The Statistical Trends of Protein Evolution: A Lesson from AlphaFold Database

Qian-Yuan Tang [1]*, Weitong Ren [2], Jun Wang [3], Kunihiko Kaneko [4, 5] †

[1] Lab for Neural Computation and Adaptation, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama, 351-0106, Japan.
[2] Theoretical Molecular Science Laboratory, RIKEN Cluster for Pioneering Research, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.
[3] School of Physics, National Laboratory of Solid State Microstructure, and Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing, 210093, China.
[4] Center for Complex Systems Biology, Universal Biology Institute, University of Tokyo, Komaba, Meguro, Tokyo 153-8902, Japan.
[5] The Niels Bohr Institute, University of Copenhagen, Copenhagen, Blegdamsvej 17, 2100-DK, Denmark.

Corresponding Authors:
* Qian-Yuan Tang, tangqianyuan@gmail.com
† Kunihiko Kaneko, kaneko@complex.c.u-tokyo.ac.jp

Abstract

The recent development of artificial intelligence provides us with new and powerful tools for studying the mysterious relationship between organism evolution and protein evolution. In this work, based on the AlphaFold Protein Structure Database (AlphaFold DB), we perform comparative analyses of the proteins of different organisms. The statistics of AlphaFold-predicted structures show that, for organisms with higher complexity, their constituent proteins will have larger radii of gyration, higher coil fractions, and slower vibrations, statistically. By conducting normal mode analysis and scaling analyses, we demonstrate that higher organismal complexity correlates with lower fractal dimensions in both the structure and dynamics of the constituent proteins, suggesting that higher functional specialization is associated with higher organismal complexity. We also uncover the topology and sequence bases of these correlations. As the organismal complexity increases, the residue contact networks of the constituent proteins will be more assortative, and these proteins will have a higher degree of hydrophilic-hydrophobic segregation in the sequences. Furthermore, by comparing the statistical structural proximity across the proteomes with the phylogenetic tree of homologous proteins, we show that, statistical structural proximity across the proteomes may indirectly reflect the phylogenetic proximity, indicating a statistical trend of protein evolution in parallel with organism evolution. This study provides new insights into how the diversity in the functionality of proteins increases and how the dimensionality of the manifold of protein dynamics reduces during evolution, contributing to the understanding of the origin and evolution of lives.

Introduction

The evolution of organisms takes place over time, as Darwin noted in his *On the Origin of Species*. It is usually naively discussed that the complexity of life has increased throughout evolution, whereas its validity is not always confirmed (McShea 1996; Adami et al. 2000; Furusawa and Kaneko 2000). Generally, it is widely accepted that complexity should be characterized as the difficulty of reducing a system to constitutional components. If the system consists of more different internal components, the complexity is generally higher. According to this idea, one can simply assume that eukaryote cells are more complex than prokaryotes, and multicellular organisms with more distinct cell types are more complex than unicellular organisms, as is also adopted in *The Major Transitions in Evolution* (Maynard Smith and Szathmary 1997). Parallel to the increasing complexity of organisms, proteins, as the basic building blocks of organisms, are also undergoing continuous evolution. Previously, there were evolutionary studies elaborating on the phylogenetic analysis of proteomes (Gerstein et al. 1994; Caetano-Anollés et al. 2009; Caetano-Anollés et al. 2021) or the proteins that share the common ancestral sequence or structure (Labas et al. 2002; Pin et al. 2003; Zardoya 2005; Morcos et al. 2011; Finnigan et al. 2012; Espada et al. 2015). To uncover the connection between organism evolution and protein evolution (Liu and Rost 2001; Koonin et al. 2002; Pál et al. 2006; Zeldovich and Shakhnovich 2008; Sikosek and Chan 2014), combining the view of evolution from two perspectives may be necessary. In the microscopic view (molecule level), the evolution of a protein does not necessarily follow the evolutionary path as the species evolve (Choi and Kim 2006); while in the macroscopic view (organism level), if considering an ensemble of thousands of proteins within the same organism, just as thermodynamic laws and collective order can emerge from the random motions of molecules, it may be possible to discover a "collective" trend consistent with the increase of organismal complexity.

Notably, the recent development of artificial intelligence (AI) provides us with new and powerful tools to help us elucidate the trends in protein evolution on the macroscopic scale. AlphaFold, an AI system developed by DeepMind, which makes full use of the coevolutionary information to predict protein structure, has already won an unprecedented and overwhelming success in protein structure predictions (Senior et al. 2020; Jumper et al. 2021). Despite the limitations noted in previous research (Bagdonas et al. 2021; Pak et al. 2021; Ruff and Pappu 2021), AlphaFold 2 is acknowledged to provide high-accuracy predictions of protein structures, even for sequences with relatively fewer homologous sequences (Jumper et al. 2021). The exceeding accuracy and the speed of AlphaFold 2 provide the possibility of generating an extensive database of structure predictions. AlphaFold Protein Structure Database (AlphaFold DB) provides now provides free access to about 200 million predicted protein structures, which covers the complete proteome of various organisms ranging from bacteria, archaea, unicellular and multicellular eukaryotes to humans (Varadi et al. 2022), and it keeps expanding. AlphaFold DB not only offers the potential to answer critical questions in medical and biological sciences (Robertson et al. 2021; Bayly-Jones and Whisstock 2022), but also shows new possibilities in the study of protein evolution. Instead of focusing only on specific families of proteins (Bayly-Jones and Whisstock 2022), we can now perform

comparative structural analyses for the proteins in different organisms. Combining with other essential evolutionary analyses, the statistics of the full-proteome protein structures may help us uncover the hidden connection between protein evolution and organism evolution.

In this work, based on the AlphaFold-predicted structures of the proteomes of 48 organisms, we perform comparative analyses of the constituent proteins of different organisms. The statistical results indicate a correlation between the flexibility of constituent proteins and the complexity of organisms; that is to say, for organisms with higher complexity, their constituent proteins will have larger radii of gyration, higher coil fractions, and slower vibrations, statistically. By conducting normal mode analysis and scaling analyses, we demonstrate that higher organismal complexity correlates with lower fractal dimensions in both the structure and dynamics of the constituent proteins, suggesting that higher functional specialization is associated with higher organismal complexity. We also uncover the topology and sequence bases of these correlations. As the organismal complexity increases, the residue contact networks of the constituent proteins will be more assortative, and these proteins will have a higher degree of hydrophilic-hydrophobic segregation in the sequences. Furthermore, by comparing the statistical structural proximity across the proteomes with the phylogenetic tree of homologous proteins, we show that, statistical structural proximity across the proteomes may indirectly reflect the phylogenetic proximity of homologous proteins among organisms. Such a result suggests a statistical trend of protein evolution in parallel with organism evolution. This study provides new insights into how the diversity of protein functionality increases and how the dimensionality of the protein dynamics manifold reduces in the evolution, contributing to the understanding of the origin and evolution of lives.

Results

The statistics of AlphaFold-predicted structures indicate a correlation between the flexibility of constituent proteins and the complexity of organisms.

We perform a comparative analysis of the predicted protein structures of the 48 organisms from the AlphaFold DB. The details of the full dataset are listed in Supplementary Information (SI) Table S1. Here, let us first focus on the proteins from the 16 model organisms with similar chain lengths. Namely, for the proteomes of various organisms, we always select the protein structures with chain lengths $N \approx 250$, calculate their structural characteristics, and conduct statistical analyses. As shown in Fig. 1A, despite having similar chain lengths, the constituent proteins in different organisms have different distributions of radii of gyration R_g. We perform two-sample Kolmogorov–Smirnov (KS) test to compare the R_g distributions of the proteins in different organisms (details listed in SI Methods). It is observed that some organisms have very similar R_g distributions, and there are no statistically significant differences between them (e.g., *M. jannaschii* vs. *E. coli*, and mouse vs. rat). However, for some distinct organisms, for example, prokaryotes vs. eukaryotes (e.g., *E. coli* vs. yeast), unicellular vs. multicellular organisms (e.g.,

yeast vs. mouse), or multicellular organisms with significantly different numbers of cell types (e.g., *C. elegans* vs. human), there are statistically significant differences between them. More interestingly, if we sort the organisms according to the median R_g, it is observed that larger median R_g correlate with increasing organismal complexity. To evaluate such a correlation, for a given organism proteome, we introduce the total number of proteins and the total chain length of the proteins as measures of organismal complexity. As shown in Fig. 1B, both complexity measures are proportional to the median R_g for proteins with similar chain lengths, demonstrating the correlation between protein flexibility and organismal complexity. Note that such a correlation is robust. For example, if one considers the proteins with other given chain lengths or selects other structural descriptors such as the solvent-accessible surface areas (SASA) to quantify the flexibility of the proteins, similar correlations can also be observed (SI Fig. S2).

