The geometric properties of the molecular surfaces of protein crystal structures and AlphaFold predicted models

(I) A hypothesis on protein structure and folding from a molecular surface perspective

(II) Large differences between the AlphaFold models for prokaryotic and eukaryotic proteomes

Lincong Wang*

The College of Computer Science and Technology and the Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China

Abstract

Previous analyses of protein structures have focused primarily on three-dimensional folds, their secondary structures, and binding or active sites while their molecular surfaces have received much less attention, due possibly to the lack of accurate and robust programs for their computation.

Using SESA we have analyzed the molecular surfaces of three mutually exclusive sets, $\mathbf{G}, \mathbf{S}$ and $\mathbf{M}$, of protein crystal structures. $\mathbf{G}$ and $\mathbf{S}$ include only non-membrane proteins with the latter having only monomers while $\mathbf{M}$ has only membrane proteins. The analyses show that SAS area per atom μ_s decreases while probe area per atom μ_p increases with the number of atoms in a molecule n. Most interestingly, the fitted power laws for μ_s intersect with those for μ_p at $n = 957$ for $\mathbf{G}$, $n = 875$ for $\mathbf{S}$ and $n = 1,061$ for $\mathbf{M}$. They correspond respectively to $60, 57$ and 64 amino acid residues. The power laws and their intersections provide an explanation for protein structural integrity and stability in general and the transition in particular from peptides typically with random conformations in solution to proteins usually with a dominant conformation.

We have also analyzed the molecular surfaces of the AlphaFold models for twenty seven proteomes. The analyses show that the molecular surfaces for thirteen prokaryotic proteomes resemble those for the crystal structures while those for fourteen eukaryotic ones differ largely from both of them. The variation may have significant implication in theory in that there exist genuine differences between prokaryotic and eukaryotic proteomes, and in application in that the current AlphaFold models for eukaryotic proteomes are likely not adequate for structure-based drug design in particular.

Molecular surface | Solvent-excluded surface | Peptide | Protein | Protein structure | Protein folding | Proteome | Neural network | AlphaFold.

1 Introduction

The experimentally-derived structures of proteins have been analyzed in depth to rationalize their physical properties such as solvation and folding, and their biochemical properties such as binding affinity [1] and catalytic efficiency [2]. Previous analyses though extensive focus largely on three-dimensional (3D)[1] folds, secondary structure content, and the physical-chemical and geometric properties of ligand-protein interface. By contrast, protein surface has not been as well investigated. For example, the contributions of surface to the solvation, folding and structural stability of proteins have not been accurately quantified [3, 4, 5, 6, 7, 8, 9].

*Corresponding author: Lincong Wang, Email: wlincong@outlook.com, wlincong@gmail.com, wanglincong@jlu.edu.cn

[1]Abbreviations: 2D, two-dimensional; 3D, three-dimensional; SES, solvent-excluded surface; SAS, solvent-accessible surface; SASA, solvent-accessible surface area; RMSD, root-mean square deviation; RMSE: root-mean squared error; PDB, Protein Data Bank; TM-score, template modelling score; GDT, global distance test; pLDDT, predicted local distance difference test; SI, Supporting Information.

Significance statement

A newly-developed analytic and robust program, SESA*, has been applied to three mutually exclusive sets,* $\mathbf{G}, \mathbf{S}$ *and* $\mathbf{M}$*, of protein crystal structures and the AlphaFold models for twenty seven proteomes to compute their exterior solvent-excluded surface (SES) areas. The results show that for the crystal structures the areas per atom for SAS* μ_s*, probe* μ_p *and toroidal* μ_t *patches each follows a power law with* n*, the number of atoms in a structure or model. Specifically,* μ_s *decreases while* μ_p *increases with* n*. Most interestingly, the power laws for* μ_s *intersect with those for* μ_p *at* $n = 957$ *for* $\mathbf{G}$*,* $n = 875$ *for* $\mathbf{S}$ *and* $n = 1,061$ *for* $\mathbf{M}$*. They correspond respectively to* $60, 57$ *and* 64 *residues. A SAS patch is convex while a probe one concave, thus a power law for* μ_s *intersects with that for* μ_p *when the total area of the patches with a negative curvature equals that with a positive curvature if one ignores toroidal patches. The points of intersection for* $\mathbf{G}$ *and* $\mathbf{S}$ *are close to the number of residues required for a polypeptide to adopt a dominant conformation in solution, and thus provide an explanation for why a chain with* < 50 *residues, that is, a peptide, has in general only random conformations in solution. In addition, the SESs of the AlphaFold models for thirteen prokaryotic proteomes resemble those for the crystal structures. However, in stark contrast with the crystal structures and the models for prokaryotic proteomes, the SESs for fourteen eukaryotic proteomes differ largely from both of them. The differences likely have significant implications for structural biology and the applications of AlphaFold models.*

There exist three main models for the surface of a molecule called, respectively, van der Waals (VDW) surface, solvent accessible surface (SAS) [10, 11] and molecular surface (also called solvent-excluded surface (SES)[2]) [11, 12]. A SES is a two-dimensional (2D) manifold impenetrable to solvent molecules, and is composed of three types of 2D patches: convex spherical polygon on a solvent-accessible atom, saddle-shaped toroidal patch determined by two accessible atoms, and concave spherical polygon on a probe by three accessible atoms. Their surface areas are called, respectively, SAS area, toroidal area and probe area.

To evaluate SES's contributions to protein solvation, folding and structural stability, and to obtain clue to the adaptation of native proteins to their environments, we have applied our robust and analytic SES area computation program SESA [13] to three mutually-exclusive sets of crystal structures, $\mathbf{G}, \mathbf{S}$ and $\mathbf{M}$. $\mathbf{G}$ and $\mathbf{S}$ include only non-membrane proteins with the former having both monomers and multimers and the latter only monomers, while $\mathbf{M}$ includes only membrane proteins. Importantly, the structures in $\mathbf{G}$ have no missing internal residues[3] and a very low percentage of missing atoms per structure. Out of several SES geometric properties examined we found that SAS, probe and toroidal areas per atom, μ_s, μ_p and μ_t, each changes with the number of atoms in a structure n by following a power law. Specifically, μ_s decreases while μ_p increases with n. Most interestingly, the power laws for μ_s intersect with those for μ_p at $n = 957$ for $\mathbf{G}$, $n = 875$ for $\mathbf{S}$ and $n = 1,061$ for $\mathbf{M}$. They correspond, respectively, to $60, 57$ and 64 residues. An intersection occurs when $\mu_s(n) = \mu_p(n)$, i.e. when the total SAS area of a molecule equals its total probe area. A SAS patch is convex while a probe one concave, thus a power law for μ_s intersects with that for μ_p when the total area of the patches with a negative curvature equals that with a positive curvature (if one ignores the toroidal patches, each of which includes regions of positive and negative curvatures). The points of intersection for $\mathbf{G}$ and $\mathbf{S}$ are close to the number of residues required for a polypeptide to have a dominant conformation in solution. The variations of μ_s, μ_p and μ_t with molecular size (n) together with the power laws and their intersections provide an explanation for why a polypeptide of < 50 residues, that is, a peptide, has in general only random conformations in solution. In other words, the power laws and their intersections may explain why the transition from the broad conformational distribution of a typical peptide to the narrow one of a regular protein occurs between $50 - 60$ residues. Furthermore, we hypothesize that the ratio $\gamma = \frac{\mu_p}{\mu_s}$ is closely related to protein stability and folding in the sense that $\gamma > 1$ is a necessary condition for the folding of a typical water-soluble protein.

The neural network-based program, AlphaFold [14, 15, 16, 17], represents a leap-forward in protein structure prediction in terms of structural coverage, and accuracy as judged by the RMSDs for both backbone and all atoms, global distance test (GDT), TM-score, electron density map [18], and other criteria [19, 20, 21, 22, 23, 24]. However, to our best knowledge neither the SASs nor the SESs of the predicted models have been analyzed. Our analyses of the SESs of the $350,974$ AlphaFold models for thirteen prokaryotic and fourteen eukaryotic proteomes show that the SESs for the former resemble those for crystal structures while the SESs for the latter differ largely from them. Specifically,

[2] In this paper, SES and molecular surface are used interchangeably.

[3] An internal residue means a residue between the N-terminal and C-terminal residues.

the μs[4] of the models for the eukaryotic proteomes are broadly distributed, hardly change with n and only minimally follow a power law. The differences may have significant implication in theory in that there exist genuine differences between prokaryotic and eukaryotic proteomes, and in application in that the current AlphaFold models for eukaryotic proteomes are likely not adequate for structure-based drug design in particular.

2 Materials and Methods

In this section we first describe the three sets of crystal structures and the twenty seven AlphaFold proteomes, and then briefly present the computation of SES area and the definitions of SES patch areas per surface atom (μs). Finally, we describe the fittings to power law of the changes with n of μs, and to linear equation of the change with n of the number of surface atom n_s of a molecule.

2.1 Data sets and the preprocessing of the PDB files for SES area computation

From the PDB we downloaded three mutually exclusive sets of protein crystal structures, $\mathbf{S}, \mathbf{G}$ and $\mathbf{M}$. $\mathbf{S}$ has $4,147$ structures of monomeric proteins with a resolution range of $0.64 - 3.50$Å and mean of 1.68Å, a R_{free} range of $0.05 - 0.350$ and mean of 0.207. The percentage of missing atoms[5] ranges from $0.0 - 2.0\%$ with a mean of 0.299% and a median of 0.051%. $\mathbf{G}$ has $8,616$ structures with a resolution range of $0.68 - 3.50$Å and mean of 1.91Å, a R_{free} range of $0.078 - 0.30$ and mean of 0.224. $\mathbf{G}$ includes 695 monomeric proteins. The percentage of missing atoms ranges from $0.0 - 2.0\%$ with a mean of 0.218% and a median of 0.046%. Neither $\mathbf{S}$ nor $\mathbf{G}$ includes any membrane proteins. $\mathbf{M}$ has $2,487$ structures of membrane proteins with a resolution range of $1.0 - 5.0$Å and mean of 2.70Å, a R_{free} range of $0.096 - 0.458$ and mean of 0.258. The percentage of missing atoms ranges from $0.0 - 2.0\%$ with a mean of 0.253% and a median of 0.052%. Among the three sets, only the structures in $\mathbf{G}$ have no missing internal residues.

From the AlphaFold Protein Structure Database [25] we downloaded $350,974$ models for twenty seven proteomes including thirteen prokaryotic ones, *Escherichia coli, Haemophilus influenzae, Helicobacter pylori, Klebsiella pneumoniae, Methanocaldococcus jannaschii, Mycobacterium leprae, Mycobacterium tuberculosis, Neisseria gonorrhoeae, Pseudomonas aeruginosa, Salmonella typhimurium, Staphylococcus aureus, Streptococcus pneumoniae, Shigella dysenteriae*; and fourteen eukaryotic ones. The latter consists of three yeasts, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Candida albicans*; four plants, *Arabidopsis thaliana, Glycine max, Oryza sativa* and *Zea mays*; and an amoeba, *Dictyostelium discoideum*; an insect, *Drosophila melanogaster*; a worm, *Caenorhabditis elegans*; a fish, *Danio rerio*; and three mammals, *Mus musculus, Rattus norvegicus* and *Homo sapiens*.

The PDB files are preprocessed for area computation as described previously [26]. Specifically, if a PDB file includes multiple models, only the first one is selected, and if an atom has multiple conformations, only the first one is chosen. Protons are added using REDUCE [27] to any PDB structure or model that lacks them. Only protein atoms are used for area computation.

2.2 The computation of SES area and the definitions of patch areas per surface atom

The program for analytic SES area computation, SESA[6], has been described previously [13]. The salient features of SESA relevant to the present study are the algorithmic separation of the exterior surface of a molecule from its interior surfaces, and the analytic and individual computations of the three types of SES patches. An interior surface is composed of surface atoms inaccessible to the bulk solvent without the rearrangement of their neighboring atoms. The analyses presented in this paper use only the SES of the exterior surface of a structure or a model. Specifically, we analyze the following geometric properties of SES: the number of surface atoms, and the SAS, probe and toroidal areas per surface atom defined as follows.

[4] For brevity, the symbol"μs" (SES patch areas per atom) means μ_ss, μ_ps and μ_ts.

[5] The number of missing atoms for a crystal structure is computed as the difference between the number of atoms detected and that expected from the protein sequence.

