THE UNIVERSITY *of* EDINBURGH

Hallmarks of cotranslational protein complex assembly and its relationship to the dominant-negative effect

Mihaly Badonyi

August 2023, Edinburgh

Doctor of Philosophy

The University of Edinburgh

Declaration

I hereby declare that this thesis was composed by myself and has not been submitted for any previous degree or professional qualification. Chapters 2, 3, and 4 of the thesis represent the findings and views of first-author papers published in open access journals:

[1] Badonyi, M. and Marsh, J.A., 2022. Large protein complex interfaces have evolved to promote cotranslational assembly. *eLife*, ***11***, p.e79602

[2] Badonyi, M. and Marsh, J.A., 2023. Buffering of genetic dominance by allele-specific protein complex assembly. *Science Advances*, ***9***(22), p.eadf9845

[3] Badonyi, M. and Marsh, J.A., 2023. Hallmarks and evolutionary drivers of cotranslational protein complex assembly. *The FEBS Journal*

Note on the use of published material

Chapter 1 includes material used in review [3]. Chapters 2-4 are minimally adjusted versions of the original publications; changes were made to the content and the layout to improve consistency throughout this thesis. Chapter 5 is a work-in-progress manuscript, but it includes material published in research paper [2].

Mihaly Badonyi
22nd of August, 2023

Acknowledgement

I would like to thank my supervisor Joseph A Marsh for his support, and for allowing me sufficient freedom to follow my own research path. Although the ideas in this thesis were within me, I could not have developed any of them without his mentorship. Joe helped me adopt a mindset that made me a more careful thinker and a better problem solver, for which I am eternally grateful.

I am forever indebted to my wife Tamina Lebek, with whom I spent countless hours discussing my silly ideas. It is also thanks to her suggestions and creativity that my line art figures look the way they do. Tamina, I appreciate your relentless support and encouragement – you are my favourite biochemical reaction.

I would also like to express my gratitude to the beautiful people in the Marsh Lab, especially but without any particular order, to Lukas Gerasimavicius, Benjamin Livesey, Marcin Plech, and Ankit Pathak, for their helpful feedback and thought exchange.

Ultimately, I owe the greatest respect and appreciation to my family, my parents Ilona & Mihaly and my sister Boglarka, who had created an environment in which I could thrive, supporting me with my every decision and throughout this journey.

Hoc illud est quod est.

Abstract

Proteins carry out most of the biochemical phenomena necessary for life as we know it. The majority of proteins do not function alone in the cell, but are instead subunits that assemble into complexes with copies of themselves and other proteins. For decades, due to limited evidence to support otherwise, the textbook model was that subunits have to be fully synthesised before they diffuse away and collide randomly with their partners to from a complex. More recently, however, increasing evidence has accumulated, revealing that this model is incomplete. We now understand that many subunits begin the assembly process during their translation on the ribosome. This phenomenon has important implications for the structure, function, and evolution of protein complexes, as well as for the understanding and the prediction of the mechanisms by which genetic mutations cause disease.

The first chapter provides an overview of our current understanding of how and why proteins assemble into complexes. Two classes of complexes are discussed: homomers, which consist of genetically identical copies of a protein and exhibit structural symmetry, and heteromers, which involve the assembly of non-identical proteins and are more common in human cells. I review the historical experiments that contributed to the discovery of cotranslational assembly, including recent breakthroughs that have made its proteome-wide detection possible, which is of tremendous value to this thesis. I provide an overview of genetic mutations in the context of human disease, as the present work has considerable clinical applications beyond its contribution to fundamental biology.

In the second chapter, I investigate the properties of subunit interfaces that influence cotranslational assembly using a combination of proteomic, structural, and computational approaches. I show that cotranslational assembly is particularly common between subunits that form large intermolecular interfaces. To test whether large interfaces have evolved to promote cotranslational assembly, as opposed to cotranslational assembly being a non-adaptive consequence of large interfaces, I compare the sizes of first and last translated interfaces of heteromeric subunits in the proteomes of three evolutionary distant species. This analysis reveals that N-terminal interfaces, on average, tend to be larger than C-terminal interfaces. Notably, the trend is significant in ancient subunits or those organised into operons in bacteria, suggesting that large N-terminal interfaces may have been selected for to seed the assembly pathway cotranslationally.

The third chapter explores an important hypothesis regarding cotranslational assembly: can it counter the dominant-negative effect, whereby the co-assembly of mutant and wild-type subunits impairs the activity of a protein complex? First, I show that cotranslationally assembling subunits are much less likely to be associated with autosomal dominant relative to recessive disorders. Second, I observe that subunits with dominant-negative disease mutations are significantly depleted in cotranslational assembly compared to those associated with loss-of-function mutations. Additionally, I find that complexes with known dominant-negative effects tend to expose their interfaces later during translation, lessening the likelihood of cotranslational assembly. Altogether, I find strong support for the hypothesis that the allele-specific nature of cotranslational assembly can buffer the effect of certain dominant mutations.

In the fourth chapter, I synthesize the hallmarks of cotranslational assembly and discuss their mechanistic interpretations, highlighting the differences between neutralist and selectionist perspectives regarding their functional importance. Finally, in the fifth chapter, by combining a diverse range of gene-level features, I train a computational model for predicting proteins likely to be associated with non-loss-of-function (non-LOF)

disease mechanisms, with the aim of accelerating the discovery of novel disease variants. I first generate a model that utilizes protein complex structural data and showcase its ability to detect properties explicitly absent from the model but are linked to proteins that give rise to non-LOF disease mechanisms. Although the results reflect the idea that LOF and non-LOF mechanisms can be captured at the protein-level, the predictor is strongly limited by the availability of protein complex structural data. Due to this limitation, I introduce a new model architecture with a spectrum of surrogate features, notably excluding those based on experimental protein complex structure data.

The resulting models enable the estimation of probabilities for a protein exhibiting loss-of-function, gain-of-function, and dominant-negative molecular disease mechanisms across the entire proteome. In preliminary results, I demonstrate the practical applications of these models, including the prioritization of mutations with non-LOF-like properties in population genetic data and the detection of cryptic de novo dominant-negative mutations in developmental disorders. This thesis offers fresh insights into the molecular and evolutionary aspects of cotranslational assembly and its role in human disease.

Lay summary

Our DNA encodes proteins that are the workhorses of our cells, performing a wide range of essential tasks, from breaking down food to fighting infections. To help put things into perspective, scientists often say that if we were to stretch out the DNA from a single human cell, it would be as long as two meters. Even more riveting is the fact that all the proteins in a single cell would span many kilometres if laid out end to end. They do not like to be stretched out, though. They curl up into compact shapes and often join forces with other proteins to form complexes, functioning like interlocking Lego pieces. If one protein were to be scaled up to the size of a person, the cellular space would be equivalent to hundreds of football fields densely packed with people. In such a vast expanse, if both you and your friend started walking, it is very unlikely that you would meet within a reasonable timeframe. However, there is a clever way to increase the chances of meeting: enter the football fields together. Similarly, recent discoveries have unveiled that protein complex formation begins even before proteins are fully synthesized by ribosomes, which are the cell's protein factories. This process, known as cotranslational assembly, enables proteins to efficiently and accurately form complexes.

My research explores the world of protein complexes, specifically, how and why proteins come together and what factors determine the nature of their association. I review the experiments that revealed how proteins can form complexes during their synthesis, and how new methods can help us discover more examples of this process in different organisms. Thanks to these experiments, we now know that at least 4,000 proteins, a fifth of all the proteins encoded by our genes, use cotranslational assembly. I also explain how genetic mutations, which are changes in the DNA code that tells the cell how to make proteins, can affect the formation and function of protein complexes and lead to disease. Moreover, I aim to answer unresolved questions about cotranslational assembly.

What makes some proteins more likely to form complexes during their synthesis than others? It turns out that proteins with larger contact areas between their subunits tend to join together earlier in the assembly process. Imagine building a Lego structure composed of pieces of various sizes. Larger pieces have broad knobs that seamlessly fit into the corresponding slots of other pieces. You find yourself attaching these together earlier in the building process, making sturdy connections that lay the foundation for the entire structure. Conversely, smaller pieces with narrower knobs demand greater precision and careful alignment to find their designated places, leading to their attachment later in the building process. This principle also applies to a single protein that interacts with multiple others. The larger contact area is usually at one end of the protein where its synthesis starts, which helps it form complexes sooner. I have found this pattern in bacteria, Baker’s yeast, and human cells, which goes to show that cotranslational assembly is an important process that has been conserved throughout evolution.

Can cotranslational assembly reduce the likelihood of certain mutations leading to disease? In our DNA, we usually have two copies of each gene, one from each of our parents. If one copy carries a mutation and the assembly occurs randomly, approximately three-quarters of the complexes will contain at least one mutant, potentially leading to disease. These mutations, known as dominant-negative variants, can disrupt complex function. However, the fascinating nature of cotranslational assembly lies in its potential to counteract such mutations. When assembly is cotranslational, both gene products form complexes independently of each other, reducing the proportion of mutant-containing complexes to 50% and in turn decreasing the likelihood of disease. Through my research, I have analysed cotranslationally assembling proteins and human disease genes with dominant-negative variants, providing robust support for this hypothesis. Interestingly, complex-forming

proteins that do not start their synthesis with a large contact area seem to be more prone to such dominant-negative variants. This is because they form complexes after being made and because of this they can mix with other subunits in the cell, increasing the chance of being disrupted by mutants.

Lastly, I explore the practical applications of my research in understanding and predicting the effects of disease mutations with specific molecular mechanisms. I develop computational models that utilize various protein properties, such as the contact area, to estimate the likelihood of a protein causing a disease by altering or negating its function. Through extensive testing on comprehensive datasets of genetic variants, the model shows its ability to identify proteins with properties that indicate function change or disruption upon mutation. My hope is that this work provides insights into the captivating process of protein complex formation, casting light on its impact on structure, function, and evolution. Moreover, it exemplifies how this knowledge can serve as a valuable tool in uncovering new mechanisms behind genetic diseases.

Contents

Abbreviations

AD – autosomal dominant

AR – autosomal recessive

BTB – broad complex, Tramtrack, and Bric-à-brac (a protein/domain family)

Cryo-EM – cryo-electron microscopy

DiSP – disome selective ribosome profiling

DN – dominant-negative

DNA – deoxyribonucleic acid

DP – dominant-positive

DMS – deep mutational scanning

DQC – dimerisation quality control

EDC – extent of disease clustering

ESI-MS – electrospray-ionisation mass spectrometry

F1 – harmonic mean of the precision and recall

GOF – gain of function

LOF – loss of function

MAVE – multiplexed assay of variant effect

MCC – Matthews correlation coefficient

NED – non-exponentially decaying (protein)

NMR – nuclear magnetic resonance

NPV – negative predictive value

pLDDT – predicted local distance difference test

PPV – positive predictive value

RNA – ribonucleic acid

PR – precision-recall

RBP – RNA binding protein

ROC – receiver operating characteristic

SeRP – selective ribosome profiling

SNV – single nucleotide variant

VUS – variant(s) of uncertain significance

VEP – variant effect predictor

ΔΔG – change in the Gibbs free energy (of a process)

Databases and resources

ClinGen – The Clinical Genome Resource is a National Institutes of Health funded initiative that defines the clinical relevance of genes and variants for clinical research and precision medicine applications.

ClinVar – A database of human genetic variants annotated with a clinical significance classification according to a set of assertion criteria, as well as the confidence of clinical interpretation.

CORUM complexes – The Comprehensive Resource of Mammalian protein complexes is a database that provides a curated repository of experimentally characterised complexes from mammalian organisms.

Complex Portal – The EMBL-EBI Complex Portal is a curated, encyclopaedic resource of macromolecular complexes from a number of key model organisms.

AlphaFold database – The AlphaFold Protein Structure Database is a database that provides open access to millions of AlphaFold protein structure predictions to accelerate scientific research.

gnomAD – The Genome Aggregation Database is a resource developed by an international coalition of investigators seeking to aggregate and harmonise exome and genome sequencing data from a variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

hu.MAP 2.0 – hu.MAP 2.0 is a human protein complex map that provides a comprehensive view of protein complexes in human cells.

Membranome 3.0 – Membranome 3.0 is a database of single-pass membrane proteins with structures.

OMIM – Online Mendelian Inheritance in Man is a continuously updated catalogue of human genetic disorders and traits, with particular focus on the molecular relationship between genetic variation and phenotypic expression.

OperomeDB – OperomeDB is a database that provides an ensemble of all the predicted operons for bacterial genomes using available RNA-sequencing datasets across a range of experimental conditions.

PANTHER – Protein ANalysis THrough Evolutionary Relationships is a database of protein families based on evolutionary relationships.

PaxDB – PaxDB is a comprehensive absolute protein abundance database that contains whole-genome protein abundance information across organisms and tissues.

PDB – The RCSB Protein Data Bank is a comprehensive resource that provides access to the 3D structures of biological macromolecules.

PubMed – PubMed is a free search engine that provides access to MEDLINE, the National Library of Medicine's database of citations and abstracts in the fields of medicine, nursing, dentistry, veterinary medicine, health care systems, and preclinical sciences.

SwissModel repository – The SWISS-MODEL Repository is a database of annotated 3D protein structure models generated by a homology-modelling pipeline based upon conservation of the interface.

UniProt Knowledgebase – The UniProt Knowledgebase is a comprehensive resource for protein sequence and functional information with extensive cross-references to more than 200 external databases.

List of figures and tables

Main figures

Tables

Appendices

1 Introduction

1.1 Principles of protein complex assembly

1.1.1 Protein complex structural data

The intracellular environment is a bustling realm where proteins and biological macromolecules constantly interact. These interactions vary greatly in terms of frequency, specificity, and duration. Some proteins come together to form stable assemblies with well-defined functions, allowing for experimental characterization. However, numerous transient and promiscuous interactions also take place within cells, often driven by the crowded nature of the intracellular space. While many of these fleeting interactions may lack biological relevance and evolutionary selection, some are vital for certain processes, particularly cell signalling. Given the dynamic nature of cellular activities and the multitude of interactions occurring, defining a protein complex becomes a challenging task. Are all interactions considered complexes, regardless of their transient or nonspecific nature? Moreover, if we aim to differentiate transient interactions from stable protein complexes, how do we establish a threshold (Marsh and Teichmann 2015)? Since obtaining a comprehensive picture of all interactions within a cell is not feasible (yet), our understanding of protein complexes heavily relies on the experimental methods employed for their characterization. Although defining a protein complex as a collection of proteins that consistently copurify in high-throughput experiments provides a starting point, gaining a deeper understanding of the structure-function relationships and the mechanisms of subunit assembly necessitates the use of experimental methods that offer structural data.

The determination of protein structures has predominantly relied on x-ray crystallography, which has provided a vast collection of structural data. In this method, a crystals of a protein are grown in solution and its structure is determined by analysing the diffraction pattern of x-rays interacting with the crystal lattice. Depending on the definition of the asymmetric unit, the structures can be classified into three categories: monomers, which consist of one polypeptide chain; homomers, which are assemblies formed by multiple identical polypeptide chains; and heteromers, which are complexes comprising multiple distinct polypeptide chains. Interestingly, there was a notable surge in the discovery of new monomers and homomers in the early 2000's, attributed to the proliferation of structural genomics projects focused on novel protein folds (**Figure 1.1**) (Marsh and Teichmann 2015). As these projects had successfully characterized a good portion of easily crystallisable protein folds, the annual discovery of new monomers and homomers had slowly plateaued. Meanwhile, exploration of the heteromeric quaternary structure space continues to be very active – in parallel with the technological advancements in the field.

Two decades ago, nuclear magnetic resonance (NMR) emerged as the second most common method for protein structure determination. NMR structures primarily consist of monomers because large protein complexes pose challenges for the method (**Figure 1.1**). This is because they often give spectra with extensive signal overlap, leading to difficulties in resolving individual resonances. Additionally, the low molecular tumbling rates and inherent dynamics of larger complexes can further hinder the acquisition of high-quality data (Mainz et al. 2009). However, new methods are being developed to overcome these challenges and obtain atom-level structural information on larger systems. For example, magic-angle spinning NMR, which improves the

spectral resolution of samples by spinning them at a high frequency and a specific angle to the magnetic field, has been successfully applied to complexes as large as a 10 megadalton HIV-1 capsid tube (Lu et al. 2020).

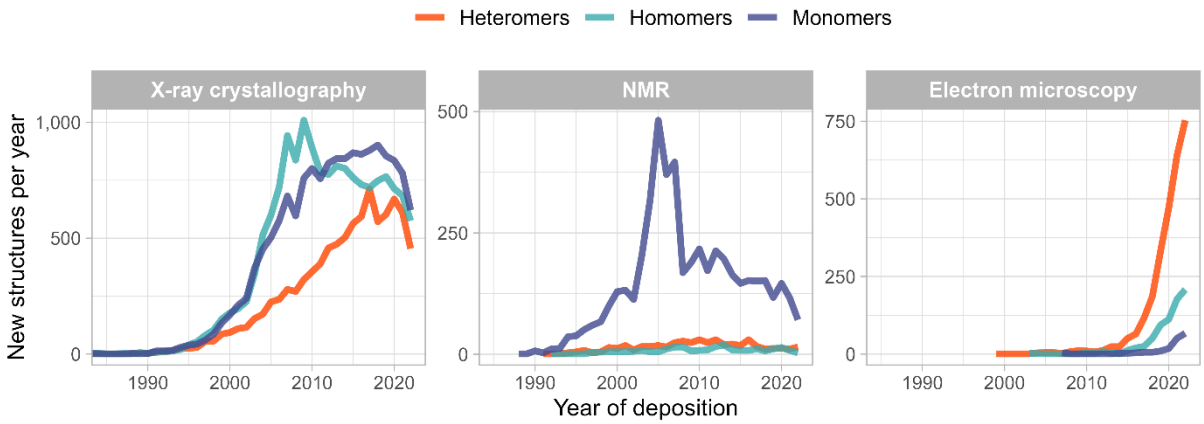


Figure 1.1 Number of yearly releases in the PDB by oligomeric state and experimental method. The plots are based on a recent snapshot of the PDB (July 20, 2023) and only consider the nonredundant set of structures below 50% sequence identity.

In recent years, cryo-electron microscopy (cryo-EM) has revolutionized our ability to visualize protein complexes with exceptional precision and in their near-native state. The technique is based on the idea that cooling the sample to the temperature of liquid nitrogen increases its tolerance to higher electron doses. However, directly resolving details from raw micrographs remained problematic for a long time. In a pioneering breakthrough, Henderson and Unwin tackled this issue by introducing a crystallographic approach: averaging images obtained from 2D crystals composed of many identical proteins (Unwin and Henderson 1975). Since then, through addressing other limitations observed in traditional electron microscopy, cryo-EM has expanded our capacity to study larger assemblies, including ribosomes, whole viruses, and molecular machines. Additionally, the emergence of cryo-tomography allows the reconstruction of protein complexes within intact tissues, unveiling their spatial organization and dynamics (L. N. Young and Villa 2023). Given the success of cryo-EM techniques (**Figure 1.1**), it is highly likely that these structures will soon dominate the space of heteromeric protein complexes in the Protein Data Bank (Berman et al. 2000).

These structure determination techniques, each with its strengths and limitations, have given us a relatively large compendium of protein structural diversity. The deposition and analysis of thousands of structures by scientists worldwide played an invaluable role in advancing the field. Collectively, these structures helped lay down the foundations of how protein complex subunits recognise one another (Chothia and Janin 1975) and made possible the exhaustive enumeration of protein complex topologies (Sebastian E. Ahnert et al. 2015). Very recently, a deep learning-based protein structure prediction method, AlphaFold, has swept through the community for its accurate predictions. The method exploits both the available protein structural data and residue coevolution in related sequences contained in the exponentially growing protein sequence universe. AlphaFold's architecture relies on the *Evoformer* unit (Jumper et al. 2021), which jointly updates the representation of residue pairs in the sequence being "folded", with a representation of many, evolutionarily similar sequences. Because a coevolutionary signal can be detected between residue pairs of subunits, AlphaFold and methods alike hold great promise for protein complex prediction (Humphreys et al. 2021).

1.1.2 Why do proteins assemble into complexes?

The primary sequence of a protein is determined by the order of twenty proteinogenic amino acids. It, for the most part, assumes the secondary and tertiary structures. The former refers to the local conformation of the protein backbone (the alpha carbon trace) stabilised by main chain hydrogen bonding patterns, with the most common structural elements being alpha helices and beta strands. The latter refers to the protein's 3D structure, which forms via further folding and stabilisation of secondary structure elements by noncovalent interactions, such as hydrogen bonds, salt bridges and van der Waals forces, occurring between amino acid side chains. It seems evolution has proteins quite well figured out – then why do so many proteins have a quaternary structure, too? The quaternary structure of a protein refers to the way in which two or more individual protein chains (subunits) are arranged in a complex with respect to each other. Remarkably, the majority of proteins in any given cell exhibit a quaternary structure (Lynch 2012). Several mutually inclusive arguments have been raised as to why they are so common, with some of the most compelling ones as follows.

Oligomer formation builds larger proteins with less information (Crick and Watson 1957). Take, for example, hemolysin E, which is a homomeric toxin responsible for the haemolytic phenotype of several bacterial strains. Twenty-four subunits, 303 amino acids each, form one large pore-like complex (Mueller et al. 2009). Having a separate gene for each subunit or one large protein consisting of 7,272 amino acids would require 24-times the space in the genome as that of the single-copy subunit. This effect would be exacerbated in higher order eukaryotes, where intronic gene structure represents additional complexity to gene organization.

Oligomer formation improves the chance of error-free protein synthesis (Crane 1950). All biological processes are susceptible to error, even RNA-polymerases and ribosomes make mistakes. In bacteria, they do so at a rate of approximately 10^{-5} and 10^{-4}, respectively (W. Li and Lynch 2020; Parker 1989). Thus, factoring in processivity errors, over 97% of the hemolysin E subunits are correctly synthesized, with only about 3% potentially carrying a substitution. By contrast, if the whole complex was instead encoded by a single gene coding for a protein of 7,272 amino acids, the probability of having no missense mutation in any of the protein copies would be about 30%. Notably, these back-of-the-envelope calculations assume independence between transcription and translation, which is generally not the case. A single mRNA copy can, and usually does, template a protein multiple times. The point is that selective degradation of many large, faulty proteins would entail a far larger metabolic cost than that of occasionally faulty, short proteins.

Oligomer formation allows subunit recycling. Many protein complexes assemble reversibly and their functions benefit hugely from this property. The most commonly used example is cytoskeletal proteins, such as tubulin. Assembly of tubulin subunits into microtubules is characterised by a process known as dynamic instability (Mitchison and Kirschner 1984), where microtubule ends can grow and shrink rapidly with newly depolymerized subunits continuously replenishing the subunit pool. The process obviates the need for synthesizing new proteins at the rate polymerization would require, which would take more time, energy, and resources. Synthesis of an equivalently large structure occupying the same functional niche by a single polypeptide chain would be an absurdly wasteful and probably unsuccessful strategy for the cell. Yet, this strategy is not always a concern for some cytoskeletal proteins that fulfil more static structural roles. For example, the giant muscle protein titin consists of 3,450 amino acids and almost 300 immunoglobulin- and fibronectin-like domain repeats. Titin also seems to defy the error-free synthesis argument, as each copy is

estimated to contain around 17 missense errors (Allan Drummond and Wilke 2009), meaning that, on average, human sarcomeres contain no error-free titin molecules at all.

Oligomer formation is favoured because of biophysical reasons (Lynch 2013). Arguably, it is generally easier to fold multiple small proteins rather than a large one. Furthermore, complexation reduces the surface area to volume ratio of a protein, which can in turn reduce its vulnerability to denaturation and its propensity to engage in nonspecific interactions. In enzymes, oligomerisation could reduce the sensitivity of catalytic sites to internal fluctuations and enhance substrate specificity. Oligomerisation of enzymes can also create multiple catalytic sites, increasing the frequency of productive enzyme-substrate encounters. Finally, complexation presents opportunities for allosteric activity regulation (Jacques Monod, Changeux, and Jacob 1963), defined as cooperation (information transfer) between distant sites within or between proteins propagated by the residue network.

In addition to the intuitive but predominantly adaptive arguments above, there is a nonadaptive theory for the existence of protein complexes, pioneered by Michael Lynch. He proposed a model whereby transitions between multimeric states of protein complexes exist in equilibrium in deep time (Lynch 2012). In his words: "*Although we do not have the luxury of making such observations, provided enough evolutionary time has elapsed for the tree of life to have reached the steady-state distribution [...] the number of transitions from a monomeric to a dimeric state should equal that in the opposite direction*" (Lynch 2013). These transitions stochastically arise via the joint processes of mutation, random genetic drift (Kimura 1968), and constant directional selection. While there are clear examples that complexation can provide adaptive benefits, neutral processes are just as important and cannot be dismissed, especially in the light of accumulating evidence (Esin et al. 2018; Garcia-Seisdedos et al. 2017; Hochberg et al. 2020; Schulz et al. 2022).

1.1.3 Quaternary structure topology

When talking about protein complexes, topology refers to the spatial arrangement of the subunits. The complex can be made up of sequence-identical subunits, like in homomers, or of genetically distinct subunits, like in heteromers. The majority of homomeric complexes are symmetric and even most heteromers can be related to a simpler symmetric topology (Goodsell and Olson 2000; Marsh and Teichmann 2015). This is in part explained by the observation that many heteromeric interactions trace back to homomeric interactions of one gene that underwent duplication with the paralogue subsequently diverging in sequence but maintaining the interface (Pereira-Leal et al. 2007). Symmetry can be considered at the level of individual subunit interfaces. A twofold axis of rotational symmetry arises from dimerisation, creating an isologous (symmetric or head-to-head) interface between two identical surfaces on homomeric subunits. On the other hand, cyclization leads to higher-order rotational symmetry and produces a heterologous (asymmetric or head-to-tail) interface between two different surfaces. Therefore, by definition, interfaces of heteromers are also heterologous.

Symmetry can also be considered globally at the level of the protein complex (**Figure 1.2**). Homomeric symmetries are typically denoted with the Schönflies point group naming convention (Downward 1973), which are symmetry operations that can be performed without changing the initial subunit arrangement. In biology, because most molecules have an intrinsic handedness (e.g., left-handed amino acids and right-handed sugars), only two types of symmetry operators are relevant: rotation and translation. According to this, twofold dimeric complexes (C_2) possess twofold rotational symmetry. Cyclic complexes ($C_{n\,[n>2]}$) display higher-order rotational

symmetry, like threefold (C_3) or fourfold (C_4), forming closed ring structures with heterologous interfaces. Dihedral complexes ($D_{n\,[n>1]}$) combine an n-fold axis of rotation with n perpendicular twofold axes, forming, for example, "dimer of dimers" from C_2 homomers (D_4) or "dimer of trimers" from C_3 homomers (D_6). Cubic symmetries include tetrahedral (*T*), octahedral (*O*), and icosahedral (*I*) complexes, which combine a number of higher-order rotational axes. Helical complexes (*H*) exhibit a screw-like symmetry composed of both rotational and translational axes. Helically symmetric complexes are topologically open and thus their subunits would polymerise endlessly under ideal conditions, explaining why they are commonly found in fibre-forming proteins, such as microtubules. While infinite assembly can have functional roles in cells, it appears to be negatively selected against (Garcia-Seisdedos et al. 2017), reflecting the aversion of random mutations that could create new interfaces with unlimited assembly potential. Lastly, asymmetric complexes (C_1) lack significant rotational symmetry, with subunits found in topologically nonequivalent, i.e. non-bijective (Sebastian E. Ahnert et al. 2015) positions. Most of these non-bijective homomers are due to quaternary structure misassignment in the PDB, often as a result of the biological assembly, the physiologically relevant topology of the complex, being erroneously set to the asymmetric unit (Bergendahl and Marsh 2017).

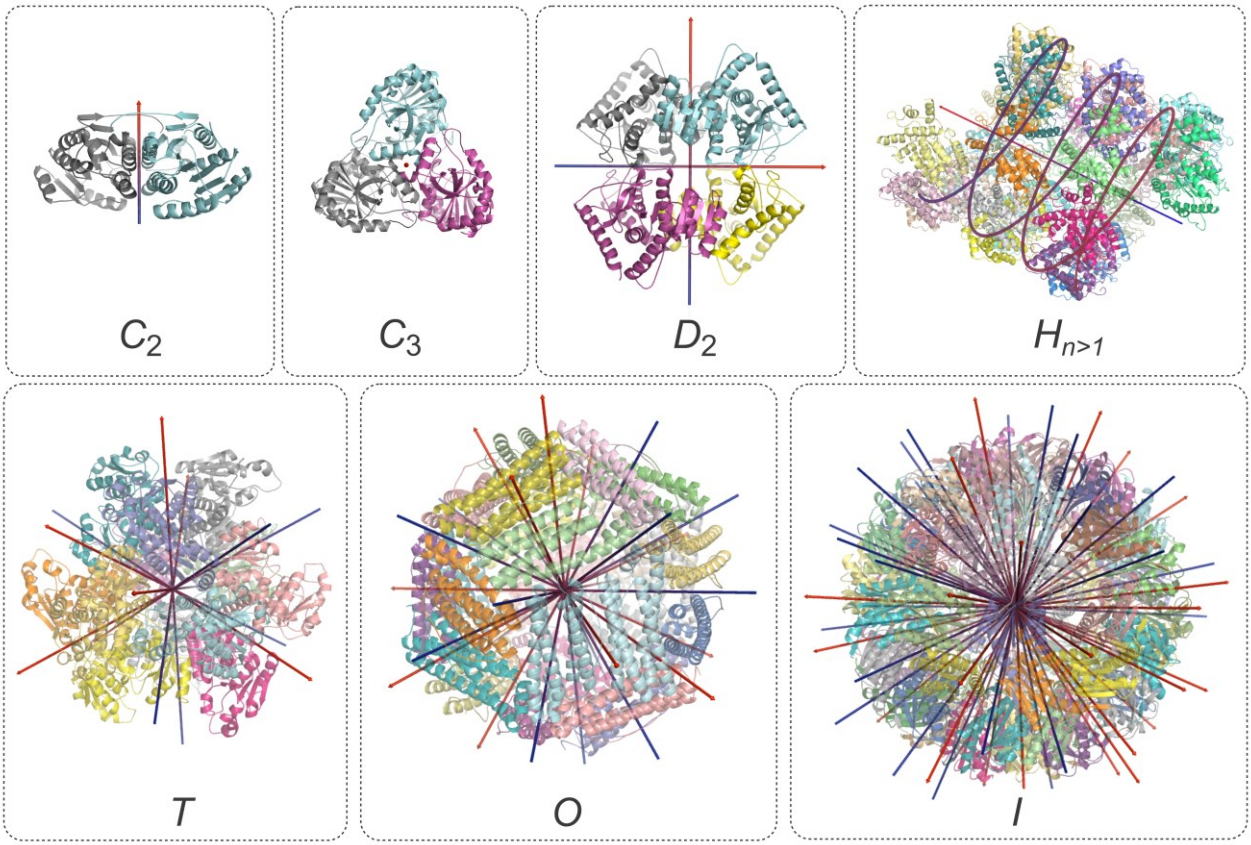


Figure 1.2 Quaternary structure topologies of symmetric homomers.
PDB codes for the structures are: 1mg5 (C_2), 1a9q (C_3), 1a5z (D_2), 601f (*H*), 1a1s (*T*), 1aew (*O*), and 1hqk (*I*). Symmetry axes were drawn with the AnAnaS software (Pagès, Kinzina, and Grudinin 2018) and the CGO arrow PyMOL plugin.

What explains the strong tendency of protein quaternary structure topology to be symmetric? Monod, Wyman, and Changeux noted that each interaction in the interface of a symmetric dimer occurs twice (Jacque Monod, Wyman, and Changeux 1965), causing the net mutational effect to be larger and increasing in turn the rate with which evolution can act on the complex. André et al. have tested this hypothesis using docking simulations

and found that, while the energy associated with mutations is $\sqrt{2}$-fold greater for symmetric interfaces, there in an opposing effect: asymmetric interfaces can contain twice as many unique residues, which increases the mutational space available for evolutionary innovation (André et al. 2008). It has also been argued before that the lowest energy state of an assembly is a symmetrical one (Blundell and Srinivasan 1996; Cornish-Bowden and Koshland 1971). Lukatsy et al. investigated random interactions between two planar disks and showed that the energy is lowest if the disks interact with equivalent surfaces (symmetric) rather than different ones (asymmetric) when aligned along the same axis (Lukatsky, Zeldovich, and Shakhnovich 2006). Likewise, simulations by André et al. suggest that evolution has a higher chance of finding symmetric interfaces, because they appear to populate most of the energetically favourable interactions it can choose from (André et al. 2008). More recently, an information theoretic approach to thinking about symmetry was developed (Johnston et al. 2022). To explain, the authors invoke the infinite monkey theorem, which claims that a monkey randomly hitting keys on a keyboard for an infinite amount of time will eventually type any given text, even the complete works of William Shakespeare (Isaac 1995). On a keyboard with m keys, the probability of randomly typing a text of length n is $1/m^n$. However, if the text is a computer script with instructions to repeat a character many times, the computer script will probably be more compact than all characters typed out, thus have a higher probability of being typed by chance. The authors estimated the Kolmogorov complexity (a measure of the shortest algorithmic expression) of biological assemblies in the PDB, and found a strong bias for low-complexity high-symmetry topologies. Nature's code remains the envy of any computer programmer.

1.1.4 Theory of interface size

Proteins collapse into their folded state because hydrophobic energy is gained by the reduction of surface in contact with water (Kauzmann 1959). This concept was later extended to interfaces of protein complex subunits. Chothia and Janin found that by far the largest contribution to the free energy of binding is made by hydrophobic bonding, accounting for approximately 25 cal/mol for every 1 Å^2 of buried surface (Chothia and Janin 1975). Later, once enough protein complex structural data had been amassed, they showed that large interfaces are usually more hydrophobic (Janin and Chothia 1990). This can be explained by a more recent observation that larger interfaces tend to have more core residues, which grant most of their surface area to the interface and are more hydrophobic than rim residues (Levy 2010). While the size of the interface, that is, the average surface area removed from contact with water between subunits, can be used to estimate binding affinity (Janin 1995), it has limitations. A frequently raised argument is that interface size correlates less with affinity in subunits that undergo large conformational changes upon binding (Kastritis et al. 2011) and in heteromeric subunits (J. Chen, Sawyer, and Regan 2013). Although these limitations emphasize that other interfacial properties must not be overlooked (Kastritis and Bonvin 2013), interface size remains an easily calculable and versatile property of protein complexes with tremendous utility for hypothesis testing.

To use interface size in structural, functional, and evolutionary analyses of protein complexes, one first has to study their dynamic assembly processes and assess how effective it is for their prediction. Advancements in electrospray ionisation mass spectrometry (ESI-MS) allowed for the monitoring of quaternary structure disassembly and the detection of intermediate subcomplexes at each step (Hernández and Robinson 2007). Application of ESI-MS to a set of homomers unveiled an interesting trend: the assembly intermediates could be inferred from the crystal structures by preserving the largest interfaces and breaking the smaller ones (Levy et al. 2008). Later, the method was performed on heteromeric complexes, with an iterative model in which

subunits are disassembled in a way that exposes the least amount of buried interface area (Marsh et al. 2013). Again, the majority of assembly intermediates were consistent with the interface size hierarchy predicted from available structural data. Moreover, the authors mapped gene fusion events onto complex assembly pathways, and found that they had a significant tendency to preserve subunit assembly order, suggesting evolutionary selection for ordered protein complex assembly in cells. This result was recapitulated in an analysis of bacterial operons, which found that gene order corresponds well to the predicted subunit assembly order (Wells, Bergendahl, and Marsh 2016). It seems evolution tends to stabilise interfaces that are functionally beneficial. Indeed, evolutionarily ancient subunits of protein complexes are larger than those that emerged in a lineage-specific manner, and are thus more recent (Dayhoff et al. 2010).

Essentially, these results suggest that the largest intersubunit interfaces are formed first during assembly, while the smaller ones are formed later, assuming the order in which new subunits are gained during the course of evolution. Why is interface size so effective at predicting assembly order given its imperfect agreement with experimentally measured binding affinity? One possibility is that assembly can only proceed by forming a limited number of possible interfaces, whose sizes can vary considerably. Therefore, only a weak correlation between affinity and interface size would be needed to account for the successful prediction of most assembly pathways (Marsh and Teichmann 2014a). Recently, it has been shown that the number of coevolving residues between subunits is an equally good predictor of assembly order (Mallik and Kundu 2017). The reason for this might be that interface core residues are often the most conserved (Levy 2010), and, since larger interfaces have a greater proportion of interface core residues, the number of coevolving residues serves simply as a proxy for interface size. However, while interface size requires protein complex structural data, residue coevolution requires sufficiently deep evolutionary signal from multiple sequence alignments; thus both methods have their limitations.

1.2 Cotranslational protein complex assembly

1.2.1 A brief history of cotranslational assembly

In the past sixty years, strong evidence has been laid down supporting the idea that the assembly of protein complexes frequently occurs cotranslationally, while at least one of the component subunits is in the process of being translated. Proteins from both prokaryotes and eukaryotes, spanning a wide variety of folds, have been observed to undergo such assembly, for example, the bacterial beta-galactosidase (Zipser 1963), the viral outer capsid protein sigma-1 (Gilmore et al. 1996), the yeast fatty acid synthase (Shiber et al. 2018), the plant photosystem II protein D1 subunit (L. Zhang et al. 1999), and many mammalian nuclear transcription complexes (Kamenova et al. 2019).

Over time, different experimental approaches have been designed to probe proteomes for their ability to produce cotranslationally assembled complexes. Duncan and Mata employed RNA immunoprecipitation to examine 31 proteins from fission yeast (Duncan and Mata 2011). The premise of the experiment was that, given a pair of interacting subunits, one would associate indirectly to the mRNA encoding the other. The association could be tested by purifying one protein and identifying the bound mRNA via a chip-based hybridisation technique (Keene, Komisarow, and Friedersdorf 2006) (**Figure 1.3**). Of the 31 proteins tested using this approach, 12 showed reproducible binding to other mRNAs, indicating potential cotranslational interactions. Although the target proteins were carefully selected to exclude known RNA-binding domains, and the authors

later validated the method's ability to predict genuine protein-protein interactions (Duncan and Mata 2014), the result provided only inferential evidence for the widespread nature of cotranslational assembly.

Our ability to interrogate cotranslational phenomena drastically increased with the development of ribosome profiling and its derivatives (Ingolia 2014). Shiber et al. applied selective ribosome profiling (SeRP) in budding yeast to analyse 12 heteromeric complexes for their assembly mode (Shiber et al. 2018). The method involves comparing the distribution of ribosome-protected mRNA fragments (footprints) from two fractions: one containing those of a selected set of ribosomes co-purified with a tagged interaction partner (selected translatome) and the other containing the remaining footprints of the target mRNA (total translatome) (**Figure 1.3**). The build-up of footprints in the selected relative to the total translatome reveals if cotranslational assembly occurs between two preselected proteins. The authors demonstrated that 9 out of the 12 complexes underwent cotranslational assembly, providing the first direct indication that cotranslational subunit binding may be a common assembly mechanism.

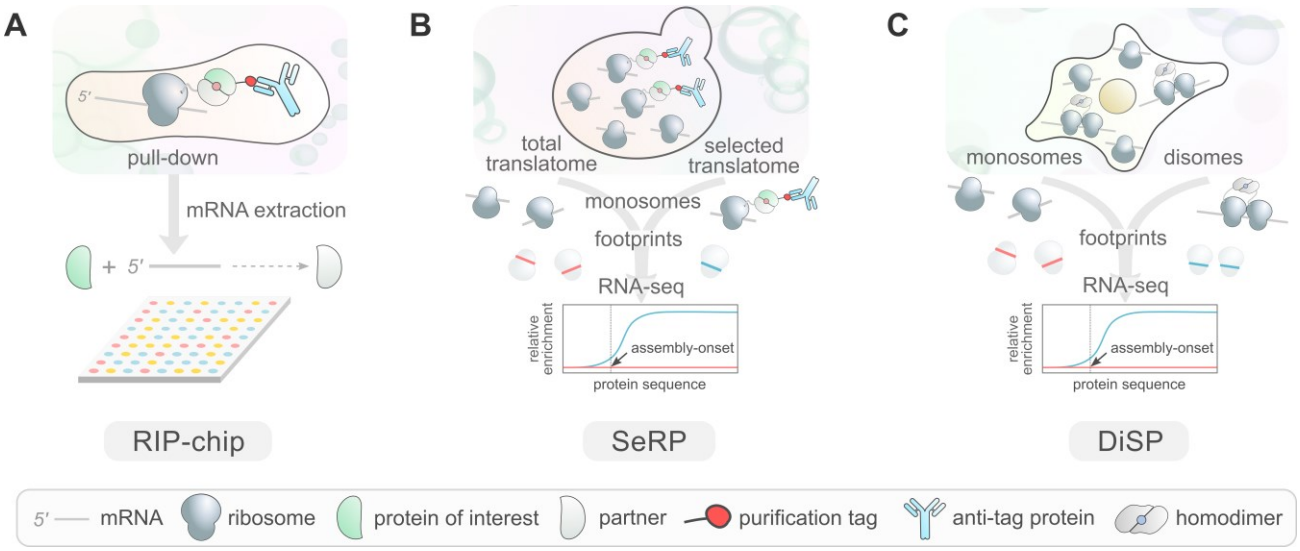


Figure 1.3 Methods to detect cotranslational assembly.

(**A**) RIP-chip assay (ribonucleoprotein immunoprecipitation analysed with DNA chips) performed by Duncan and Mata (Duncan and Mata 2011). (**B**) SeRP performed by multiple studies for the detection of cotranslational assembly (Seidel et al. 2022; Shiber et al. 2018; Shieh et al. 2015). (**C**) DiSP performed by Bertolini et al. (Bertolini et al. 2021).

Even more recently, the high-throughput survey of cotranslational assembly became possible via disome-selective ribosome profiling (DiSP) (Bertolini et al. 2021) (**Figure 1.3**). In this, single ribosomes (monosomes) as well as double ribosomes tethered together via interacting nascent chains (disomes) are isolated from the total translatome. By sequencing the footprints contained in both fractions, it is possible to generate a monosome-to-disome enrichment profile for every transcript, which pinpoints the position of the assembly onset in the protein sequence. Using this powerful method, Bertolini et al. have identified over 4,000 human proteins with evidence for cotranslationally assembly, further strengthening the support for the phenomenon.

1.2.2 Generalized mechanisms of cotranslational assembly

Protein synthesis depends on an mRNA transcript that is translated vectorially from the N to the C terminus. The mRNA may be translated by multiple ribosomes at the same time in both prokaryotes (Barondes and

Nirenberg 1962) and eukaryotes (Warner, Rich, and Hall 1962). In the simplest case, considering binary interactions, this model can result in two different mechanisms of cotranslational assembly, which can be divided further into two categories based upon the maturity of the interaction partner (**Figure 1.4**). First, assembly can happen through cis- or trans-acting mechanisms; the former involves subunits from the same transcript binding to each other, while latter involves subunits from different transcripts. Second, the assembly can be categorised as either simultaneous (also referred to as "co-co" assembly), with both proteins being translated at the same time, or sequential ("co-post" assembly), with the partner protein having been released from the ribosome at the time of binding (Bertolini et al. 2021; Kamenova et al. 2019). Additionally, high complexity translation-coupled assembly modes involving multiple protein chains are known to exist (Halbach et al. 2009; Khan et al. 2022; Redick and Schwarzbauer 1995; Seidel et al. 2022), such as that of the major vault protein (**Figure 1.4**) (Mrazek et al. 2014), but their detection requires case-specific techniques, limiting the extent to which they have been studied so far.

The four different modes of cotranslational assembly that result from binary protein interactions have all been observed in vivo. The first case of cotranslational assembly assumed to be in cis was identified in studies on the bacterial β-galactosidase (Kiho and Rich 1964; Zipser 1963), which assembles into a complex composed of four identical subunits, with two being necessary for the formation of an active site. By using sucrose density gradient centrifugation, researchers were able to isolate partially active polysome-bound protein, suggesting the presence of an assembly intermediate. This was the earliest indication that bacterial polysomes, which contain 70-80% of the ribosome pool (Kiho and Rich 1964), are not only protein factories but also potential sites of complex assembly.

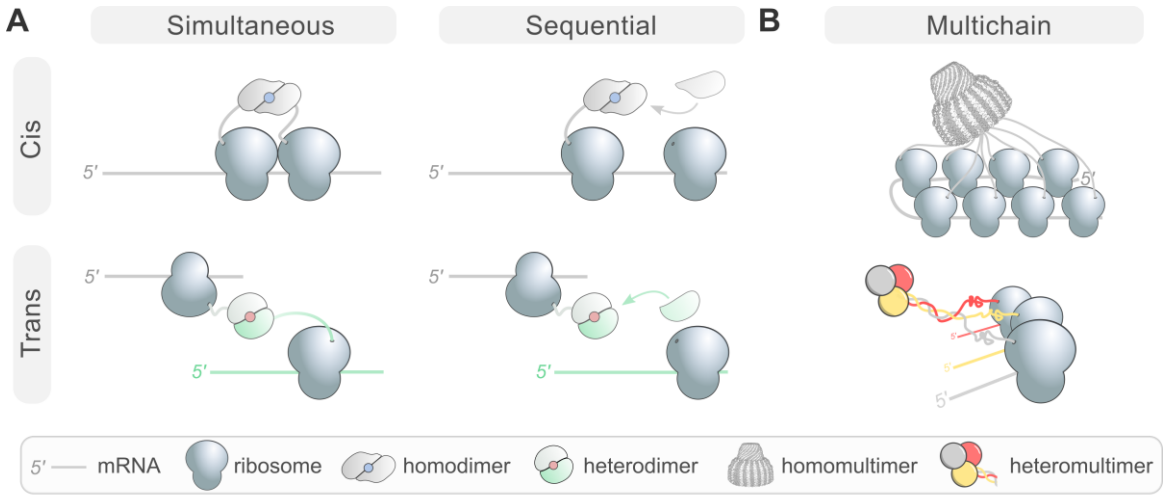


Figure 1.4 Cotranslational assembly modes.

(**A**) Categories of binary interactions grouped by assembly mechanisms (cis and trans) and the maturity of the assembly partner (simultaneous and sequential). (**B**) High complexity cotranslational assembly modes involving multiple nascent chains. Top: Assembly of the major vault protein (Mrazek et al. 2014). Bottom: Hypothetical simultaneous assembly of three heteromeric subunits.

Continuing the trend, researchers later identified a trans cotranslationally assembling subunit pair in chicken muscle cells (Isaacs et al. 1992). Using tritiated puromycin to label the protein titin, they found that native contacts with myosin had formed during its translation. However, these experiments could not determine if the assembly took place simultaneously or sequentially. By contrast, Redick and Schwarzbauer used pulse-

chase labelling to follow the assembly of the protein tenascin into a six-armed structure known as the hexabrachion (Redick and Schwarzbauer 1995). They observed no assembly intermediates, suggesting that nascent tenascin polypeptides had associated before translation was complete, in a simultaneous manner (**Figure 1.4**). Finally, Young and Andrews introduced pause sites into the mRNA of the signal recognition particle receptor alpha subunit (SRα), which allowed for the identification of its sequential cotranslational assembly with the membrane-anchored SRβ subunit (J. C. Young and Andrews 1996).

Currently, DiSP and SeRP are the most effective techniques for the detection of simultaneous and sequential assembly, respectively. Homomers, subunits that form complexes with other copies of themselves, are uniquely suited for DiSP, because multiple active ribosomes on their mRNA make an interaction between two nascent chains more likely, leading to the formation of a disome, which is the target of DiSP. This notion has been confirmed by Bertolini et al., who found that homomers are enriched in simultaneous assembly relative to heteromers (Bertolini et al. 2021). In sharp contrast to this, detection of sequential assembly by SeRP requires tagging a protein of interest to allow for its purification, making it less feasible for large-scale experiments. This technical bottleneck means that the proteome-wide detection of sequential assembly, which may be more common among heteromers (Shiber et al. 2018), has not yet been accomplished.

1.2.3 Spatiotemporal and stoichiometric regulation of protein complex assembly

Evolutionary selection for assembly order and the existence of cotranslational assembly in both prokaryotes and eukaryotes suggest it matters that subunits of complexes find each other at the right place and time. There are, however, two outstanding differences in eukaryotes: they lack operonal genome organisation and mRNA molecules have to leave the nucleus for translation. Hence, mRNAs of subunits belonging to the same complex should colocalise in order for the assembly to be less dependent on random diffusion. Colocalisation is even more important for subunits with large interfaces, because the large number of unburied contacts will increase their conformational entropy and render them more likely to favour the unfolded state (Miller, Lesk, et al. 1987), which can lead to aggregation.

The classical mechanism for partitioning mRNAs for localised translation is targeting to membrane-bounded organelles, like the ER or mitochondria. At the site, the protein can be cotranslationally translocated by an appropriate translocon (Rapoport, Li, and Park 2017) or become membrane-immersed by an insertion apparatus (Hegde and Keenan 2022). While the foregoing strategies typically require a signal peptide to emerge from the ribosome tunnel, there are processes capable of transporting the mRNA itself. Active transport of mRNA as part of ribonucleoprotein granules along the cytoskeleton is the most widely observed mechanism (de Heredia and Jansen 2004). Interestingly, the mRNA cargo can be translationally active, like in the case of pericentrin, whose mRNA is translated during its journey to the site of centrosome assembly (Sepulveda et al. 2018). In neurons, where protein activity is required at distant axonal locations, on-site production becomes particularly crucial for rapid synaptic transmission. This is achieved by long-range transport via dyneins and kinesins on microtubules, as opposed to short-range transport via myosins on actin filaments (Das, Singer, and Yoon 2019). There is now also abundant evidence for cis-regulatory elements in the 3’ UTR region of mRNAs (Mendonsa et al. 2023) being targeted by RNA binding proteins (RBPs), which, functionally analogous to bacterial operons, may give rise to RNA regulons (Keene 2007). Chen and Mayr proposed a model for how RBPs might orchestrate protein complex assembly (X. Chen and Mayr 2022). According to this, high valency RBPs create a mesh to form translationally active membraneless organelles in

the cytoplasm and recruit functionally related mRNAs. These RBPs can transiently retain proteins, thereby ensuring their colocalisation and facilitating their assembly into complexes with each other or with nascent chains of partner subunits. For example, TIGER domains are ER-associated granules that may fulfil this functional niche (Ma and Mayr 2018).

