 14.suppl_2 (2005): R171-R181.
232. Aloy, P. and Russell, R.B., 2006. Structural systems biology: modelling protein interactions.
233. Pollet, L., Lambourne, L. and Xia, Y., 2022. Structural Determinants of Yeast Protein-Protein Interaction Interface Evolution at the Residue Level. *Journal of Molecular Biology*, *434*(19), p.167750.
234. Pollet, L. and Xia, Y., 2024. Structure-guided evolutionary analysis of interactome network rewiring at single residue resolution in yeasts. *Journal of Molecular Biology*, p.168641.
235. Vidal, M., 2016. How much of the human protein interactome remains to be mapped?. *Science signaling*, *9*(427), pp.eg7-eg7.
236. Hakes, L., Pinney, J.W., Robertson, D.L. and Lovell, S.C., 2008. Protein-protein interaction networks and biology—what's the connection?. *Nature biotechnology*, *26*(1), pp.69-72.
237. Yates, A.D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Azov, A.G., Bennett, R. and Bhai, J., 2020. Ensembl 2020. *Nucleic acids research*, *48*(D1), pp.D682-D688.
238. Dey, S., Ritchie, D.W. and Levy, E.D., 2018. PDB-wide identification of biological assemblies from conserved quaternary structure geometry. *Nature methods*, *15*(1), pp.67-72.
239. Hart, G.T., Ramani, A.K. and Marcotte, E.M., 2006. How complete are current yeast and human protein-interaction networks?. *Genome biology*, *7*, pp.1-9.
240. Perkins, J.R., Diboun, I., Dessailly, B.H., Lees, J.G. and Orengo, C., 2010. Transient protein-protein interactions: structural, functional, and network properties. *Structure*, *18*(10), pp.1233-1243.

# Appendix 1

**Supplementary information for Chapter 3**

**Structural Determinants of Yeast Protein-Protein interaction Interface Evolution at the Residue Level**

Léah Pollet [1], Luke Lambourne [2,3,4*] and Yu Xia [1*]

1 - Department of Bioengineering, Faculty of Engineering, McGill University, Montreal, QC, Canada

2 - Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA, USA

3 - Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA, USA

4 - Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA

Correspondence to Luke Lambourne and Yu Xia: Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA, USA (L. Lambourne).

luke_lambourne@dfci.harvard.edu (L. Lambourne), brandon.xia@mcgill.ca (Y. Xia)

https://doi.org/10.1016/j.jmb.2022.167750

Edited by Michael Sternberg

*Modeling the relationship between individual structural properties and ConSurf score*

We studied the relationship between individual structural property and *ConSurf* scores in PPI interfaces using a weighted least-square regression technique that takes the error associated with calculating *ConSurf* for each residue bin into account. For each structural property of interest, the regression model follows the equation:

$$y(x) = w_0 + w_1 x_1 + e_x$$

Where $y(x)$ is the average *ConSurf* score of residues in bin $x$ (residues are binned in 10% intervals over the range of each structural property), $x_1$ is the center value of bin $x$ for the structural property investigated, $w_0, w_1$ are the intercept, and the weight associated with the structural feature in the regression model, and $e_x$ is a random variable ("noise term") following a Gaussian distribution with zero mean and standard deviation equal to the standard error associated with the *ConSurf* score for bin $x$. One model was trained for each structural property and the resulting linear fits can be seen in **Figure S1** and **Figure S2**.

*Supplementary analysis performed with additional species included in evolutionary rate calculations*

**Analysis S1** includes additional species in dN/dS calculations.
**Analysis S2** includes additional species in ConSurf score calculations using the rate4site program.
**Analysis S3** uses pre-computed ConSurf scores downloaded from the ConSurf database.

The set related species considered in **Analysis S1** and **S2** is composed of the 8 original species: *Saccharomyces paradoxus (S. paradoxus)*, *Saccharomyces mikatae (S. mikatae)*, *Saccharomyces bayanus (S. bayanus)*, *Naumovozyma castellii (N. castellii)*, *Candida glabrata (C. glabrata), Eremothecium gossypii (E. gossypii), Kluyveromyces lactis (K. lactis)* and *Candida albicans (C. albicans)*; and 8 additional, and more distantly related species: *Fusarium graminearum (F. graminearum), Neurospora crassa (N. crassa), Aspergillus nidulans (A. nidulans), Schizosaccharomyces pombe (S. pombe), Neolecta irregularis (N. irregularis), Protomyces lactucaedebilis (P. lactucaedebilis), Pneumocystis jirovecii (P. jirovecii)* and *Pneumocystis murina (P. murina)*. The phylogenetic tree used in evolutionary rate calculations is as follows:
[[[[[[[*S. cerevisiae*, *S. paradoxus*], [*S. mikatae*, *S. bayanus*]], *C. glabrata*], *N. castellii*], [[*E. gossypii*, *K. lactis*], *C. albicans*]], [[*F. graminearum*, *N. crassa*], *A. nidulans*]], [*S. pombe*, [*N. irregularis*, [*P. lactucaedebilis*, [*P. jirovecii*, *P. murina*]]]]];.

**Analysis S3**, uses downloaded evolutionary rates from the ConSurf database. The ConSurf database provides pre-computed evolutionary rates for structures on the PDB. ConSurf database evolutionary rate calculations are performed using the rate4site program on MSAs constructed using PSI-BLAST with an e-value cutoff of to $10^{-3}$ to find potential homologs in the UniProtKB/SwissProt database for proteins on the PDB.

*Supplementary analysis performed with low percent sequence identity PPIs excluded*

**Analysis S4** excludes PPIs for which either of the two partner proteins has sequence identity lower than 50% between their yeast protein sequence and the PDB protein sequence used to compute structural properties. Instead, structural property calculations, evolutionary rate calculations, and all subsequent analysis were performed for 428 high-percent-sequence-identity PPIs.

| | **ΔRSA** | **InterRRC** | **dCenter** | **dEdges** |
|---|---|---|---|---|
| Correlation with *dN/dS* ratio | -0.889 ** | 0.490 | 0.828 ** | -0.545 |
| Correlation with *ConSurf* score | -0.044 ** | -0.075 ** | 0.078 ** | -0.106 ** |
| Correlation with *dN/dS* ratio (weighted) | -0.987 ** | 0.272 | 0.897 ** | -0.014 |
| Correlation with *ConSurf* score (weighted) | -0.032 ** | -0.052 ** | 0.046 ** | -0.063 ** |
| Correlation with *ConSurf* score (*ConSurf* DB) | -0.063 ** | -0.138 ** | 0.192 ** | -0.070 ** |

**Table S1.** Results of a Pearson product-moment correlation test between structural measures of interface involvement and evolutionary rate estimates (computed from a larger set of aligned related species - **Analysis S1, S2 and S3**) for interfacial residues. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**), values significant at the P-value < 0.05 level are denoted with a single asterisk (*).

| | **Linear regression with ConSurf score** | **Linear regression with ConSurf DB score** |
|---|---|---|
| **Structural properties included in the model** | $R^2$ | $R^2$ |
| Monomer RSA | 8.79% | 3.9% |
| Monomer RSA + ΔRSA | 15.12% | 7.04% |
| Monomer RSA + interRRC | 12.46% | 5.5% |
| Monomer RSA + dCenter | 10.66% | 4.02% |
| Monomer RSA + dEdges | 9% | 3.52% |
| Monomer RSA + ΔRSA + interRRC + dCenter + dEdges | 16.61% | 7.79% |

**Table S2.** Regression results for different models aiming to predict residue evolutionary rate (*ConSurf* score, computed from a larger set of aligned related species and *ConSurf* score from *ConSurf* DB – **Analysis S1, S2 and S3**) from structural properties in PPI interfaces. All models were trained using 10-fold cross-validation, and the results shown here are average adjusted $R^2$ values across all cross-validation trials.

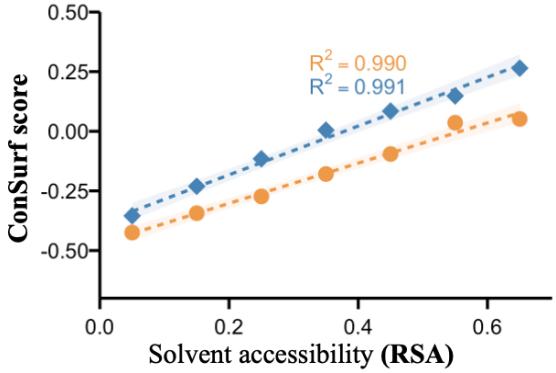
| | **ΔRSA** | **InterRRC** | **dCenter** | **dEdges** |
|---|---|---|---|---|
| Correlation with *dN/dS* ratio | -0.716 ** | -0.486 | 0.827 ** | -0.821 * |
| Correlation with *ConSurf* score | -0.041 ** | -0.069 ** | 0.077 ** | -0.111 ** |
| Correlation with *dN/dS* ratio (weighted) | -0.902 ** | -0.253 | 0.855 ** | -0.893 ** |
| Correlation with *ConSurf* score (weighted) | -0.021 ** | -0.043 ** | 0.040 ** | -0.057 ** |

**Table S3.** Results of a Pearson product-moment correlation test between structural measures of interface involvement and evolutionary rate estimates for interfacial residues (for high-sequence-identity PPIs – **Analysis S4**). Values significant at the P-value < 0.01 level are denoted with a double asterisk (**), values significant at the P-value < 0.05 level are denoted with a single asterisk (*).

| | **Linear regression with ConSurf score** |
|---|---|
| **Structural properties included in the model** | $R^2$ |
| Monomer RSA | 8.71% |
| Monomer RSA + ΔRSA | 14.9% |
| Monomer RSA + interRRC | 11.45% |
| Monomer RSA + dCenter | 9.57% |
| Monomer RSA + dEdges | 8.9% |
| Monomer RSA + ΔRSA + interRRC + dCenter + dEdges | 16.15% |

**Table S4.** Regression results for different models aiming to predict residue evolutionary rate from structural properties in PPI interfaces (for high-sequence-identity PPIs – **Analysis S4**). All models were trained using 10-fold cross-validation, and the results shown here are average adjusted $R^2$ values across all cross-validation trials.

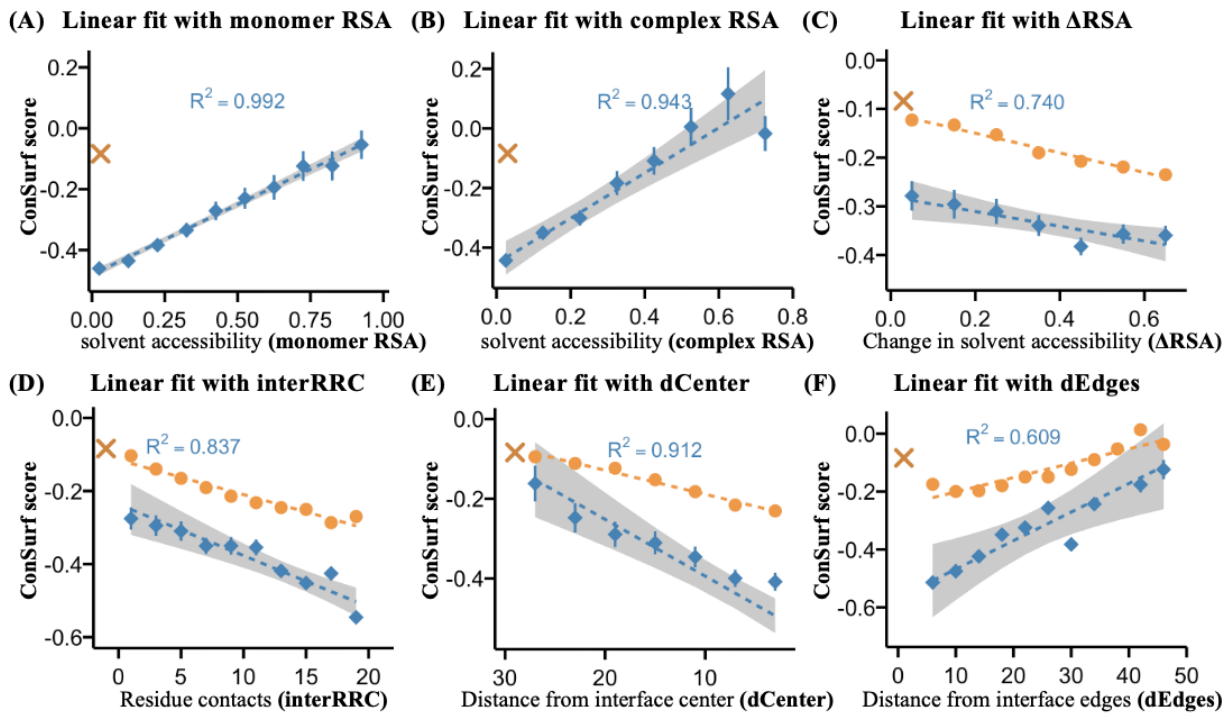
**(A) Linear fits with solvent accessibility**




**(B) Statistics for the linear fits**

| | Difference in intercept | Difference in slope |
|---|---|---|
| Interfacial residues fit vs. non-interfacial residues fit | 0.0838 ** | 0.1793 |


**Figure S1. The difference in evolutionary rate between interfacial and non-interfacial residues. (A)** Linear fits between binned measures of solvent accessibility (complex RSA) and evolutionary rate (*ConSurf* score) for interfacial and non-interfacial residues. Weighted linear regression lines, 95% confidence interval, and $R^2$ values of the fits are also shown. **(B)** Results of t-tests for differences in slope and intercept between the two fits in **(A)**. Values significant at the P-value < 0.05 level are denoted with a single asterisk (**).




