Natural protein structures have evolved exceptional robustness to mutations

S.H.A. von der Dunk, K. Dingle, A.A. Louis, B. Snel and P. Hogeweg

August 27, 2025

Abstract

Protein structures are often conserved across widely divergent sequences, suggesting high mutational robustness. However, how such robustness emerges through evolution, and how it relates to the underlying sequence–structure map, remains poorly understood. In contrast, the mutational profiles (distribution of structures upon point mutation) of RNA secondary structures are well characterised, exhibiting both high mutational robustness and high evolvability through mutational access to diverse folds. The recent revolution in protein structure prediction now enables analagous large-scale analyses for proteins. Here, we use the structure prediction algorithm ESMFold to systematically investigate the mutational profiles of natural, random, and *de novo* proteins. Unlike RNA, where functional and random sequences share similar mutational profiles, natural proteins are substantially more robust than random amino acid sequences, suggesting an evolutionary drive toward robustness. They also exhibit limited structural variation among close sequence neighbours, potentially constraining access to new folds. Interestingly, many *de novo* proteins do resemble random sequences in their mutational profiles, with low robustness relative to established proteins. These findings reveal how gene duplication and *de novo* gene birth follow distinct evolutionary trajectories toward functional proteins and highlight a potential role for large-effect mutations in the emergence of structural complexity.

Introduction

How functional biopolymers such as RNA and proteins evolve is a fundamental biological question. The sequence–structure–function paradigm assigns a central role to structure, which is specified by sequence through the folding process and, in turn, carries out a function through its three-dimensional physicochemical properties. For proteins, conservation of structure appears to be the norm based on evolutionary data, with many highly divergent sequences still showing functional conservation and structural homology (Orengo et al., 1994; Gabaldón and Koonin, 2013; Lesk, 2010). In rare cases, structural similarity is found in sequences with putative independent evolutionary origins (Orengo et al., 1994; Medvedev et al., 2019; Wright, 2025). These observations show that protein structures have the capacity to be encoded by large numbers of sequences. At the same time, examples of proteins adapting towards new structures are relatively rare (e.g. Dishman et al., 2021). In fact, most extant protein domains likely emerged prior to the appearance of the Last Universal Common Ancestor (LUCA; Ranea et al., 2006).

As a consequence of the complexity of protein folding, many insights have been derived from related but simpler maps. RNA secondary structure is an intermediate level of abstraction that can be studied more easily and therefore has served as a paradigm for molecular genotype–phenotype maps (Schuster et al., 1994; Hogeweg, 2012; Manrubia et al., 2021). Using well-established tools for secondary structure prediction, computational analyses have revealed that RNA structures are both robust and evolvable (Huynen et al., 1996; Huynen, 1996; Wagner, 2005, 2008), and this is supported by experimental data (Schultes and Bartel, 2000). Briefly, sequences encoding the same secondary structure tend to form large neutral networks that can be traversed by mutations. These neutral networks grant RNA immediate robustness to mutations, and also long-term evolvability by allowing neutral sequence divergence to explore large regions of sequence space (Huynen et al., 1996; Huynen, 1996; Wagner, 2005, 2008). Mutational robustness and evolvability facilitate structural adaptation and can be further increased by adaptive forces (Wilke, 2001; Sanjuán et al., 2007; Wagner, 2023), phenotypic bias in the sequence–structure map (Dingle et al., 2018; Johnston et al., 2022) and neutral evolution (Van Nimwegen et al., 1999). Similar properties have been discovered in other genotype–phenotype maps—including gene regulatory networks

(Ciliberti et al., 2007), metabolic networks (Catalán et al., 2018), and simplified models of protein folding (e.g. Lipman and Wilbur, 1991; Greenbury et al., 2016)—suggesting that robustness and evolvability are fundamental to genotype–phenotype maps (Wagner, 2008; Ahnert, 2017). Yet for protein structure, the complex folding process has so far prevented large-scale investigation of evolutionary dynamics.

The recent revolution in protein structure prediction (Jumper et al., 2021) grants the opportunity to evaluate the sequence–structure relationship for proteins in a large-scale fashion. We use ESMFold, an already well-established algorithm for protein structure prediction directly from single sequence (e.g. Meier et al., 2021; Xu et al., 2021; Verkuil et al., 2022; Lin et al., 2023; Sahakyan et al., 2025, see also Appendix D), to explore the mutational profile and robustness of proteins at the backbone structure level across sequence space. Our main finding is that natural proteins have evolved backbone structures that are robust to point mutations. We provide support for this claim by exploring different protein datasets and assessing correlations of robustness with taxonomy, function and a suite of structural features. The largest difference in mutational robustness is between natural proteins and random sequences of fixed length, but viral proteins and *de novo* proteins also show limited robustness similar to random sequences. Thus, using the predictions of ESMFold, we uncover a unique quality of protein evolution with striking implications for molecular evolution and genome evolution.

Results

Mutational profile of natural RNA structures is close to random

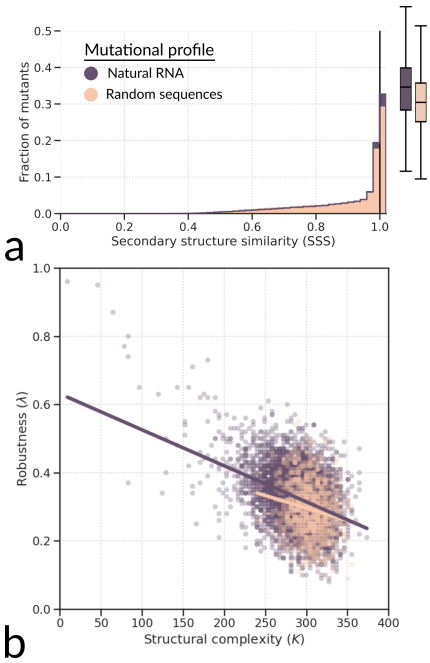

Figure 1 – The mutational profile of natural RNA ($N = 5451$) resembles that of random nucleotide sequences ($N = 1000$) in terms of mutational robustness at secondary structure level. (**a**) Most RNA molecules have both identical and very different secondary structures in their mutational neighbourhood (as different as two unrelated random structures, see *i.e.* random wildtype similarity). In Fig. S15a, individual mutational profiles of several natural RNA molecules and random sequences are shown. (**b**) The number of neutral neighbours (λ, corresponding to the far-right bar in panel **a**) summarizes the mutational robustness for each molecule and correlates with structural complexity, both for natural and random RNA. The variation in the fraction of neutral neighbours (λ) across RNA molecules is shown with boxes to the right of panel **a** denoting the median and interquartile range.

We analyze the mutational profiles—the distributions of phenotypes one mutation away from the wildtype—of RNA secondary structures and protein backbone structures with the same methodological framework, focusing on polymers of length $L = 300$ (a typical length for both types; see Fig. S4 in Appendix A for results on $L = 100$). For RNA, we use the popular ViennaRNA package to predict secondary structures from sequences, *i.e.* for wildtypes and their sampled point mutants. Using dot-bracket notation, the secondary structures of mutants are aligned to the wildtype and secondary

structure similarity (SSS) is calculated for each as the global Needleman-Wunsch alignment score (using 1 for match, 0 for mismatch and -1 for gap Needleman and Wunsch, 1970). Thus SSS= 1 corresponds to complete alignment, and SSS= 0 to no similarity. It is known that a sizable fraction λ of single point mutants ($\sim 30\%$) is typically neutral (SSS=1), folding into the same secondary structure as the wildtype and providing mutational robustness (Schuster et al., 1994; Von der Dunk et al., 2025). Concurrently, some mutants fold into very different structures, providing the wildtype RNA with phenotypic variation for adaptation. We here recover such broad mutational profile for natural RNA sequences from the RNAcentral database (Fig. 1a; The RNAcentral Consortium, 2021). Strikingly, we find a similar mutational profile for randomly generated nucleotide sequences, as previously observed for shorter RNA of $L = 55$ (Dingle et al., 2015) and in line with other studies showing that natural RNA structures are similar to the structures of random sequences (Fontana et al., 1993; Jörg et al., 2008; Dingle et al., 2015, 2022a; Ghaddar and Dingle, 2023; Von der Dunk et al., 2025). Furthermore, previous work shows that high-frequency RNA structures are more robust (Greenbury et al., 2016; Mohanty et al., 2023) and less complex as measured by compression of their secondary structure representation (Dingle et al., 2018; Von der Dunk et al., 2025, see Methods). Together, these points suggest that a negative correlation exists between mutational robustness and structural complexity, and our analysis indeed uncovers such relationship for RNA (Fig. 1b). In sum, the highly comparable mutational profiles of natural and random RNA structures, which aligns with the structural similarities between these two sets previously described in the literature, shows that RNA structures are inherently robust and evolvable.