One question naturally arises in the light of these precise colocalisation and cotranslation mechanisms: how do eukaryotic cells regulate protein complex stoichiometry? The balance hypothesis proposes that there is a non-negligible fitness cost to stoichiometric imbalances of protein complex subunits, making them more likely to be dosage sensitive (Papp, Pál, and Hurst 2003). Most protein complex components are produced at rates proportional to their stoichiometry, meaning that their abundances are precisely tuned at the synthesis-level to the amount required based on the number of copies in their complexes (Taggart and Li 2018). Exceptions include peripheral subunits, which have less effect on complex stability than core subunits (Bray and Lay 1997; Oberdorf and Kortemme 2009), or those part of multiple complexes, that is, have moonlighting activities (Matalon, Horovitz, and Levy 2014). Consequences of dosage control are evident through various phenomena, for example, the rapid degradation of surplus ribosomal subunits (Sung et al. 2016) and depletion of heteromers in regions with high copy number variation (Schuster-Böckler, Conrad, and Bateman 2010). Ample evidence points towards protein degradation as the primary mechanism for correcting subunit levels at the final step of gene expression, rather than at earlier stages (Ishikawa et al. 2017), although it has been argued that transcription feedback could influence protein synthesis rates (Veitia and Potier 2015). A notable exception is the condition of aneuploidy, where protein aggregation can be almost as effective as degradation in reducing the levels of excess proteins (Brennan et al. 2019). Surprisingly, some proteins are produced in excess by design. A recent analysis of protein age-dependent degradation revealed that proteins with non-exponential degradation (NED) kinetics are enriched in heteromeric subunits and often become more stable with age (McShane et al. 2016). NED proteins show stronger coexpression with other subunits' mRNA and assemble earlier during the complex assembly pathway. This suggests a model whereby only a portion of newly synthesized NED proteins is stabilised via complex formation, while the rest undergo degradation.

Given the ribosome's role as a central hub that brings together mRNA, proteins, and hundreds of accessory factors, it is likely that translation-level quality control would evolve as a strategic checkpoint for the cell. Similar to mRNA control processes like nonsense-mediated decay, no-go decay, and non-stop decay in eukaryotes, we can expect translation-coupled regulatory mechanisms for proteins. For example, misfolded proteins on the ribosome can be destined for degradation via cotranslational ubiquitination or N-terminal modification (Varshavsky 2019). The extent to which nascent polypeptides are ubiquitinated was initially debated, arguing it may be as high as 30%. However, recent studies utilizing puromycin labelling in mammalian cells provided a more refined estimate of 12 to 15% (Duttler, Pechmann, and Frydman 2013; F. Wang, Durfee, and Huibregtse 2013). In a single cell, with just 50% of the 5 million ribosomes actively translating an mRNA encoding a 500-amino-acid protein at a rate of 5 amino acids per second, a staggering 1.5 million proteins will be synthesized every minute (Harper and Bennett 2016). It is daunting to consider that more than a tenth of these may be degraded before they could even begin to function, but it goes to show how important quality, i.e. the intended function, is for the cell. As for protein complexes, the newly discovered mechanism of dimerisation quality control (DQC) permits detection and selective degradation of aberrantly paired subunits on the ribosome, leaving the native counterparts intact (Mena et al. 2020). Considering that DQC was discovered through the analysis of BTB (broad complex, Tramtrack, and Bric-à-brac) domains, which

cotranslationally assemble at high frequency (Bertolini et al. 2021), it is possible that cotranslational assembly and DQC are co-occurring phenomena. Thus, ribosomes are necessarily the most dynamic entities in the spatiotemporal regulation of protein complex assembly.

1.3 The nature of genetic disease

1.3.1 Mechanisms of genetic dominance

Each of us inherits two alleles from our parents at any given autosomal locus, which may not be identical and carry different mutations relative to each other or to a reference sequence. If the two alleles are identical, the individual is considered homozygous for the allele, and heterozygous otherwise. The classical Mendelian definition for dominance only permits discrete traits where the phenotypes of the heterozygous state and the homozygous state of the dominance-causing allele are the same. In contrast, modern medical genetics defines dominant and recessive based upon the clinical consequences of the heterozygote, which is more suitable for rare diseases, as it is inclusive for intermediate phenotypes (Zschocke, Byers, and Wilkie 2023). However, strictly speaking, mutations are not in themselves dominant or recessive; these terms are more appropriate to describe the *effect* of a mutation on a given trait, which will determine the pattern of its inheritance.

Autosomal recessive conditions occur when both alleles of a gene have a pathogenic variant, which can be homozygous or compound heterozygous, i.e. two different variants on each allele. This means that for an autosomal recessive condition to arise, either one of the parents has to have the disease and the other carry the variant, or both parents have to be carriers. In rare cases, even if only one parent is a carrier and the other is homozygous wild type, a de novo (spontaneously) occurring mutation in the parents' germline or in the post-zygotic embryo could lead to a rare compound heterozygous recessive disorder (Acuna-Hidalgo et al. 2015; Saito et al. 2017). Recessive variants on the X-chromosome are defined based on how the trait behaves in females, thus a homozygous pathogenic variant in females would be hemizygous in males but likewise considered recessive. When these recessive variants are missense, the disease typically results from loss of protein function, as both protein copies will become inactivated (Veitia, Caburet, and Birchler 2018). In rare cases, recessive variants can cause gain of function, such as the p.Arg544Gln substitution in the extracellular calcium-sensing receptor, which induces hyperactivity due to loss of a cation-π interaction (Cavaco et al. 2018).

In contrast, autosomal dominant diseases only require one mutant allele to manifest. When the dominant variant is missense and its effect is loss of protein function, the disease will be due to haploinsufficiency, i.e. dominant loss-of-function (LOF). As the name implies, the phenotype arises because the level of the protein provided by the nonmutant allele is not enough to maintain the intended function. It was recognised early by Sewall Wright (S. Wright 1934) that enzymes are largely resilient to dominant LOF mutations, which he correctly attributed to the law of diminishing returns in metabolic flux. In other words, because enzymes tend to be highly abundant and substrate concentrations are almost never saturating, halving the dose of an enzyme will not equate to halving the product concentration. On the other end of the spectrum are transcription factors, which are overly sensitive to dominant LOF mutations. The reason for this is thought to be that in order for transcription factors to attain transcriptional synergy (Veitia 2003), they must be expressed at specific levels, act cooperatively by binding to multiple sites on the target promoter, form cross-regulatory networks by binding to other transcription factors, and often autoregulate their own expression; all this amounts to substantial non-linearity and thus dose sensitivity (Pan et al. 2006; Seidman and Seidman 2002).

Dominant mutations can also exert their effects through other means, including gain-of-function (GOF) and dominant-negative (DN) effects, collectively referred to as non-LOF mechanisms. A dominant GOF mutation modifies an existing function or introduces a novel one, while a DN mutation directly or indirectly impairs the function of their wild-type counterpart. Many DN and GOF mutations can be referred to as "assembly-mediated" (Backwell and Marsh 2022), where the molecular phenotype depends on complex formation with the mutant protein physically interacting with the wild type, as explored in the next section. However, some classical examples of GOF and DN mutations extend beyond complexes. For example, with our collaborators, we have described de novo GOF mutations in hexokinase 1, which are linked to a neurodevelopmental disorder (Poole et al. 2023). These mutations potentially affect the allosteric glucose 6-phosphate binding site, leading to the loss of autoregulatory function and therefore accumulation of glycolytic intermediates in the brain. Additionally, a non-assembly-mediated but competitive form of DN mutation is observed in the transcription factor PAX8, associated with congenital hypothyroidism. The p.Ser48Phe substitution is believed to neither destabilise the protein nor affect its affinity for DNA but instead competes with wild-type PAX8 for DNA binding, thereby disrupting the synergism of the wild type (Grasberger et al. 2005).

Genetic diseases caused by mutations in a single gene are termed monogenic, often exhibiting a characteristic pattern in a pedigree consistent with Mendel's laws of inheritance. However, the aetiology of genetic diseases can be more intricate, categorized as oligo- or polygenic, and influenced by multiple genes or environmental factors, giving rise to what are known as complex traits (Badano and Katsanis 2002). Surprisingly, even when the genetic basis of a monogenic disease is well understood, its phenotype can be as complex as those seen in complex traits (Scriver et al. 1999). Moreover, mutations in a gene associated with a specific condition in one individual can lead to variable expressivity, resulting in sub-symptomatic phenotypes in other individuals. This phenomenon, known as incomplete penetrance, is thought to arise from randomness in biological processes, such as gene expression, contributing to both genetic and environmental diversity (Cooper et al. 2013; Raj et al. 2010). A recent analysis of 589,306 genomes identified asymptomatic individuals harbouring variants linked to highly penetrant Mendelian diseases (R. Chen et al. 2016), hinting at the presence of modifiers and protective variants. These findings underscore the complex interplay between genetic factors and the environment in shaping the landscape of human genetic diseases, which can now be investigated using genome-wide population genetics approaches.

1.3.2 Protein complex assembly-mediated genetic effects

The effect of a dominant mutation may depend on whether the protein assembles into a complex with more than one copy of itself. In such cases, the mutation is said to have an assembly-mediated effect (Backwell and Marsh 2022). Homomers are particularly vulnerable to such effects, as they interact with identical copies of themselves, but proteins present in heteromeric complexes with at least two copies can also exhibit these effects. Two relatively common effects include the assembly-mediated GOF, also known as the dominant-positive (DP) effect (Backwell & Marsh, 2022), and the assembly-mediated DN effect. Competitive DN effects are likely less common than assembly-mediated DN effects, while simple GOF mechanisms are possibly just as, if not more, frequent than assembly-mediated DP effects. The reason behind this is that in the DN effect, the mutant protein must directly impact the function of the wild-type protein, which is most easily achieved through a physical contact, such as co-assembly (Veitia 2007). On the other hand, simple GOF mutations that affect autoregulatory binding sites, like in the case of hexokinase 1, or those altering substrate specificity in

catalytic binding sites, will inherently lead to GOF at the protein level, irrespective of whether the protein is part of a complex. If the protein is part of a complex and these mutations transmit their effects allosterically to adjacent subunits, then we observe an assembly-mediated DP effect.

A unifying theme of the assembly-mediated DP effect may lie in its connection to complexes whose subunits collectively contribute to function. These complexes are characterized by their reliance on the combined efforts of all subunits to achieve their intended function, rather than each subunit functioning independently. Two examples of such complexes are membrane channels, where the central pore forms only upon cyclization of all subunits (Forrest 2015), and homomers with multi-chain binding sites, where at least two adjacent subunits form one ligand binding pocket (Abrusán and Marsh 2019). It is straightforward to rationalize an assembly-mediated DP effect in these complexes, because in order for a new or altered function to arise, the effect of the mutation has to be intimately liked to function, unlike in a DN effect. For example, the DN mutation p.Glu504Lys in the mitochondrial aldehyde dehydrogenase (ALDH2), cause of the alcohol flush response and present at an allele frequency of 8% in Asian populations (C.-H. Chen, Kraemer, and Mochly-Rosen 2022), is thought to destabilise the holoenzyme (Crabb et al. 1989). Yet, the destabilization does not have to directly affect the catalytic function of the enzyme. Instead, it may happen through a generic mechanism whereby complexes that have incorporated mutant subunits undergo proteasomal degradation at a higher frequency, constitutively depleting the available wild-type subunits.

Assuming that a single mutant subunit can confer DN or DP effects onto the wild-type complex upon co-assembly, the mutant phenotype gets amplified with greater number of subunits in the complex (Bergendahl et al. 2019). Without cotranslational assembly, a homodimer will retain 25% wild-type activity in the cell after random pairing of mutant and wild-type subunits. In case of a homotrimer, the wild-type activity will theoretically decrease to only 12.5%. It is one of the key questions of this thesis whether cotranslational assembly is able to improve these ratios in favour of the wild type, as will be discussed in Chapter 3. Additionally, it is possible that function loss/gain lies on a continuous spectrum as a function of the number of subunits in the complex. This assumption has been recently tested in vivo using single-channel patch-clamp, an electrophysiology technique that allows quantification of the gating properties of individual membrane channels (Geng et al. 2023). The study focused on the p.GLy375Arg substitution, associated with Liang-Wang syndrome, in the calcium-activated potassium channel subunit alpha-1 protein that forms a tetrameric membrane channel. Indeed, hybrid wild-type:mutant complexes had an intermediate effect on potassium conductance, suggesting that mutant subunits incrementally titrate down the wild-type function.

Our understanding of the properties of mutations associated with assembly-mediated effects is still incomplete and largely limited to missense substitutions. Generally, these mutations are expected to have milder structural consequences compared to LOF mutations since they should permit at least partial folding of the protein for successful assembly. This intuition is supported by a study on DN mutations causing Gillespie Syndrome in the inositol 1,4,5-trisphosphate receptor type 1 protein (McEntagart et al. 2016) as well as by a recent structural analysis from our group involving dozens of proteins associated with GOF and DN effects (Gerasimavicius, Livesey, and Marsh 2022). Exceptions to this trend are cases where the mutation affects a domain that folds independently from the domain responsible for forming the intersubunit interface. For example, the DN mutation p.Gly163Cys in the complement C1q tumor necrosis factor-related protein 5, associated with late-onset retinal degeneration, is predicted to be highly destabilizing to the folding of the C1q domain. Yet, because

the protein's assembly into trimers is driven by the collagen-like domain, the mutation may not entirely prevent assembly (Stanton et al. 2017).

Another notable characteristic of non-LOF mechanisms, including assembly-mediated effects, is the tendency of these mutations to exhibit spatial clustering within the protein structure. Initially reported for cancer genes, where mutations in tumour suppressor genes are often associated with LOF and those in oncogenes frequently lead to over-activation (GOF) (Stehr et al. 2011), this spatial clustering has also been observed beyond cancer. An analysis from our group has revealed that spatial clustering is an emergent property of non-LOF mechanisms (Gerasimavicius, Livesey, and Marsh 2022), and it has proven valuable for identifying de novo mutations with alternative dominant molecular mechanisms (Lelieveld et al. 2017). Understanding the properties of non-LOF, assembly-mediated or otherwise, mechanisms at both the gene- and the variant-level is key for advancing in this field.

1.3.3 Contextualizing the human genetic variation

The DNA of all people alive today contain almost every possible 9.3 billion single nucleotide variants (SNVs). Additionally, there are also numerous insertions and deletions, but small structural variants of 1-10,000 bases only account for about a fifth of the genetic variation (Mullaney et al. 2010). It is estimated that unrelated individuals differ from each other by as much as 20 million bases, with 3.6 million attributed to SNVs (Auton et al. 2015). Although much of this variation has been passed down over many generations, each person carries around 70 mutations not present in their parent's germline DNA, known as de novo mutations (Sasani et al. 2019). Based on cDNA sequences from the CCDS database that correspond to UniProt canonical sequences, 75 million missense mutations are reachable by SNVs. Notably, the genome aggregation database (gnomAD), an international consortium aiming to harmonize human genetic sequencing data, catalogues about 5.6 million missense SNVs (Karczewski et al. 2020), representing roughly 7% of the total. Nevertheless, as technological advancements continue and population sequencing studies expand globally, the fraction of known tolerated missense mutations in healthy humans is expected to increase.

The majority of the possible SNVs are believed to be benign, ascribed to the selection to minimize deleterious effects of mutations during the early evolution of the universal genetic code (Haig and Hurst 1991). Some missense SNVs will remain unobserved, i.e. “incompatible with life” or “embryonic lethal”. These terms are commonly used in the literature to describe mutations that impair the function of cell-essential genes or those indispensable for instituting the developmental programme during the course of gestation. In many cases, mutations may not cause miscarriage but instead result in severe developmental phenotypes at an early age (McRae et al. 2017). Rarely, developmental errors arise when both parents are carriers of the variant, leading to a recessive disorder (H. C. Martin et al. 2018). However, approximately half the time, these errors are caused by de novo mutations, occurring primarily in dominant genes (Kaplanis et al. 2020). On average, of the 70 de novo SNVs per generation, up to 2 will be located in protein-coding genes, resulting in developmental disorders with a prevalence of approximately 1 in 300 births (McRae et al. 2017). Despite tremendous efforts over the last decade in utilizing trio sequencing data to establish a causal link between de novo missense variants and the associated phenotypes, 50% of patients remain undiagnosed (C. F. Wright et al. 2023). Because LOF mutations are easier to identify by their generally more damaging effect on protein structure (Gerasimavicius, Livesey, and Marsh 2022), there could be many cryptic de novo variants with alternative dominant molecular mechanisms. An important hypothesis to test in the future is the extent to which de novo mutations in biallelic

genes cause disease via assembly-mediated effects. Particularly, DN mutations are known to match the severity of the recessive phenotype (McEntagart et al. 2016), thus raising the question whether a fraction of variants is being missed on the account of the inheritance pattern typically associated with the gene.

Genetic diseases beyond developmental disorders can lead to life-long illness, have a predisposing effect to certain conditions such as cancer, or exhibit their phenotype later in life. By extensively cataloguing mutations causally linked to diseases, it may be possible in the future to better predict interactions between them or utilize machine learning to anticipate the effects of yet-unobserved variants. A valuable resource for this purpose is ClinVar (Landrum et al. 2018), a database that aggregates information about genomic variation and its relation to human health. As of July 2023, ClinVar records about 50,000 missense variants that are "pathogenic" or "likely pathogenic" according to the American College of Medical Genetics and Genomics guidelines (Richards et al. 2015). With advances in laboratory and computational methods, the number of pathogenic missense mutations, just like the fraction of known tolerated missense mutations, is expected to increase. However, a pressing concern arises from "variants of uncertain significance" (VUS), which are variants identified in a patient's DNA, but whose clinical significance is unknown. Currently, missense VUS outnumber clinically relevant variants by 17-fold. Although a significant portion of these is likely to be benign, the ones that are not highlight the disparity between the efficiency with which we discover new variants and the ability of our methods to identify potentially pathogenic ones.

Another source of genetic variation linked to disease are somatic mutations. Normal aging induces various changes that increase susceptibility to cancer and create a tissue microenvironment that fosters the growth of malignant cells (De Magalhães and Pedro 2013). Somatic cells tend to accumulate mutations 4 to 25 times faster than germline cells (Lynch 2010). For example, the intestinal epithelium has around 10^6 independent stem cells, each of which produces transient daughter cells every one to two weeks. As a result, the intestinal lining of a 60-year-old is predicted to contain more than 10^9 independent mutations, implying that nearly every genomic site in this single organ is likely to mutate in at least one cell by the age of 60 (Lynch 2010). Depending on the spectrum of mutations and their sequence of occurrence in a cell lineage, cancer initiation will occur at different times and the resultant cells will have different metabolic and proliferative potential. The question arises: can we predict the likelihood of cancer emerging in a tissue based on the frequency and fitness of somatic mutations? A recent study quantified the fitness of somatic mutations in haematopoietic stem cells from older individuals over a 12-year period (Robertson et al. 2022). Some mutations were found to confer different fitness advantages depending on the gene and the variant and the effects could be used to predict the future growth of clones as well as the time required for clinical monitoring. Understanding these dynamics could offer new insights into cancer prevention and personalized treatment strategies.

The challenges posed by genetic diseases arising from de novo, inherited, or somatic mutations necessitate the establishment of a new coalition dedicated to collecting, standardising, and distributing the functional effect of mutations. This need has been realised through the Atlas of Variant Effects (AVE) Alliance, whose primary objective is to comprehensively understand the genome at the nucleotide level (Fowler et al. 2023). To achieve this goal, the alliance integrates data from independent laboratories conducting multiplexed assays of variant effect (MAVE) experiments. MAVEs comprise various methods, such as deep mutational scanning (DMS) experiments on proteins and massively parallel reporter assays on gene regulatory sequences. DMS, a relatively new fusion of deep sequencing and saturation mutagenesis, offers quantitative measurements of variant fitness, potentially providing values for all possible variants in a protein (Fowler and Fields 2014). Notably,

DMS technology has witnessed rapid advancements over the past decade, leading to several studies accurately recapitulating the effects of clinically validated variants (Findlay et al. 2018; Mighell, Evans-Dutson, and O'Roak 2018).

While DMS is costly and laborious to perform for the entire proteome, computational variant effect predictors (VEPs) are showing promise as an accurate and scalable alternative (Livesey and Marsh 2023). By harnessing an increasing number of experimentally measured variant functional scores and enhancing the specificity of VEPs to identify pathogenic variants, a powerful synergy between the two methodologies is likely to emerge in the near future. DMS will be strategically employed for challenging genes and variants of clinical significance, while VEPs will continue to refine and complement by imputing variant scores that might have been missed by the experimental method. In this endeavour, gene-level predictors of alternative dominant molecular disease mechanisms, such as those introduced in this thesis, will be potentially very valuable, as they will allow us to identify DN or GOF variants with greater specificity, which can in turn help understand the properties of those variants better. This symbiotic approach holds the potential to contextualise the disease relevance of the human genetic variation and fuel a new era of biological discoveries and medical breakthroughs.

2 Large protein complex interfaces have evolved to promote cotranslational assembly

2.1 Introduction

The majority of proteins across all domains of life function as part of multimeric complexes. Although we have a comprehensive understanding of the diverse quaternary structure space occupied by complexes (Sebastian E. Ahnert et al. 2015), much less is known about where, when, and how their component subunits assemble. Continuing advances in cryo-electron microscopy, mass photometry, and genetic interaction mapping (Braberg et al. 2020) are facilitating a transition towards a structural view of proteomes (Levy and Vogel 2021). While, at the present, our structural analyses are primarily limited to structures of complexes in the Protein Data Bank (Berman et al. 2000), there is a new generation of multiscale protein complex modelling approaches (Evans et al. 2021; Gao et al. 2022; Humphreys et al. 2021) promising to fill the gap and accelerate structure-based discovery. Our understanding of proteomes has also been dramatically improved by the development of ribosome profiling, which has provided us with quantitative measurements at the level of translation. Alterations of the technique revealed the cotranslational action of chaperones (Oh et al. 2011; Shiber et al. 2018; Stein, Kriel, and Frydman 2019), shed light on the role of collided ribosomes in proteostasis (Arpat et al. 2020; Han et al. 2020; T. Zhao et al. 2021), and supported the view of the ribosome as a signalling hub (D'Orazio and Green 2021). To the present work, however, it is of outstanding relevance that ribosome profiling has laid down strong evidence that the assembly of protein complexes often starts on the ribosome (Bertolini et al. 2021; Kamenova et al. 2019; Natan et al. 2017; Panasenko et al. 2019; Shiber et al. 2018).

Two factors appear to be particularly important for cotranslational assembly: the proximity of nascent chains on adjacent (cis) or between juxtaposed (trans) ribosomes, and the localisation of interface residues towards the N-terminus of a protein, which allows more time for an interaction to occur during translation (Kamenova et al. 2019; Natan et al. 2018; Shieh et al. 2015). Recent findings demonstrated that homomers, formed from multiple copies of a single type of polypeptide chain, frequently assemble on the same transcript via the interaction of adjacent elongating ribosomes (Bertolini et al. 2021). This mechanism is highly effective because it takes advantage of the fact that homomeric subunits are identical; therefore nascent chains are essentially colocalised by the nature of their synthesis. Although homomers may benefit from polysome-driven assembly, it requires allocation of cellular resources to ensure at least two ribosomes are actively translating the same mRNA at any one time (Yansheng Liu, Beyer, and Aebersold 2016). On the other hand, heteromers, products of different genes that physically interact, can only employ the trans assembly mode in eukaryotes, providing mechanisms that colocalise their transcripts exist (X. Chen and Mayr 2022; F. Liu et al. 2016; Pizzinga et al. 2019; G. Wang et al. 2020). In contrast to homomers, cotranslational assembly of heteromers may only require a single ribosome on each mRNA, which could allow lowly abundant regulatory proteins to cotranslationally assemble (Biever et al. 2020; Heyer and Moore 2016). Alternate ribosome usage and translation-coupled assembly can explain how cells achieve efficient construction of complexes with uneven stoichiometry, accounting for a substantial fraction of heteromeric complexes (Marsh et al. 2015).

Despite growing evidence supporting the importance of cotranslational assembly, far less is known about the properties of the interfaces involved. It has been observed that cotranslationally binding subunits have a

tendency to fall out of solution or become degraded by orphan subunit surveillance mechanisms in the absence of their partner subunits (Choe et al. 2016; Juszkiewicz and Hegde 2018; Kamenova et al. 2019; Natan et al. 2018; Shiber et al. 2018). This observation may be explained under two assumptions: N-terminal interfaces are aggregation prone due to interference with cotranslational folding (Ciryam et al. 2013; Jacobs and Shakhnovich 2017; G. Kramer, Shiber, and Bukau 2019; Kudva et al. 2018), and/or that cotranslationally forming interfaces possess unique structural properties that predispose them to aggregation in the absence of binding partners. Whilst there is evidence for the former (Natan et al. 2018), interfaces involved in nascent chain assembly have not been systematically studied before. Therefore, we cannot exclude the possibility that they have structural features that make them more susceptible to a cotranslational route.

Hydrophobic surfaces play a key role in nucleation theory (Chandler 2005; Hermann 1972; Tanford 1978) and protein folding (Chothia 1975; Gething and Sambrook 1992; Privalov and Khechinashvili 1974), but more importantly, hydrophobicity remains the founding principle of protein-protein recognition theory (Chothia and Janin 1975; Kauzmann 1959). Defined as the buried surface area between subunits, interface size shows correspondence to hydrophobic area because larger interfaces contain more interface core residues (Levy 2010). Conveniently, interface area is relatively simple to compute from structural data (Hubbard SJ 1993; Kleinjung and Fraternali 2005; Mitternacht 2016; Winn et al. 2011). Whilst the relationship between interface area and measured affinity is non-linear (Brooijmans, Sharp, and Kuntz 2002; Eisenberg and Mclachlan 1986; Horton and Lewis 1992; Vangone and Bonvin 2015), interface area shows remarkable correspondence with subunit dissociation energy, and is reflective of the evolutionary history of subunits within complexes (Levy et al. 2008).

We hypothesised that cotranslational interactions may be distinguished from others based upon the areas of the interfaces involved. The size hierarchy of interfaces in protein complexes can be used to predict the order in which their subunits assemble, in good agreement with experimental data (Levy et al. 2008; Marsh et al. 2013; Wells, Bergendahl, and Marsh 2016). According to this theory, the largest interfaces in a complex correspond to the earliest forming subcomplexes within the assembly pathway, irrespective of the binding mode. While specific contacts that increase affinity would introduce compositional biases into the sequence space, exerting undue selection pressure on proteomes, variability in interface size can emerge from nonadaptive processes as the organising principle of cotranslational assembly (S. E. Ahnert et al. 2010; Conant 2009; Gray et al. 2010; Hochberg et al. 2020; Leonard and Ahnert 2019; Lynch 2013).

In the present study, we address this idea by analysing experimental data on cotranslationally assembling human proteins (Bertolini et al. 2021). Our results establish a strong correspondence between cotranslational assembly and subunit interface size. To test whether large interfaces could represent an evolutionary adaptation to cotranslational assembly, we took advantage of the many protein complex subunits that have more than one interface. We compared the areas of first and last translated interfaces in bacterial, yeast, and human heteromeric subunits and found a clear tendency for the first interface to be larger across all species. This finding suggests that large protein complex interfaces have evolved to promote cotranslational assembly.

2.2 Results

2.2.1 Cotranslationally assembling subunits are characterized by large interfaces

In a recent study, a novel ribosome profiling method was used to identify over 4000 cotranslationally assembling human proteins (Bertolini et al. 2021). By design, the method can identify subunits that undergo cotranslational assembly when both subunits are in the process of translation. As recently proposed (Kamenova et al. 2019), we refer to this mode of binding as "simultaneous" assembly (**Figure 2.1A/B**).

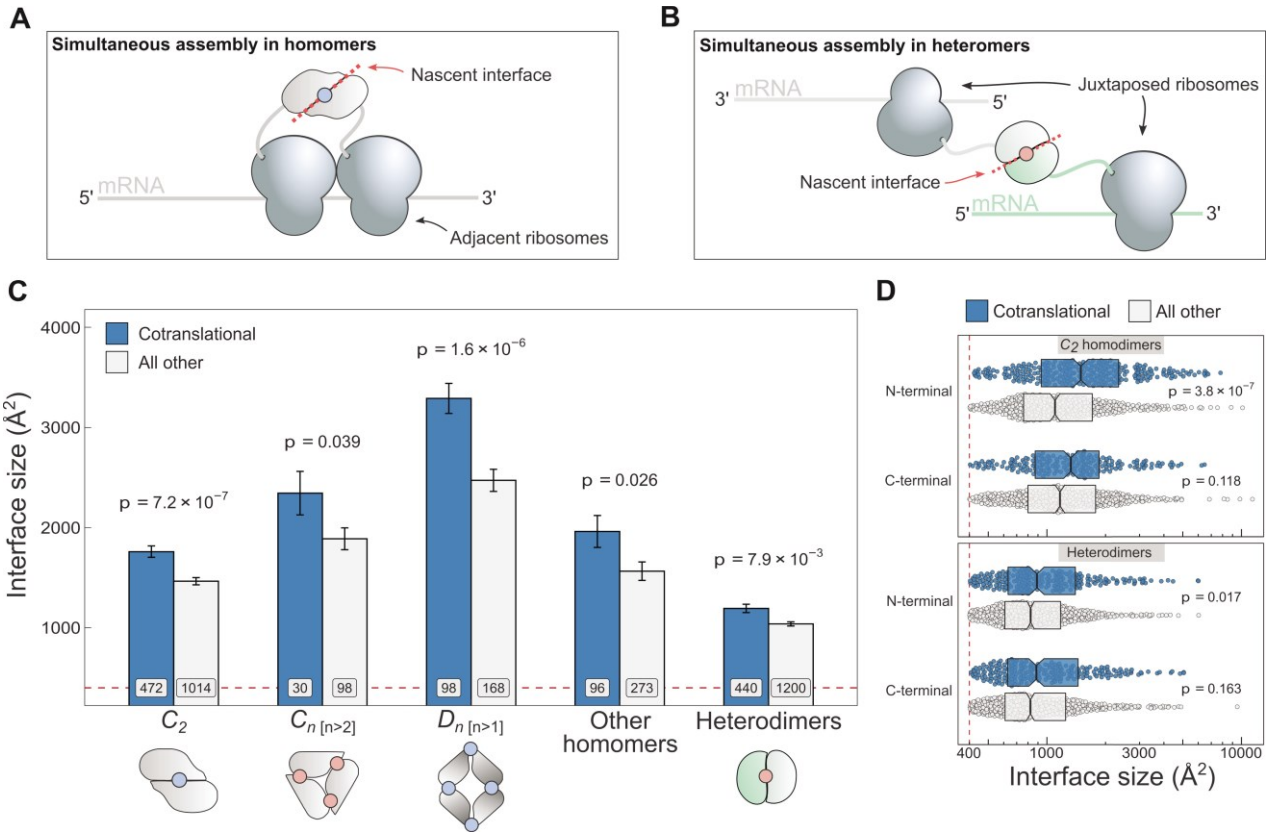


Figure 2.1 Cotranslationally assembling subunits are characterised by large interfaces.

(**A**) Schematic representation of (cis) simultaneous cotranslational assembly in homomers. (**B**) Schematic representation of (trans) simultaneous cotranslational assembly in heteromers. (**C**) Interface size differences between cotranslationally assembling and all other subunits of homomeric symmetry groups and heterodimers. Error bars represent standard error of the mean (SEM) and labels on bars show the number of proteins in each group. The p-values were calculated with two-sided Wilcoxon rank-sum tests. Pictograms show the basic structure of symmetry group members, with the blue dots representing isologous and red dots representing heterologous and heteromeric interfaces. (**D**) Interface size distributions of cotranslationally assembling and all other subunits of C_2 homodimers and heterodimers, subset by the terminal location of the interface. The p-values were calculated with two-sided Wilcoxon rank-sum tests.

To investigate if interface area correlates with simultaneous assembly, we computed the buried surface areas of homomeric and heterodimeric subunits and subset the results by whether or not the protein was detected to cotranslationally assemble (**Figure 2.1C**). The arrangement of homomeric subunits with respect to one or more rotational axes allows their classification into symmetry groups. The three most common groups are the twofold symmetric (Schönflies notation, C_2), cyclic ($C_{n>2}$), and dihedral ($D_{n>1}$) complexes, which all have distinct structural and functional characteristics (Bergendahl and Marsh 2017; Goodsell and Olson 2000; Levy

and Teichmann 2013) and should therefore be considered separately. For example, members of the cyclic and dihedral symmetry tend to have larger buried surfaces, because they interface with more than one subunit, while C_2 symmetric homodimers tend to have larger interfaces than heterodimers (Jones and Thornton 1996).

C_2 homodimers represent the most highly populated symmetry group and their single isologous interface (i.e. symmetric or head-to-head) makes the analysis simple to perform. In line with our expectation, C2 symmetric subunits that assemble during translation expose 20% larger areas than those that do not (**Figure 2.1C**; $p = 7.2\times10^{-7}$, Wilcoxon rank-sum test). Considering that the cotranslational assembly annotations are derived from a laboratory technique that uses extensive biochemical fractionation, it is important consider the possibility that larger interfaces would be more persistent to these procedures. We therefore controlled for the potential confounding effect of larger interfaces by setting incrementally higher interface area cutoffs (**Appendix 2.1A**), to which the trend appears robust. Note that we have not tested the effects of similar interface cutoffs for the other symmetry groups, due to their much smaller dataset size.

Higher-order cyclic complexes are centred on a rotational axis so that every subunit has two distinct interfaces, each with an adjacent protomer. Both interfaces are heterologous (i.e. asymmetric or head-to-tail) and approximately the same size. Cyclic symmetry is potentially confounded by its tendency to form ring-like structures, which are ubiquitous components of biological membranes (Forrest 2015). As a result, membrane-bound complexes are enriched in non-polar amino acids that form the interface with the alkane core of the lipid bilayer. We focused on the analysis of cyclic homomers that do not localise to the plasma membrane, owing to competing hydrophobic forces exerted by protein-lipid interactions. Despite the limited number of structures available, we detect a significant difference in interface area among soluble members of the cyclic symmetry group (**Figure 2.1C**), with the mean of cotranslationally forming subunits being 24% larger ($p = 0.039$, Wilcoxon rank-sum test). Notably, we did not observe a trend in plasma membrane-localised cyclic complexes (**Appendix 2.1B**).

Dihedral symmetry can be thought of as the stacking of a dimeric or cyclic complex through the acquisition of a twofold axis. All dihedral complexes have isologous interfaces, and those with at least six subunits can have both isologous and heterologous interfaces (e.g., D_3 dimers of cyclic trimers). We find that cotranslationally assembling dihedral complexes have on average 33% larger interfaces than those assumed to assemble after their complete synthesis (**Figure 2.1C**; $p = 1.6\times10^{-6}$, Wilcoxon rank-sum test). A dihedral complex is likely to have evolved from a C_2 homodimer if its largest interface is isologous and, conversely, when its heterologous interface is largest, the complex probably arose via a cyclic intermediate (Levy et al. 2008; Marsh and Teichmann 2014b). When dihedral complexes are grouped by their evolutionary history, the trend is present in both groups (**Appendix 2.1B**), consistent with that observed in C_2 homodimers and cyclic complexes.

We pooled all remaining homomers, including those with helical and cubic symmetry, and those that are asymmetric, into a single “other” category, due to their relatively low representation in the human proteome. Altogether, cotranslationally assembling subunits in this heterogeneous category present 25% larger interface areas than other members (**Figure 2.1C**; $p = 0.026$, Wilcoxon rank-sum test). Thus, the interface size trend in cotranslationally assembling complexes appears to hold up across all types of homomers.

Because in heteromers simultaneous assembly requires two different transcripts positioned in trans (**Figure 2.1B**), we were curious if they, too, showed a correspondence between cotranslational assembly and interface size. Due to the diverse quaternary structures and assembly pathways associated with heteromeric complexes

(Sebastian E. Ahnert et al. 2015), we focused on the simplest cases, the heterodimers, which form a single heterologous interface via the interaction of two different proteins. When compared, heterodimers that simultaneously assemble reveal a 15% larger interface area on average than those not detected to cotranslationally assemble (**Figure 2.1C**; $p = 7.9\times10^{-3}$, Wilcoxon rank-sum test). Similar to C_2 homodimers, the trend in heterodimers is also robust to incremental interface area cutoffs (**Appendix 2.1A**), making it unlikely to be an artefact.

The weaker effect size in heterodimers relative to C_2 homodimers may be explained by the combination of two factors. First, previous experimental evidence suggests that heteromers commonly employ the "sequential" mode of assembly, whereby a subunit in the process of translation recruits a fully synthesised and folded subunit (Kamenova et al. 2019; Shiber et al. 2018). This mode of assembly has not yet been experimentally probed on a proteome-wide scale, and it is possible that many heterodimers lacking cotranslational assembly annotations in our dataset employ sequential assembly. As a result, assuming that interface size plays a role in sequential assembly as well, these unannotated proteins weaken the effect we can detect. Second, it is plausible that another biological process, yet uncharacterised in detail, is responsible for the colocalisation of transcripts and the subsequent subunit assembly (X. Chen and Mayr 2022; G. Wang et al. 2020), which could make assembly in heteromers less reliant on interface area.

We considered three potentially confounding variables of the ribosome profiling method, which was used to detect cotranslationally assembling proteins. First is protein length, in part because long polypeptide chains take more time to translate, making it more likely for a cotranslational interaction to come about, and partly because bigger proteins tend to form larger interfaces. Long proteins are also encoded by long transcripts on which structures called di-ribosomes (two ribosomes connected by interacting nascent chains) may persist for a longer time period, potentially leading to their survivorship bias to the observer. In **Appendix 2.1C**, we present an analysis where both C_2 homodimers and heterodimers are binned by their length into bins containing equal number of structures and subset by cotranslational assembly. With the exception of long heterodimers, all bins follow the expected interface size trend. More importantly, for both types of complexes, the middle bin, which contains approximately 350-720 residue long proteins and thus covers a large fraction of the human proteome, shows the strongest effect size.

The second variable we accounted for is the confidence-based classification of the cotranslational assembly data set. Bertolini et al., 2021 employed an elaborate strategy to assign high or low confidence to the protein candidates (details in (Bertolini et al. 2021)). However, these high confidence proteins only make up a fifth of the data, which prohibits their exclusive use in our analyses. To address this, we leveraged homology models from the SWISS-MODEL repository (Bienert et al. 2017) to increase the number of available structures for analysis. With this supplemented structural data set, we found the difference between high confidence and all other subunits (excluding low confidence) to be statistically significant at symmetry level for all homomers (**Appendix 2.1D**), but not for heterodimers, consistent with the weaker effect size for heterodimers observed earlier. There are no significant differences observed between high and low confidence proteins for any of the groups. Although we might expect that high confidence proteins should, on average, have larger interfaces than low confidence proteins (assuming that a greater fraction of them represent true cases of cotranslational assembly), we note that the size of the high confidence set is relatively small, especially when split by symmetry group, and that the average interface size of the high confidence proteins is larger for all homomers, except in the very small dihedral group. Overall, we think that the small differences between high and low confidence

sets, as well as the small size of the high confidence set, justify our use of the combined sets throughout this study.

The third potential confounder is the location of the interface relative to protein termini. Interactions via N-terminal interfaces are translated first, therefore increasing the time available for cotranslational assembly. Given that cotranslationally forming interfaces identified by ribosome profiling are known to be significantly enriched towards the N terminus of proteins (Bertolini et al. 2021), but that overall, homomeric interfaces tend to be enriched towards the C terminus (Natan et al. 2018), we wished to control for interface location. We classified all interfaces as occurring on either the N- or C-terminal halves of proteins, based on the position of the interface midpoint, which is the residue at which half of the buried surface area of an interface is reached. This comparison is presented for C_2 homodimers and heterodimers in **Figure 2.1D**. In all groups, there is a clear interface size trend wherein cotranslationally assembling subunits have a larger interface area. More interestingly, however, the trend is only significant and much larger in effect between N-terminally localised interfaces. In fact, N-terminal interfaces are significantly larger than C-terminal interfaces in cotranslationally assembling homodimers ($p = 0.021$, Wilcoxon rank-sum test). One possible explanation for this is that N-terminally localised interfaces are far more likely to represent cases of genuine cotranslational assembly.

2.2.2 Interface area is more important than other interfacial contact-based properties for explaining cotranslational assembly

To rule out that interface size is masking a more important property of cotranslationally assembling subunits, we explored other interface features using the same set of C_2 symmetric homodimers (n = 1,486) and heterodimers (n = 1,640) as shown in **Figure 2.1C**, which are abundant in the structural data and possess only a single interface, making the results simple to interpret. Because hydrophobicity is essential to protein-protein interactions (Chothia and Janin 1975), we calculated the apolar interface area and compared it to the total area (**Figure 2.2A/B**). We found that, while the difference in apolar interface area shows a stronger effect among homodimers than the total size (Wilcoxon effect size 0.24, $p = 1.8\times10^{-7}$, *vs* 0.228, $p = 7.2\times10^{-7}$), the opposite is observed among heterodimers (Wilcoxon effect size 0.054, $p = 0.262$, *vs* 0.127, $p = 7.9\times10^{-3}$). The origin of this sharp contrast is likely the fact that heteromeric interfaces are less hydrophobic (Jones and Thornton 1996), and thus complexation is less likely to be primarily driven by the size of the hydrophobic patch of the interface. We also looked at the absolute number of residue-residue contacts within a 5.5 Å radius (Vangone and Bonvin 2015), which echoes the results we obtained with total interface size, but with weaker effects (**Figure 2.2C**; Wilcoxon effect size 0.211 for C_2, $p = 4.4\times10^{-6}$, and 0.083 for heterodimers, $p = 0.08$). Next, we employed a contact-based model to estimate binding affinity from the number and character of residue-residue contacts (Vangone and Bonvin 2015). As expected, this analysis revealed that cotranslationally assembling subunits have higher predicted affinities, or lower ΔG of binding (**Figure 2.2D**), among both homo- and heterodimers (Wilcoxon effect size 0.166 for C_2, $p = 3.1\times10^{-4}$, and 0.122 for heterodimers, $p = 0.011$), although these differences are also weaker than those observed with interface size.

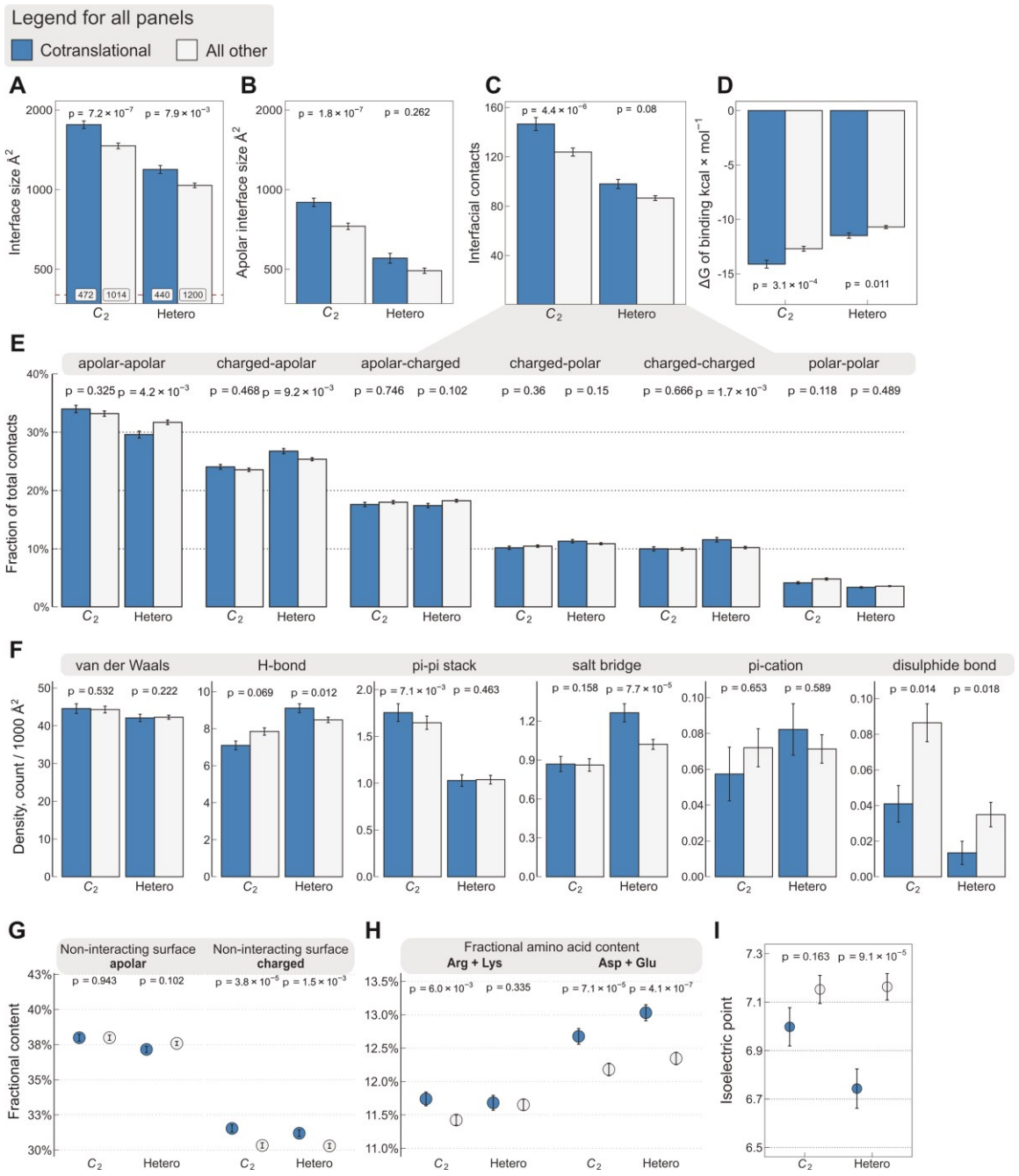


Figure 2.2 Interface area is more important than other interfacial contact-based properties for explaining cotranslational assembly.

All panels show the sample mean ± standard error of the mean (SEM) for cotranslationally assembling and all other subunits of human C_2 symmetric homodimers and heterodimers. The p-values were derived from two-sided Wilcoxon rank-sum tests in panels (**A–D**) and (**I**), and from two-sided Dunn's test of multiple comparisons in panels (**E–H**). Labels on bars in panel (**A**) represent sample sizes. The following parameters are shown. (**A**) Total interface size (Å²). (**B**) Apolar interface size (Å²). (**C**) The absolute number of interfacial contacts. (**D**) Predicted Gibbs free energy (ΔG) of binding (kcal/mol). (**E**) Fraction of residue-residue contacts by chemical character, in descending order of prevalence. (**F**) Specific interaction density (count/1000 Å²), in descending order of prevalence. (**G**) Non-interacting surface apolar (NIS_a) and charged (NIS_c) residue per cent. (**H**) Fractional content of positively (Arg+Lys) and negatively (Asp+Glu) charged amino acids in the full-length sequence. (**I**) Protein isoelectric point determined with continuum electrostatics on the full monomeric structures.

We further dissected interfacial contacts based upon chemical character (**Figure 2.2E**) and interaction type of the residues (**Figure 2.2F**). The differences within these categories are again weaker than with interface size, the only exception being the contribution of salt bridges to the cotranslational assembly of heterodimers (Dunn's test effect size 0.173, $p = 7.7\times10^{-5}$), which is similarly reflected in the differences between the fraction of charged-apolar and charged-charged contacts. By contrast, the only non-negligible contribution to binding in the cotranslational assembly of homodimers is that of pi–pi interactions (Dunn's test effect size 0.115, $p = 7.1\times10^{-3}$). While these observations are interesting, they are broadly explained by the apolar interface areas presented in **Figure 2.2B**. On the one hand, this trend suggests that homodimeric interfaces are made up of more hydrophobic residues, which in turn leave less space for other types of residues, for example charged amino acids that may form salt bridges. This compositional bias in isologous interfaces can, by exclusion, naturally increase the frequency of pi–pi stacked aromatic side chains, a phenomenon that impacts the mutational and evolutionary landscape of symmetric homomers (Goodsell and Olson 2000; Jacque Monod, Wyman, and Changeux 1965; Ponstingl et al. 2005). On the other hand, heterodimers, whose interfaces are less hydrophobic, accommodate more polar and charged residues that may give rise to specific interactions, in accord with the higher frequency of hydrogen bonds and salt bridges. This notion is consistent with the weak correlation between interface size and binding affinity in heteromers (Brooijmans, Sharp, and Kuntz 2002), because in heteromeric protein-protein recognition, interface complementarity is likely to play a more important role. Nevertheless, as demonstrated here, interface size shows tremendous utility in discriminating cotranslationally assembling subunits because it is easily calculable and it is a fundamental property of protein-protein interactions, unlike salt bridges or pi-pi interactions, which do not necessarily occur at every interface.