**(G) Statistics for the linear fits**

| | Difference in intercept between observed and expected fit | Difference in slope between observed and expected fit |
|---|---|---|
| ΔRSA | 0.1706 ** | 0.0497 |
| interRRC | 0.1214 ** | 0.0046 |
| dCenter | 0.2745 ** | 0.0075 ** |
| dEdges | 0.3160 * | 0.0049 |

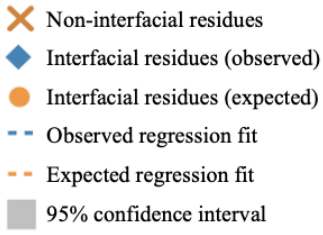



**Figure S2. The relationship between structural properties and evolutionary rate for residues in PPI interfaces. (A)-(F)** Linear fits in blue between binned structural properties and evolutionary rate (*ConSurf* score), for relative solvent accessibility in monomer state (monomer RSA), relative solvent accessibility in complex state (complex RSA), change in burial upon complex formation (ΔRSA), inter-protein residue-residue contacts (interRRC), distance from interface center (dCenter), and distance from interface edges (dEdges) respectively. Weighted linear regression lines, 95% confidence interval, and $R^2$ values of the fits are also shown. In addition to the observed fit in blue, we also show the expected fit in red for the structural properties that measure interfacial involvement (**(C)-(F)**), assuming that interfacial burial and non-interfacial burial are selectively equivalent and that interfacial residues are subject to the same evolutionary constraints as non-interfacial residues with the same total burial in PPIs. For each panel, the average *ConSurf* score for non-interfacial residues (average *ConSurf* score = -0.083) is marked by a yellow "X". **(G)** Results of t-tests for differences in slope and intercept between observed and expected fits in **(C)-(F)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**), values significant at the P-value < 0.05 level are denoted with a single asterisk (*).

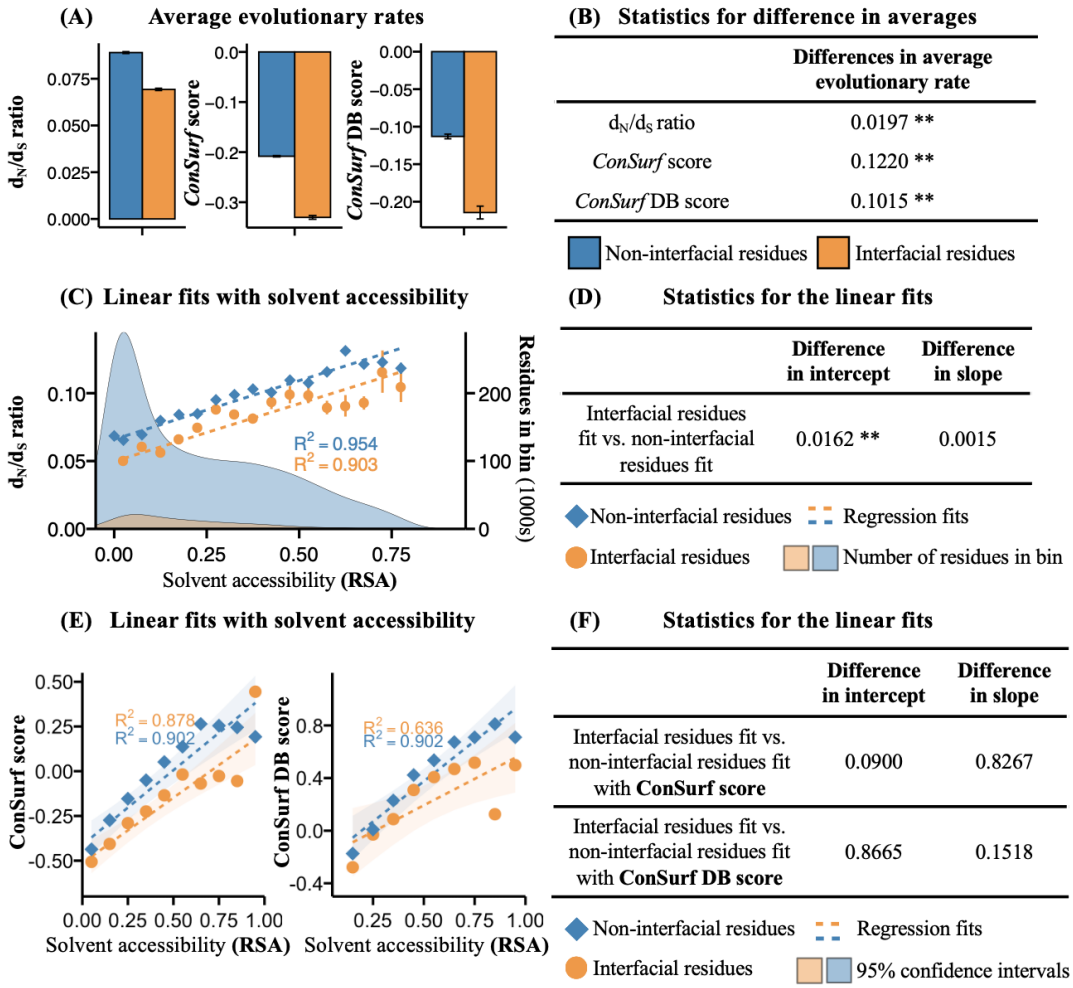



**Figure S3. The difference in evolutionary rate between interfacial and non-interfacial residues with additional species included in evolutionary rate calculations – Analysis S1, S2 and S3. (A)** Average evolutionary rates (estimated using *dN/dS* ratio and *ConSurf* score computed from a larger set of aligned related species and *ConSurf* score from *ConSurf* DB), plotted for interfacial and non-interfacial residues in our data. Standard errors for the average values of each group of residues are also shown. **(B)** Results of t-tests for differences in average evolutionary rates between interfacial and non-interfacial residues using the three evolutionary rate estimates. Differences significant at the P-value < 0.01 level are denoted with a double asterisk (**). **(C)** Linear fits between binned measures of solvent accessibility (complex RSA) and evolutionary rate (*dN/dS*, computed from a larger set of aligned related species) for interfacial and non-interfacial residues. Distributions of the number of residues per bin, weighted linear regression lines, and $R^2$ values of the fits are also shown. **(D)** Results of t-tests for differences in slope and intercept between the two fits in **(C)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**). **(E)** Linear fits between binned measures of solvent accessibility (complex RSA) and evolutionary rate (*ConSurf* score, computed from a larger set of aligned related species and *ConSurf* score from *ConSurf* DB) for interfacial and non-interfacial residues. Weighted linear regression lines, 95% confidence interval, and $R^2$ values of the fits are also shown. **(F)** Results of t-tests for differences in slope and intercept between the fits in **(E)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**).

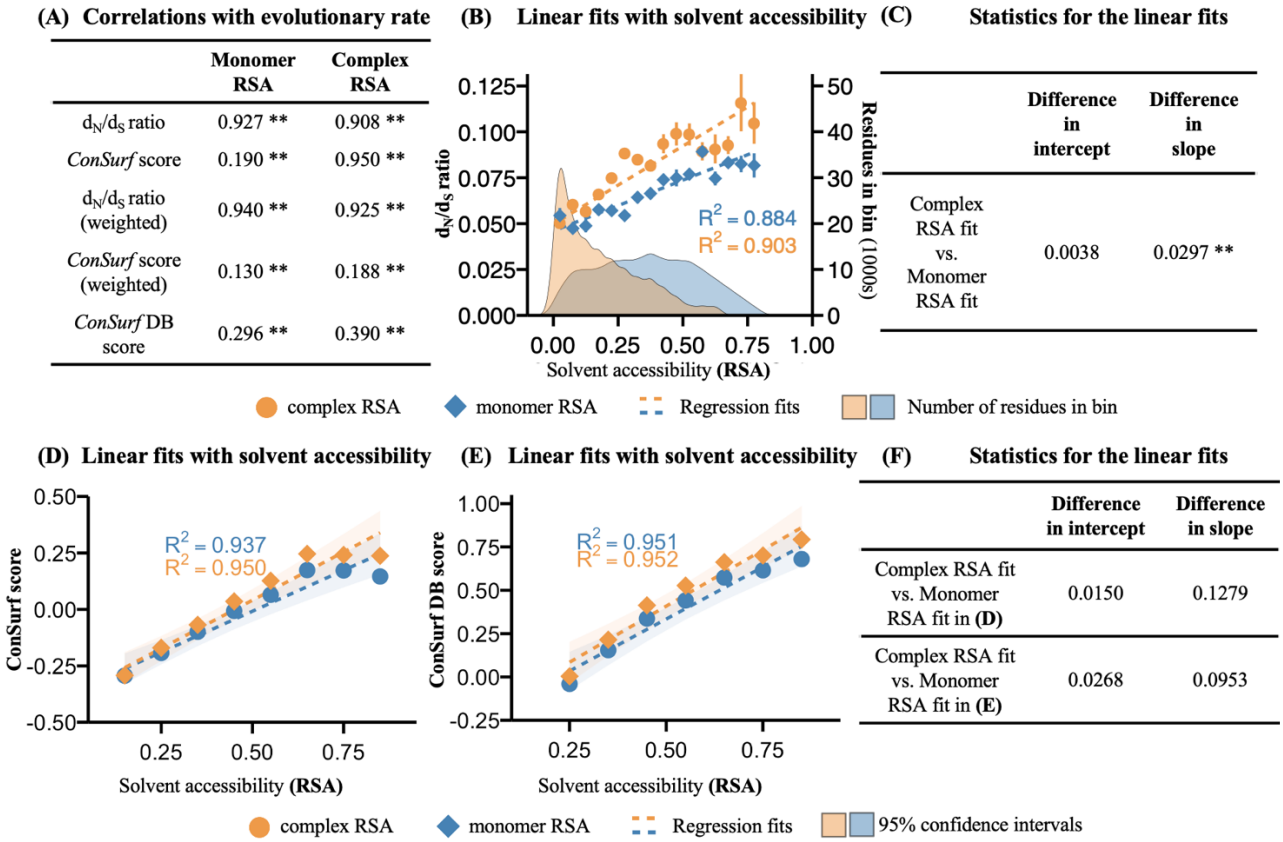
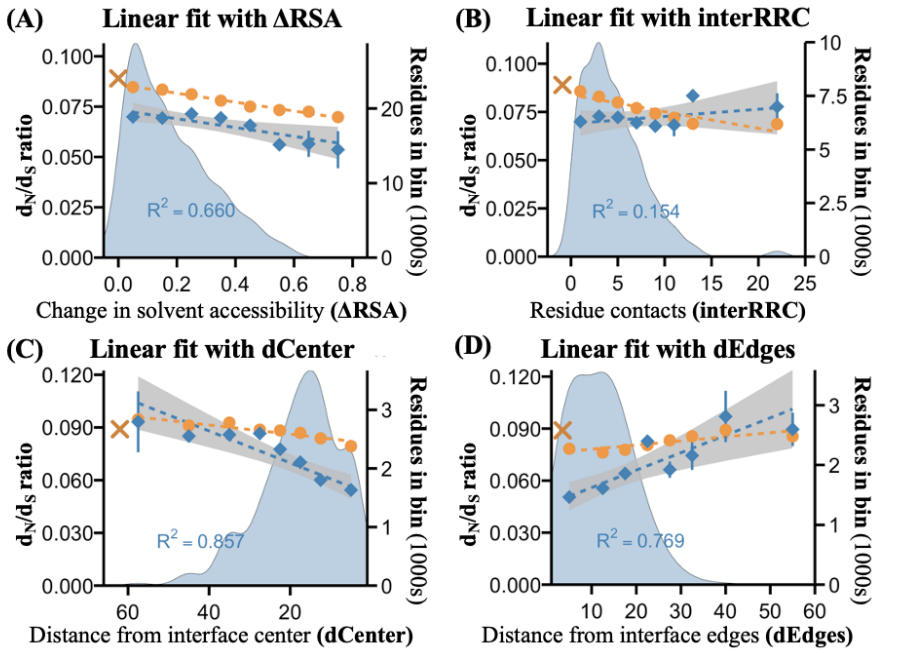