[6] SESA is freely available at www.github/wlincong/sesA.

2.2.1 The definitions of SAS, toroidal and probe areas per surface atom

Let $\mathbf{A}$ be the set of all the atoms in a molecule. To surface atom $i \in \mathbf{A}$ we assign a nonzero SES area $a(i)$:

$$a(i) = a_s(i) + a_t(i) + a_p(i);$$
$$a_t(i) = \frac{1}{2}\sum_j a_t(i,j), \quad a_p(i) = \frac{1}{3}\sum_{j,k} a_p(i,j,k) \tag{1}$$

where $a_s(i), a_t(i)$ and $a_p(i)$ are respectively the SAS, toroidal and probe areas for atom i; $a_t(i,j)$ is the total area of the toroidal patches determined by atom i and j, and $a_p(i,j,k)$ the total area of spherical polygons on a probe determined by atom i, j and k.

Given the individual patch areas for $\mathbf{A}$, the total SES area A, the SAS, toroidal and probe areas per atom, μ_s, μ_t and μ_p, are computed as follows.

$$A = A_s + A_t + A_p \tag{2}$$
$$\mu_s = \frac{A_s}{n}, \qquad \mu_t = \frac{A_t}{n}, \qquad \mu_p = \frac{A_p}{n} \tag{3}$$
$$A_s = \sum_{i=1}^{n_s} a_s(i), \quad A_t = \sum_{i=1}^{n_s} a_t(i), \quad A_p = \sum_{i=1}^{n_s} a_p(i) \tag{4}$$

where n and n_s are, respectively, the numbers of total atoms and (exterior) surface atoms in a molecule, and A_s, A_t and A_p are, respectively, SAS, toroidal and probe areas. The three summations are all over n_s. SAS area A_s, as a part of A (Eq. 2), differs from solvent-accessible surface area (SASA) with the latter being computed using a sphere radius of $r_a + r_w$ where r_a and r_w are respectively the radii of a molecular atom and the probe [13]. The analyses presented in this paper focus on μ_s, μ_t, μ_p and n_s for the exterior surface of a molecule.

2.3 Fitting to power law and linear equation

The changes with n of the μs for $\mathbf{G}, \mathbf{S}$ and $\mathbf{M}$ and for each of the twenty seven AlphaFold proteomes are best-fitted using MATLAB to a power law with three coefficients, a, b and c.

$$\mu = an^b + c \tag{5}$$

The point of intersection between the power laws for μ_s and μ_p is denoted as (n_{sp}, μ_{sp}). When the power laws are expressed in n_r, the number of residues in a structure or a model, i.e., $\mu = an_r^b + c$, the point of intersection is denoted as (N_{sp}, μ_{sp}). The increase of n_s with n is fitted to a linear equation, $n_s = \tau n + b$, with a slope $\tau = \frac{dn_s}{dn}$. The changes of μs with n, the power laws and their intersections, and the increase of n_s with n and its ratio τ are the main topics of this paper.

3 Results

In this section we describe the geometric properties of the molecular surfaces of both crystal structures and AlphaFold proteomes.

3.1 The geometric properties of the molecular surfaces of crystal structures

We first describe the changes of μs with n, the power laws and their intersections, and then the increase of n_s with n and the linear equation.

3.1.1 The changes of μs with n, the power laws and their intersections for $\mathbf{G}$, $\mathbf{S}$ and $\mathbf{M}$

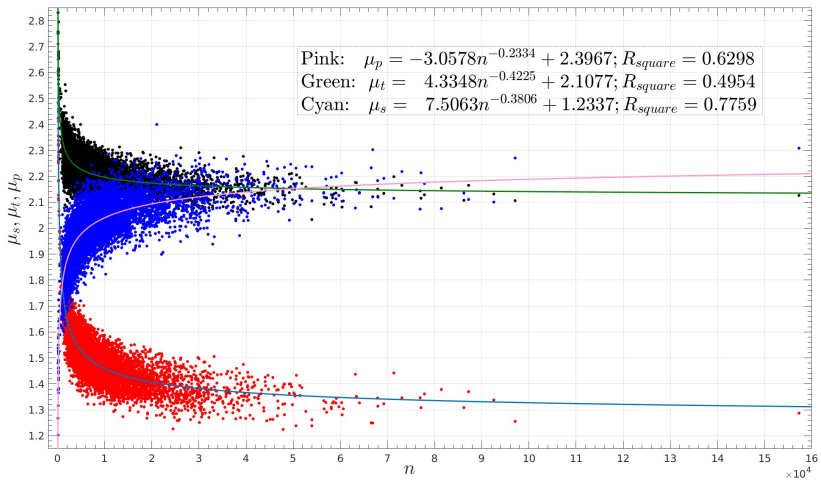
$\mathbf{G}$ has $8,616$ structures. The smallest and the largest ones are, respectively, 3FTR with 6 residues (76 atoms) and 2Y26 with $10,080$ residues ($157,300$ atoms). As shown in Fig. 1a, both μ_s and μ_t decrease with n with the former faster

than the latter while μ_p increases with n. The power laws for μ_s and μ_p are, respectively,

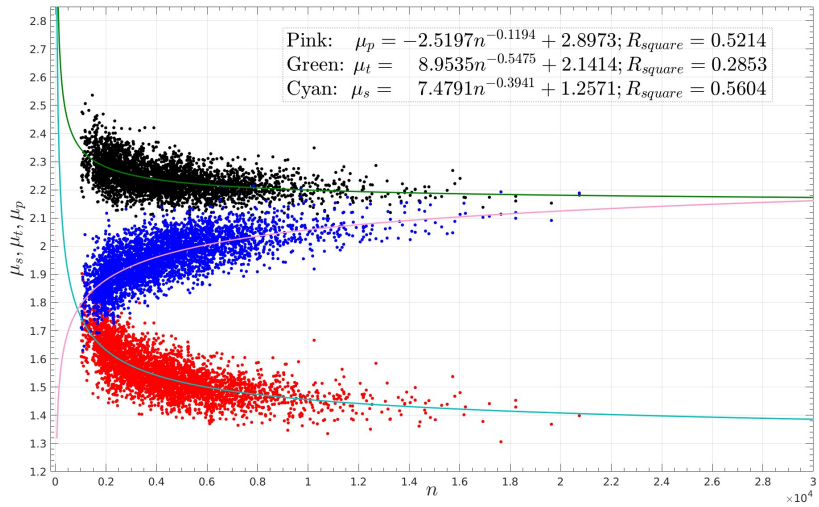
$$\mu_s = \quad 7.5063n^{-0.3806} + 1.2337 \tag{6}$$

$$\mu_p = -3.0578n^{-0.2334} + 2.3967 \tag{7}$$

They intersect at $n_{sp} = 967$ with $\mu_{sp} = 1.782$. It equals 60 residues, i.e., $N_{sp} = 60$, using the fitted linear equation, $n_r = 0.0647n - 2.9278$, where n_r is the number of residues in a protein (Fig. S1 of SI). In addition, the power law for μ_p intersects with that for μ_t at $n = 50,369$ ($3,256$ residues). However, the μ_ts could not be fitted to a power law as well as either the μ_ss or μ_ps (Fig. 1a insert).



(a) **G**



(b) **S**

Figure 1: **The μ_ss, μ_ts and μ_ps of G and S and the power laws**. Figs. (**a**) and (**b**) depict, respectively, the μs and power law curves for **G** and **S**. The μ_ss, μ_ts and μ_ps are colored, respectively, in red, black and blue with the power law curves colored respectively in cyan, green and pink. The inserts list the power laws and MATLAB fitting scores R_{square}s. The x-axes are n and the y-axes are μ_s, μ_t and μ_p in $\mathring{A}^2$.

S has $4,147$ monomers. The smallest and largest ones are, respectively, 3ZR8 with 65 residues ($1,245$ atoms) and 7MHM with $13,158$ residues ($201,246$ atoms). As shown in Fig. 1b and as with **G**, both μ_s and μ_t decrease with n with the former faster than the latter while μ_p increases with n. The power laws for μ_s and μ_p are, respectively, $\mu_s = 7.4791n^{-0.3941} + 1.2571$ and $\mu_p = -2.5197n^{-0.1194} + 2.8973$. They intersect at $n_{sp} = 875$ with $\mu_{sp} = 1.775$. Using the linear equation between n_r and n, $n_r = 0.0643n + 0.9134$ with a $R_{square} = 0.9951$, it equals $N_{sp} = 57$

residues, a purely theoretical number since all the monomers have $n_r \geq 65$ residues. Furthermore, the differences in μs between **S** and **G** are very small and their power laws are very close to each other (Table 1 and Fig. S2). In addition, the power law for μ_p intersects with that for μ_t at $n = 33,540$ ($2,158$ residues). However, the μ_ts could not be fitted to a power law as well as either the μ_ss or μ_ps (Fig. 1b inserts).

Set	$\Delta\mu_s$ (mean / std) ($\mathring{A}^2$)	$\Delta\mu_p$ (mean / std) ($\mathring{A}^2$)	$P_s(\%)$	$P_p(\%)$
S	-0.0121 / 0.0535	0.0070 / 0.0557	5.5	2.5
M	-0.0335 / 0.0606	-0.0155 / 0.0680	12.2	6.4

Table 1: **The differences in μs between G and S, and between G and M.** Column 1 lists set names. The differences of the $\mu_s(i)$ and $\mu_p(i)$ of structure i of either **S** or **M** from the μ_ss and μ_ps of **G** are quantified as $\Delta\mu_s(i) = \mu_s(i) - \mu_s^g(i)$ and $\Delta\mu_p(i) = \mu_p(i) - \mu_p^g(i)$, where $\mu_s^g(i), \mu_p^g(i)$ are computed using, respectively, Eqn. (6) and Eqn. (7) with n being the number of atoms in structure i. The smaller a Δ is, the closer the two power laws are. Column 2 lists the means and standard deviations of the lists of the $\Delta\mu_s$s for **S** and **M**. Column 3 lists the means and standard deviations of the lists of the $\Delta\mu_p$s for **S** and **M**. Column 4 lists P_ss, the percentages of the structures in **S** and **M** whose $\Delta\mu_s > 2 \times \text{RMSE}_{G,s}$, where $\text{RMSE}_{G,s}$ is the RMSE (0.0532) of fitting to a power law the change with n of the μ_ss of **G**. Column 5 lists P_ps, the percentages of the structures in **S** and **M** whose $\Delta\mu_p > 2 \times \text{RMSE}_{G,p}$, where $\text{RMSE}_{G,p}$ is the RMSE (0.0631) of fitting to a power law the change with n of the μ_ps of **G**. Fig. S2 visualizes the P_s and P_p for **S**, and Figs. S3b and S3c visualize the P_s and P_p for **M**. Assuming a normal distribution, only 4.5% of the μ_ss for **G** have a $\Delta\mu_s > 2 \times \text{RMSE}_{G,s}$, and only 4.5% of the μ_ps for **G** have a $\Delta\mu_p > 2 \times \text{RMSE}_{G,p}$. The smaller a P_s or a P_p is, the closer the two distributions are.

M has $2,487$ structures. The smallest and the largest ones are, respectively, 5N9H with 16 residues (244 atoms), and 3OAA with $13,028$ residues ($198,458$ atoms). As shown in Fig. S3a and as with **S** and **G**, both μ_s and μ_t decrease with n with the former faster than the latter while μ_p increases with n. The power laws for μ_s and μ_p are, respectively, $\mu_s = 10.9135n^{-0.4565}+1.2606$ and $\mu_p = -4.1148n^{-0.2491}+2.4396$. They intersect at $n_{sp} = 1,061$ with $\mu_{sp} = 1.714$. It equals $N_{sp} = 64$ residues using the linear equation $n_r = 0.0644n - 4.7048$ with a $R_{square} = 0.9992$. The R_{square} for μ_s is worse than those for both **G** and **S** (Fig. S3a). However, the differences in μs between **M** and **G** remain small (Table 1 and Figs. S3b and S3c).

3.1.2 The increases of n_s with n, and the linear equations for G, S and M

The linear equations between n and n_s for **G**, **S** and **M** are, respectively, $n_s = 0.3770n + 499.3710$ with a $R_{square} = 0.9771$ (Fig. S4a); $n_s = 0.3952n + 297.2236$ with a $R_{square} = 0.9586$; and $n_s = 0.4234n + 559.2468$ with a $R_{square} = 0.9667$. Their τs (slopes), $0.3770, 0.3952$ and 0.4234, are very close to each other.