A linear model developed for the estimation of binding affinity from residue-residue contacts at the interface (Vangone and Bonvin 2015) incorporates coefficients for the terms "non-interfacial surface apolar/charged" (NIS_a and NIS_c, respectively), which are the fraction of apolar or charged surface residues of a subunit in complex, and they have been shown to influence binding affinity (Kastritis et al. 2014). We found significant differences in the NIS_c parameter between cotranslationally assembling and all other subunits among both homo- and heterodimers (**Figure 2.2G**), suggesting that cotranslationally assembling subunits possess a larger proportion of charged residues on the surface than other subunits. To investigate this further, we calculated the fractional content of positively (Arg+Lys) and negatively (Asp+Glu) charged amino acids from protein sequences to see if one charge group in particular is responsible for the trend. This analysis is possible because other than the relatively rare internal charges (Hendsch and Tidor 1994; Kajander et al. 2000), most charges tend to be on the surface, an assumption that is supported by our structural data, in which nearly 83% of charged residues have more than 25% relative accessible surface area in the monomeric protein (Levy 2010). We found that negatively charged amino acids in cotranslationally assembling subunits are overrepresented relative to other subunits (**Figure 2.2H**; Dunn's test of multiple comparisons, C_2, $p = 7.1\times10^{-5}$ and heterodimers, $p = 4.1\times10^{-7}$), and we also detect a small but significant enrichment in positively charged amino acids, but only in homodimers ($p = 6.0\times10^{-3}$). We further support these observations with isoelectric points calculated in continuum electrostatics at physiological pH and salt concentration from the human AlphaFold structures (Tunyasuvunakool et al. 2021), which represent the full-length monomers translated on ribosomes. This analysis revealed that the net enrichment in charged amino acids on the surfaces of cotranslationally assembling heterodimeric subunits results in a lower isoelectric point than in other subunits (**Figure 2.2I**; Wilcoxon rank-sum test, $p = 9.1\times10^{-5}$).

What may explain the finding that cotranslationally assembling subunits display more negative surface charges than other proteins? We believe there are four mutually non-exclusive hypotheses that are compatible with the observation. The first is based on the work of Kastritis et al., 2014, who proposed based on alanine scanning mutagenesis experiments that polar and charged residues of non-interacting surfaces contribute to binding affinity. Second, one might argue that the role of charged residues on the surface is to counteract the strong water-orientation forces exerted at large interfaces by supporting protein solubility through favourable interactions with water molecules (R. M. Kramer et al. 2012) and ions (Linse et al. 1988). The third idea concerns the ribosome:nascent chain interaction, where negative charges could help avoid unproductive interactions with the ribosome surface (Cassaignau et al. 2021; Deckert et al. 2021), thus facilitating cotranslational folding and assembly. The fourth scenario would be attributable to a proteome-wide effect, whereby the higher the abundance of a protein, the more its surface has been shaped by evolution for optimal "stickiness" to combat non-specific interactions upon molecular crowding (Levy, De, and Teichmann 2012). While further analysis of this effect is out of the scope of this study, using pooled homo- and heterodimers, we detect a weak but significant Spearman correlation of 0.18 ($p = 1.7\times10^{-23}$) between the NIS_c parameter and HEK293-specific active ribosome count (Clamer et al., 2018), which corroborates the fourth hypothesis.

2.2.3 Larger and earlier-assembling interfaces tend to form cotranslationally in heteromeric subunits with multiple interfaces

Having confirmed that subunit interface size correlates with cotranslational assembly, we next wanted to see if this trend applies within single subunits that have more than one interface. In other words, do multi-interface heteromeric subunits also employ their largest interface during the course of simultaneous assembly? A multi-interface heteromeric subunit forms at least two distinct interfaces with two other proteins in a complex that contains at least three subunits. Because of the interface hierarchy that exists within protein complexes (Levy et al. 2008; Marsh et al. 2013), we hypothesised that the largest interface, which is most likely to assemble earliest, should also be more likely to cotranslationally assemble.

To perform an analysis at the multi-interface level, we made use of the assembly-onset positions determined for every protein in the cotranslational assembly dataset (Bertolini et al. 2021). An assembly-onset is a single residue in the protein sequence, whose codon is being decoded by the ribosome at the time of cotranslational assembly. In order to identify which interface the assembly-onsets belongs to, we mapped them to the closest interface midpoint in the linear protein sequence, as illustrated in **Figure 2.3A**. This is to avoid biases from large interfaces, which have many more interface residues and therefore a higher probability that an assembly-onset would map to them if the interface was not compressed into a single midpoint residue.

Using this method, we identified 281 interfaces of multi-interface heteromeric subunits that may form in a simultaneous cotranslational fashion. To see if these correspond to the largest interfaces within each protein, we calculated the mean interface area for all other interfaces on these subunits to be able to perform a paired statistical test. Our results show that the identified cotranslationally assembling interfaces are indeed larger by 19% than other interfaces on these subunits (**Figure 2.3B**; $p = 0.018$, Wilcoxon signed-rank test).

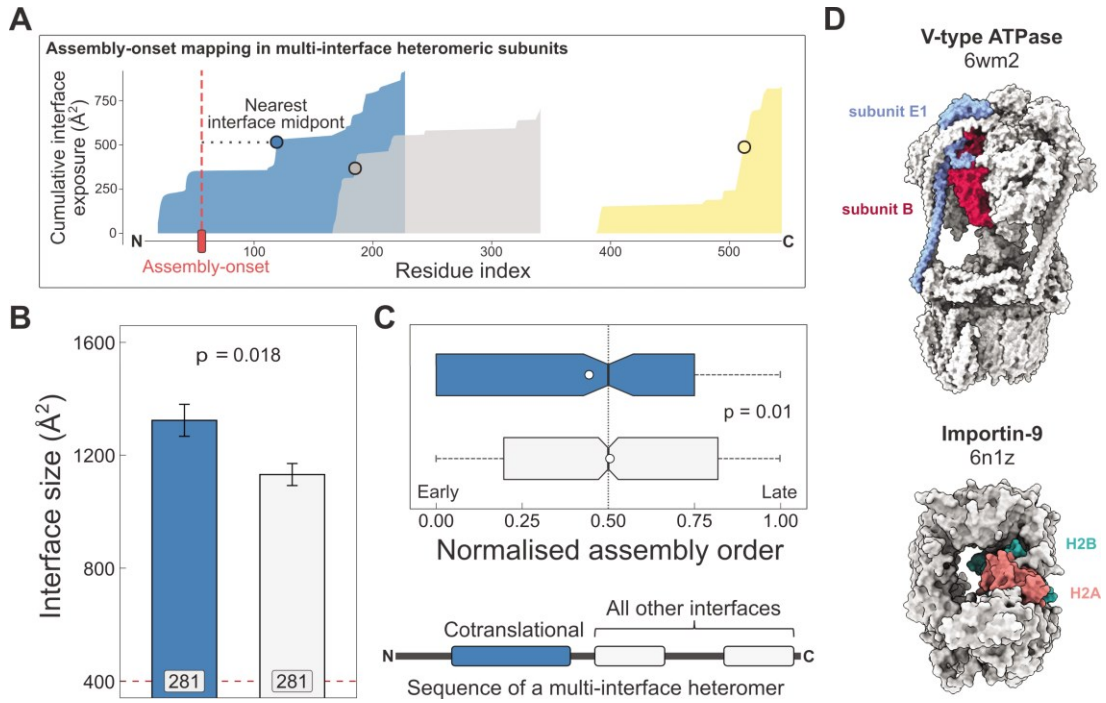


Figure 2.3 Larger and earlier-assembling interfaces tend to form cotranslationally in heteromeric subunits with multiple interfaces.

(**A**) Visual representation of the interface mapping protocol. The area plot shows the cumulative interface area build-up of individual interfaces during translation, which are shown in different colours, with dots representing their midpoints. Assembly-onsets determined by Bertolini et al. are mapped to the nearest midpoint on condition that it is not a homomeric interface. (**B**) Pairwise comparison of cotranslationally forming (in the simultaneous mode) interfaces of multi-interface heteromeric subunits to the mean of all other heteromeric interfaces on them. For visual aid, see line diagram under panel (**C**). Error bars represent standard error of the mean (SEM) and labels on bars show the number of proteins in each group. The p-value was calculated with the Wilcoxon signed-rank test. (**C**) Pairwise comparison of the normalised assembly order in 201 complexes between cotranslationally forming and all other heteromeric interfaces. The p-value was calculated with the Wilcoxon signed-rank test. (**D**) Two examples of simultaneous cotranslational assembly between subunit pairs in heteromeric complexes: the subunits E and B1 of the V-type ATPase (pdb: 6wm2), and importin-9 with histone H2A (6n1z).

We wished to put these interfaces into the context of their full complexes. Do simultaneously forming interfaces represent early forming subcomplexes that then initiate further assembly events, since the first step of a protein complex assembly pathway is the most likely to occur cotranslationally (Wells, Bergendahl, and Marsh 2015)? Although the largest interface in a complex is always predicted to assemble earliest in the assembly pathway, subsequent steps are non-trivial because they can involve multiple subunit:subunit interfaces (Sebastian E. Ahnert et al. 2015; Levy et al. 2008; Marsh et al. 2013). To answer this question, we predicted the assembly steps of the complexes on the basis of their structures (Wells et al., 2016). This analysis revealed that the identified interfaces tend to form much earlier than other heteromeric interfaces in the complexes (**Figure 2.3C**; p = 0.01, Wilcoxon signed-rank test). Another interpretation of this can be given by classifying assembly steps into “early” and “late”, depending on their normalised assembly order (McShane et al. 2016), which is a 0-to-1 scale indicating the first-to-last steps of a pathway, where we defined early steps with values less than or equal to 0.5. According to this, a simultaneously forming interface is 1.7 times more likely to form early (180 [67%] of 270 *vs* 633 [54%] of 1171; $p = 9.5\times10^{-5}$, Fisher’s exact test).

Some of the identified interfaces belong to complexes that have been shown to use cotranslational assembly routes, such as the proteasome (Panasenko et al. 2019) and subunits of the transcription initiation complex (Kamenova et al. 2019). However, many are not yet described in the literature, for example, the loading of histone H2A onto importin-9 (**Figure 2.3D**), which has been reported to act as a storage chaperone while transporting a histone dimer to the nucleus (Padavannil et al. 2019). More recently, cotranslational binding of importins to cargo has been suggested to play a role in protecting nascent nuclear proteins from degradation (Seidel et al. 2023). Another example is the V-type ATPase (**Figure 2.3D**), whose catalytic A and B subunits have been tested for their ability to assemble in the sequential mode with a negative result (Shiber et al., 2018), but our structural approach using the assembly-onset identified the E1 subunit to form in the simultaneous mode with the catalytic B subunit. Although these two subunits can undergo major structural rearrangements in the complex, the B subunits stay in contact with the same E1 subunits across all the observed conformational states (Vasanthakumar et al. 2022). In fact, such large post-translational conformational rearrangements may be common in cotranslationally forming interfaces, given that proteins with larger interfaces will have an inherent tendency to be more flexible (Marsh and Teichmann 2014b)

2.2.4 Evolutionarily ancient subunits are more likely to undergo cotranslational assembly

Protein complexes are under evolutionary selection to minimise misassembly (Leonard and Ahnert 2019; Marsh et al. 2013), meaning that over evolutionary timescales, ordered subunit assembly has been prioritised in cells (Wells, Bergendahl, and Marsh 2016). These findings have led to the formulation of the interface size hypothesis, which posits that the assembly pathway of a protein complex parallels with its evolutionary history. A simple proxy for predicting the steps of such pathways is interface size, which demonstrates exceptional correspondence with in vitro complex assembly-disassembly analyses (Levy et al. 2008; Marsh et al. 2013).

Naturally related to the interface size hypothesis is another trend that reflects the time it takes for a newly emerged interface to become strengthened by evolution, either because of a functional association between the proteins, or via entrenchment in a neutral ratchet (Dayhoff et al. 2010; Hochberg et al. 2020; W. K. Kim et al. 2006). To address this, we grouped heteromeric protein complex subunits based on the phylogenetic class in which a protein's encoding gene first appeared, that is protein age (Liebeskind, McWhite, and Marcotte 2016). In addition to the human proteins, we also employed a dataset of yeast complexes. Due to the smaller size of the yeast dataset, we supplemented it with heteromeric models computed for yeast core complexes, which were determined using residue coevolution inferred from paired multiple sequence alignments, followed by deep learning-based structure prediction of subunit pairs with experimental evidence to interact (Humphreys et al. 2021). In **Figure 2.4A**, we show that the average interface size of the groups, for both human and yeast, is ordered almost perfectly according to evolutionary age of both homo- and heteromers. When we ask what percentage of each protein age group was found to assemble cotranslationally by Bertolini et al. there is a remarkable agreement between older proteins with larger interfaces having a higher frequency of cotranslational assembly (**Figure 2.4B**). Older subunits are also more likely to be found in more than one complex than younger subunits (Drew, Wallingford, and Marcotte 2021; Saeed and Deane 2006); hence cotranslationally assembling proteins are expected to be enriched in moonlighters. Indeed, it has recently been shown among components of the nuclear pore complex that some subunits preferentially assemble cotranslationally with one partner, but not with a different partner from another complex (Seidel et al. 2023).

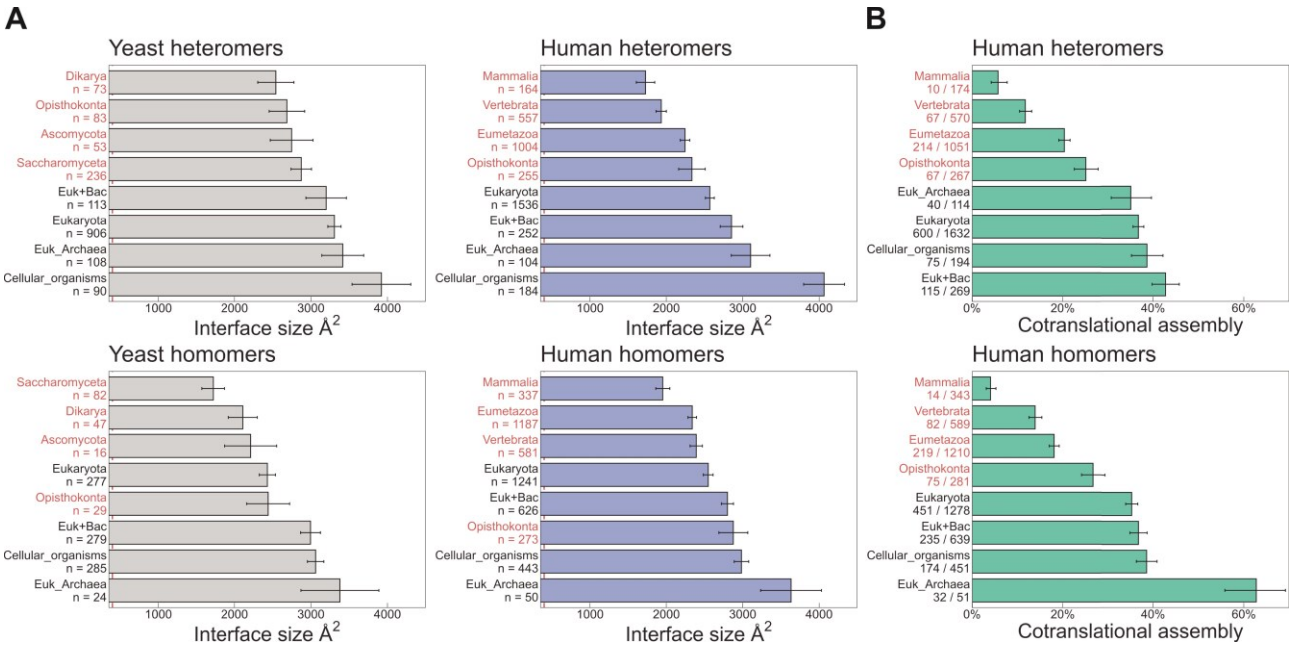


Figure 2.4 Evolutionarily more ancient subunits of complexes are more likely to undergo cotranslational assembly.

(**A**) Average (mean ± SEM) interface sizes of yeast and human homo- and heteromeric subunits grouped by the evolutionary age of the protein. Age group labels coloured in red are defined as "more recent" proteins, while those in black represent "ancient" proteins. Numbers under labels represent the number of distinct proteins in the given age group. Homomeric interface sizes are a pool of experimentally determined structures and SWISS-MODEL homology models. (**B**) The frequency (%) of cotranslational assembly, as detected by Bertolini et al. in the different protein age groups, split into homo- and heteromers. Heteromer annotations were supplemented with those contained in hu.MAP2.0 (Drew et al., 2021). Fractions under labels denote the number of cotranslationally assembling proteins out of the total in the given age group. Error bars are 68% Jeffrey's binomial credible intervals.

2.2.5 N-terminal interfaces tend to be larger than C-terminal interfaces supporting evolutionary selection for cotranslational assembly

There are two possible explanations for the observation that cotranslationally forming interfaces tend to be larger. First, larger interfaces may be inherently more likely to form cotranslationally because their assembly is more energetically favourable. In this scenario, cotranslational assembly has not been evolutionarily selected for; instead, the larger interfaces are simply more likely to be formed while the protein is still in the process of being translated, without providing any functional benefit. Alternatively, cotranslational assembly may have been selected for, for example, because it increases the efficiency of assembly and avoids potentially damaging non-specific interactions. Here, large interfaces have evolved to increase the level of functionally beneficial cotranslational assembly.

One way to distinguish between these two scenarios is to compare the sizes of N- and C-terminal interfaces. Regardless of whether cotranslational assembly occurs simultaneously (**Figure 2.1A/B**) or sequentially (**Figure 2.5A**), due to vectorial synthesis on the ribosome, N-terminal regions of proteins are more likely to be involved in binding events during translation. Therefore, if cotranslational assembly is adaptive, we would expect that N-terminal interfaces in multi-interface heteromeric subunits, which will be translated first, should show a significant tendency to be larger than C-terminal interfaces, as illustrated in **Figure 2.5B**.

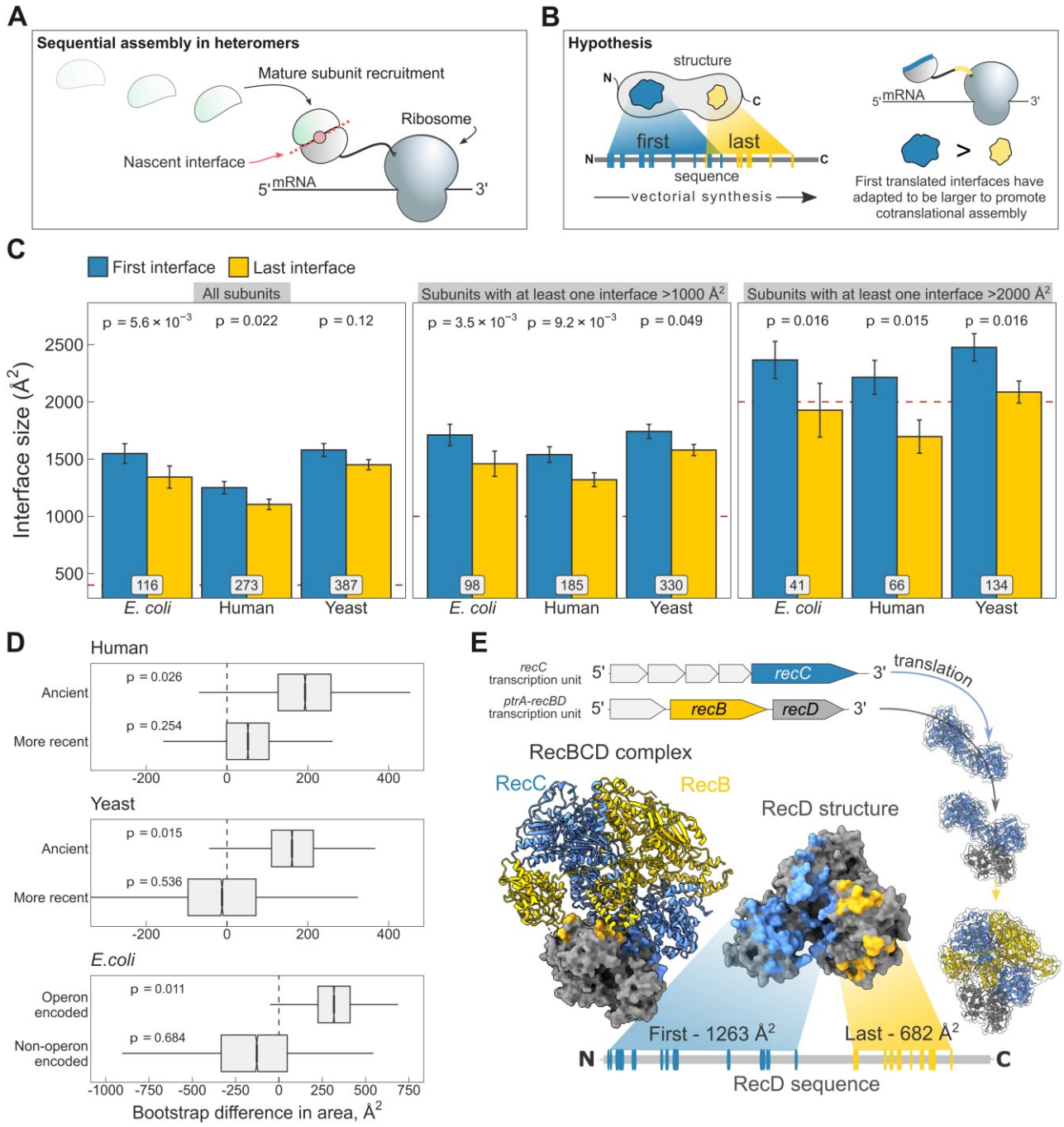


Figure 2.5 N-terminal interfaces tend to be larger than C-terminal interfaces supporting evolutionary selection for cotranslational assembly.

(**A**) Schematic representation of sequential cotranslational assembly in homomers. (**B**) Diagrammatic representation of the hypothesis test of the adaptive model of cotranslational assembly. (**C**) Area differences between the first and last translated interfaces in multi-interface heteromeric subunits across the species. Panels are ordered by the area cutoffs, 400, 1,000, and 2,000 Å², which are satisfied if either the first or the last interface is larger than the given cutoff. Error bars represent standard error of the mean (SEM) and labels on bars show the number of proteins in each group. The p-values were calculated with Wilcoxon signed-rank tests. (**D**) Bootstrap distributions of the area difference between the first and the last translated interfaces. The p-values were calculated from 10^4 bootstrap resamples with correction for finite testing. (**E**) Example of an operon-encoded complex, the RecBCD nuclease (pdb: 5ld2).

To address the question, we selected experimentally determined heteromeric complex structures from three model proteomes: *Escherichia coli*, *Saccharomyces cerevisiae* (yeast), and Homo sapiens (human). For each heteromeric subunit, we defined the first interface as the one that exposes the most N-terminal interface residue in the linear protein sequence, and, to treat the termini symmetrically, the last interface was defined as the one that exposes the last interface residue, that is the first interface residue from the C-terminal direction.

When we compare the areas of the first and last translated interfaces in heteromeric subunits across species, we find the first interface to be larger (**Figure 2.5C**). The strongest effect is measured in *E. coli*, where the first interface is larger than the last interface in 60% of cases, and on average is 205 Å^2 (15%) larger ($p = 5.6\times10^{-3}$, Wilcoxon signed-rank test). In humans, the trend is weaker, with the first interface larger in 52% of cases, being 147 Å^2 (13%) larger on average ($p = 0.022$). In yeast, although the first interfaces are 130 Å^2 (9%) larger on average, the difference is not significant. Averaging over the three species, we observe first interface to be larger than the last interface in 54% of multi-interface subunits.

The above analysis includes all multi-interface heteromeric subunits. However, some of these contain only very small interfaces, and would therefore be unexpected to undergo cotranslational assembly via either interface, thus reducing the observed tendency for N-terminal interfaces to be larger. Indeed, when we filter the subunits to exclude those with only small interfaces, including only those where the size of at least one of the two interfaces is larger than 1,000 or 2,000 Å2, we find that the trend gets much larger across all three species (**Figure 2.5C,** full distribution in **Appendix 2.2A**). When using the 2,000 Å2 filter, we observe that N-terminal interfaces are on average 23%, 31%, and 19% larger for *E. coli*, human, and yeast, respectively, significant for all. Averaging over the three species, we find the first interface to be larger than the last interface in 64% of cases. Thus, when considering multi-interface heteromers where at least one interface is >2,000 Å^2, the N-terminal interface is larger than the C-terminal interface nearly two-thirds of the time.

Building on our earlier observation that cotranslationally assembling proteins tend to be older in evolutionary terms (**Figure 2.4B**), we asked if a protein's age could also influence the size of the N-terminal interface, assuming that over time selection would work to minimise misassembly by favouring larger first translated interfaces and ensuring cotranslational assembly between the correct subunit pair. We split the multi-interface subunits from human and yeast into "ancient" and "more recent" categories. We then generated bootstrap distributions of the area difference between the first and the last translated interfaces in yeast and human multi-interface heteromeric subunits within the respective age categories. The analysis revealed that the difference between the areas of the first and last translated interfaces is significantly larger in ancient proteins in both yeast and human (**Figure 2.5D**; mean difference of 163 Å^2, $p = 0.015$ and 192 Å^2, $p = 0.026$, respectively), but not in younger heteromers. We speculate that this is because there has not been enough time for selection to considerably act on the assembly of these subunits, or because newer subunits tend to be required less for cotranslational interactions due their later assembly or functional roles they play. Thus, these analyses support that large, typically ancient, interfaces have evolved to promote cotranslational assembly.

Our attention was next drawn to *E. coli*, which demonstrated the strongest first versus last interface size trend. In prokaryotes, many heteromeric complexes are encoded by operons, where the different subunits are translated off of the same polycistronic mRNA molecules. Early studies in bacteria indicated that operon gene order is correlated to physical interactions between the encoded proteins (Dandekar et al. 1998; Mushegian

and Koonin 1996). Further investigation laid down theoretical, mechanistic, and evolutionary evidence in support of this (Shieh et al. 2015; Sneppen et al. 2010; Wells, Bergendahl, and Marsh 2016). Cotranslational assembly is likely to be particularly common in operon-encoded heteromers, given that the translation of different subunits is inherently colocalised. We hypothesised that the tendency for N-terminal interfaces to be larger should be stronger in operon-encoded *E. coli* heteromers, compared to those that are not.

We illustrate the example of the RecBCD nuclease in **Figure 2.5E**. Genes of the subunits are located in adjacent loci encoding transcriptional units for RecC and RecB/D. One study reported that purification of RecD is complicated by the formation of inclusion bodies, while the other two subunits remain in the soluble fraction (Masterson et al. 1992). Moreover, a genetic analysis suggested that partially folded RecC and RecD might interact during translation, or that RecC forms a complex with RecB first, onto which RecD is then assembled (Amundsen, Taylor, and Smith 2002). The regulatory subunit RecD has two interfaces well separated in the sequence, where the interface with RecC is translated first. One might imagine that the nascent chain of RecD forms a complex with mature subunit of RecC, having double the interface area to accommodate RecC than that for RecB. In this scenario, the assembly efficiency is not only maximised by gene order reducing stochasticity, but also by cotranslational assembly minimising the need for post-translational association.

To further test whether large interfaces could have been selected to promote cotranslational assembly, we acquired annotations derived from RNA sequencing datasets (Chetal and Janga 2015) to group heteromers from *E. coli* according to whether or not they are encoded by operons. We generated bootstrap distributions of the area difference between the first and the last translated interfaces to visualise and derive a probability (**Figure 2.5D**). In agreement with the above idea, we found that the size difference between the two interfaces is significantly larger in operon-encoded multi-interface heteromeric subunits, favouring the first interface (mean area difference of 317 $Å^2$, $p = 0.011$).

We speculate that the first versus last interface trend may be the hallmark of sequential cotranslational assembly (**Figure 2.5A**) rather than that of the simultaneous mode (**Figure 2.1B**). Bertolini et al. have suggested that simultaneous assembly is predominantly employed for the formation of homomeric protein complexes, which could mean sequential assembly is the more common cotranslational assembly mode in heteromers. The strong trend in *E. coli* also supports this idea, because polycistronic gene structure is more compatible with sequential cotranslational assembly (Shieh et al. 2015). In eukaryotes, large complexes and subunits of lowly abundant complexes may require an additional biological process to ensure their transcripts are colocalised for simultaneous assembly and to facilitate further assembly steps (X. Chen and Mayr 2022). Sequential assembly, on the other hand, may have evolved to exploit large interface areas for the recruitment of partner subunits. This can be conceptualised as the "bait and prey strategy" of cotranslational assembly, in which a nascent interface represents the "bait" bound by a fully folded "prey" subunit. Although a proteome-scale dataset of cotranslationally assembling proteins is not available for yeast, we have identified case studies of five multi-interface heteromers in yeast that use the sequential assembly mode, and found that all five subunits follow the first versus last interface size trend (**Table 2.1**). To substantiate the model further, we removed from the human dataset those proteins that were identified by Bertolini et al. to simultaneously assemble. Strikingly, removal of these proteins increases the size difference between the first and the last translated interfaces from 13% to 18% (**Appendix 2.2B**; $p = 0.014$, Wilcoxon signed-rank test). One explanation consistent with the model is that the remaining heteromers are enriched in sequential assembly.

Study	Nascent chain	Mature partner	First (Å²)	Last (Å²)	Source
Halbach et al., 2009	Set1	uncertain	1,467	1,025	6bx3
Shiber et al., 2018; Fischer et al., 2020	Fas2	Fas1	4,226	484	6ql9
Shiber et al., 2018	Pfka1	Pfk2	7,796	5,749	3o80
Shiber et al., 2018	Gcn3	Gcd2	1,210	687	Humphreys et al., 2021
Panasenko et al., 2019	Rpt2	Rpt1	4,226	484	6fvt

Table 2.1 Table of yeast multi-interface heteromeric subunits, which have been shown to utilise the sequential mode of cotranslational assembly.

The yeast Set1, part of the COMPASS complex, binds multiple partners during its translation process (Halbach et al., 2009), but the order of these assembly steps is uncertain. In a partial structure of the histone methyltransferase complex, Set1 is found to have two biologically significant interfaces, the first with Swd3 and the last with Swd1. For the rest of the cases, the mature partners are known and the available structural data support a model where the first translated interface is larger relative to the last translated interface, likely to promote cotranslational subunit recruitment.

A property that would be consistent with the above model is interface separation. The later the translation of the last interface starts relative to the first, the higher the chance that assembly of the first interface will be undisturbed, free of competition with the partner subunit of the last interface. Therefore, we hypothesised that the distance between translation start points of the first and last interfaces, which are the earliest emerged interface residues of each, should correlate with the size difference in favour of the first interface. Because of large variances in protein length and interface size, we normalised the translational distance between the first and last interfaces as the percentage of the protein's sequence length, and scaled the area difference by the sum of both interfaces.

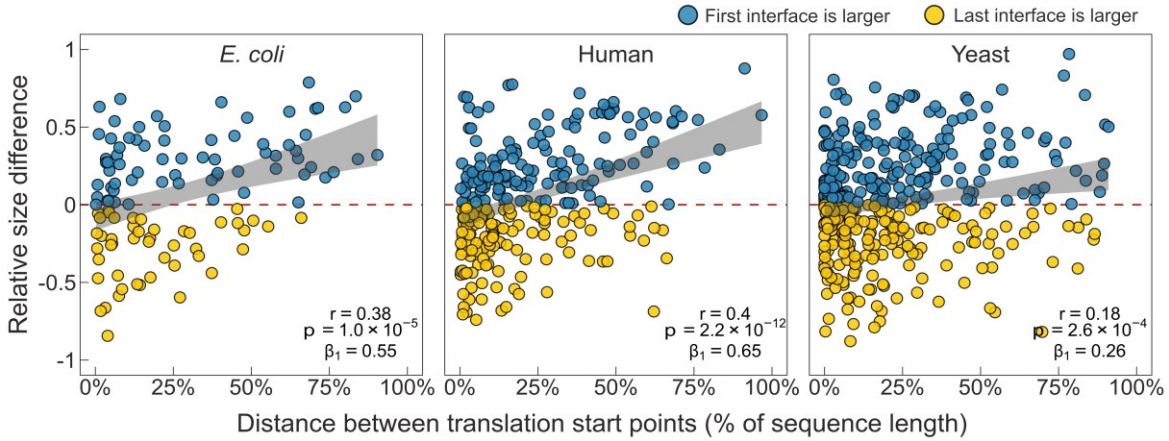


Figure 2.6 Separation between the first and last translated interfaces.

Correlation between the relative distance of translational start points and the relative area difference of the first and last translated interfaces. Shaded lines represent the 95% confidence interval of the regression line. The Pearson's correlation coefficient r, its p-value, and the regression coefficient β_1 are shown in the panels.

Figure 2.6 shows the correlation between the separation of translation start points and the area differences of the first and the last interfaces (absolute values in **Appendix 2.2C**). As expected, increasing the distance between the interfaces monotonically increases the extent of the area difference across all species. To rule out that the interface separation metric is confounded by sequence length, we split the structures into less and more than 400 amino acids, which is the pan-species mean of sequence lengths. In both subsets, there is a pronounced preference for a larger first interface when the separation is high (**Appendix 2.2D**). The causal

direction of this effect, whether it reflects that cotranslational assembly happens more often in high degrees of interface separation, or that separation is driven by selection for cotranslational assembly, remains to be addressed.

2.3 Discussion

It has long been understood that interface area is important for assembly, but the capacity in which it shapes the hierarchy of individual interfaces on subunits remained elusive. In this study, we first combined information from ribosome profiling with structural data on protein complexes to probe the importance of interface area in the process of cotranslational assembly. Our results demonstrate a strong correspondence between interface size and cotranslational subunit binding. Inspired by this, we set out to test an important question about the biological significance of cotranslational assembly: do large interfaces give rise to cotranslational assembly because of simple energetic reasons or do they reflect an evolutionary adaptation for a functional benefit? We found a clear trend across three evolutionarily distant species for the first translated interface of heteromeric subunits to be larger, suggesting that large interfaces have evolved to promote cotranslational assembly.

While our results support the adaptive hypothesis of interface size, it is not entirely inconceivable that cotranslational assembly represents a ratchet-like mechanism of constructive neutral evolution (Gray et al. 2010; Hochberg et al. 2020), whereby a drift in interface properties creates ideal conditions for assembly on the ribosome. Reversion to the post-translational route is prevented by the accumulation of mutations that are neutral in the subunit entrenched in cotranslational assembly, but would otherwise be deleterious in the ancestor. A similar neutral process may sustain the differences in interface area presented in this study.

In light of current knowledge of bacterial operon structure and translation regulation, it is not surprising that we observed the strongest trend among *E. coli* heteromers with respect to the size of the interface that first emerges from the ribosome. Supposedly, the effect is attributable to the widespread sequential assembly between mature subunits and nascent chains, reflecting the mechanism of effective subunit recruitment by large N-terminal “bait” interfaces. An interesting question for laboratory experiments is whether operon-encoded heteromers can assemble in the simultaneous mode, providing that the structural organisation of the bacterial polysome allows for such a precise coordination (Brandt et al. 2009).

How does a large interface area help translating ribosomes find one another? Its benefit in homomers for facilitating cotranslational assembly is clear, because the subunits are localised to the same mRNA, and large interfaces are more likely to form interactions before translation is complete. In heteromers, one hypothesis argues the involvement of RNA-binding proteins that orchestrate transcript colocalisation (X. Chen and Mayr 2022; Keene 2007), from where similar rules may apply to heteromer assembly as for homomers. A more parsimonious hypothesis is formed on the observation that cotranslational assembly can result in transcript colocalisation, which is ablated when subunit affinity is decreased (Heidenreich et al. 2020). This may suggest that affinity, which correlates strongly though imperfectly with interface size (Brooijmans, Sharp, and Kuntz 2002; Vangone and Bonvin 2015), can play a role in the colocalisation of transcripts belonging to the same complex.

Many more topics of inquiry remain open for future studies. Analogous to protein folding, cotranslational assembly can be thought of as a hydrophobic collapse that shapes the quaternary structure of the complex. As

with folding in the cell, the involvement of other factors must not be overlooked. A wide array of ribosome-associated chaperones are vital for nascent chain homeostasis (Döring et al. 2017; Koldewey, Horowitz, and Bardwell 2017; G. Kramer, Shiber, and Bukau 2019) and the degree to which they choreograph assembly steps is yet to be elucidated.

Attention should be paid to the far-reaching genetic consequences of cotranslational assembly (Natan et al. 2017). How much does allele-specific assembly buffer the dominant-negative effect, and what does it mean in the context of human genetic disease (Bergendahl et al. 2019)? This effect requires mutant subunits to be stable enough to assemble into complexes, and thus the impact of the mutations tends to be milder at the structural level than of other pathogenic mutations (McEntagart et al. 2016), making them exceptionally difficult to detect using the existing variant effect predictors (Gerasimavicius, Livesey, and Marsh 2022). Interestingly, cotranslational assembly should actually make the dominant-negative effect less common in homomers, because it can limit the mixing that occurs between wild-type and mutant subunits. It remains to be seen whether dominant-negative mutations are in fact less common in cotranslationally assembling complexes.

Finally, our results build on evidence from the past decade and emphasise the importance of protein complex assembly at the translatome level. Although evolutionary selection against N-terminal interface contacts to avoid premature assembly was previously found in homomers (Natan et al. 2018), here we report an opposite phenomenon in which proteins that do cotranslationally assemble sustain large N-terminal interfaces in order to promote cotranslational subunit recruitment. We expect our observation to be supported by experiments once the proteome-wide detection of sequentially assembling heteromers is made possible.

2.4 Methods

Protein structural datasets

Starting from the entire set of structures in the Protein Data Bank (PDB, (Berman et al. 2000)) on 2021-02-18, we searched for all polypeptide chains longer than 50 residues with greater than 90% sequence identity to *H. sapiens*, *S. cerevisiae*, and *E. coli* canonical protein sequences. When proteins mapped to multiple chains, we selected a single chain sorting by sequence identity, then by the number of unique subunits in the complex, and then by the number of atoms present in the chain. Only biological assemblies were used and symmetry assignments were taken directly from the PDB. Polypeptides formed by cleavage were excluded. In the generation of the multi-interface heteromeric subunit datasets, to exclude proteins with yet uncharacterised interfaces, chains with an at least 70% complete structure were considered and only included if they formed interface pairs >800 Å^2 with at least two different subunits. To supplement the smaller yeast dataset, computed structures of yeast core complexes were downloaded from the ModelArchive link provided by (Humphreys et al. 2021). For downstream analysis, mmCIF files were converted into standard Brookhaven PDB format and the chains were mapped to genes using the table provided on ModelArchive. Homology models of yeast and human homomeric complexes were obtained from the SWISS-MODEL repository (version 2022_02) (Bienert et al., 2017; Waterhouse et al., 2018). When a protein's UniProt accession number mapped to multiple homology models, we selected a single model ranking by the number of subunits in the complex, followed by the length of the modelled chain. Symmetry groups of the homology models were assigned with the software AnAnaS (Pagès and Grudinin 2018; Pagès, Kinzina, and Grudinin 2018). In the analysis of N- versus C-terminal interface sizes, we excluded very large heteromeric complexes, defined as those containing ≥10 subunits. This is because of the previous evidence that predicting assembly order based on

interface size in very large complexes is not as accurate (Sebastian E. Ahnert et al. 2015; Marsh et al. 2013), likely because of the many intersubunit interfaces these complexes possess.

Calculation of interface area-related properties

Interface areas of SWISS-MODEL homology models were calculated with FreeSASA (Mitternacht 2016) using the default surface probe radius of 1.4 Å. Residue-level pairwise interfaces in complexes derived from the PDB and from (Humphreys et al. 2021) were calculated between all pairs of subunits using AREAIMOL from the CCP4 suite (Winn et al. 2011) with a probe radius of 1.4 Å. The interface was defined as the difference between the solvent accessible surface area of each subunit in isolation and within the context of the full complex. Apolar interface area was calculated from the residue-level data by classifying A, F, G, I, L, V, M, P, and Y amino acids as apolar (Vangone and Bonvin 2015). An interface area cutoff of >400 Å^2 was used for homomeric subunits and multi-interface heteromeric subunits derived from the PDB to exclude potential crystallographic interfaces and restrict the analysis to biologically significant interfaces. Assembly order was computed by predicting the assembly pathway assuming additivity of pairwise interfaces in each complex (Marsh et al. 2013), and implemented with the *assembly-prediction* Perl package (Wells, Bergendahl, and Marsh 2016).

The relative interface location was calculated according to the formula:

$$Relative\ interface\ location = (i - 1) / (L - 1)$$

where i marks the residue at which half of the cumulative buried surface area of the interface is passed (*i.e.* interface midpoint), and L is the sequence length.

The normalised distance between translational start points of two interfaces was calculated as:

$$Relative\ translational\ distance = (f_{\mathbf{last}} - f_{\mathbf{first}}) / L$$

where f marks the first residue of the given interface and L is the sequence length.

Area differences between the first and the last interface were normalised according to the equation:

$$Relative\ size\ difference = (BSA_{\mathbf{first}} - BSA_{\mathbf{last}}) / (BSA_{\mathbf{first}} + BSA_{\mathbf{last}})$$

where BSA is the buried surface area of the corresponding interface.

Calculation of interfacial contact-related properties

To make sure all software runs without errors, we converted the human structure files of homo- and heterodimers obtained by our pipeline from the Protein Data Bank to standard Brookhaven PDB format by taking the first atom locations in case of multiple AltLoc entries, converting non-canonical amino acids into equivalent standard names, renaming chain pairs with identical chain identifiers, and stripping files to only contain ATOM, TER and END lines, thus excluding heteroatoms. Interfacial residue contacts were determined with the software PRODIGY (Vangone and Bonvin 2015; Xue et al. 2016), using default settings. The software RING 3.0 (Clementel et al. 2022) was used to compute residue interactions between the subunits. The network policy was set to “closest” and all interactions were returned for a contact using the flag `--all_edges`.

Determining isoelectric point from structure

Predicted structures of human proteins were acquired from the AlphaFold database (Tunyasuvunakool et al. 2021) in PDB format, which were converted into PQR files with PDB2PQR (Dolinsky et al. 2004) using the

PARSE force field. The PQR files were piped into the software BLUUES (Walsh et al. 2012), and the .ddg output was kept, containing the pH ~ charge titration data. Then, each protein's isoelectric point (pI) was calculated by interpolating from the pH ~ charge curve for charge = 0, using the `approx()` function in R. For proteins that are longer than 2700 amino acids we took the mean pI across the fragments.

Protein localisation

We obtained annotations for plasma membrane, cytoplasmic, and nuclear localisations directly from the UniProt FTP site (Bateman et al. 2021). Canonical UniProt entries with the gene ontology terms plasma membrane (GO:0005886), cytoplasm (GO:0005737), and nucleus (GO:0005634) were considered.

HEK293 active ribosome count

Normalised ribosome protected fragments of actively translating ribosomes specific to Human Embryonal Kidney 293 lineage were determined by (Clamer et al. 2018); the data is available at the NCBI Gene Expression Omnibus under the accession GSE112353. We used averages from two biological replicates.

Protein age

The ages of proteins were obtained from the work of (Liebeskind, McWhite, and Marcotte 2016), at the link https://github.com/marcottelab/Gene-Ages/tree/master/Main. The main_HUMAN and the main_YEAST comma separated value files were parsed and the modeage column was used in our analyses. We combined the mode age into two categories, where "ancient" proteins are those whose genes are common to all cellular life ("Cellular_organisms"), whose genes were transferred horizontally from bacteria after eukaryotes diverged from archaea ("Euk+Bac"), and whose genes emerged in the clades of eukaryotes and archaea ("Eukaryota" and "Euk_Archaea"). The other four age groups, ranging from genes emerged in the classes Opisthokonta to Mammalia in the human proteome, and from Ascomycota to Saccharomyceta in the yeast proteome, were classified as "more recent".

hu.MAP2.0 heteromers

Drew et al. integrated large-scale affinity purification mass spectrometry datasets, biochemical fractionation data, proximity labelling and RNA hairpin pulldown data to generate a complex map with >7,000 complexes (Drew, Wallingford, and Marcotte 2021), freely available at http://humap2.proteincomplexes.org/.

Mapping bacterial subunits to operons

Operon annotations were downloaded from OperomeDB (Chetal and Janga 2015). Genes were mapped to UniProt identifiers using *E.coli* proteome-specific mapping from the UniProt FTP site (Bateman et al. 2021).

Molecular graphics

Visualisation of structures was performed with UCSF ChimeraX version 1.1 (Pettersen et al. 2021).

Statistical analysis

Data exploration and statistical analyses were carried out in RStudio (Rstudio 2022) version 2022.02.0+443 "Prairie Trillium" Release, using R version 4.2.1 (R Core Team 2023). The R packages used for analyses are `tidyverse`, `tidytable`, `rsample`, `rstatix`, `scales`, and `ggbeeswarm`. The Wilcoxon rank-sum or signed-rank tests were used for A/B testing of interface size distributions, because although they appear log-normal, they are also left-bounded because of the minimum interface size cutoff, thus a non-parametric test was required. Wilcoxon signed-rank tests were one-tailed, and their main assumption that the data are symmetric around

the median was supported by boxplot distributions. In Dunn's test of multiple comparisons the Holm-Bonferroni method (Holm, Sture 1979) was used to correct for family wise error rate. The effect sizes were defined as the Z-score computed from the *p*-value over the square root of sample size (Tomczak and Tomczak 2014). In the bootstrap analyses, data were stratified for protein age or operonal localisation in 10^4 resamples. The *p*-value was calculated by determining the fraction of point estimates (difference in area) greater than 0, with correction for finite sampling (Buckland, Davison, and Hinkley 1998). In the regression analysis, conditions for statistical inference, including linearity of the relationship between variables, the independence and normality of the residuals, and homoscedasticity were met; validations can be found in the analysis script deposited with all data on OSF at https://osf.io/x5b2n/.

3 Buffering of genetic dominance by allele-specific protein complex assembly

3.1 Introduction

Almost half of the proteins with experimentally determined structures interact with other copies of themselves to form homomeric complexes (Bergendahl and Marsh 2017) and more than one-third of heteromeric complexes with known structures contain sequence identical repeated subunits (Marsh et al. 2015). Considering the human proteome, about one-fifth of proteins have been detected to cotranslationally assemble in a simultaneous fashion (Bertolini et al. 2021), whereby two subunits interact while still being translated on the ribosome. Cotranslationally assembling homomers are thought to predominantly undergo cis-assembly, yielding allele-specific complexes (Bertolini et al. 2021; Gilmore et al. 1996; Nicholls et al. 2002). Repeated subunits in a heteromeric complex are also more likely to have come from the same transcript, especially when their assembly is seeded cotranslationally, reducing the chance of a single complex containing subunits from two alleles. We are beginning to understand the properties that make subunits more likely to undergo assembly on the ribosome. These include N-terminally exposed interface residues, a large interface area, a high alpha helix content and presence of coiled-coil motifs (Bertolini et al. 2021) or domain invasion motifs (Seidel et al. 2022). While many studies have shed light on the functional and evolutionary aspects of cotranslational assembly, our understanding of its allele-specific nature and its impact on genetics is very limited.

Previously, a potential genetic consequence was proposed on theoretical grounds (Natan et al. 2017; Nicholls et al. 2002; Perica et al. 2012). According to the hypothesis, cotranslational assembly should reduce the likelihood of dominant-negative (DN) disease mechanisms. A DN effect occurs when expression of a mutant allele disrupts the activity of the wild-type allele (Herskowitz 1987; Veitia, Caburet, and Birchler 2018), causing disproportionate function loss and thus a dominant mode of inheritance. Observational evidence has long suggested that DN effects are common in homomers (Veitia 2007), likely because incorporation of a mutant subunit into a complex along with wild-type subunits is enough to “poison” it. This assembly-mediated DN effect can lead to a reduction in functional activity exceeding the 50% that would be expected for a simple heterozygous loss-of-function (LOF) mutation. However, cotranslational assembly can result in complexes whose subunits are allele specific, i.e. made up entirely of either wild-type or mutant subunits, potentially reducing the harmful effects of an otherwise DN mutations (**Figure 3.1**). This ability of cotranslational assembly can be considered its "buffering capacity" against DN mutations.

Some gain-of function (GOF) mutations have a molecular mechanism similar to the assembly-mediated DN effect. At the protein-level, the phenotypic effect of GOF mutations is the consequence of the mutant protein functioning differently from the wild type, e.g., through increased protein activity. However, formation of mixed wild-type:mutant complexes can lead to GOF in a similar manner to the DN effect, but instead of the mutant blocking the activity of the wild-type, the GOF is conferred to the whole complex. An example is the L171R mutation in the G protein-activated inward rectifier potassium channel 2, implicated in the Keppen-Lubinsky syndrome, which reduces ion selectivity, thus allowing sodium and calcium to pass the channel (Horvath et al. 2018). This mechanism can be referred to the assembly-mediated dominant-positive effect (Backwell and Marsh 2022), and, just like the DN effect, should be subject to genetic buffering via allele-

specific assembly. However, there are far fewer reports of this phenomenon, so most of this study will be focused on DN effects.