Natural protein structures are exceptionally robust to mutations

Adapting our mutational profiling approach to proteins, we use ESMFold for structure prediction from sequences. Although mutational effects can be measured directly with the tertiary structures provided by ESMFold (e.g. using TM-align; Zhang and Skolnick, 2005), we find that coarse-graining to backbone structures (using DSSP; Touw et al., 2015) followed by alignment is more suitable for interpreting changes in disordered loops whose three-dimensional structure is inherently unstable. Still, systematic comparisons show that our results as presented below are consistent under diverse distance measures with different levels of structural detail, including tertiary structure (see Appendix D). The map from sequence to backbone structure, in which each residue is represented by one of 8 possible conformations, is very similar to the sequence to secondary structure map in RNA. In both cases, a sequence of length L (with alphabets of sizes 4 and 20 for RNA and proteins, respectively) is mapped onto a structure description of the same length L (with alphabets of sizes 3 and 8), facilitating comparison between these two types of molecules. Hence, we obtain the mutational profile in a similar manner, *i.e.* using global Needleman-Wunsch alignment of the backbone structures to obtain backbone structure similarity (BBS) for each mutant.

Unlike for RNA, the mutational profiles of natural proteins are very different from those of random sequences (Fig. 2a). Natural protein structures are generally robust to mutations, in line with high levels of sequence divergence frequently observed between orthologous proteins. Roughly a quarter of single point mutants has an identical structure as the wildtype, but the majority is slightly different. Mutants with very different predicted structures (e.g. < 0.8 backbone similarity) are extremely rare for natural proteins, unlike for natural RNA. In contrast, proteins with random amino acid sequences are not robust: on average, mutants share roughly 0.8 of their backbone with the wildtype, and no neutral mutants were found in the entire dataset. Thus, natural proteins appear to form a special subset of all possible sequences.

Both natural proteins and random sequences have a much narrower predicted mutational profile than observed for RNA. As a consequence, we find that the mutational profile of proteins is best characterized by the mean similarity of mutants to the wildtype (see Fig. S15b), while that of RNA molecules is best characterized by the fraction of identical mutants (λ). The qualitatively distinct mutational profiles of RNA and protein structures are further supported by very different spatial patterns of mutation-induced structural change (Appendix B). We are able to find substantial variation in the mean similarity of mutants among natural proteins (see boxes in Fig. 2a–b). In particular, bacterial proteins are more robust than eukaryotic proteins, and viral proteins are least robust in our dataset. Further investigation buttressed these newly uncovered high-level taxonomic differences for protein structures (see Appendix A), suggesting that selection on mutational profiles is different in organisms with different macro-evolutionary patterns. Thus, in addition to revealing large differences between natural proteins and random sequences, our analysis uncovers meaningful trends within natural proteins, where the predictive power of ESMFold has been established (Notin et al., 2023).

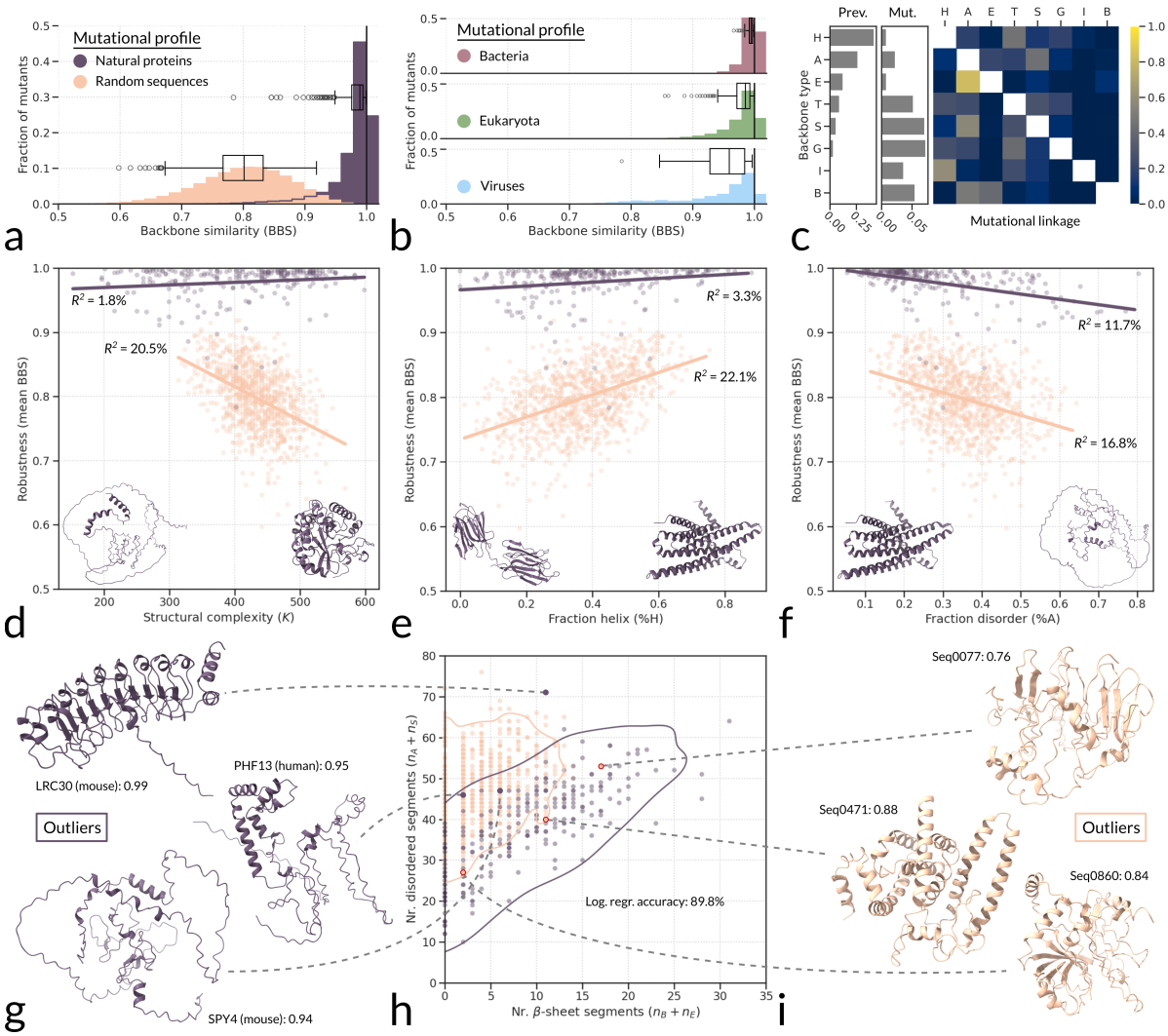

Figure 2 – Natural proteins ($N = 249$) and random amino acid sequences ($N = 1000$) differ significantly in mutational robustness at the structural level, considered at a fixed length of $L = 300$. (**a**) In contrast to RNA (Fig. 1), natural proteins are robust to mutations while random sequences are not robust. In Fig. S15b, individual similarity distributions of several natural proteins and random sequences are shown. (**b**) Mean backbone similarity of mutants of natural proteins from different taxonomic origins (bacteria, $N = 76$; eukaryota, $N = 159$, viruses $N = 14$, see Appendix A). Boxplots in **a** and **b** show the distribution of the mean backbone similarities per protein. (**c**) Distribution of backbone types in natural proteins (Prev: prevalence), their vulnerability to mutation (Mut: mutated frequency) and the likely outcome of mutation (mutational linkage, cf. Fig. S16). Robustness, here represented by the mean backbone similarity of mutants, correlates strongest with (**d**) structural complexity and (**e**) helical fraction within random sequences, and with (**f**) disorder within natural proteins. (**h**) Between the two sets—natural proteins and random sequences—the two most distinguishing features are the number of disordered segments and the number of beta-sheet segments. Structural outliers of (**g**) natural proteins resembling random sequences and (**i**) random sequences resembling natural proteins still feature high and low mutational robustness, respectively (values indicated with the labels).