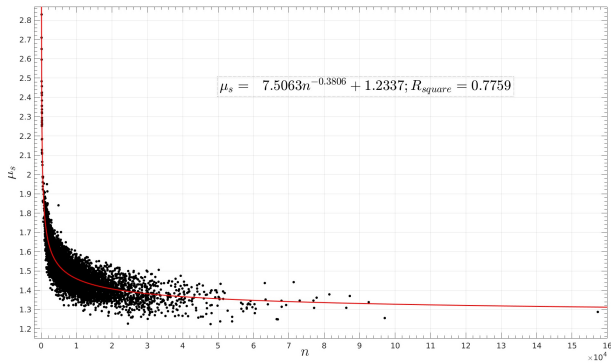
3.2 The geometric properties of the molecular surfaces of AlphaFold models

We first describe the changes of μs with n, the power laws and their intersections. We then describe the increases of n_s with n and the linear equations.

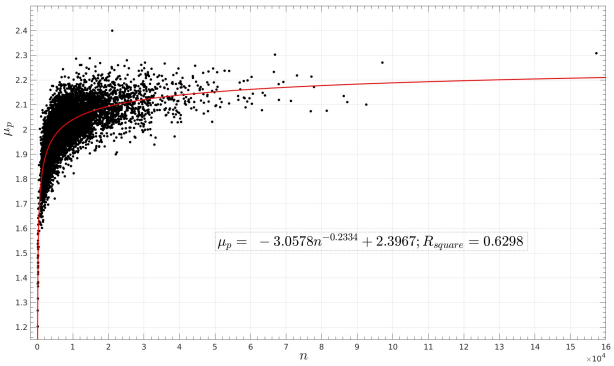
3.2.1 The changes of μs with n and the power laws and their intersections

The changes with n of the μs of the AlphaFold models for each proteome are fitted to a power law as described above. We illustrate the variations, the power laws and their intersections using the *E.coli* proteome as a representative of the thirteen prokaryotic proteomes and the human proteome as a representative of the fourteen eukaryotic ones. Please refer to Table S1 for the R_{square}s and points of intersection (N_{sp}s) for the twenty seven proteomes. In addition, Figs. 4 and 5 depict the power laws and their n_{sp}s, and Figs. S9–S35 depict the changes with n of their μs and n_ss.

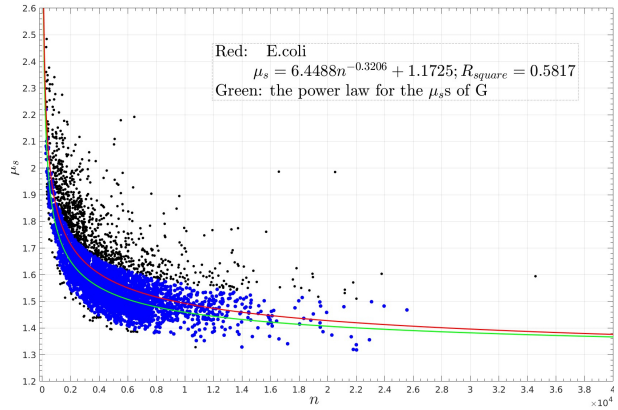
AlphaFold predicted $4,363$ models for *E.coli* proteome. The smallest and largest ones are, respectively, 257 atoms (16 residues), and $34,552$ atoms ($2,358$ residues). As with **S**, **G** and **M**, both μ_s and μ_t decrease with n with the former faster than the latter (Figs. 2c, 2d and S9) while μ_p increases with n. The power laws for μ_s and μ_p are, respectively, $\mu_s = 6.4488n^{-0.3261} + 1.1725$ and $\mu_p = -3.8938n^{-0.2619} + 2.3657$. They intersect at $n_{sp} = 1,393$ ($N_{sp} = 87$ residues (the legend of Fig. S1)). In addition, the power law for μ_p intersects with that for μ_t at $n = 42,897$, corresponding to $2,681$ residues, larger than the n_r for any *E.coli* protein.



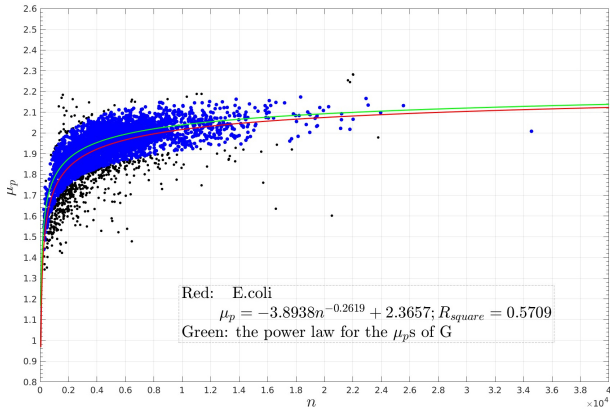
(a) The μ_ss of **G**



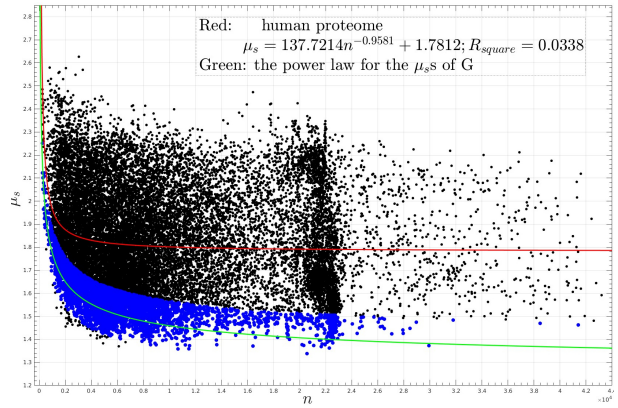
(b) The μ_ps of **G**



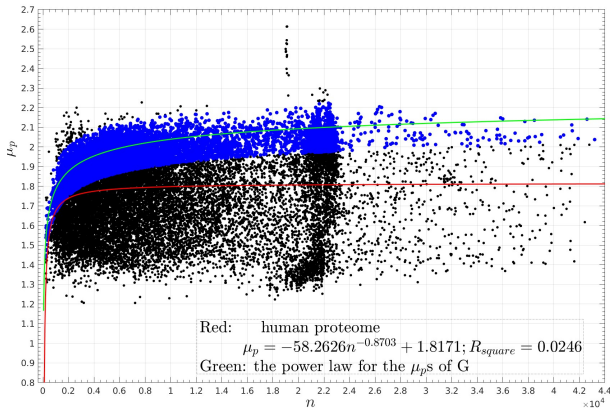
(c) The μ_ss of the *E.coli* proteome



(d) The μ_ps of the *E.coli* proteome



(e) The μ_ss of the human proteome



(f) The μ_ps of the human proteome

Figure 2: **The changes with n of the μ_ss and μ_ps for G, and the *E.coli* and human proteomes**. Each point represents either a μ_s (**a, c, e**) or a μ_p (**b, d, f**) value for a structure or a model. The power law curves are shown in red. In Figs. (**c, d, e, f**), the corresponding power law curves for **G** are shown in green. The insert in each figure lists the power law and R_{square}. About 24% of the models for the *E.coli* proteome have a $\Delta\mu_s > 2 \times \text{RMSE}_{G,s}$, where $\text{RMSE}_{G,s} = 0.0532$, and about 13% have a $\Delta\mu_p > 2 \times \text{RMSE}_{G,p}$, where $\text{RMSE}_{G,p} = 0.0631$ (Table 2). In stark contrast, about 80% of the models for the human proteome have a $\Delta\mu_s > 2 \times \text{RMSE}_{G,s}$, and about 61% have a $\Delta\mu_p > 2 \times \text{RMSE}_{G,p}$. In Figs. (**c, e**) the μ_ss with $\Delta\mu_s < 2 \times \text{RMSE}_{G,s}$, and in Figs. (**d, f**) the μ_ps with $\Delta\mu_p < 2 \times \text{RMSE}_{G,p}$ are colored in blue, and the rest in black. In Table 2, the percentages of the black data points over the total number of data points are listed, respectively, as P_s and P_p. Assuming a normal distribution, only 4.5% of the μ_ss for **G** have a $\Delta\mu_s > 2 \times \text{RMSE}_{G,s}$, and only 4.5% of the μ_ps for **G** have a $\Delta\mu_p > 2 \times \text{RMSE}_{G,p}$. The x-axes are n. The y-axes are, respectively, μ_s (**a, c, e**) and μ_p (**b, d, f**) in $\mathring{A}^2$.

AlphaFold predicted $23,391$ models for the human proteome. The smallest and largest ones are, respectively, 238 atoms (16 residues) and $43,190$ atoms ($2,699$ residues). As shown in Figs. 2e and 2f, in stark contrast to the μs for both the crystal structures (Figs. 1, 2a, 2b and S3) and the models for the thirteen prokaryotic proteomes (Figs. 2c, 2d and S9–S21), the μs for the fourteen eukaryotic proteomes (Figs. S22–S35) differ largely from both of them in the sense that when plotted with n as x-axis and μ as y-axis, the μs are broadly distributed and hardly change with n. The changes of the μs with n follow only minimally a power law with a $R_{square} = 0.0338$ for μ_s and $R_{square} = 0.0246$ for μ_p. The power laws for μ_s and μ_p intersect at $n_{sp} = 11,100$ ($N_{sp} = 694$ residues (the legend of Fig. S1)). However, the power law for μ_p *does not* intersect with that for μ_t.

3.2.2 The increases of n_s with n and the linear equations

The linear equations for the twenty seven proteomes are shown in Figs. 3, S5a and S5b with their τs listed in their inserts and Table S1. As in the last section, we illustrate the variations using the *E.coli* proteome as a representative of the thirteen prokaryotic proteomes and the human proteome as a representative of the fourteen eukaryotic proteomes.

The line for the *E.coli* proteome, $n_s = 0.4659n + 263.2263$, has a $\tau = 0.4659$, which is only 1.2-fold larger than that for **G** (0.3770, Figs. 3 and S4b). In contrast, the line for the human proteome (Figs. 3, S4c and S5b), $n_s = 0.6742n - 38.9199$, has a $\tau = 0.6742$, 1.8-fold larger than that for **G**. As shown in Fig. S4c, the increase with n of the n_ss for the human proteome looks artificial.

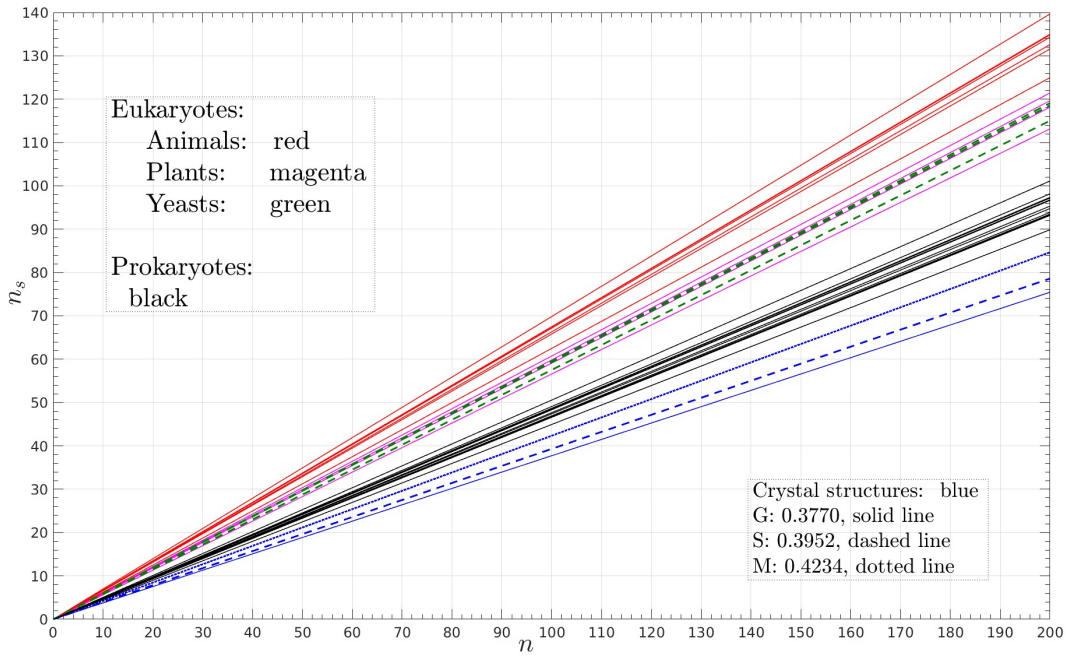


Figure 3: **The linear equations for the increase of n_s with n.** The lines for the fourteen eukaryotic proteomes are colored respectively in red (animals), magenta (plant) and yeast (green). The lines for the thirteen prokaryotic proteomes are colored in black, and those for the crystal structures in blue. The lines all pass through the origin to highlight their differences in τ. Table S1 and the inserts in Fig. S5 list their τ values. The τs for the crystal structures are all smaller than those for the prokaryotic proteomes and the latter are all smaller than those for eukaryotic proteomes. The x-axis is n, and the y-axis is n_s.