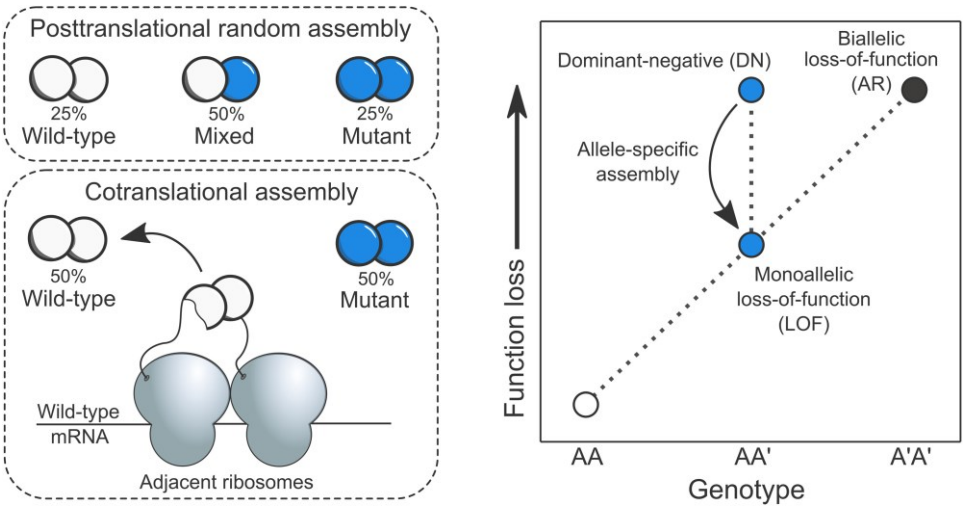


Figure 3.1 Genetic consequence of allele-specific protein complex assembly.

Left: Consider a homodimer with one allele of its gene carrying a heterozygous mutation with DN properties. When complex assembly occurs after the subunits have been fully translated and folded (posttranslational random assembly), the maximum entropy configuration of the subunits dictates that mixed complexes will make up half of all complexes. This means that pure wild-type and mutant complexes form only 25% of the time. However, when the homodimer cotranslationally assembles, both complexes will form independently in an allele-specific manner, increasing the ratio of fully functional complexes to virtually 50%. *Right*: This relationship is illustrated on a phenotype versus function loss landscape diagram. Allele-specific assembly of homomers and repeated subunit heteromers may alleviate the effects of heterozygous LOF mutations by reducing the mixing between the products of wild-type and mutant genes.

The detectability of genetic buffering against assembly-mediated DN and dominant-positive effects may be influenced by two molecular phenomena. First, peri-translational or localised assembly, which occurs when subunits assemble shortly after translation near the parent mRNA (Natan et al. 2017), could also make it more likely that subunits of a complex are allele specific. This effect is likely to be more common in highly abundant proteins, whose transcripts have high ribosome densities and are translated more efficiently than those of lowly abundant proteins (Riba et al. 2019; Schwanhäusser et al. 2011). Second, subunit exchange may increase the entropy of subunit stoichiometry post-assembly via shuffling of wild-type and mutant subunits (Tusk, Delalez, and Berry 2018), resulting in proportions expected from random posttranslational assembly. However, because the likelihood of subunit exchange is determined by the dissociation constants of the subunits involved, it should be less likely to occur in cotranslationally assembling complexes, which tend to have larger interfaces (Badonyi and Marsh 2022) and thus generally higher binding affinities.

Overall, cotranslational assembly may counter the DN effect, which can have an appreciable impact on how genes with inherited and de novo missense variants are prioritised in clinical sequencing pipelines. To address this idea, we used a set of experimentally determined cotranslationally assembling proteins and formulated two hypotheses based upon the above lines of thought. First, genes with an autosomal dominant (AD) disease inheritance pattern should be less likely to assemble cotranslationally compared to autosomal recessive (AR) genes, given that a large fraction of them are likely to be associated with DN effects. Second, protein subunits with known DN disease mutations should have the lowest rate of cotranslational assembly compared to other genes with autosomal dominant inheritance. Here, we show that both hypotheses are upheld. Examination of

the structural properties of complexes associated with DN mutations suggests that their interfaces are exposed relatively late during translation, which should strongly disfavour cotranslational assembly. Using a knowledge-based approach, we trained a regression model to prioritise genes whose mutations are expected to be associated with non-LOF disease mechanisms. We hope that our work will be of interest to clinical geneticists and accelerate the prediction and discovery of variant-level molecular mechanisms.

3.2 Results

3.2.1 AD genes are depleted in cotranslationally assembling subunits

We started with a set of 9,053 human proteins, of which 6,562 (72%) physically interact with copies of themselves to form homomeric complexes. The remaining 2,491 (28%) are repeated subunits of heteromers, meaning that they are present in heteromeric complexes in more than one copy. Both types of proteins have the potential to be associated with assembly-mediated DN or dominant-positive effects, as the mutant and wild-type proteins can co-assemble within the same complex. We obtained genetic inheritance modes from the OMIM database (Amberger et al. 2015) and defined a gene as AD if it had any disease inherited in an AD pattern, which could possibly be caused by assembly-mediated dominant-negative or positive mutations. We defined a gene as AR if it had mutations inherited exclusively in an AR pattern, which are almost certain to be associated only with loss of function.

Our initial hypothesis was that, amongst human disease-associated genes that encode homomers or repeated subunits of heteromers, those known to exhibit AD inheritance would have lower levels of cotranslational assembly compared to those with exclusively AR inheritance. Although using AD inheritance as a proxy for the DN effect is a simplification, assembly-mediated DN effects are believed to play an important role in AD disorders (Bergendahl et al. 2019). Our analysis shows that 24% of AD subunits undergo cotranslational assembly compared to 35.6% with AR inheritance ($p = 2\times10^{-10}$; hypergeometric test). We calculated the odds ratio (OR) to assess the strength of the difference between the groups. Overall, the OR of 0.57 implies that the odds of cotranslational assembly for a randomly selected AD subunit are almost half as that for an AR subunit. In **Figure 3.2A**, we show this analysis grouped by the three main sources of the subunits: homomers with experimentally characterised structures [Protein Data Bank (PDB) homomers], homomers with non-structural evidence (other, which includes SWISS-MODEL homology models and evidence for homo-oligomerisation from different databases) and repeated subunits of heteromers (see **Methods**). The strongest effect was found in PDB homomers (OR = 0.46, $p = 7.5\times10^{-7}$), followed by repeated subunits (OR = 0.6, $p = 4.4\times10^{-4}$) and other homomers (OR = 0.61, $p = 1.5\times10^{-3}$). We speculate that the stronger trend in PDB homomers is due to their enrichment in biologically important interfaces, making a higher fraction of this group compatible with the buffering of assembly-mediated effects. Nonetheless, the results show that the trend is consistent across all groups, despite slight variations in effect size.

Protein abundance can bias both cotranslational assembly and its detection. Ribosome density tends to be higher in abundant proteins (Riba et al. 2019), resulting in more ribosome footprints for sequencing. Proteins that are more abundant might exhibit a higher level of peri-translational assembly, which could also affect the degree of allele-specific complex formation. Therefore, it is important to control for abundance in our analysis, in case the enrichment of highly abundant proteins in the cotranslationally assembling group is affecting our results. We examined the median abundance of AD and AR homomers and repeated subunits, finding no

significant difference between them (**Appendix 3.1A**, p = 0.658; Wilcoxon rank-sum test). We then divided subunits into quartiles based on their approximate intracellular concentration, ranging from 0.005 nanomolar to 180 micromolar, and found that the trend of AD subunits having lower cotranslational assembly rates than AR subunits held across all quartiles, with the strongest effect in the highest abundance bin (OR = 0.43, $p = 3.2 \times 10^{-7}$) (**Figure 3.2B**). These results were mirrored by a complementary analysis using active ribosome-protected fragment counts specific to HEK293 cells (**Appendix 3.1B**) (Clamer et al. 2018), employed by (Bertolini et al. 2021) for the detection of cotranslational assembling proteins.

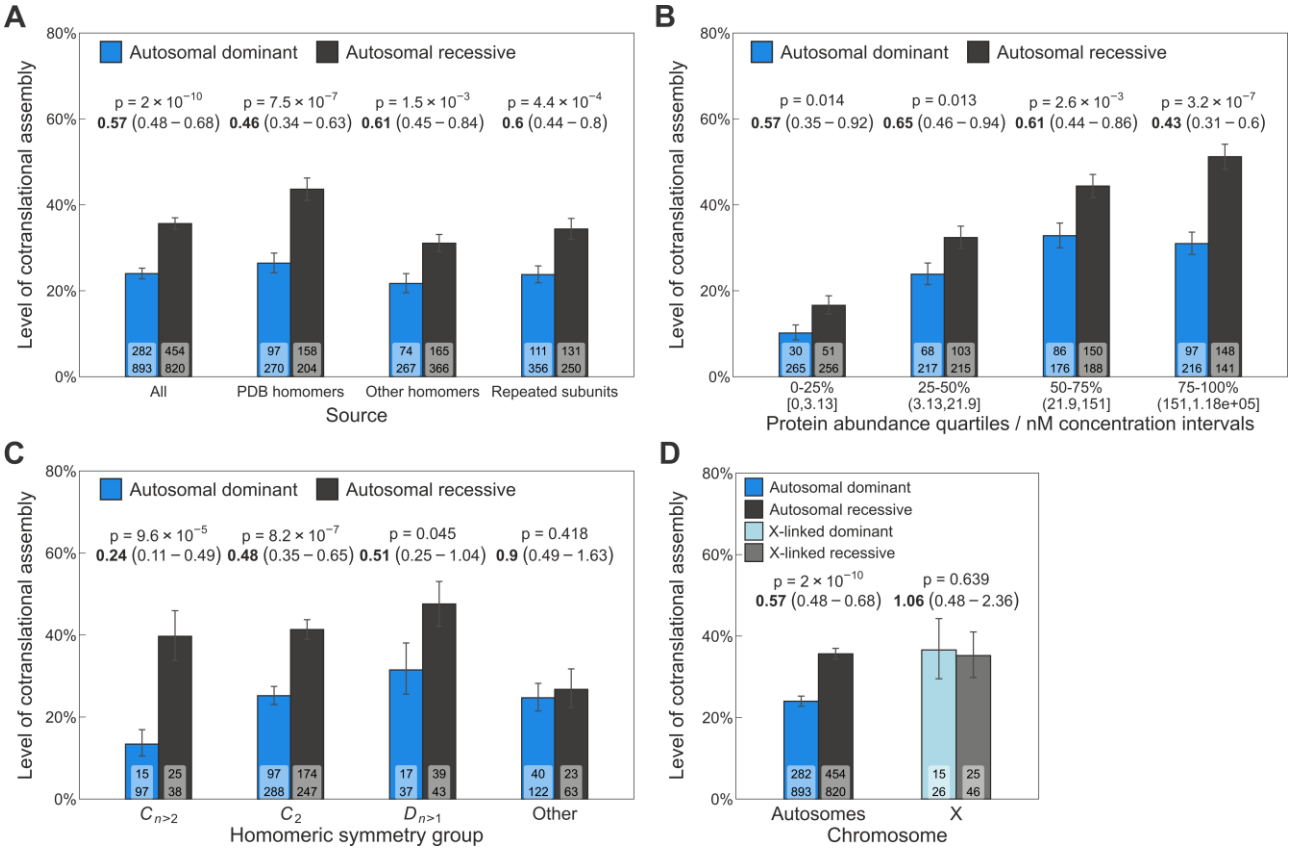


Figure 3.2 AD genes are depleted in cotranslationally assembling subunits.

(**A**) Level of cotranslational assembly in homomers and repeated subunits among AD versus AR genes grouped by subunit source (see **Methods**). Bar values are percent level of cotranslational assembly; error bars are Jeffrey's 68% binomial credible intervals. The p-value from the hypergeometric test and the OR (in bold) and its 95% confidence interval are shown above the bars. Labels on bars are the count of cotranslationally assembling subunits (top) and all other subunits (bottom). (**B**) to (**D**) have the same parameters. (**B**) Level of cotranslational assembly binned into protein abundance quartiles. Each bin corresponds to 25% of proteins by count, and the corresponding approximate nanomolar concentration intervals are shown in brackets. (**C**) Level of cotranslational assembly in genes of homomers and repeated subunits with AD and AR disease inheritance split by symmetry groups: cyclic ($C_{n>2}$), twofold (C_2), dihedral ($D_{n>1}$), and other. (**D**) Comparison of the level of cotranslational assembly in genes of homomers and repeated subunits on autosomes and the X chromosome.

Symmetry plays an important role in the formation of homomers, with each symmetry group exhibiting unique sequence and structural features that influence their functional roles (Bergendahl and Marsh 2017; Goodsell and Olson 2000). The three most common symmetry groups in the human proteome are twofold (Schönflies notation: C_2), higher-order cyclic ($C_{n>2}$), and dihedral symmetry ($D_{n>1}$). In a previous study, we showed that cotranslationally forming complexes tend to have large interfaces across the symmetry groups (Badonyi and

Marsh 2022). However, there are also symmetry-level differences, with members of the cyclic symmetry being the least likely to undergo cotranslational assembly (**Appendix 3.1C**). When we divided the AD and AR homomers based upon their symmetry group, we observed substantial variation in the level of cotranslational assembly (**Figure 3.2C**). For instance, if we randomly select a cyclic complex with AD inheritance, the odds of cotranslational assembly are 0.24 times lower compared to an AR cyclic complex ($p = 9.6\times10^{-5}$). In contrast, the odds of cotranslational assembly for a dihedral complex are only 0.51 times lower ($p = 0.045$). Symmetries that have a low representation in the human proteome, such as helical, cubic and asymmetric homomers, were grouped into the "other" category. We did not find a significant trend in this group, which may be due to the heterogeneous properties of their members.

For example, torsin-A is a chaperone involved in synaptic vesicle recycling and it forms a helical complex with 8.5 subunits per turn (Demircioglu et al. 2019). The DN mutation E303del in torsin-A results in the deletion of a glutamate near the C terminus involved in early-onset torsion dystonia (Torres et al. 2004). Because complexes with helical symmetry are topologically open, they would undergo unlimited polymerisation under ideal conditions (Marsh and Teichmann 2015). Despite torsin-A having been detected to cotranslationally assemble (Bertolini et al. 2021), the large number of subunits and the nature of fibre assembly will likely render it susceptible to the DN effect, analogous to tubulin subunits that make up helical microtubules (Attard, Welburn, and Marsh 2022). This contrasts with comparably large complexes with a closed topology that could benefit from cotranslational assembly, such as the dihedral (D_{39}) major vault protein, which has been observed to be readily assembled by the polyribosome (Mrazek et al. 2014).

We also considered the genetic dominance of mutations on autosomes and the X chromosome. Autosomal genes typically exist in two copies, either homozygous or heterozygous, with one allele on each chromosome. Genes on the X chromosome, on the other hand, are hemizygous in males and present in two copies in females, where one allele is usually silenced, excluding a subset of genes that escape X-inactivation in a tissue-specific manner. This means that cotranslational assembly is unlikely to be effective in buffering X-linked dominance, since there is no wild-type allele to counteract the phenotype. Our findings support this idea, as we did not find a significant difference in the level of cotranslational assembly between X-linked dominant and recessive genes of homomers and repeated subunits (**Figure 3.2D**; OR = 1.06, $p = 0.639$).

Finally, we examined a range of confounding variables to ensure the robustness of the results, including protein length, the confidence-based classification of cotranslationally assembling proteins (Bertolini et al. 2021), and presence of coiled-coil motifs. Similar to the analysis of protein abundance in **Figure 3.2B**, we controlled for protein length, because AD homomers and repeated subunits tend to be longer than AR (**Appendix 3.1D**, $p = 7.1\times10^{-10}$; Wilcoxon rank-sum test). Since longer proteins take more time to translate and could form larger subunit interfaces than shorter proteins, the difference in length between AD and AR subunits would actually favour the cotranslational assembly of AD subunits. When we split the subunits into length quartiles, we found that AD subunits have a significantly lower proportion of cotranslational assembly across the four bins (**Appendix 3.1E**), suggesting that the overall trend is not confounded by protein length.

Bertolini et al. have provided a confidence-based classification for the cotranslationally assembling proteins (Bertolini et al. 2021). However, relying entirely on the high confidence candidates is prohibitive to these analyses, as they make up only one-fifth of the detected proteins. To address potential biases coming from low confidence candidates, we divided the subunits into high and low confidence groups (**Appendix 3.1F**). The

high confidence group excludes all low confidence proteins, and similarly, the low confidence group excludes high confidence proteins altogether. The results showed a significant reduction in the level of cotranslational assembly in AD compared to AR subunits in both the high confidence group (OR = 0.66, $p = 7.7\times10^{-3}$) and the low confidence group (OR = 0.55, $p = 5\times10^{-10}$). One possible explanation for the stronger trend in the low confidence group is that they are enriched in membrane-bound complexes. These complexes tend to adopt cyclic symmetry, which has the strongest buffering capacity among the symmetry groups, as demonstrated in **Figure 3.2C**. On the other hand, high confidence candidates are limited to exclusively cytoplasmic or nuclear proteins by design (Bertolini et al. 2021), and thus may not reflect the full diversity of protein complexes.

Coiled-coil motifs are highly enriched among cotranslationally assembling proteins (Bertolini et al. 2021). Our analysis showed that homomers and repeated subunits participating in cotranslational assembly are significantly enriched in alpha helices (effect size = 0.161, $p = 1.5\times10^{-52}$; Wilcoxon rank-sum test), and this remains unchanged even after the removal of coiled-coil motif containing proteins from the data (effect size = 0.155, $p = 9.6\times10^{-44}$). To account for a potential bias from alpha helix content, we divided the subunits into four quartiles and re-examined the trend (**Appendix 3.1G**). The results support the trend across the bins, although with a lack of statistical significance in the quartile with the lowest alpha helix content.

Thus, our analyses suggest that none of the potential confounding factors have a significant impact on the trend, reinforcing the idea that the allele-specific assembly of protein complex constituents can act as a buffer for certain dominant mutations.

3.2.2 Subunits with DN disease mutations are less likely to assemble cotranslationally than subunits with heterozygous LOF mutations

Given the lower occurrence of cotranslational assembly in AD genes compared to AR genes, we sought to investigate further the molecular mechanisms underlying this trend. Nearly all mutations with dominant inheritance cause disease via one of three broad mechanisms: heterozygous loss of function (LOF, or haploinsufficiency), gain of function (GOF) and the dominant-negative effect (DN) (Backwell and Marsh 2022). We hypothesized that subunits with DN disease mutations should show a reduction in cotranslational assembly compared to genes with LOF mutations. This is because cotranslational assembly reduces the mixing of wild-type and mutant subunits, therefore lessening the likelihood that the mutant will interfere with the function of the wild type and inflict a DN effect.

To begin, we classified 1,185 AD genes (66% of known AD genes) into LOF, GOF and DN mechanisms using text-mining approaches and manual curation of the corresponding evidence (detailed in **Methods**). For example, a DN mutation in the ferritin light chain complex, which stores iron in a readily available form, has been linked to neurodegenerative disorders associated with iron accumulation in the brain (Curtis et al. 2001). Specifically, the F167SerfsX26 mutation replaces a C-terminal short helix with a stretch of disordered residues, which is thought to have a DN effect by creating large pores in the complex (**Figure 3.3A**), thus affecting its iron storage ability. Although the mutation has severe functional consequences, it does not impede assembly. By contrast, the LOF mutation L2067P in neurofibromin 1, observed in spinal neurofibromatosis (Kaufmann et al. 2001), affects the protein's folded core and dimer interface (**Figure 3.3B**). Mutations like these may escape nonsense-mediated decay, and either lead to aggregation of the protein before assembly can occur or render its interface incompatible with assembly, hence have no effect on the wild type.

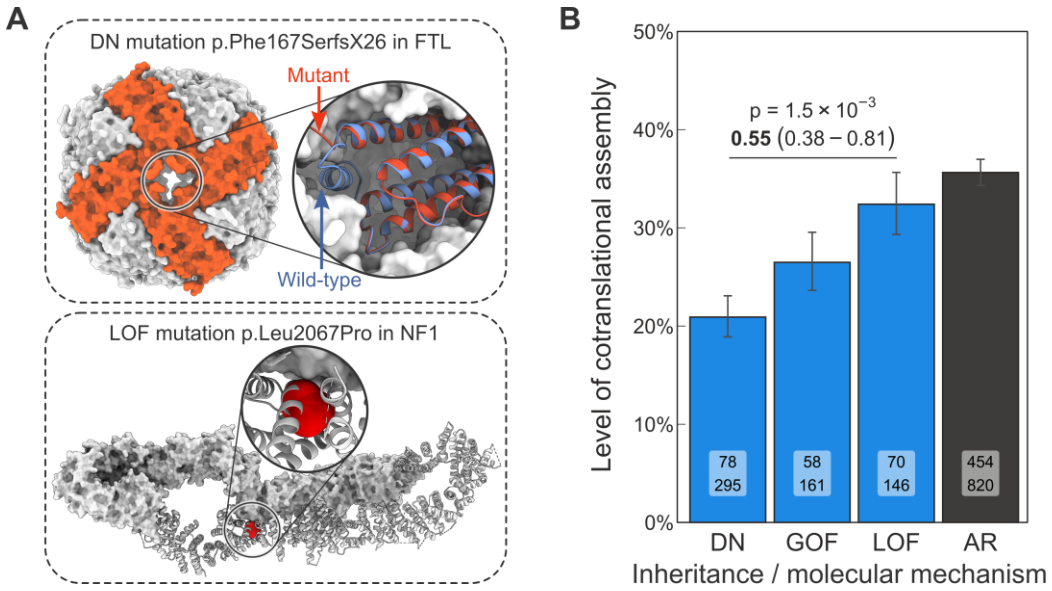


Figure 3.3 Subunits with DN disease mutations are less likely to assemble cotranslationally than subunits with heterozygous LOF mutations.

(**A**) Known examples of protein-level genetic disease mechanisms. Top: Structure of the p.Phe167SerfX26 mutant ferritin light chain complex (PDB ID: 4v6b) overlaid on the wild type (2ffx). Bottom: Structure of the wild-type neurofibromin 1 dimer (7mp6) shown as surface (top subunit) and as cartoon (bottom subunit), with the LOF mutation–associated residue Leu2067 highlighted. (**B**) Level of cotranslational assembly in homomers and repeated subunit heteromers according to dominant molecular mechanisms and AR inheritance. Bar values are percent level of cotranslational assembly; error bars are Jeffrey's 68% binomial credible intervals. The p-value from the hypergeometric test and the OR (in bold) and its 95% confidence interval are shown for the DN versus LOF comparison. Labels on bars are the count of cotranslationally assembling subunits (top) and all other subunits (bottom).

We first evaluated the suitability of the gene sets for analysis by examining properties known to be associated with the different molecular mechanisms, such as the change in Gibbs free energy (ΔΔG) upon pathogenic mutations and their clustering in 3D space (Gerasimavicius, Livesey, and Marsh 2022). It was previously observed in a subset of membrane proteins that DN mutations tend to have low predicted ΔΔG values, consistent with the fact that the mutant protein needs to remain stable enough to assemble into complexes (McEntagart et al. 2016). In agreement with this, we found that DN mutations in homomers and repeated subunits have significantly lower predicted ΔΔG values compared to LOF mutations (**Appendix 3.2A**; $p = 1.3\times10^{-16}$; Wilcoxon rank-sum test). In terms of 3D clustering, non-LOF mutations are often concentrated in specific regions of a protein, such as interfaces and functional sites, while LOF mutations tend to be more dispersed throughout the structure (Gerasimavicius, Livesey, and Marsh 2022). Consistent with this, DN mutations in our data exhibit higher 3D clustering than LOF mutations (**Appendix 3.2B**; $p = 4.7\times10^{-4}$; Wilcoxon rank-sum test) and are enriched at homomeric interfaces (**Appendix 3.2C**; $p = 1.3\times10^{-18}$; hypergeometric test).

We next addressed our hypothesis by calculating the fraction of cotranslationally assembling subunits in each molecular mechanism group, shown in **Appendix 3.2D**. As expected, the fraction is markedly lower among DN (20.9%) compared to LOF subunits (32.4%, OR = 0.55, $p = 1.5\times10^{-3}$, hypergeometric test) and AR subunits (35.6%, OR = 0.49, $p = 6.9\times10^{-8}$). At the molecular level, AR and heterozygous LOF mutations are very similar in their effect. Recessive disorders are almost always due to biallelic (homozygous or compound heterozygous) loss of function, with a few rare examples of biallelic gain of function (Cavaco et al. 2018; Drutman et al. 2019;

Wishner et al. 1975). Our results indicate that the level of cotranslational assembly in subunits with monoallelic and biallelic LOF mutations is similar, but subunits with DN mutations are observed to assemble cotranslationally less frequently. Therefore, allele-specific protein complex assembly may prevent some mutations from exhibiting a clinical phenotype caused by certain heterozygous variants. Interestingly, while there is no statistically significant difference in cotranslational assembly between the GOF and LOF classes (26.5% *vs* 32.4%, $p = 0.11$), we observed a significant depletion in the GOF class relative to AR (26.5% *vs* 35.6%, OR = 0.65, $p = 4.7\times10^{-3}$). We speculate that this discrepancy may be due to the assembly-mediated dominant-positive effect that often underlies GOF mutations (Backwell and Marsh 2022).

Cyclic symmetry is found at a much higher frequency in GOF homomers than the LOF class (**Appendix 3.2D**; 21.3% *vs* 6.4%, $p = 3.9\times10^{-3}$; Fisher's exact test). Interestingly, this symmetry group has the lowest level of cotranslational assembly (18.4%) in comparison to twofold homodimers (27.3%) and dihedral complexes (31.2%) (**Appendix 3.1C**). For this reason, we investigated the potential confounding effect of structural symmetry, and found distinct preferences among the molecular mechanisms (**Appendix 3.2E**). DN homomers, similar to GOF, are enriched in cyclic symmetry relative to LOF (20.4% *vs* 6.4%, $p = 4\times10^{-3}$; Fisher's exact test). However, dihedral symmetry is highly enriched in AR homomers compared to DN homomers (12.6% *vs* 4.8%, $p = 1.3\times10^{-3}$; Fisher's exact test). We suspect that these symmetry group compositions are reflective of biases in protein function, because the relationship between disease mechanisms and protein function is well established. For example, disorders caused by genes encoding enzymes are primarily recessive (Jimenez-Sanchez, Childs, and Valle 2001), genes encoding transcription factors are more likely to be haploinsufficient (Seidman and Seidman 2002) and those of membrane channels commonly give rise to non-LOF disease mechanisms (Celesia 2001). These admittedly broad functional classes have been linked to structural properties, with metabolic enzymes being enriched in dihedral symmetry, transcription factors in twofold symmetry, and membrane channels in cyclic symmetry (Bergendahl and Marsh 2017; Forrest 2015; Goodsell and Olson 2000). Our investigation into protein functional classification confirmed these assumptions (**Appendix 3.2E**). Metabolic enzymes are overrepresented in AR subunits (AR *vs* all other, OR = 4.27, $p = 3.9\times10^{-40}$; Holm-Bonferroni corrected Fisher's exact test), membrane transporters among GOF and DN subunits (OR = 2.96, $p = 4.8\times10^{-9}$ and OR = 1.99, $p = 4.8\times10^{-6}$, respectively), and transcription factors have 6.9-fold higher odds to be associated with the LOF class than a subunit sampled randomly from the disease gene pool ($p = 2.15\times10^{-19}$). When homomers in the different molecular mechanism classes are grouped by their symmetry, the level of cotranslational assembly is consistently lower among DN subunits than in LOF or AR subunits (**Appendix 3.2F**). The only exception is the LOF class with cyclic symmetry, where the rate of cotranslational assembly cannot be reliably estimated because no cotranslationally assembling member has been identified.

Lastly, we split the analysis into homomers and repeated subunits of heteromers. We found that repeated subunits with DN mutations exhibit a non-significant reduction in cotranslational assembly relative to the LOF class (**Appendix 3.2G**; OR = 0.74, $p = 0.188$) and a significant reduction relative to the AR class (OR = 0.55, $p = 3.3\times10^{-3}$). As expected, homomers with DN disease mutations are more strongly depleted in cotranslational assembly compared to these classes (OR = 0.44, $p = 1.1\times10^{-3}$ and OR = 0.43, $p = 2\times10^{-6}$, respectively). Overall, the results demonstrate that the genetic buffering capacity of cotranslational assembly, despite evident differences in structural symmetry, is neither confounded by symmetry nor is it exclusive to homomers, but extends to repeated subunits of heteromeric complexes.

3.2.3 Interfaces of homodimers with DN disease mutations are C-terminally shifted

To further understand the observation that subunits with DN disease mutations are less likely to undergo cotranslational assembly, we investigated the impact of interface area, which we have previously established as an important correlate of cotranslational assembly (Badonyi and Marsh 2022). Homomeric complexes with larger subunit contact areas are more hydrophobic and experience a stronger drive to assemble early on the ribosome. Due to the known confounds of structural symmetry, we performed the analysis split by homomeric symmetry groups. The analysis revealed that subunits associated with DN mutations do not have smaller interfaces than other disease-related subunits (**Appendix 3.3A**). On the contrary, the interfaces of LOF homomers are significantly smaller relative to DN subunits across the main symmetry groups. This finding is consistent with the enrichment of pathogenic mutations at interfaces of DN subunits (**Appendix 3.2C**), which, assuming a random mutation model, would be less likely to occur if the interfaces were small. However, larger interface areas for GOF and DN subunits could also reflect biological importance, given that their molecular mechanisms depend on complex formation. Conversely, subunits in the LOF group may have a higher proportion of crystallographic interfaces, which are typically smaller (Prasad Bahadur et al. 2004).

We next examined the interface location of the subunits, because the idea that N-terminal regions of proteins are more likely to be involved in cotranslational interactions has received strong experimental support (Bertolini et al. 2021; Kamenova et al. 2019; Natan et al. 2018; Shiber et al. 2018). We hypothesised that interfaces of homomeric subunits with DN disease mutations should be C-terminally shifted, reflective of their lower tendency to assemble cotranslationally. To test this, we calculated the relative interface location for all homomeric subunits and the average interface location for the different symmetry groups (Badonyi and Marsh 2022) (**Appendix 3.3B**). We found that the interfaces of homodimers with DN disease mutations are significantly more C-terminal compared to what is expected from the symmetry group ($p = 0.025$; Holm-Bonferroni corrected Wilcoxon rank-sum test against basemean). To quantify this difference, we resampled the homodimer dataset with replacement and calculated confidence intervals.

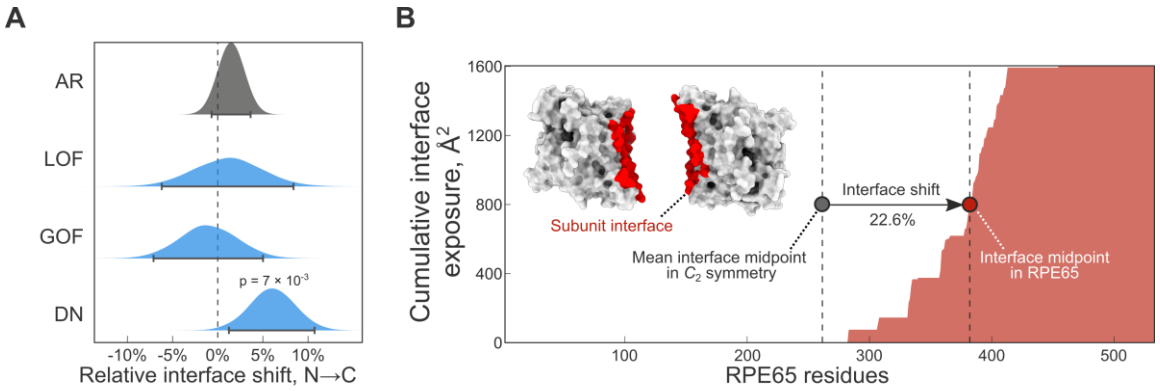


Figure 3.4 Interfaces of homodimers with DN disease mutations are C-terminally shifted.
(**A**) Bootstrap distributions of the difference between the relative interface location of the symmetry mean and the observed value for C_2 symmetric homodimers in the different classes. Error bars represent 95% confidence intervals of the percentile method, and the p-value was calculated from the resamples. (**B**) Cumulative interface exposure of the enzyme RPE65 during the translation process. Half of its final interface area (1,604 Å²) is reached 22.6% later than what is expected from in the C_2 symmetry group.

Figure 3.4A shows the bootstrap distribution of the relative interface shift, i.e. distance from the symmetry mean (raw relative interface location shown in **Appendix 3.3C**). The relative interface shift can be interpreted as a percentage, where +5% indicates that the interface is C-terminal to its expected value by 5% of the protein's length. According to this, subunits of homodimers with DN disease mutations are 6% more C-terminal compared to the symmetry group mean ($p = 7 \times 10^{-3}$, resampling p-value). We exemplify this finding in **Figure 3.4B**. The retinoid isomerohydrolase (RPE65) is an enzyme critical for phototransduction in the retinal pigment epithelium. Retinitis pigmentosa has been linked to both recessive and dominant mutations in RPE65, making heterozygous mutations in the gene less likely to be caused by simple LOF mechanisms. The DN mutation D477G was found to exert a DN effect and delay chromophore regeneration (Bowne et al. 2011). Notably, the interface of RPE65 during translation is exposed 22.6% later than what would be expected from an average homodimer, which creates a condition that disfavours cotranslational assembly. The observation is consistent with the absence of cotranslational assembly in RPE65 (Bertolini et al. 2021) and supports a model whereby subunits that expose their interfaces later in the translation process are less likely to assemble cotranslationally, and are in turn more susceptible to DN mutations.

3.3 Discussion

Cotranslational assembly of homomers is thought to result in complexes whose subunits originate from the same allele (Bertolini et al. 2021; Gilmore et al. 1996; Mrazek et al. 2014). A possible consequence of this mechanism is that subunits harbouring pathogenic heterozygous mutations may be sequestered into half of the protein complex pool rather than mixing with the wild type and inflicting functionally harmful effects. By comparing the fraction of cotranslationally assembling subunits associated with Mendelian diseases, we showed that genes of homomers and repeated subunits with mutations inherited in an AD pattern are significantly depleted in this mode of assembly compared to AR genes. Moreover, among AD genes of homomers, those that exert a DN effect are the least likely to cotranslationally assemble compared to other protein-level molecular mechanisms of disease, but importantly, to those that predominantly harbour heterozygous LOF mutations. Our results therefore reveal a previously hypothesised genetic buffering mechanism (Natan et al. 2017; Perica et al. 2012), whereby complexes undergoing cotranslational assembly are to some extent protected from the deleterious consequences of DN mutations.

We observe AR complexes to have consistently high levels of cotranslational assembly regardless of their structural symmetry. It was first proposed by Wright (S. Wright 1929) and Haldane (Haldane 1930), whose ideas were developed further by Hurst and Randerson (Hurst and Randerson 2000), that recessivity is a consequence of selection for larger amounts of protein, because the high abundance of enzymes is a "safety factor" (Kacser and Burns 1981) that increases their robustness to dominant mutations. It is possible that the abundance and the structural properties of metabolic enzymes, such as their preference for dihedral symmetry (Bergendahl and Marsh 2017), necessarily lead to frequent cotranslational assembly events, representing an additional safety factor against the deleteriousness of dominant, especially DN mutations. Although the evolution of protein oligomeric state can arise from nonadaptive processes (Hochberg et al. 2020; Lynch 2013), it is not implausible that biological phenomena such as this impose weak selection.

Our results hint at the extraordinary regulation of protein complex assembly within cells. Allele-specific assembly in homomers may emerge from the inherent colocalisation of their nascent chains, although certain protein structural features appear to predispose subunits to the process. Interestingly, we also observed

repeated subunits of heteromeric complexes to exhibit genetic buffering by cotranslational assembly. According to one hypothesis, subunits may combine information in their mRNAs and protein sequences to increase the efficiency of assembly mediated by RNA-binding proteins (Bourke, Schwarz, and Schuman 2023; X. Chen and Mayr 2022; K. C. Martin and Ephrussi 2009). A range of membraneless compartments have been put forward as putative sites of intense protein complex assembly under physiological conditions, including TIS granules (Ma and Mayr 2018), assemblysomes (Panasenko et al. 2019) and translation factories (Chouaib et al. 2020), which may well represent the same type of organelles (reviewed in (Morales-Polanco et al. 2022)).

Across diverse proteomes, interface contacts of homomers are enriched towards the C terminus, which is thought to be the product of evolutionary pressure on folding to happen before assembly (Natan et al. 2018). By contrast, N-terminal protein interfaces have been found to favour cotranslational assembly (Badonyi and Marsh 2022; Bertolini et al. 2021; Kamenova et al. 2019; Natan et al. 2018; Shiber et al. 2018). Interestingly, our structural analysis suggests that interfaces of homodimers with DN disease mutations are significantly shifted towards the C terminus relative to what is expected from their symmetry group. As a possible consequence, their interfaces become exposed in nascent polypeptides relatively late during translation, strongly reducing the likelihood of cotranslational assembly, as illustrated in **Figure 3.5**. This observation represents a survivorship bias, so that we tend to observe subunits cause disease via a DN mechanism when they "escape" cotranslational assembly and co-assemble with wild-type subunits.

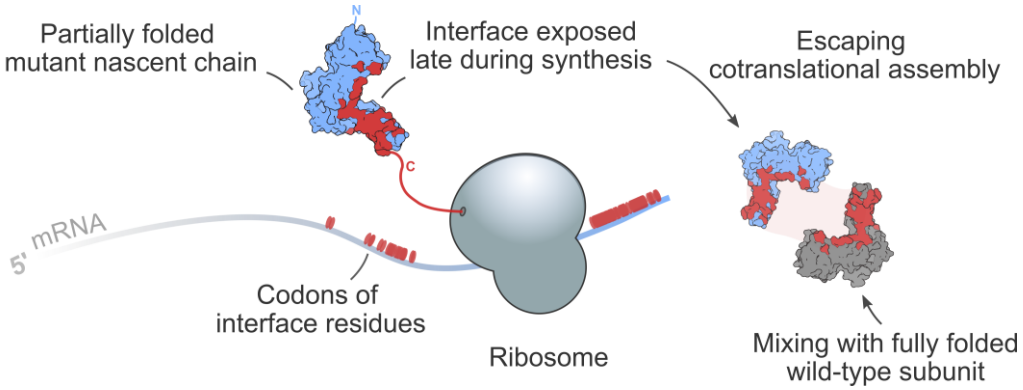


Figure 3.5 Mechanistic interpretation of C-terminally shifted interfaces in homodimers.
A mutant subunit is more likely to assemble posttranslationally when it exposes its interface residues late in the translation process, which can increase the level of mixing with the wild-type subunit.

Ongoing efforts to develop variant effect predictors focusing on the molecular consequences of protein-coding variants should consider whether the subunit assembles cotranslationally or, providing that structural data is available, the properties of interfaces in order to prioritise those with a possible DN effect. As demonstrated in this study, a substantial fraction of subunits with mutations inherited in an AD pattern display a DN phenotype, likely including many of those that have not yet been characterised and classified under one of the molecular mechanism classes. Additionally, current clinical sequencing pipelines frequently identify inherited and de novo heterozygous variants in recessive genes, which are ranked lower under the assumption that they would not be pathogenic in a heterozygous state (Birgmeier et al. 2020; Eilbeck, Quinlan, and Yandell 2017; Paila et al. 2013; G. T. Wang, Peng, and Leal 2014; K. Wang, Li, and Hakonarson 2010; Zemojtel et al. 2014). However, a DN effect is possible if the gene encodes a homomer or repeated subunit heteromer (Backwell and Marsh 2022), and especially if its complex does not assemble cotranslationally.

Ultimately, our results shine light onto the fascinating connection between inheritance, which determines the genetic traits of an individual, and protein complex assembly, which takes place only after the genetic information has been decoded. Further research is needed to measure directly the effect of allele-specific assembly on wild-type and mutant subunits in vivo.

3.4 Methods

Structural data

We searched the Protein Data Bank (PDB, on 2021-02-18 for all polypeptide chains >50 amino acids and >90% sequence identity to human canonical sequences in the UniProt proteome UP000005640. For genes that map to multiple chains, we selected a single chain ranking by sequence identity, the number of unique subunits in the complex, and by the number of atoms present in the chain. In every case, we used the first biological assembly and its symmetry assignment was taken from the PDB. The interface area was calculated at residue-level between all pairs of subunits with AREAIMOL from the CCP4 suite (Winn et al. 2011), using a probe radius of 1.4 Å. The interface was defined as the difference between the solvent accessible surface area of the subunit in isolation and within the context of the full complex. Subunits with interfaces >400 Å^2 were considered for analysis to exclude potentially crystallographic interfaces.

We extended the PDB dataset with homology models of human homomeric complexes in the SWISS-MODEL repository (Bienert et al. 2017) (UniProt release 2022_02). Models based on isoform sequences were excluded. The software AnAnaS (Pagès and Grudinin 2018; Pagès, Kinzina, and Grudinin 2018) was run on default settings to determine the number of subunits and the symmetry group of the complexes. In rare cases when symmetry was not detected, we assigned the symmetry group of the PDB template used to model the complex. If a protein was found in multiple homology models, we selected the one with the largest number of subunits followed by the length of the modelled chain. The interface area was calculated at residue-level between all pairs of subunits with FreeSASA 2.0.3 (Mitternacht 2016) using a probe radius of 1.4 Å. We performed pairwise alignments between the modelled chain and the paired UniProt sequence to confirm residue correspondence to the canonical sequence, because any mismatch in residue numbering could influence the relative interface location metric. Similarly to the PDB structure data, only subunits with interfaces >400 Å^2 were included in the analyses. When the SWISS-MODEL dataset was pooled with the PDB dataset, we prioritised the homomeric subunit with the larger interface area.

Relative interface location

The relative interface location is a value between 0 and 1 indicating the location of the interface relative to the protein termini (N=0 and C=1), and it was calculated as previously described (Badonyi and Marsh 2022). To ensure the analysis is not biased by homologous proteins, we generated a distance matrix based on the sequences of the chains from the structures using Clustal Omega 1.2.4 (Sievers et al. 2011).The distances were converted to per cent identities and the matrix was filtered to below 50% using a redundancy-filtering algorithm. Only those structures were included in the analysis that passed the homology cutoff.

FoldX free energy calculation

FoldX 5.0 (Delgado et al. 2019) was used to calculate the change in Gibbs free energy of ClinVar (Landrum et al. 2018) missense mutations in AlphaFold predicted structures of human proteins (Tunyasuvunakool et al. 2021). The `RepairPDB` command was first run to minimise structures followed by the `BuildModel` command

on the repaired structures. The final Gibbs free energy change was calculated as the average of ten replicates, and in subsequent analyses residues with pLDDT < 50, which are predicted to be disordered in solution (Akdel et al. 2022), were excluded.

3D clustering of missense pathogenic mutations

The extent of disease clustering (EDC) metric expresses the proximity of every disease non-associated protein residue to a known disease-associated residue, and it was calculated as previously described from AlphaFold predicted structures (Gerasimavicius, Livesey, and Marsh 2022). Briefly, for each residue with pLDDT > 50, we calculated the alpha carbon distance to all other residues with a known ClinVar disease mutation, selecting the shortest distance. The final metric is derived as the ratio of the common logarithm of non-disease and disease average distances. Values ≤1 indicate that the mutations are dispersed and those >1 suggest a degree of spatial clustering. Only proteins with at least 3 pathogenic or likely pathogenic mutations in ClinVar were included.

Alpha helix content

The percentage of alpha helix residues was calculated from the AlphaFold predicted structures of human proteins using DSSP 2.2.1 (Kabsch and Sander 1983).

Gene-level inheritance patterns

Gene-disease inheritance relationships were obtained from OMIM (Amberger et al. 2015). Gene-specific XML files were retrieved via the OMIM API in four batches over consecutive days ending on 2022-07-07. Inheritances were extracted from the “phenotypeInheritance” node of each XML file.

Gene set of homomers and repeated subunits

We extended the gene set of homomers identified by our structural mapping pipeline with genes that have non-structural evidence to form homo-oligomers or are present in >1 copy in a complex. For homomers, we used UniProt (Bateman et al. 2021), EMBL-EBI ComplexPortal (Meldal et al. 2019), CORUM (Giurgiu et al. 2019), the OmniPath database (Dénes et al. 2021), as well as single spanning membrane homodimers from the Membranome 3.0 database (Lomize et al. 2022). For repeated subunits, we extracted protein chains that appear in multiple copies in the biological units of complexes in the PDB (Marsh et al. 2015), and included proteins that have a stoichiometry >1 in the OmniPath database. Homomers were removed from the repeated subunit list to create a non-redundant dataset.

Gene-level classification of dominant molecular mechanisms

We classified autosomal dominant genes into molecular disease mechanisms via text-mining PubMed (https://pubmed.ncbi.nlm.nih.gov/) titles and abstracts and OMIM XML gene entries. We searched PubMed using the keywords “dominant negative” for the dominant-negative (DN) mechanism, “gain of function” OR “activating mutation” for the gain-of-function (GOF) mechanism, and “haploinsufficiency” OR “haploinsufficient” OR “dosage sensitivity” OR “dosage sensitive” OR “heterozygous loss of function” for the loss-of-function (LOF) mechanism. The same workflow was applied to OMIM entries. The resulting corpus was tokenised into sentences and, to facilitate downstream data curation, we filtered for lines that explicitly mention the keywords, thus keeping the most descriptive lines for each gene. The LOF class was appended with genes annotated in the ClinGen database (Rehm et al. 2015) as “Sufficient evidence for dosage pathogenicity” (as of 2022-07-07), and a supporting evidence was added from the ClinGen entry. The raw

evidence lines were manually reviewed and obvious false positives were removed. For example, in the LOF class a line may be: "… individuals harboring a heterozygous deletion in ATAD3A are unaffected suggesting a dominant-negative pathogenic mechanism or a gain-of-function mechanism for de novo missense variants rather than haploinsufficiency" (Harel et al. 2016), which explicitly dismisses haploinsufficiency as a molecular mechanism. Importantly, we noticed that a significant proportion of GOF and DN evidence lines pertained to artificial constructs employed in biological research and were not linked to human disease, requiring additional manual curation for verification. In overlap cases, when genes belong to multiple categories, we employed a hierarchical strategy to create a non-redundant gene list: DN>GOF>LOF.

Protein functional classes

Functional classification of proteins were retrieved from PANTHER version 17.0 (Mi et al. 2021). In the category "Transporter" we included the classes "transporter", "transmembrane signal receptor" and "membrane traffic protein". In the category "Metabolic enzymes" we grouped "nucleic acid metabolism protein", and "metabolite interconversion enzyme"**.** Finally, the category "TF/chromatin regulator" represents the combined classes of "gene-specific transcriptional regulator" and "chromatin/chromatin-binding, or "regulatory protein".

Protein abundance

Protein abundances were obtained from the integrated human dataset (version 2021) of PAXdb (M. Wang et al. 2015). Parts per million (ppm) values were converted to molar concentration based on the equation given by (Dubreuil, Matalon, and Levy 2019).

HEK293 active ribosome profile

Normalised active ribosome protected fragments in the Human Embryonal Kidney 293 lineage were determined by (Clamer et al. 2018). The data is available via the NCBI Gene Expression Omnibus (Edgar, Domrachev, and Lash 2002) under accession GSE112353. Values were averaged over the two biological replicates and transcripts with values <1 were excluded from the analysis.

Cotranslationally assembling proteins in HEK293 cells

The gene set was downloaded from the supplemental material of reference (Bertolini et al. 2021).

Coiled coil motif containing proteins

Coiled-coil motif containing proteins were retrieved from UniProt, using the search terms: `(keyword:KW-0175) AND (organism_id:9606) AND (reviewed:true)`.

Position of genes on chromosomes

Genes were mapped to chromosomes using the consensus coding sequence (CCDS) database (Pujar et al. 2018) downloaded via the NCBI FTP site.

Molecular graphics

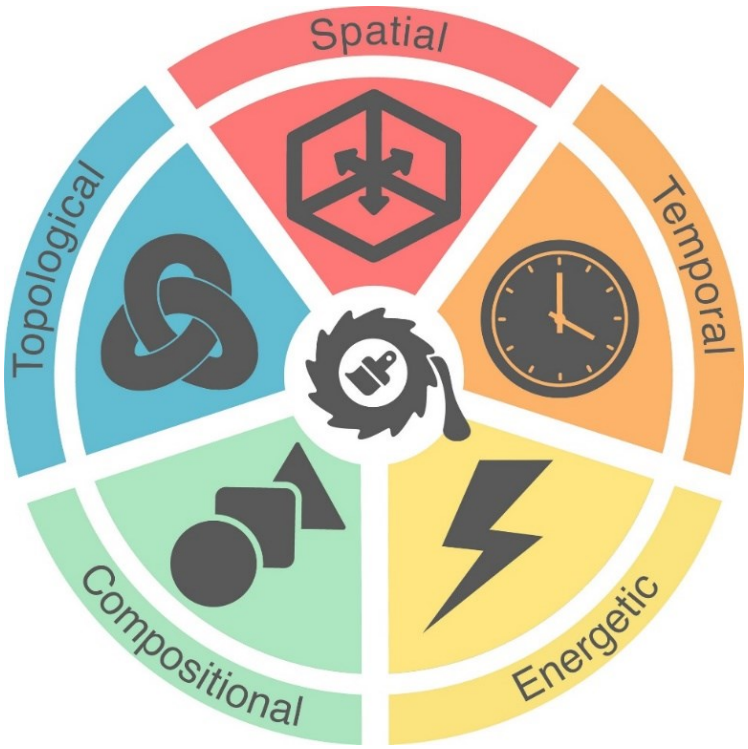
Visualisation of structures was performed with UCSF ChimeraX version 1.5 (Pettersen et al. 2021).