To better understand what causes the variation in predicted mutational robustness, we first asked whether robustness is related to specific structural characteristics. Even though we use a different measure for robustness between RNA and proteins (see definitions above), robustness correlates negatively with structural complexity in random protein sequences just as found for RNA (Fig. 2d) (see also Von der Dunk et al. (2025)). By contrast, natural proteins have high robustness irrespective of complexity, suggesting that evolution can overcome the robustness–complexity tradeoff that applies to most of protein sequence space as revealed by random sampling. For random sequences, robustness correlates positively with the α-helical fraction (Fig. 2e), which is the most prevalent backbone type and more robust to mutations than other backbone conformations including disorder, the second-most prevalent type (Fig. 2c, S16). For natural proteins, the fraction of disorder is the strongest correlate with robustness (Fig. 2f), and likely explains the higher robustness observed in bacteria (mean disorder 17.9%) compared to eukaryotes and viruses (means 31.7% and 32.0%, respectively). The negative correlation between disorder and mutational robustness could suggest that disordered proteins are inherently unrobust, or that their function is established through different structural constraints.

None of the structural characteristics that correlate with robustness can on their own explain the large difference in robustness *between* natural proteins and random sequences, despite explaining variation within each set to a certain degree (Fig. 2d–f). In most features—including structural complexity, α-helical fraction and disordered fraction—both datasets overlap substantially in range and natural proteins cannot be separated from random sequences. However, using logistic regression, two composite features—the number of β-sheets (including isolated β-bridges) and the number of disordered segments (including bends)—can discriminate most natural proteins from random sequences (89.8% accuracy for balanced sample sizes of $N = 249$). Specifically, natural proteins have more β-sheets and fewer disordered segments than random sequences. Still, these structural differences do not explain the disparity in robustness. For example, three natural protein outliers that resemble random sequences in the two relevant features show remarkably high robustness (Fig. 2g). Conversely, three random sequence outliers that resemble natural proteins in the two relevant features and which even appear to form compact globular folds, show remarkably low robustness (Fig. 2i).

Natural proteins are structurally diverse

The exceptionally high mutational robustness predicted for natural proteins could suggest that they occupy a limited region of sequence and structure space. Indeed, at the sequence level, we find that natural proteins are slightly more similar to each other than random sequences pointing to a low degree of clustering (Fig. 3a, S18a). In contrast, at the level of backbone structure, natural proteins are more diverse than random sequences (Fig. 3b, S18b). The higher structural diversity of natural proteins is even more pronounced when projected into two or more principal components based on 176 structural features (Fig. 3c–d; Suppl. Data S1). Finally, natural proteins occupy a much larger range of structural complexity than random sequences (Fig. 3e), which is opposite from natural RNA whose structural complexity distribution is very similar to that of random nucleotide sequences (Von der Dunk et al., 2025). All these observations indicate that evolution has shaped extant proteins to display a wide range of structural phenotypes compared to random sequences, in addition to selecting for mutational robustness.

To get a sense of the proximity between natural proteins and random sequences in sequence and structure space, we performed several iterations in which natural proteins were mutated at 50 additional sites, *i.e.* resulting in natural proteins with 50, 100, 150, 200 and 250 sites mutated. It takes as many as 150 mutations (*i.e.* randomizing half of the protein sequence) to arrive at a mutational profile that resembles that of random sequences—in particular, where neutral mutants are rarely observed anymore (Fig. 3f). This implicates dramatic change during the evolutionary construction of natural proteins and suggests that regions of high robustness and regions of low robustness (*i.e.* the bulk of the protein folding landscape) are far apart in sequence space.

The genetic code promotes robustness slightly

The relative abundance of amino acids in natural proteins differs markedly from uniform, so we asked whether amino acid frequencies could help to explain the exceptional robustness of natural proteins. The genetic code impacts the relative abundance of amino acids—by the number of codons assigned to each—as well as their mutational linkage—by the similarity between codons of different amino acids. Previous

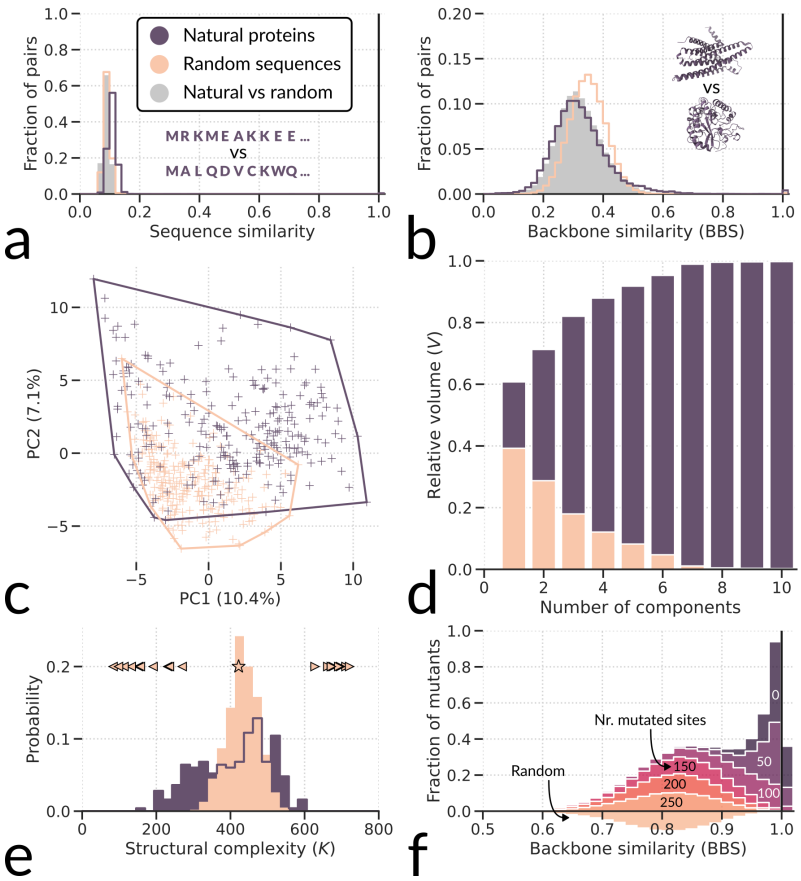

Figure 3 – Natural proteins are structurally more diverse than random amino acid sequences. The distribution of (**a**) sequence similarity and (**b**) backbone similarity between pairs of unrelated natural proteins and pairs of unrelated random sequences (see also Fig. S18 for other random sequence sets). In grey, the similarity distributions of randoms pairs between sets are shown. (**c**) The first two principal components produced from 176 structural features (Suppl. Data S1) suggest that natural proteins are more diverse than random sequences, as they occupy a larger volume (convex hull). (**d**) With additional principal components, the relative volume taken up by natural proteins becomes even larger (first 10 components account for 42.2% of variation in the data). (**e**) Structural complexity of natural proteins covers a larger range than that of random sequences. Triangles show the end result of *in silico* evolution (for $t = 1000$ generations) with two random proteins towards low or high complexity (5 replicates per direction per protein, see Appendix C). (**f**) Mutational neighbourhood of natural proteins slowly converges to that of random sequences after accumulating mutations, shown per 50 sites.