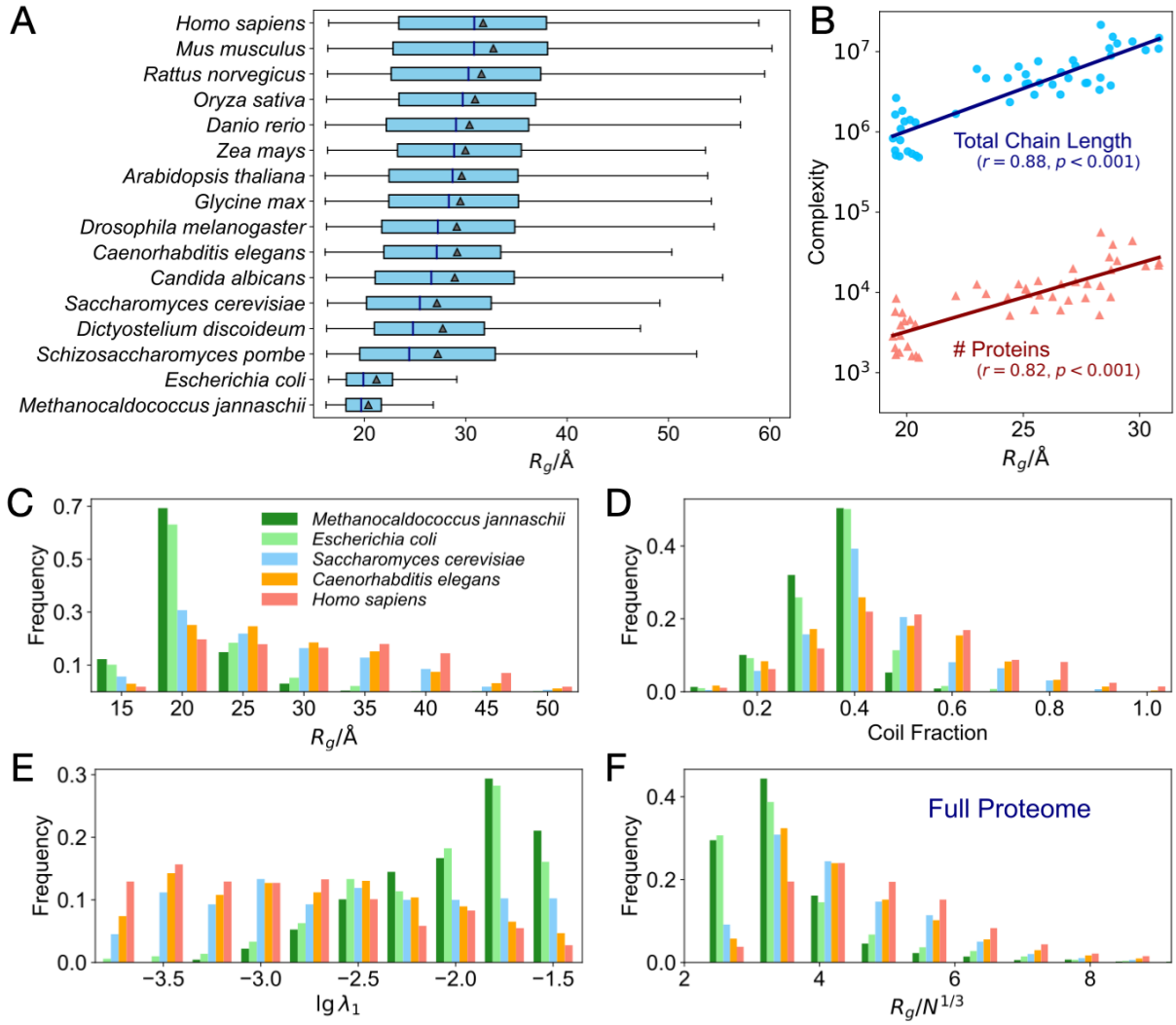

Figure 1. The statistics of AlphaFold-predicted structures indicate the correlation between the flexibility of constituent proteins and the complexity of organisms. (A) For the proteins from 16 model organisms with similar chain lengths $N \approx 250$ ($225 \leq N < 275$), the distributions of the radii of gyration R_g are

shown as the box-and-whiskers (extreme values not shown). Here, the triangle and the vertical bar in the box denote the mean value and the median of the R_g, respectively. (B) For all the 48 organisms in AlphaFold DB, the measures of organismal complexity (the total number of proteins and the total chain length of the proteins in the organism proteome) vs. median R_g of the proteins with chain lengths $N \approx 250$. For proteins from the five selected organisms with chain length $N \approx 250$, the histogram of the (C) radii of gyration R_g, (D) coil fraction, and (E) the slowest mode eigenvalue λ_1 at the logarithmic scale. (F) The distribution of the normalized R_g (i.e., $R_g/N^{1/3}$) for all the proteins in the proteome of the five selected organisms.

To demonstrate such a correlation clearly, we select five species with different organismal complexity for detailed examinations. The five selected species include archaea (*M. jannaschii*), bacteria (*E. coli*), unicellular eukaryotes (budding yeast, *S. cerevisiae*), multicellular invertebrates (*C. elegans*), and human (*Homo sapiens*). Based on the proteomes of these organisms, we select the proteins with chain lengths $N \approx 250$, conduct structural analysis, and perform comparative analysis. The histograms shown in Figs. 1C-E demonstrate how the shape, secondary structures, and equilibrium dynamics of the proteins vary in the selected organisms.

First, as shown in Fig. 1C, for the proteins with similar chain lengths, the mean value, as well as the standard deviation of the radius of gyration R_g gradually increases for organisms with higher complexity. Then, we apply the DSSP algorithm (Kabsch and Sander 1983; Joosten et al. 2011) to assign secondary structures (helices, sheets, or coils) to the proteins and investigate how the structural flexibility of the constituent proteins varies in different organisms. We find that the average fraction of coils increases as the organismal complexity increases (Fig. 1D). Next, we examine the dynamical flexibility of the proteins according to their vibrations around the native structure. These vibrations are closely related to the functional dynamics of the proteins, and can be predicted by the elastic network model (ENM) (Haliloglu et al. 1997; Bahar et al. 1998; Bahar et al. 2010). Based on ENM, one can conduct the normal mode analysis and obtain the eigenvalues corresponding to the vibration modes (Case 1994; Hayward et al. 1995; Wako and Endo 2017). Among these eigenvalues, the smallest nonzero eigenvalue λ_1, which corresponds to the square of the slowest mode frequency ($\lambda_1 \sim \omega_1^2$) and is proportional to the inverse square of the amplitude of the motion ($\lambda_1 \sim A_1^{-2}$), can be recognized as a measure of dynamical flexibility. As shown in Fig. 1E, statistically, for organisms with higher complexity, their constituent proteins will exhibit slower vibrations (i.e., smaller λ_1) around the equilibrium. Usually, the slow vibrations are closely related to the high modularity of the residue contact networks. We also show that for the proteins with similar chain lengths, the mean value and the standard deviation of modularity increase as the organismal complexity increases (see SI Figs. S3 and S4). In short, the above results clearly demonstrate the correlation between the flexibility of constituent proteins and the complexity of organisms. Further analyses show that even if we remove the proteins with low prediction confidence from the dataset, a similar correlation can still be observed (see Discussions and SI Fig. S5).

Furthermore, similar structural statistics can be extended to the analysis of the full proteomes. To compare the proteins with different chain lengths, we normalize the radii of gyration as $R_g/N^{1/3}$. When a protein is highly ordered and densely packed into a globular shape, then the normalized R_g will be small; in contrast, when a protein is highly disordered or deviates from a globular shape, then the normalized R_g will be large. The statistics of the normalized R_g are shown in Fig. 1F. For the five selected organisms, the statistics of the full proteome show a similar trend as the statistics of the proteins with comparable chain lengths. For organisms with higher complexity, there will be larger mean values and standard deviations of the normalized R_g. In SI Fig. S2, we conduct similar statistics for the full proteomes of all the 48 organisms in AlphaFold DB, compare the differences between the normalized R_g distribution of different organisms, and show the correlations between the normalized R_g and the complexity measures. These results further suggest that the increasing organismal complexity is accompanied by higher flexibility of the proteins in the full proteome. Further analyses (see SI Fig. S4) also show that structural diversity (measured by the standard deviation of the normalized R_g) of constituent proteins also correlates with organismal complexity.

The scaling analyses suggest a decreasing fractal dimension of constituent proteins as the organismal complexity increases.

Based on the Protein Data Bank (PDB) (Berman et al. 2000), recent scaling analyses (Tang et al. 2017; Tang and Kaneko 2020) revealed the power laws related to the size dependence of the protein structure and dynamics. Due to the limited number of experimentally determined protein structures, it had been unable to compare the scaling coefficients for proteomes of different organisms separately. AlphaFold DB, however, compensates for the lack of experimental data, enables us to perform accurate scaling analyses of proteins across different organisms, compare scaling coefficients, and reveal the correlation between the organism's complexity and the structural dynamics of their constituent proteins.

In this work, based on the predicted structures from AlphaFold DB, we apply the scaling analyses to the proteins from different species. For proteins within an organism, we first divide the proteins into bins according to their chain lengths. Then, for the proteins in the same bin (i.e., with similar chain length *N*), we calculate the mean value, as well as standard error, of the length of the shortest principal semi-axis L_C and the slowest mode eigenvalue λ_1. In this way, we obtain the size dependence of proteins' shape and dynamics. Figure 2A and B show the results of the scaling analysis for the five selected organisms.