Statistical analyses

Data exploration and statistical analyses were carried out in RStudio "Elsbeth Geranium" release, using R version 4.2.2. The R packages used for analyses were: `tidyverse`, `tidytable`, `rsample`, `rstatix`, `scales`,

ggridges, and ggbeeswarm. Error bars in bar charts represent 68% Jeffrey's binomial credible intervals and the probabilities between the proportions of cotranslationally assembling subunits were calculated from the hypergeometric distribution. The 95% confidence interval for the odds ratio was calculated with the standard error method, where the value of the 97.5th percentile point of the normal distribution (~1.96) was derived as `stats::qnorm(0.975)` in R. In Wilcoxon rank-sum tests the effect size was defined as the z-score computed from the p-value over the square root of sample size. In multiple comparisons, the Holm-Bonferroni method was used to correct for familywise error rate. In the bootstrap analysis, data were stratified for molecular mechanisms in 10,000 resamples. The p-value was calculated by determining the fraction of point estimates indicating a C-terminal interface shift, with correction for finite sampling. The 95% confidence intervals of the bootstrap estimates were derived using the percentile method (Jung et al. 2019).

4 Hallmarks and evolutionary drivers of cotranslational protein complex assembly



Graphical abstract of the *FEBS Journal* review article (Badonyi and Marsh 2023b)**.**

4.1 The hallmarks of cotranslational assembly

In the previous chapters, I showed that protein complex subunits with larger interfaces tend to assemble cotranslationally more often, and that mutations in proteins that cotranslationally assemble give rise to disease through the dominant-negative effect less often. Aside from these important attributes, just how much do we know about the properties of this assembly mechanism? Based upon our current understanding of the process, cotranslational assembly can be distinguished by five hallmarks that encompass spatial, temporal, energetic, compositional and topological aspects (**Figure 4.1**). Some molecular features can be clearly assigned to a single hallmark, while others may exhibit characteristics that are shared by two or more hallmarks, depending on how we interpret its contribution to the assembly process. Additionally, many features are supported by independent experimental observations, whereas others, however intuitive they may be, remain weakly supported or speculative. In this section, I delve into each of the hallmarks and the evidence supporting their role in cotranslational assembly.

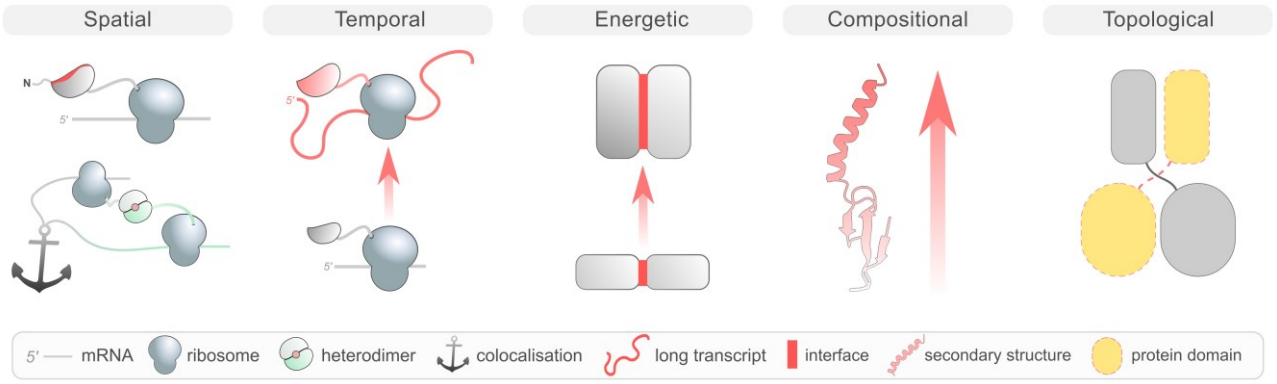


Figure 4.1 Hallmarks of cotranslational assembly.

The illustration depicts selected features in each hallmark. (**1**) *Spatial*: The spatial hallmark of cotranslational assembly involves the synergistic organization of mRNAs, ribosomes and nascent chains in the confined translational milieu. N-terminal interface location (top) and transcript colocalisation (bottom) are shown. (**2**) *Temporal*: The temporal hallmark of cotranslational assembly includes factors that delay or accelerate a process, such as the length of the ORF (as shown), cotranslational chaperone activity, mRNA structure effects on translation, local codon context-dependent translational slowdowns and functional ribosome collisions. (**3**) *Energetic*: The energetic hallmark of cotranslational assembly involves the subunit affinity, which is determined by the strength of the interactions between two proteins and can be approximated by the size of the interface (as shown) between two subunits. (**4**) *Compositional*: The compositional hallmark of cotranslational assembly involves the higher frequency of cotranslational assembly in proteins with coiled-coil motifs, and the tendency for cotranslationally assembling subunits to have a higher proportion of alpha helices than other proteins (as shown). (**5**) *Topological*: The topological hallmark of cotranslational assembly includes the influence of structural symmetry and may include intertwined associations (as shown).

4.1.1 Spatial

The spatial hallmark of cotranslational assembly comprises the supramolecular organisation of mRNAs, ribosomes and nascent chains in the cell, which synergistically create optimal conditions in the confined space of the translational milieu (Natan et al. 2017). In prokaryotes, genes encoding heteromeric subunits are often organised into operons, which are transcribed into a single polycistronic mRNA, warranting their cotranslation. Because higher order eukaryotes lack operonal genome structure, other processes have evolved to facilitate cotranslation of functionally related mRNAs when necessary (Bourke, Schwarz, and Schuman 2023; X. Chen and Mayr 2022; K. C. Martin and Ephrussi 2009). Evidence suggests that colocalisation of mRNAs can be mediated by specific cis-regulatory elements, or “zip codes”, in their sequences (Mendonsa et al. 2023), which can become physically linked in the cell through RNA binding proteins to form so-called RNA regulons (Keene 2007). Another interesting phenomenon is that of densely packed polysomes in both bacteria and humans (Brandt et al. 2009, 2010), which appear to adopt configurations that maximise the distance between ribosome exit tunnels. This organisational level may be in place to prevent non-productive interactions between nascent chains, but it is possible that they coordinate their folding and assembly in parallel. A third spatial feature of cotranslational assembly in our classification is an N-terminally localised interface (Wells, Bergendahl, and Marsh 2015), which has been extensively corroborated by experiments (Bertolini et al. 2021; Kamenova et al. 2019; Natan et al. 2018; Seidel et al. 2022; Shiber et al. 2018). While an N-terminal interface can spatially favour assembly on the ribosome, it also serves as a temporal

determinant by allowing more time for an interaction to develop. We will discuss N-terminal interface location and its bearings on evolution in the following section.

4.1.2 Temporal

A property may be considered a temporal hallmark of cotranslational assembly if its functional contribution is realised through either delaying or accelerating a process. For example, the length of the open reading frame (ORF) in the mRNA, i.e. the length of the eventual polypeptide, is a temporal determinant, because a cotranslational interaction is more likely to occur if the ribosome dwell time is long. This is reflected in the earliest discovered cotranslationally assembling proteins, which, being filamentous components of the cytoskeleton, are some of the longest proteins in the human proteome (Isaacs et al. 1992; Redick and Schwarzbauer 1995; Veis and Kirk 1989). Similarly, cotranslational chaperone activity may be considered a temporal determinant. Ribosome-associated chaperones, such as the bacterial trigger factor and the yeast Ssb, are thought to increase the efficiency of folding by binding to hydrophobic patches emanating from the exit tunnel and thereby transiently halting their folding (Agashe et al. 2004; Döring et al. 2017; Oh et al. 2011). Shiber et al. proposed that Ssb binding may be coordinated with assembly so that interface residues, which are generally hydrophobic, are shielded from premature assembly until folding of the full domain is completed (Shiber et al. 2018). Other temporal determinants may include mRNA structure effects (Faure et al. 2016; Nissley and O'Brien 2014), local codon context-dependent translational slowdowns (Bertolini et al. 2021; Panasenko et al. 2019), and functional ribosome collisions unrelated to mRNA decay pathways (Arpat et al. 2020; Han et al. 2020; T. Zhao et al. 2021). These factors can influence cotranslational assembly by modulating the elongation rate, the folding pathway of the nascent chain, and ultimately the time available for partner protein association. However, very little is known about how translation dynamics control the cotranslational assembly process, making it an area that requires more attention.

4.1.3 Energetic

The energetic hallmark of cotranslational assembly incorporates the strength of the interactions between two proteins. This subunit affinity is the primary determinant of the rate at which complexation will occur, although other factors, such as chaperone activity can overlap with it through kinetic control of the energy landscape. While the factors that determine affinity can be complex and difficult to quantify (Brooijmans, Sharp, and Kuntz 2002; Erijman, Rosenthal, and Shifman 2014; Vangone and Bonvin 2015), previous research has found that the size of the interface between two subunits corresponds well with the energy of hydrophobic bonding (Chothia 1974; Chothia and Janin 1975), which is a key contributor to overall affinity in protein complexes (Fersht et al. 1985). As a result, using interface size as an approximation of the hydrophobic bonding energy has been successful at predicting assembly and disassembly pathways of protein complex subunits (Levy et al. 2008; Marsh et al. 2013; Wells, Bergendahl, and Marsh 2016). Recently, we proposed that cotranslationally assembling proteins should have large interfaces, thus higher affinities, to explain their strong drive to assemble early on the ribosome, and found the hypothesis to be upheld for both homomers and heteromers detected by Bertolini et al. (Badonyi and Marsh 2022; Wells, Bergendahl, and Marsh 2015). This supports the idea that a large subunit interface can contribute to creating energetically favourable conditions for cotranslational assembly.

4.1.4 Compositional

A compositional hallmark of cotranslational assembly has also emerged from the data. Bertolini et al. have observed proteins with coiled-coil motifs to frequently assemble on the ribosome (Bertolini et al. 2021). This potentially helps to explain why many cytoskeletal proteins containing these motifs were among the first cases of cotranslational assembly to be discovered. Coiled-coil motifs are common protein structural elements, characterised by two or more alpha helices wrapped around one another and held together by hydrophobic interactions (Burkhard, Stetefeld, and Strelkov 2001). One possible explanation for the higher frequency of cotranslational assembly in proteins with coiled-coil motifs is that the alignment of heptad repeats – the regularly alternating patterns of amino acids that form the basis of these motifs – would not occur correctly by chance if their assembly relied on random collisions in the cell. Interestingly, our analysis suggests that cotranslationally assembling subunits, irrespective of coiled-coil motifs, tend to have a higher proportion of alpha helices than other proteins (Badonyi and Marsh 2023a). This may be due to two factors. First, for alpha helices, the loss of accessible surface area upon folding is smaller compared to beta strands (Chothia 1976), which will result in a larger contribution of buried surface area at the interface. Second, because alpha helices form more short-range intramolecular contacts than beta strands, they can reach their tertiary structure more rapidly (Plaxco, Simons, and Baker 1998) and form folded domains earlier on the ribosome, favouring productive cotranslational interactions.

4.1.5 Topological

When considering the topological hallmark of cotranslational assembly, it is important to account for structural symmetry. Homomers typically form symmetric complexes, with different symmetry groups having unique sequence and structural features that influence their functional roles (Bergendahl and Marsh 2017; Goodsell and Olson 2000). The three most commonly observed symmetry groups in homomers are twofold (Schönflies notation, C_2), cyclic ($C_{n>2}$) and dihedral ($D_{n>1}$) (Marsh and Teichmann 2015). These groups are named according to the arrangement of subunits in relation to n axes, imaginary lines around which the subunits can be rotated while maintaining the same pattern. Subunits of cyclic and dihedral symmetries form multiple interfaces with adjacent protomers, resulting in larger interfaces than homodimers, which might make them more likely to assemble cotranslationally. However, the specific cellular functions associated with different symmetry groups may also play a role in determining which complexes undergo cotranslational assembly. For example, dihedral complexes have a higher rate of cotranslational assembly compared to twofold dimers, but the same trend does not hold true for cyclic complexes (Badonyi and Marsh 2022). This discrepancy could be due to the fact that cyclic complexes are enriched in channel proteins (Forrest 2015), which require different assembly mechanisms for stable membrane immersion (Hegde and Keenan 2022). Overall, the relationship between symmetry and cotranslational assembly is complex and depends on a variety of factors.

A potential topological hallmark of cotranslational assembly that may apply to heteromeric subunits is "intertwined" associations (Schwarz and Beck 2019), where two proteins interact in such a way that a segment of one protein extends into the other. The reason why intertwined associations may be more likely to form cotranslationally follows the analogy of knotted proteins, whose folding trajectories can be enhanced by folding on the ribosome (Jackson, Suma, and Micheletti 2017). For example, the yeast fatty acid synthase complex, composed of alpha and beta subunits, forms an intertwined association shortly after the alpha subunit emerges

from the ribosome exit tunnel (Fischer et al. 2020; Seidel et al. 2022; Shiber et al. 2018). The N-terminal domain of the alpha subunit becomes deeply embedded in the beta subunit, leading to the formation of the malonyl/palmitoyl-transferase functional module, the most stable interface in the complex [92]. Domain invasion motifs may represent another form of intertwined cotranslational association (Hsia et al. 2007; Seidel et al. 2022). Seidel et al. examined two WD repeat-containing proteins of the nuclear pore complex, Seh1 and Sec13, which possess incomplete beta-propellers. They have shown that, upon cotranslational assembly, the missing blades are completed by the invading blades of Nup85 and Nup145C, respectively. Cotranslational domain-swapping events could also contribute to intertwined associations. Bertolini et al. found that BTB domains commonly employed cotranslational assembly, and BTBs are dimerisation domains in which mutually swapped segments can make up 60% of the interface (Yanshun Liu and Eisenberg 2002; Wodak, Malevanets, and MacKinnon 2015). Because domain-swapped systems are thought to arise through single deletion events (Hashimoto and Panchenko 2010; Lynch 2012), by exploring the potential link between intertwinedness and cotranslational assembly, we could gain a better understanding of its evolution.

4.2 Evolutionary impetus for cotranslational assembly

It is important to stress that certain conditions inherently favour the occurrence of cotranslational assembly in cells. All things being equal, it is more likely in homomers because of colocalisation of multiple active ribosomes on their mRNA (Natan et al. 2017). This is especially relevant in bacteria, where at least half of the proteome assemble into homomeric complexes (Kühner et al. 2009; Lynch 2012). When sequencing of bacterial genomes became possible, analysis of operons suggested a strong positive selection for clustering of genes that encode physically interacting proteins (Mushegian and Koonin 1996). This led to the hypothesis that operons may play a role in coordinating heteromeric complex assembly. Further research revealed that cotranscription of protein complex genes can reduce fluctuations in subunit level (Sneppen et al. 2010), and that operon gene order and the assembly order of the encoded subunits is highly coherent (Wells, Bergendahl, and Marsh 2016), supporting the idea. Shieh et al. have shown that inserting the genes for a protein complex into the same operon significantly improves assembly efficiency compared to when the genes are located at distant sites in the genome (Shieh et al. 2015). Through the example of a luciferase heterodimer, they demonstrated that only the operon-encoded complex assembled cotranslationally in a sequential manner. Without operons, eukaryotes may rely on transcript colocalisation mechanisms for efficient assembly. For example, mRNAs encoding seven subunits of the actin-related protein 2/3 complex colocalise in fibroblast filopodia (Mingle et al. 2005), and paralogues of this protein family were found to assemble cotranslationally by Duncan and Mata (Duncan and Mata 2011). Altogether, these observations emphasise that proximity of a nascent chain to its interaction partner can locally confine subunits and render their assembly less stochastic.

Cotranslational assembly is also naturally more likely to occur when the residues that will form the interface are located in the N-terminal half of the protein. However, because interface residues are mostly hydrophobic, the absence of the binding partner can cause misfolding and aggregation (Kamenova et al. 2019; Shiber et al. 2018). There is some evidence of evolutionary selection to mitigate such negative consequences by favouring the location of interface residues towards the C terminus in homomers, ensuring that protein folding occurs before assembly and creating, in turn, conditions that do not promote cotranslational interactions (Natan et al. 2018). Bertolini et al. have found that simultaneous assembly often occurs when N-terminal dimerisation

domains are exposed, suggesting that, despite some selection against these in homomers, those that do have them tend to assemble cotranslationally (Bertolini et al. 2021). Although there is a lack of a similar evolutionary constraint in heteromers (Natan et al. 2018), Kamenova et al. have established through engineering of TATA-binding protein-associated factors that the exact sequence of a heterodimerisation domain is less important for cotranslational assembly than its N-terminal position (Kamenova et al. 2019). Thus, N-terminal interface location appears to be an important factor in determining which proteins undergo cotranslational assembly for both homomers and heteromers.

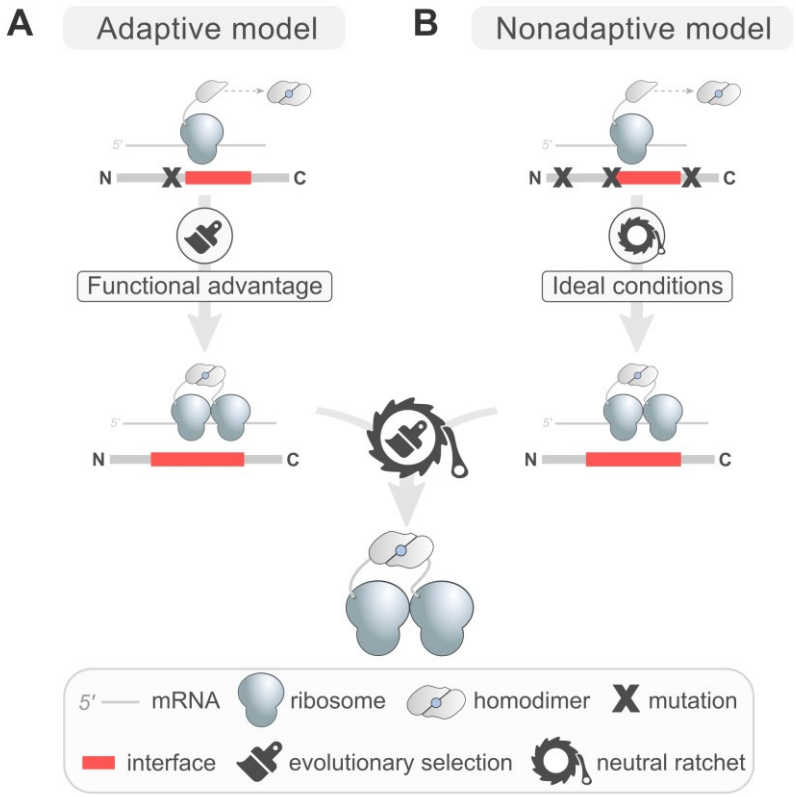


Figure 4.2 Adaptive and nonadaptive models of cotranslational assembly evolution.
(**A**) The adaptive model proposes that cotranslational assembly has evolved because it provides a functional advantage. Here, the cotranslationally assembling subunit is selected from the sequence space for its fitness benefit, and reversion to the previous assembly mode in the future is prevented by either purifying selection or a neutral ratchet. (**B**) The non-adaptive model suggests that cotranslational assembly has emerged simply because it is energetically favourable and optimal conditions for it were fortuitously created. While it does not provide a fitness benefit, subsequent genetic drift "locks" the subunit in this assembly mode through a ratchet-like evolutionary mechanism.

Considering that selection acts to minimise misassembly of protein complexes (Marsh et al. 2013), it is tempting to view cotranslational assembly as a process that has evolved because it provides an adaptive advantage (**Figure 4.2A**). A deluge of benefits to cotranslational assembly have been hypothesised, including but not limited to, avoidance of misinteractions and faster assembly (Natan et al. 2017; Schwarz and Beck 2019; Wells, Bergendahl, and Marsh 2015), ability to integrate into quality control pathways (Bertolini et al. 2021; Juszkiewicz and Hegde 2018; Mena et al. 2020; Schwarz and Beck 2019), isoform specificity (Bertolini et al. 2021; Schwarz and Beck 2019), prevention of transcription factor and cell cycle component toxicity (Schwarz and Beck 2019), faithful assembly of paralogous complexes (Schwarz and Beck 2019), and an increased tolerance to dominant-negative mutations via allele-specific assembly (Badonyi and Marsh

2023a; Natan et al. 2017; Nicholls et al. 2002; Perica et al. 2012). However, experimental evidence supporting these benefits is still very limited. In an adaptive scenario where cotranslational assembly is functionally beneficial, the hallmarks, such as an N-terminal interface and its size, could provide the mechanistic and structural basis for selection. In a recent study, we put the adaptive model of cotranslational assembly to the test (Badonyi and Marsh 2022). We hypothesised that if interfaces have evolved to increase the level of functionally beneficial cotranslational assembly, then N-terminal interfaces should have a tendency to be larger than C-terminal interfaces. Our analysis of heteromeric complexes with subunits that form multiple interfaces in bacteria, yeast and human supported the adaptive model. Importantly, the trend was significant in evolutionarily ancient subunits or those organised into operons in bacteria, suggesting that large N-terminal interfaces may have been selected for to seed the assembly pathway cotranslationally.

While selection for efficiency can be strong, particularly in prokaryotes (Lynch 2006, 2012), there is no basis to assume that adaptive forces explain all cases of cotranslational assembly. In some instances, seemingly needless biological complexity may be explained by constructive neutral evolution, which suggests that certain functions we observe as essential may not have been so when they first appeared (Gray et al. 2010; Lukeš et al. 2011; Pillai, Hochberg, and Thornton 2022; Stoltzfus 1999). In recent years, evidence supporting the theory has been growing particularly in the context of protein complexes (Emlaw et al. 2021; Fernández and Lynch 2011; Hochberg et al. 2020; Schulz et al. 2022). Building on these ideas, we proposed that that cotranslational assembly could represent the workings of a ratchet-like neutral process (Badonyi and Marsh 2022). In this model, ideal conditions for assembly on the ribosome spontaneously arise through random genetic drift, deletion or fusion events (**Figure 4.2B**). The ideal conditions refer to the physical and chemical properties of the ribosome and the nascent polypeptides that fortuitously allow for interactions between subunits. These interactions are energetically favourable in the sense that they lower the free energy of the system, but they do not necessarily provide a fitness benefit for the cell. In other words, the interactions are thermodynamically rather than biologically driven. The functional benefit of cotranslational assembly for the protein is not important, as reversion to the previous assembly mode will be blocked by the accumulation of mutations that are tolerated in the extant protein but not in the ancestor (Gray et al. 2010). Consequently, cotranslational assembly can become essential for the subunit, creating interdependence and irreversibility in the system, and its mode will be conserved through purifying selection.

Ultimately, experiments should be explicitly designed to answer if selective forces have driven two subunits to assemble cotranslationally. As previously noted (Schulz, Sendker, and Hochberg 2022), destroying the function of a biological process does not distinguish between neutral and adaptive reasons for its existence. Disruptive mutagenesis of a cotranslationally assembling protein pair shows only that the assembly mode of the complex is essential today, and not that there was a functional benefit behind its emergence. To gain insight, researchers could follow the approach of Schulz et al. who investigated the evolutionary path of Rubisco's small subunit and found that the historical interface gain with the large subunit coincided with improved catalytic efficiency and enabled mutations that would not have been possible otherwise (Schulz et al. 2022). By applying a similar strategy to a suitable model system, we can better understand the rationale behind the emergence of cotranslational assembly in complex assembly pathways.

4.3 Concluding remark

In this chapter, I have synthesised the hallmarks of cotranslational assembly into a simple framework that can help researchers interested in the phenomenon navigate the field. Although tremendous community effort has enabled this categorisation, many hallmarks remain incomplete and insufficiently supported. Open questions about the regulation and evolution of cotranslational assembly represent outstanding challenges. How does translation dynamics control the cotranslational assembly process? How common is sequential assembly in heteromers, and what methodology will lead to their proteome-wide discovery? Is chaperone activity synchronised on a global scale with the course of cotranslational assembly? What is the true extent to which adaptive forces have driven the evolution of cotranslationally assembling complexes? What is the full array of functional benefits that cotranslational assembly can provably offer? Could the allele-specific nature of cotranslational assembly have affected the evolution of diploid genomes by serving as a buffer against dominant mutations? Answering these questions will require case studies and systems-level investigations into proteomes, but will undoubtedly advance our understanding of the role of protein complexes in cells.

5 Proteome-scale prediction of molecular mechanisms underlying dominant genetic diseases

5.1 Introduction

Genetic diseases are often caused by protein-altering mutations that act via different protein-level mechanisms. One such mechanism is dominant loss of protein function (LOF), whereby the mutation ablates a function that cannot be compensated by the nonmutant allele, leading to haploinsufficiency (Veitia 2002). Another mechanism is gain of function (GOF), which is characterized by an altered or newly appeared function in the mutant protein. Mutations may also have a dominant-negative (DN) effect, provided that the mutant allele can directly or indirectly affect the function of the wild type. DN effects are commonly "assembly-mediated" (Backwell and Marsh 2022), i.e. inflicted upon a protein through complex assembly, so that the mutant subunit physically interacts with the wild type. Moreover, some GOF mutations may also be assembly-mediated if the altered function of the mutant is conferred to wild-type subunits (Backwell and Marsh 2022). Understanding how missense mutations exert their effects is essential for better diagnoses and treatments of genetic disorders. Recent years have seen major improvements in experimental methods capable of probing a large number of variants at once. In particular, deep mutational scanning and its modifications hold huge promise for deciphering the effect of variants and, by extension, their mechanisms (Fowler et al. 2023). For now, relative to LOF, our understanding of the properties of variants that act via DN or GOF (collectively, non-LOF) mechanisms remains limited.

In many cases, the different dominant molecular mechanisms are not mutually exclusive in a gene. For example, a type of cardiomyopathy can be influenced by both titin haploinsufficiency (LOF) and truncated titin peptides that seem to "poison" wild-type complexes (DN) (Fomin et al. 2021). Although molecular mechanisms are indisputably variant-level phenomena, it has become evident that mutations in many genes are more likely to exhibit a specific mechanism (Gerasimavicius, Livesey, and Marsh 2022). This concept has shed light on certain structural and functional properties associated with non-LOF proteins at the systems level (Badonyi and Marsh 2023a; Gerasimavicius, Livesey, and Marsh 2022). Moreover, these analyses have called attention to a limitation of state-of-the-art variant effect predictors (VEPs), as they struggle to accurately predict the pathogenicity of non-LOF variants. This is an important skill to improve, otherwise there is a strong possibility that we could miss variants on the account of difficulties in computationally predicting their effects. Until a new generation of VEPs more sensitive to non-LOF variants is available, a gene-level predictor could potentially be very useful.

Here, I first demonstrate that non-LOF molecular mechanisms can be predicted at protein-level with the help of protein complex structural data (Badonyi and Marsh 2023a). A necessary limitation of this model is its reliance on experimentally resolved structures of protein complexes. I present a solution to this problem by introducing surrogate features, compensating for the incomplete structural and functional attributes across the proteome. To expand the model's predictive capacity, I build three binary classifiers, adding up to a "tripartite model", each tasked with identifying one of the molecular mechanism classes over another unique class (DN *vs* LOF and GOF *vs* LOF) or pooled classes (LOF vs non-LOF). This design maximizes the number of cases available for training with the resulting probabilities enabling a more robust classification regime.

Although the models individually have average performance, which is expected given the ultimately variant-level nature of the problem, together they can be combined into predictions that have biological and clinical utility. I demonstrate its capabilities by constructing an unbiased analysis set devoid of training and recessive genes and examining properties previously linked to protein-level molecular mechanism classes.

It is important to stress that the models do not predict disease involvement. A non-disease associated protein with a predicted mechanism does not imply that it contributes to disease, but rather that the protein's properties are most consistent with the given mechanism. These predictions offer valuable insight when assessing novel variants in genes with no prior disease linkage or pinpoint the likely mechanism of variants in a dominant gene with no previous association to a molecular mechanism. For example, it could help prioritise genes in clinical and population genetics data for laboratory studies and allow researchers to explore the functional, structural, and evolutionary properties inherent to the mechanisms. Predictions for the human UniProt reference proteome are available at https://osf.io/z4dcp/.

5.2 Methods

Software and databases

In this study, we rely on a number of properties derived from the AlphaFold-predicted monomeric structures of human proteins (Tunyasuvunakool et al. 2021). Solvent accessible surface area is calculated at residue level with AREAIMOL from the CCP4 programme suite (Agirre et al. 2023). FoldX version 5.0 (Delgado et al. 2019) is run to determine the predicted Gibbs free energy of folding (ΔΔG) of missense substitutions by first calling the `RepairPDB` command followed by `BuildModel`, with ΔΔG computed as the average of 10 replicates. Residue-level SCRIBER scores (J. Zhang and Kurgan 2019) for the human proteome are extracted from the 9606_database.json file provided by the DescribePROT database (B. Zhao et al. 2021). Protein-level UniProt ProtNLM embeddings (Gane, A. et al. 2022) are obtained via the UniProt FTP for the UP000005640_9606 reference proteome (2023_02 release). All 1024 embedding dimensions are extracted from the per-protein.h5 file using the `h5dump` command from the HDF5 command line tool. Clustal Omega version 1.2.4 (Sievers et al. 2011) is used to create a distance matrix of protein sequences for homology filtering. Scores from VEPs ESM-1v (Meier et al. 2021), EVE (Frazer et al. 2021), MetaRNN (C. Li et al. 2022), and VARITY_R (Wu et al. 2021) are gathered using an in-house pipeline described in (Livesey and Marsh 2023). Machine learning methods are implemented in R version 4.3.0 (R Core Team 2023) using `tidymodels` (https://www.tidymodels.org/) with extension packages `themis`, `probably`, and `DALEX`. Statistical tests are performed with the `rstatix` package. Bootstrap confidence intervals are calculated with the percentile method from 1,000 resamples (Jung et al. 2019).

ClinVar and gnomAD mapping to UniProt reference proteome

Genomic coordinates of pathogenic and likely pathogenic (hereinafter "pathogenic") missense variants are extracted from the ClinVar (Landrum et al. 2018) variant calling file (VCF) (accessed on 2023-06-15) with BCFtools (Danecek et al. 2021). Putatively benign variants (hereinafter "benign") with a "PASS" filter are extracted from gnomAD v2.1.1 (Karczewski et al. 2020) using the exomes all chromosomes VCF file. Mapping is performed with Ensembl VEP 107 (McLaren et al. 2016) using the `--uniprot` and `--canonical` flags. For each variant, we select either the canonical transcript or the first UniProt isoform on condition that the mapped amino acid corresponds to that in the sequence.

Lasso regression – feature selection

To prioritise genes that mainly give rise to non-LOF mutations over those that harbour LOF mutations, the following variables were included in the model:

1. gnomAD mutational constraint metrics (Karczewski et al. 2020):
    - pLI – Probability that transcript falls into the distribution of haploinsufficient genes.
    - pRec – Probability that transcript falls into distribution of recessive genes.
    - oe_lof – Observed over expected ratio for predicted loss-of-function variants in transcript.
2. Sequence-derived or evolutionary variables:
    - Protein-length – As per UniProt canonical isoform.
    - Number of paralogues (Huang et al. 2010) – Paralogues of human genes were called from Ensembl (Cunningham et al. 2022) via the biomaRt R package.
    - Maximum identity to paralogue (Huang et al. 2010) – We calculated the protein sequence identity of each gene to every one of its paralogues using Clustal Omega version 1.2.4 (Sievers et al. 2011) and the maximum identity was kept.
    - Human-macaque (*Macaca mulatta*) dN/dS (Huang et al. 2010) – The ratios of nonsynonymous to non-synonymous substitutions between human-macaque orthologues were called from Ensembl (Cunningham et al. 2022) via the biomaRt R package.
    - ncGERP++ – The phylogenetic conservation values of the genes' regulatory sequences as derived from GERP++ scores were acquired from (Petrovski et al. 2015).
    - Number of domains (Huang et al. 2010) – The number of domains were taken from the human proteome-specific Pfam release 2021-11-15 (Mistry et al. 2021).
    - Membrane propensity – We calculated from protein sequences the mean value of the scale NAKH900110 "Normalized composition of membrane proteins" (Nakashima, Nishikawa, and Ooi 1990) from the AA index database (Kawashima et al. 2008).
3. Interaction network-based property:
    - Betweenness centrality (Huang et al. 2010) – This measure was calculated from the human protein interaction network of the STRING database version 11.5 (Szklarczyk et al. 2021) at the default score threshold of 400 using the STRINGdb and igraph R packages.
4. Structural properties:
    - Structural symmetry (monomer, heteromer, and homomeric symmetry groups: C_2, $C_{n>2}$, $D_{n>1}$, and other homomeric symmetry), interface size and number of subunits.
    - Structure isoelectric point – as previously described in (Badonyi and Marsh 2022).
    - Absolute contact order (Plaxco, Simons, and Baker 1998) – We calculated the contact order from the AlphaFold predicted human structures using the perl script written by Eric Alm available at https://depts.washington.edu/bakerpg/contact_order/contactOrder.pl. A copy of the script can be found in the OSF repository linked to this manuscript.
    - Mean structure pLDDT and alpha helix content.
5. Functional properties:
    - Protein functional classification from this study. Proteins with functions other than those introduced earlier were classified as "known other function" and those lacking a functional annotation were classed as "unknown function".

6. Experimental data:
    - Protein abundance.
    - Cotranslational assembly annotations.
    - RNA expression variance – We accessed the rna_tissue_consensus.tsv.zip file from the Human Protein Atlas (Uhlen et al. 2010) on 2022-09-01 and calculated the variance in expression per gene across the 54 tissues.

Lasso regression – data preprocessing

We assembled a dataset of 9,051 genes with the above features. Genes of monomers in the PDB were assigned an interface size of 0, a relative interface location of 0, a number of subunits of 1, and were assumed not to undergo cotranslational assembly even if they were detected by (Bertolini et al. 2021). Missing data in ten variables were imputed using five nearest neighbours (Gower 1971). The missing value rate were the following (per cent missing in brackets): ncGERP++ (9.7); human-macaque dN/dS (9.3); pLI, pRec, and oe_lof (7.3); betweenness centrality (5.6); protein abundance (4.2); number of domains (2.7); structure isoelectric point (2.1); RNA expression variance (1.3). Lastly, all nominal variables were one-hot encoded and numeric data was normalised to have a standard deviation of one and a mean of zero.

Lasso regression – model building and performance evaluation

In an initial model screen, we evaluated the performance of more complex statistical learning methods, including a random forest, a support vector machine, and a single layer neural network. However, no performance gain was observed relative to a simpler and more interpretable regression model (data not shown). Logistic regression with lasso (least absolute shrinkage and selection operator) is a solution to fitting a model in which only certain variables play a role. The algorithm applies increasingly larger penalties to multivariable regression coefficients, shrinking those of less important variables to zero, causing their sequential drop-out (L1 regularisation) and thus retaining only informative features. First, to avoid inflating the model's performance by presence of homologues, we performed redundancy filtering at 50% sequence identity on canonical protein sequences. This procedure removed 88 from the 879 genes with experimentally available structure data and dominant molecular mechanism classifications. The remaining 791 genes, of which 543 (69%) are non-LOF and 248 (31%) are LOF, were split into 75% training and 25% test sets with 10-fold cross-validation performed on the training set and repeated 3 times. The model was tuned using 18 values of the λ parameter, generated to be log-linearly distributed between 0 and 1. The final value of $\lambda = 0.00501$ was chosen on the basis that it yielded the highest prediction accuracy in the assessment-folds of the cross-validation. The model was finalised on the entire training set and evaluated on the test set. The similarity of ROC AUCs measured on the cross-validation folds (0.735) versus on the test set (0.743) suggested that the model had not been overfitted. Variable importance was computed as the absolute values of the β coefficients scaled to the [0,1] interval. Model building and evaluation was performed using the `tidymodels` R metapackage. Thresholds T1 and T2 were derived using the `threshold_perf()` function from the R package probably.

Support vector machines – feature selection and engineering

We use gene and protein-level features previously described in (Badonyi and Marsh 2023a), notably excluding subunit interface size, which is derived from experimentally determined structures of protein complexes, and the PANTHER functional classification of proteins (Mi et al. 2021), which is incomplete for the human

proteome. We sought to rationally select and design surrogate features that make up for the loss of protein structural and functional information. To replace interface size, we calculate the median SCRIBER score of residues that have >5% relative surface accessible surface area (RSA) in the corresponding AlphaFold predicted structure. RSA is defined as the ratio of the maximum solvent accessible surface area determined from Gly-X-Gly tripeptides (where X denotes the amino acid) (Miller, Janin, et al. 1987) and that of the residue in the context of the monomeric protein. To substitute protein functional classification, we use UniProt ProtNLM embeddings that capture information about Pfam domains and functional sequence features and transfer those to unannotated and manually unreviewed entries. In each model, to avoid overfitting on a large number of uncorrelated features, we select the top 20 ProtNLM embeddings by Wilcoxon effect size between the relevant binary outcomes. In addition, we introduce a pair of features related to the likelihood of randomly drawing a missense substitution with the characteristics of a particular type of molecular mechanism. This heuristic metric is derived as the median of the ratio of ESM-1v score and the FoldX-computed ΔΔG of all missense mutations based on possible amino acid substitutions in the canonical sequence. We use both raw ΔΔG and its absolute value (|ΔΔG|), which has been shown to improve the performance of pathogenic variant prediction (Gerasimavicius, Liu, and Marsh 2020). Finally, we complement existing population genetics constraint metrics with s_{het} (Zeng et al. 2023), which relates the frequency of LOF mutations in a gene to the strength of selection against them with higher values indicating stronger selection and lower tolerance to LOF mutations.

Support vector machines – data acquisition

Evidence for DN, GOF, or LOF molecular disease mechanism observed in a gene was collected using a combination of text-mining and semi-manual curation, as previously described (Badonyi and Marsh 2023a). We updated the gene set using newly added autosomal dominant genes from the Online Inheritance of Man (OMIM) database (Amberger et al. 2015) (accessed via the API on 2023-05-25) and developmental disorder genes with monoallelic inheritance from the Deciphering Developmental Disorders Genotype-to-Phenotype (DDG2P) database (https://www.deciphergenomics.org/ddd/ddgenes, accessed on 2023-05-25). The updated list contains 1,270 genes with one or more molecular mechanism class based upon available evidence. Of the genes, 874 are assigned to a unique class, 318 are assigned to two classes, and 78 genes intersect with all three classes.

Support vector machines – data preprocessing

In each model, we prioritise the first event level of the binary outcome. For example, in the DN *vs* LOF model, we treat any gene with a DN class annotation as DN even if it has other class assignments; this strategy is likewise applied to the GOF *vs* LOF and the LOF *vs* non-LOF models (where non-LOF represents pooled DN and GOF gene sets). The rationale is to maximise the available cases for the primary outcome and ameliorate the class imbalance. We next remove features that have a Spearman correlation >0.9 with another, which eliminates the gnomAD oe_lof metric (observed over expected ratio for predicted LOF variants in transcript) in all three models due to its high correspondence with s_{het}. The features are normalised and missing values are imputed based upon all other features with five nearest neighbours (Gower 1971). We use a distance matrix of the protein sequences to create a non-redundant dataset with proteins sharing <50% sequence identity within each outcome, e.g., within DN or LOF genes, but not across, as it may be important to learn differences between homologues assigned to different classes. To improve the signal-to-noise ratio, when possible, instead of randomly removing one protein from a pair of proteins above the sequence identity cutoff, we remove the one that overlaps with another class. This procedure results in the following class proportions across the

models (protein count, class percentage): DN (498, 52%) *vs* LOF (451, 48%), GOF (479, 51%) *vs* LOF (463, 49%) and LOF (610, 52%) *vs* non-LOF (559, 48%).

Support vector machines – initial model screen

To assess which statistical learning approach suits the data and the classification problem the best, we carried out a model screen. We chose five diverse modelling methods (tuned hyperparameters): lasso logistic regression (penalty), a multilayer perceptron (penalty, hidden units, epochs), a polynomial support vector machine (SVM) (cost, degree, scale factor, margin), and two tree-based methods, a random forest (trees, mtry [minimum number of randomly sampled features at each split], min_n [minimum number of data points a node must contain]) and the LightGBM algorithm (trees, tree depth, mtry, min_n, loss reduction) (Ke et al. 2017). Maintaining class ratios, we split the data into 75% training and 25% test sets and ran 10-fold cross-validation, similarly divided into 75% analysis and 25% assessment sets. The ROSE (random over-sampling examples) upsampling algorithm (Menardi and Torelli 2014) was applied to perfectly balance the classes. An efficient grid search was performed via the ANOVA race tuning method (Kuhn 2014) on a random grid of size $k = 10\times$(number of hyperparameters), which eliminated configurations unlikely to have been the best after 3 resamples using repeated measures ANOVA.

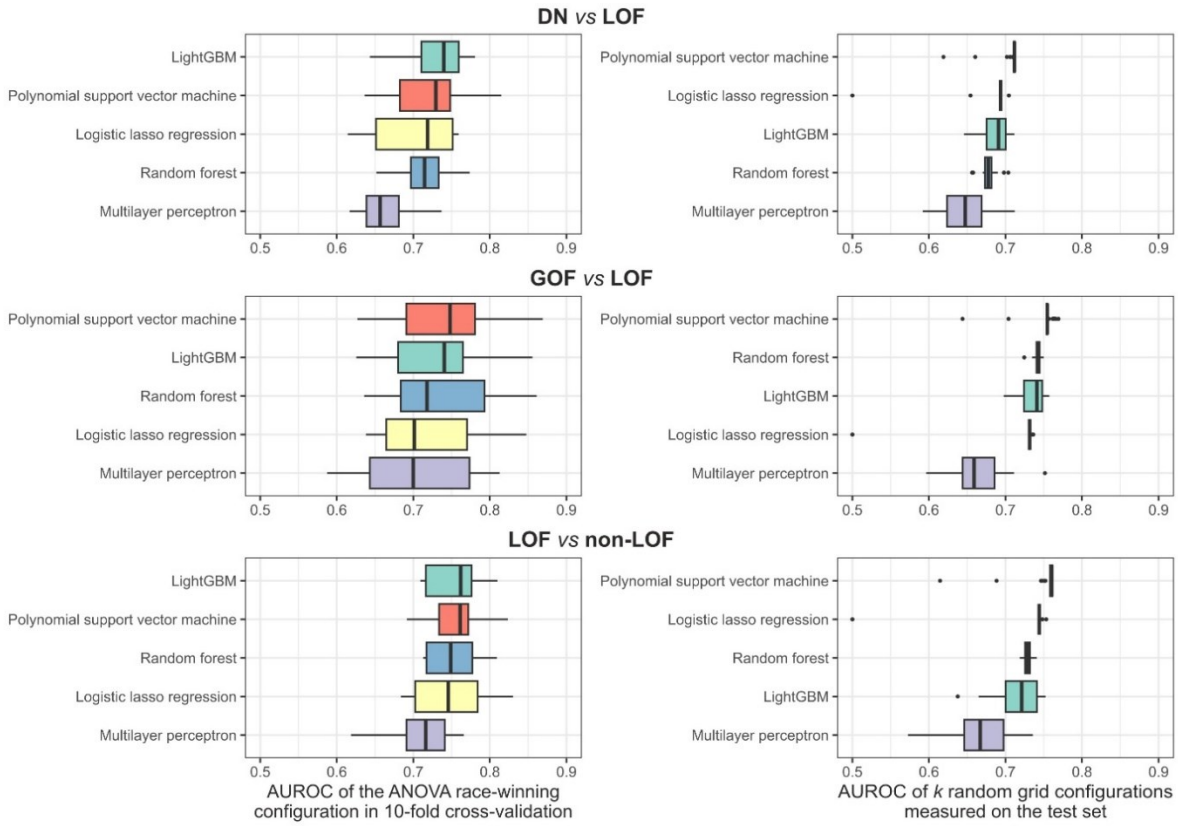


Figure 5.1 Results of the initial model screen.

Plots on the left show the AUROC of the best hyperparameter set of each model on the cross-validation folds. Plots on the right show the AUROC of k randomly generated parameter combinations as measured on the test sets, where $k = 10\times$(hyperparameters). Boxes denote data within 25th and 75th percentiles and the middle line represents the median. Whiskers extend from the box to 1.5× the interquartile range.

We compared the models by both the area under the receiver operating characteristics (AUROC) curve of the ANOVA race-winning configuration measured on the cross-validation folds as well as the AUROC of the *k* random grid configurations of each model on the test set (**Figure 5.1**). By the former, the top ranking models were LightGBM for DN *vs* LOF and LOF *vs* non-LOF models, and polynomial SVM for the GOF *vs* LOF model. However, because the *k* random grid configurations of the polynomial SVM consistently gave the best performance on the test sets, suggesting it performs best on unseen data without hyperparameter tuning, it was chosen to reduce the risk of overfitting and to maintain a similar prediction profile across the three models.

Support vector machines – model building

The final DN *vs* LOF, GOF *vs* LOF, and LOF *vs* non-LOF models are built using an SVM with a polynomial kernel. The data are divided into 75% training and 25% test sets with 3-times repeated 10-fold cross-validation, which itself is split into 75% analysis and 25% assessment sets. We perform Bayesian hyperparameter optimisation on the tunable parameters of the model: cost, degree, scale factor, and margin. Initially, a random grid of size 10 is created for the Gaussian process model and a maximum of 20 search iterations are performed. The models are finalised based upon the parameters that achieve the highest AUROC.

Support vector machines – model evaluation

We compute threshold-agnostic performance measures for the models based upon the test sets, including AUROC and area under the precision recall curves (AUPRC) (**Figure 5.2**).

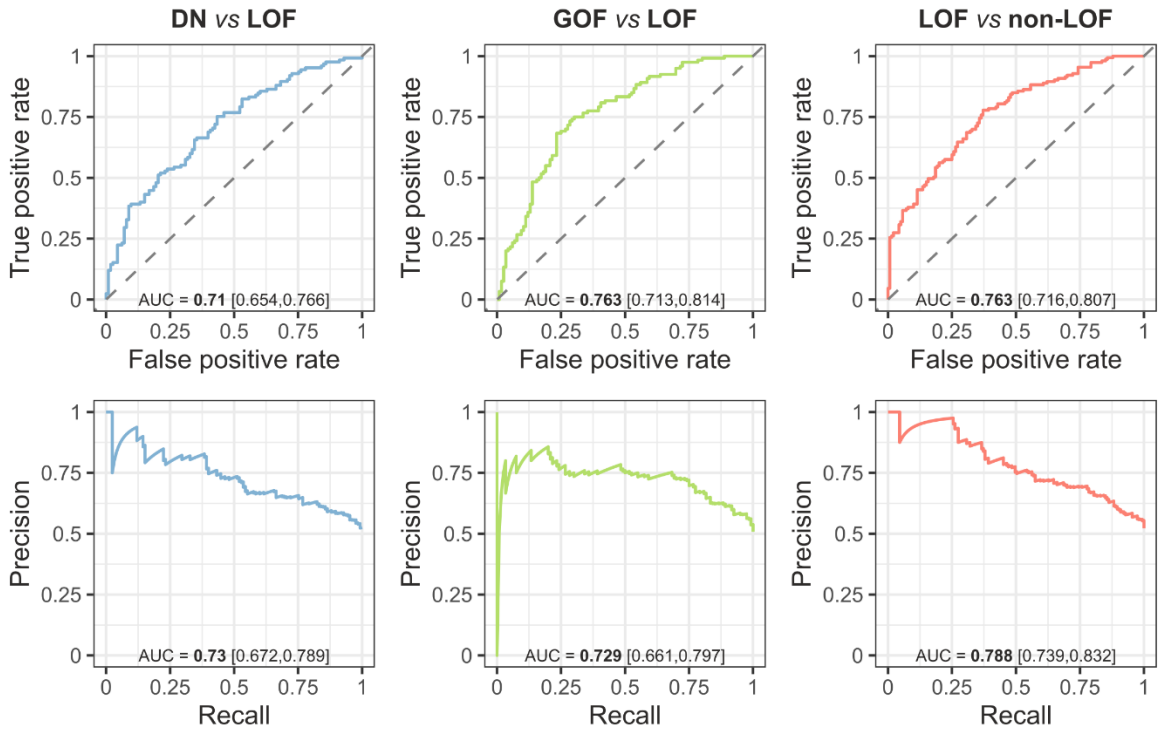


Figure 5.2 ROC and PR curves of the models measured on the test sets.

Numbers in bold denote point estimates and those in brackets represent 90% bootstrap confidence intervals.

To assess the models in further analyses, we determine the probability thresholds at which the models reach 50% sensitivity on their respective test sets. At these thresholds, which we refer to as t_{50}, the models are able to correctly identify half of the positive cases and approximately 80% of the negative cases. Threshold-

dependent performance measures at t_{50}, including accuracy, Matthews correlation coefficient (MCC) and the F1 score, are summarised in **Table 5.1**.

Model	t_{50}	Sensitivity	Specificity	Accuracy	MCC	F1
DN *vs* LOF	0.61	**0.519** [0.448,0.584]	**0.778** [0.717,0.841]	**0.642** [0.592,0.693]	**0.306** [0.208,0.408]	**0.603** [0.539,0.664]
GOF *vs* LOF	0.63	**0.519** [0.45,0.592]	**0.818** [0.759,0.879]	**0.666** [0.619,0.712]	**0.353** [0.258,0.449]	**0.612** [0.547,0.676]
LOF *vs* non-LOF	0.64	**0.519** [0.451,0.588]	**0.815** [0.764,0.871]	**0.66** [0.618,0.703]	**0.348** [0.263,0.435]	**0.614** [0.556,0.672]

Table 5.1 Threshold-dependent performance metrics.

Metrics were derived for the probability thresholds at which the models have approximately 50% specificity on the test sets (t_{50}). Numbers in bold denote point estimates and those in brackets represent 90% bootstrap confidence intervals.