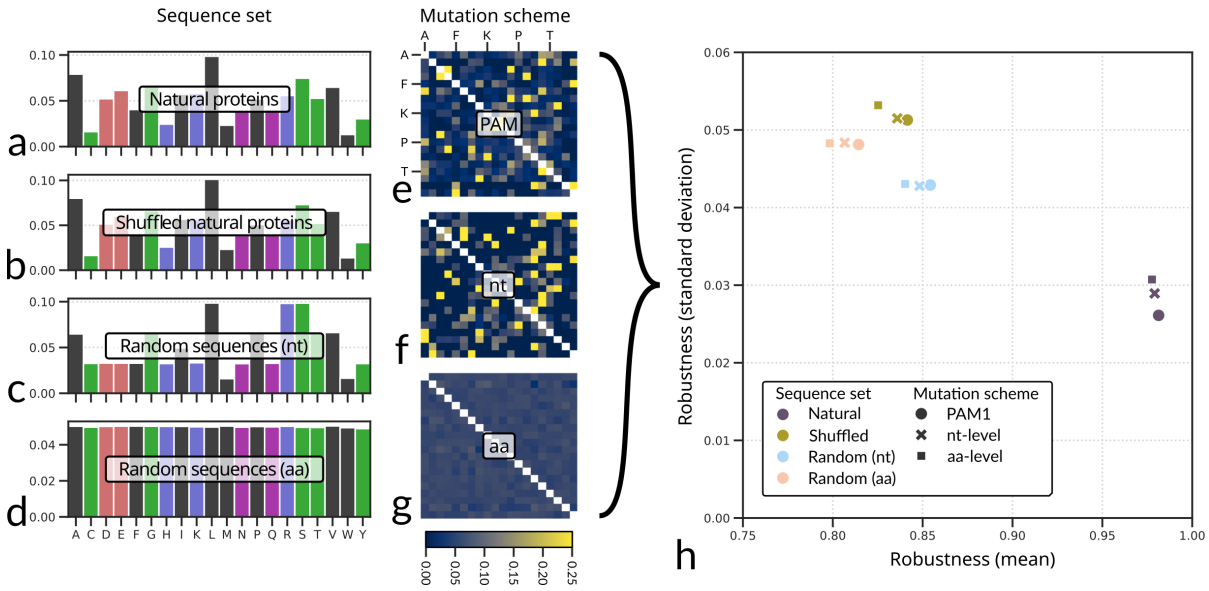

Figure 4 – The genetic code favours mutational robustness in protein sequences of length $L = 300$. (**a**–**d**) Frequency distributions of amino acids in natural proteins and different random sequence sets. (**e**–**g**) Amino acid substitution rates realized by different mutational schemes. (**h**) Comparison of mean and standard deviation of robustness distributions from all 12 mutation assays (4 sequence sets times 3 mutation schemes).

work suggests that the standard genetic code promotes mutational robustness of protein structures by limiting transitions between residues with very different polar requirement (Freeland and Hurst, 1998). Our framework grants an opportunity to test this hypothesis by using protein structure prediction to measure the impact of the genetic code on mutational robustness.

First, we created two additional sets of "random sequences" of length $L = 300$: (i) translated random DNA sequences with equal nucleotide frequencies, such that amino acid frequencies reflect redundancies in the genetic code, and (ii) shuffled natural proteins—two scrambled sequences for each natural protein—with the same amino acid frequencies as natural proteins, reflecting redundancies in the genetic code as well as the effects of selection. Second, we used two additional mutation schemes to sample mutations and measure robustness: (i) non-synonymous nucleotide transitions incorporating the relative purine–pyrimidine bias, which captures the mutational linkage of the genetic code, and (ii) amino acid transitions as observed in closely related proteins defined by the PAM1 substitution matrix (Dayhoff, 1972), which captures the linkage of the genetic code as well as the effects of selection.

To facilitate comparison between 12 mutational profiles, we focus on their means and standard deviations (Fig. 4h). These two measures are related across sequence sets. In natural proteins, the mutational profiles are sequeezed against the upper limit in backbone similarity (*i.e.* 1), resulting in high means and corresponding low standard deviations.

Protein sequences constructed from random nucleotide sequences display slightly higher mutational robustness than random amino acid sequences, indicating that the genetic code promotes robustness by directing evolution towards a particular set of sequences (Fig. 4a–d,h). As shuffled proteins are on average slightly less robust than proteins built from random nucleotide sequences, it is not just the amino acid frequency distribution of natural proteins that explains their high mutational robustness. For instance, a high frequency of aspartic acid (D) and glutamic acid (E) relative to random nucleotide sequences (compare red bars in Fig. 4b–c) is not guaranteed to increase robustness, even though it apparently does so in the specific context found in natural proteins.

In line with previous work (Freeland and Hurst, 1998), the genetic code buffers against structurally harmful mutations (Fig. 4e–h). Non-synonymous nucleotide mutants are on average more similar to the wildtype than random amino acid mutants. Nevertheless, the effect of the genetic code in linking structurally similar amino acids is very small (e.g. increasing robustness from 0.84 to 0.85 in random nucleotide sequences) compared to the effect of the genetic code in sampling more robust genotypes in the first place (e.g. increasing robustness from 0.81 to 0.85 from random amino acid sequences to random nucleotide sequences). The application of PAM1 transition probabilities increases robustness slightly further, suggesting that some of the possible non-synonymous sequence variations have been selected out from the PAM1 homologs based on their deleterious effect on structure. In sum, while there are small

effects of the genetic code on mutational robustness of protein structures, these are far from sufficient to explain the high mutational robustness demonstrated by natural proteins.

De novo yeast proteins resemble random sequences

Our results seem to suggest that substantial evolution is required to increase the mutational robustness of proteins relative to that of randomly generated and non-evolved sequences. Following this reasoning, we might ask how robust *de novo* proteins are which have a hypothesized recent origin from non-coding DNA (Levine et al., 2006; Begun et al., 2007). *De novo* proteins were once thought to be rare, but as more and more candidates are being identified, *de novo* gene birth may actually represent a major route for the origin of new proteins (Magadum et al., 2013; McLysaght and Guerzoni, 2015). Thus, we now turn to the *Saccharomyces cerevisiae* proteome, a large and diverse set of proteins with a length range of $50 \leq L \leq 2014$ (the upper boundary being the limit for ESMFold) and a small number of annotated *de novo* proteins (Vakirlis et al., 2020). Mutational robustness is measured at the nucleotide level (*i.e.* sampling unique non-synonymous mutations) to incorporate the subtle effects of the genetic code discovered in the previous section. Our findings as described below for yeast are consistent with proteome analyses of *Oryza sativa* and *Drosophila melanogaster* for which *de novo* annotation is also available (see Appendix A).

Established (*i.e.* non-*de novo*) yeast proteins, ranging more than an order of magnitude in length, feature similar high mutational robustness as natural proteins of length $L = 300$ (Fig. 5a). Yeast proteins with a molecular function in binding, catalytic activity or transporter activity are in particular associated with high mutational robustness, as determined by GO analysis (Fig. 5b–d). In contrast, *de novo* proteins vary much more in robustness, with some displaying robustness similar to that of established proteins and others displaying robustness similar to random sequences. Moreover, *de novo* proteins generally follow the robustness versus complexity correlation that was also observed for random sequences (Fig. 5e), unlike established proteins for which disorder is the best correlate (Fig. 5f). Thus, in agreement with Edgell et al. (2020), we find that most proteins that recently emerged from untranslated DNA are not yet robust and still resemble random sequences.

To see if *de novo* proteins have unique structural characteristics that underlie their low robustness, we searched for structural analogs using FoldSeek. Only 10 out of 188 *de novo* proteins produced a significant hit (E-value < 0.05) in either the ESM Metagenomic Atlas or a subset of 13 eukaryotic species from the AlphaFold database (Fig. 5h). For two of these (`YOL164W` and `YLL060C`), structural similarity with an established protein was unexpectedly high. Further investigation of these two proteins revealed a large number of statistically significant bacterial homologs (pHMMer $\log(E) < -100$, see Suppl. Data S2, S3), indicating that they likely originate from horizontal gene transfer and are wrongly annotated as *de novo* proteins. Interestingly, this conclusion is supported by our predicted high robustness, especially given their high structural complexity (Fig. 5e). Three *de novo* proteins (`YJR108W`, `YER188C-A` and `YLR036C`) hit other yeast *de novo* proteins, which also show limited robustness. For the remaining five proteins however (bold labels in Fig. 5h), the structural analog is more robust to mutations than the *de novo* protein. These data suggest that there are no inherent physical constraints that prevent *de novo* proteins from achieving higher robustness. Rather, it appears that *de novo* proteins have not had enough time to evolve high mutational robustness to the same level of older established proteins (see Appendix C).