Let us first analyze the scaling relations between protein shape and chain length *N* for different species. As the protein shape is described by the length of the shortest principal semi-axis L_C, there will be a scaling relation: $L_C \sim N^{1/d}$, where d is the average fractal dimension of the proteins. For example, when a protein is densely packed into a globular shape in three-dimensional space, there will be $L_C \sim R_g \sim N^{1/3}$ (i.e.,

$d = 3$). The scaling analyses shown in Fig. 2A indicate that the average fractal dimensions d vary across different organisms. Besides, for all the organisms, it is observed that $d \leq 3$, displaying that the proteins are not always folded into densely packed globules. Previous studies on the fractal structure of the proteins (Lewis and Rees 1985; Liang and Dill 2001; Reuveni et al. 2008), as well as the statistics of the normalized R_g (see Fig. 1F), can support such a result.

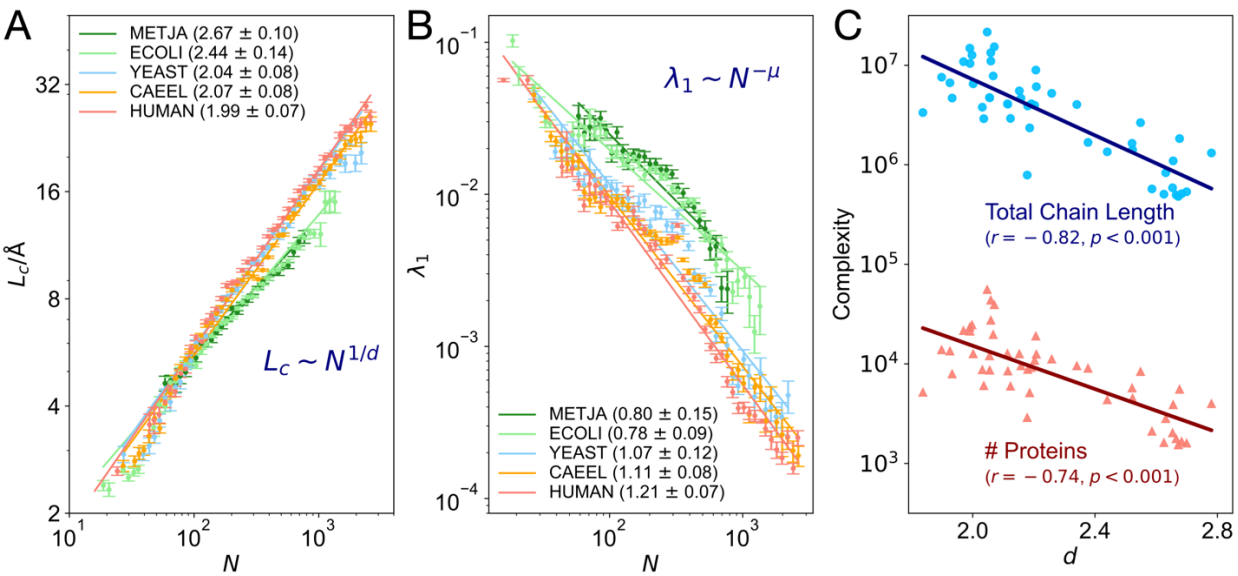

Figure 2. The scaling analyses suggest a decreasing fractal dimension of constituent proteins as the organismal complexity increases. For the proteins from five selected organisms, the size (chain length N) dependence of (A) length of the shortest semi-axis L_c and (B) the slowest mode eigenvalue λ_1. Here, for the data in all the bins, the average standard error of L_c is below 3.5% of the mean value, and the average standard error of λ_1 is below 13.1% of the mean value. The fitted scaling coefficients d (fractal dimension) and μ are listed in the legends of the subplots (A) and (B), respectively. These scaling coefficients are obtained by robust fittings with Theil–Sen estimators (95% confidence interval). (C) For all the 48 organisms in AlphaFold DB, the measures of complexity (the total number of proteins and the total chain length of the proteins in the organism proteome) vs. the estimated fractal dimension d.

Then, we investigate the size dependence of the equilibrium dynamics of the native proteins. Previous research based on the PDB had shown that as the protein chain length N increases, the vibrations of the protein would become slower (Tang and Kaneko 2020), i.e., the slowest mode eigenvalue λ_1 versus chain length N obeys the scaling relation: $\lambda_1 \sim N^{-\mu}$, where $\mu \approx 1$. Notably, based on the AlphaFold DB, as shown in Fig. 2B, detailed analyses reveal that the scaling coefficients μ vary across different organisms. Next, let us take a closer look at the scaling coefficients d in Fig. 2A, one can find that, for two organisms with significant complexity differences (e.g., *E. coli* vs. human), there are significant differences in the corresponding fractal dimension d. The average fractal dimension of the constituent proteins of human is

significantly lower than that of *E. coli*. According to such a result, one may conjecture that the average fractal dimension of constituent proteins is negatively correlated with the organismal complexity. To validate such a conjecture, we estimate the average fractal dimension d for all the 48 species in AlphaFold DB and evaluate their correlation with the complexity measures (total kinds of proteins and total chain lengths of proteins) of the organisms. The fittings in Fig. 2C confirm such a conjecture. The negative Pearson correlation coefficients indicate that for the species with higher organismal complexity, their constituent proteins will be lower average fractal dimension d. Such a correlation can also be validated by the statistics of the fractal dimension (i.e., the average packing dimension calculated with the box-counting method) for all the proteins (see SI Fig. S1C). Besides, for a given proteome, the average fractal dimension of the proteins can also be estimated by the size dependence of the modularity Q of the proteins' residue contact networks (Guimerà et al. 2004; Tang and Kaneko 2020). Although the average fractal dimension $\tilde{d}$ (estimated from N vs. Q) may be different from d (estimated from N vs. L_C), it is also observed that a decreasing fractal dimension $\tilde{d}$ correlates with the increase of organismal complexity (see SI Fig. S3B). Similarly, we evaluate the correlation between the scaling coefficient μ and organismal complexity. As shown in SI Fig. S3D, the scaling coefficient μ is correlated with the measures of organismal complexity, suggesting that the constituent proteins of organisms with higher complexity will have slower vibration frequencies.

Higher organismal complexity correlates with higher functional specialization and lower dimensionality in protein dynamics.

According to the "structure-dynamics-function" paradigm (Friedman 1985; Agarwal 2006; Haliloglu and Bahar 2015), the dynamics encoded by the three-dimensional structure determine the functionality of the protein. As reported by previous research, for proteins with ordered native structures, the dynamics are suggested to be constrained to low-dimensional manifolds and could be described by a few 'slow modes' which correspond to low vibration frequency, low excitation energies, and large amplitudes motions. These slow modes are robustly encoded in the native structure of proteins. One may conjecture that the slow mode distributions of the constituent proteins are also correlated with the organismal complexity. To validate such a conjecture, for every protein, we construct the corresponding elastic network and conduct the normal mode analysis. In this way, the vibration spectrum can be obtained (see Method). In the spectrum, the eigengap between the leading two eigenvalues (λ_1 and λ_2) reflects how much the equilibrium dynamics of the protein will be dominated by the slowest normal mode (Togashi et al. 2007). For example, if $\lambda_2/\lambda_1 = 10$, then the ratio of the square amplitudes of the two modes will be: $A_1^2/A_2^2 = \lambda_2/\lambda_1 = 10$, that is, the motion along the direction of the first mode is an order of magnitude greater than the motion along the direction of the second mode, reflecting a high functional specialization of the protein, or say, low dimensionality in protein dynamics. In contrast, if $\lambda_2/\lambda_1 \approx 1.01$, the protein has almost the same tendency to move along the directions associated with the first and second slowest modes, indicating a higher dimensionality in dynamics and a lower functional specialization of the proteins. In

short, the eigengaps between the leading eigenvalues and dimensionality in protein dynamics can act as measures of the functional specialization of the proteins.