Biologically interpretable evaluation is performed by examining the models' capacity to identify emergent properties known to be associated with dominant molecular disease mechanisms (Badonyi and Marsh 2023a; Gerasimavicius, Livesey, and Marsh 2022). Specifically, we evaluate 3 properties:

1) The energetic impact of pathogenic missense mutations in the structure, approximated by the FoldX-predicted Gibbs free energy of folding (ΔΔG). Missense mutations that map to residues with a predicted local distance difference test (pLDDT) value (Jumper et al. 2021) less than 70 are excluded. This is because residues with low pLDDT are less likely to be ordered and thus ΔΔG is less interpretable for them.
2) The degree to which pathogenic missense mutations spatially cluster in the structure, approximated by the extent of disease clustering (EDC) metric. EDC is calculated as previously described (Gerasimavicius, Livesey, and Marsh 2022), however, alpha carbon atoms with a pLDDT<70 are excluded from the calculation and only proteins with at least 5 pathogenic variants after this procedure are used for the analysis. This is because pathogenic missense mutations are enriched in structured regions of proteins (Porta-Pardo et al. 2022), hence AlphaFold structures with high disorder could exhibit pseudo-clustering.
3) The performance of pathogenic missense variant prediction in a binary classification task, using two unsupervised (ESM1-v and EVE) and two supervised (MetaRNN and VARITY_R) VEPs, measured by AUROC. Only proteins with at least 5 pathogenic and 5 benign variants, and all variants mutually shared across the VEPs (that is, have predicted values), are considered for the analysis to increase per-protein AUROC confidence.

Statistical overrepresentation tests for molecular function

Functional enrichment of proteins that are subsets of predicted DN or GOF proteins is performed by comparing the protein lists against each other via PANTHER (Mi et al. 2019) using Fisher's exact test and the Gene Ontology release 2023-05-10. We consider functions with a false discovery rate <1%, a sample size (number of proteins) >50, and a fold-enrichment of >1.5. To avoid redundancy, we select a single term with the largest sample size for groups of similar terms.

5.3 Results and discussion

5.3.1 Prediction of proteins associated with non-LOF disease mechanisms using protein complex structural properties

We first reviewed the literature to identify properties of LOF genes and then trained a logistic regression model with lasso penalty using a range of diverse features, including cotranslational assembly (Bertolini et al. 2021), the functional and structural determinants investigated in this study, as well as population-level mutational constraints (Karczewski et al. 2020), evolutionary-, sequence- and interaction network-based properties (Huang et al. 2010; Shihab et al. 2017; Steinberg et al. 2015), and experimental data (Uhlen et al. 2010; M. Wang et al. 2015) (detailed in **Methods**). Measured on the test set, the classifier achieves a receiver operating characteristics area under the curve of 0.74 (**Figure 5.3A**), an F_1 score of 0.8, and a Matthews correlation coefficient of 0.24 (Chicco, Tötsch, and Jurman 2021) (detailed performance profile in **Appendix 5.1**).

Cotranslational assembly was found to be a discriminating feature in the model, ranking 12th out of 30 features and being roughly one-fourth as important as the top predictor, which is the ratio of nonsynonymous-to-synonymous substitutions (dN/dS) in the coding sequence of human relative to macaque genes (Huang et al. 2010) (**Figure 5.3B**). Notably, the second and third most important predictors in the model are transporter/channel function and the number of paralogues of the gene. It has been observed before that haplosufficient genes have higher average sequence identity to the closest paralogue than LOF genes (Huang et al. 2010), suggesting functional compensation by closely related proteins. It is possible that due to this functional redundancy, autosomal dominant genes with a high number of paralogues are simply more likely to be associated with non-LOF mechanisms. For example, ion channel genes are known to have undergone multiple gene duplication events (Liebeskind, Hillis, and Zakon 2015), which is consistent with their enrichment among DN and GOF subunits (**Appendix 3.2E**). In cyclic complexes with more than one unique subunit, paralogous copies typically sequester in the same complex (Mallik, Tawfik, and Levy 2022), suggesting that information on paralogues is a valuable proxy for non-LOF mechanisms in homomers as well as in heteromers.

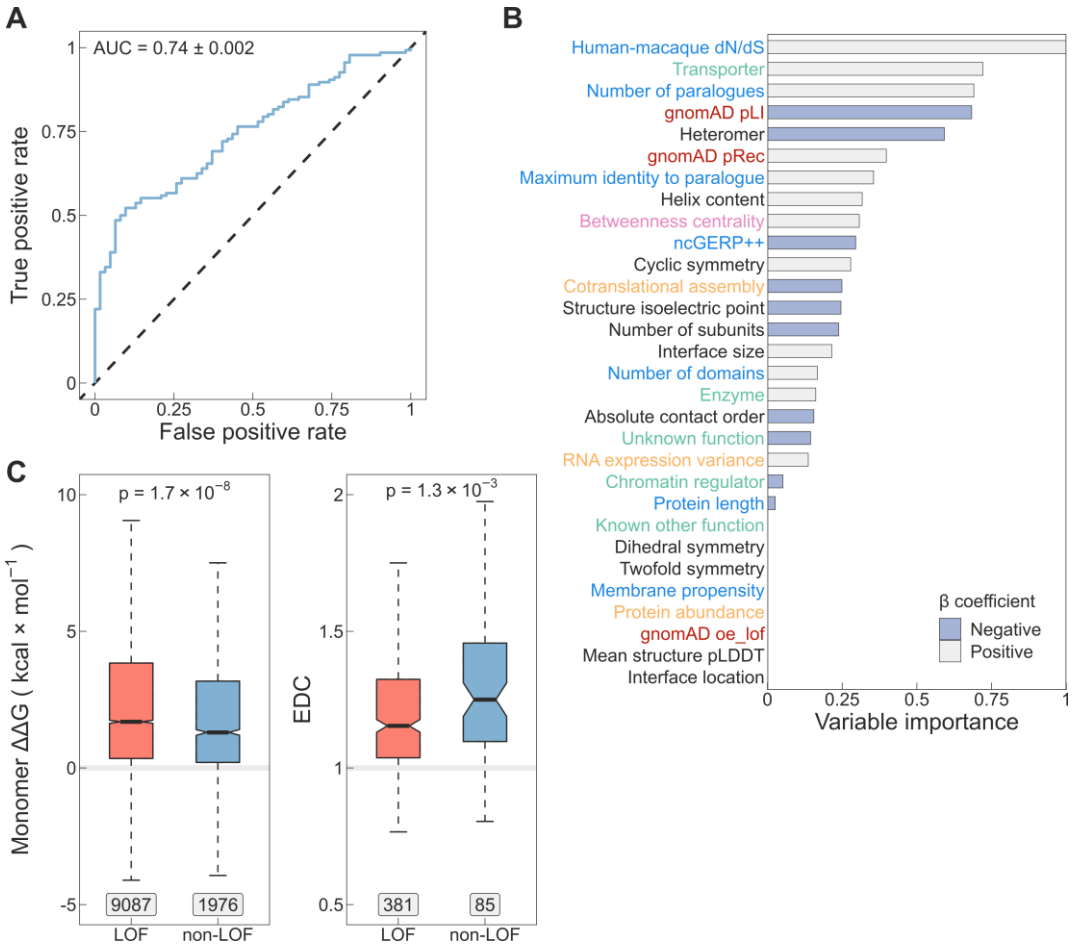


Figure 5.3 A computational model for identifying genes most likely to be associated with non-LOF molecular mechanisms.

(**A**) Receiver operating characteristic (ROC) curve of the lasso regression model measured on the test set. AUC ± bootstrap (n = 1,000) SE is shown. (**B**) Variable importance calculated as the absolute values of the β coefficients scaled to the [0,1] interval. The y-axis labels are colored according to the type of the variable: sequence-derived or evolutionary variables (blue), functional annotations (green), mutational constraint metrics (red), structural properties (black), interaction network–based property (pink), and experimental data (orange). Bars are colored based on the sign of β. (**C**) Differences in Gibbs free energy change (ΔΔG, left) and EDC (right) of pathogenic mutations between genes predicted to be non-LOF versus all other genes at threshold T2. Genes that were used for training the model as well as known AR genes were excluded. Boxes denote data within 25th and 75th percentiles, the middle line represents the median, the notch contains the 95% confidence interval of the median, and the whiskers extend from the upper and lower quartiles to a distance of 1.5 times the interquartile range. Labels indicate the number of variants (for ΔΔG) or the number of genes (for EDC) in the groups. The p-values were calculated with the Wilcoxon rank sum test.

We derived two probability thresholds (**Appendix 5.1**). The threshold of p = 0.82 (T1) was selected on the basis of the maximum value of Youden's J statistic (Youden 1950) (test set confusion matrix: 68/68 [50%] non-LOF vs 5/57 [8%] LOF). A second threshold of p = 0.92 (T2) was chosen as the value at which the specificity of the model reaches 100%, i.e. no ground truth LOF genes are classified as non-LOF at the cost of classifying more ground truth non-LOF genes as LOF (29/107 [21%] non-LOF vs 0/62 [0%] LOF). We provide predictions for 9,051 proteins covering ~44% of the proteome (available via **Science Advances**) that have at least partial structures in the PDB. Of these, 880 (9.7%) are above T2 and 3,315 (36.6%) are above T1. Of the latter, 2,840 have no dominant disease association recorded in OMIM.

As an unbiased approach to assess the model, we analysed the ΔΔG of pathogenic mutations and their extent of disease clustering (EDC) after removing genes used for training, and AR genes, whose biallelic LOF mutations would bias the trend. In **Figure 5.3C** we show the result of this analysis at threshold T2, demonstrating that missense mutations in predicted non-LOF genes exhibit a milder impact on protein structure (Gerasimavicius, Livesey, and Marsh 2022; McEntagart et al. 2016). Moreover, pathogenic variants in predicted non-LOF genes show strong 3D clustering in their respective protein structures, consistent with previous observations (Gerasimavicius, Livesey, and Marsh 2022; Lelieveld et al. 2017). In **Appendix 5.1F/G**, we provide further support that, between thresholds T1 and T2, both metrics exhibit the trend with an increasing effect size.

5.3.2 Global and local feature importance evaluation of the tripartite model

The tripartite model represents the three support vector machines built without protein complex structural and manually annotated functional features. In total, 21 interpretable features were used in the models, including properties derived from protein sequences, structures, networks, and gene mutational constraints (Badonyi and Marsh 2023). Additionally, 20 language model-based embeddings were also included, which are thought to represent protein function in their latent space (Ziegler et al. 2023). As a measure of feature importance, we calculated the loss in AUROC relative to the full model (**Figure 5.4**). While we cannot draw conclusions on why a particular ProtNLM embedding is useful for a given model, their relatively high ranks in the GOF *vs* LOF model may reflect that genes whose dominant disease mutations act via GOF are functionally more dissimilar or diverse than DN or LOF genes.

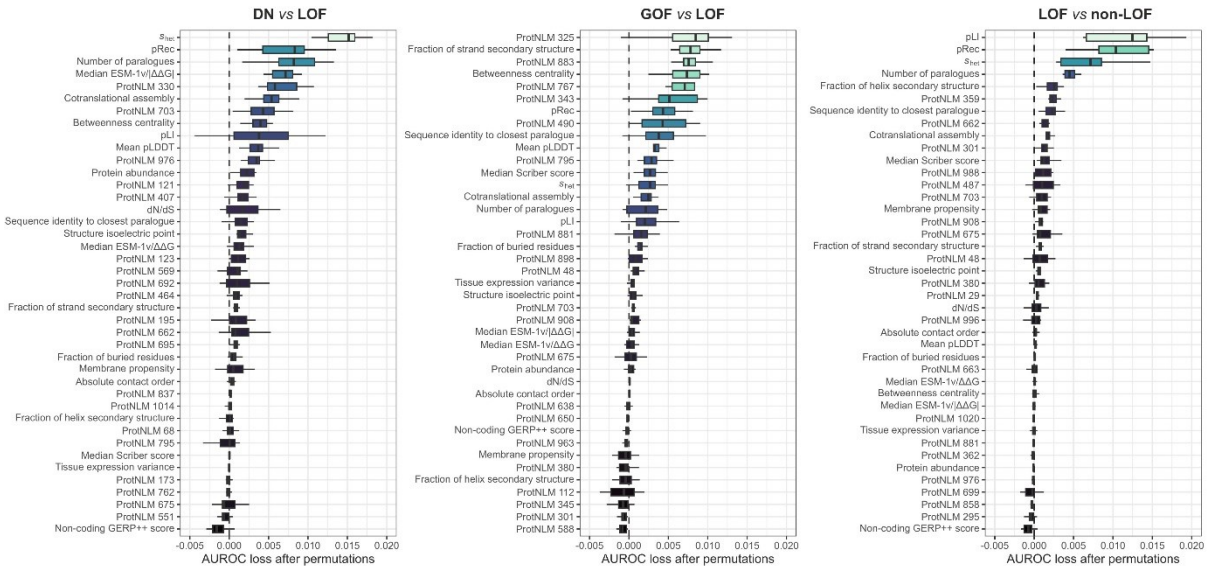


Figure 5.4 Feature importance of the models.

Feature importance is estimated by loss in AUROC upon removal of the feature in 10 permutations. Boxes denote data within 25th and 75th percentiles and the middle line represents the median. Whiskers extend from the box to 1.5× the interquartile range. Dashed line is at AUROC loss = 0.

In terms of the interpretable features, s_{het} ranks 1st and 3rd in the DN *vs* LOF and LOF *vs* non-LOF models, which emphasizes the importance of eliminating length-bias from selective constraint metrics to improve their power (Zeng et al. 2023). Moreover, the number of paralogues a protein has in the genome seems similarly important for the latter two models, ranking 3rd and 4th, respectively, consistent with our previous model

(Badonyi and Marsh 2023a). Other notable features are gnomAD metrics pLI and pRec, which is unsurprising given that genes with high pLI should be enriched in genuinely haploinsufficient genes and those with non-LOF disease mechanisms will tend to have higher pRec values, i.e. are more "recessive-like". Interestingly, the median ESM-1v/|ΔΔG| metric ranks 4th in the DN *vs* LOF model, suggesting that DN genes may possess a missense variant repertoire that is biased for functionally important but structurally less damaging effects.

We next looked at the worst and best predicted genes from the training sets of each model and calculated Shapley values for the features (Lundberg and Lee 2017). These values indicate the average contributions of different feature orderings, with positive values suggesting a tendency to predict a protein towards the primary outcome and vice versa. In the training set of the DN *vs* LOF model, the DNA-binding protein SATB2 has the lowest probability of beloning to the DN class (pDN), while the type I cytoskeletal keratin 14 (KRT14) has the highest (**Figure 5.5**). Shapely values for SATB2 suggest that primarily s_{het}, but also, for example, pLI and cotranslational assembly have contributed to its low pDN. Indeed, SATB2 has a particularly high s_{het} (0.652, which is 5.2 standard deviations above the proteome mean), a pLI of 1, and has been found to cotranslationally assemble in a recent study (Bertolini et al. 2021), consistent with a decreased likelihood of observing its role in disease through a DN mechanism (Badonyi and Marsh 2023a). In our source data, evidence for SATB2 dominant-negativity came from a report that describes a frameshift mutation in the gene likely to escape nonsense-mediated decay, whose clinical phenotype is more severe than glass syndrome, attributed to SATB2 haploinsufficiency (Boone et al. 2016). However, SATB2 is recognised by the ClinGen review panel to have sufficient evidence for haploinsufficiency (Rehm et al. 2015). Thus, it is possible that SATB2 either represents a false positive case or, due to regression to the mean, the model fails to identify it as its features are too consistent with the secondary outcome. In contrast, Shapely values for KRT14 indicate congruence with known characteristics of DN proteins, e.g., the high number of paralogues (68 in the proteome, only 2 in the training set). Furthermore, keratin disorders are considered a classical group of DN diseases (McLean and Moore 2011), reinforcing the prediction of the model.

Similar conclusions can be drawn for the other two models. AF4/FMR2 family member 4 (AFF4), implicated in CHOPS syndrome, is a GOF case in the source data with the lowest pGOF value among genes of the training set (**Figure 5.5**). Three heterozygous missense mutations in its gene were recently found to cause GOF due to decreased clearance of the protein by the ubuiquitin proteasomal system, leading to transcriptional overactivation (Izumi et al. 2015). Since transcription factors are frequently haploinsufficient (Seidman and Seidman 2002), their gene-level features in our data may generally align more with a LOF mechanism, which could hinder the identification of edge-cases like AFF4. Interestingly, the worst predicted positive case by the LOF *vs* non-LOF model is apolipoprotein E (APOE) (**Figure 5.5**), which does lack sufficient evidence for dosage sensitivity in ClinGen. On further review, evidence in our source data revealed that a heterozygous null mutation of APOE had been linked to protection against a type of amyloidosis observed in Alzheimer's disease, rather than being implicated in disease causation (J. Kim et al. 2011). Hence, the model is able to correctly down-weight false positive cases.

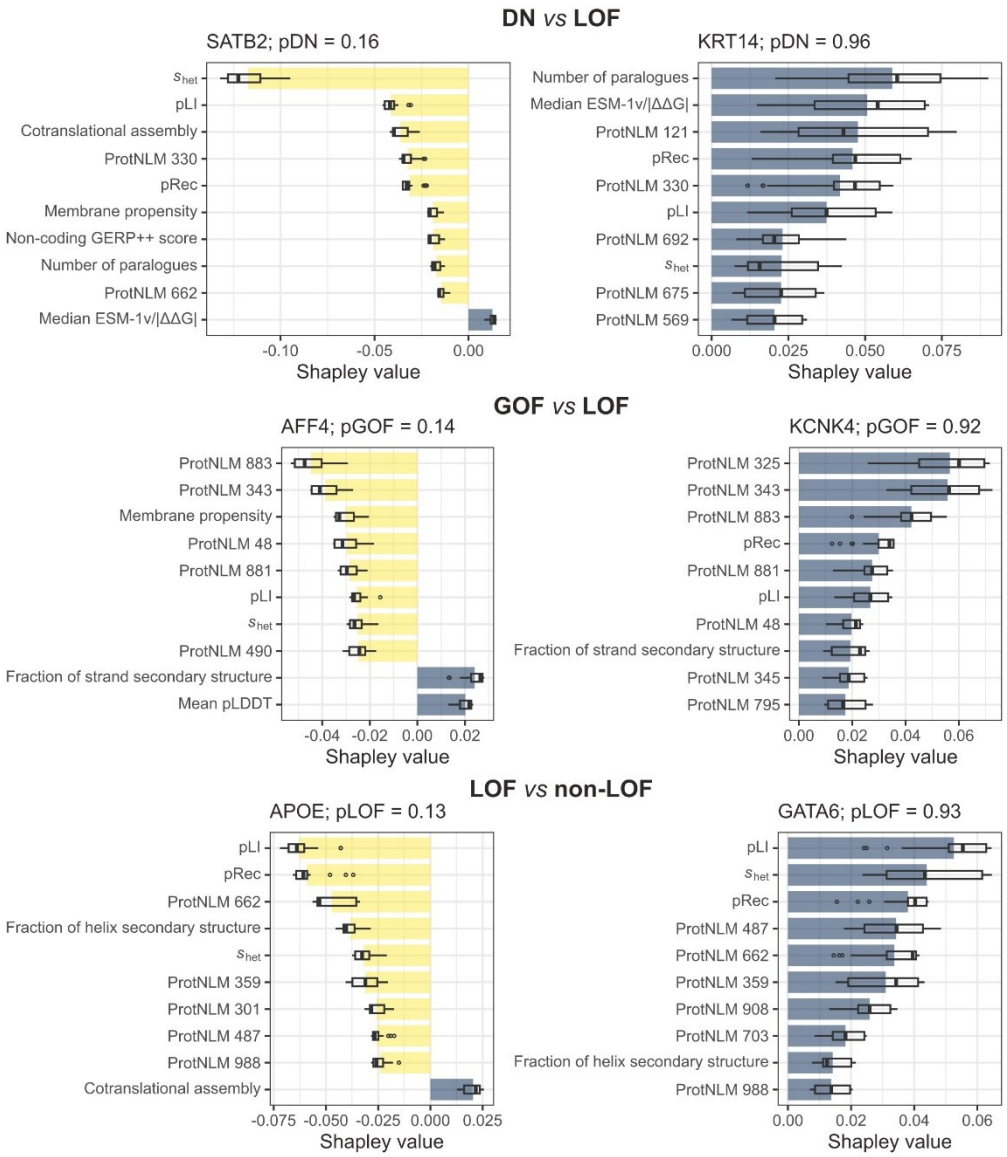


Figure 5.5 Local interpretations of model class probabilities.

Plots show the worst (left) and best (right) predicted genes of the training sets and Shapley values of the top 10 most influential features in 20 permutations. Bars show the mean, boxes denote data within 25th and 75th percentiles and the middle line represents the median. Whiskers extend from the box to 1.5x the interquartile range.

5.3.3 Proteome-scale molecular mechanism prediction

From the test sets we derived probability threshold t_{50} for each model, which represents a classification performance where approximately half of the cases belonging to the primary outcome and 80% of the secondary outcome are correctly predicted (**Table 5.1**). This threshold constitutes a relatively good balance between sensitivity and specificity while being more stringent than what is commonly considered the optimal threshold, i.e. minimum distance from the [0,1] corner of the ROC curve (red vertical line in **Figure 5.6A**) and being more lenient than the maximum positive predictive value. In **Figure 5.6B**, we show t_{50} in the context of model probability distributions for unique and overlapping cases in the test sets. Intriguingly, both the DN *vs* LOF (pDN) and the GOF *vs* LOF (pGOF) models assign higher probabilities to mixed LOF/GOF and LOF/DN classes, respectively, relative to the unique secondary outcomes, despite the fact that the models are

blind to the overlapping classification of these cases. This suggests that these genes may have intermediary characteristics or exhibit a higher false positive rate for the secondary class in the ground truth data.

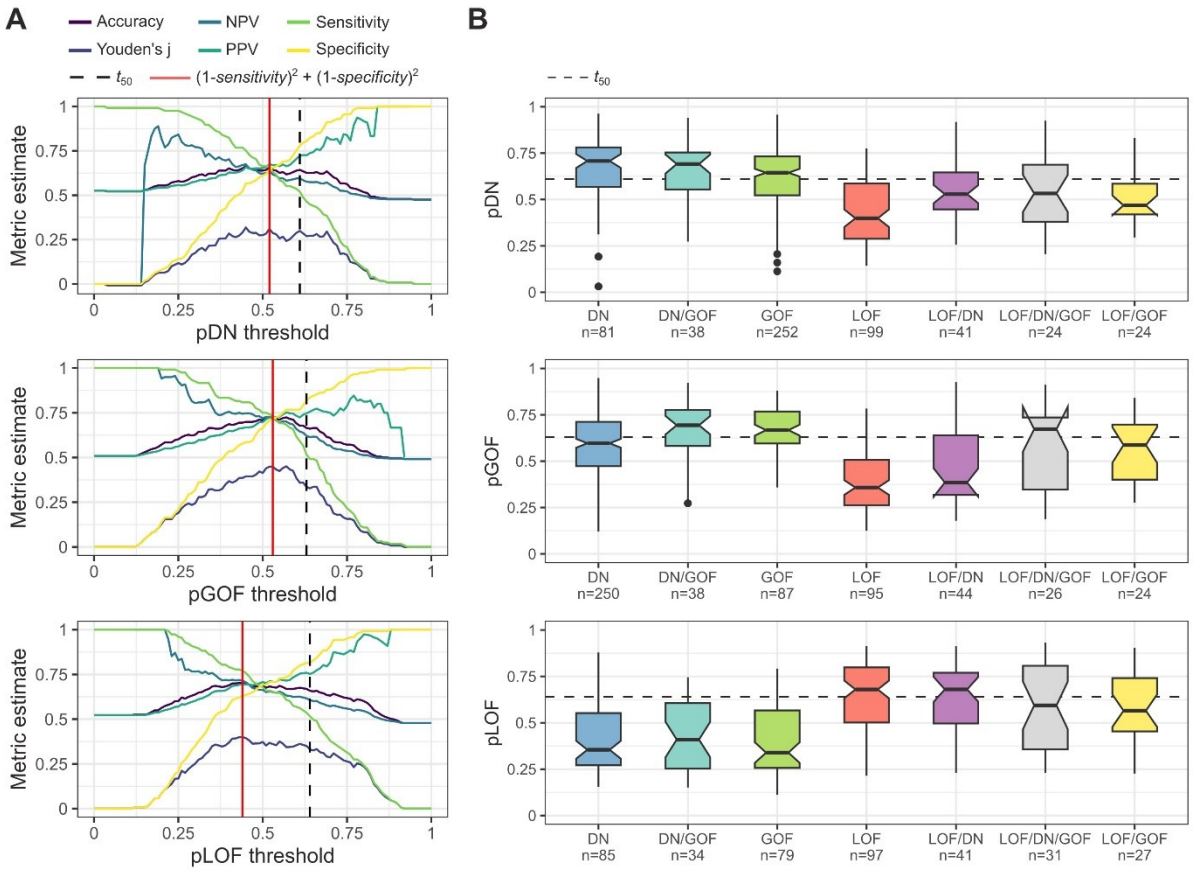


Figure 5.6 Threshold plots and test set class probabilities.

(**A**) Commonly used threshold metrics derived from the test sets for pDN, pGOF, and pLOF. NPV and PPV denote negative and positive predictive values, respectively. (**B**) Model probabilities mapped to the test sets, classified by their ground truth classes, both unique and overlapping. Sample sizes indicate the number of proteins. Boxes denote data within 25th and 75th percentiles and the middle line represents the median. Whiskers extend from the box to 1.5x the interquartile range. Dashed lines run at the t_{50} probability thresholds.

We chose a distribution-independent approach to classify the human proteome into a single molecular mechanism class. By ranking the class probabilities rather than comparing their raw values, the approach is less sensitive to differences in the models' class probability distributions. First, class probabilities are computed and ranked for all proteins and then each protein is assigned to the mechanism with the highest rank relative to the proteome, on condition that its class probability is above the t_{50} value. Based upon this strategy, in the 2023_02 UniProt reference proteome, 6,058 proteins are predicted to be DN, 5,287 GOF, and 2,580 LOF, with 6,440 proteins lacking classification. It is critical to emphasize again that the models do not predict disease involvement, i.e. the proportions should not be misconstrued as implying a twofold prevalence of DN over LOF proteins contributing to disease. The high number of predicted DN and GOF proteins may be ascribed to a higher number of paralogous proteins in these mechanism classes relative to LOF (Badonyi and Marsh 2023a), which results in larger fraction of the proteome aligning with their characteristics. When we consider only known dominant disease genes, the proportion of proteins predicted to be DN, GOF, and LOF are 27%, 28%, and 44%, respectively. We provide the class probabilities, the rank-based classification results, and labels for training set genes, for the 2023_02 UniProt reference proteome (n = 20,365).

5.3.4 Biologically and clinically relevant validation of the models

We wished to test whether the models are able to recapitulate properties of dominant molecular mechanisms we and others have previously characterised (Badonyi and Marsh 2023a; Gerasimavicius, Livesey, and Marsh 2022; Stehr et al. 2011). These properties include the energetic impact of missense mutations in the protein structure, their degree of spatial clustering, and their current predictability with state-of-the-art VEPs (see **Methods**). None of these properties were explicitly used for training, thus making them a simple and powerful mean to assess the usefulness of the predictions. We created an unbiased analysis set where we removed genes used for training and those associated with autosomal recessive inheritance. We first looked at the predicted ΔΔG of pathogenic missense mutations across the different predicted classes in this dataset (**Figure 5.7A**).

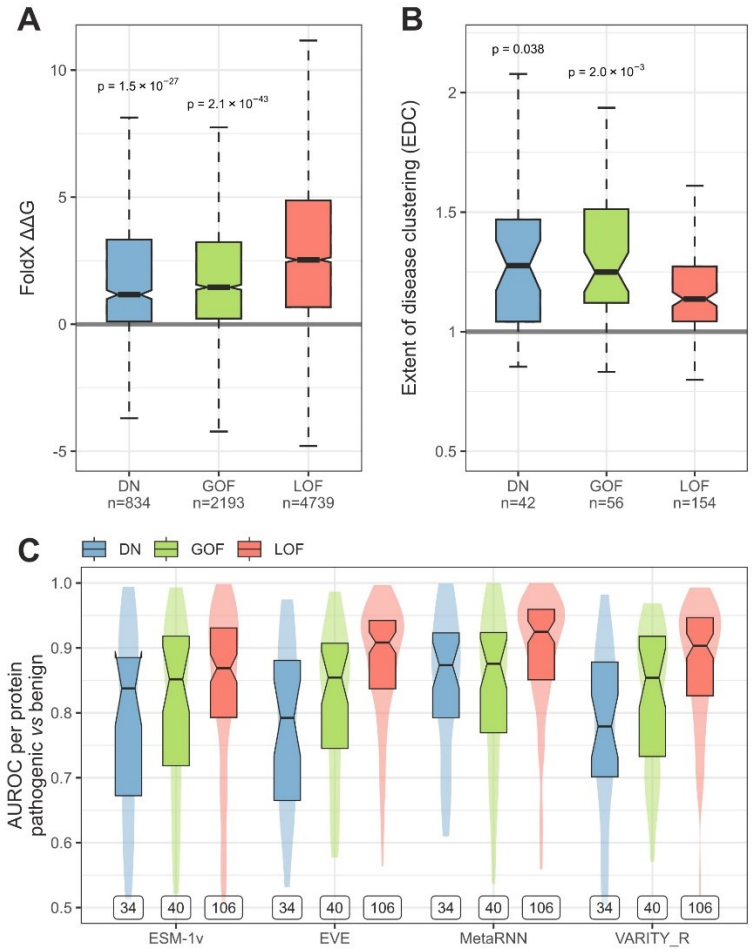


Figure 5.7 Validation of the models through model-independent metrics on an unbiased analysis set. (**A**) FoldX-predicted ΔΔG of pathogenic missense mutations. Numbers below classes denote the number of mutations. Holm-Bonferroni corrected p-values above DN and GOF boxes are relative to the LOF group and were determined by one-sided Wilcoxon rank-sum test. Sample sizes indicate the number of variants. (**B**) Class probabilities of the analysis set *vs* EDC. Holm-Bonferroni corrected p-values above DN and GOF boxes are relative to the LOF group and were determined by one-sided Wilcoxon rank-sum test. Sample sizes indicate the number of proteins in each class. (**C**) Aggregated AUROC analysis of pathogenic *vs* benign variants in predicted molecular mechanism classes. Labels indicate the number of proteins in each class. Boxes denote data within 25th and 75th percentiles and the middle line represents the median. Violins show area-normalized distributions.

The results suggests mutations in DN and GOF proteins are significantly less damaging than in LOF proteins. This can be explained by the idea that destabilisation is one of the signature mechanisms of LOF mutations (Gerasimavicius, Livesey, and Marsh 2022). By contrast, GOF mutations should not be too damaging in order to alter a function, and most DN and many GOF genes are in fact assembly-mediated (Backwell and Marsh 2022). Thus, the effect of their mutations should not preclude protein complex assembly, which makes them necessarily less damaging.

Next, we looked at how much pathogenic missense mutations cluster in the structures of proteins predicted to be DN or GOF. Our prior expectation is that both classes should exhibit higher degree of clustering than LOF, as measured by the EDC metric. We found this assumption to hold up, with both DN and GOF proteins having significantly higher clustering values than LOF proteins (**Figure 5.7B**). This observation agrees with the concept that LOF mutations are generally more sparsely distributed in the protein structure, but non-LOF mutations tend to be concentrated at protein interfaces and functional sites (Gerasimavicius, Livesey, and Marsh 2022). Lastly, we took advantage of an important bottleneck of contemporary VEPs, which is that they less well predict pathogenic missense mutations associated with DN and GOF genes (Gerasimavicius, Livesey, and Marsh 2022). In **Figure 5.7C**, we show a per-protein aggregated AUROC analysis of pathogenic *vs* benign missense variants, evaluated by two unsupervised (ESM1-v and EVE) and two supervised (MetaRNN and VARITY_R) VEPs, which we recently showed to be the top performing VEPs of their category (Livesey and Marsh 2023). Expectedly, missense mutations in proteins classified as LOF are much better predicted than those in DN or GOF proteins.

Importantly, the above trends are observed with the raw rank-based classification results without the t_{50} probability cutoff (**Appendix 5.2**), demonstrating that they are not an artefact of a careful threshold selection. We conclude that the models effectively reproduce biologically and clinically relevant properties of dominant molecular mechanisms.

5.3.5 The functional landscape of predicted DN and GOF proteins

Given that pDN and pGOF class probabilities are positively correlated by design (**Figure 5.8**), our focus was on discerning the differences between proteins predicted exclusively as DN or GOF. We therefore conducted statistical overrepresentation tests to explore the molecular functions associated with these proteins. It is well-established that DN effects are common in homomers, which are proteins that form complexes with copies of themselves (Badonyi and Marsh 2023a; Veitia 2007). Correspondingly, we identified functions closely related to homomeric symmetry groups in exclusively DN proteins, including “organic cyclic compound binding”, “oxidoreductase activity”, and “hydrolase activity” (Bergendahl and Marsh 2017). Interestingly, a strongly enriched, large group of proteins possess “nucleic acid binding” function, potentially indicating a higher prevalence of DN effects among transcription factors and other DNA/RNA-binding proteins. This underscores the importance, and difficulty, of distinguishing between mutations that act through DN effects versus haploinsufficiency. In contrast, exclusively GOF proteins exhibit functions less directly associated with protein complexes but more susceptible to the impact of overactivation events. These functions include “molecular transducer activity”, “signaling receptor activity”, “kinase activity”, “enzyme regulator activity”, and “phosphotransferase activity”.

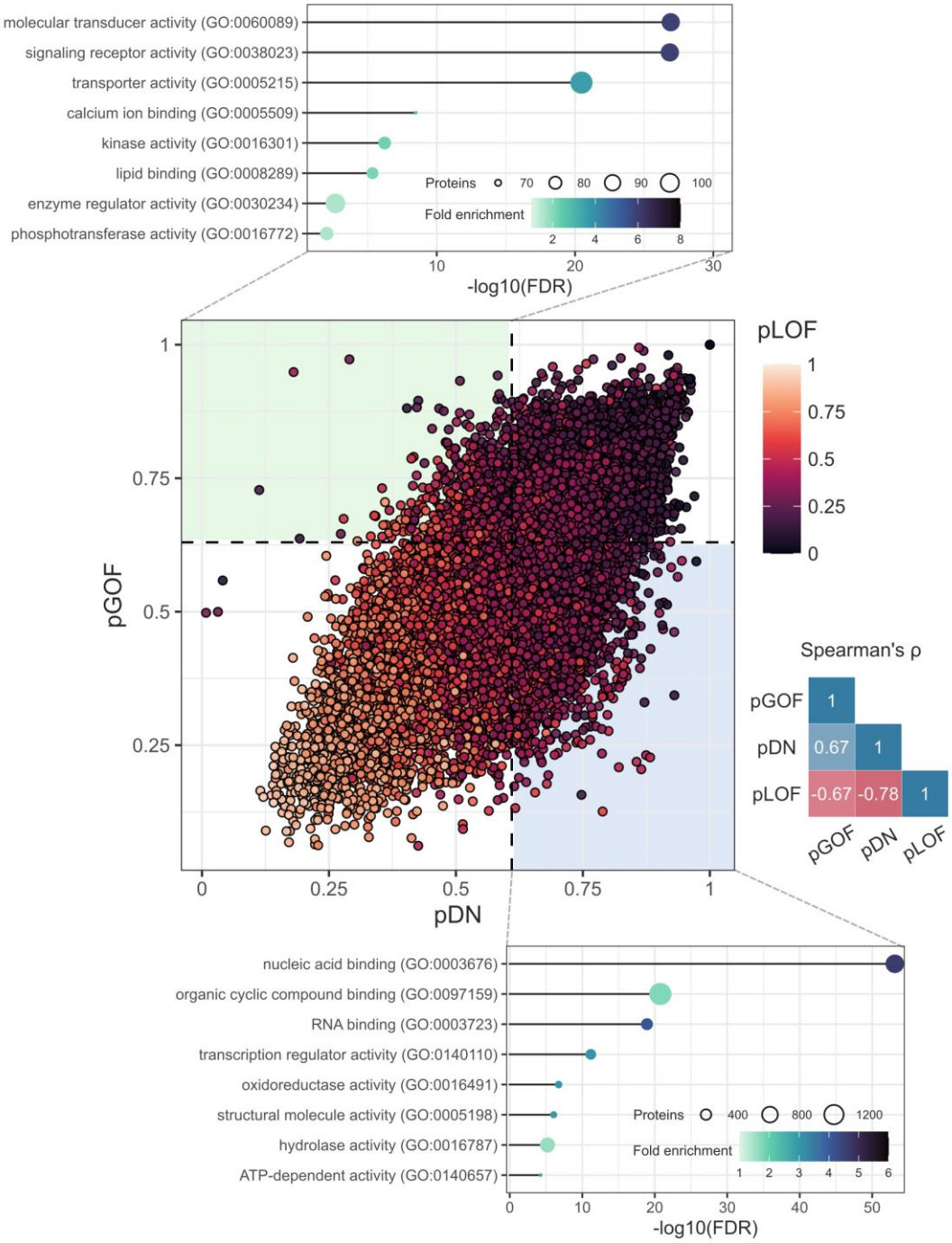


Figure 5.8 The functional landscape of proteins predicted to be exclusively DN or GOF.
Middle: pDN vs pGOF scatter plot coloured by pLOF. Correlation triangle on the right shows Spearman correlations of the class probabilities. *Top*: The lollipop chart shows the most enriched molecular functions for genes predicted to be GOF below the t_{50} threshold for pDN (exclusively GOF). *Bottom*: Likewise, the chart shows the most enriched functions for DN genes below the t_{50} threshold for pGOF (exclusively DN). The *x*-axis shows the negative $\log_{10}$ false discovery rate.

Altogether, these findings reinforce that the predictions are consistent with known and expected properties of non-LOF molecular mechanisms. Thus, they can be used to test hypotheses about the functional attributes of the proteins associated with the molecular mechanisms, which could facilitate the discovery of novel features for the next generation of VEPs.

5.4 Conclusion

We constructed three binary classifiers to predict DN, GOF, and LOF dominant molecular disease mechanisms by substituting structural and functional features of protein-coding genes with limited coverage. These models have a similar performance to our previous non-LOF vs LOF classifier, which was based upon such limited but interpretable features. We combined the models' output into a single class prediction using a proteome rank-based approach and examined the properties of proteins assigned to the different molecular mechanisms. Our analyses provide strong evidence that the models reproduce the properties linked to molecular mechanisms at systems level. Notably, our findings underscore the predictability of pathogenic mutations in DN and GOF proteins – a problem recognised to be more challenging than in LOF proteins. This observation was consistently upheld in our unbiased analysis set, further stressing the need for diversifying the predictive capabilities of future VEPs. We have made available the predictions for the human reference proteome, and we hope researchers will use the data to prioritise novel or challenging variants and test hypotheses about the functional roles fulfilled by these proteins. It is important to emphasize that the models do not predict the likelihood of disease involvement by a gene. Instead, a non-disease associated gene with a predicted mechanism implies that the protein's properties are most consistent with the given mechanism, which we expect will be particularly useful for researchers seeking to determine the molecular mechanisms underlying novel disease mutations.

6 Summary and future directions

There are several conclusions relevant to both biology and translational medicine that can be drawn from the studies presented in this thesis. In this final chapter, I will highlight a few of these that I consider particularly important or interesting and discuss how they may lead to further investigation or be used to help solve outstanding challenges.

In light of what we had established about the assembly of protein complexes solely by using interface size for inference, it is not at all surprising that I have found cotranslationally assembling subunits to have larger interfaces. A large interface has an inherent drive to assemble early due to the many yet unburied contacts, whose interface with water is less favourable than that with the partner subunit. It is surprising to observe, however, that multi-interface heteromers tend to expose larger interfaces on the N terminus, reflecting that the interface hierarchy is shaped by the process of translation. Perhaps this trend is suggestive of the adaptive mechanisms occurring over the course of evolution to increase assembly efficiency and confer incremental fitness benefit to the cell. We do not know what proportion of these evolutionary steps arise fortuitously, offering no benefit at first but then becoming essential via entrenchment. Although addressing this is a non-trivial task, as we do not have the luxury to exactly replicate historical systems in which these changes first occurred, case studies employing ancestral sequence reconstruction and constructive biochemistry can at least give us an idea about which process is the default. We do not know either if large N-terminal interfaces are a signature of the widespread sequential cotranslational assembly of heteromers in the cell, although I have presented some evidence and intuitive arguments for this prediction. I believe this hypothesis will be simple to test once the proteome-scale detection of sequential cotranslational assembly has become possible.

Analysis of the inheritance of Mendelian disease genes and their likely molecular mechanisms has revealed that protein complex subunits with dominant-negative disease mutations cotranslationally assemble at a lower frequency than other disease-linked subunits. A suitable explanation of this is that cotranslational assembly spatiotemporally separates the assembly of wild-type and mutant subunits in the cell, essentially preventing the mutant from poisoning functional complexes. Under this assumption, it is easy to see how C-terminally exposed interfaces, which lessen the likelihood of cotranslational assembly, would be associated more often with the dominant-negative effect. I find it thought-provoking that proteins with recessively inherited gene mutations have a high level of cotranslational assembly. In the introduction of this thesis, I listed popular adaptive arguments for the emergence of quaternary structure, including enhancing substrate specificity, increasing the frequency of substrate encounters, and providing opportunities for allosteric activity regulation in enzymes. Could cotranslational assembly represent a safety factor against dominant-negative and positive mutations contributing to the persistent recessivity of enzymes? If so, is it driven by an adaptive process or is it a necessary consequence of their quaternary structure topology biased for dihedral symmetry? Considering that many core metabolic enzymes had evolved to a dihedral state in monoploid prokaryotes, it may well be that this benefit of cotranslational assembly is merely an epiphenomenon. It is my hope that our work will motivate evolutionary biologists to look further into this puzzle.

I mentioned in the introduction that a common theme of complexes prone to assembly-mediated effects may be their reliance on the combined efforts of all subunits to achieve their intended function, rather than each subunit functioning on its own. A simple proxy for function is ligand binding, which can often be “multi-chain”

binding, i.e. two subunits having residues located at the same ligand interface. Multi-chain ligand binding residues in homomers should be more sensitive to assembly-mediated effects, because these residues are functionally contributed by two different alleles. In preliminary results, I have found support for this phenomenon, in that disease-linked homomers tend to have a higher fraction of multi-chain ligand binding residues. This relates to the likelihood of a random process, such as mutation, being more likely to affect these sites by chance in these complexes. A key confounder to control for is interface size, because homomers with multi-chain ligand binding sites are more likely to represent cases where complex topology has evolved to support function and therefore more likely to possess large, biologically significant interfaces. If this result held up to all relevant confounders, we could envisage a feature derived from multi-chain binding residues that could be important in future variant-level predictors of mutation effects.

My research exploring the properties of proteins with dominant-negative, and generally non-LOF, disease mutations has culminated in a practical application. I developed a relatively simple but versatile model to distinguish between proteins whose mutations tend to act via one of the dominant molecular mechanisms. To illustrate how the proteome-scale mechanism prediction could enable variant prioritisation, consider the SG10K_Health data (SG10K), which is the output of the Singapore National Precision Medicine programme, comprised of 10,000 whole-genome sequences from healthy Chinese, Indian, and Malay individuals. If we hypothesise that there are dominant-negative variants in SG10K, much like p.Glu504Lys in ALDH2 with an allele frequency of 8% in Asian populations, how do we find them? I performed an analysis on missense variants predicted to be pathogenic (VARITY_R score >0.95) and at the same time predicted to have a relatively mild impact on protein structure ($|\Delta\Delta G| < 2$), i.e. prime candidates for dominant-negative effects (DN candidates). To identify significantly enriched variants in SG10K, I calculated for every disease-associated gene the fraction of the population that carries a putatively dominant-negative allele and compared it to gnomAD, which is biased towards European ancestry. **Figure 6.1** shows the end-point of this analysis, where I contrast the pDN (probability of the gene being associated with a dominant-negative mechanism) of DN candidates to the background and find that the former indeed have higher pDN on average.

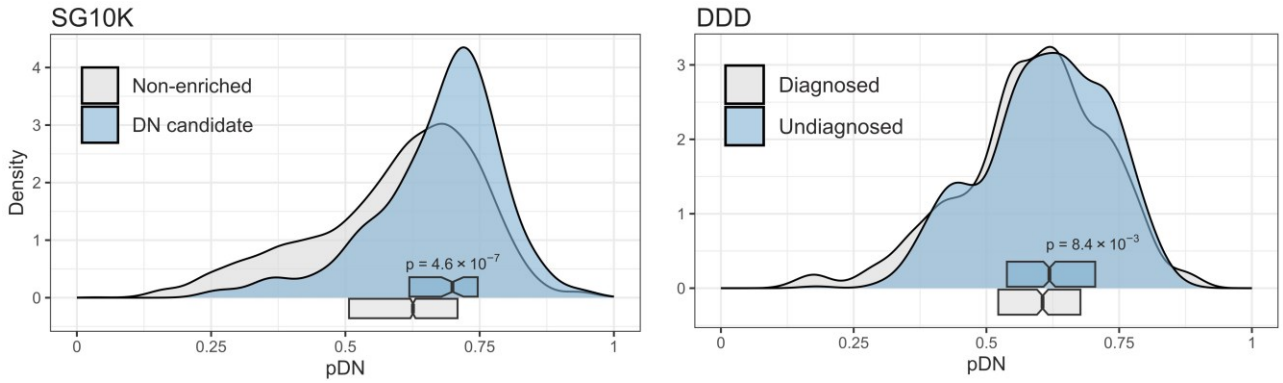


Figure 6.1 Clinically relevant use-cases of the DN *vs* LOF model.

Left: Disease genes with significantly enriched variants with dominant-negative properties in the SG10K population (DN candidate) have higher pDN than the background. *Right*: De novo mutations in recessive genes of undiagnosed developmental disorder patiens are more likely to map to genes with higher pDN.

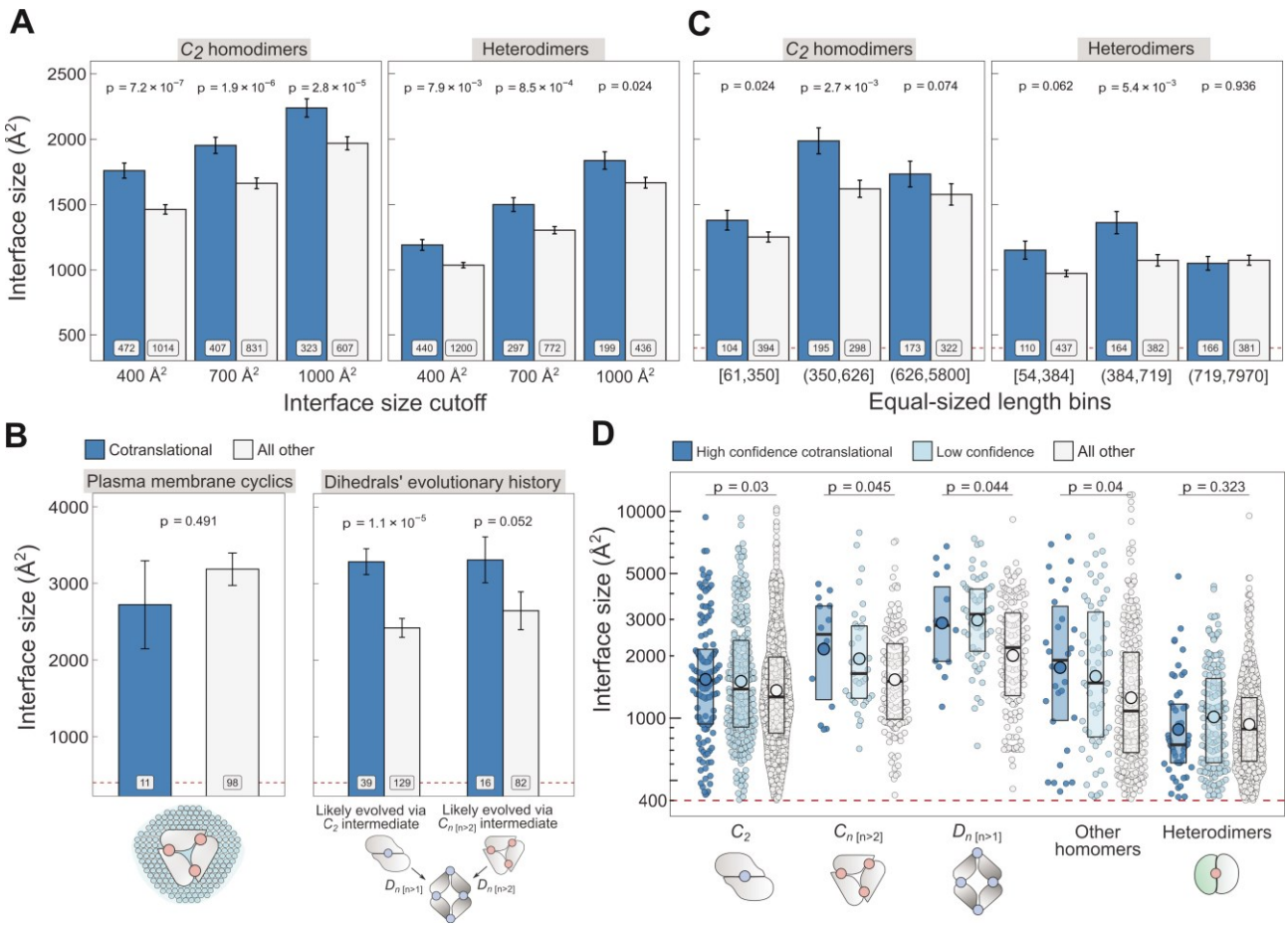
Encouragingly, one of the proteins pulled down by the analysis is ALDH2, lending confidence to the strategy. Our hope is that a similar screen can help decide which variants are worth studying further by combined

laboratory and computational techniques, increasing the chance of finding clinically or epidemiologically relevant variants.