4 Discussion

We start with the previous analyses of experimentally-derived protein structures and their surfaces. Next we describe the meanings of the geometric properties μ_s, μ_p and τ and follow up with the discussion of the geometric properties of the SESs for both crystal structures and AlphaFold models. Specifically, we discuss first the geometric properties of the crystal structures and their implications for protein structure and folding, and then those of the AlphaFold models and

their implications for the quality and application of AlphaFold prediction. Finally we discuss the possible limitations of this study.

4.1 Previous analyses of protein structures and their surfaces

Previous analyses of experimental protein structures have focused primarily on their folds, secondary structures, binding or active sites and structural quality [28], while their surfaces have received much less attention. Out of the three surface types, previous analyses have largely been performed on SASA and at residue-level rather than on SAS (as a part of SES) or SES and at atom-level, even though it has been shown that SES could be better suited for quantifying the contribution of surface to protein structure and function [29], possibly due to the lack of robust programs for accurate SES computation. Specifically, except for SESA no robust and accurate programs are available for individual computation of the three SES patch types. Furthermore, to our best knowledge none of the three SES types have been studied *separately and at atom-level* for their possible relevance to protein structure, folding and function.

4.2 The geometric properties μ_s, μ_p and τ describe the packing of surface atoms

A SAS patch of surface atom i is a spherical polygon with up to a few tens of sides [13]. Each side corresponds to an intersection between i and a neighboring atom with each of them modeled as a sphere with a radius of $r_a + r_p$ where r_a and r_p are respectively atomic and probe radii. A toroidal patch exists when two neighboring atoms are close enough to form an impenetrable barrier to the probe, and a probe patch exists when a triple of them are close enough to form such a barrier. Thus the μs are averages over the spatial distribution of surface atoms in a molecule. Specifically, a large μ_s means that either n_s is large, i.e., many atoms are on surface, or a_s per surface atom is large (Eqs. (3, 4)). A large μ_p means that most of the surface atoms are close enough to form an impenetrable barrier to the probe. Since the vast majority of the surface atoms of a polypeptide come from its side chains, a large μ_s but a small μ_p indicates that the surface side chains are loosely packed with large space among them, i.e., most side chain atoms are largely exposed, a hallmark of the structure of a typical peptide. By contrast, a large μ_p but a small μ_s indicates that the surface side chains are tightly packed with the space among them largely impenetrable to the probe, a hallmark of a folded and spherical-shaped protein. Thus we could use the geometric property of SES, $\gamma = \frac{\mu_p}{\mu_s}$, to characterize the spatial distribution of surface atoms.

The three power laws and their points of intersection could be looked upon as an average over either a set of crystal structures or an AlphaFold proteome. The power law for μ_s intersects with that for μ_p when $\gamma = 1$ or $n = n_{sp}$, i.e., when the total SAS area of a molecule equals its total probe area. A SAS patch is convex while a probe one concave, thus the two power laws intersect when the total area of the patches with a negative curvature equals that with a positive curvature (if one ignores the toroidal patches, each of which includes regions of positive and negative curvatures). A typical peptide has a $\gamma < 1$ while a folded protein usually has $\gamma > 1$.

The τ for a set of crystal structures or a proteome describes, on average, how fast n_s increases with n. As shown in Fig. 3, the τ for **G** (0.3770), is very close to that for **S** (0.3952). Both of them are smaller than that for **M** (0.4243). A small τ together with a large γ is a feature of a set of well-packed crystal structures of soluble proteins. The simultaneous increase of n_s and decrease of a_s occur in the transmembrane region of a membrane protein where hydrogen atoms are more abundant than in the rest of the protein (Fig. S6).

4.3 The geometric properties of the molecular surfaces of crystal structures

We have analyzed the changes of A_s, A_t, A_p and A (Eq. 2) with n, n_r and n_s, and with secondary structure content and surface atom type. Among them we find that the changes with n of μs, particularly μ_s and μ_p, are most characteristic (Figs. 1 and 2) and thus likely most relevant to protein structure, folding and function. The μs for the crystal structures are fitted to n slightly better than to either n_s or n_r (Fig. S7), though there exist excellent linear relationships among them (Figs. S1 and S4a). For example, the number of atoms per residue, i.e., 16 atoms per residue, could be determined accurately from the linear equation between n_r and n for **G** (Fig. S1). The fitting to n is better than to n_r likely because of the existence of missing atoms in many crystal structures, and to n_s likely because of the diversity in surface shape among the structures. Furthermore, the μs change only slightly with secondary structure content and the composition of surface atom types (data not shown). In the following we focus our discussion on the changes with n of μ_s and μ_p, and N_{sp}.

4.3.1 The characteristic changes of the μs with n and their geometric and biophysical meanings

The characteristic changes of the μs with n could be (1) a global geometrical feature, (2) the optimization of solvent-protein interaction, and (3) the minimization of misinteraction.

4.3.1.1 The change of μs with n is unlikely a global geometric feature

SES is computed with the molecule modeled as a set of possibly intersecting spheres having different radii. With an increasing number of spheres, as a whole such a set is expected to become more and more similar to a large sphere with a small positive curvature. This is in contrast with the simultaneous increase of μ_p and decrease of μ_s with n, and the fact that $N_{sp} \leq 65$ for $\mathbf{S}$, $\mathbf{G}$ and $\mathbf{M}$. The power laws and the N_{sp} values show that with an increasing number of atoms (spheres) and particularly after the point of intersection, the total area of the regions of negative curvature becomes larger than that of positive curvature. In other words, the results show that both μ_s and μ_p are only local geometric properties of SES, and as such they should not change with molecular size.

4.3.1.2 The optimization of solvent-protein interaction

If the representation of solvent molecules as spheres is a decent approximation, the increase of μ_p with n means that the surface area where solvent molecules could interact more optimally with the protein increases with n. The solvent could be either the aqueous solvent for a water-soluble protein or both the lipids and aqueous solvent for a membrane protein. In other words, the increase in μ_p suggests that the disruption to solvent's intrinsic structure is reduced. For a water-soluble protein this means that the disruption to water hydrogen-bond network is reduced, the entropic increase of the whole system reduced, and the inter-molecular hydrogen bonding and VDW interaction optimized. For a membrane protein, the disruption to the structural integrity of the membrane is reduced.

4.3.1.3 The minimization of protein-protein misinteraction

A naturally-evolved protein must guarantee that the disruptions to and by all its neighbors are minimum in terms of free energy [30]. In other words, in a crowded intracellular environment, except for its cognate binding partners, a naturally-evolved protein should avoid binding to other molecules. A possible way to prevent noncognate binding is to have a molecular surface with a negative area-weighted average curvature. With a larger probe area, solvent-protein interaction becomes stronger, and that in turn makes it less likely the occurrence of intermolecular hydrophobic interactions between hydrophobic surface side chains, and that in turn reduces the chance for two proteins to stick together and for them to further form inclusion bodies inside a cell.

4.3.2 A hypothesis on protein structure and folding from a molecular surface perspective

Up to now, no satisfactory explanations have been given to the occurrence of the transition at about 50 residues from a peptide with a broad conformational distribution to a protein with a narrow one. A polypeptide chain with < 50 residues is typically called a peptide. It is not known whether the transition is determined by geometric properties, biophysical properties, solvent-protein interaction, or something else. It is suggested to be of geometric origin [31] based purely on theoretical calculation: the surface to volume ratio of a globular protein does not resemble that of a prolate ellipsoid until its chain length is ≥ 50 residues, and of biophysical origin since a polypeptide chain has to be long enough to enclose an inside in which hydrophobic residues could be sequestered. However, even for membrane proteins (set $\mathbf{M}$), μ_s decreases and μ_p increases with n, and their power laws intersect at 64 residues. Furthermore, if the sequester of hydrophobic residues is the sole driving force, then after the point of intersection, μ_s and μ_p should remain constant. Thus the sequester of hydrophobic residues could provide only a partial explanation.

The N_{sp}s for $\mathbf{G}$ and $\mathbf{S}$ are respectively 60 and 57, close to the number of residues required for a polypeptide to adopt a dominant conformation in solution. It suggests that a surface with a zero, or better, a negative average curvature is required for the folding of a polypeptide into a 3D structure with a dominant conformation in solution. In other words, the geometric properties of the SESs of the crystal structures provide a plausible explanation for the transition from peptide to protein. In turn, it suggests that $\gamma = \frac{\mu_p}{\mu_s}$ is an important parameter for characterizing the transition, a type of phase transition, from a form resembling liquid to a form resembling solid. We hypothesize that γ is closely related to the structural stability of proteins [32] and protein folding in the sense that $\gamma > 1$ is a necessary condition for protein folding.

4.4 The geometric properties of the molecular surfaces of AlphaFold models

The neural-network based AlphaFold [14, 15, 16, 17] has been hailed as a breakthrough in protein structure prediction based on several criteria for accuracy and structural quality [20, 21, 22, 23, 24, 28]. To objectively assess AlphaFold's accuracy, some models have been compared directly with crystallographic maps [18]. In addition, the topology of the models has been evaluated [33]. However, to our best knowledge, neither their SASs nor SESs have been analyzed.

The AlphaFold models studied here are for monomeric proteins containing entire protein sequence. For a fair comparison with them the crystal structures of non-membrane proteins have been separated into two sets $\mathbf{S}$ and $\mathbf{G}$ with the former including only monomeric proteins. In addition, the structures in $\mathbf{G}$ have no missing internal residues and an average of 0.218% missing atoms per structure. As shown in Table 1 and Figs. 4, 5 and S2, the power laws and τs for $\mathbf{S}$ and $\mathbf{G}$ are very close to each other. In addition, the power laws for the SESs of the set of the individual chains of the multimers in $\mathbf{G}$ are almost identical to those for $\mathbf{G}$ itself (section S8 and Fig. S8), and the τ for the set of chains is 0.4087, only 8.4% larger than that for $\mathbf{G}$ itself. In other words, though $\mathbf{G}$ consists mainly of multimers, it introduces only negligent errors when it is used to represent the crystal structures of monomeric proteins. Consequently, the comparisons with the AlphaFold models are presented with references to the geometric properties of the SESs of $\mathbf{G}$.

4.4.1 The similarity in SES between the crystal structures and AlphaFold models for prokaryotic proteomes

The changes with n of the μs of the AlphaFold models for the thirteen prokaryotic proteomes resemble those for the crystal structures (Tables 2 and S1, and Figs. 4, 5 and S9–S21). Among them the models of the *E.coli* proteome have geometric properties quite similar to those of the crystal structures (Figs. 2c, 2d and S9). In addition, the τs for the thirteen proteomes range from $0.4497 - 0.5057$ (Figs. 3 and S5a, and Table S1), all larger than any of the τs for the crystal structures, but smaller than any for the fourteen eukaryotic proteomes. This shows that though the SESs of the models for prokaryotic proteomes differ only slightly from those for crystal structures, their surface side chains are still packed less tightly than those of crystal structures.

The sensitivity of neural-network to training data [34] suggests that the resemblances result likely from (1) the abundance of sequenced prokaryotic genomes, (2) the similarity among prokaryotic genomes [35], (3) the high experimental coverage of at least one of the thirteen prokaryotic proteomes [35], (4) prokaryotic proteins are generally smaller in size and simpler in structure than their counterparts in eukaryotic proteomes, (5) the low content of disordered regions in prokaryotic proteomes [36], and (6) the low level of glycosylation during post-translation modification. In other words, the abundance of high quality training data somewhat compensates for the lack of surface features for neural-network model training.