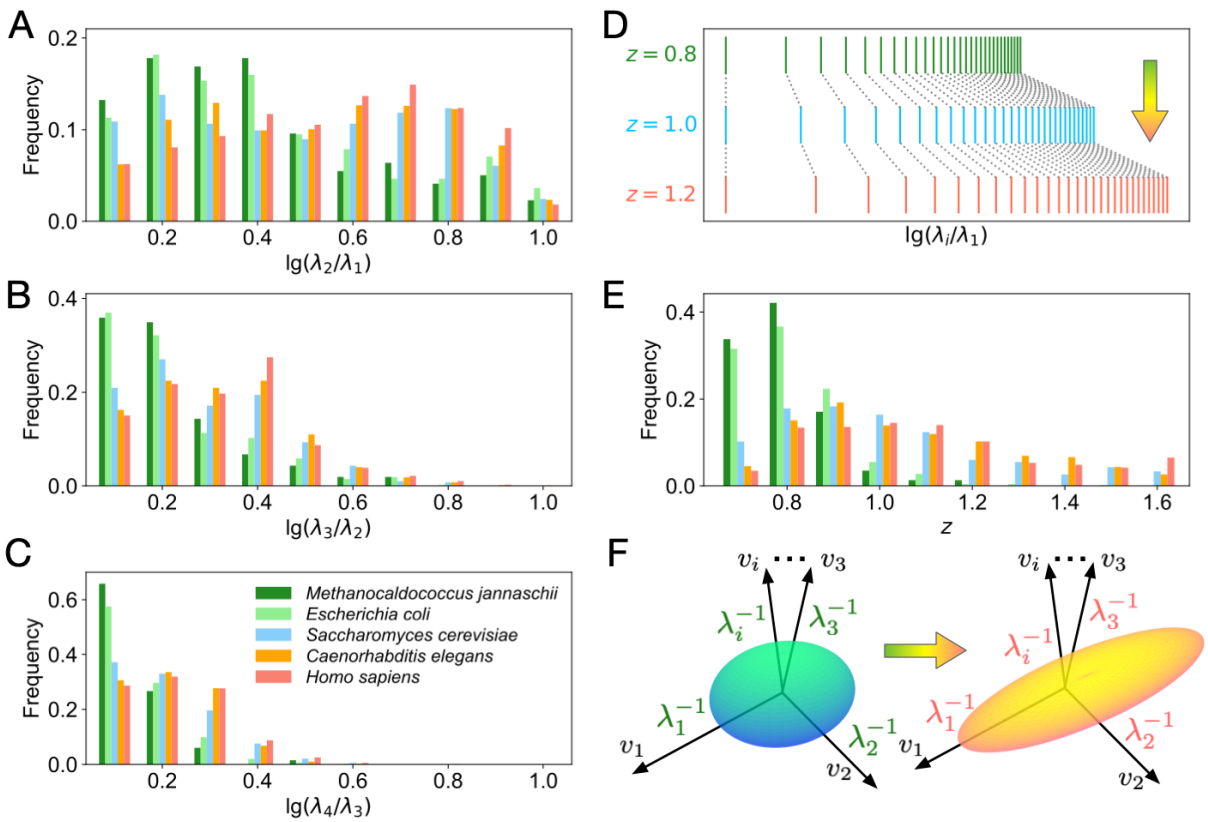

Figure 3. The increasing eigengaps in the vibration spectra imply that higher organismal complexity correlates with higher functional specialization and lower dimensionality in protein dynamics. For proteins with similar chain lengths N ($225 \leq N < 275$) in the five selected organisms, the distribution of the eigengaps (A) λ_2/λ_1, (B), λ_3/λ_2 and (C) λ_4/λ_3 at the logarithmic scale. (D) Illustration of power-law spectra at the logarithmic scale with Zipf's coefficients $z = 0.8$, 1.0, and 1.2. (E) For proteins with similar chain lengths in different species, the distribution of the Zipf's coefficients z. (F) Illustration of the trend of dimensional reduction in the conformational spaces. The arrows in subplots (D) and (F) indicate the increase in organismal complexity.

Similar to previous sections, for the proteins with similar chain lengths $N \approx 250$ from the five selected organisms, we conduct normal mode analysis based on the elastic network model and calculate the corresponding eigenvalues. Then, the distributions of the eigengaps λ_2/λ_1, λ_3/λ_2, and λ_4/λ_3 (at logarithmic scale) are obtained. As shown in Figs. 3A-C, all the average ratios between neighboring eigenvalues increase as the organismal complexity increases, clearly demonstrating the increasing scale separation in the vibrational frequencies of the constituent proteins. This result indicates that, for organisms with higher complexity, their constituent proteins will have higher functional specialization and lower dimensionality. Moreover, it is worth noting that similar correlations can happen on every scale.

As the organismal complexity increases, the constituent proteins have a statistical trend that the motions along the direction of the first modes have evolved to be much greater than the second, the motions along the second modes have evolved to be much greater than the third, etc.

To quantify such a scale-free property, it is necessary to introduce the power laws to describe the vibrations of the proteins. Previous studies have shown that the vibrational spectrum of proteins obeys a power-law distribution (Reuveni et al. 2008), and the rank-size distribution of the inverse of eigenvalues follows a Zipf-like distribution (Tang et al. 2020; Tang and Kaneko 2021; Xie et al. 2022). In SI Fig. S6, several proteins are provided as examples. These proteins have similar chain lengths, but their vibration spectra are different. The rank-size distribution of the slow-mode eigenvalues shows a power-law-like behavior, which can be described as the Zipf's law, i.e., $\lambda_i^{-1} \sim i^{-z}$, where z is known as the Zipf's coefficient.

Generally, for the proteins with similar sizes, a larger z corresponds with larger eigengaps (Fig. 3D). In Fig. 3E, the statistics of Zipf's coefficients z show that, for organisms with higher complexity, z will be larger, i.e., there will be larger eigengaps in the vibrational spectra of constituent proteins. Such a correlation is illustrated in Fig. 3F. In the illustration, the conformational space (the space encompassing all possible structures results from thermal fluctuations) of a protein around its native structure is represented as an ellipsoid. The axes of the ellipsoids denote the normal modes, and their semi-length represents the square magnitudes of the motions in the directions of the normal modes. The functional specialization and low dimensionality correspond to the anisotropy of the conformational space. As the organismal complexity increases, both the structure and dynamics of the constituent proteins show a statistical trend towards dimensional reduction.

Organismal complexity correlates with the assortativity of the residue contact networks and the hydrophilic-hydrophobic segregation in protein sequences.

In previous studies, residue contact networks of native proteins have been shown to play a dominating role in determining protein dynamics and functions (Bahar et al. 2010; Atilgan et al. 2012). The residue contacts are, in turn, largely determined by the protein sequences. Therefore, we may use the residue contact network as a bridge to investigate how organismal complexity correlates with the sequences of the constituent proteins.

In the previous subsection, we show that higher organismal complexity correlates with larger scale separation in the vibrational frequencies of the proteins. In fact, such a correlation in frequency space can correspond to the changes in the distribution of the residues' local packing density in the real space. According to the local density model (Halle 2002), there is a nearly linear relationship between the residues' square of vibration frequency and the local packing density. As a result, the increasing eigengaps in vibrational frequencies corresponds to the increasing variances of the residues local packing density. In protein molecules, the residue packing density is mainly determined by the hydrophobic effects. As

illustrated in Fig. 4A, by avoiding exposure to water, hydrophobic residues are buried in the interior of the protein, forming the hydrophobic cores with high local packing density. In contrast, the hydrophilic (polar) residues are likely to exposure to the solvent and have low local packing density. It is such hydrophobic effects that serve as the major driving force of protein folding (Dill 1990; Dill and MacCallum 2012).

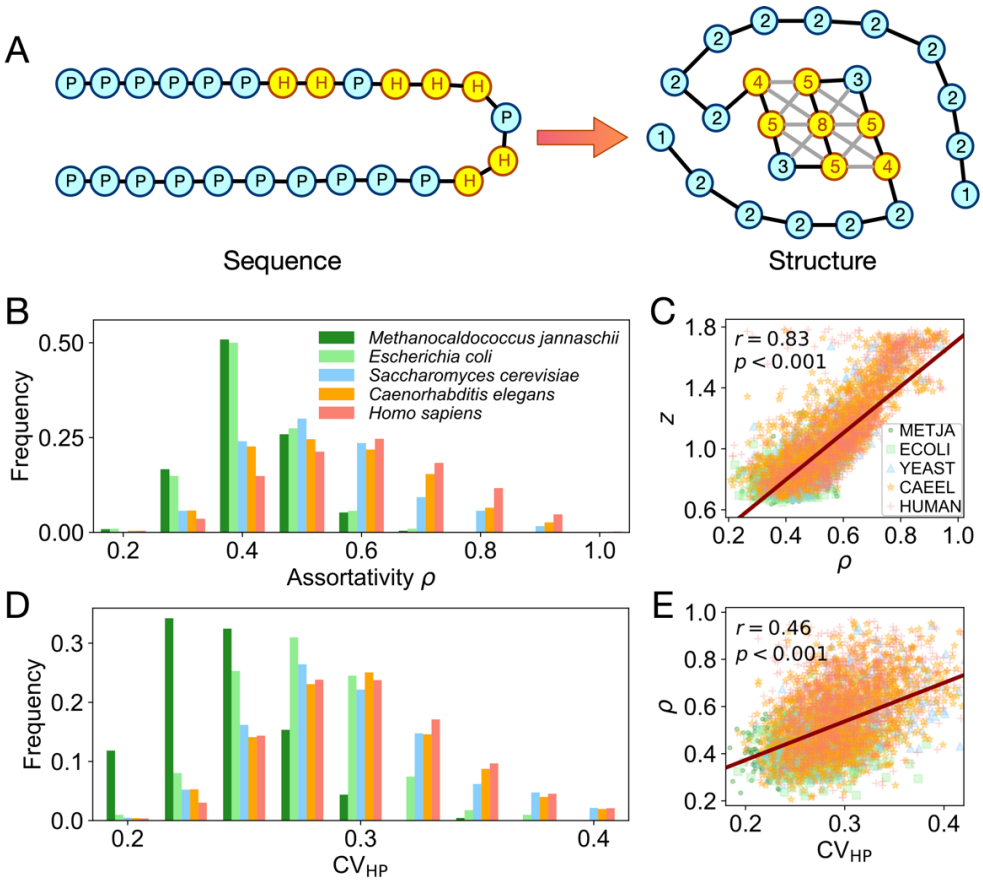

Figure 4. Organismal complexity correlates with the assortativity of the residue contact networks and the hydrophilic-hydrophobic segregation in protein sequences. (A) Illustration of a protein sequence (left) folds into a native structure described as an assortative residue contact network (right). The yellow and cyan nodes denote the hydrophobic (H) and polar (P) amino acid residues, respectively. When the protein folds into the native structure, the hydrophobic residues tend to aggregate into a densely connected hydrophobic core. In the subplot (right), the numbers represent the node's degree (number of connections with other nodes). (B) For proteins with similar chain length N ($225 \leq N < 275$) in the five selected organisms, the histogram of residue contact network assortativity ρ. (C) The scattering plot and the fitted trend line of the assortativity ρ vs. Zipf's coefficient z. (D) For proteins with similar chain lengths N in different organisms, the histogram of the hydropathy variations CV_{HP}. (E) The scattering plot and the fitted trend line of hydropathy variation CV_{HP} vs. residue contact network assortativity ρ.