**Figure S4. The relationship between solvent accessibility and evolutionary rate in PPI interfaces with additional species included in evolutionary rate calculations – Analysis S1, S2 and S3. (A)** Results of a Pearson product-moment correlation test between values of solvent accessibility and measure of evolutionary rates (computed from a larger set of aligned related species) for interfacial residues in our models. Rows 1,2 and 5 in the table show standard Pearson correlations, whereas rows 3 and 4, show weighted Pearson correlations, taking the standard error on evolutionary rate estimates into consideration. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**). **(B)** Linear fits between binned measures of solvent accessibility and evolutionary rate (*dN/dS*, computed from a larger set of aligned related species), for monomer solvent accessibility (monomer RSA) and complex solvent accessibility (complex RSA). Distributions of the number of residues per bin, weighted linear regression lines, and $R^2$ values of the fits are also shown. **(C)** Results of t-tests for differences in slope and intercept between the two fits in **(B)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**). **(D), (E)** Linear fits between binned measures of solvent accessibility and evolutionary rate (*ConSurf* score, computed from a larger set of aligned related species and *ConSurf* score from *ConSurf* DB), for monomer solvent accessibility (monomer RSA) and complex solvent accessibility (complex RSA). Weighted linear regression lines, 95% confidence interval, and $R^2$ values of the fits are also shown. **(F)** Results of t-tests for differences in slope and intercept between the two fits in **(D)** and **(E)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**).




**(D) Statistics for linear fits**

| | Difference in intercept between observed and expected fit | Difference in slope between observed and expected fit |
|---|---|---|
| ΔRSA | 0.0127 ** | 0.0002 |
| interRRC | 0.0150 ** | 0.0012 * |
| dCenter | 0.0292 ** | 0.0006 ** |
| dEdges | 0.0298 ** | 0.0008 * |

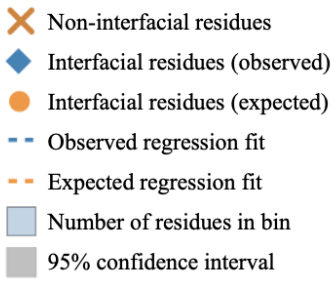
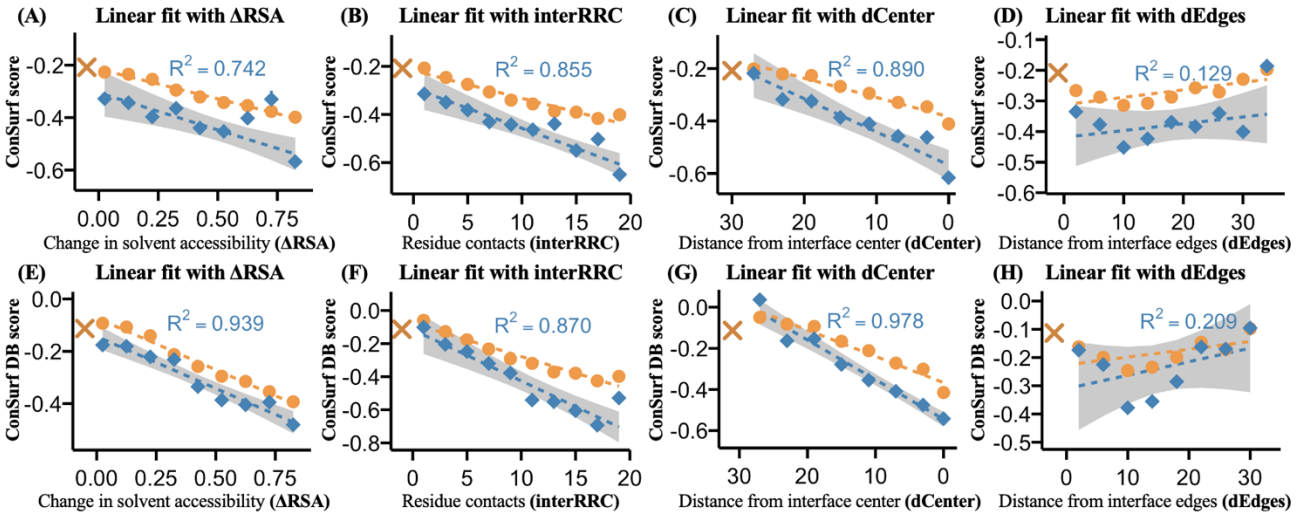



**Figure S5. The relationship between interface involvement and evolutionary rate for residues in PPI interfaces with additional species included in evolutionary rate calculations – Analysis S1. (A)-(D)** Linear fits in blue between binned measures of interface involvement and evolutionary rate (*dN/dS*, computed from a larger set of aligned related species), for change in burial upon complex formation (ΔRSA), inter-protein residue-residue contacts (interRRC), distance from interface center (dCenter), and distance from interface edges (dEdges) respectively. Distributions of the number of residues per bin, weighted linear regression lines, and $R^2$ values of the fits are also shown. In addition to the observed fit in blue, we also show the expected fit in red, assuming that interfacial burial and non-interfacial burial are selectively equivalent and that interfacial residues are subject to the same evolutionary constraints as non-interfacial residues with the same total burial in PPIs. For each panel, the average *dN/dS* value for non-interfacial residues (average *dN/dS* = 0.089) is marked by a red "X". **(E)** Results of t-tests for differences in slope and intercept between observed and expected fits in **(A)-(D)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**), values significant at the P-value < 0.05 level are denoted with a single asterisk (*).




**(I) Statistics for linear fits**

| | Difference in **intercept** between observed and expected fit with **ConSurf score** | Difference in **slope** between observed and expected fit with **ConSurf score** | Difference in **intercept** between observed and expected fit with **ConSurf DB score** | Difference in **slope** between observed and expected fit with **ConSurf DB score** |
|---|---|---|---|---|
| ΔRSA | 0.0901 * | 0.0564 | 0.0705 ** | 0.0015 |
| interRRC | 0.0726 | 0.0052 | 0.0379 | 0.0110 * |
| dCenter | 0.1841 ** | 0.0053 * | 0.1781 ** | 0.0065 ** |
| dEdges | 0.1062 | 0.0002 | 0.0846 | 0.0020 |

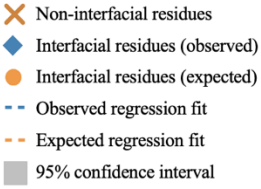



**Figure S6. The relationship between interface involvement and evolutionary rate for residues in PPI interfaces with additional species included in evolutionary rate calculations – Analysis S2, S3. (A)-(D)** Linear fits in blue between binned measures of interface involvement and evolutionary rate (*ConSurf* score, computed from a larger set of aligned related species), for change in burial upon complex formation (ΔRSA), inter-protein residue-residue contacts (interRRC), distance from interface center (dCenter), and distance from interface edges (dEdges) respectively. Weighted linear regression lines, 95% confidence interval, and $R^2$ values of the fits are also shown. In addition to the observed fit in blue, we also show the expected fit in yellow, assuming that interfacial burial and non-interfacial burial are selectively equivalent and that interfacial residues are subject to the same evolutionary constraints as non-interfacial residues with the same total burial in PPIs. For each panel, the average *ConSurf* score value for non-interfacial residues (average *ConSurf* score = -0.2082) is marked by a yellow "X". **(E)-(H)** Identical plots to **(A)-(D)**, using *ConSurf* scores from *ConSurf* DB, and with average *ConSurf* score value for non-interfacial residues = -0.1130. **(I)** Results of t-tests for differences in slope and intercept between observed and expected fits in **(A)-(H)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**), values significant at the P-value < 0.05 level are denoted with a single asterisk (*).

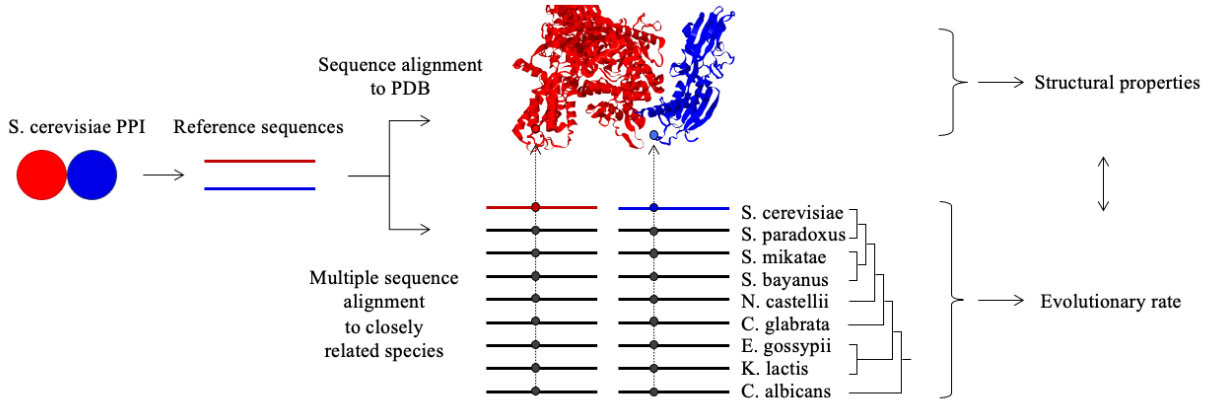
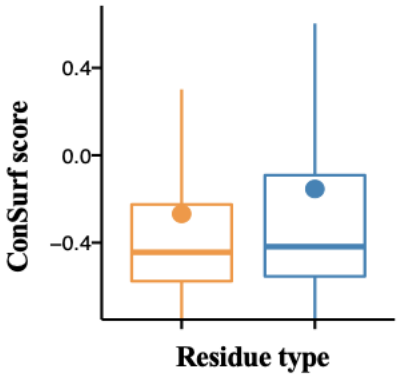



**Figure S7.** Graphical representation of the homology-based structural annotation transfer and evolutionary sequence analysis portion of our data curation pipeline.

**(A) Distribution of evolutionary rates**




**(B) Statistics for difference in averages**

| | Differences in average evolutionary rate |
|---|---|
| *ConSurf* score | 0.1140 ** |

Interfacial residues
Non-interfacial residues

**Figure S8. The distribution of evolutionary rate for interfacial and non-interfacial residues.** **(A)** Distribution of evolutionary rates (estimated using *ConSurf* score), plotted for interfacial and non-interfacial residues in our data. The mean of both distributions is also shown as a filled circle. **(B)** Results of t-tests for differences in average *ConSurf* score between interfacial and non-interfacial residues. Differences significant at the P-value < 0.01 level are denoted with a double asterisk (**).