4.4.2 The large differences in SES between the crystal structures and AlphaFold models for eukaryotic proteomes

As illustrated in Figs. 2e and 2f, and shown in Tables 2 and S1 and Figs. S22–S35, the μs of the AlphaFold models for each of the fourteen eukaryotic proteomes are broadly distributed, hardly change with n and only minimally follow a power law. Among them the μs for the human proteome (Figs. 2e and 2f) are the furthest from those for crystal structures. In addition, their τs range from $0.5752 - 0.6984$ (Figs. 3 and Table S1); all are larger than any of the τs for the thirteen prokaryotic proteomes, and much larger than any for the crystal structures. This means that for the same n, the models for the eukaryotic proteomes have more surface atoms per model than either the crystal structures or the models for prokaryotic proteomes. In other words, the surface side chains of these models are peptide-like. Among them the human proteome has the largest $N_{sp} = 694$, $P_s = 80.2\%$ and $P_p = 60.8\%$ (Table 2 and Figs. 2e and 2f), and the third largest $\tau = 0.6742$ (Table S1). The differences suggest that in terms of SES, either the models are accurate and consequently there exist genuine variations between them, or the models have errors. If the latter is true, it means that the current AlphaFold models for eukaryotic proteomes are not adequate for structure-based drug design in particular.

4.4.2.1 Possible reasons for the genuine differences

It is estimated that there exist 30% intrinsically-disordered regions in a eukaryotic proteome. The SES of a disordered region resembles the SES of a typical peptide with a $\gamma < 1.0$. However, as shown in Figs. 2e and 2f and Table 2,

Organism	$\Delta\mu_s$ (mean / std) ($\mathring{A}^2$)	$\Delta\mu_p$ (mean / std) ($\mathring{A}^2$)	$P_s(\%)$	$P_p(\%)$
Crystal structures				
S	**-0.0121 / 0.0535**	**0.0070 / 0.0557**	**5.5**	**2.5**
M	**-0.0335 / 0.0606**	**-0.0155 / 0.0680**	**12.2**	**6.4**
Prokaryotic proteomes				
E.coli	*0.0512/ 0.1011*	*-0.0357 / 0.0825*	*23.7*	*12.8*
haeIn	0.0505 / 0.0994	-0.0318 / 0.0811	21.2	11.7
helpy	0.0609 / 0.1119	-0.0409 / 0.0894	26.7	16.6
klePh	0.0557 / 0.1047	-0.0453 / 0.0873	25.1	16.5
metJa	***0.0310 / 0.0850***	***-0.0156 / 0.0724***	***16.4***	***8.6***
mycLe	0.0960 / 0.1362	-0.0645 / 0.1090	37.0	25.0
mycTu	0.0859 / 0.1391	-0.0752 / 0.1098	35.7	26.9
neiG1	0.0761 / 0.1276	-0.0511 / 0.1027	32.3	21.8
pseAr	0.0529 / 0.1117	-0.0521 / 0.0858	25.2	16.2
salTy	0.0569 / 0.1046	-0.0384 / 0.0846	25.7	14.0
shiDs	0.0604 / 0.1067	-0.0429 / 0.0883	27.0	16.7
staA8	0.0553 / 0.1120	-0.0236 / 0.0914	23.1	12.5
strR6	0.0549 / 0.0997	-0.0296 / 0.0837	24.4	13.7
Yeast proteomes				
schpo	0.2043 / 0.1976	-0.1220 / 0.1601	62.5	41.4
yeast	0.2295 / 0.2063	-0.1366 / 0.1647	66.2	46.1
canAl	0.2360 / 0.2129	-0.1348 / 0.1655	66.6	44.0
Plant proteomes				
araTh	0.2539 / 0.2087	-0.1511 / 0.1699	71.8	50.3
soyBn	0.2555 / 0.2112	-0.1508 / 0.1755	71.8	50.3
maize	0.2900 / 0.2140	-0.1844 / 0.1764	77.6	58.5
orySj	0.3075 / 0.2251	-0.1896 / 0.1854	77.9	60.9
Eukaryotic proteomes				
celEg	0.2203 / 0.2067	-0.1240 / 0.1637	64.0	41.6
dicDi	0.2556 / 0.2125	-0.1251 / 0.1808	70.7	46.0
danre	0.2809 / 0.2153	-0.1793 / 0.1858	76.3	53.2
drome	0.2883 / 0.2335	-0.1882 / 0.1946	73.0	54.0
mouse	0.2798 / 0.2290	-0.1917 / 0.1910	73.0	54.7
rat	0.2763 / 0.2272	-0.1896 / 0.1894	72.8	55.0
human	***0.3181 / 0.2287***	***-0.2219 / 0.2025***	***80.2***	***60.8***

Table 2: **The differences in SES between the** $350,974$ **AlphaFold models for twenty seven proteomes and crystal structures**. Column 1 lists organism names in abbreviation with their full names listed in Table S3. The differences in μ_s and μ_p of model i of a proteome from the μ_ss and μ_ps of $\mathbf{G}$ are quantified as $\Delta\mu_s(i) = \mu_s^\alpha(i) - \mu_s^g(i)$ and $\Delta\mu_p(i) = \mu_p^\alpha(i) - \mu_p^g(i))$ where $\mu_s^\alpha(i), \mu_p^\alpha(i)$ are the μ_s, μ_p values for model i, and $\mu_s^g(i), \mu_p^g(i)$ are computed using, respectively, Eqn. (6) and Eqn. (7) with n being the number of atoms of model i. Column 2 lists the mean and standard deviation of the list of the $\Delta\mu_s$s for a proteome, and column 3 lists the mean and standard deviation of the list of the $\Delta\mu_p$s for a proteome. Columns 4 and 5 list, respectively, the P_ss and P_ps for $\mathbf{S}$, $\mathbf{M}$ and the twenty seven proteomes. They have the same meanings as those in Table 1. Fig. 2 visualizes the P_ss and P_ps for the *E.coli* and human proteomes. The rows for $\mathbf{S}$ and $\mathbf{M}$ are repeated here for easy comparison. As shown in the last row, the μs, P_s and P_p for the human proteome are the furthest from those for $\mathbf{G}$.

80.2% of the models for the human proteome have a $\Delta\mu_s > 2 \times \text{RMSE}_{G,s}$. In **G** itself only 4.5% structures have a $\Delta\mu_s > 2 \times \text{RMSE}_{G,s}$, thus about 75.7% ($75.7 = 80.2 - 4.5$) of the AlphaFold models have a μ_s outside of the μ_s range[7] for **G**. In other words, the estimated 30% disordered regions is not large enough to account for the 75.7% difference in μ_s between the crystal structures and the models for the human proteome. Among them *Schizosaccharomyces pombe* proteome has the smallest $P_s = 62.5\%$, meaning that 58.0% of its models have a μ_s outside of the μ_s range for **G**. It also has the smallest $|\Delta\mu_s|$ and $|\Delta\mu_p|$ (Table 2).

A eukaryotic proteome includes about 30% membrane proteins. However, the power laws for the μs, N_{sp} and τ of **M** are all much closer to those for **G** than to any of the eukaryotic proteomes (Figs. 3, 4, 5 and S3).

The low-confidence regions of an AlphaFold model, as marked by a low pLDDT score, and often corresponding to disordered regions, are included in our calculation. If the differences are attributed solely to such regions, then it means that about 75.7% of the surface residues of the models for the human proteome have a low pLDDT score, which seems to disagree largely with the current quality assessment of AlphaFold models [17, 18]. It has been claimed that the high-confidence regions of an AlphaFold model are usually very close to experimental structures [15, 17].

4.4.2.2 The differences are due possibly to the lack of surface features in AlphaFold model training

The models are predicted using an AI model trained with the currently available crystal structures and other data. Thus we are puzzled by the up to 75.7% difference in SES between the models for eukaryotic proteomes and the crystal structures. At present we think that the large differences are due possibly to the lack of surface features in AI model training, which seems to affect the eukaryotic models much more than the prokaryotic models.

4.4.3 The large differences in SES between prokaryotic and eukaryotic proteomes

It is somewhat surprising that there exist large differences in SES between the models for prokaryotic and eukaryotic proteomes (Figs. 2-6, and Tables 2, S1 and S2) since they are all predicted using the same AI model. In the following we discuss two possible explanations: (1) the differences result from genuine structural variations between them, and (2) there exist errors in the models for eukaryotic proteomes.

4.4.3.1 The differences reflect the genuine variations between prokaryotic and eukaryotic proteomes

A eukaryotic proteome has a higher percentage of disordered regions than a prokaryotic one. It is estimated that on average a eukaryotic protein contains 32% disordered residues while a prokaryotic one only 12% [36]. However, the 20% difference is unlikely to be large enough to rationalize 53.2% difference[8] in μ_s between the models for the human and the *E.coli* proteomes (Fig. 6 and Table S2), even though it has been shown that the disordered regions of several AlphaFold models agree well with experimental results [24].

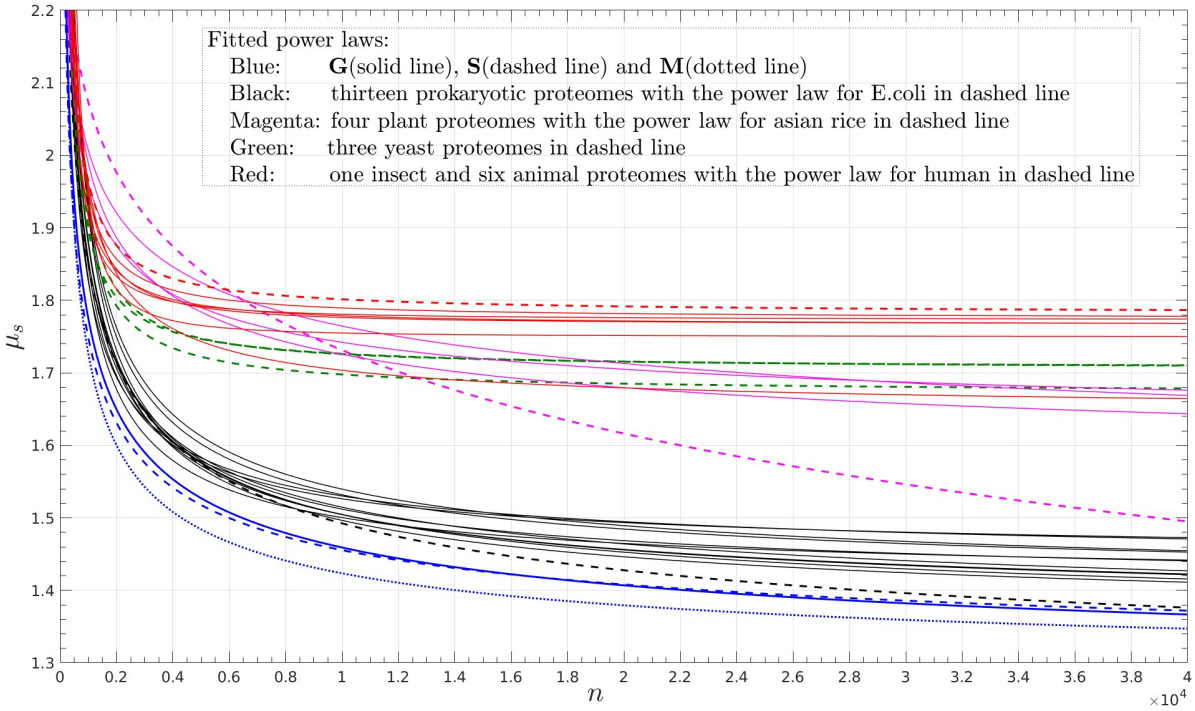
A typical eukaryotic protein has a high level glycosylation during past-translation modifications. However, glycosylation occurs in both prokaryotic and eukaryotic proteins. Though 50%−70% of eukaryotic proteins are possibly glycosylated, the percentage of glycosylated surface side chains of a eukaryotic protein is unknown but likely low. The μs are an average over a whole surface while the power laws are an average over a set of crystal structures or models. In addition, glycosylation has not been considered in AlphaFold2 and the vast majority of the structures in **G** are not glycosylated. So it is unlikely that glycosylation could account for the large P_ss (37.7%−57.7%) and P_ps (29.9%−48.5%) computed with the power laws for the *E.coli* proteome as references (Fig. 6 and Table S2).