For the proteins with similar chain lengths, the variance of the local packing density is closely related to the assortativity ρ of the residue contact network. The assortativity of a network is defined as the Pearson correlation coefficient of degrees between pairs of linked nodes (Newman 2002; Newman 2003). According to the definition, if the residues (nodes) with high packing density (degrees) are more likely to have contact with each other, then there will be a higher assortativity ρ and a larger variance in the packing density distribution of residues. Remarkably, as shown in Fig. 4B, as the organismal complexity increases, the constituent proteins with similar chain lengths will have a larger mean value of assortativity ρ. Notably, such a correlation is in line with the trend related to the increasing eigengap of vibration spectra. As shown in Fig. 4C, for the proteins in our dataset, regardless of the species they belong to, the assortativity ρ is proportional to Zipf's coefficient z. This result indicates that we have found the topological descriptor associated with the functional specialization of the protein.

Next, let us discuss how organismal complexity correlates with the sequences of the constituent proteins. As illustrated in Fig. 4A, the spatial segregation of hydrophilic and hydrophobic residues is closely related to their segregation in sequence. When hydrophilic and hydrophobic residues are uniformly mixed and randomly distributed in the sequence, they will be less likely to have profound spatial segregation. Conversely, when there are longer hydrophilic or hydrophobic fragments in a sequence, the hydrophobic cores can be formed more easily. Here, we introduce the hydropathy variation CV_{HP} to quantify the hydrophilic-hydrophobic segregation in protein sequences, where CV_{HP} is defined as the coefficient of variation (i.e., standard deviation divided by the mean value) of the filtered hydropathy profile of a sequence (see Method). As shown in Fig. 4D, for proteins with similar chain lengths, as the complexity of the organism increase, there will be a larger average CV_{HP}, which corresponds to a more significant hydrophilic-hydrophobic segregation. Further evaluation of such a correlation is shown in SI Fig. S8. As shown in Fig. 4E, for proteins in our dataset, regardless of the organisms they belong to, the hydropathy variation CV_{HP} is proportional to the assortativity of the residue contact network, implying that the spatial segregations of hydrophilic and hydrophobic residues do have statistically significant correlation with the sequence segregations.

The sequence analysis result also uncovers the differences between the constituent proteins from archaea (*M. jannaschii*) and bacteria (*E. coli*). Although their structures (e.g., radii of gyration) do not show significant differences (see Fig. 1C), their sequences do show significant differences (Fig. 4D). These results do not mean that *E. coli* has a significantly higher complexity than *M. jannaschii*. It is their living environments that determine the major difference in sequence. As a thermophilic methanogenic archaean, *M. jannaschii* lives in extreme environments such as hypothermal vents at the bottom of the oceans. The optimal growth temperature of *M. jannaschii* is about 85°C (Jones et al. 1983; Bult et al. 1996), which is even higher than the melting temperature (~50°C) of most proteins in *E. coli* (Ghosh and Dill 2010). The low hydropathy segregation in the protein sequences of *M. jannaschii* contributes to the stabilization of the surface residues of the proteins, thus leading to the organism's adaptation to a hot environment.

Phylogenetic proximity of homologous proteins correlates with the statistical structural proximity of the proteomes.

In previous sections, our statistical structural analysis of proteomes has shown that the organismal complexity can be reflected by the structural properties of the constituent proteins. Since the complexity of organisms has emerged through evolution, one may conjecture that the phylogenetic proximity can be reflected in the structural statistics of the constituent proteins. To validate such a conjecture, here we define the statistical structural proximity between two proteomes based on the Kolmogorov-Smirnov (KS) statistic D (Massey 1951) between the distributions of the radii of gyration R_g for proteins with similar chain length $N \approx 250$. Intuitively, the KS statistic D can be understood as the largest absolute difference between two cumulative distribution functions. When two proteomes have very similar R_g distributions, then their distance (KS statistic D) will be small, and the statistical structural proximity between the two proteomes will be high. By calculating the KS statistic between every two organisms, a statistical structural distance matrix can be constructed. Interestingly, the phylogenetic tree (Fig. 5A) generated with the sequences of a homological gene shows correspondence with the statistical structural distance matrix. The lineages that are close in the phylogenetic tree also have a small distance (i.e., small KS statistic D) between the proteomes. Such a result not only demonstrates that phylogenetic proximity correlates with the statistical structural proximity of the proteomes, but also suggests that there is a statistical trend of protein evolution in parallel with organism evolution.

Notably, the matrix shown in Fig. 5B is based on a simplistic definition of statistical structural proximity between proteomes, say, only R_g for proteins with similar chain lengths are considered. One may take into account proteins with other chain lengths or introduce other structural descriptors (e.g., SASA, secondary structures, modularity etc.) to refine the definition. Further analysis (see SI Fig. S9) shows that, even though statistical structural proximity based on other descriptors may vary in values, we can generally observe that the lineages that are close in the phylogenetic tree show high statistical structural proximity. In all, the statistical structural proximity across the proteomes may indirectly reflect the evolutionary relationships among species.

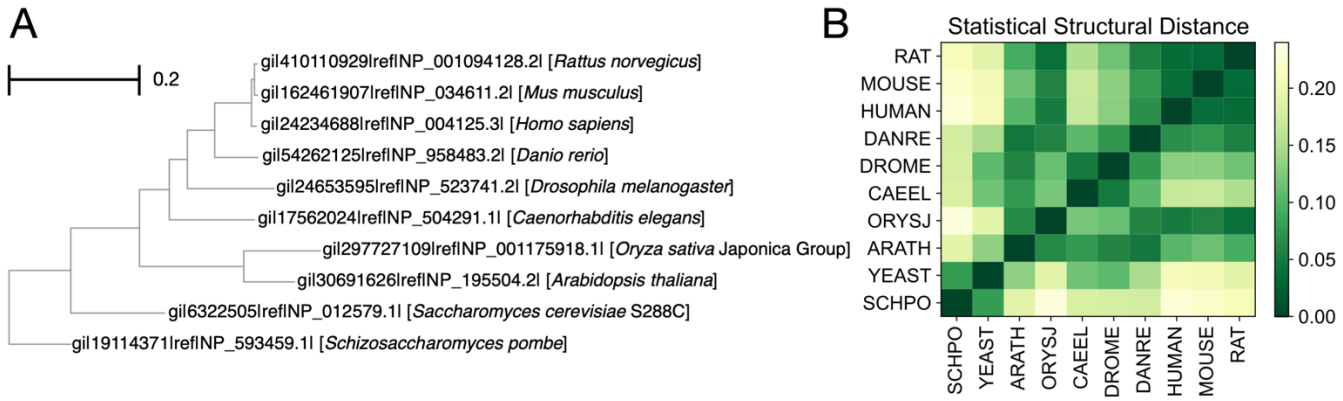

Figure 5. Phylogenetic proximity of homologous proteins correlates with the statistical structural proximity of the proteomes. (A) A phylogenetic tree generated with the sequences of the heat shock protein (HSPA9, HomoloGene: 39452) by NCBI protein BLAST (Gish and States 1993; McGinnis and Madden 2004). (B) The statistical structural distance matrix with entries of the KS statistic D between R_g distributions of the organisms' constituent proteins with similar chain length $N \approx 250$.

Discussion

In this work, based on the protein structures predicted by AlphaFold, we propose a statistical framework for proteome analysis by comparing the protein structures in different organisms. Rather than studying the evolution of specific protein families or superfamilies, this study aims to reveal the correlations between organismal complexity and structural or dynamical properties of the constituent proteins. It is observed that, statistically, the constituent proteins of higher-complexity organisms will have higher flexibility, lower fractal dimensions in both structure and dynamics, and higher degrees of hydrophobicity segregation in both their structures and sequences. Note that these correlations do not depend on the definition of complexity. Although the mathematical definition of complexity is still controversial (Lloyd 2001), different measures of organismal complexity (e.g., number of cell types, genome size, proteome size, etc.) are correlated (Markov et al. 2010; Niklas et al. 2014). Besides, it is worth noting that the structure prediction confidence of AlphaFold will not affect our main conclusions. For the AlphaFold-predicted structures, lower prediction confidence (which can be quantified as lower pLDDT values, see SI) usually arrives for proteins with long disordered regions (Jumper et al. 2021). Interestingly, even if we remove those proteins with lower prediction confidence (i.e., with long intrinsic disorder regions) from the dataset, similar correlations can still be observed (see Fig. S2). Moreover, our study focuses primarily on the global dynamics (e.g., slowest modes) that are robust to the local variations in native structures (Bahar et al. 2010; Tang et al. 2020). Such insensitivity to prediction confidence can further strengthen our conclusions.