De novo proteins are known to be relatively short, and we find that their length partly explains their low robustness. Namely, for proteins below a certain length ($L \approx 250$), mutational robustness decreases as protein length further decreases (Fig. 5g; see Appendix A). The broad predicted mutational profile of individual *de novo* proteins (see Fig. S19) seems to arise from two processes: the relative impact of a point mutation is larger in short proteins (promoting large mutational effects), while at the same time short proteins have fewer residues increasing the probability that they are all conserved (see also Fig. S4). Thus, short random proteins are more robust and evolvable, like RNA, and present a more suitable starting point for the evolution of novel structures than longer sequences. In line with this, we find that ancient proteins are longer and their structures more robust to mutations (see Appendix A, Fig. S20).

Mutational robustness buffers transcriptional and translational errors

We have established that natural protein structures are highly robust to mutations, as compared to random sequences. Yet, there is substantial variation in robustness among yeast proteins, depending on protein length, disorder and *de novo* status. Based on the sequence–structure–function paradigm,

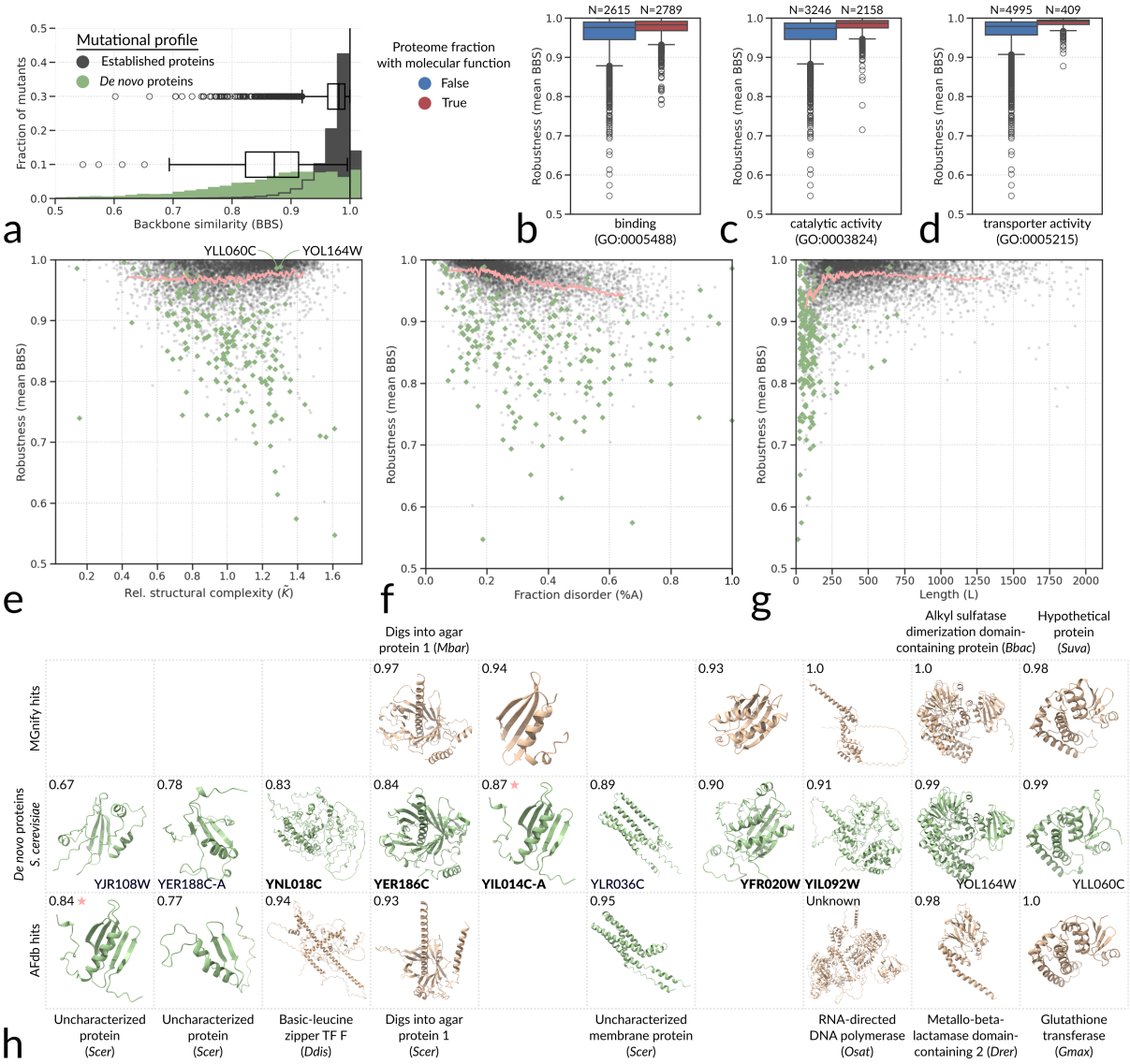

Figure 5 – *De novo* proteins in *S. cerevisiae* have relatively low robustness similar to random sequences. Prediction of *de novo* origin was obtained from Vakirlis et al. (2020). (**a**) *De novo* proteins ($N = 188$) are less robust than established proteins ($N = 5217$). (**b**–**d**) Three top-level GO terms from the molecular function namespace are positively associated with high robustness. (**e**) Robustness correlates with relative structural complexity in *de novo* proteins. In established proteins, robustness correlates (**f**) negatively with disorder and (**g**) positively with length. Red lines in **e**–**g** represent rolling means for established proteins. (**h**) Proteins with structural similarity to *de novo* proteins found in MGnify or AFdb as detected by FoldSeek have higher robustness (values indicated in top-left corners). The red star marks the *de novo* protein `YIL014C-A`, which is hit by another *de novo* protein (`YJR108W`), and also hits an unknown protein in MGnify; robustness was independently assayed for both cases, resulting in two distinct yet similar values. The FoldSeek hits of two highly robust *de novo* proteins which are also very complex, `YOL164W` and `YLL060C` (see also panel **e**), are highly suspicious, and suggest horizontal gene transfer instead of *de novo* origin.

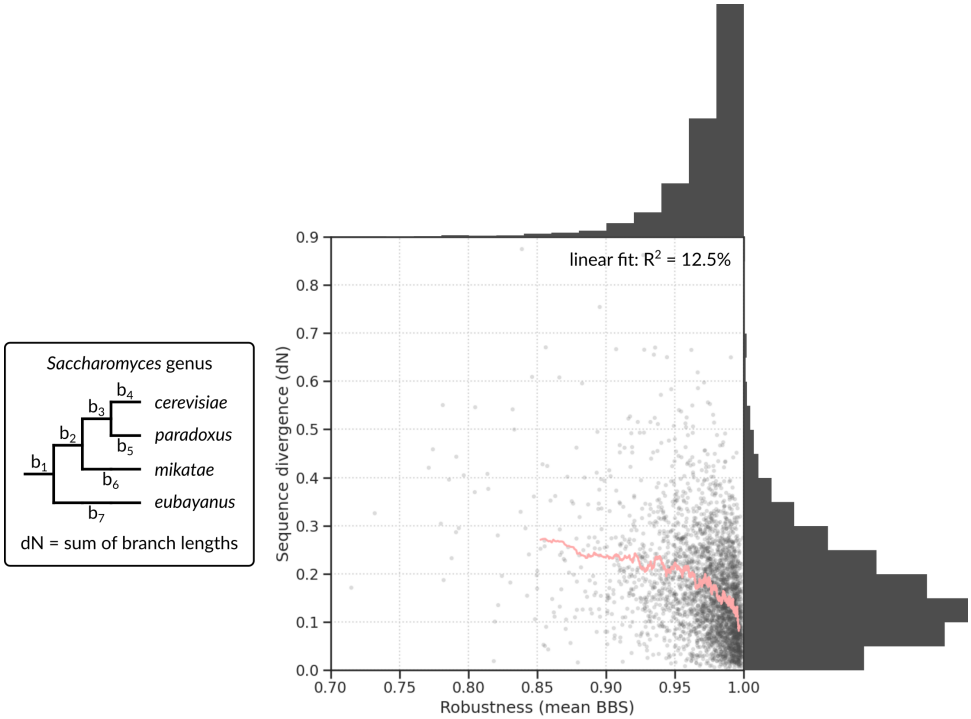

Figure 6 – Correlation between mutational robustness and rate of sequence divergence in the *Saccharomyces* genus. Sequence divergence was inferred using CodeML (Yang, 2007) for genes with clean 1-to-1 orthology between the four depicted species ($N = 3220$, $\sim 62\%$ of the proteins shown in Fig. 5).