Finally, at multiple junctions I eluded to the idea that dominant-negative mutations in recessive genes may be an overlooked cause of disease, especially developmental disorders (DD). When de novo mutations occur in recessive genes, they are often ruled out to be disease-causing as the gene is *thought to be* robust to the effect of dominant mutations. Although this may be the case with simple LOF mechanisms, it does not exclude a dominant-negative effect. I would like to estimate the contribution of cryptic dominant-negative mutations to DD, by means yet to be determined. However, for over 10,000 patients with trio-sequencing data (Kaplanis et al. 2020), we obtained diagnosed/undiagnosed proband labels from a recent Deciphering Developmental Disorder Consortium study (C. F. Wright et al. 2023). An analysis of genes with biallelic contribution to DD suggests that a randomly selected de novo mutation is more likely to come from a gene with high pDN in undiagnosed patients relative to diagnosed patients (**Figure 6.1**), consistent with the hypothesis. Even if the overall dominant-negative burden turns out to be relatively small, providing a potential diagnosis can make an immeasurable impact in the patients' lives.

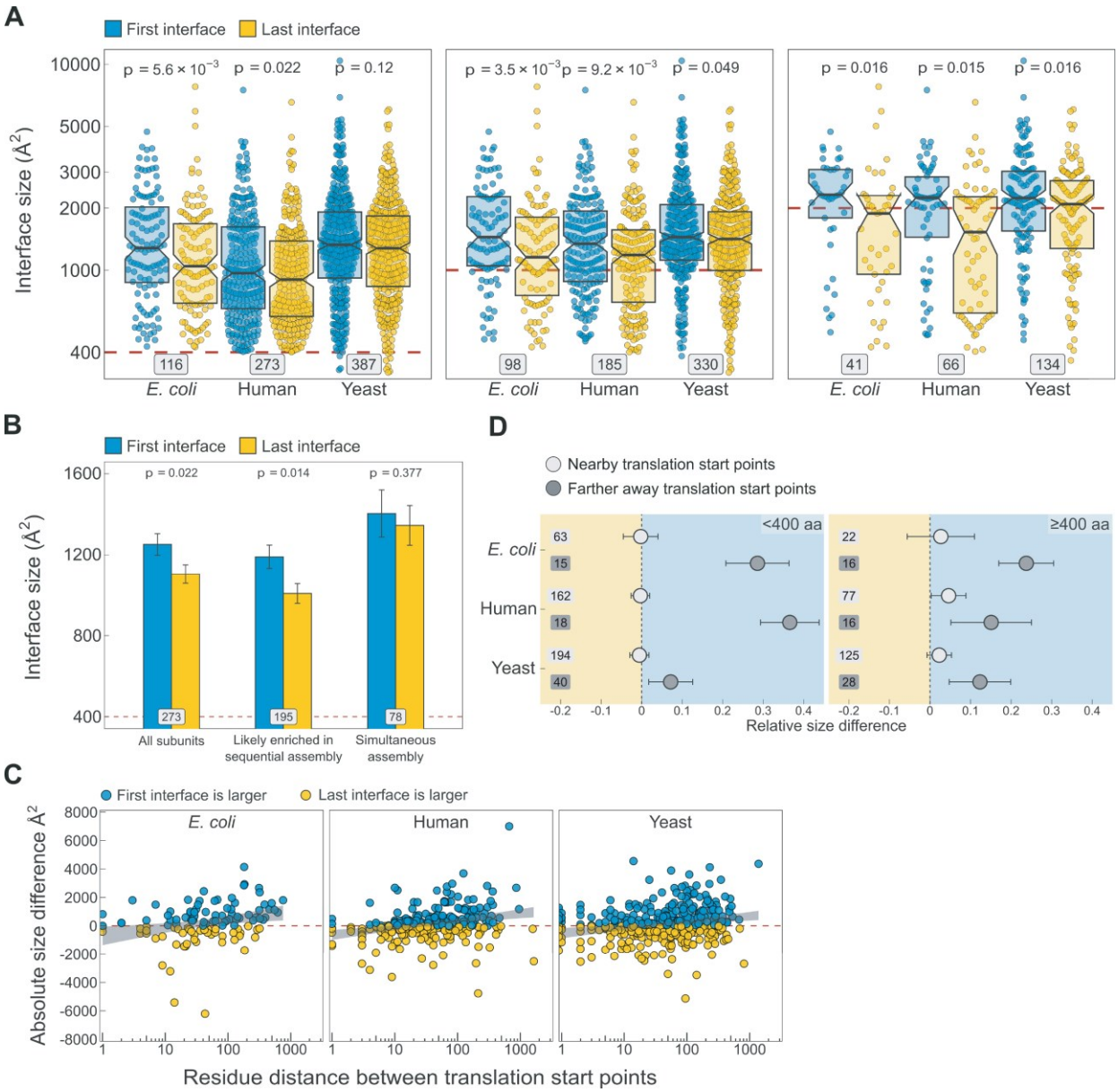
The value of this thesis is in highlighting the importance of protein complex assembly in human disease. We are fortunate to live in a time when high-throughput experiments and complex statistical methods joined with increasing computational power are accelerating our knowledge about structural biology faster than ever before. This knowledge will continue to draw attention to protein complexes as it is their assembly that is often the key to understanding their function. While much remains to be accomplished, the rapidly improving field of protein complex and structure prediction will be fundamental to exploring molecular disease mechanisms other than simple loss of function.

Appendix



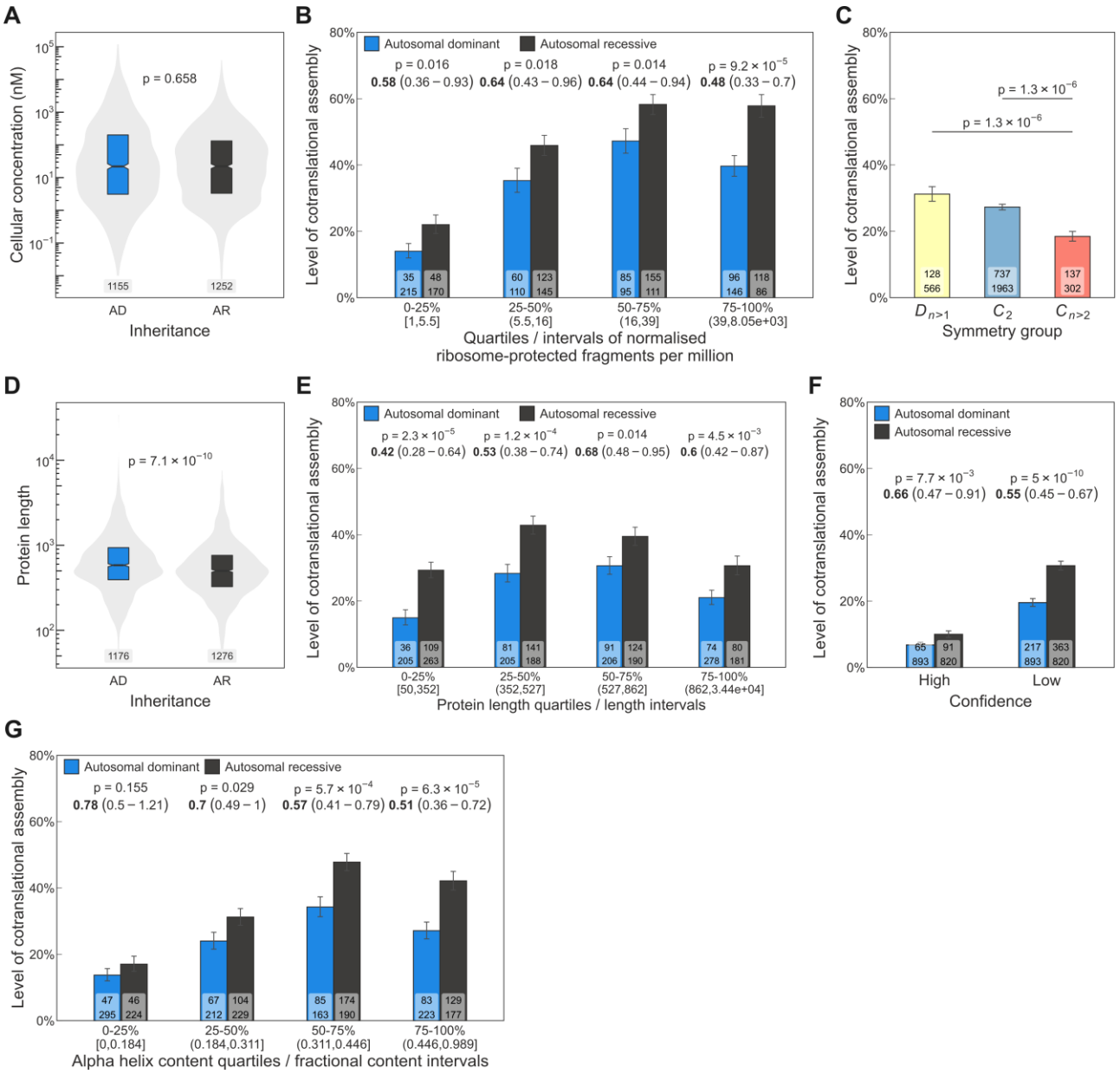
Appendix 2.1 Controlling for potential confounders of the cotranslational assembly data.

(**A**) Interface size differences between cotranslationally assembling and all other subunits of C_2 homodimers and heterodimers, measured at incremental interface area cutoffs. Error bars represent standard error of the mean (SEM) and labels show the number of proteins in each group. The p-values were calculated with two-sided Wilcoxon rank-sum tests. (**B**) *Left*: Interface size differences between cotranslationally assembling and all other subunits of plasma membrane localised higher-order cyclic symmetry members. *Right*: Interface size differences between cotranslationally assembling and all other subunits of dihedral complexes grouped by their probable evolutionary history. The p-values were calculated with two-sided Wilcoxon rank-sum tests. (**C**) Interface size differences between cotranslationally assembling and all other subunits of C_2 homodimers and heterodimers, binned into three approximately equal-sized bins of sequence length. Error bars represent SEM and labels show the number of proteins in each group. The p-values were calculated with two-sided Wilcoxon rank-sum tests. (**D**) Interface size differences subset by confidence in cotranslational assembly. Only proteins with cytoplasmic and nuclear localisations are included. Pictograms show the basic structure of symmetry group members, with the blue dots representing isologous and red dots representing heterologous and heteromeric interfaces. The p-values were calculated with two-sided Wilcoxon rank-sum tests between high confidence and all other subunits. Differences between high confidence and low confidence subunits are not significant. Larger dots within boxes represent the sample mean.



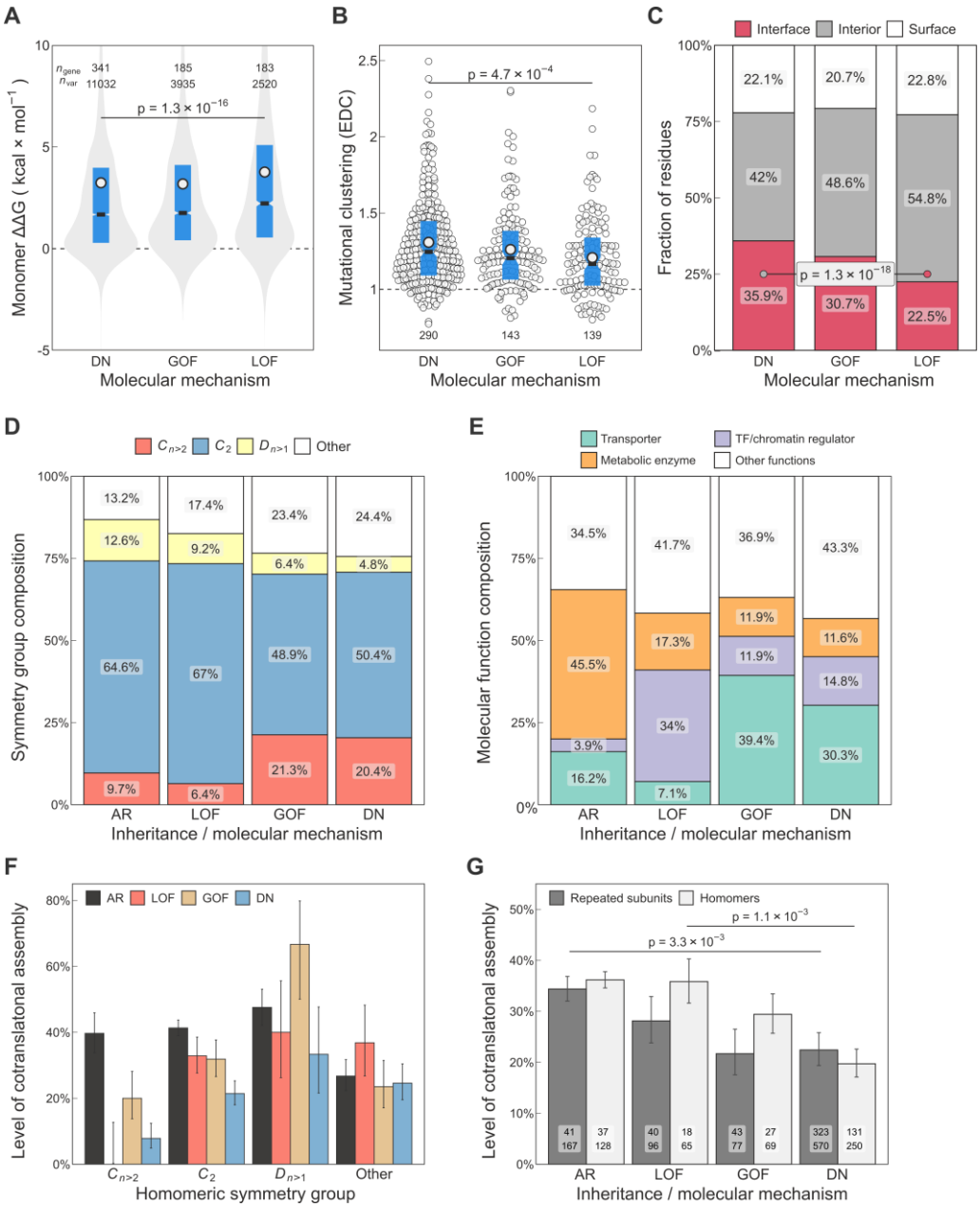
Appendix 2.2 Additional analyses supporting the results shown in Figure 2.5.

(**A**) Full distribution of the first versus last translated interface size differences for all subunits, subunits with at least one interface >1,000 Å², and subunits with at least one interface >2,000 Å², in the first, second, and third panels, respectively. (**B**) Interface size differences between the first and last translated interfaces in human multi-interface heteromeric subunits, with those identified to have simultaneously forming interfaces shown as a separate group. Error bars are standard error of the mean (SEM) and labels on bars show the number of proteins in each group. The p-values were calculated with Wilcoxon signed-rank tests. (**C**) Scatter plots showing the absolute distance in amino acids between translational start points of the first and last translated interfaces and the absolute area difference across species. Shaded lines represent the 95% confidence interval of the regression line. (**D**) TIE-fighter plots demonstrating that interface separation increases the area difference in favour of the first interface independent of protein length. For all species, the relative translational distance interval was split at the mean, and the plot is divided into less and more than 400 amino acid long sequences. Dots represent the mean and error bars are SEM. Labels are the number of proteins in each group. Background colour reflects the direction of the size difference: blue – first interface larger, yellow – last interface larger.



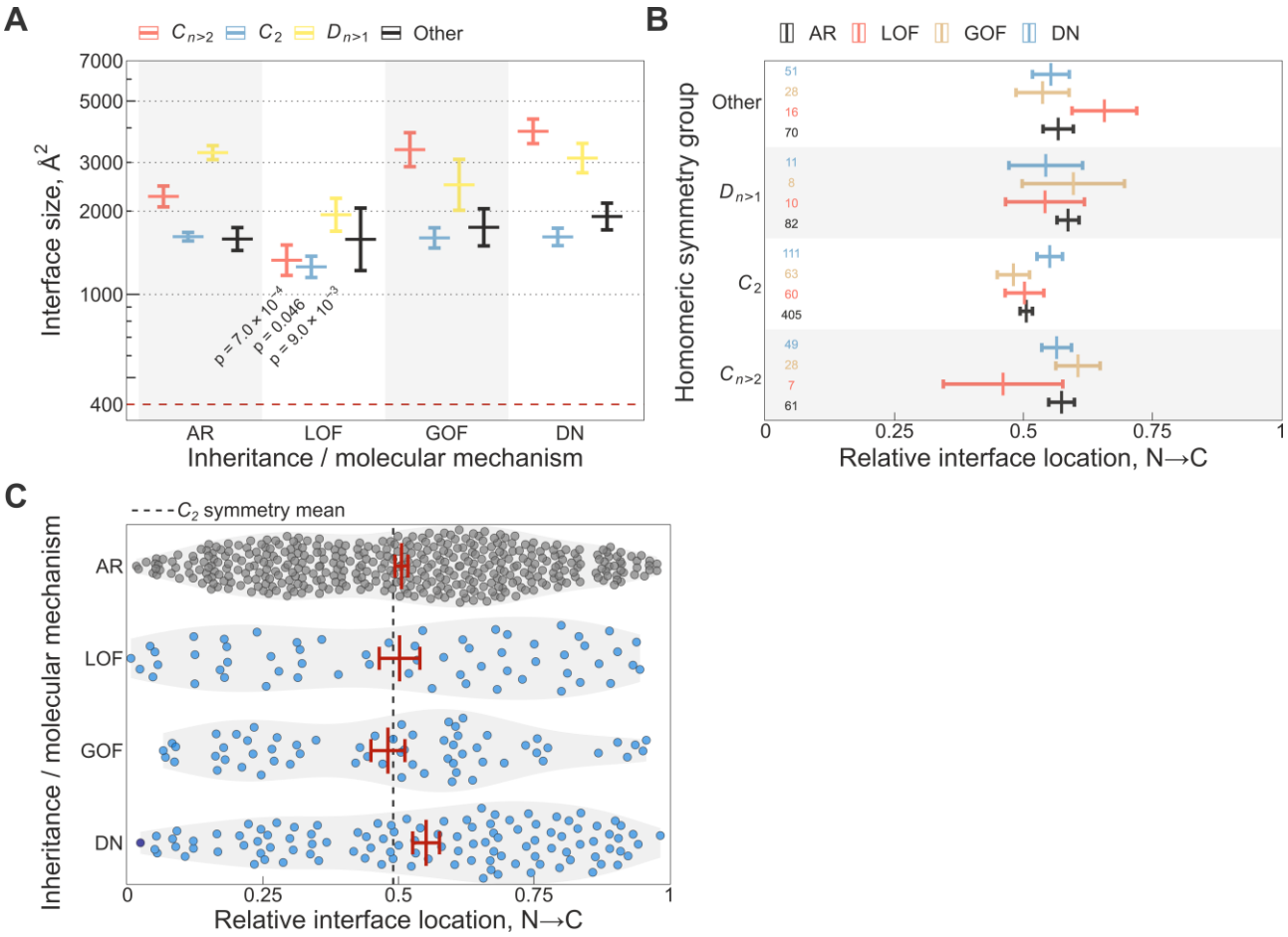
Appendix 3.1 Controls of potential confounders of the inheritance-level analysis.

(**A**) Box-violin plot comparison of the abundance distribution of AD and AR homomers and repeated subunits. Boxes denote data within 25th and 75th percentiles, the middle line represents the median and the notch contains the 95% confidence interval of the median. Numbers are sample size and the p-value was calculated with the Wilcoxon rank-sum test. (**B**) The level of cotranslational assembly among AD vs AR genes binned into quartiles of active ribosome protected fragment counts measured in HEK293 cells. Each bin corresponds to 25% of proteins by count and the fragment per million intervals are displayed in brackets. Bar values are per cent level of cotranslational assembly, error bars are Jeffrey's 68% binomial credible intervals. The p-value from the hypergeometric test and the odds ratio (in bold) and its 95% confidence interval is shown above the bars. Labels on bars are the count of cotranslationally assembling subunits (top) and all other subunits (bottom). Panels **E-G** have the same parameters. (**C**) Level of cotranslational assembly in homomeric symmetry groups. (**D**) Box-violin plot comparison of the length distribution of AD and AR subunits. Numbers at the bottom represent sample size and the p-value was calculated with the Wilcoxon rank-sum test. (**E**) The level of cotranslational assembly binned by protein length. (**F**) The level of cotranslational assembly grouped by the confidence in their identification. (**G**) The level of cotranslational assembly binned by fractional helix content.



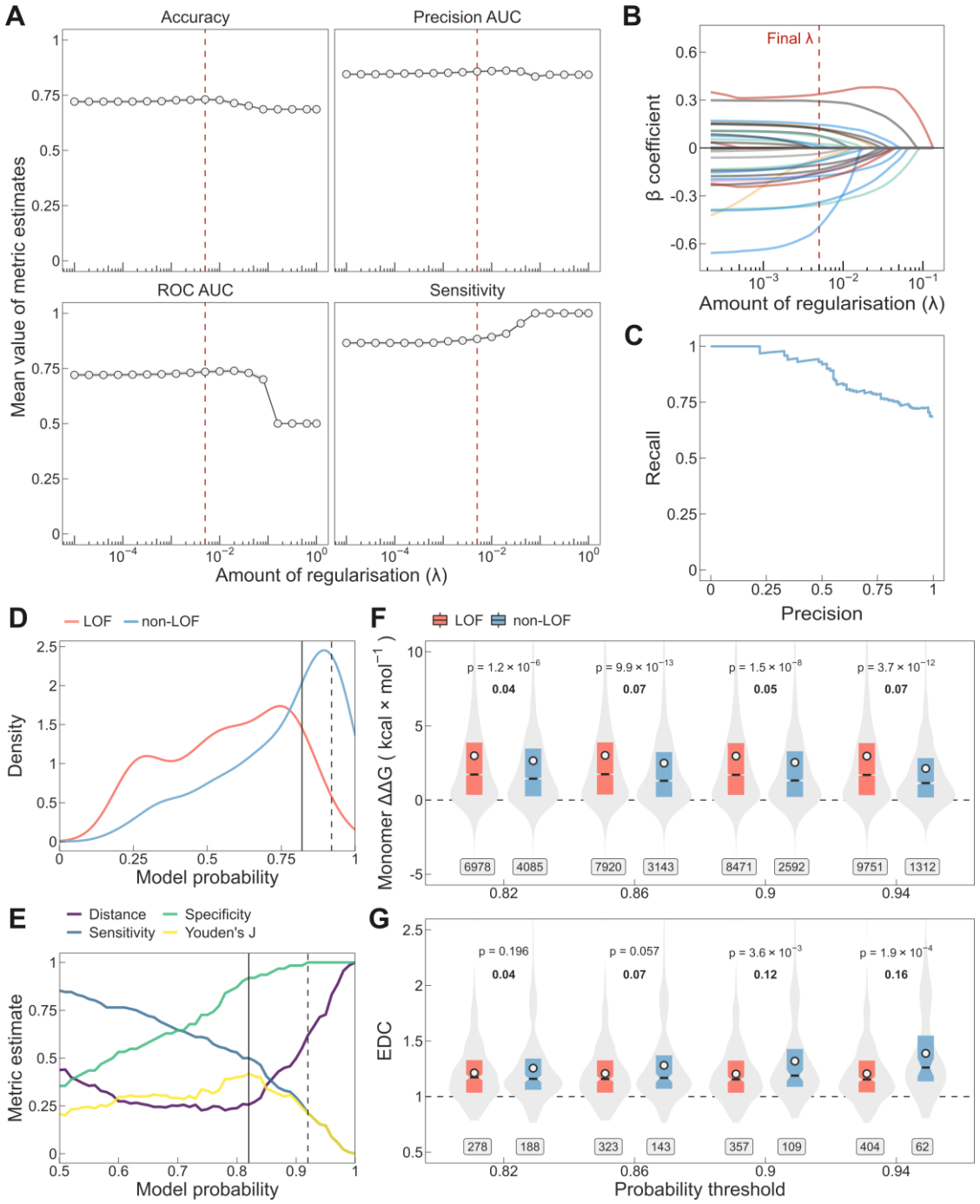
Appendix 3.2 Controls of potential confounders of the molecular mechanism-level analysis.

(**A**) Box-violin plot comparison of the predicted ΔΔG of pathogenic mutations in homomers and repeated subunits, grouped by molecular mechanisms. Boxes denote data within 25th and 75th percentiles, the middle line represents the median, the notch contains the 95% confidence interval of the median and white dots are the mean. Numbers on top are sample sizes for genes (n_{gene}) and missense variants (n_{var}) within the groups. The p-value was calculated with the Wilcoxon rank-sum test. (**B**) Box-beeswarm plot comparison of the extent of disease clustering (EDC) metric that measures the extent to which pathogenic mutations cluster in 3D space. Numbers show the number of genes in each group. The p-value was calculated with the Wilcoxon rank-sum test. (**C**) Stacked bar chart showing the interface residue enrichment of missense pathogenic mutations in the DN group relative to LOF. The p-value was calculated with the hypergeometric test. (**D**) Stacked bar chart of the symmetry group composition of homomeric subunits with different inheritance and molecular mechanisms. (**E**) Stacked bar chart of the molecular function composition of homomers and repeated subunits with different inheritance and molecular mechanisms. (**F**) Level of cotranslational assembly within the different inheritance and molecular mechanism classes subset by homomeric symmetry groups. Bar values are per cent level of cotranslational assembly, error bars are Jeffrey's 68% binomial credible intervals. (**G**) Level of cotranslational assembly split into homomers and repeated subunit heteromers. Bar values are per cent level of cotranslational assembly, error bars are Jeffrey's 68% binomial credible intervals. The p-values were calculated from a hypergeometric test.



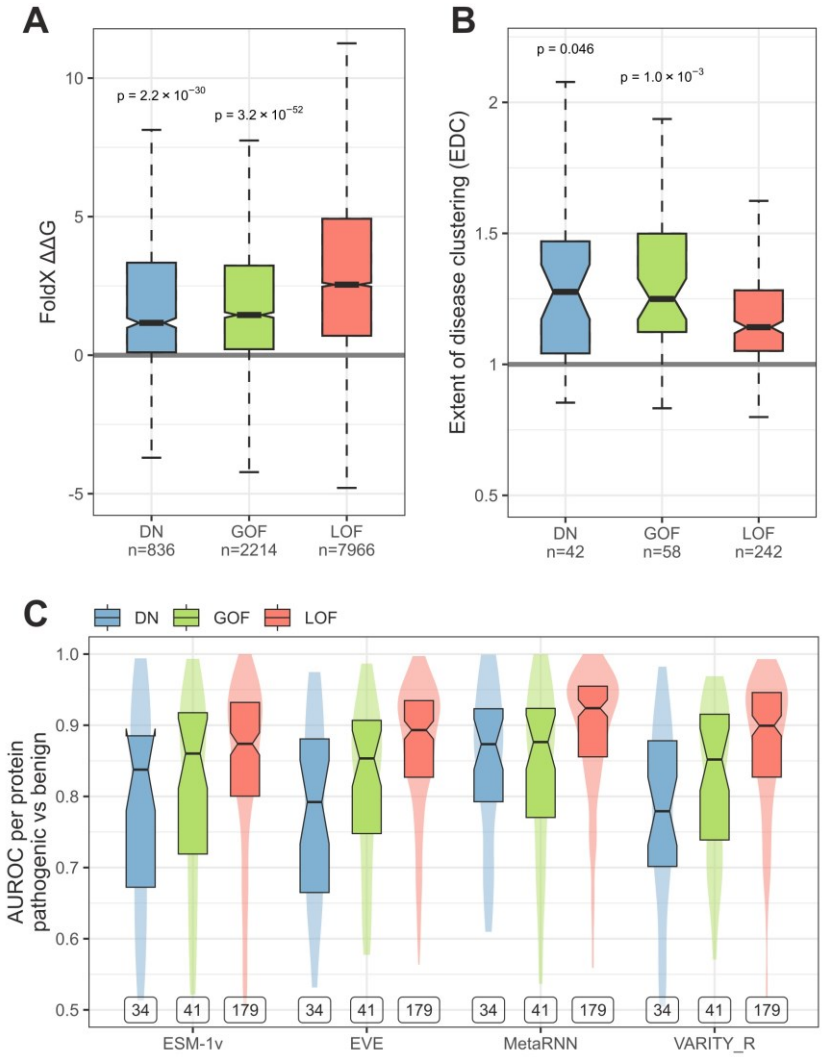
Appendix 3.3 Supplemental analyses using interface size and relative interface location.

(**A**) Interface area differences in homomers with different inheritance and molecular mechanisms. Crossbars are mean ± SEM. The p-values were calculated with Dunn's test using Holm-Bonferroni correction. Sample sizes are shown in panel B. (**B**) Relative interface location of homomers with different inheritance and molecular mechanisms. Crossbars are mean ± SEM. Sample sizes are shown on the left. (**C**) Relative interface location of C_2 symmetric homodimers with different inheritance and molecular mechanisms. Violins show the density distribution of the data and the crossbars are mean ± SEM. Dashed line shows the symmetry mean measured in all human C_2 dimers with structural data.



Appendix 5.1 Performance evaluation of the lasso regression model.

(**A**) Performance of the lasso regression model as a function of the penalty parameter estimated from the cross-validation folds. Dashed line is at λ = 0.00501, the penalty value in the final model. (**B**) Lasso penalty (λ) vs the regression coefficient (β). The lines are coloured according to the type of the variable: sequence-derived or evolutionary variables (blue), functional annotations (green), mutational constraint metrics (red), structural properties (black), interaction network-based property (pink), and experimental data (orange). (**C**) Precision-recall curve of the lasso regression model measured on the test set. (**D**) Distribution of model probabilities for non-LOF and LOF genes in the test set. The solid and dashed vertical lines mark threshold T1 and T2, respectively. (**E**) Performance metrics measured on the test set as a function of the model probability threshold. (**F**) Differences in ΔΔG and EDC (**G**) of pathogenic mutations between proteins predicted to be non-LOF versus all other proteins measured at different thresholds. Boxes denote data within 25th and 75th percentiles with the middle line representing the median, the notch containing the 95% confidence interval of the median and the white dots are the mean. Labels indicate the number of variants (ΔΔG) or the number of proteins (EDC) in the groups. The p-values were calculated with Wilcoxon rank-sum tests and effect sizes are shown in bold.



Appendix 5.2 Supplemental analysis to Figure 5.7: Validation of the models through model-independent metrics on an unbiased analysis set, using rank-based classification without t_{50} cutoff.

(**A**) FoldX-predicted ΔΔG of pathogenic missense mutations. Numbers below classes denote the number of mutations. Holm-Bonferroni corrected p-values above DN and GOF boxes are relative to the LOF group and were determined by one-sided Wilcoxon rank-sum test. Sample sizes indicate the number of variants. (**B**) Class probabilities of the analysis set *vs* EDC. Sample sizes indicate the number of proteins in each class. Holm-Bonferroni corrected p-values above DN and GOF boxes are relative to the LOF group and were determined by one-sided Wilcoxon rank-sum test. (**C**) Aggregated AUROC analysis of pathogenic *vs* benign variants in predicted molecular mechanism classes. Labels indicate the number of proteins in each class. Boxes denote data within 25th and 75th percentiles, the middle line represents the median and the notch contains the 95% confidence interval of the median. Violins show area-normalized distributions.

Bibliography

Abrusán, György, and Joseph A. Marsh. 2019. “Ligand Binding Site Structure Shapes Folding, Assembly and Degradation of Homomeric Protein Complexes.” *Journal of Molecular Biology* 431(19): 3871–88.

Acuna-Hidalgo, Rocio et al. 2015. “Post-Zygotic Point Mutations Are an Underrecognized Source of De Novo Genomic Variation.” *The American Journal of Human Genetics* 97(1): 67–74.

Agashe, Vishwas R et al. 2004. “Function of Trigger Factor and DnaK in Multidomain Protein Folding: Increase in Yield at the Expense of Folding Speed.” *Cell* 117(2): 199–209.

Agirre, J. et al. 2023. “The CCP4 Suite: Integrative Software for Macromolecular Crystallography.” *Acta Crystallographica Section D: Structural Biology* 79(6): 449–61.

Ahnert, S. E. et al. 2010. “Self-Assembly, Modularity, and Physical Complexity.” *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 82(2): 026117.

Ahnert, Sebastian E. et al. 2015. “Principles of Assembly Reveal a Periodic Table of Protein Complexes.” *Science* 350(6266): aaa2245–aaa2245.

Akdel, Mehmet et al. 2022. “A Structural Biology Community Assessment of AlphaFold2 Applications.” *Nature Structural & Molecular Biology*: 1–12.

Allan Drummond, D., and Claus O. Wilke. 2009. “The Evolutionary Consequences of Erroneous Protein Synthesis.” *Nature Reviews Genetics* 10(10): 715–24.

Amberger, Joanna S. et al. 2015. “OMIM.Org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders.” *Nucleic Acids Research* 43(D1): D789–98.

Amundsen, Susan K, Andrew F Taylor, and Gerald R Smith. 2002. “A Domain of RecC Required for Assembly of the Regulatory RecD Subunit into the Escherichia Coli RecBCD Holoenzyme.” *Genetics* 161(2): 483–92.

André, Ingemar et al. 2008. “Emergence of Symmetry in Homooligomeric Biological Assemblies.” *Proceedings of the National Academy of Sciences* 105(42): 16148–52.

Arpat, Alaaddin Bulak et al. 2020. “Transcriptome-Wide Sites of Collided Ribosomes Reveal Principles of Translational Pausing.” *Genome Research* 30(7): 985–99.

Attard, Thomas J., Julie P. I. Welburn, and Joseph A. Marsh. 2022. “Understanding Molecular Mechanisms and Predicting Phenotypic Effects of Pathogenic Tubulin Mutations.” *PLOS Computational Biology* 18(10): e1010611.

Auton, Adam et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526(7571): 68–74.

Backwell, Lisa, and Joseph A. Marsh. 2022. “Diverse Molecular Mechanisms Underlying Pathogenic Protein Mutations: Beyond the Loss-of-Function Paradigm.” *Annual review of genomics and human genetics* 23(1).

Badano, Jose L., and Nicholas Katsanis. 2002. “Beyond Mendel: An Evolving View of Human Genetic Disease Transmission.” *Nature Reviews Genetics* 3(10): 779–89.

Badonyi, Mihaly, and Joseph A. Marsh. 2022. “Large Protein Complex Interfaces Have Evolved to Promote Cotranslational Assembly.” *eLife* 11: e79602.

Badonyi, Mihaly, and Joseph A Marsh. 2023a. “Buffering of Genetic Dominance by Allele-Specific Protein Complex Assembly.” *Science Advances* 9(22): eadf9845.

Badonyi, Mihaly, and Joseph A. Marsh. 2023b. “Hallmarks and Evolutionary Drivers of Cotranslational Protein Complex Assembly.” *The FEBS Journal*. https://onlinelibrary.wiley.com/doi/abs/10.1111/febs.16869.

Barondes, Samuel H., and Marshall W. Nirenberg. 1962. “Fate of a Synthetic Polynucleotide Directing Cell-Free Protein Synthesis II. Association with Ribosomes.” *Science* 138(3542): 813–17.

Bateman, Alex et al. 2021. “UniProt: The Universal Protein Knowledgebase in 2021.” *Nucleic Acids Research* 49(D1): D480–89.

Bergendahl, L. Therese et al. 2019. “The Role of Protein Complexes in Human Genetic Disease.” *Protein Science* 28(8): 1400–1411.

Bergendahl, L. Therese, and Joseph A. Marsh. 2017. “Functional Determinants of Protein Assembly into Homomeric Complexes.” *Scientific Reports* 7(1).

Berman, Helen M. et al. 2000. “The Protein Data Bank.” *Nucleic Acids Research*.

Bertolini, Matilde et al. 2021. “Interactions between Nascent Proteins Translated by Adjacent Ribosomes Drive Homomer Assembly.” *Science* 371(6524): 57–64.

Bienert, Stefan et al. 2017. “The SWISS-MODEL Repository-New Features and Functionality.” *Nucleic Acids Research* 45(D1): D313–19.

Biever, Anne et al. 2020. “Monosomes Actively Translate Synaptic MRNAs in Neuronal Processes.” *Science* 367(6477). http://science.sciencemag.org/ (March 30, 2021).

Birgmeier, Johannes et al. 2020. “AMELIE Speeds Mendelian Diagnosis by Matching Patient Phenotype and Genotype to Primary Literature.” *Science Translational Medicine* 12(544): eaau9113.

Blundell, Tom L., and N. Srinivasan. 1996. “Symmetry, Stability, and Dynamics of Multidomain and Multicomponent Protein Systems.” *Proceedings of the National Academy of Sciences* 93(25): 14243–48.

Boone, Philip M. et al. 2016. “Increased Bone Turnover, Osteoporosis, Progressive Tibial Bowing, Fractures, and Scoliosis in a Patient with a Final-Exon SATB2 Frameshift Mutation.” *American Journal of Medical Genetics Part A* 170(11): 3028–32.

Bourke, Ashley M., Andre Schwarz, and Erin M. Schuman. 2023. “De-Centralizing the Central Dogma: MRNA Translation in Space and Time.” *Molecular Cell* 83(3): 452–68.

Bowne, Sara J. et al. 2011. “A Dominant Mutation in RPE65 Identified by Whole-Exome Sequencing Causes Retinitis Pigmentosa with Choroidal Involvement.” *European Journal of Human Genetics* 19(10): 1074–81.

Braberg, Hannes et al. 2020. “Genetic Interaction Mapping Informs Integrative Structure Determination of Protein Complexes.” *Science* 370(6522). https://doi.org/10.1126/science.aaz4910 (May 16, 2021).

Brandt, Florian et al. 2009. “The Native 3D Organization of Bacterial Polysomes.” *Cell* 136(2): 261–71.

———. 2010. “The Three-Dimensional Organization of Polyribosomes in Intact Human Cells.” *Molecular Cell* 39(4): 560–69.

Bray, Dennis, and Steven Lay. 1997. “Computer-Based Analysis of the Binding Steps in Protein Complex Formation.” *Proceedings of the National Academy of Sciences* 94(25): 13493–98.

Brennan, Christopher M. et al. 2019. “Protein Aggregation Mediates Stoichiometry of Protein Complexes in Aneuploid Cells.” *Genes & Development* 33(15–16): 1031–47.

Brooijmans, Natasja, Kim A. Sharp, and Irwin D. Kuntz. 2002. “Stability of Macromolecular Complexes.” *Proteins: Structure, Function and Genetics*.

Buckland, S. T., A. C. Davison, and D. V. Hinkley. 1998. “Bootstrap Methods and Their Application.” *Biometrics* 54(2).

Burkhard, Peter, Jörg Stetefeld, and Sergei V Strelkov. 2001. “Coiled Coils: A Highly Versatile Protein Folding Motif.” *Trends in Cell Biology* 11(2): 82–88.

Cassaignau, Anaïs M.E. et al. 2021. “Interactions between Nascent Proteins and the Ribosome Surface Inhibit Co-Translational Folding.” *Nature Chemistry* 13(12): 1214–20.

Cavaco, Branca M et al. 2018. “Homozygous Calcium-Sensing Receptor Polymorphism R544Q Presents as Hypocalcemic Hypoparathyroidism.” *The Journal of Clinical Endocrinology & Metabolism* 103(8): 2879–88.

Celesia, Gastone G. 2001. “Disorders of Membrane Channels or Channelopathies.” *Clinical Neurophysiology* 112(1): 2–18.

Chandler, David. 2005. “Interfaces and the Driving Force of Hydrophobic Assembly.” *Nature* 437(7059): 640–47.

Chen, Che-Hong, Benjamin R. Kraemer, and Daria Mochly-Rosen. 2022. "ALDH2 Variance in Disease and Populations." *Disease Models & Mechanisms* 15(6): dmm049601.

Chen, Jieming, Nicholas Sawyer, and Lynne Regan. 2013. "Protein–Protein Interactions: General Trends in the Relationship between Binding Affinity and Interfacial Buried Surface Area." *Protein Science* 22(4): 510–15.

Chen, Rong et al. 2016. "Analysis of 589,306 Genomes Identifies Individuals Resilient to Severe Mendelian Childhood Diseases." *Nature Biotechnology* 34(5): 531–38.

Chen, Xiuzhen, and Christine Mayr. 2022. "A Working Model for Condensate RNA-Binding Proteins as Matchmakers for Protein Complex Assembly." *RNA* 28(1): 76–87.

Chetal, Kashish, and Sarath Chandra Janga. 2015. "OperomeDB: A Database of Condition-Specific Transcription Units in Prokaryotic Genomes." *BioMed Research International* 2015. /pmc/articles/PMC4620388/ (March 23, 2021).

Chicco, Davide, Niklas Tötsch, and Giuseppe Jurman. 2021. "The Matthews Correlation Coefficient (MCC) Is More Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-Class Confusion Matrix Evaluation." *BioData Mining* 14(1): 13.

Choe, Young Jun et al. 2016. "Failure of RQC Machinery Causes Protein Aggregation and Proteotoxic Stress." *Nature* 531(7593): 191–95.

Chothia, Cyrus. 1974. "Hydrophobic Bonding and Accessible Surface Area in Proteins." *Nature* 248(5446): 338–39.

———. 1975. "Structural Invariants in Protein Folding." *Nature* 254(5498): 304–8.

———. 1976. "The Nature of the Accessible and Buried Surfaces in Proteins." *Journal of Molecular Biology* 105(1): 1–12.

Chothia, Cyrus, and Joël Janin. 1975. "Principles of Protein-Protein Recognition." *Nature* 256(5520): 705–8.

Chouaib, Racha et al. 2020. "A Dual Protein-MRNA Localization Screen Reveals Compartmentalized Translation and Widespread Co-Translational RNA Targeting." *Developmental Cell* 54(6): 773-791.e5.

Ciryam, Prajwal et al. 2013. "In Vivo Translation Rates Can Substantially Delay the Cotranslational Folding of the Escherichia Coli Cytosolic Proteome." *Proceedings of the National Academy of Sciences of the United States of America* 110(2).

Clamer, Massimiliano et al. 2018. "Active Ribosome Profiling with RiboLace." *Cell Reports* 25(4): 1097-1108.e5.

Clementel, Damiano et al. 2022. "RING 3.0: Fast Generation of Probabilistic Residue Interaction Networks from Structural Ensembles." *Nucleic Acids Research*. https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkac365/6584780 (June 6, 2022).

Conant, Gavin C. 2009. "Neutral Evolution on Mammalian Protein Surfaces." *Trends in Genetics* 25(9): 377–81.

Cooper, David N. et al. 2013. "Where Genotype Is Not Predictive of Phenotype: Towards an Understanding of the Molecular Basis of Reduced Penetrance in Human Inherited Disease." *Human Genetics* 132(10): 1077–1130.

Cornish-Bowden, Athel J., and D.E. Koshland. 1971. "The Quaternary Structure of Proteins Composed of Identical Subunits." *Journal of Biological Chemistry* 246(10): 3092–3102.

Crabb, D. W., H. J. Edenberg, W. F. Bosron, and T. K. Li. 1989. "Genotypes for Aldehyde Dehydrogenase Deficiency and Alcohol Sensitivity. The Inactive ALDH2(2) Allele Is Dominant." https://www.jci.org/articles/view/113875/pdf (July 23, 2023).

Crane, H. R. 1950. "Principles and Problems of Biological Growth." *The Scientific Monthly* 70(6): 376–89.

Crick, F. H. C., and J. D. Watson. 1957. "Virus Structure: General Principles." In *Ciba Foundation Symposium - Steroid Hormones and Enzymes (Book II of Colloquia on Endocrinology)*, John Wiley & Sons, Ltd, 5–18. https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470715239.ch1 (July 6, 2023).

Cunningham, Fiona et al. 2022. "Ensembl 2022." *Nucleic Acids Research* 50(D1): D988–95.

Curtis, Andrew R.J. et al. 2001. "Mutation in the Gene Encoding Ferritin Light Polypeptide Causes Dominant Adult-Onset Basal Ganglia Disease." *Nature Genetics* 28(4): 350–54.

Dandekar, Thomas, Berend Snel, Martijn Huynen, and Peer Bork. 1998. “Conservation of Gene Order: A Fingerprint of Proteins That Physically Interact.” *Trends in Biochemical Sciences* 23(9): 324–28.

Danecek, Petr et al. 2021. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10(2): giab008.

Das, Sulagna, Robert H Singer, and Young J Yoon. 2019. “The Travels of MRNAs in Neurons: Do They Know Where They Are Going?” *Current Opinion in Neurobiology* 57: 110–16.

Dayhoff, Judith E., Benjamin A. Shoemaker, Stephen H. Bryant, and Anna R. Panchenko. 2010. “Evolution of Protein Binding Modes in Homooligomers.” *Journal of Molecular Biology* 395(4): 860–70.

De Magalhães, and João Pedro. 2013. “How Ageing Processes Influence Cancer.” *Nature Reviews Cancer* 13(5): 357–65.

Deckert, Annika et al. 2021. “Common Sequence Motifs of Nascent Chains Engage the Ribosome Surface and Trigger Factor.” *Proceedings of the National Academy of Sciences of the United States of America* 118(52).

Delgado, Javier, Leandro G. Radusky, Damiano Cianferoni, and Luis Serrano. 2019. “FoldX 5.0: Working with RNA, Small Molecules and a New Graphical Interface.” *Bioinformatics* 35(20): 4168–69.

Demircioglu, F. Esra et al. 2019. “The AAA + ATPase TorsinA Polymerizes into Hollow Helical Tubes with 8.5 Subunits per Turn.” *Nature Communications* 10(1): 3262.

Dénes, Türei et al. 2021. “Integrated Intra- and Intercellular Signaling Knowledge for Multicellular Omics Analysis.” *Molecular Systems Biology* 17(3): e9923.

Dolinsky, Todd J., Jens E. Nielsen, J. Andrew McCammon, and Nathan A. Baker. 2004. “PDB2PQR: An Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations.” *Nucleic Acids Research* 32(WEB SERVER ISS.): W665–67.

D’Orazio, Karole N., and Rachel Green. 2021. “Ribosome States Signal RNA Quality Control.” *Molecular Cell* 81(7): 1372–83.

Döring, Kristina et al. 2017. “Profiling Ssb-Nascent Chain Interactions Reveals Principles of Hsp70-Assisted Folding.” *Cell* 170(2): 298-311.e20.

Downward, M. J. 1973. “A System for the Description of Point Groups.” *Journal of Chemical Education* 50(8): 553.

Drew, Kevin, John B Wallingford, and Edward M Marcotte. 2021. “Hu.MAP 2.0: Integration of over 15,000 Proteomic Experiments Builds a Global Compendium of Human Multiprotein Assemblies.” *Molecular Systems Biology* 17(5): e10016.

Drutman, Scott B. et al. 2019. “Homozygous NLRP1 Gain-of-Function Mutation in Siblings with a Syndromic Form of Recurrent Respiratory Papillomatosis.” *Proceedings of the National Academy of Sciences* 116(38): 19055–63.

Dubreuil, Benjamin, Or Matalon, and Emmanuel D. Levy. 2019. “Protein Abundance Biases the Amino Acid Composition of Disordered Regions to Minimize Non-Functional Interactions.” *Journal of Molecular Biology* 431(24): 4978–92.

Duncan, Caia D.S., and Juan Mata. 2011. “Widespread Cotranslational Formation of Protein Complexes.” *PLoS Genetics* 7(12): e1002398.

———. 2014. “Cotranslational Protein-RNA Associations Predict Protein-Protein Interactions.” *BMC Genomics* 15(1): 298.

Duttler, Stefanie, Sebastian Pechmann, and Judith Frydman. 2013. “Principles of Cotranslational Ubiquitination and Quality Control at the Ribosome.” *Molecular cell* 50(3): 10.1016/j.molcel.2013.03.010.

Edgar, Ron, Michael Domrachev, and Alex E. Lash. 2002. “Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository.” *Nucleic acids research* 30(1): 207–10.

Eilbeck, Karen, Aaron Quinlan, and Mark Yandell. 2017. “Settling the Score: Variant Prioritization and Mendelian Disease.” *Nature Reviews Genetics 2017 18:10* 18(10): 599–612.

Eisenberg, David, and Andrew D. Mclachlan. 1986. “Solvation Energy in Protein Folding and Binding.” *Nature* 319(6050): 199–203.

Emlaw, Johnathon R. et al. 2021. “A Single Historical Substitution Drives an Increase in Acetylcholine Receptor Complexity.” *Proceedings of the National Academy of Sciences* 118(7): e2018731118.

Erijman, Ariel, Eran Rosenthal, and Julia M. Shifman. 2014. “How Structure Defines Affinity in Protein-Protein Interactions.” *PLOS ONE* 9(10): e110085.

Esin, Alexander et al. 2018. “The Genetic Basis and Evolution of Red Blood Cell Sickling in Deer.” *Nature Ecology & Evolution* 2(2): 367–76.

Evans, Richard et al. 2021. “Protein Complex Prediction with AlphaFold-Multimer.” *bioRxiv*: 2021.10.04.463034.

Faure, Guilhem, Aleksey Y. Ogurtsov, Svetlana A. Shabalina, and Eugene V. Koonin. 2016. “Role of MRNA Structure in the Control of Protein Folding.” *Nucleic Acids Research* 44(22): 10898–911.

Fernández, Ariel, and Michael Lynch. 2011. “Non-Adaptive Origins of Interactome Complexity.” *Nature* 474(7352): 502–5.

Fersht, Alan R. et al. 1985. “Hydrogen Bonding and Biological Specificity Analysed by Protein Engineering.” *Nature* 314(6008): 235–38.

Findlay, Gregory M. et al. 2018. “Accurate Classification of BRCA1 Variants with Saturation Genome Editing.” *Nature* 562(7726): 217–22.