Our topological and sequence analyses show that higher organismal complexity correlates with the assortativity of the residue contact networks and the hydrophilic-hydrophobic segregation in protein sequences. These correlations are consistent with what has been suggested by previous evolutionary studies (Phillips 2009; Phillips 2012; Hemery and Rivoire 2015; Foy et al. 2019; Moret et al. 2019; Phillips 2020), and may guide the design or modification of proteins. For example, one may enhance the flexibility or functional sensitivity of a protein by greater segregation of hydrophobic and hydrophilic amino acids in the sequence. Besides, hydrophilic-hydrophobic segregation can also prevent the dysfunctional aggregation of proteins (Foy et al. 2019). Notably, the sequence analysis does not rely on AlphaFold predictions in any way, which acts as essential support that our main conclusions do not arise from a systematic bias inherent in structural prediction methods, but reflect the natural tendency.

Our analysis also shows that the phylogenetic proximity of homologous proteins correlates with the statistical structural proximity of the proteomes, indicating the evolutionary roots of the statistical trend discussed in this paper. Although not all the constituent proteins in an organism will evolve in the same direction, statistically, as the organismal complexity increases, the constituent proteins show higher flexibility. The evolution of many protein families is in line with this statistical trend (Brocchieri and Karlin 2005; Kasho et al. 2006). Besides, such a trend is also consistent with the fact that the intrinsic disorder is more abundant in organisms with higher complexity (Niklas et al. 2014; Basile et al. 2019). Not only will the intrinsic disorder proteins (or regions) exhibit high dynamical plasticity, as shown in previous research (Meier and Özbek 2007; Tokuriki et al. 2008; Tokuriki and Tawfik 2009; Marsh and Teichmann 2014), but they may also exhibit high evolvability (towards more specialized function and towards new folds). Still, it is worth noting that, as thermostability becomes a selection pressure, the proteins can evolve to be less flexible (Berezovsky and Shakhnovich 2005). This statistical trend is also parallel with the phenomenon of evolutionary dimensional reduction observed in other experimental and theoretical studies (Hemery and Rivoire 2015; Dutta et al. 2018; Furusawa and Kaneko 2018; Eckmann et al. 2019; Sakata and Kaneko 2020; Sato and Kaneko 2020; Eckmann and Tlusty 2021). In protein dynamics, it is also implied in the emergence of the funnel-like energy landscape (Onuchic et al. 1997), and may contribute to the efficiency and robustness of the protein function. For allosteric proteins, it is observed that evolution designs the sequence and shapes the structure of the proteins, leading toward more specific transition pathways (Togashi et al. 2007; Li et al. 2011).

Moreover, the statistical correlation between proteins' functional specialization and organismal complexity is in line with the experimental observations that ancestral enzymes are likely to have high promiscuity, as they may be able to catalyze a wide variety of chemical reactions (O'Loughlin et al. 2006; Khersonsky and Tawfik 2010; Takano et al. 2013; Petrovic et al. 2018; Gardner et al. 2020; Modi et al. 2021). It is widely observed that ancient generalist proteins tend to evolve toward specialists (Soskine and Tawfik 2010). Previous research in designing thermally stable and promiscuous enzymes with ancestral sequence reconstruction can act as additional support of such a trend (Wheeler et al. 2016; Trudeau and Tawfik 2019; Pinto et al. 2022). Notably, the correlation between protein flexibility and functional specialization is also statistical. There are counterexamples that proteins with higher conformational plasticity show higher promiscuity (Campbell et al. 2018).

The high thermal stability (which is usually associated with low flexibility and high promiscuity) of ancestral enzymes may in turn be compatible with the low complexity of the ancestral species. Organisms with lower complexity have relatively smaller genomes and fewer kinds of enzymes (Gagler et al. 2022). Despite a small genome size, promiscuous enzymes help these organisms achieve a variety of life activities. Conversely, larger genomes can encode more proteins capable of performing highly specialized functions and coping with more complex and diverse cellular environments. The specialization and diversification of proteins enable them to function in a more complex and diverse cellular environment. Consequently, complex organisms can perform their biological functions more effectively, acquiring the

plasticity to adapt to complex and diverse external environments. The compatibility between organismal complexity and functional specialization of constituent proteins suggests the interdependence between the whole and parts for biological systems and other kinds of complex systems, shedding light on the design of complex systems. As a system becomes more complex, its components or elements should change their properties (e.g., become more plastic or modularized).

In summary, based on the AlphaFold DB, we establish a framework for the comparative analyses of the constituent proteins of different organisms. The statistics of AlphaFold-predicted structures have revealed the correlations between the complexity of organisms and features related to the structures (topology), dynamics, and sequences of constituent proteins. Moreover, we show that the phylogenetic proximity is correlated with the statistical structural proximity of the organisms' proteomes, indicating the connections between protein evolution and organism evolution. Our analysis suggests a statistical trend in protein evolution, that is, as organisms evolve toward higher complexity, their constituent proteins may evolve toward higher flexibility and structural diversity, statistically. In the future, the proteome analysis based on AI-predicted protein structures, integrated with other kinds of bioinformation such as protein-protein interactions (Maslov and Sneppen 2002; Zhang et al. 2008), expression levels (Drummond et al. 2005), evolutionary speed (Agozzino and Dill 2018), etc., will definitely offer us new insight into the behaviors and evolution of cells and organisms.

Materials and Methods

Data Availability. All the protein structures used in this study can be downloaded from AlphaFold Protein Structure Database (Varadi et al. 2022). A detailed description of the dataset is listed in SI Table S1. The curated data underlying this article are available in Github, at https://github.com/qianyuantang/stat-trend-protein-evo.

Residue contact network. We construct the residue contact networks based on the native structure of the proteins. The residues (represented as their C_α atoms) are modeled as the nodes of the network. Our calculations of the modularity and assortativity are based on the residue contact networks corresponding to the native structure. When the mutual distance between two residues (nodes) is smaller than the cutoff distance r_C, then the two residues will be connected with an edge. In this work, we take $r_C = 8$ Å.

Elastic network model (ENM). By modeling the edges in the residue contact networks as linear springs, the elastic network models describe the equilibrium fluctuations of proteins as vibrations around the native conformation. These fluctuations are closely related to the functional dynamics of the proteins. In this work, our discussions are based on the Gaussian network model (GNM), the simplest form of the ENM, where the residue fluctuations are assumed to be Gaussian variables distributed around the equilibrium coordinates. The dynamics predicted by GNM can be well matched to experimental or simulation results (Haliloglu et al. 1997; Bahar et al. 2010). With GNM, the potential energy of a protein with chain length

N is given as: $V_{\mathrm{GNM}} = \frac{\kappa}{2}\sum_{i,j=1}^{N} \Delta r_i \Gamma_{ij} \Delta r_j$, where κ is a uniform force constant; Δr_i and Δr_j denote the displacement of residues i and j, respectively; and Γ_{ij} is the element of Kirchhoff matrix Γ. For residues i and j ($i \neq j$), if their mutual distance $r_{ij} \leq r_C$, then $\Gamma_{ij} = -1$; if $r_{ij} > r_C$, then $\Gamma_{ij} = 0$; and for the diagonal elements, $\Gamma_{ii} = -k_i = -\sum_{j \neq i} \Gamma_{ij}$, where k_i denote the degree of node i. Here we take $r_C = 8$ Å.

Normal mode analysis. The eigendecomposition of matrix Γ gives the eigenvalues and the corresponding eigenvectors related to the motions of normal modes (Case 1994; Hayward et al. 1995; Wako and Endo 2017). To compare the eigenvalues for the proteins with different chain lengths, the diagonal elements of matrices Γ are normalized as 1. The normalized matrix is also known as the symmetric normalized graph Laplacian (Atilgan et al. 2012): $L = K^{-1/2}\Gamma K^{-1/2}$, where K is a diagonal matrix $K = \mathrm{diag}[k_1, k_2, \cdots k_N]$ describing the local packing density of the residues. Diagonalizing matrix L, we have $L = U\Lambda U^T$, in which matrix $\Lambda = \mathrm{diag}[\lambda_0, \lambda_1, \lambda_2, \cdots \lambda_{N-1}]$ ($0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{N-1}$) represents the eigenvalues, and matrix $U = [u_0, u_1, u_2, \cdots, u_{N-1}]^T$ denotes to the eigenvectors. The zeroth eigenvector u_0 corresponds to the translational and rotational motions, other nonzero modes correspond to the vibrations of the proteins. The nonzero eigenvalue λ_i is proportional to the square of the vibrational frequency ω_i, i.e., $\lambda_i \sim \omega_i^2$. According to the equipartition of energy, the vibration amplitude A_i is inversely proportional to the frequency: $A_i \sim 1/\omega_i$. Thus, for two nonzero modes i and j, there is $A_i^2/A_j^2 = \lambda_j/\lambda_i$. We use the Python package ProDy to calculate the normal modes of the proteins (Bakan et al. 2011).