4.4.3.2 The differences are due possibly to the lack of high quality training data for eukaryotic proteomes

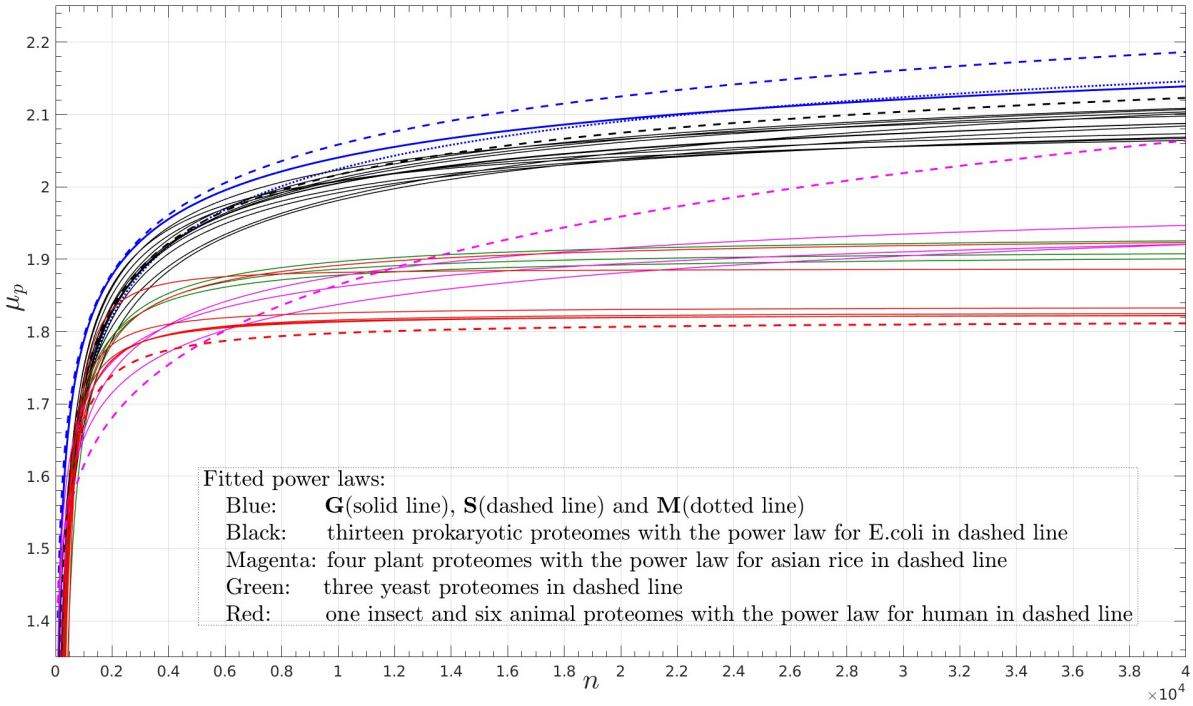
As discussed in the last section, the percentages of disordered residues and glycosylation are not large enough to account for the differences in SES between the prokaryotic and eukaryotic proteomes. At present we think that the differences are attributed at least partially to the lack of high quality training data for eukaryotic proteomes.

[7] The μ_s range here means the range of the μ_ss with approximately the same n.

[8] The 53.2% difference is calculated as follows: $53.2 = 57.7 - 4.5$ where 57.7 is the P_s value for the human proteome (Table S2). It means that 53.2% of the μ_ss for the same n for the human proteome are out of the μ_s range for the *E.coli* proteome.



(a) The power laws for the changes of μ_s with n



(b) The power laws for the changes of μ_p with n

Figure 4: **The power laws for $\mathbf{G}, \mathbf{S}, \mathbf{M}$ and the twenty seven proteomes**. Figs. (**a**) and (**b**) show, respectively, the power law curves for μ_s and μ_p. From bottom up, the three blue curves in (**a**) are, respectively, the power laws for **M**, **S** and **G**. Right above them are the power law curves for the thirteen prokaryotic proteomes colored in black. Above them are the power law curves for the fourteen eukaryotic proteomes, each of them is well separated from the curves for both the crystal structures and prokaryotic proteomes. The thirty power law curves for μ_p (Fig. (**b**) are similarly distributed but in a top down manner. Please see the inserts for coloring scheme. The x-axes are n, and the y-axes are μ_s (**a**) and μ_p (**b**) in $\mathring{A}^2$.

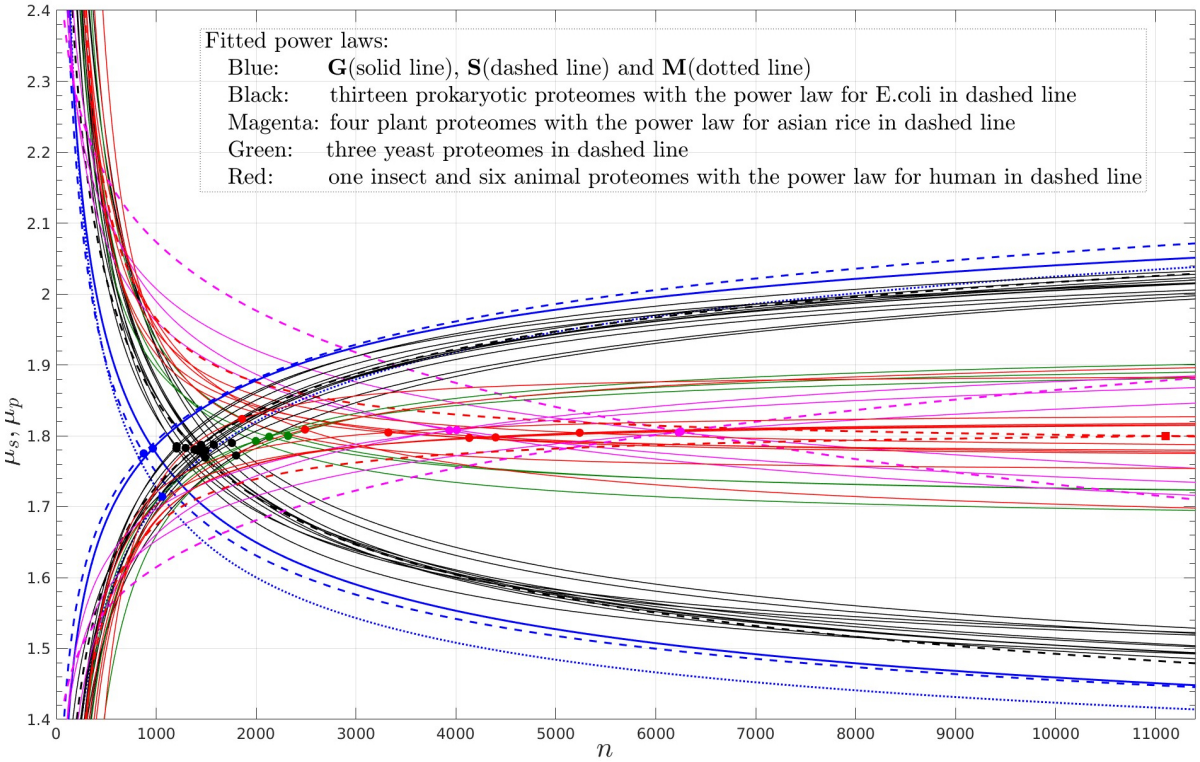
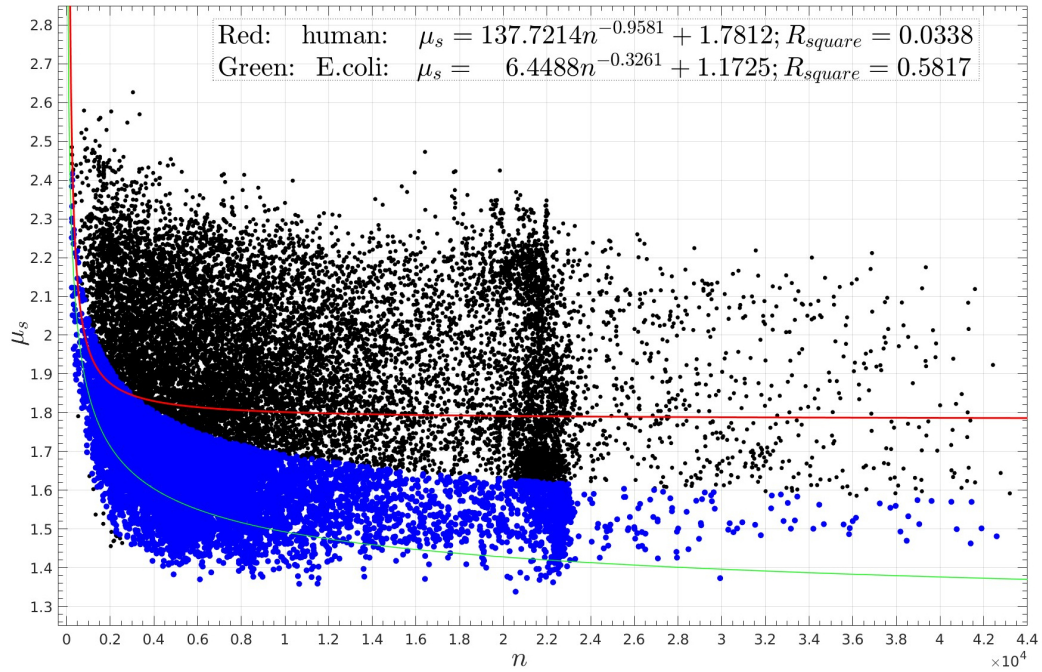
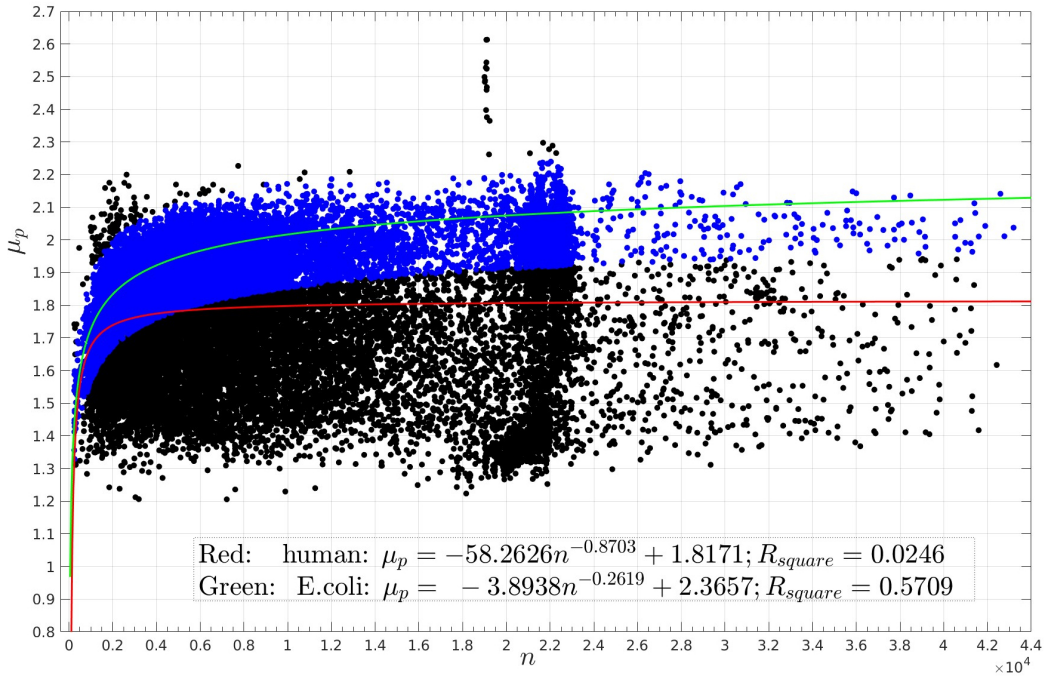


Figure 5: **The n_{sp}s for $\mathbf{G}, \mathbf{S}, \mathbf{M}$ and the twenty seven proteomes**. Except for the human proteome whose n_{sp} is depicted as a square, the rest are depicted as a filled circle. The n_{sp}s for the crystal structures are all smaller than any of those for the twenty seven proteomes, and the n_{sp}s for the prokaryotic proteomes are all smaller than any for the eukaryotic proteomes (Table S1). **S** has the smallest $n_{sp} = 875$ while the human proteome the largest $n_{sp} = 11,000$. The x-axis is n and the y-axis is μ_s, μ_p in $\mathring{A}^2$.