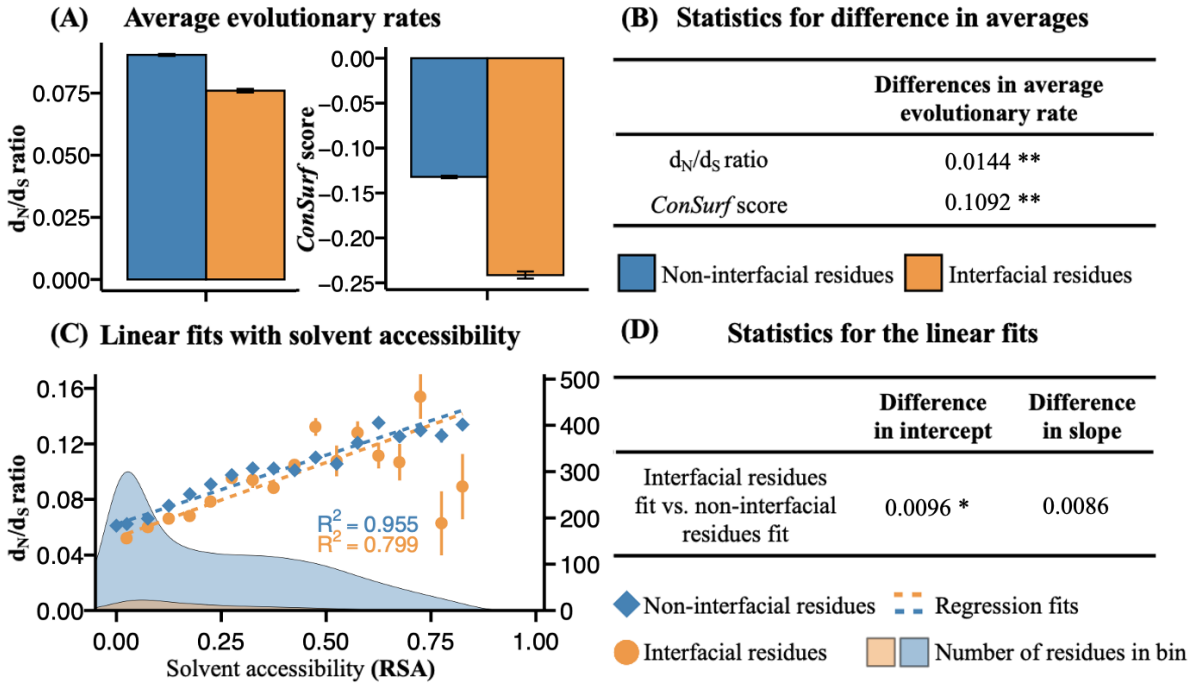
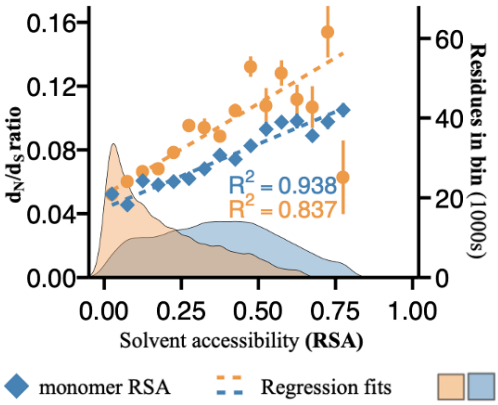



**Figure S9. The difference in evolutionary rate between interfacial and non-interfacial residues** (for high-sequence-identity PPIs– **Analysis S4**). **(A)** Average evolutionary rates (estimated using both *dN/dS* ratio and *ConSurf* score), plotted for interfacial and non-interfacial residues in our data. Standard errors for the average values of each group of residues are also shown. **(B)** Results of t-tests for differences in average evolutionary rates between interfacial and non-interfacial residues using both evolutionary rate estimates. Differences significant at the P-value < 0.01 level are denoted with a double asterisk (**). **(C)** Linear fits between binned measures of solvent accessibility (complex RSA) and evolutionary rate (*dN/dS*) for interfacial and non-interfacial residues. Distributions of the number of residues per bin, weighted linear regression lines, and $R^2$ values of the fits are also shown. **(D)** Results of t-tests for differences in slope and intercept between the two fits in **(C)**. Values significant at the P-value < 0.05 level are denoted with a single asterisk (*).

**(A) Correlations with evolutionary rate**

| | Monomer RSA | Complex RSA |
|---|---|---|
| $d_N/d_S$ ratio | 0.476 ** | 0.958 ** |
| *ConSurf* score | 0.219 ** | 0.163 ** |
| $d_N/d_S$ ratio (weighted) | 0.748 ** | 0.966 ** |
| *ConSurf* score (weighted) | 0.125 ** | 0.089 ** |

**(B) Linear fits with solvent accessibility**



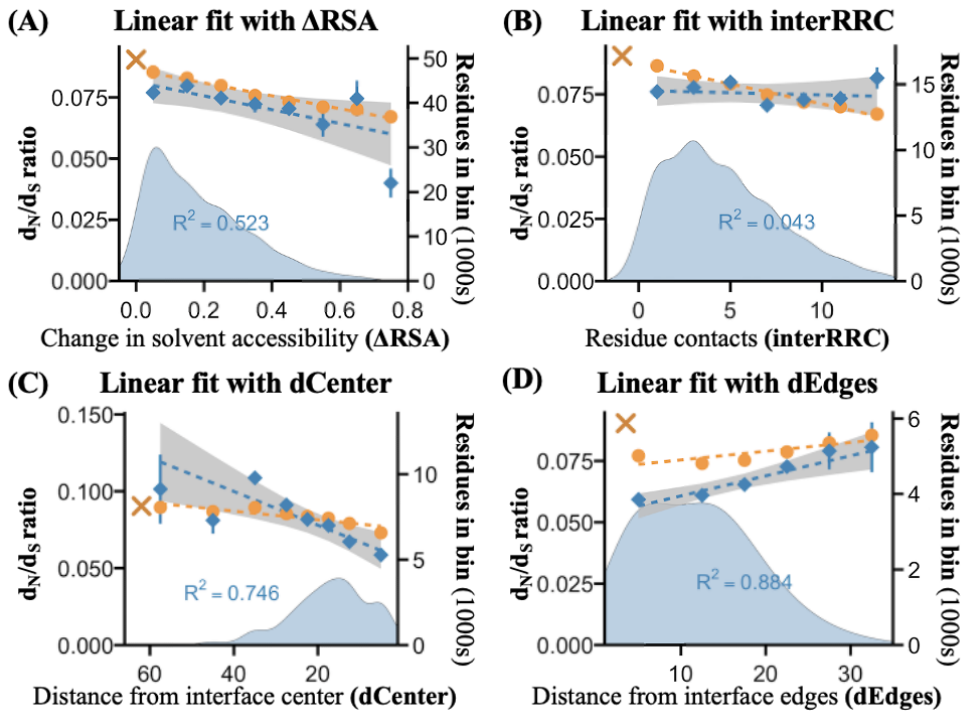




**(C) Statistics for the linear fits**

| | Difference in intercept | Difference in slope |
|---|---|---|
| Complex RSA fit vs. Monomer RSA fit | 0.008 | 0.034 * |

**Figure S10. The relationship between solvent accessibility and evolutionary rate in PPI interfaces** (for high-sequence-identity PPIs – **Analysis S4**). **(A)** Results of a Pearson product-moment correlation test between values of solvent accessibility and measure of evolutionary rates for interfacial residues in our models. The first two rows in the table show standard Pearson correlations, whereas the two last rows, show weighted Pearson correlations, taking the standard error on evolutionary rate estimates into consideration. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**). **(B)** Linear fits between binned measures of solvent accessibility and evolutionary rate (*dN/dS*), for monomer solvent accessibility (monomer RSA) and complex solvent accessibility (complex RSA). Distributions of the number of residues per bin, weighted linear regression lines, and $R^2$ values of the fits are also shown. **(C)** Results of t-tests for differences in slope and intercept between the two fits in **(B)**. Values significant at the P-value < 0.05 level are denoted with a single asterisk (*).




**(D) Statistics for linear fits**

| | Difference in intercept between observed and expected fit | Difference in slope between observed and expected fit |
|---|---|---|
| ΔRSA | 0.0048 | 0.0021 |
| interRRC | 0.0107 ** | 0.0014 * |
| dCenter | 0.0199 * | 0.0008 * |
| dEdges | 0.0193 ** | 0.0005 * |

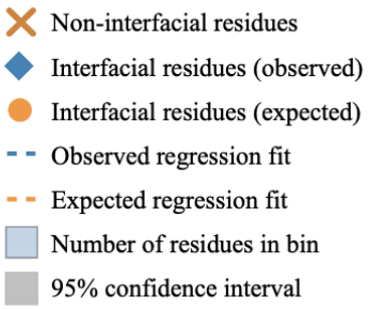



**Figure S11. The relationship between interface involvement and evolutionary rate for residues in PPI interfaces** (for high-sequence-identity PPIs – **Analysis S4**). **(A)-(D)** Linear fits in blue between binned measures of interface involvement and evolutionary rate (*dN/dS*), for change in burial upon complex formation (ΔRSA), inter-protein residue-residue contacts (interRRC), distance from interface center (dCenter), and distance from interface edges (dEdges) respectively. Distributions of the number of residues per bin, weighted linear regression lines, and $R^2$ values of the fits are also shown. In addition to the observed fit in blue, we also show the expected fit in yellow, assuming that interfacial burial and non-interfacial burial are selectively equivalent and that interfacial residues are subject to the same evolutionary constraints as non-interfacial residues with the same total burial in PPIs. For each panel, the average dN/dS value for non-interfacial residues (average dN/dS = 0.0904) is marked by a yellow "X". **(E)** Results of t-tests for differences in slope and intercept between observed and expected fits in **(A)-(D)**. Values significant at the P-value < 0.01 level are denoted with a double asterisk (**), values significant at the P-value < 0.05 level are denoted with a single asterisk (*).

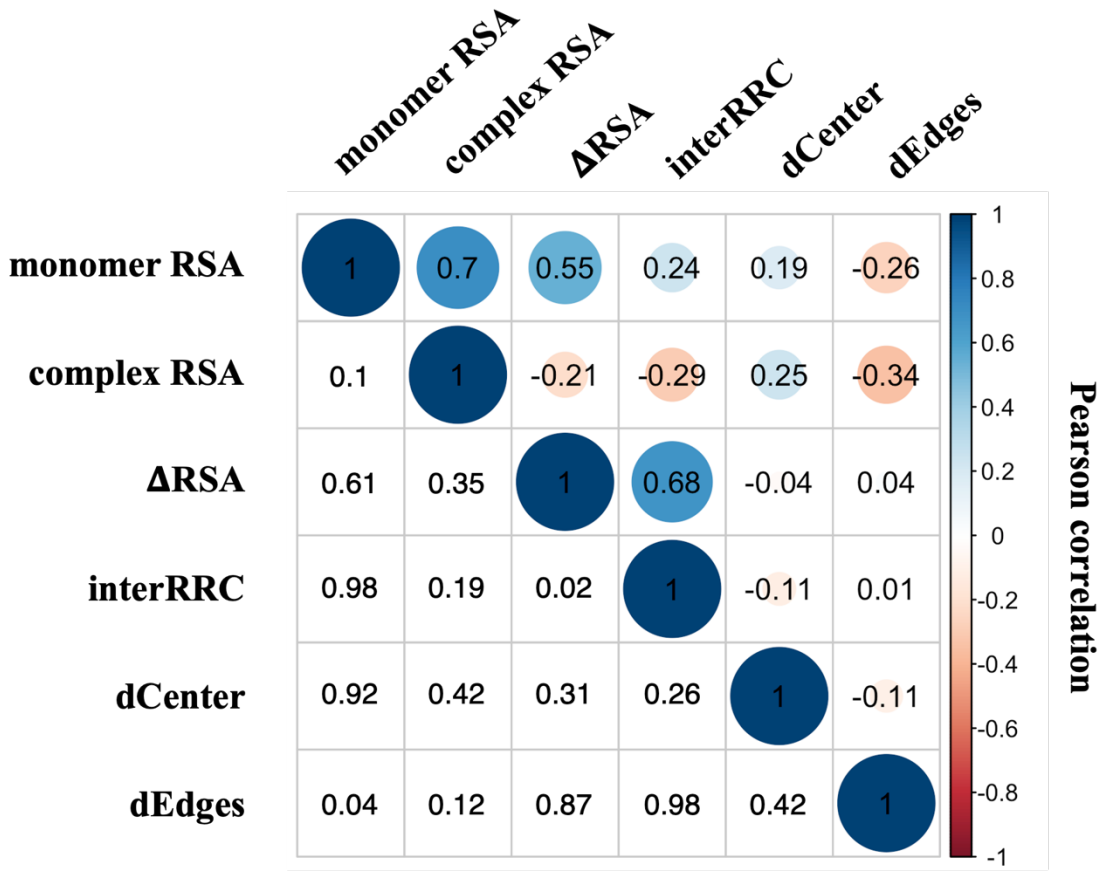



**Figure S12. Correlation between structural properties of a residue's microenvironment**. Pairwise Pearson correlation matrix for the structural properties of interest in this study computed for all interfacial residues. Pearson correlations are listed in the upper triangle as well as illustrated using circles or various shades. The p-values associated with each correlation are listed in the lower triangle.