Power-law fitting and the Zipf's coefficient. To fit the Zipf's law $\lambda_i^{-1} \sim i^{-z}$ and obtain the coefficient z, in the calculation, we select the top 25% of the eigenvalues to perform the power-law fitting (Alstott et al. 2014). Note that the observed evolutionary trend as indicated by Zipf's coefficient z are robust to the details related to the normal mode calculation and power law fitting (see SI Fig. S7).

Hydropathy variation. In the calculation, we first obtain the original hydropathy data of a sequence according to the Zimmerman hydrophobicity scale of amino acid residues (Zimmerman et al. 1968). Then, a moving average with window length $l_w = 7$ is introduced to calculate the hydropathy profile. If there are no significant hydrophobic-hydrophilic segregations in a sequence, the moving average will smooth out the differences between hydrophobic and hydrophilic amino acids, which will lead to little variation in the filtered hydropathy profile. In contrast, when there are significant hydropathy segregations in a sequence, there will be large variations in the filtered hydropathy profile. Thus, we introduce the coefficient of variation $\mathrm{CV_{HP}}$ (defined as the standard deviation divided by the mean value of the filtered hydropathy profile) to quantify the sequential segregation of hydrophobic and hydrophilic residues.

Acknowledgments

We gratefully thank Taro Toyoizumi, Xiangze Zeng, Hisao Moriya, Haobo Wang, Haiguang Liu, Zhengqi He, Sida Chen, Weiyi Qiu, Wenfei Li, Zeke Xie, Yingnan Li, Xinhong Liu, Lei-Han Tang and Xuefei Li

for participating in stimulating discussions. This work was supported by Brain/MINDS from Japan Agency for Medical Research and Development (grant number JP21dm0207001), National Natural Science Foundation of China (grant number 11774157), a Grant-in-Aid for Scientific Research on Innovative Areas (grant number 17H06386) from the Ministry of Education, Culture, Sports, Science and Technology of Japan, a Grant-in-Aid for Scientific Research (grant number (A)20H00123) from the Japanese Society for the Promotion of Science, and Novo Nordisk Fonden (to K.K.).

References

Adami C, Ofria C, Collier TC. 2000. Evolution of biological complexity. *Proc. Natl. Acad. Sci. USA* 97:4463–4468.

Agarwal PK. 2006. Enzymes: An integrated view of structure, dynamics and function. *Microb Cell Fact* 5:1–12.

Agozzino L, Dill KA. 2018. Protein evolution speed depends on its stability and abundance and on chaperone concentrations. *Proc. Natl. Acad. Sci. USA* 115:9092–9097.

Alstott J, Bullmore E, Plenz D. 2014. Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS One* 9:e95816.

Atilgan C, Okan OB, Atilgan AR. 2012. Network-based models as tools hinting at nonevident protein functionality. *Annu Rev Biophys* 41:205–225.

Bagdonas H, Fogarty CA, Fadda E, Agirre J. 2021. The case for post-predictional modifications in the AlphaFold Protein Structure Database. *Nat Struct Mol Biol* 28:869–870.

Bahar I, Atilgan AR, Demirel MC, Erman B. 1998. Vibrational Dynamics of Folded Proteins: Significance of Slow and Fast Motions in Relation to Function and Stability. *Phys Rev Lett* 80:2733.

Bahar I, Lezon TR, Yang LW, Eyal E. 2010. Global dynamics of proteins: Bridging between structure and function. *Annu Rev Biophys* 39:23–42.

Bakan A, Meireles LM, Bahar I. 2011. ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* 27:1575–1577.

Basile W, Salvatore M, Bassot C, Elofsson A. 2019. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput Biol* 15:e1007186.

Bayly-Jones C, Whisstock JC. 2022. Mining folded proteomes in the era of accurate structure prediction. *PLoS Comput Biol* 18:e1009930.

Berezovsky IN, Shakhnovich EI. 2005. Physics and evolution of thermophilic adaptation. *Proc. Natl. Acad. Sci. USA* 102:12742–12747.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res* 28:235–242.

Brocchieri L, Karlin S. 2005. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res* 33:3390–3400.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, Fitzgerald LM, Clayton RA, Gocayne JD, et al. 1996. Complete Genome Sequence of the Methanogenic Archaeon, Methanococcus jannaschii. *Science (1979)* 273:1058–1073.

Caetano-Anollés G, Aziz MF, Mughal F, Caetano-Anollés D. 2021. Tracing protein and proteome history with chronologies and networks: folding recapitulates evolution. *Expert Rev Proteomics* 18:863–880.

Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE. 2009. The origin, evolution and structure of the protein world. *Biochem J* 417:621–637.

Campbell EC, Correy GJ, Mabbitt PD, Buckle AM, Tokiriki N, Jackson CJ. 2018. Laboratory evolution of protein conformational dynamics. *Curr Opin Struct Biol* 50:49–57.

Case DA. 1994. Normal mode analysis of protein dynamics. *Curr Opin Struct Biol* 4:285–290.

Choi I-G, Kim S-H. 2006. Evolution of protein structural classes and protein sequence families. *Proc. Natl. Acad. Sci. USA* 103:14506–14061.

Dill KA. 1990. Dominant forces in protein folding. *Biochemistry* 29:7133–7155.

Dill KA, MacCallum JL. 2012. The protein-folding problem, 50 years on. *Science (1979)* 338:1042–1046.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. USA* 102:14338–14343.

Dutta S, Eckmann JP, Libchaber A, Tlusty T. 2018. Green function of correlated genes in a minimal mechanical model of protein evolution. *Proc. Natl. Acad. Sci. USA* 115:E4559–E4568.

Eckmann JP, Rougemont J, Tlusty T. 2019. Colloquium: Proteins: the physics of amorphous evolving matter. *Rev Mod Phys* 91:031001.

Eckmann JP, Tlusty T. 2021. Dimensional reduction in complex living systems: Where, why, and how. *BioEssays* 43:2100062.

Espada R, Parra RG, Mora T, Walczak AM, Ferreiro DU. 2015. Capturing coevolutionary signals in repeat proteins. *BMC Bioinformatics* 16:207.

Finnigan GC, Hanson-Smith V, Stevens TH, Thornton JW. 2012. Evolution of increased complexity in a molecular machine. *Nature* 481:360–364.

Foy SG, Wilson BA, Bertram J, Cordes MHJ, Masel J. 2019. A Shift in Aggregation Avoidance Strategy Marks a Long-Term Direction to Protein Evolution. *Genetics* 211:1345.

Friedman JM. 1985. Structure, Dynamics, and Reactivity in Hemoglobin. *Science (1979)* 228:1273–1280.

Furusawa C, Kaneko K. 2000. Origin of Complexity in Multicellular Organisms. *Phys Rev Lett* 84:6130.

Furusawa C, Kaneko K. 2018. Formation of dominant mode by evolution in biological systems. *Phys Rev E* 97:42410.

Gagler DC, Karas B, Kempes CP, Malloy J, Mierzejewski V, Goldman AD, Kim H, Walker SI. 2022. Scaling laws in enzyme function reveal a new kind of biochemical universality. *Proc. Natl. Acad. Sci. USA* 119.

Gardner JM, Biler M, Risso VA, Sanchez-Ruiz JM, Kamerlin SCL. 2020. Manipulating Conformational Dynamics to Repurpose Ancient Proteins for Modern Catalytic Functions. *ACS Catal* 10:4863–4870.

Gerstein M, Sonnhammer ELL, Chothia C. 1994. Volume changes in protein evolution. *J Mol Biol* 236:1067–1078.

Ghosh K, Dill K. 2010. Cellular proteomes have broad distributions of protein stability. *Biophys J* 99:3996–4002.

Gish W, States DJ. 1993. Identification of protein coding regions by database similarity search. *Nat Genet* 3:266–272.

Guimerà R, Sales-Pardo M, Amaral LAN. 2004. Modularity from fluctuations in random graphs and complex networks. *Phys Rev E* 70:025101(R).

Haliloglu T, Bahar I. 2015. Adaptability of protein structures to enable functional interactions and evolutionary implications. *Curr Opin Struct Biol* 35:17–23.

Haliloglu T, Bahar I, Erman B. 1997. Gaussian Dynamics of Folded Proteins. *Phys Rev Lett* 79:3090–3093.

Halle B. 2002. Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA* 99:1274–1279.

Hayward S, Kitao A, Gō N. 1995. Harmonicity and Anharmonicity in Protein Dynamics: A Normal Mode Analysis and Principal Component Analysis. *Proteins: Structure, Function, and Genetics* 23:177–186.

Hemery M, Rivoire O. 2015. Evolution of sparsity and modularity in a model of protein allostery. *Phys Rev E* 91:042704.