(a) The change of μ_s with n and the power law



(b) The change of μ_p with n and the power law

Figure 6: **The changes with n of the μs for the human proteome and the comparisons with those for the *E.coli* proteome**. Figs. (**a**) and (**b**) depict, respectively, the μ_ss and μ_ps for the human proteome. The power law curves for the human and *E.coli* proteomes are colored respectively in red and green. In Figs. (**a**) the μ_ss with $\Delta\mu_s < 2 \times \text{RMSE}_{E,s}$, where $\text{RMSE}_{E,s} = 0.1002$ is the RMSE of fitting the μ_ss of the *E.coli* proteome to a power law, and in Figs. (**b**) the μ_ps with $\Delta\mu_p < 2 \times \text{RMSE}_{E,p}$, where $\text{RMSE}_{E,p} = 0.0816$ is the RMSE of fitting the μ_ps of the *E.coli* proteome to a power law, are colored in blue while the rest in black. In Table S2, the percentages of the black data points over the total number of data points are listed, respectively, as P_s and P_p. The x-axes are n, and the y-axes are μ_s (**a**) and μ_p (**b**) in $\mathring{A}^2$.

4.5 Possible limitations

SES is computed by modeling both the probe and protein atoms as hard spheres. This means that SES itself is only an approximation to the physical surface of a protein. Consequently, both the power laws and their intersections are only an approximation to a likely more complicated physical relationship between protein surface and molecular size.

The surface of a soluble protein is inherently flexible in solution [2] while crystal structure is static in nature. However, if one assumes that a crystal structure freezes one conformation out of an ensemble of conformations in solution, then by ergodic hypothesis the SES geometric properties described in this paper should have real physical meanings since they all are averages over a set of structures or models with each of the individual properties being an average over a whole surface. In addition, though we have not quantified the contribution to the SES geometric properties of the crystal packing for the structures in $\mathbf{G}$ and $\mathbf{M}$, such contributions, if exist, should be minor (section S8 and Fig. S36).

References

[1] Lincong Wang, Y. Li, and H. Yan. Human cellular retinoic acid binding proteins: Quantitative analysis of the ligand binding properties of the wild-type and site-directed mutants. *J. Biol. Chem.*, 272:1541–1547, 1997.

[2] Lincong Wang, Y. Pang, T. Holder, J. Brender, A. V. Kurochkin, and E. R. P. Zuiderweg. Functional dynamics in the active site of the ribonuclease binase. *Proc. Natl. Acad. Sci. USA*, 98:7684–7689, 2001.

[3] C. Tanford. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *Journal of the American Chemical Society*, 84(20):4240–4247, 1962.

[4] D. Eisenberg and A. D. Andrew D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.

[5] W. Kauzmann. Thermodynamics of unfolding. *Nature*, 325:763–764, February 1987.

[6] T. Y. Lin and S. N. Timasheff. On the role of surface tension in the stabilization of globular proteins. *Protein science*, 5(2):372–381, 1996.

[7] Y. Levy and J. N. Onuchic. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.*, 35:389–415, 2006.

[8] M. Bellissent-Funel, A. Hassanali, M. Havenith, R. Henchman, P. Pohl, F. Sterpone, D. van der Spoel, Y. Xu, and A. E Garcia. Water determines the structure and dynamics of proteins. *Chemical Reviews*, 116(13):7673–7697, 2016. PMID: 27186992.

[9] W. Blokzijl and Jan B. F. N. Engberts. Hydrophobic effects. opinions and facts. *Angewandte Chemie International Edition in English*, 32(11):1545–1579, 1993.

[10] R. B. Hermann. Theory of hydrophobic bonding. II. correlation of hydrocarbon solubility in water with solvent cavity surface area. *The Journal of Physical Chemistry*, 76(19):2754–2759, 1972.

[11] F. M. Richards. Areas, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering*, 6(1):151–176, 1977. PMID: 326146.

[12] J. Greer and B. L. Bush. Macromolecular shape and surface maps by solvent exclusion. *Proc. Natl. Acad. Sci. USA*, 75(1):303–307, 1978.

[13] Lincong Wang. SESA: A program for the analytic computation of solvent-excluded surface areas. *ChemistryOpen*, n/a(n/a):e202400172.

[14] A. W. Senior, R. Evans, J Jumper, and et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577, 2020.

[15] J. Jumper, R. Evans, A. Pritzel, and et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596, 2021.

[16] J. Abramson, J. Adler, J. Dunger, and et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630, 2024.

[17] K. Tunyasuvunakool, J. Adler, Z. Wu, and et al. Highly accurate protein structure prediction for the human proteome. *Nature*, 596, 2021.

[18] T.C. Terwilliger, D. Liebschner, T.I. Croll, and et al. Alphafold predictions are valuable hypotheses and accelerate but do not replace experimental structure determination. *Nat Methods*, 21, 2024.

[19] A. Kryshtafovych, T. Schwede, M. Topf, K. Fidelis, and J. Moult. Critical assessment of methods of protein structure prediction (casp)—round xiv. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617, 2021.

[20] M. Akdel, D.E.V. Pires, E.P. Pardo, and et al. A structural biology community assessment of alphafold2 applications. *Nat Struct Mol Biol*, 29, 2022.

[21] C. Millán, R. M. Keegan, J. Pereira, M. D. Sammito, A. J. Simpkin, A. J. McCoy, A. N. Lupas, M. D. Hartmann, D. J. Rigden, and R. J. Read. Assessing the utility of casp14 models for molecular replacement. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1752–1769, 2021.

[22] G. Grzegorz. Sequence-assignment validation in cryo-EM models with *checkMySequence*. *Acta Crystallographica Section D*, 78(7):806–816, Jul 2022.

[23] R. Tejero, Y. Huang, T. A. Ramelot, and G. T. Montelione. Alphafold models of small proteins rival the accuracy of solution nmr structures. *Frontiers in Molecular Biosciences*, 9, 2022.

[24] T. R. Alderson, I. Pritisanac, D. Kolaric, A. M. Moses, and J. D. Forman-Kay. Systematic identification of conditionally folded intrinsically disordered regions by alphafold2. *Proceedings of the National Academy of Sciences*, 120(44):e2304302120, 2023.

[25] M. Varadi, D. Bertoni, P. Magana, and et al. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, 11 2023.

[26] Lincong Wang, Y. Zhang, and S. Zou. The characterization of pc-polylines representing protein backbones. *Proteins: Structure, Function, and Bioinformatics*, 88(2):307–318, 2020.

[27] J. M. Word, S. M. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1. *Journal of Molecular Biology*, 285(4):1735–1747, 1999.

[28] A. Bhattacharya, R. Tejero, and G. T. Montelione. Evaluating protein structures determined by structural genomics consortia. *Proteins: Structure, Function, and Bioinformatics*, 66(4):778–795, 2007.

[29] K. A. Sharp, A. Nicholls, R. F. Fine, and B. Honig. Reconciling the magnitude of the microscopic and macroscopic hydrophobic effects. *Science*, 252(5002):106–109, 1991.

[30] J. Echave and C. O. Wilke. Biophysical models of protein evolution: Understanding the patterns of evolutionary sequence divergence. *Annual Review of Biophysics*, 46(Volume 46, 2017):85–103, 2017.

[31] G. Rose and D. Wetlaufer. The number of turns in globular proteins. *Nature*, 268:769–770, 1977.

[32] A. D. Robertson and K. P. Murphy. Protein structure and the energetics of protein stability. *Chemical Reviews*, 97(5):1251–1268, 1997. PMID: 11851450.

[33] P. Dabrowski-Tumanski and A. Stasiak. Alphafold blindness to topological barriers affects its ability to correctly predict proteins' topology. *Molecules*, 28(22), 2023.

[34] Lincong Wang, C. Cao, and S. Zuo. Protein secondary structure assignment using pc-polyline and convolutional neural network. *Proteins: Structure, Function, and Bioinformatics*, 89(8):1017–1029, 2021.

[35] E. V. Koonin and Y. I. Wolf. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*, 36(21):6688–6719, 10 2008.

[36] W. Basile, M. Salvatore, C. Bassot, and A. Elofsson. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLOS Computational Biology*, 15(7):1–20, 07 2019.

Supporting Information

Section S1 describes the increase of the number of residues n_r with the number of atoms n for $\mathbf{G}$, and the best-fitted linear equation between n_r and n. Section S2 presents the differences in μs between $\mathbf{S}$ and $\mathbf{G}$. Section S3 describes the changes of the μs with n, the power laws and point of intersection for $\mathbf{M}$. Section S4 describes the increases with n of the n_ss for $\mathbf{G}$, the *E.coli* and human proteomes. Section S5 presents the τs for the crystal structures and AlphaFold proteomes. Section S6 describes the SES of the transmembrane region of a membrane protein. Section S7 presents the changes of the μs with both n_r and the number of surface atoms n_s for $\mathbf{G}$. Section S8 describes the SESs of the individual chains of the multimers in $\mathbf{G}$. Section S9 presents the R_{square}s, N_{sp}s and τs for the crystal structures and AlphaFold models. Section S10 presents the differences in SES between the *E.coli* proteome and the other twenty six proteomes. Section S11 describes the changes with n of the μs and n_ss of the $350,974$ AlphaFold models for twenty seven proteomes. Section S12 lists the biological names of the twenty seven species and their abbreviations. Section S13 describes an example of a protein-protein interface formed by crystal packing.

S1: The increase of n_r with n for G

As shown in Fig. S1, the increase of n_r with n for $\mathbf{G}$ could be fitted to a linear equation with a MATLAB $R_{square} = 0.9982$. Because some structures in $\mathbf{G}$ have missing atoms, it is more accurate to use this equation to compute n_r from n than to use directly the fitted power laws for the variations of the μs with n_r (Fig. S6).

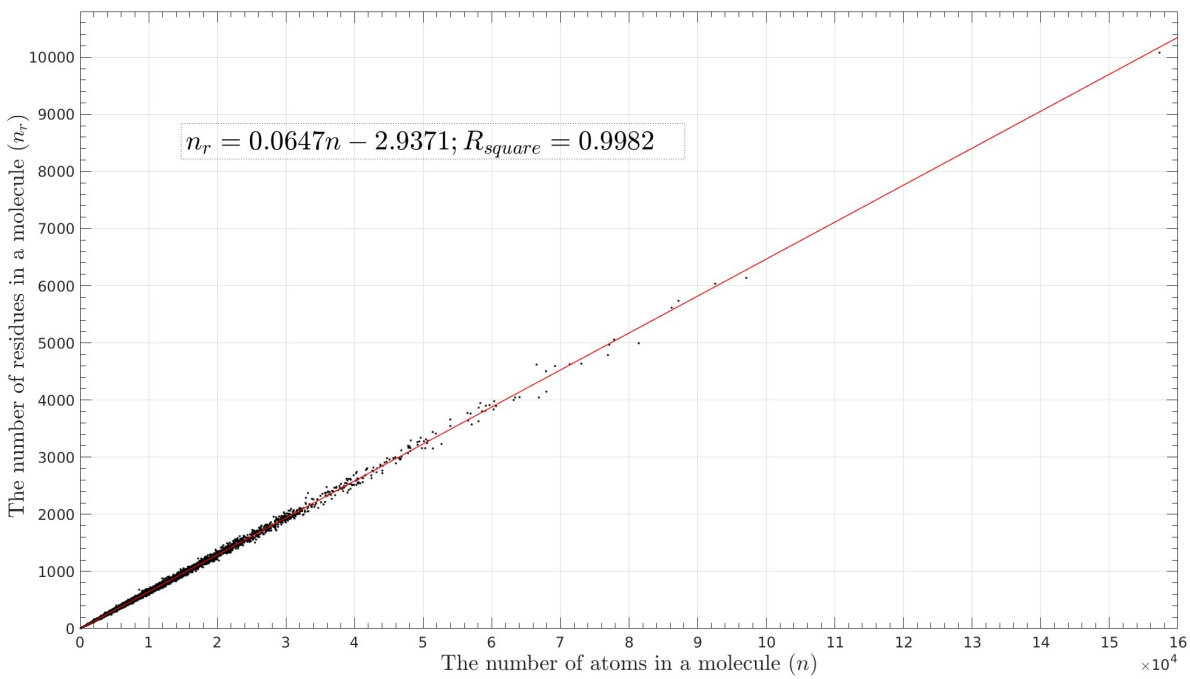
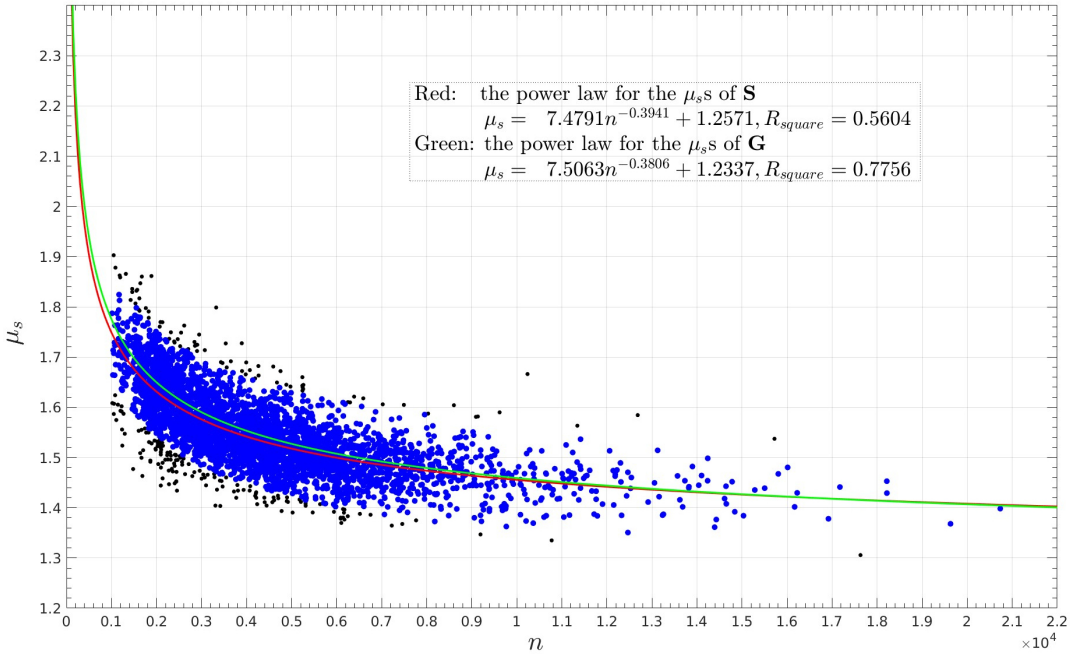