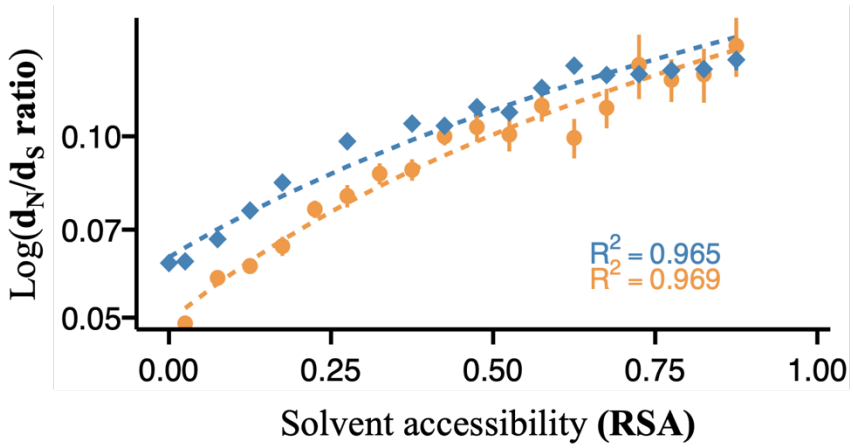



**Figure S13. The difference in evolutionary rate between interfacial and non-interfacial residues.** Linear fits between binned measures of solvent accessibility (complex RSA) and evolutionary rate (*dN/dS*) for interfacial and non-interfacial residues plotted on a log-10 scale. Weighted linear regression lines, and $R^2$ values of the fits are also shown.

**Supplementary information for Chapter 4**

# Structure-guided evolutionary analysis of interactome network rewiring at single residue resolution in yeasts

Léah Pollet and Yu Xia *

Department of Bioengineering, Faculty of Engineering, McGill University, Montreal, QC, Canada

Correspondence to Yu Xia: brandon.xia@mcgill.ca (Y. Xia)

https://doi.org/10.1016/j.jmb.2024.168641

Edited by Michael Sternberg

**Analysis S1. The difference in evolutionary rate between "interfacial" residues and "pseudo-interfacial" residues.**

For each rewired *S. cerevisiae* and *S. pombe* PPIs in our data, PSI-BLAST protein sequence alignments were used to transfer interface annotations from the protein in the species with the PPI (query), to the orthologous single protein in the species without the PPI (subject). Gapped positions in the query were ignored. Non-interfacial residues in one species that align to the functional interface in the other species were labeled as "pseudo-interfacial residues". *ConSurf-rate4site* scores and *ConSurf-DB* scores were then computed for all interfacial and pseudo-interfacial residues as described in the main paper. Comparisons of average evolutionary rates (computed using both *ConSurf-rate4site* and *ConSurf-DB* scores) for all interfacial residues and all pseudo-interfacial residues within a species (*S. cerevisiae* and *S. pombe*) can be seen in **Figure S1**. Direct comparisons between an individual interfacial residue evolutionary rate (in one species) and its corresponding pseudo-interfacial residue (in the other species) using both *ConSurf-rate4site* and *ConSurf-DB* scores can be seen in **Figure S2**.

**Analysis S2. The difference in average evolutionary rate for non-interfacial surface and interior residues and interfacial rim support and core residues.**

For each residue in our structurally-resolved interactome networks for *S. pombe* and *S. cerevisiae*, solvent Accessible Surface Area (SASA) was calculated using the DSSP program with hydrogen atoms excluded and SASA values were normalized to produce Relative Solvent Accessibility (RSA). For each residue, two values of RSA were computed: monomer RSA, which was calculated using the structure of monomeric proteins (discarding the chain mapped to the partner protein in a structure), and complex RSA, which was obtained from the co-complexed structure of both protein partners (PPI structure). ΔRSA, the change in residue burial upon complex formation, was computed as the difference between monomeric and co-structured RSA values for each residue in the structural models ($\Delta RSA$ = monomer RSA − complex RSA). ΔRSA was subsequently used in the definition of interfaces: any residue with a change in burial upon complex formation ($\Delta RSA \neq 0$) was defined as an interfacial residue, and any residue with no change in burial upon complex formation ($\Delta RSA = 0$) was defined as a non-interfacial residue. Non-interfacial residues were further subdivided into non-interfacial surface residues (complex RSA > 25% and $\Delta RSA = 0$) and non-interfacial interior residues (complex RSA < 25% and $\Delta RSA = 0$). Interfacial residues were further subdivided into interfacial rim residues (complex RSA > 25% and $\Delta RSA \neq 0$), interfacial support residues (monomer RSA < 25% and $\Delta RSA \neq 0$) and interfacial core residues (monomer RSA > 25% and complex RSA < 25% and $\Delta RSA \neq 0$). Comparisons of average evolutionary rate (computed using both *ConSurf-rate4site* and *ConSurf-DB* scores) for non-interfacial surface and interior residues and interfacial rim support and core residues (overall results for all PPI types combined) can be seen in **Figure S3**. Comparisons of average evolutionary rate (computed using both *ConSurf-rate4site* and *ConSurf-DB* scores) for non-interfacial surface and interior residues and interfacial rim support and core residues in preserved, missing ortholog, and rewired PPIs (separate results for each PPI type) can be seen in **Figure S4**.

**Analysis S3. Repeat analysis without homology-based PPI structural models.**

Any PPI in our structurally-resolved interactome networks for *S. pombe* and *S. cerevisiae* modeled using the structure of a closely related PPI in another species was discarded, keeping only PPIs for which a high-resolution experimental 3D structure was solved in the species. All analyses from the main paper were then repeated. Comparisons of average evolutionary rate (computed using both *ConSurf-rate4site* and *ConSurf-DB* scores) for interfacial and non-interfacial residues in PPIs with experimentally determined protein complex structures can be seen in **Figure S5**. Comparisons of average evolutionary rate (computed using both *ConSurf-rate4site* and *ConSurf-DB* scores) for interfacial and non-interfacial residues in preserved, missing ortholog, and rewired PPIs with experimentally determined protein complex structures can be seen in **Figure S6**. Comparisons of average evolutionary rate (computed using both *ConSurf-rate4site* and *ConSurf-DB* scores) for buried and exposed residues outside of PPI interfaces in PPIs with experimentally determined protein complex structures can be seen in **Figure S7**. Comparisons of average evolutionary rate (computed using both *ConSurf-rate4site* and *ConSurf-DB* scores) for buried and exposed residues outside of PPI interfaces in preserved, missing ortholog, and rewired PPIs with experimentally determined protein complex structures can be seen in **Figure S8**.

| | ***S. cerevisiae* PPIs** | ***S. pombe* PPIs** | **Total PPIs** | **Interfacial residues** | **Non-interfacial residues** |
|---|---|---|---|---|---|
| **Preserved PPIs** | 51 | 51 | 102 | 6 875 | 63 814 |
| **Missing ortholog PPIs** | 437 | 10 | 447 | 42 626 | 335 789 |
| **Rewired PPIs** | 61 | 141 | 202 | 9 481 | 128 226 |
| **Total** | 549 | 202 | 751 | 58 982 | 527 829 |

**Table S1. PPI dataset summary.** Summary of the number of PPIs, interfacial residues, and non-interfacial residues curated for each PPI type in this analysis. Numbers of PPIs and residues in our data for preserved PPIs, missing ortholog PPIs and rewired PPIs, as well as the number of PPIs of each type obtained from both *S. cerevisiae* and *S. pombe* are included.

| | *S. cerevisiae* | | | *S. pombe* | | |
|---|---|---|---|---|---|---|
| | **Average interface size** | **Number of interfaces** | **# rim residues / # core residues** | **Average interface size** | **Number of interfaces** | **# rim residues / # core residues** |
| Interfaces in **preserved PPIs** | 42 residues (SE: 4.29) | 96 | 1230/1523 = 0.81 | 35 residues (SE: 3.91) | 80 | 1096/1013 = 1.08 |
| Interfaces in **missing ortholog PPIs** | 35 residues (SE: 1.00) | 1216 | 13230/15908 = 0.83 | 16 residues (SE: 3.58) | 20 | 179/165 = 1.08 |
| Interfaces in **rewired PPIs** | 32 residues (SE: 2.90) | 122 | 1266/1468 = 0.86 | 20 residues (SE: 1.19) | 282 | 2460/1956 = 1.26 |

**Table S2. Interface size comparison between PPI types.** Summary of the average size (and the associated standard error, SE), number of preserved, missing ortholog and rewired PPI interfaces in *S. cerevisiae* and *S. pombe.* For each interface type in both species the ratio of the overall number of rim residues to overall number of core residues is also shown.

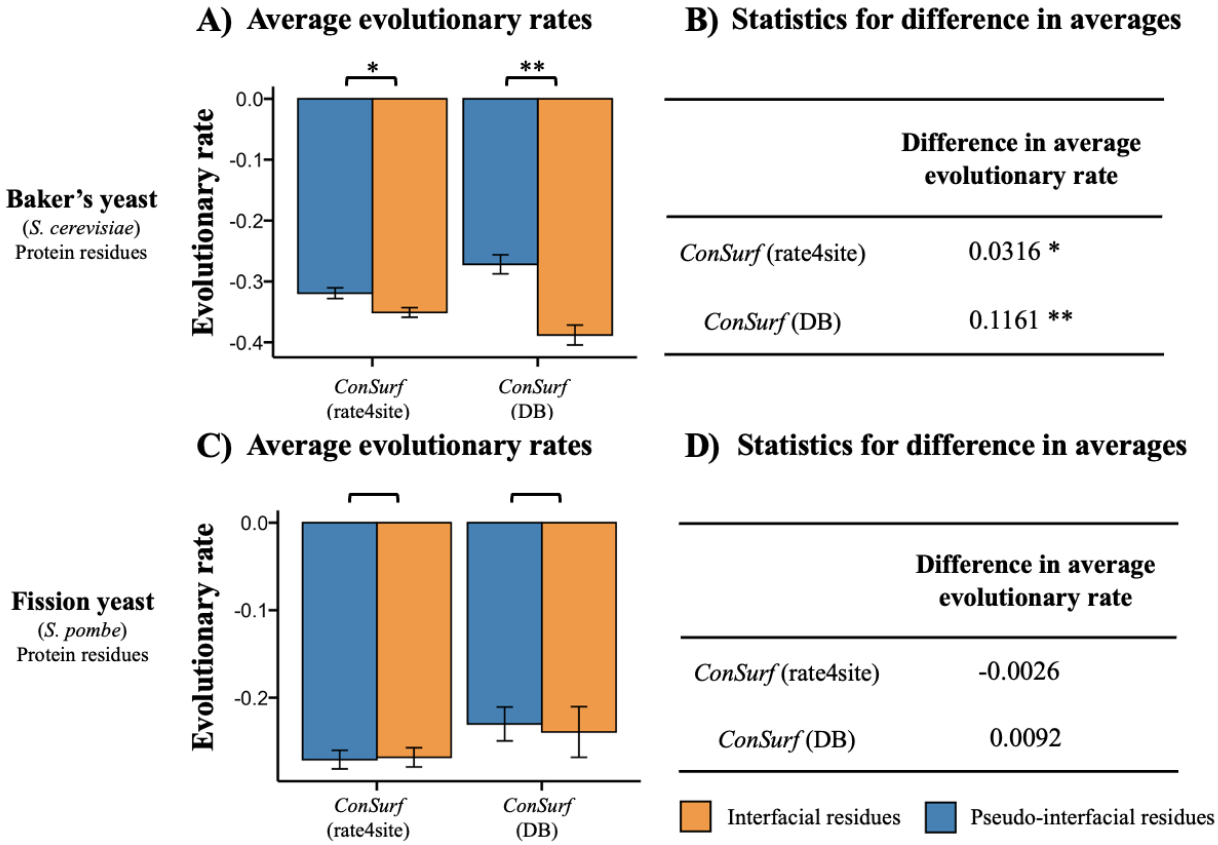



**Figure S1. The difference in evolutionary rate between interfacial and pseudo-interfacial residues within a species.** **(A)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for *S. cerevisiae* interfacial residues in rewired PPIs, and *S. cerevisiae* pseudo-interfacial residues in single proteins that do not interact but have a corresponding functional interface in *S. pombe*. Standard errors for the average values of each group of residues are also shown. **(B)** Results of t-tests for differences in average evolutionary rates between interfacial and pseudo-interfacial residues in *S. cerevisiae* using both evolutionary rate measures. **(C)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for *S. pombe* interfacial residues in rewired PPIs, and *S. pombe* pseudo-interfacial residues in single proteins that do not interact but have a corresponding functional interface in *S. cerevisiae*. Standard errors for the average values of each group of residues are also shown. **(D)** Results of t-tests for differences in average evolutionary rates between interfacial and pseudo-interfacial residues in *S. pombe* using both evolutionary rate measures. Comparisons significant at the P-value < 0.05 level are denoted with a single asterisk (*) and comparisons significant at the P-value < 0.01 level are denoted with a double asterisk (**).