Jones WJ, Leigh JA, Mayer F, Woese CR, Wolfe RS. 1983. Methanococcus jannaschii sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. *Arch. Microbiol.* 136:254–261.

Joosten RP, te Beek TAH, Krieger E, Hekkelman ML, Hooft RWW, Schneider R, Sander C, Vriend G. 2011. A series of PDB related databases for everyday needs. *Nucleic Acids Res* 39:D411.

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

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637.

Kasho VN, Smirnova IN, Kaback HR. 2006. Sequence Alignment and Homology Threading Reveals Prokaryotic and Eukaryotic Proteins Similar to Lactose Permease. *J Mol Biol* 358:1060.

Khersonsky O, Tawfik DS. 2010. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annu Rev Biochem* 79:471–505.

Koonin EV, Wolf YI, Karev GP. 2002. The structure of the protein universe and genome evolution. *Nature* 420:218–223.

Labas YA, Gurskaya NG, Yanushevich YG, Fradkov AF, Lukyanov KA, Lukyanov SA, Matz M v. 2002. Diversity and evolution of the green fluorescent protein family. *Proc. Natl. Acad. Sci. USA* 99:4256–4261.

Lewis M, Rees DC. 1985. Fractal Surfaces of Proteins. *Science (1979)* 230:1163–1165.

Li W, Wolynes PG, Takada S. 2011. Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins. *Proc. Natl. Acad. Sci. USA* 108:3504–3509.

Liang J, Dill KA. 2001. Are Proteins Well-Packed? *Biophys J* 81:751–766.

Liu J, Rost B. 2001. Comparing function and structure between entire proteomes. *Protein Science* 10:1970–1979.

Lloyd S. 2001. Measures of Complexity: A Nonexhaustive List. *IEEE Control Syst* 21:7–8.

Markov A v., Anisimov VA, Korotayev A v. 2010. Relationship between genome size and organismal complexity in the lineage leading from prokaryotes to mammals. *Paleontological Journal 2010 44:4* 44:363–373.

Marsh JA, Teichmann SA. 2014. Protein Flexibility Facilitates Quaternary Structure Assembly and Evolution. *PLoS Biol* 12:e1001870.

Maslov S, Sneppen K. 2002. Specificity and stability in topology of protein networks. *Science (1979)* 296:910–913.

Massey FJ. 1951. The Kolmogorov-Smirnov Test for Goodness of Fit. *J Am Stat Assoc* 46:68–78.

Maynard Smith J, Szathmary E. 1997. The Major Transitions in Evolution. New York: Oxford University Press

McGinnis S, Madden TL. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res* 32:W20.

McShea DW. 1996. Perspective: Metazoan Complexity and Evolution: Is There a Trend? *Evolution (N Y)* 50:492.

Meier S, Özbek S. 2007. A biological cosmos of parallel universes: Does protein structural plasticity facilitate evolution? *BioEssays* 29:1095–1104.

Modi T, Risso VA, Martinez-Rodriguez S, Gavira JA, Mebrat MD, van Horn WD, Sanchez-Ruiz JM, Banu Ozkan S. 2021. Hinge-shift mechanism as a protein design principle for the evolution of β-lactamases from substrate promiscuity to specificity. *Nat Commun* 12:1852.

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. 2011. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA* 108:E1293–E1301.

Moret MA, Zebende GF, Phillips JC. 2019. Hydropathic wave ordering of alpha crystallin—Membrane interactions enhances human lens transparency and resists cataracts. *Physica A* 514:573–579.

Newman MEJ. 2002. Assortative Mixing in Networks. *Phys Rev Lett* 89:208701.

Newman MEJ. 2003. Mixing patterns in networks. *Phys Rev E* 67:026126.

Niklas KJ, Cobb ED, Dunker AK. 2014. The number of cell types, information content, and the evolution of complex multicellularity. *Acta Societatis Botanicorum Poloniae* 83:337–347.

O'Loughlin TL, Patrick WM, Matsumura I. 2006. Natural history as a predictor of protein evolvability. *Protein Engineering, Design and Selection* 19:439–442.

Onuchic JN, Luthey-Schulten Z, Wolynes PG. 1997. Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* 48:545–600.

Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, Maksimova ES, Kondrashov FA, Ivankov DN. 2021. Using AlphaFold to predict the impact of single mutations on protein stability and function. *bioRxiv*:2021.09.19.460937.

Pál C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet* 7:337–348.

Petrovic D, Risso VA, Kamerlin SCL, Sanchez-Ruiz JM. 2018. Conformational dynamics and enzyme evolution. *J R Soc Interface* 15.

Phillips JC. 2009. Scaling and self-organized criticality in proteins II. *Proc. Natl. Acad. Sci. USA* 106:3113–3118.

Phillips JC. 2012. Hydropathic Self-Organized Criticality: A Magic Wand for Protein Physics. *Protein Pept Lett* 19:1089–1093.

Phillips JC. 2020. Self-organized networks: Darwinian evolution of dynein rings, stalks, and stalk heads. *Proc. Natl. Acad. Sci. USA* 117:7799–7802.

Pin JP, Galvez T, Prézeau L. 2003. Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacol Ther* 98:325–354.

Pinto GP, Corbella M, Demkiv AO, Kamerlin SCL. 2022. Exploiting enzyme evolution for computational protein design. *Trends Biochem Sci* 47:375–389.

Reuveni S, Granek R, Klafter J. 2008. Proteins: Coexistence of stability and flexibility. *Phys Rev Lett* 100:208101.

Robertson AJ, Courtney JM, Shen Y, Ying J, Bax A. 2021. Concordance of X-ray and AlphaFold2 Models of SARS-CoV-2 Main Protease with Residual Dipolar Couplings Measured in Solution. *J Am Chem Soc* 143:19306–19310.

Ruff KM, Pappu R v. 2021. AlphaFold and Implications for Intrinsically Disordered Proteins. *J Mol Biol* 433.

Sakata A, Kaneko K. 2020. Dimensional Reduction in Evolving Spin-Glass Model: Correlation of Phenotypic Responses to Environmental and Mutational Changes. *Phys Rev Lett* 124:218101.

Sato TU, Kaneko K. 2020. Evolutionary dimension reduction in phenotypic space. *Phys Rev Res* 2:013197.

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, et al. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577:706–710.

Sikosek T, Chan HS. 2014. Biophysics of protein evolution and evolutionary protein biophysics. *J R Soc Interface* 11:20140419.

Soskine M, Tawfik DS. 2010. Mutational effects and the evolution of new protein functions. *Nat Rev Genet* 11:572–582.

Takano K, Aoi A, Koga Y, Kanaya S. 2013. Evolvability of thermophilic proteins from archaea and bacteria. *Biochemistry* 52:4774–4780.

Tang Q-Y, Hatakeyama TS, Kaneko K. 2020. Functional sensitivity and mutational robustness of proteins. *Phys Rev Res* 2:033452.

Tang Q-Y, Kaneko K. 2020. Long-range correlation in protein dynamics: Confirmation by structural data and normal mode analysis. *PLoS Comput Biol* 16:e1007670.

Tang Q-Y, Kaneko K. 2021. Dynamics-Evolution Correspondence in Protein Structures. *Phys Rev Lett* 127:098103.

Tang Q-Y, Zhang Y-Y, Wang J, Wang W, Chialvo DR. 2017. Critical Fluctuations in the Native State of Proteins. *Phys Rev Lett* 118:088102.

Togashi Y, Mikhailov AS, Ertl G. 2007. Nonlinear relaxation dynamics in elastic networks and design principles of molecular machines. *Proc. Natl. Acad. Sci. USA* 104:8697–8702.

Tokuriki N, Stricher F, Serrano L, Tawfik DS. 2008. How protein stability and new functions trade off. *PLoS Comput Biol* 4:e1000002.

Tokuriki N, Tawfik DS. 2009. Protein dynamism and evolvability. *Science (1979)* 324:203–207.

Trudeau DL, Tawfik DS. 2019. Protein engineers turned evolutionists—the quest for the optimal starting point. *Curr Opin Biotechnol* 60:46–52.

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

Wako H, Endo S. 2017. Normal mode analysis as a method to derive protein dynamics information from the Protein Data Bank. *Biophys Rev* 9:877–893.

Wheeler LC, Lim SA, Marqusee S, Harms MJ. 2016. The thermostability and specificity of ancient proteins. *Curr Opin Struct Biol* 38:37–43.

Xie Z, Tang Q-Y, Cai Y, Sun M, Li P. 2022. On the Power-Law Spectrum in Deep Learning: A Bridge to Protein Science. *Arxiv: 2201.13011*.

Zardoya R. 2005. Phylogeny and evolution of the major intrinsic protein family. *Biol Cell* 97:397–414.

Zeldovich KB, Shakhnovich EI. 2008. Understanding Protein Evolution: From Protein Physics to Darwinian Selection. *Annu Rev Phys Chem* 59:105–127.

Zhang J, Maslov S, Shakhnovich EI. 2008. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. *Mol Syst Biol* 4:210.

Zimmerman JM, Eliezer N, Simha R. 1968. The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol* 21:170–201.