we would predict that proteins with higher mutational robustness can accommodate more sequence divergence over evolutionary time while still conserving structure and function. Sequence divergence has been studied extensively in yeast (Pál et al., 2001; Wall et al., 2005; Drummond et al., 2005a, 2006; Serohijos et al., 2012, 2013), allowing us to explore its correlation with our mutational robustness predictions.

Following previous studies, we calculated sequence divergence for a large number of established proteins as the rate of non-synonymous substitutions dN between four closely related yeast species—*S. cerevisiae*, *S. paradoxus*, *S. mikatae* and *S. eubayanus*. Drummond et al. (2006) pinpointed gene expression level as the single strongest correlate with sequence divergence, accounting for 28.8% of variation. We here find that mutational robustness is able to explain a small but significant fraction of 12.1% of variation in sequence divergence (Fig. 6; $p \approx 10^{-85}$). Surprisingly though, the correlation between mutational robustness and sequence divergence is negative rather than positive: proteins with the most robust structures show the least sequence divergence. The negative correlation between mutational robustness and sequence divergence suggests that highly expressed proteins which, as pointed out by Drummond et al. (2006), evolve slower at the sequence level, also experience stronger selection for mutational robustness of their structure. A possible explanation for this is that errors in transcription and translation of highly expressed proteins pose a burden for the cellular machinery that protects against misfolded proteins, and thereby could drive selection for mutational robustness. Given that errors in transcription and translation occur at much higher rates than errors in replication (Mordret et al., 2019), the former may in fact be what drives evolution towards stable protein structures with limited evolutionary access to structural variation.

Discussion

In order to understand how proteins evolve, we have explored the effect of mutations on protein structure. We found that the predicted protein structure landscape is very different from the RNA structure landscape, with proteins showing evolution towards mutational robustness and limited access to structural variation. We pinpointed large differences in mutational robustness between natural proteins and folds from random sequences, but also variation among natural proteins linked to variation in structural

properties, taxonomic origin, molecular function and *de novo* status. The low mutational robustness of many *de novo* proteins is especially striking and supports our results for random sequences where the predictive power of ESMFold has not been proven.

We began our analysis with a direct comparison between RNA and proteins. As mentioned, RNA folding combines mutational robustness with evolvability and as a consequence has been dubbed an "ideal evolvable molecule" (Schuster et al., 1994; Hogeweg, 2012). We recovered the broad mutational profile at the secondary structure level, and corroborated the similarity between natural RNA molecules and random sequences described in previous work (Fontana et al., 1993; Smit et al., 2006; Jörg et al., 2008; Stich et al., 2008; Dingle et al., 2015, 2022a; Von der Dunk et al., 2025). In contrast, protein structures seem to face a tradeoff between robustness and evolvability: natural proteins feature high robustness but limited structural variation in their mutational neighborhood, while random sequences lack robustness but show large structural variation in their mutational neighborhood. Differences between natural proteins and random sequences have been found before (e.g. De Lucrezia et al., 2012; Ferrada and Wagner, 2012), although recent work instead highlighted similarities (Tretyachenko et al., 2017). Our findings indicate that intuitions drawn from the RNA sequence-to-structure relationship might not readily apply to protein backbone structures and that, contrary to earlier assertions (Hogeweg, 2012; Wagner, 2013), the evolutionary dynamics of RNA folding are not universally applicable across genotype–phenotype maps.

New protein structures may evolve via two main routes: gene duplication followed by neo-functionalization and *de novo* gene birth. Our observations indicate that both evolutionary routes face substantial challenges. Gene duplication creates a copy of a pre-existing robust structure with limited access to new structures through mutation. With most mutations yielding slightly different structures, duplicated proteins would have to accumulate substantial sequence divergence to evolve a new structure. Along the way, intermediate structures may be disfavored by selection preventing such adaptation. As a result, folds of duplicated proteins may be more likely to be conserved and repurposed for a new cellular niche rather than being substantially modified (Cardarelli et al., 2010). In contrast, *de novo* gene birth directly creates novel structures with additional access to new structural variants through mutation. However, these structures all feature low mutational robustness and would therefore be difficult to conserve over evolutionary timespans.

Based on our findings, we propose that the high robustness of protein structures to point-mutations is the result of selection against transcriptional and translational errors. Specifically, mutational robustness correlates negatively with sequence divergence in *S. cerevisiae*, whereas selection for mutational robustness would have shown a positive correlation. This suggests that evolvability of protein structures, *i.e.* access to novel structures through point-mutations, is limited by constraints on their faithful expression in large copy numbers, unlike for RNA structures which are inherently robust to mutations.

One possibility for evolving new and robust protein structures suggested by our work is to start with a short protein, as in the case of *de novo* and unconserved proteins (Lipman et al., 2002). Short proteins have a relatively broad mutational profile, which allows them to navigate the sequence–structure map more efficiently, similar to RNA. Once a suitable structure is found, protein length may be gradually increased to increase mutational robustness. Interestingly, this scenario fits with a recent study on evolutionary simulations of short random sequences selected for stability (Sahakyan et al., 2025). In this study, stable proteins of length $L \approx 50$ rapidly evolved from short peptides of length $L = 12$. Large mutations such as partial duplications played an important role, allowing evolution to duplicate stable substructures to form complete globular domains. In the field of protein design, large mutations are also leveraged to efficiently search the enormous sequence space without getting stuck in local optima (Drummond et al., 2005b; Romero and Arnold, 2009; Fram et al., 2024). Thus, through large mutations provided by nature and exploitation of modularity, protein structures may be substantially more evolvable than predicted from the effects of point mutations.

We conclude that, based on the current state-of-the-art methods for protein structure prediction, protein structures seem to have unique evolutionary properties. While RNA structures are inherently robust to mutations due to the properties of RNA folding, protein structures as found in the natural world have had to evolve such robustness. The distinct sequence–structure maps of RNA and proteins as well as their evolved properties yield distinct modes of molecular evolution, presenting new avenues for further investigation. In particular, where RNA can efficiently explore large sequence and structure spaces, evolution of protein structures may proceed through slow and gradual change, or through evolution of sequence length and large mutations allowing modular evolutionary construction.

Methods

Mutational robustness of RNA

To measure the mutational robustness of an RNA molecule, we randomly sample 100 point mutants. We then predict the secondary structure of the wildtype and all mutants, stored as dot-bracket structures representing the bonded state of each nucleotide: '`(`' for opening a bond, '`)`' for closing a bond, and '`.`' for no bond. Thus, RNA folding maps 4^L genotypes (4 possible characters per site, for length L) to 3^L possible phenotypes (3 possible characters per site), although physical constraints further limit the number of possible phenotypes to approximately 0.13×1.76^L (Dingle et al., 2015). We then align each mutant secondary structure to that of the wildtype, calculating similarity as the global Needleman-Wunsch alignment score (using 1 for match, 0 for mismatch, and -1 for gap; Needleman and Wunsch, 1970). The resulting distribution of secondary structure similarities represent the mutational profile (in practice, $N < 100$ due to sampling with replacement), which can be combined across multiple wildtypes as in Fig. 1a. We also summarize the robustness for a single wildtype as the fraction of neutral mutants (*i.e.* $SSS = 1.0$; as in Fig. 1b).