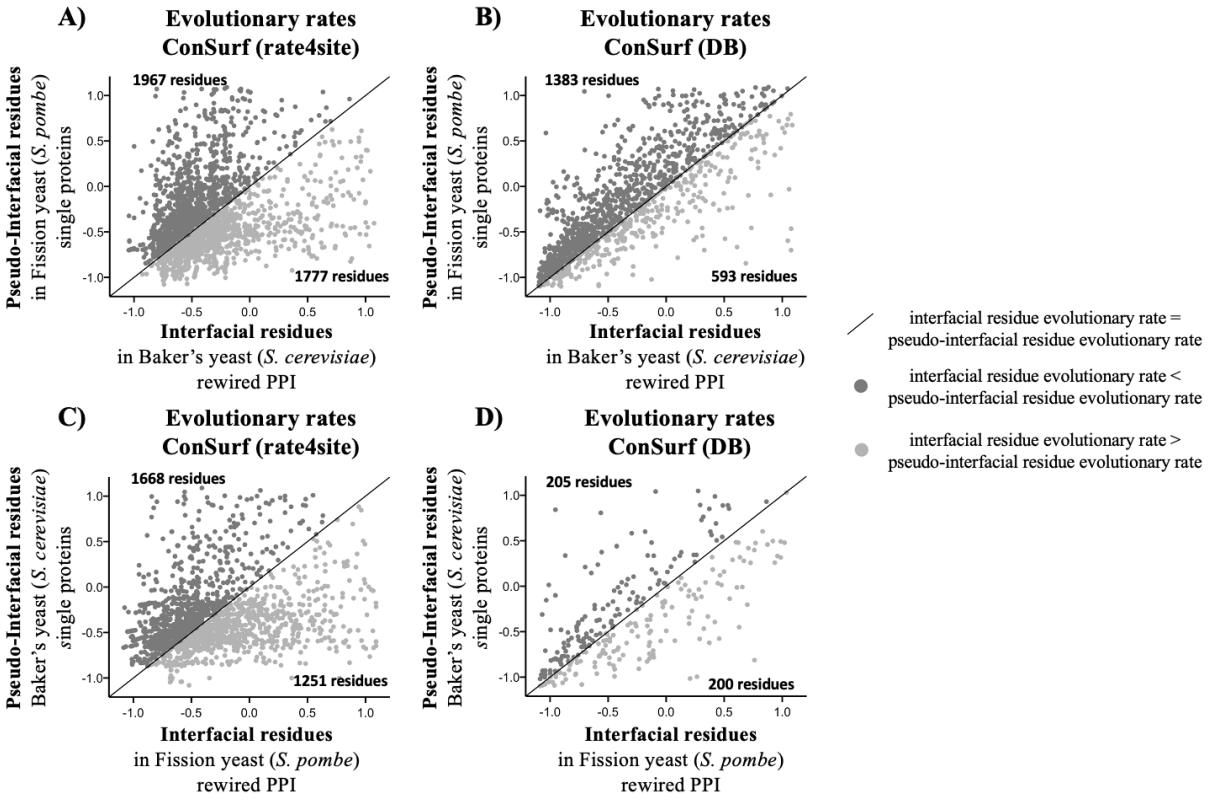



**Figure S2. The difference in evolutionary rate between interfacial residues and pseudo-interfacial residues across species. (A)** Evolutionary rates (as measured by *ConSurf-rate4site* score), for each *S. cerevisiae* interfacial residue in rewired PPIs and its corresponding *S. pombe* pseudo-interfacial residue in a single protein in our datasets. **(B)** Evolutionary rates (as measured by *ConSurf-DB* score), for each *S. cerevisiae* interfacial residue in rewired PPIs and its corresponding *S. pombe* pseudo-interfacial residue in a single protein in our datasets. **(C)** Evolutionary rates (as measured by *ConSurf-rate4site* score), for each *S. pombe* interfacial residue in rewired PPIs and its corresponding *S. cerevisiae* pseudo-interfacial residue in a single protein in our datasets. **(D)** Evolutionary rates (as measured by *ConSurf-DB* score), for each *S. pombe* interfacial residue in rewired PPIs and its corresponding *S. cerevisiae* pseudo-interfacial residue in a single protein in our datasets. The diagonal line (interfacial residue evolutionary rate = pseudo-interfacial residue evolutionary rate), as well as counts for the number of points above the diagonal line (interfacial residue more conserved than pseudo-interfacial residue) and below the diagonal line (pseudo-interfacial residue more conserved than interfacial residue) are also included for each panel.

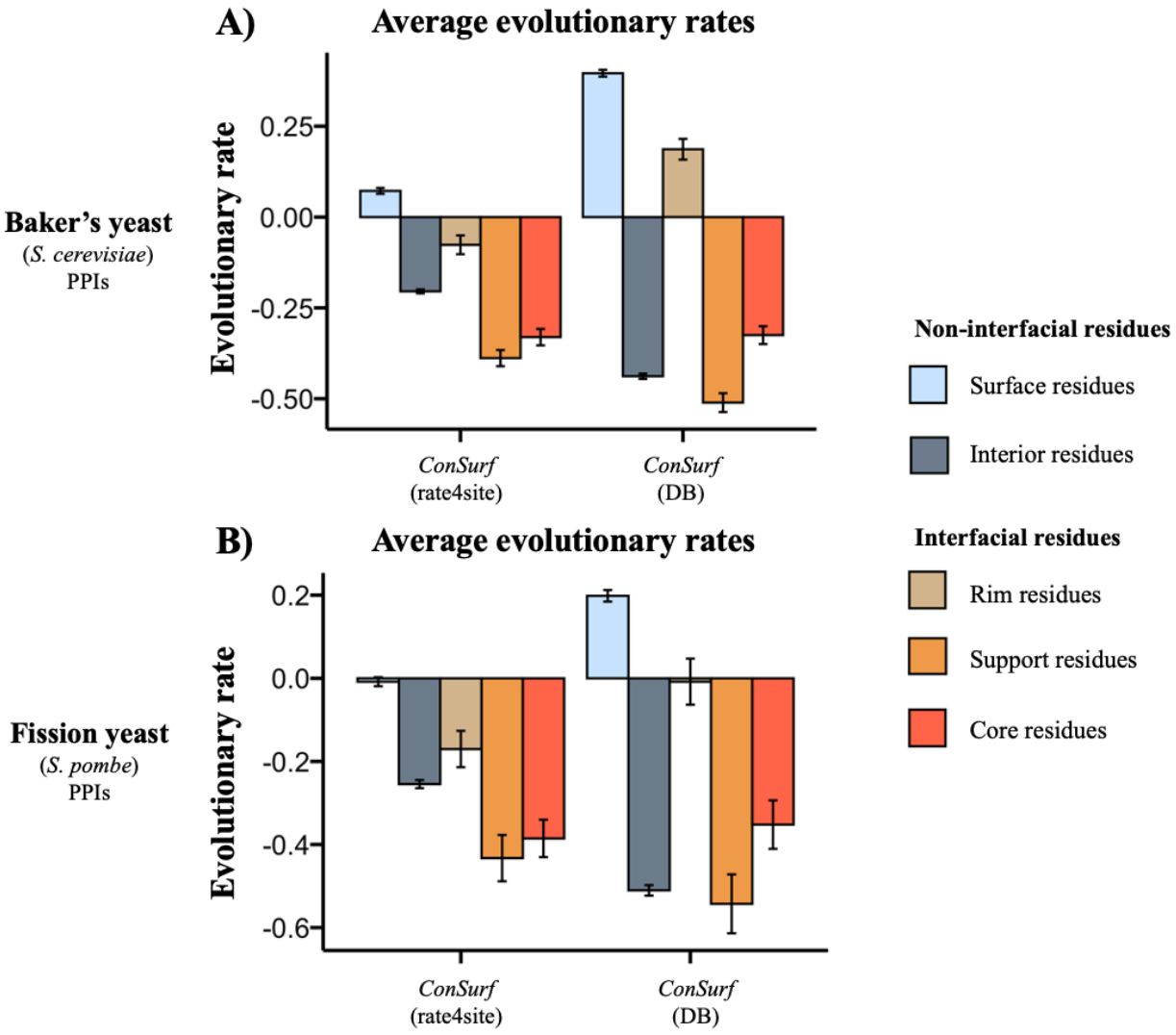



**Figure S3. The difference in average evolutionary rate for non-interfacial surface and interior residues and interfacial rim support and core residues. (A)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), for *S. cerevisiae* residues divided into non-interfacial surface residues (complex RSA > 25%), non-interfacial interior residues (complex RSA < 25%), interfacial rim residues (complex RSA > 25%), interfacial support residues (monomer RSA < 25%) and interfacial core residues (monomer RSA > 25% and complex RSA < 25%). Standard errors for the average values of each group of residues are also shown. **(B)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), for *S. pombe* residues divided into non-interfacial surface residues (complex RSA > 25%), non-interfacial interior residues (complex RSA < 25%), interfacial rim residues (complex RSA > 25%), interfacial support residues (monomer RSA < 25%) and interfacial core residues (monomer RSA > 25% and complex RSA < 25%). Standard errors for the average values of each group of residues are also shown.