Fischer, Manuel et al. 2020. “Analysis of the Co-Translational Assembly of the Fungal Fatty Acid Synthase (FAS).” *Scientific Reports* 10(1).

Fomin, Andrey et al. 2021. “Truncated Titin Proteins and Titin Haploinsufficiency Are Targets for Functional Recovery in Human Cardiomyopathy Due to TTN Mutations.” *Science Translational Medicine* 13(618): eabd3079.

Forrest, Lucy R. 2015. “Structural Symmetry in Membrane Proteins∗.” *Annual Review of Biophysics* 44(1): 311–37.

Fowler, Douglas M. et al. 2023. “An Atlas of Variant Effects to Understand the Genome at Nucleotide Resolution.” *Genome Biology* 24(1): 147.

Fowler, Douglas M., and Stanley Fields. 2014. “Deep Mutational Scanning: A New Style of Protein Science.” *Nature Methods* 11(8): 801–7.

Frazer, Jonathan et al. 2021. “Disease Variant Prediction with Deep Generative Models of Evolutionary Data.” *Nature* 599(7883): 91–95.

Gane, A. et al. 2022. “ProtNLM: Model-Based Natural Language Protein Annotation.” https://storage.googleapis.com/brain-genomics-public/research/proteins/protnlm/uniprot_2022_04/protnlm_preprint_draft.pdf.

Gao, Mu, Davi Nakajima An, Jerry M. Parks, and Jeffrey Skolnick. 2022. “AF2Complex Predicts Direct Physical Interactions in Multimeric Proteins with Deep Learning.” *Nature Communications 2022 13:1* 13(1): 1–13.

Garcia-Seisdedos, Hector, Charly Empereur-Mot, Nadav Elad, and Emmanuel D. Levy. 2017. “Proteins Evolve on the Edge of Supramolecular Self-Assembly.” *Nature* 548(7666): 244–47.

Geng, Yanyan et al. 2023. “BK Channels of Five Different Subunit Combinations Underlie the de Novo KCNMA1 G375R Channelopathy.” *The Journal of General Physiology* 155(5): e202213302.

Gerasimavicius, Lukas, Xin Liu, and Joseph A. Marsh. 2020. “Identification of Pathogenic Missense Mutations Using Protein Stability Predictors.” *Scientific Reports* 10(1): 15387.

Gerasimavicius, Lukas, Benjamin J. Livesey, and Joseph A. Marsh. 2022. “Loss-of-Function, Gain-of-Function and Dominant-Negative Mutations Have Profoundly Different Effects on Protein Structure.” *Nature Communications 2022 13:1* 13(1): 1–15.

Gething, Mary Jane, and Joseph Sambrook. 1992. “Protein Folding in the Cell.” *Nature* 355(6355): 33–45.

Gilmore, Ross et al. 1996. “Co-Translational Trimerization of the Reovirus Cell Attachment Protein.” *EMBO Journal* 15(11): 2651–58.

Giurgiu, Madalina et al. 2019. "CORUM: The Comprehensive Resource of Mammalian Protein Complexes - 2019." *Nucleic Acids Research* 47(D1): D559–63.

Goodsell, David S., and Arthur J. Olson. 2000. "Structural Symmetry and Protein Function." *Annual Review of Biophysics and Biomolecular Structure* 29(1): 105–53.

Gower, J. C. 1971. "A General Coefficient of Similarity and Some of Its Properties." *Biometrics* 27(4): 857–71.

Grasberger, Helmut et al. 2005. "Thyroid Transcription Factor 1 Rescues PAX8/P300 Synergism Impaired by a Natural PAX8 Paired Domain Mutation with Dominant Negative Activity." *Molecular Endocrinology* 19(7): 1779–91.

Gray, Michael W. et al. 2010. "Irremediable Complexity?" *Science* 330(6006): 920–21.

Haig, David, and Laurence D. Hurst. 1991. "A Quantitative Measure of Error Minimization in the Genetic Code." *Journal of Molecular Evolution* 33(5): 412–17.

Halbach, André et al. 2009. "Cotranslational Assembly of the Yeast SET1C Histone Methyltransferase Complex." *EMBO Journal* 28(19): 2959–70.

Haldane, J. B. S. 1930. "A Note on Fisher's Theory of the Origin of Dominance, and on a Correlation between Dominance and Linkage." *The American Naturalist* 64(690): 87–90.

Han, Peixun et al. 2020. "Genome-Wide Survey of Ribosome Collision." *Cell Reports* 31(5): 107610.

Harel, Tamar et al. 2016. "Recurrent De Novo and Biallelic Variation of ATAD3A, Encoding a Mitochondrial Membrane Protein, Results in Distinct Neurological Syndromes." *American Journal of Human Genetics* 99(4): 831–45.

Harper, J. Wade, and Eric J. Bennett. 2016. "Proteome Complexity and the Forces That Drive Proteome Imbalance." *Nature* 537(7620): 328–38.

Hashimoto, Kosuke, and Anna R. Panchenko. 2010. "Mechanisms of Protein Oligomerization, the Critical Role of Insertions and Deletions in Maintaining Different Oligomeric States." *Proceedings of the National Academy of Sciences* 107(47): 20352–57.

Hegde, Ramanujan S., and Robert J. Keenan. 2022. "The Mechanisms of Integral Membrane Protein Biogenesis." *Nature Reviews Molecular Cell Biology* 23(2): 107–24.

Heidenreich, Meta et al. 2020. "Designer Protein Assemblies with Tunable Phase Diagrams in Living Cells." *Nature Chemical Biology* 16(9): 939–45.

Hendsch, Zachary S., and Bruce Tidor. 1994. "Do Salt Bridges Stabilize Proteins? A Continuum Electrostatic Analysis." *Protein Science* 3(2): 211–26.

de Heredia, Miguel López, and Ralf-Peter Jansen. 2004. "MRNA Localization and the Cytoskeleton." *Current Opinion in Cell Biology* 16(1): 80–85.

Hermann, Robert B. 1972. "Theory of Hydrophobic Bonding. II. The Correlation of Hydrocarbon Solubility in Water with Solvent Cavity Surface Area." *Journal of Physical Chemistry* 76(19): 2754–59.

Hernández, Helena, and Carol V. Robinson. 2007. "Determining the Stoichiometry and Interactions of Macromolecular Assemblies from Mass Spectrometry." *Nature Protocols* 2(3): 715–26.

Herskowitz, Ira. 1987. "Functional Inactivation of Genes by Dominant Negative Mutations." *Nature* 329(6136): 219–22.

Heyer, Erin E., and Melissa J. Moore. 2016. "Redefining the Translational Status of 80S Monosomes." *Cell* 164(4): 757–69.

Hochberg, Georg K.A. et al. 2020. "A Hydrophobic Ratchet Entrenches Molecular Complexes." *Nature* 588(7838): 503–8.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6(2).

Horton, Nancy, and Mitchell Lewis. 1992. "Calculation of the Free Energy of Association for Protein Complexes." *Protein Science*: 169–81.

Horvath, Gabriella A. et al. 2018. “Gain-of-Function KCNJ6 Mutation in a Severe Hyperkinetic Movement Disorder Phenotype.” *Neuroscience* 384: 152–64.

Hsia, Kuo-Chiang, Pete Stavropoulos, Günter Blobel, and André Hoelz. 2007. “Architecture of a Coat for the Nuclear Pore Membrane.” *Cell* 131(7): 1313–26.

Huang, Ni, Insuk Lee, Edward M. Marcotte, and Matthew E. Hurles. 2010. “Characterising and Predicting Haploinsufficiency in the Human Genome” ed. Mikkel H. Schierup. *PLoS Genetics* 6(10): e1001154.

Hubbard SJ, Thornton JM. 1993. “NACCESS.”

Humphreys, Ian R. et al. 2021. “Computed Structures of Core Eukaryotic Protein Complexes.” *Science* 374(6573). https://www.science.org/doi/abs/10.1126/science.abm4805 (January 9, 2022).

Hurst, Laurence D., and James P. Randerson. 2000. “Dosage, Deletions and Dominance: Simple Models of the Evolution of Gene Expression.” *Journal of Theoretical Biology* 205(4): 641–47.

Ingolia, Nicholas T. 2014. “Ribosome Profiling: New Views of Translation, from Single Codons to Genome Scale.” *Nature Reviews Genetics* 15(3): 205–13.

Isaac, Richard. 1995. “The Idea of Independence, with Applications.” In *The Pleasures of Probability*, Undergraduate Texts in Mathematics, ed. Richard Isaac. New York, NY: Springer, 41–53. https://doi.org/10.1007/978-1-4612-0819-8_5 (July 10, 2023).

Isaacs, W B, I S Kim, A Struve, and A. B. Fulton. 1992. “Association of Titin and Myosin Heavy Chain in Developing Skeletal Muscle.” *Proceedings of the National Academy of Sciences of the United States of America* 89(16): 7496–7500.

Ishikawa, Koji et al. 2017. “Post-Translational Dosage Compensation Buffers Genetic Perturbations to Stoichiometry of Protein Complexes.” *PLoS Genetics* 13(1).

Izumi, Kosuke et al. 2015. “Germline Gain-of-Function Mutations in AFF4 Cause a Developmental Syndrome Functionally Linking the Super Elongation Complex and Cohesin.” *Nature Genetics* 47(4): 338–44.

Jackson, Sophie E, Antonio Suma, and Cristian Micheletti. 2017. “How to Fold Intricately: Using Theory and Experiments to Unravel the Properties of Knotted Proteins.” *Current Opinion in Structural Biology* 42: 6–14.

Jacobs, William M., and Eugene I. Shakhnovich. 2017. “Evidence of Evolutionary Selection for Cotranslational Folding.” *Proceedings of the National Academy of Sciences of the United States of America* 114(43): 11434–39.

Janin, J. 1995. “Principles of Protein-Protein Recognition from Structure to Thermodynamics.” *Biochimie* 77(7): 497–505.

Janin, J, and C Chothia. 1990. “The Structure of Protein-Protein Recognition Sites.” *Journal of Biological Chemistry* 265(27): 16027–30.

Jimenez-Sanchez, G., B. Childs, and D. Valle. 2001. “Human Disease Genes.” *Nature* 409(6822): 853–55.

Johnston, Iain G. et al. 2022. “Symmetry and Simplicity Spontaneously Emerge from the Algorithmic Nature of Evolution.” *Proceedings of the National Academy of Sciences* 119(11): e2113883119.

Jones, Susan, and Janet M. Thornton. 1996. “Principles of Protein-Protein Interactions.” *Proceedings of the National Academy of Sciences of the United States of America* 93(1): 13–20.

Jumper, John et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596(7873): 583–89.

Jung, Kwanghee, Jaehoon Lee, Vibhuti Gupta, and Gyeongcheol Cho. 2019. “Comparison of Bootstrap Confidence Interval Methods for GSCA Using a Monte Carlo Simulation.” *Frontiers in Psychology* 10. https://www.frontiersin.org/articles/10.3389/fpsyg.2019.02215 (August 30, 2022).

Juszkiewicz, Szymon, and Ramanujan S. Hegde. 2018. “Quality Control of Orphaned Proteins.”

Kabsch, Wolfgang, and Christian Sander. 1983. “Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features.” *Biopolymers* 22(12): 2577–2637.

Kacser, H., and J. A. Burns. 1981. “The Molecular Basis of Dominance.” *Genetics* 97(3–4): 639–66.

Kajander, Tommi et al. 2000. "Buried Charged Surface in Proteins." *Structure* 8(11): 1203–14.

Kamenova, Ivanka et al. 2019. "Co-Translational Assembly of Mammalian Nuclear Multisubunit Complexes." *Nature Communications* 10(1).

Kaplanis, Joanna et al. 2020. "Evidence for 28 Genetic Disorders Discovered by Combining Healthcare and Research Data." *Nature* 586(7831): 757–62.

Karczewski, Konrad J. et al. 2020. "The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans." *Nature* 581(7809): 434–43.

Kastritis, Panagiotis L et al. 2011. "A Structure-Based Benchmark for Protein–Protein Binding Affinity." *Protein Science : A Publication of the Protein Society* 20(3): 482–91.

Kastritis, Panagiotis L. et al. 2014. "Proteins Feel More than They See: Fine-Tuning of Binding Affinity by Properties of the Non-Interacting Surface." *Journal of Molecular Biology* 426(14): 2632–52.

Kastritis, Panagiotis L, and Alexandre M.J.J. Bonvin. 2013. "On the Binding Affinity of Macromolecular Interactions: Daring to Ask Why Proteins Interact." *Journal of the Royal Society Interface* 10(79).

Kaufmann, D. et al. 2001. "Spinal Neurofibromatosis without Café-Au-Lait Macules in Two Families with Null Mutations of the NF1 Gene." *American Journal of Human Genetics* 69(6): 1395–1400.

Kauzmann, W. 1959. "Some Factors in the Interpretation of Protein Denaturation." *Advances in Protein Chemistry* 14(C): 1–63.

Kawashima, Shuichi et al. 2008. "AAindex: Amino Acid Index Database, Progress Report 2008." *Nucleic Acids Research* 36(suppl_1): D202–5.

Ke, Guolin et al. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." In *Advances in Neural Information Processing Systems*, Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (July 29, 2023).

Keene, Jack D. 2007. "RNA Regulons: Coordination of Post-Transcriptional Events." *Nature Reviews Genetics* 8(7): 533–43.

Keene, Jack D., Jordan M. Komisarow, and Matthew B. Friedersdorf. 2006. "RIP-Chip: The Isolation and Identification of MRNAs, MicroRNAs and Protein Components of Ribonucleoprotein Complexes from Cell Extracts." *Nature Protocols* 1(1): 302–7.

Khan, Krishnendu et al. 2022. "Multimodal Cotranslational Interactions Direct Assembly of the Human Multi-TRNA Synthetase Complex." *Proceedings of the National Academy of Sciences* 119(36): e2205669119.

Kiho, Y., and A. Rich. 1964. "Induced Enzyme Formed on Bacterial Polyribosomes." *Proceedings of the National Academy of Sciences of the United States of America* 51(1): 111–18.

Kim, Jungsu et al. 2011. "Haploinsufficiency of Human APOE Reduces Amyloid Deposition in a Mouse Model of Amyloid-β Amyloidosis." *Journal of Neuroscience* 31(49): 18007–12.

Kim, Wan Kyu, Andreas Henschel, Christof Winter, and Michael Schroeder. 2006. "The Many Faces of Protein-Protein Interactions: A Compendium of Interface Geometry." *PLoS Computational Biology* 2(9): 1151–64.

Kimura, Motoo. 1968. "Evolutionary Rate at the Molecular Level." *Nature* 217(5129): 624–26.

Kleinjung, Jens, and Franca Fraternali. 2005. "POPSCOMP: An Automated Interaction Analysis of Biomolecular Complexes." *Nucleic Acids Research*.

Koldewey, Philipp, Scott Horowitz, and James C A Bardwell. 2017. "Chaperone-Client Interactions: Non-Specificity Engenders Multi-Functionality." *Journal of Biological Chemistry*. http://www.jbc.org/cgi/doi/10.1074/jbc.R117.796862 (December 16, 2019).

Kramer, Günter, Ayala Shiber, and Bernd Bukau. 2019. "Mechanisms of Cotranslational Maturation of Newly Synthesized Proteins." *Annual Review of Biochemistry* 88(1): 337–64.

Kramer, Ryan M. et al. 2012. "Toward a Molecular Understanding of Protein Solubility: Increased Negative Surface Charge Correlates with Increased Solubility." *Biophysical Journal* 102(8): 1907–15.

Kudva, Renuka et al. 2018. "The Shape of the Bacterial Ribosome Exit Tunnel Affects Cotranslational Protein Folding." *eLife* 7.

Kuhn, Max. 2014. "Futility Analysis in the Cross-Validation of Machine Learning Models." http://arxiv.org/abs/1405.6974 (July 29, 2023).

Kühner, Sebastian et al. 2009. "Proteome Organization in a Genome-Reduced Bacterium." *Science* 326(5957): 1235–40.

Landrum, Melissa J. et al. 2018. "ClinVar: Improving Access to Variant Interpretations and Supporting Evidence." *Nucleic Acids Research* 46(D1): D1062–67.

Lelieveld, Stefan H. et al. 2017. "Spatial Clustering of de Novo Missense Mutations Identifies Candidate Neurodevelopmental Disorder-Associated Genes." *American Journal of Human Genetics* 101(3): 478–84.

Leonard, Alexander S., and Sebastian E. Ahnert. 2019. "Evolution of Interface Binding Strengths in Simplified Model of Protein Quaternary Structure" ed. Charlotte M Deane. *PLoS Computational Biology* 15(6): e1006886.

Levy, Emmanuel D. 2010. "A Simple Definition of Structural Regions in Proteins and Its Use in Analyzing Interface Evolution." *Journal of Molecular Biology* 403(4): 660–70.

Levy, Emmanuel D., Subhajyoti De, and Sarah A. Teichmann. 2012. "Cellular Crowding Imposes Global Constraints on the Chemistry and Evolution of Proteomes." *Proceedings of the National Academy of Sciences of the United States of America* 109(50): 20461–66.

Levy, Emmanuel D., Elisabetta Boeri Erba, Carol V. Robinson, and Sarah A. Teichmann. 2008. "Assembly Reflects Evolution of Protein Complexes." *Nature* 453(7199): 1262–65.

Levy, Emmanuel D., and Sarah Teichmann. 2013. "Structural, Evolutionary, and Assembly Principles of Protein Oligomerization." In *Progress in Molecular Biology and Translational Science*, Elsevier B.V., 25–51.

Levy, Emmanuel D., and Christine Vogel. 2021. "'Structuromics': Another Step toward a Holistic View of the Cell." *Cell* 184(2): 301–3.

Li, Chang, Degui Zhi, Kai Wang, and Xiaoming Liu. 2022. "MetaRNN: Differentiating Rare Pathogenic and Rare Benign Missense SNVs and InDels Using Deep Learning." *Genome Medicine* 14(1): 115.

Li, Weiyi, and Michael Lynch. 2020. "Universally High Transcript Error Rates in Bacteria" eds. Christian R Landry, Patricia J Wittkopp, and Joanna Masel. *eLife* 9: e54898.

Liebeskind, Benjamin J., David M. Hillis, and Harold H. Zakon. 2015. "Convergence of Ion Channel Genome Content in Early Animal Evolution." *Proceedings of the National Academy of Sciences of the United States of America* 112(8): E846–51.

Liebeskind, Benjamin J., Claire D. McWhite, and Edward M. Marcotte. 2016. "Towards Consensus Gene Ages." *Genome Biology and Evolution* 8(6): 1812–23.

Linse, Sara et al. 1988. "The Role of Protein Surface Charges in Ion Binding." *Nature* 335(6191): 651–52.

Liu, Fang, David K. Jones, Willem J. De Lange, and Gail A. Robertson. 2016. "Cotranslational Association of MRNA Encoding Subunits of Heteromeric Ion Channels." *Proceedings of the National Academy of Sciences of the United States of America* 113(17): 4859–64.

Liu, Yansheng, Andreas Beyer, and Ruedi Aebersold. 2016. "On the Dependency of Cellular Protein Levels on MRNA Abundance." *Cell* 165(3): 535–50.

Liu, Yanshun, and David Eisenberg. 2002. "3D Domain Swapping: As Domains Continue to Swap." *Protein Science* 11(6): 1285–99.

Livesey, Benjamin J, and Joseph A Marsh. 2023. "Updated Benchmarking of Variant Effect Predictors Using Deep Mutational Scanning." *Molecular Systems Biology* n/a(n/a): e11474.

Lomize, Andrei L. et al. 2022. "Membranome 3.0: Database of Single-Pass Membrane Proteins with AlphaFold Models." *Protein Science* 31(5): e4318.

Lu, Manman et al. 2020. "Atomic-Resolution Structure of HIV-1 Capsid Tubes by Magic-Angle Spinning NMR." *Nature Structural & Molecular Biology* 27(9): 863–69.

Lukatsky, D. B., K. B. Zeldovich, and E. I. Shakhnovich. 2006. “Statistically Enhanced Self-Attraction of Random Patterns.” *Physical Review Letters* 97(17): 178101.

Lukeš, Julius et al. 2011. “How a Neutral Evolutionary Ratchet Can Build Cellular Complexity.” *IUBMB Life* 63(7): 528–37.

Lundberg, Scott M., and Su-In Lee. 2017. “A Unified Approach to Interpreting Model Predictions.” In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, Red Hook, NY, USA: Curran Associates Inc., 4768–77.

Lynch, Michael. 2006. “Streamlining and Simplification of Microbial Genome Architecture.” *Annual Review of Microbiology* 60(1): 327–49.

———. 2010. “Rate, Molecular Spectrum, and Consequences of Human Mutation.” *Proceedings of the National Academy of Sciences* 107(3): 961–68.

———. 2012. “The Evolution of Multimeric Protein Assemblages.” *Molecular Biology and Evolution* 29(5): 1353–66.

———. 2013. “Evolutionary Diversification of the Multimeric States of Proteins.” *Proceedings of the National Academy of Sciences of the United States of America* 110(30): E2821–28.

Ma, Weirui, and Christine Mayr. 2018. “A Membraneless Organelle Associated with the Endoplasmic Reticulum Enables 3′UTR-Mediated Protein-Protein Interactions.” *Cell* 175(6): 1492-1506.e19.

Mainz, Andi et al. 2009. “Large Protein Complexes with Extreme Rotational Correlation Times Investigated in Solution by Magic-Angle-Spinning NMR Spectroscopy.” *Journal of the American Chemical Society* 131(44): 15968–69.

Mallik, Saurav, and Sudip Kundu. 2017. “Coevolutionary Constraints in the Sequence-Space of Macromolecular Complexes Reflect Their Self-Assembly Pathways.” *Proteins: Structure, Function, and Bioinformatics* 85(7): 1183–89.

Mallik, Saurav, Dan S Tawfik, and Emmanuel D Levy. 2022. “How Gene Duplication Diversifies the Landscape of Protein Oligomeric State and Function.” *Current Opinion in Genetics & Development* 76: 101966.

Marsh, Joseph A. et al. 2013. “Protein Complexes Are under Evolutionary Selection to Assemble via Ordered Pathways.” *Cell* 153(2): 461–70.

Marsh, Joseph A., Holly A. Rees, Sebastian E. Ahnert, and Sarah A. Teichmann. 2015. “Structural and Evolutionary Versatility in Protein Complexes with Uneven Stoichiometry.” *Nature Communications* 6(1): 1–10.

Marsh, Joseph A., and Sarah A. Teichmann. 2014a. “Parallel Dynamics and Evolution: Protein Conformational Fluctuations and Assembly Reflect Evolutionary Changes in Sequence and Structure.” *BioEssays* 36(2): 209–18.

———. 2014b. “Protein Flexibility Facilitates Quaternary Structure Assembly and Evolution” ed. Gregory A. Petsko. *PLoS Biology* 12(5): e1001870.

Marsh, Joseph A, and Sarah A Teichmann. 2015. “Structure, Dynamics, Assembly, and Evolution of Protein Complexes.” *Annual Review of Biochemistry* 84: 551–75.

Martin, Hilary C. et al. 2018. “Quantifying the Contribution of Recessive Coding Variation to Developmental Disorders.” *Science* 362(6419): 1161–64.

Martin, Kelsey C., and Anne Ephrussi. 2009. “MRNA Localization: Gene Expression in the Spatial Dimension.” *Cell* 136(4): 719–30.

Masterson, C. et al. 1992. “Reconstitution of the Activities of the RecBCD Holoenzyme of Escherichia Coli from the Purified Subunits.” *Journal of Biological Chemistry* 267(19): 13564–72.

Matalon, Or, Amnon Horovitz, and Emmanuel D Levy. 2014. “Different Subunits Belonging to the Same Protein Complex Often Exhibit Discordant Expression Levels and Evolutionary Properties.” *Current Opinion in Structural Biology* 26: 113–20.

McEntagart, Meriel et al. 2016. “A Restricted Repertoire of de Novo Mutations in ITPR1 Cause Gillespie Syndrome with Evidence for Dominant-Negative Effect.” *American Journal of Human Genetics* 98(5): 981–92.

McLaren, William et al. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17(1): 122.

McLean, W.H. Irwin, and C.B. Tara Moore. 2011. “Keratin Disorders: From Gene to Therapy.” *Human Molecular Genetics* 20(R2): R189–97.

McRae, Jeremy F. et al. 2017. “Prevalence and Architecture of de Novo Mutations in Developmental Disorders.” *Nature* 542(7642): 433–38.

McShane, Erik et al. 2016. “Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation.” *Cell* 167(3): 803-815.e21.

Meier, Joshua et al. 2021. “Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function.” https://www.biorxiv.org/content/10.1101/2021.07.09.450648v2 (July 29, 2023).

Meldal, Birgit H.M. et al. 2019. “Complex Portal 2018: Extended Content and Enhanced Visualization Tools for Macromolecular Complexes.” *Nucleic Acids Research* 47(D1): D550–58.

Mena, Elijah L. et al. 2020. “Structural Basis for Dimerization Quality Control.” *Nature* 586(7829): 452–56.

Menardi, Giovanna, and Nicola Torelli. 2014. “Training and Assessing Classification Rules with Imbalanced Data.” *Data Mining and Knowledge Discovery* 28(1): 92–122.

Mendonsa, Samantha et al. 2023. “Massively Parallel Identification of MRNA Localization Elements in Primary Cortical Neurons.” *Nature Neuroscience*: 1–12.

Mi, Huaiyu et al. 2019. “Protocol Update for Large-Scale Genome and Gene Function Analysis with PANTHER Classification System (v.14.0).” *Nature protocols* 14(3): 703–21.

———. 2021. “PANTHER Version 16: A Revised Family Classification, Tree-Based Classification Tool, Enhancer Regions and Extensive API.” *Nucleic Acids Research* 49(D1): D394–403.

Mighell, Taylor L., Sara Evans-Dutson, and Brian J. O’Roak. 2018. “A Saturation Mutagenesis Approach to Understanding PTEN Lipid Phosphatase Activity and Genotype-Phenotype Relationships.” *American Journal of Human Genetics* 102(5): 943–55.

Miller, Susan, Joël Janin, Arthur M. Lesk, and Cyrus Chothia. 1987. “Interior and Surface of Monomeric Proteins.” *Journal of Molecular Biology* 196(3): 641–56.

Miller, Susan, Arthur M. Lesk, Joël Janin, and Cyrus Chothia. 1987. “The Accessible Surface Area and Stability of Oligomeric Proteins.” *Nature* 328(6133): 834–36.

Mingle, Lisa A. et al. 2005. “Localization of All Seven Messenger RNAs for the Actin-Polymerization Nucleator Arp2/3 Complex in the Protrusions of Fibroblasts.” *Journal of Cell Science* 118(Pt 11): 2425–33.

Mistry, Jaina et al. 2021. “Pfam: The Protein Families Database in 2021.” *Nucleic Acids Research* 49(D1): D412–19.

Mitchison, Tim, and Marc Kirschner. 1984. “Dynamic Instability of Microtubule Growth.” *Nature* 312(5991): 237–42.

Mitternacht, Simon. 2016. “FreeSASA: An Open Source C Library for Solvent Accessible Surface Area Calculations.” *F1000Research* 5. /pmc/articles/PMC4776673/ (February 22, 2021).

Monod, Jacque, Jeffries Wyman, and Jean Pierre Changeux. 1965. “On the Nature of Allosteric Transitions: A Plausible Model.” *Journal of Molecular Biology* 12(1): 88–118.

Monod, Jacques, Jean-Pierre Changeux, and François Jacob. 1963. “Allosteric Proteins and Cellular Control Systems.” *Journal of Molecular Biology* 6(4): 306–29.

Morales-Polanco, Fabián, Jae Ho Lee, Natália M. Barbosa, and Judith Frydman. 2022. “Cotranslational Mechanisms of Protein Biogenesis and Complex Assembly in Eukaryotes.” *Annual Review of Biomedical Data Science* 5: 67–94.

Mrazek, Jan et al. 2014. “Polyribosomes Are Molecular 3D Nanoprinters That Orchestrate the Assembly of Vault Particles.” *ACS Nano* 8(11): 11552–59.

Mueller, Marcus et al. 2009. “The Structure of a Cytolytic α-Helical Toxin Pore Reveals Its Assembly Mechanism.” *Nature* 459(7247): 726–30.

Mullaney, Julienne M., Ryan E. Mills, W. Stephen Pittard, and Scott E. Devine. 2010. "Small Insertions and Deletions (INDELs) in Human Genomes." *Human Molecular Genetics* 19(R2): R131–36.

Mushegian, Arcady R., and Eugene V. Koonin. 1996. "Gene Order Is Not Conserved in Bacterial Evolution." *Trends in genetics* 12(8): 289–90.

Nakashima, Hiroshi, Ken Nishikawa, and Tatsuo Ooi. 1990. "Distinct Character in Hydrophobicity of Amino Acid Compositions of Mitochondrial Proteins." *Proteins: Structure, Function, and Bioinformatics* 8(2): 173–78.

Natan, Eviatar et al. 2018. "Cotranslational Protein Assembly Imposes Evolutionary Constraints on Homomeric Proteins." *Nature Structural and Molecular Biology* 25(3): 279–88.

Natan, Eviatar, Jonathan N. Wells, Sarah A. Teichmann, and Joseph A. Marsh. 2017. "Regulation, Evolution and Consequences of Cotranslational Protein Complex Assembly." *Current Opinion in Structural Biology* 42: 90–97.

Nicholls, Chris D, Kevin G Mclure, Michael A Shields, and Patrick W K Lee. 2002. "Biogenesis of P53 Involves Cotranslational Dimerization of Monomers and Posttranslational Dimerization of Dimers." *Journal of Biological Chemistry* 277(15): 12937–45.

Nissley, Daniel A., and Edward P. O'Brien. 2014. "Timing Is Everything: Unifying Codon Translation Rates and Nascent Proteome Behavior." *Journal of the American Chemical Society* 136(52): 17892–98.

Oberdorf, Richard, and Tanja Kortemme. 2009. "Complex Topology Rather than Complex Membership Is a Determinant of Protein Dosage Sensitivity." *Molecular Systems Biology* 5(1): 253.

Oh, Eugene et al. 2011. "Selective Ribosome Profiling Reveals the Cotranslational Chaperone Action of Trigger Factor in Vivo." *Cell* 147(6): 1295–1308.

Padavannil, Abhilash et al. 2019. "Importin-9 Wraps around the H2A-H2B Core to Act as Nuclear Importer and Histone Chaperone." *eLife* 8.

Pagès, Guillaume, and Sergei Grudinin. 2018. "Analytical Symmetry Detection in Protein Assemblies. II. Dihedral and Cubic Symmetries." *Journal of Structural Biology* 203(3): 185–94.

Pagès, Guillaume, Elvira Kinzina, and Sergei Grudinin. 2018. "Analytical Symmetry Detection in Protein Assemblies. I. Cyclic Symmetries." *Journal of Structural Biology* 203(2): 142–48.

Paila, Umadevi, Brad A. Chapman, Rory Kirchner, and Aaron R. Quinlan. 2013. "GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations." *PLOS Computational Biology* 9(7): e1003153.

Pan, Guangjin et al. 2006. "A Negative Feedback Loop of Transcription Factors That Controls Stem Cell Pluripotency and Self-Renewal." *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* 20(10): 1730–32.

Panasenko, Olesya O. et al. 2019. "Co-Translational Assembly of Proteasome Subunits in NOT1-Containing Assemblysomes." *Nature Structural and Molecular Biology* 26(2): 110–20.

Papp, Balázs, Csaba Pál, and Laurence D. Hurst. 2003. "Dosage Sensitivity and the Evolution of Gene Families in Yeast." *Nature* 424(6945): 194–97.

Parker, J. 1989. "Errors and Alternatives in Reading the Universal Genetic Code." *Microbiological Reviews* 53(3): 273–98.

Pereira-Leal, Jose B., Emmanuel D. Levy, Christel Kamp, and Sarah A. Teichmann. 2007. "Evolution of Protein Complexes by Duplication of Homomeric Interactions." *Genome Biology* 8(4): R51.

Perica, Tina et al. 2012. "The Emergence of Protein Complexes: Quaternary Structure, Dynamics and Allostery." In *Biochemical Society Transactions*, Biochem Soc Trans, 475–91.

Petrovski, Slavé et al. 2015. "The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity" ed. Chris Cotsapas. *PLOS Genetics* 11(9): e1005492.

Pettersen, Eric F. et al. 2021. "UCSF ChimeraX: Structure Visualization for Researchers, Educators, and Developers." *Protein Science* 30(1): 70–82.

Pillai, Arvind S., Georg K.A. Hochberg, and Joseph W. Thornton. 2022. "Simple Mechanisms for the Evolution of Protein Complexity." *Protein Science* 31(11): e4449.

Pizzinga, Mariavittoria et al. 2019. "Translation Factor MRNA Granules Direct Protein Synthetic Capacity to Regions of Polarized Growth." *Journal of Cell Biology* 218(5): 1564–81.

Plaxco, K. W., K. T. Simons, and D. Baker. 1998. "Contact Order, Transition State Placement and the Refolding Rates of Single Domain Proteins." *Journal of Molecular Biology* 277(4): 985–94.

Ponstingl, Hannes, Thomas Kabir, Denise Gorse, and Janet M. Thornton. 2005. "Morphological Aspects of Oligomeric Protein Structures." *Progress in Biophysics and Molecular Biology* 89(1): 9–35.

Poole, Rebecca L. et al. 2023. "Expanding the Neurodevelopmental Phenotype Associated with HK1 de Novo Heterozygous Missense Variants." *European Journal of Medical Genetics*: 104696.

Porta-Pardo, Eduard, Victoria Ruiz-Serra, Samuel Valentini, and Alfonso Valencia. 2022. "The Structural Coverage of the Human Proteome before and after AlphaFold." *PLOS Computational Biology* 18(1): e1009818.

Prasad Bahadur, Ranjit, Pinak Chakrabarti, Francis Rodier, and Joël Janin. 2004. "A Dissection of Specific and Non-Specific Protein–Protein Interfaces." *Journal of Molecular Biology* 336(4): 943–55.

Privalov, P. L., and N. N. Khechinashvili. 1974. "A Thermodynamic Approach to the Problem of Stabilization of Globular Protein Structure: A Calorimetric Study." *Journal of Molecular Biology* 86(3): 665–84.

Pujar, Shashikant et al. 2018. "Consensus Coding Sequence (CCDS) Database: A Standardized Set of Human and Mouse Protein-Coding Regions Supported by Expert Curation." *Nucleic Acids Research* 46(D1): D221–28.

R Core Team. 2023. "R Core Team." *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www. R-project. org.*

Raj, Arjun, Scott A. Rifkin, Erik Andersen, and Alexander van Oudenaarden. 2010. "Variability in Gene Expression Underlies Incomplete Penetrance." *Nature* 463(7283): 913–18.

Rapoport, Tom A., Long Li, and Eunyong Park. 2017. "Structural and Mechanistic Insights into Protein Translocation." *Annual Review of Cell and Developmental Biology* 33(1): 369–90.

Redick, S.D., and J.E. Schwarzbauer. 1995. "Rapid Intracellular Assembly of Tenascin Hexabrachions Suggests a Novel Cotranslational Process." *Journal of Cell Science* 108(4): 1761–69.

Rehm, Heidi L. et al. 2015. "ClinGen — The Clinical Genome Resource." *New England Journal of Medicine* 372(23): 2235–42.

Riba, Andrea et al. 2019. "Protein Synthesis Rates and Ribosome Occupancies Reveal Determinants of Translation Elongation Rates." *Proceedings of the National Academy of Sciences of the United States of America* 116(30): 15023–32.

Richards, Sue et al. 2015. "Standards and Guidelines for the Interpretation of Sequence Variants: A Joint Consensus Recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology." *Genetics in Medicine* 17(5): 405–24.

Robertson, Neil A. et al. 2022. "Longitudinal Dynamics of Clonal Hematopoiesis Identifies Gene-Specific Fitness Effects." *Nature Medicine* 28(7): 1439–46.

Rstudio, Team. 2022. "RStudio: Integrated Development for R."

Saeed, Ramazan, and Charlotte M. Deane. 2006. "Protein Protein Interactions, Evolutionary Rate, Abundance and Age." *BMC Bioinformatics* 7(1): 1–13.

Saito, T. et al. 2017. "A de Novo Missense Mutation in SLC12A5 Found in a Compound Heterozygote Patient with Epilepsy of Infancy with Migrating Focal Seizures." *Clinical Genetics* 92(6): 654–58.

Sasani, Thomas A et al. 2019. "Large, Three-Generation Human Families Reveal Post-Zygotic Mosaicism and Variability in Germline Mutation Accumulation" eds. Amy L Williams, Mark I McCarthy, and Amy L Williams. *eLife* 8: e46922.

Schulz, Luca et al. 2022. "Evolution of Increased Complexity and Specificity at the Dawn of Form I Rubiscos." *Science* 378(6616): 155–60.

Schulz, Luca, Franziska L. Sendker, and Georg K. A. Hochberg. 2022. "Non-Adaptive Complexity and Biochemical Function." *Current Opinion in Structural Biology* 73: 102339.

Schuster-Böckler, Benjamin, Donald Conrad, and Alex Bateman. 2010. "Dosage Sensitivity Shapes the Evolution of Copy-Number Varied Regions." *PloS One* 5(3): e9474.

Schwanhäusser, Björn et al. 2011. "Global Quantification of Mammalian Gene Expression Control." *Nature* 473(7347): 337–42.

Schwarz, Andre, and Martin Beck. 2019. "The Benefits of Cotranslational Assembly: A Structural Perspective." *Trends in Cell Biology* 29(10): 791–803.

Scriver, Charles R., Paula J. Waters, Charles R. Scriver, and Paula J. Waters. 1999. "Monogenic Traits Are Not Simple: Lessons from Phenylketonuria." *Trends in Genetics* 15(7): 267–72.

Seidel, Maximilian et al. 2022. "Co-Translational Assembly Orchestrates Competing Biogenesis Pathways." *Nature Communications* 13(1): 1–15.

———. 2023. "Co-Translational Binding of Importins to Nascent Proteins." *Nature Communications* 14(1): 3418.

Seidman, J. G., and Christine Seidman. 2002. "Transcription Factor Haploinsufficiency: When Half a Loaf Is Not Enough." *The Journal of Clinical Investigation* 109(4): 451–55.

Sepulveda, Guadalupe et al. 2018. "Co-Translational Protein Targeting Facilitates Centrosomal Recruitment of PCNT during Centrosome Maturation in Vertebrates." *eLife* 7.

Shiber, Ayala et al. 2018. "Cotranslational Assembly of Protein Complexes in Eukaryotes Revealed by Ribosome Profiling." *Nature* 561(7722): 268–72.

Shieh, Yu Wei et al. 2015. "Operon Structure and Cotranslational Subunit Association Direct Protein Assembly in Bacteria." *Science* 350(6261): 678–80.

Shihab, Hashem A., Mark F. Rogers, Colin Campbell, and Tom R. Gaunt. 2017. "HIPred: An Integrative Approach to Predicting Haploinsufficient Genes." *Bioinformatics (Oxford, England)* 33(12): 1751–57.

Sievers, Fabian et al. 2011. "Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega." *Molecular Systems Biology* 7(1): 539.

Sneppen, Kim et al. 2010. "Economy of Operon Formation: Cotranscription Minimizes Shortfall in Protein Complexes." *mBio* 1(4): e00177-10.

Stanton, Chloe M. et al. 2017. "Novel Pathogenic Mutations in C1QTNF5 Support a Dominant Negative Disease Mechanism in Late-Onset Retinal Degeneration." *Scientific Reports* 7(1). https://pubmed.ncbi.nlm.nih.gov/28939808/ (December 28, 2021).

Stehr, Henning et al. 2011. "The Structural Impact of Cancer-Associated Missense Mutations in Oncogenes and Tumor Suppressors." *Molecular Cancer* 10(1): 54.

Stein, Kevin C., Allison Kriel, and Judith Frydman. 2019. "Nascent Polypeptide Domain Topology and Elongation Rate Direct the Cotranslational Hierarchy of Hsp70 and TRiC/CCT." *Molecular Cell* 75(6): 1117-1130.e5.

Steinberg, Julia, Frantisek Honti, Stephen Meader, and Caleb Webber. 2015. "Haploinsufficiency Predictions without Study Bias." *Nucleic Acids Research* 43(15): e101.

Stoltzfus, Arlin. 1999. "On the Possibility of Constructive Neutral Evolution." *Journal of Molecular Evolution* 49(2): 169–81.

Sung, Min-Kyung et al. 2016. "Ribosomal Proteins Produced in Excess Are Degraded by the Ubiquitin–Proteasome System." *Molecular Biology of the Cell* 27(17): 2642–52.

Szklarczyk, Damian et al. 2021. "The STRING Database in 2021: Customizable Protein–Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets." *Nucleic Acids Research* 49(D1): D605–12.

Taggart, James C., and Gene-Wei Li. 2018. "Production of Protein-Complex Components Is Stoichiometric and Lacks General Feedback Regulation in Eukaryotes." *Cell systems* 7(6): 580-589.e4.

Tanford, Charles. 1978. "The Hydrophobic Effect and the Organization of Living Matter." *Science* 200(4345): 1012–18.

Tomczak, Maciej, and Ewa Tomczak. 2014. "The Need to Report Effect Size Estimates Revisited. An Overview of Some Recommended Measures of Effect Size." *TRENDS in Sport Sciences* 1(21): 19–25.

Torres, Gonzalo E. et al. 2004. "Effect of TorsinA on Membrane Proteins Reveals a Loss of Function and a Dominant-Negative Phenotype of the Dystonia-Associated DeltaE-TorsinA Mutant." *Proceedings of the National Academy of Sciences of the United States of America* 101(44): 15650–55.

Tunyasuvunakool, Kathryn et al. 2021. "Highly Accurate Protein Structure Prediction for the Human Proteome." *Nature* 596(7873): 590–96.

Tusk, Samuel E., Nicolas J. Delalez, and Richard M. Berry. 2018. "Subunit Exchange in Protein Complexes." *Journal of Molecular Biology* 430(22): 4557–79.

Uhlen, Mathias et al. 2010. "Towards a Knowledge-Based Human Protein Atlas." *Nature Biotechnology* 28(12): 1248–50.

Unwin, P. N. T., and R. Henderson. 1975. "Molecular Structure Determination by Electron Microscopy of Unstained Crystalline Specimens." *Journal of Molecular Biology* 94(3): 425–40.

Vangone, Anna, and Alexandre M.J.J. Bonvin. 2015. "Contacts-Based Prediction of Binding Affinity in Protein–Protein Complexes." *eLife* 4(JULY2015).

Varshavsky, Alexander. 2019. "N-Degron and C-Degron Pathways of Protein Degradation." *Proceedings of the National Academy of Sciences* 116(2): 358–66.

Vasanthakumar, Thamiya et al. 2022. "Coordinated Conformational Changes in the V1 Complex during V-ATPase Reversible Dissociation." *Nature Structural & Molecular Biology 2022 29:5* 29(5): 430–39.

Veis, A., and T. Z. Kirk. 1989. "The Coordinate Synthesis and Cotranslational Assembly of Type I Procollagen." *Journal of Biological Chemistry* 264(7): 3884–89.

Veitia, Reiner A. 2002. "Exploring the Etiology of Haploinsufficiency." *BioEssays* 24(2): 175–84.

———. 2003. "A Sigmoidal Transcriptional Response: Cooperativity, Synergy and Dosage Effects." *Biological Reviews* 78(1): 149–70.

———. 2007. "Exploring the Molecular Etiology of Dominant-Negative Mutations." *Plant Cell* 19(12): 3843–51.

Veitia, Reiner A., S. Caburet, and J.A. Birchler. 2018. "Mechanisms of Mendelian Dominance." *Clinical Genetics* 93(3): 419–28.

Veitia, Reiner A., and Marie Claude Potier. 2015. "Gene Dosage Imbalances: Action, Reaction, and Models." *Trends in Biochemical Sciences* 40(6): 309–17.

Walsh, Ian et al. 2012. "Bluues Server: Electrostatic Properties of Wild-Type and Mutated Protein Structures." *Bioinformatics* 28(16): 2189–90.

Wang, Feng, Larissa A. Durfee, and Jon M. Huibregtse. 2013. "A Co-Translational Ubiquitination Pathway For Quality Control of Misfolded Proteins." *Molecular cell* 50(3): 368–78.

Wang, Gao T., Bo Peng, and Suzanne M. Leal. 2014. "Variant Association Tools for Quality Control and Analysis of Large-Scale Sequence and Genotyping Array Data." *The American Journal of Human Genetics* 94(5): 770–83.

Wang, Guiping et al. 2020. "Spatial Organization of the Transcriptome in Individual Neurons." *bioRxiv*: 2020.12.07.414060.

Wang, Kai, Mingyao Li, and Hakon Hakonarson. 2010. "ANNOVAR: Functional Annotation of Genetic Variants from High-Throughput Sequencing Data." *Nucleic Acids Research* 38(16): e164.

Wang, Mingcong et al. 2015. "Version 4.0 of PaxDb: Protein Abundance Data, Integrated across Model Organisms, Tissues, and Cell-Lines." *Proteomics* 15(18): 3163–68.

Warner, J. R., A. Rich, and C. E. Hall. 1962. "Electron Microscope Studies of Ribosomal Clusters Synthesizing Hemoglobin." *Science (New York, N.Y.)* 138(3548): 1399–1403.

Wells, Jonathan N., L. Therese Bergendahl, and Joseph A. Marsh. 2015. "Co-Translational Assembly of Protein Complexes." *Biochemical Society Transactions* 43: 1221–26.

Wells, Jonathan N, L Therese Bergendahl, and Joseph A Marsh. 2016. "Operon Gene Order Is Optimized for Ordered Protein Complex Assembly." *Cell Reports* 14(4): 679–85.

Winn, Martyn D. et al. 2011. "Overview of the CCP4 Suite and Current Developments." *Acta Crystallographica Section D: Biological Crystallography* 67(4): 235–42.

Wishner, B. C., K. B. Ward, E. E. Lattman, and W. E. Love. 1975. "Crystal Structure of Sickle-Cell Deoxyhemoglobin at 5 Å Resolution." *Journal of Molecular Biology* 98(1): 179–94.

Wodak, Shoshana J., Anatoly Malevanets, and Stephen S. MacKinnon. 2015. "The Landscape of Intertwined Associations in Homooligomeric Proteins." *Biophysical Journal* 109(6): 1087–1100.

Wright, Caroline F. et al. 2023. "Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland." *New England Journal of Medicine* 388(17): 1559–71.

Wright, Sewall. 1929. "Fisher's Theory of Dominance." *The American Naturalist* 63(686): 274–79.

———. 1934. "Physiological and Evolutionary Theories of Dominance." *The American Naturalist* 68(714): 24–53.

Wu, Yingzhou et al. 2021. "Improved Pathogenicity Prediction for Rare Human Missense Variants." *The American Journal of Human Genetics* 108(10): 1891–1906.

Xue, Li C. et al. 2016. "PRODIGY: A Web Server for Predicting the Binding Affinity of Protein-Protein Complexes." *Bioinformatics* 32(23): 3676–78.

Youden, W. J. 1950. "Index for Rating Diagnostic Tests." *Cancer* 3(1): 32–35.

Young, J. C., and D. W. Andrews. 1996. "The Signal Recognition Particle Receptor Alpha Subunit Assembles Co-Translationally on the Endoplasmic Reticulum Membrane during an MRNA-Encoded Translation Pause in Vitro." *The EMBO journal* 15(1): 172–81.

Young, Lindsey N., and Elizabeth Villa. 2023. "Bringing Structure to Cell Biology with Cryo-Electron Tomography." *Annual Review of Biophysics* 52(1): 573–95.

Zemojtel, Tomasz et al. 2014. "Effective Diagnosis of Genetic Disease by Computational Phenotype Analysis of the Disease-Associated Genome." *Science Translational Medicine* 6(252): 252ra123-252ra123.

Zeng, Tony, Jeffrey P. Spence, Hakhamanesh Mostafavi, and Jonathan K. Pritchard. 2023. "Bayesian Estimation of Gene Constraint from an Evolutionary Model with Gene Features." https://www.biorxiv.org/content/10.1101/2023.05.19.541520v1 (July 29, 2023).

Zhang, Jian, and Lukasz Kurgan. 2019. "SCRIBER: Accurate and Partner Type-Specific Prediction of Protein-Binding Residues from Proteins Sequences." *Bioinformatics* 35(14): i343–53.

Zhang, L., V. Paakkarinen, K. J. van Wijk, and E. M. Aro. 1999. "Co-Translational Assembly of the D1 Protein into Photosystem II." *The Journal of Biological Chemistry* 274(23): 16062–67.

Zhao, Bi et al. 2021. "DescribePROT: Database of Amino Acid-Level Protein Structure and Function Predictions." *Nucleic Acids Research* 49(D1): D298–308.

Zhao, Taolan et al. 2021. "Disome-Seq Reveals Widespread Ribosome Collisions That Promote Cotranslational Protein Folding." *Genome Biology* 22(1): 1–35.

Ziegler, Cheyenne, Jonathan Martin, Claude Sinner, and Faruck Morcos. 2023. "Latent Generative Landscapes as Maps of Functional Diversity in Protein Sequence Space." *Nature Communications* 14(1): 2222.

Zipser, D. 1963. "Studies on the Ribosome-Bound Beta-Galactosidase of Escherichia Coli." *Journal of Molecular Biology* 7: 739–51.

Zschocke, Johannes, Peter H. Byers, and Andrew O. M. Wilkie. 2023. "Mendelian Inheritance Revisited: Dominance and Recessiveness in Medical Genetics." *Nature Reviews Genetics*: 1–22.