Structural complexity of RNA

To calculate structural complexity of an RNA molecule, the dot-bracket representation for secondary structure is converted into a binary string: '`(`' → '`10`', '`)`' → '`01`', '`.`' → '`00`'. The binary string is then compressed using Lempel-Ziv compression (Lempel and Ziv (1976) as detailed in Dingle et al. (2018)), with the resulting number of bits representing the complexity of the original object. This measure of complexity originates from the field of algorithmic information theory, and has found a valuable application in RNA folding due to the existence of a phenotypic bias that allows an upper limit of the neutral set size of RNA phenotypes to be predicted from their structural complexity.

Mutational robustness of proteins

The procedure for measuring mutational robustness of proteins is similar to that of RNA (see above). However, we use ESMFold to predict tertiary structure followed by DSSP to extract the backbone structure, which is also often referred to as secondary structure (Kabsch and Sander, 1983; Lin et al., 2023). In these backbone structures, each residue in a protein sequence is assigned to one of eight conformational classes (`H`: α-helix, `B`: isolated β-bridge, `E`: β-sheet, `G`: 3_{10}-helix, `I`: π-helix, `T`: turn, `S`: bend, `A`: disorder). Thus, protein folding maps 20^L genotypes (amino acid sequences of length L) to 8^L naively possible phenotypes. Compared to RNA folding, the protein folding map thus has a much larger genotype space and a much larger degree of redundancy given the same polymer length, even in the most conservative scenario where all 8^L protein phenotypes are physically feasible.

Alignment of mutants to the wildtype and calculation of backbone similarities are performed in the same way as for RNA. In the case where mutations are sampled at the nucleotide level, the actual number of sampled amino acid mutants is $N \approx 75$ after synonymous mutants and mutants with early stop-codons are filtered out. Besides analysis of the full mutational profile (as in Fig. 2a), we also summarize the mutational robustness for a single wildtype as the mean backbone similarity of mutants to the wildtype (as in Fig. 2b). While previous studies have shown that the number of neutral mutants is a more informative statistic for the mutational similarity distribution of RNA, the mutational similarity distributions that we observe for proteins are much better described simply by the mean.

Structural complexity of proteins

Following Dingle et al. (2022b), structural complexity of proteins is calculated in a similar way to RNA, with the 8-letter backbone alphabet translated into binary triplets: `H` → `101`, `B` → `010`, `E` → `111`, `G` → `011`, `I` → `110`, `T` → `001`, `S` → `100`, `A` → `000`. The resulting binary string is compressed using the Lempel-Ziv algorithm (as discussed above), with the resulting number of bits representing the complexity of the protein backbone structure.

References

Ahnert, S. E. (2017). Structural properties of genotype–phenotype maps. *Journal of The Royal Society Interface*, 14(132):20170275.

Begun, D. J., Lindfors, H. A., Kern, A. D., and Jones, C. D. (2007). Evidence for de novo evolution of testis-expressed genes in the drosophila yakuba/drosophila erecta clade. *Genetics*, 176(2):1131–1137.

Cardarelli, L., Pell, L. G., Neudecker, P., Pirani, N., Liu, A., Baker, L. A., Rubinstein, J. L., Maxwell, K. L., and Davidson, A. R. (2010). Phages have adapted the same protein fold to fulfill multiple functions in virion assembly. *Proceedings of the National Academy of Sciences*, 107(32):14384–14389.

Catalán, P., Wagner, A., Manrubia, S., and Cuesta, J. A. (2018). Adding levels of complexity enhances robustness and evolvability in a multilevel genotype–phenotype map. *Journal of The Royal Society Interface*, 15(138):20170516.

Ciliberti, S., Martin, O. C., and Wagner, A. (2007). Innovation and robustness in complex regulatory gene networks. *Proceedings of the National Academy of Sciences*, 104(34):13591–13596.

Dayhoff, M. O. (1972). A model of evolutionary change in proteins. *Atlas of protein sequence and structure*, 5:89–99.

De Lucrezia, D., Slanzi, D., Poli, I., Polticelli, F., and Minervini, G. (2012). Do natural proteins differ from random sequences polypeptides? natural vs. random proteins classification using an evolutionary neural network. *PLoS One*, 7(5):e36634.

Dingle, K., Camargo, C. Q., and Louis, A. A. (2018). Input–output maps are strongly biased towards simple outputs. *Nature communications*, 9(1):761.

Dingle, K., Ghaddar, F., Šulc, P., and Louis, A. A. (2022a). Phenotype bias determines how natural rna structures occupy the morphospace of all possible shapes. *Molecular biology and evolution*, 39(1):msab280.

Dingle, K., Novev, J. K., Ahnert, S. E., and Louis, A. A. (2022b). Predicting phenotype transition probabilities via conditional algorithmic probability approximations. *Journal of the Royal Society Interface*, 19(197):20220694.

Dingle, K., Schaper, S., and Louis, A. A. (2015). The structure of the genotype–phenotype map strongly constrains the evolution of non-coding rna. *Interface focus*, 5(6):20150053.

Dishman, A. F., Tyler, R. C., Fox, J. C., Kleist, A. B., Prehoda, K. E., Babu, M. M., Peterson, F. C., and Volkman, B. F. (2021). Evolution of fold switching in a metamorphic protein. *Science*, 371(6524):86–90.

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O., and Arnold, F. H. (2005a). Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences*, 102(40):14338–14343.

Drummond, D. A., Iverson, B. L., Georgiou, G., and Arnold, F. H. (2005b). Why high-error-rate random mutagenesis libraries are enriched in functional and improved proteins. *Journal of molecular biology*, 350(4):806–816.

Drummond, D. A., Raval, A., and Wilke, C. O. (2006). A single determinant dominates the rate of yeast protein evolution. *Molecular biology and evolution*, 23(2):327–337.

Edgell, C. L., Savery, N. J., and Woolfson, D. N. (2020). Robust de novo-designed homotetrameric coiled coils. *Biochemistry*, 59(10):1087–1092.

Ferrada, E. and Wagner, A. (2012). A comparison of genotype-phenotype maps for rna and proteins. *Biophysical journal*, 102(8):1916–1925.

Fontana, W., Konings, D. A., Stadler, P. F., and Schuster, P. (1993). Statistics of rna secondary structures. *Biopolymers: Original Research on Biomolecules*, 33(9):1389–1404.

Fram, B., Su, Y., Truebridge, I., Riesselman, A. J., Ingraham, J. B., Passera, A., Napier, E., Thadani, N. N., Lim, S., Roberts, K., et al. (2024). Simultaneous enhancement of multiple functional properties using evolution-informed protein design. *Nature communications*, 15(1):5141.

Freeland, S. J. and Hurst, L. D. (1998). The genetic code is one in a million. *Journal of molecular evolution*, 47:238–248.

Gabaldón, T. and Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, 14(5):360–366.

Ghaddar, F. and Dingle, K. (2023). Random and natural non-coding rna have similar structural motif patterns but differ in bulge, loop, and bond counts. *Life*, 13(3):708.

Greenbury, S. F., Schaper, S., Ahnert, S. E., and Louis, A. A. (2016). Genetic correlations greatly increase mutational robustness and can both reduce and enhance evolvability. *PLoS computational biology*, 12(3):e1004773.

Hogeweg, P. (2012). Toward a theory of multilevel evolution: long-term information integration shapes the mutational landscape and enhances evolvability. *Evolutionary systems biology*, pages 195–224.

Huynen, M. A. (1996). Exploring phenotype space through neutral evolution. *Journal of molecular evolution*, 43:165–169.

Huynen, M. A., Stadler, P. F., and Fontana, W. (1996). Smoothness within ruggedness: the role of neutrality in adaptation. *Proceedings of the National Academy of Sciences*, 93(1):397–401.