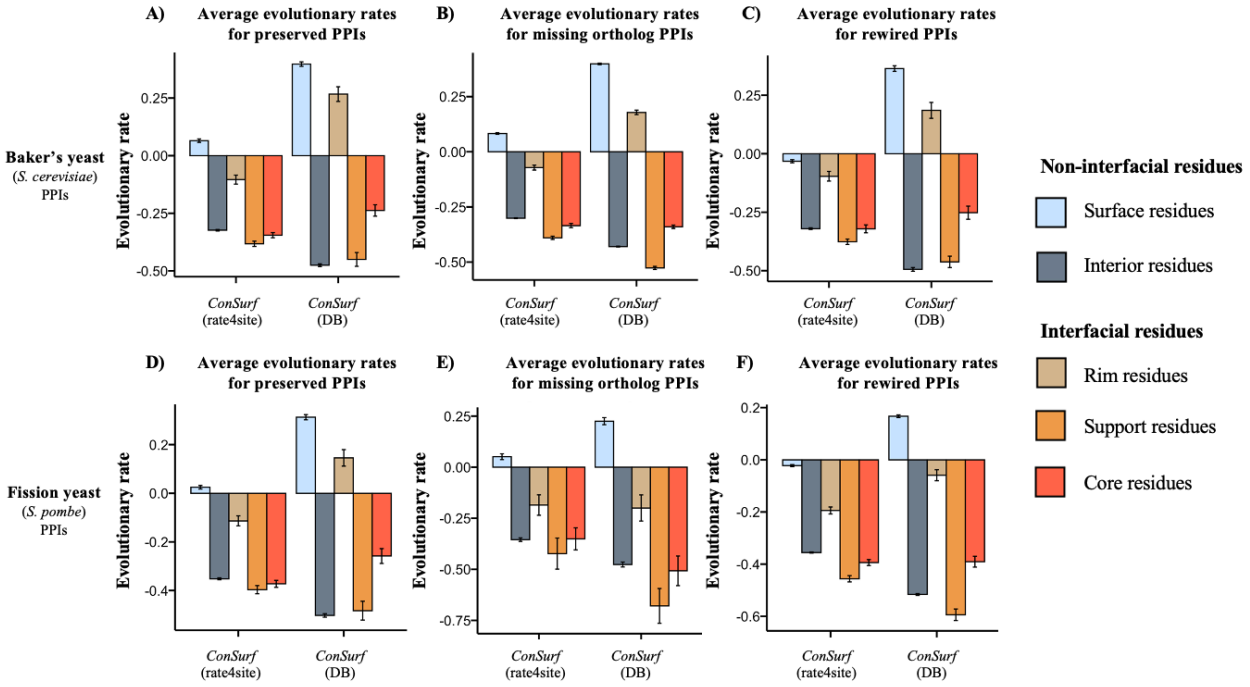



**Figure S4. The difference in average evolutionary rate for non-interfacial surface and interior residues and interfacial rim support and core residues in preserved, missing ortholog, and rewired PPIs. (A)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), for *S. cerevisiae* residues divided into non-interfacial surface residues (complex RSA > 25%), non-interfacial interior residues (complex RSA < 25%), interfacial rim residues (complex RSA > 25%), interfacial support residues (monomer RSA < 25%) and interfacial core residues (monomer RSA > 25% and complex RSA < 25%). Standard errors for the average values of each group of residues are also shown. **(B)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), for *S. pombe* residues divided into non-interfacial surface residues (complex RSA > 25%), non-interfacial interior residues (complex RSA < 25%), interfacial rim residues (complex RSA > 25%), interfacial support residues (monomer RSA < 25%) and interfacial core residues (monomer RSA > 25% and complex RSA < 25%). Standard errors for the average values of each group of residues are also shown.

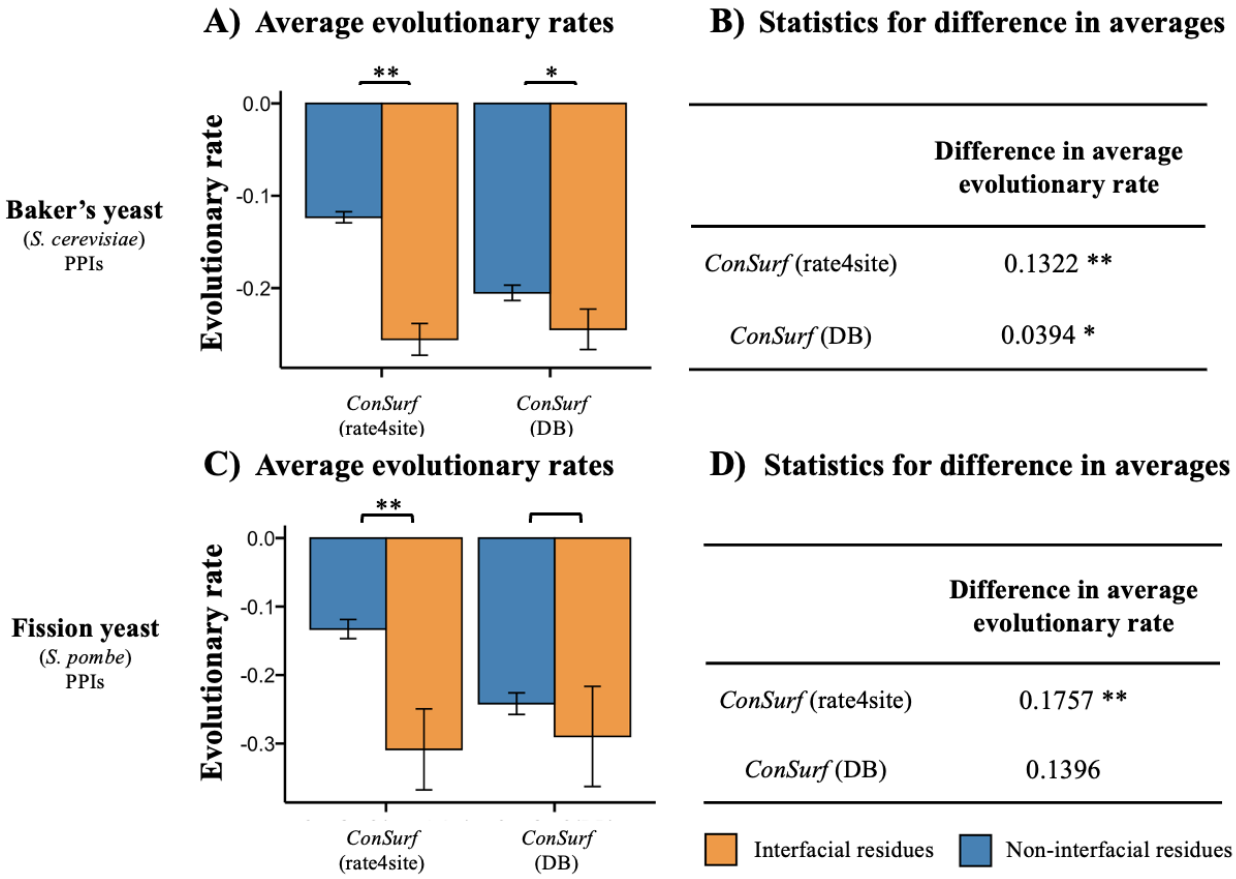



**Figure S5. The difference in evolutionary rate between interfacial and non-interfacial residues. Repeat analysis using only experimentally determined protein complex structures, with no homology-based PPI structural models. (A)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for interfacial and non-interfacial residues from all *S. cerevisiae* PPIs with experimentally determined protein complex structure in our datasets. Standard errors for the average values of each group of residues are also shown. **(B)** Results of t-tests for differences in average evolutionary rates between interfacial and non-interfacial residues in *S. cerevisiae* PPIs with experimentally determined protein complex structure using both evolutionary rate measures. **(C)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for interfacial and non-interfacial residues from all *S. pombe* PPIs with experimentally determined protein complex structure in our datasets. Standard errors for the average values of each group of residues are also shown. **(D)** Results of t-tests for differences in average evolutionary rates between interfacial and non-interfacial residues in *S. pombe* PPIs with experimentally determined protein complex structure using both evolutionary rate measures. Comparisons significant at the P-value < 0.05 level are denoted with a single asterisk (*) and comparisons significant at the P-value < 0.01 level are denoted with a double asterisk (**).

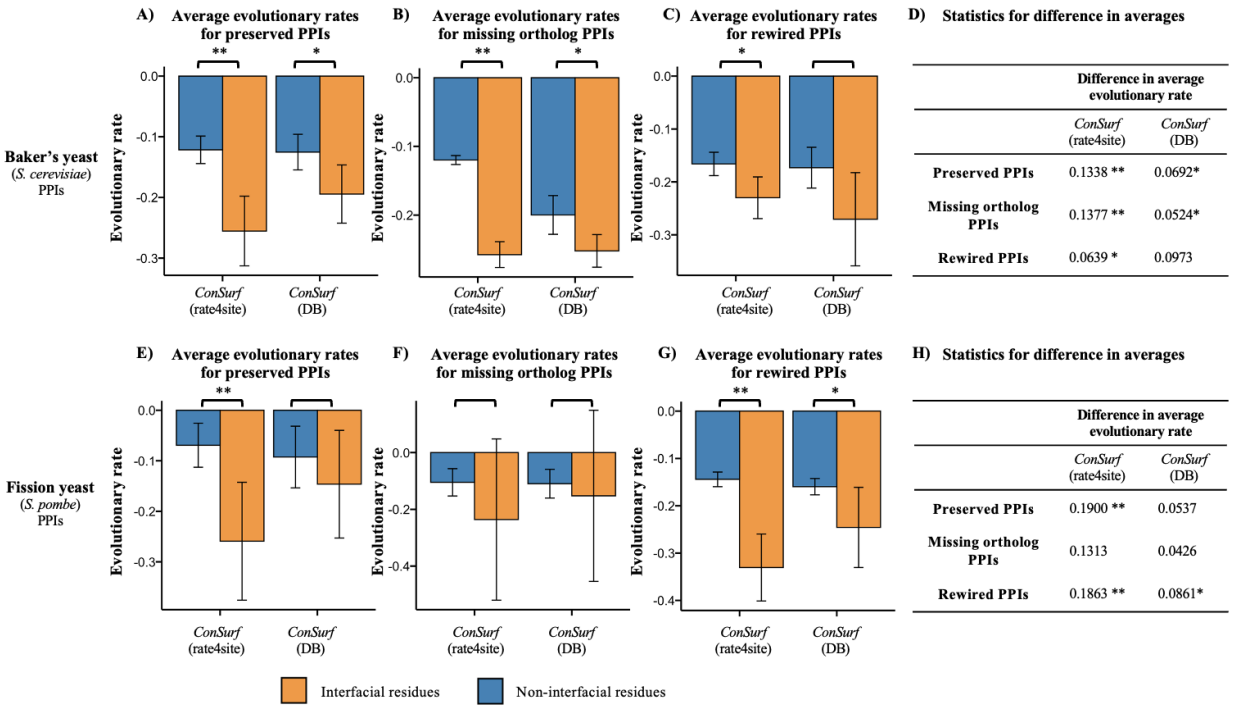



**Figure S6. The difference in evolutionary rate between interfacial and non-interfacial residues for preserved, missing ortholog, and rewired PPIs. Repeat analysis using only experimentally determined protein complex structures, with no homology-based PPI structural models. (A-C)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for interfacial and non-interfacial residues from all preserved PPIs, missing ortholog PPIs and rewired PPIs with experimentally determined protein complex structure in *S. cerevisiae* respectively. Standard errors for the average values of each group of residues are also shown. **(D)** Results of t-tests for differences in average evolutionary rates between interfacial and non-interfacial residues in preserved PPIs, missing ortholog PPIs and rewired PPIs with experimentally determined protein complex structure in *S. cerevisiae* using both measures of evolutionary rate. **(E-G)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for interfacial and non-interfacial residues from all preserved PPIs, missing ortholog PPIs and rewired PPIs with experimentally determined protein complex structure in *S. pombe* respectively. Standard errors for the average values of each group of residues are also shown. **(H)** Results of t-tests for differences in average evolutionary rates between interfacial and non-interfacial residues in preserved PPIs, missing ortholog PPIs and rewired PPIs with experimentally determined protein complex structure in *S. pombe* using both measures of evolutionary rate. Comparisons significant at the P-value < 0.05 level are denoted with a single asterisk (*) and comparisons significant at the P-value < 0.01 level are denoted with a double asterisk (**).

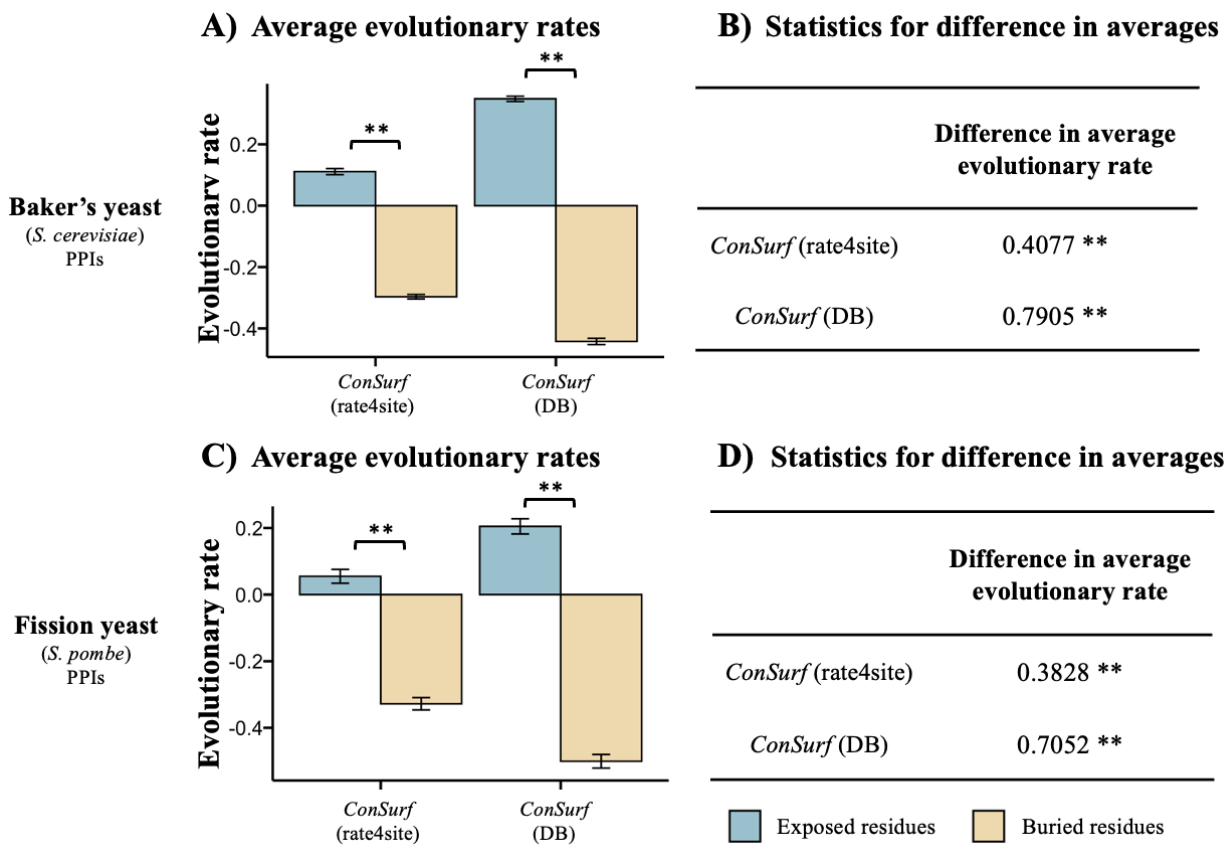



**Figure S7. The difference in evolutionary rate between buried and exposed residues outside of PPI interfaces. Repeat analysis using only experimentally determined protein complex structures, with no homology-based PPI structural models. (A)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for buried and exposed residues outside of PPI interfaces from all *S. cerevisiae* PPIs with experimentally determined protein complex structure in our data. Standard errors for the average values of each group of residues are also shown. **(B)** Results of t-tests for differences in average evolutionary rates between buried and exposed residues outside of PPI interfaces in *S. cerevisiae* PPIs with experimentally determined protein complex structure using both evolutionary rate measures. **(C)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for buried and exposed residues outside of PPI interfaces from all *S. pombe* PPIs with experimentally determined protein complex structure in our data. Standard errors for the average values of each group of residues are also shown. **(D)** Results of t-tests for differences in average evolutionary rates between buried and exposed residues outside of PPI interfaces in *S. pombe* PPIs with experimentally determined protein complex structure using both evolutionary rate measures. Comparisons significant at the P-value < 0.01 level are denoted with a double asterisk (**).

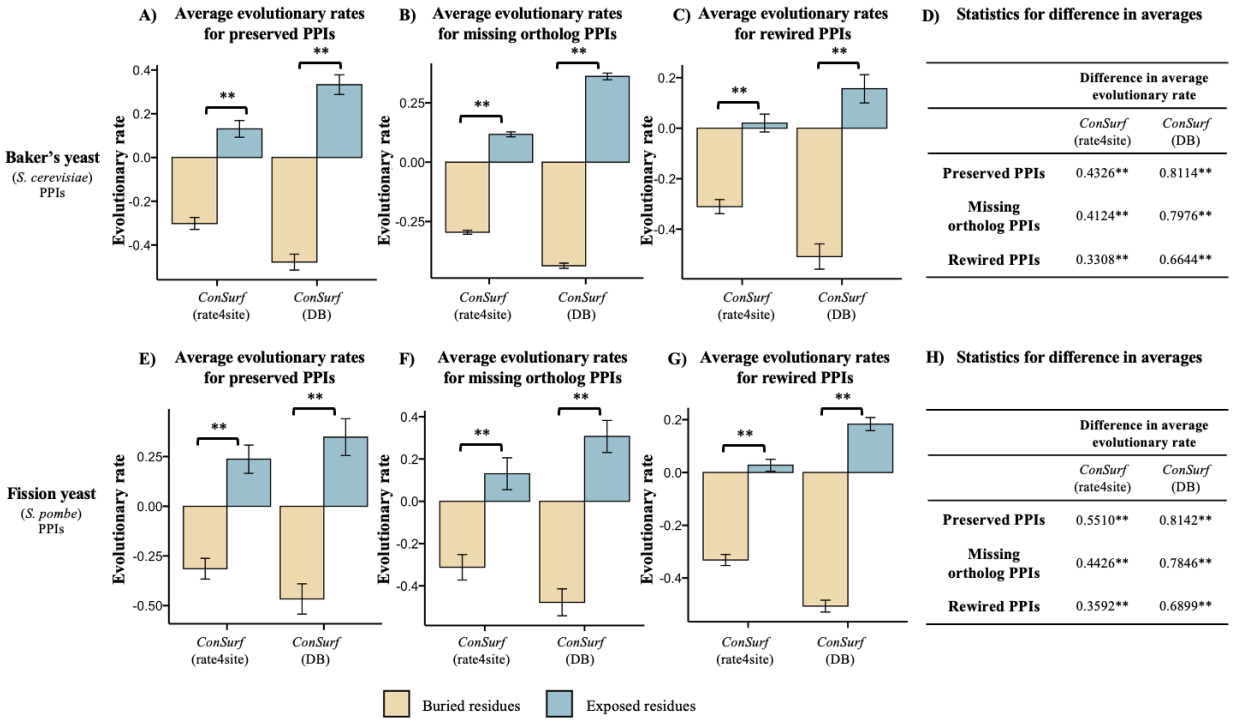



**Figure S8. The difference in evolutionary rate between buried and exposed residues outside of PPI interfaces for preserved, missing ortholog and rewired PPIs. Repeat analysis using only experimentally determined protein complex structures, with no homology-based PPI structural models. (A-C)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for buried and exposed residues outside of PPI interfaces from all preserved PPIs, missing ortholog PPIs, and rewired PPIs with experimentally determined protein complex structure in *S. cerevisiae* respectively. Standard errors for the average values of each group of residues are also shown. **(D)** Results of t-tests for differences in average evolutionary rates between buried and exposed residues outside of PPI interfaces in preserved PPIs, missing ortholog PPIs, and rewired PPIs with experimentally determined protein complex structure in *S. cerevisiae* using both measures of evolutionary rate. **(E-G)** Average evolutionary rates (as measured by *ConSurf-rate4site* score and *ConSurf-DB* score), plotted for buried and exposed residues outside of PPI interfaces from all preserved PPIs, missing ortholog PPIs, and rewired PPIs with experimentally determined protein complex structure in *S. pombe* respectively. Standard errors for the average values of each group of residues are also shown. **(H)** Results of t-tests for differences in average evolutionary rates between buried and exposed residues outside of PPI interfaces in preserved PPIs, missing ortholog PPIs, and rewired PPIs with experimentally determined protein complex structure in *S. pombe* using both measures of evolutionary rate. Comparisons significant at the P-value < 0.05 level are denoted with a single asterisk (*), and comparisons significant at the P-value < 0.01 level are denoted with a double asterisk (**).

# Appendix 2

**Supplementary information for all statistics analysis (Chapter 3 and Chapter 4)**

## Correlations

All correlations in this work were calculated as Pearson correlation coefficients. Standard Pearson correlations were obtained in R using the Feature Selection (Including Multiple Solutions) and Bayesian Networks ('MXM') package. Correlations labeled as "(weighted)" are weighted Pearson correlation, using the standard error to weigh the correlation analysis and were obtained in R with the Weighting and Weighted Statistics ('weights') package. For each Pearson correlation test throughout this work the variables of interest are using a continuous scale. Linear relationships without spurious outliers between variables of interest were established using simple scatter plots. Moreover, the structural and evolutionary data used in this analysis does not always follow a normal distribution, therefore, significance for each correlation coefficient was determined from 1,000 rounds of randomizing permutations.

## Linear fits

Linear relationships in this work were investigated using a weighted least-square regression technique that takes the error into account in the line fitting process. This technique has been used in previous literature and was adapted in R. The regression model takes the following form:

$$y(x) = w_0 + w_1 x_1 + e_x$$

where $y(x)$ is the evolutionary rate measure for residues in bin $x$, $x_1$ is the center value of bin $x$ for the structural property investigated, $w_0$, $w_1$ are the intercept, and the weight associated with the structural feature in the regression model, and $e_x$is a random variable ("noise term") following

a Gaussian distribution with zero mean and standard deviation equal to the standard error associated with the evolutionary rate measure for bin $x$. This method also reports a standard error for the slope and intercept of the resulting linear fit which we used in t-tests to compare slope and intercept across different fits. This method was chosen because errors for evolutionary rate measures are uncorrelated in our data, but the variance of the errors are not the same. Therefore, using weighted least-square regression ensures that high variability cases receive low weights, while low variability cases receive high weights.

**T-tests**

Welch two-samples t-tests were used in this work. For each t-test performed observations are independent and data from each group is approximately normally distributed with no significant outliers. However, as equality of variance and sample size cannot always be assumed Welch two-samples t-tests were used. This test is generally applied when the there is a difference between the variations of two populations or when their sample sizes are unequal. The Welch-Satterthwaite equation was, therefore used to approximate the degrees of freedom throughout.

**Integrated modeling**

We investigated the combined influences of residue structural properties on evolutionary rate using a weighted multiple linear regression technique, again aiming to take the error associated with calculating evolutionary rates into account. The regression model was implemented in R with the Classification and Regression Training ('caret') package and takes the following form:

$$y(x) = w_0 + w_1 x_1 + \cdots + w_n x_n + e_x$$

where $y(x)$ is the evolutionary rate of residue $x$, $x_1, \ldots, x_n$ are the values of each structural property investigated for residue $x$, $w_0, \ldots, w_n$ are the intercept and weights associated with the structural features, and $e_x$ is a random variable ("error term") following a Gaussian distribution with zero and standard deviation equal to the standard error associated with the evolutionary rate for residue $x$. The model is trained using a 10-fold cross-validation process: the set of residues is randomly partitioned into 10 subsamples of equal size. A single subsample is retained as validation data for testing the model and the remaining 9 subsamples are used as training data. This cross-validation process is repeated 10 times so that each of the 10 subsamples is used exactly once as validation data. The advantage of this method is that all observations are used for both training and validation and each observation is used for validation exactly once. The overall performance of a model is taken as the average performance across all cross-validation trials and compared across models including different subsets of structural properties.

# Reprint permissions

## Structure-guided Evolutionary Analysis of Interactome Network Rewiring at Single Residue Resolution in Yeasts

**Author:** Léah Pollet,Yu Xia

**Publication:** Journal of Molecular Biology

**Publisher:** Elsevier

**Date:** 15 August 2024



### Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

BACK CLOSE WINDOW

## Structural Determinants of Yeast Protein-Protein Interaction Interface Evolution at the Residue Level

**Author:** Léah Pollet,Luke Lambourne,Yu Xia

**Publication:** Journal of Molecular Biology

**Publisher:** Elsevier

**Date:** 15 October 2022



### Journal Author Rights

Please note that, as the author of this Elsevier article, you retain the right to include it in a thesis or dissertation, provided it is not published commercially. Permission is not required, but please ensure that you reference the journal as the original source. For more information on this and on your other retained rights, please visit: https://www.elsevier.com/about/our-business/policies/copyright#Author-rights

BACK CLOSE WINDOW