Johnston, I. G., Dingle, K., Greenbury, S. F., Camargo, C. Q., Doye, J. P., Ahnert, S. E., and Louis, A. A. (2022). Symmetry and simplicity spontaneously emerge from the algorithmic nature of evolution. *Proceedings of the National Academy of Sciences*, 119(11):e2113883119.

Jörg, T., Martin, O. C., and Wagner, A. (2008). Neutral network sizes of biological rna molecules can be computed and are not atypically small. *BMC bioinformatics*, 9:1–12.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637.

Lempel, A. and Ziv, J. (1976). On the complexity of finite sequences. *IEEE Transactions on information theory*, 22(1):75–81.

Lesk, A. (2010). *Introduction to protein science: architecture, function, and genomics*. Oxford University Press, USA.

Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A., and Begun, D. J. (2006). Novel genes derived from noncoding dna in drosophila melanogaster are frequently x-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences*, 103(26):9935–9939.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130.

Lipman, D. J., Souvorov, A., Koonin, E. V., Panchenko, A. R., and Tatusova, T. A. (2002). The relationship of protein conservation and sequence length. *BMC evolutionary biology*, 2:1–10.

Lipman, D. J. and Wilbur, W. J. (1991). Modelling neutral and selective evolution of protein folding. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 245(1312):7–11.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of genetics*, 92(1):155–161.

Manrubia, S., Cuesta, J. A., Aguirre, J., Ahnert, S. E., Altenberg, L., Cano, A. V., Catalán, P., Diaz-Uriarte, R., Elena, S. F., García-Martín, J. A., et al. (2021). From genotypes to organisms: State-of-the-art and perspectives of a cornerstone in evolutionary dynamics. *Physics of Life Reviews*, 38:55–106.

McLysaght, A. and Guerzoni, D. (2015). New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1678):20140332.

Medvedev, K. E., Kinch, L. N., Schaeffer, R. D., and Grishin, N. V. (2019). Functional analysis of rossmann-like domains reveals convergent evolution of topology and reaction pathways. *PLoS computational biology*, 15(12):e1007569.

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in neural information processing systems*, 34:29287–29303.

Mohanty, V., Greenbury, S. F., Sarkany, T., Narayanan, S., Dingle, K., Ahnert, S. E., and Louis, A. A. (2023). Maximum mutational robustness in genotype–phenotype maps follows a self-similar blancmange-like curve. *Journal of The Royal Society Interface*, 20(204):20230169.

Mordret, E., Dahan, O., Asraf, O., Rak, R., Yehonadav, A., Barnabas, G. D., Cox, J., Geiger, T., Lindner, A. B., and Pilpel, Y. (2019). Systematic detection of amino acid substitutions in proteomes reveals mechanistic basis of ribosome errors and selection for translation fidelity. *Molecular Cell*, 75(3):427–441.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Notin, P., Kollasch, A., Ritter, D., Van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., et al. (2023). Proteingym: Large-scale benchmarks for protein fitness prediction and design. *Advances in Neural Information Processing Systems*, 36:64331–64379.

Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilles and domain superfolds. *Nature*, 372(6507):631–634.

Pál, C., Papp, B., and Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics*, 158(2):927–931.

Ranea, J. A., Sillero, A., Thornton, J. M., and Orengo, C. A. (2006). Protein superfamily evolution and the last universal common ancestor (luca). *Journal of Molecular Evolution*, 63:513–525.

Romero, P. A. and Arnold, F. H. (2009). Exploring protein fitness landscapes by directed evolution. *Nature reviews Molecular cell biology*, 10(12):866–876.

Sahakyan, H., Babajanyan, S. G., Wolf, Y. I., and Koonin, E. V. (2025). In silico evolution of globular protein folds from random sequences. *Proceedings of the National Academy of Sciences*, 122(27):e2509015122.

Sanjuán, R., Cuevas, J. M., Furió, V., Holmes, E. C., and Moya, A. (2007). Selection for robustness in mutagenized rna viruses. *PLoS genetics*, 3(6):e93.

Schultes, E. A. and Bartel, D. P. (2000). One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science*, 289(5478):448–452.

Schuster, P., Fontana, W., Stadler, P. F., and Hofacker, I. L. (1994). From sequences to shapes and back: a case study in rna secondary structures. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1344):279–284.

Serohijos, A. W., Lee, S. R., and Shakhnovich, E. I. (2013). Highly abundant proteins favor more stable 3d structures in yeast. *Biophysical journal*, 104(3):L1–L3.

Serohijos, A. W., Rimas, Z., and Shakhnovich, E. I. (2012). Protein biophysics explains why highly abundant proteins evolve slowly. *Cell reports*, 2(2):249–256.

Smit, S., Yarus, M., and Knight, R. (2006). Natural selection is not required to explain universal compositional patterns in rrna secondary structure categories. *RNA*, 12(1):1–14.

Stich, M., Briones, C., and Manrubia, S. C. (2008). On the structural repertoire of pools of short, random rna sequences. *Journal of theoretical biology*, 252(4):750–763.

The RNAcentral Consortium (2021). RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic acids research*, 49(D1):D212–D220.

Touw, W. G., Baakman, C., Black, J., Te Beek, T. A., Krieger, E., Joosten, R. P., and Vriend, G. (2015). A series of pdb-related databanks for everyday needs. *Nucleic acids research*, 43(D1):D364–D368.

Tretyachenko, V., Vymětal, J., Bednárová, L., Kopeckỳ Jr, V., Hofbauerová, K., Jindrová, H., Hubálek, M., Souček, R., Konvalinka, J., Vondrášek, J., et al. (2017). Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Scientific Reports*, 7(1):15449.

Vakirlis, N., Acar, O., Hsu, B., Castilho Coelho, N., Van Oss, S. B., Wacholder, A., Medetgul-Ernar, K., Bowman, R. W., Hines, C. P., Iannotta, J., et al. (2020). De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nature communications*, 11(1):781.

Van Nimwegen, E., Crutchfield, J. P., and Huynen, M. (1999). Neutral evolution of mutational robustness. *Proceedings of the National Academy of Sciences*, 96(17):9716–9720.

Verkuil, R., Kabeli, O., Du, Y., Wicky, B. I., Milles, L. F., Dauparas, J., Baker, D., Ovchinnikov, S., Sercu, T., and Rives, A. (2022). Language models generalize beyond natural proteins. *BioRxiv*, pages 2022–12.

Von der Dunk, S. H., Martin, N. S., Dingle, K., and louis, a. (2025). Rna secondary structures are conserved but random. *bioRxiv*, pages 2025–08.

Wagner, A. (2005). Robustness, evolvability, and neutrality. *FEBS letters*, 579(8):1772–1778.

Wagner, A. (2008). Robustness and evolvability: a paradox resolved. *Proceedings of the Royal Society B: Biological Sciences*, 275(1630):91–100.

Wagner, A. (2013). Robustness and evolvability in living systems.

Wagner, A. (2023). Evolvability-enhancing mutations in the fitness landscapes of an rna and a protein. *Nature Communications*, 14(1):3624.

Wall, D. P., Hirsh, A. E., Fraser, H. B., Kumm, J., Giaever, G., Eisen, M. B., and Feldman, M. W. (2005). Functional genomic analysis of the rates of protein evolution. *Proceedings of the National Academy of Sciences*, 102(15):5483–5488.

Wilke, C. O. (2001). Selection for fitness versus selection for robustness in rna secondary structure folding. *Evolution*, 55(12):2412–2420.

Wright, E. S. (2025). Tandem repeats provide evidence for convergent evolution to similar protein structures. *Genome Biology and Evolution*, 17(2):evaf013.

Xu, J., Mcpartlon, M., and Li, J. (2021). Improved protein structure prediction by deep learning irrespective of co-evolution information. *Nature Machine Intelligence*, 3(7):601–609.

Yang, Z. (2007). Paml 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, 24(8):1586–1591.

Zhang, Y. and Skolnick, J. (2005). Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309.