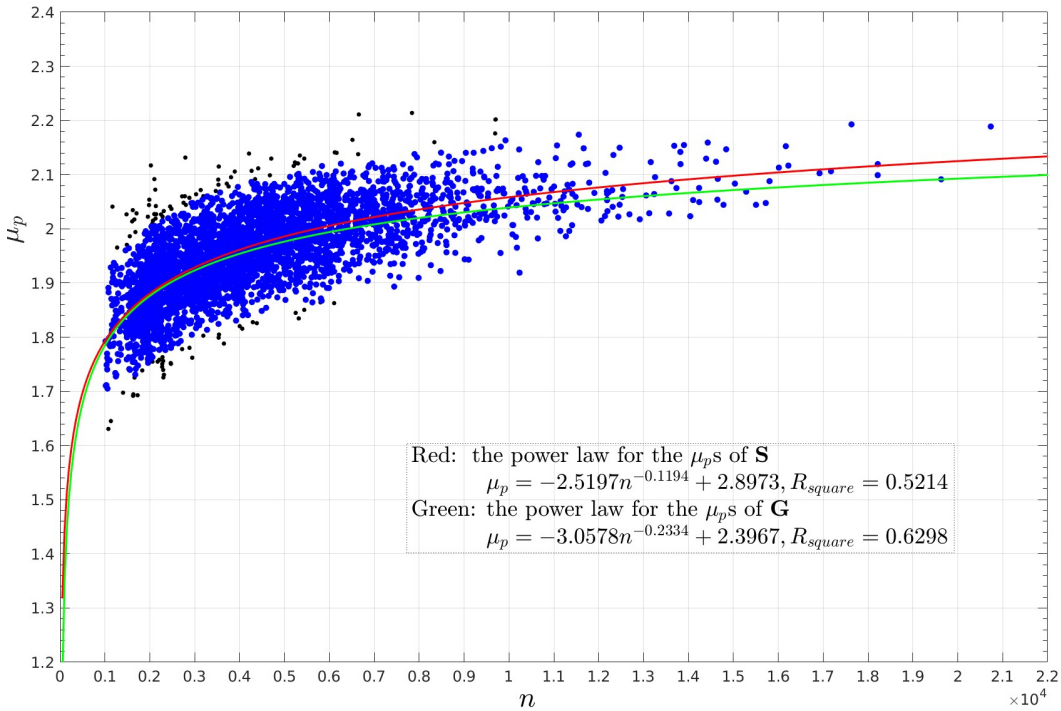
Figure S1: **The increase of n_r with n for G**. The insert is the MATLAB fitted linear equation, $n_r = 0.0647n - 2.9371$, with a $R_{square} = 0.9982$. As described in the main text, the equation is used to compute N_{sp} from n_{sp}. The number of atoms per residue, $\mathbf{16}$, estimated from its slope 0.0647, is used to compute N_{sp} from n_{sp} for the twenty seven AlphaFold proteomes (Table S1). The x-axis is n, and y-axis n_s.

S2: The differences in μs between S and G

As shown in Fig. S2, the variations with n of the μ_ss and μ_ps of **S** and **G**, and their power laws are very close in the sense that $< 6\%$ of the μ_ss of **S** has a $\Delta\mu_s > 2 \times \text{RMSE}_{G,s}$, and $< 3\%$ of the μ_ps of **S** has a $\Delta\mu_p > 2 \times \text{RMSE}_{G,p}$.



(a) The change of μ_s with n and the power law

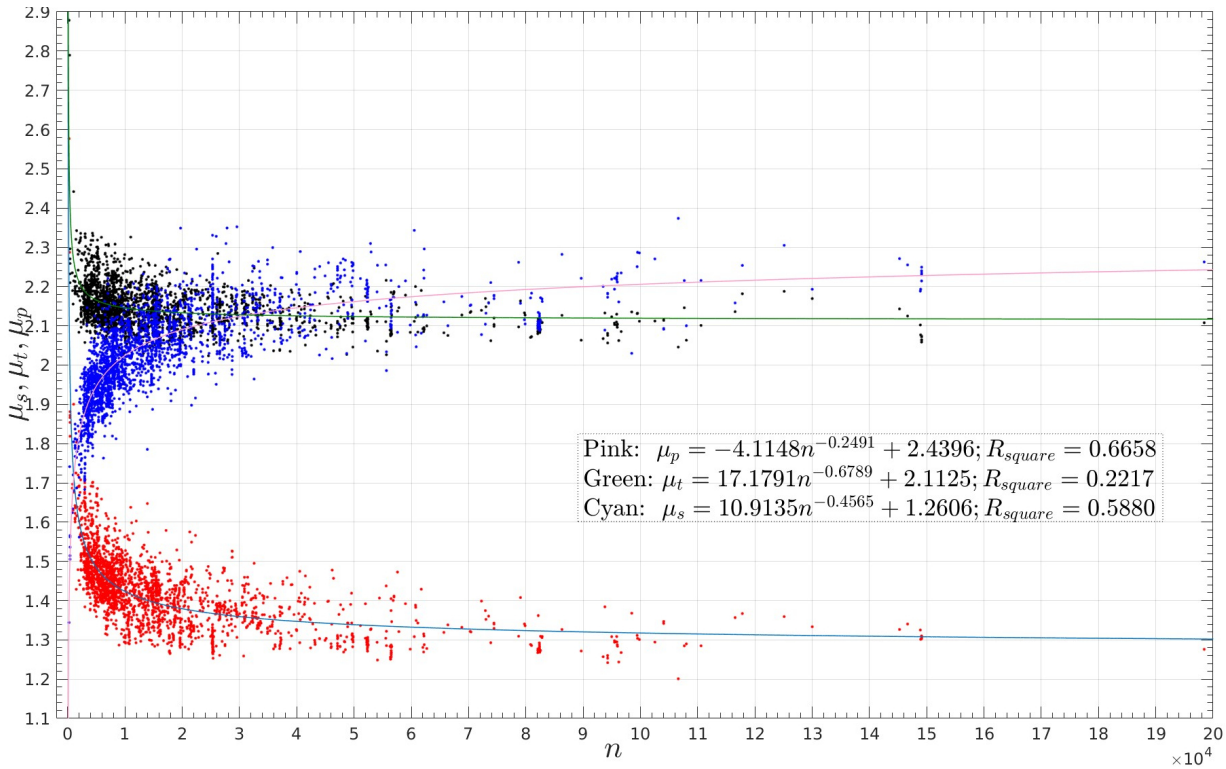


(b) The change of μ_p with n and the power law

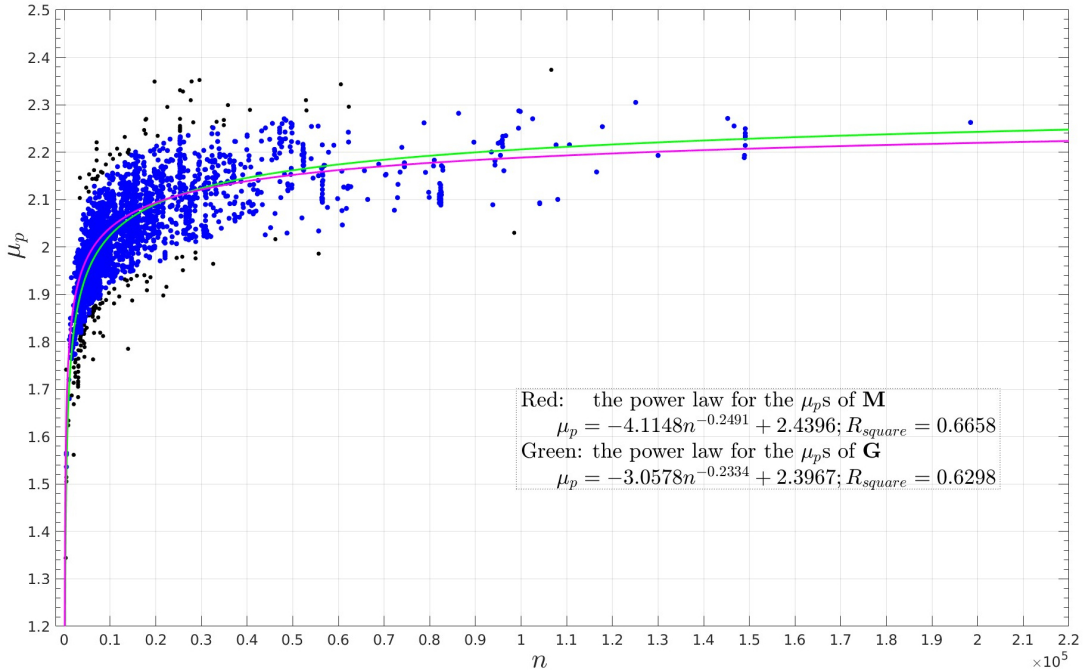
Figure S2: **The changes with n of the μs for S and the comparisons with those for G**. Figs. (**a**) and (**b**) depict, respectively, the μ_ss and μ_ps of **S** and the power law curves. In Fig. (**a**) the μ_ss with $\Delta\mu_s < 2 \times \text{RMSE}_{G,s}$, where $\text{RMSE}_{G,s} = 0.0532$ is the RMSE of fitting the μ_ss for **G** to a power law, and in Fig. (**b**) the μ_ps with $\Delta\mu_p < 2 \times \text{RMSE}_{G,p}$, where $\text{RMSE}_{G,p} = 0.0631$ is the RMSE of fitting the μ_ps for **G** to a power law, are colored in blue and the rest in black. There are only 229 structures in **S** with a $\Delta\mu_s > 2 \times \text{RMSE}_{G,s}$, and 103 with a $\Delta\mu_p > 2 \times \text{RMSE}_{G,p}$ (Table 1 of the main text). The inserts list the power laws, R_{square}s and coloring schemes. The x-axes are n, and the y-axes are μ_s (**a**) and μ_p (**b**) in $\mathring{A}^2$.

S3: The changes of the μs with n, the power laws and point of intersection for $\mathbf{M}$

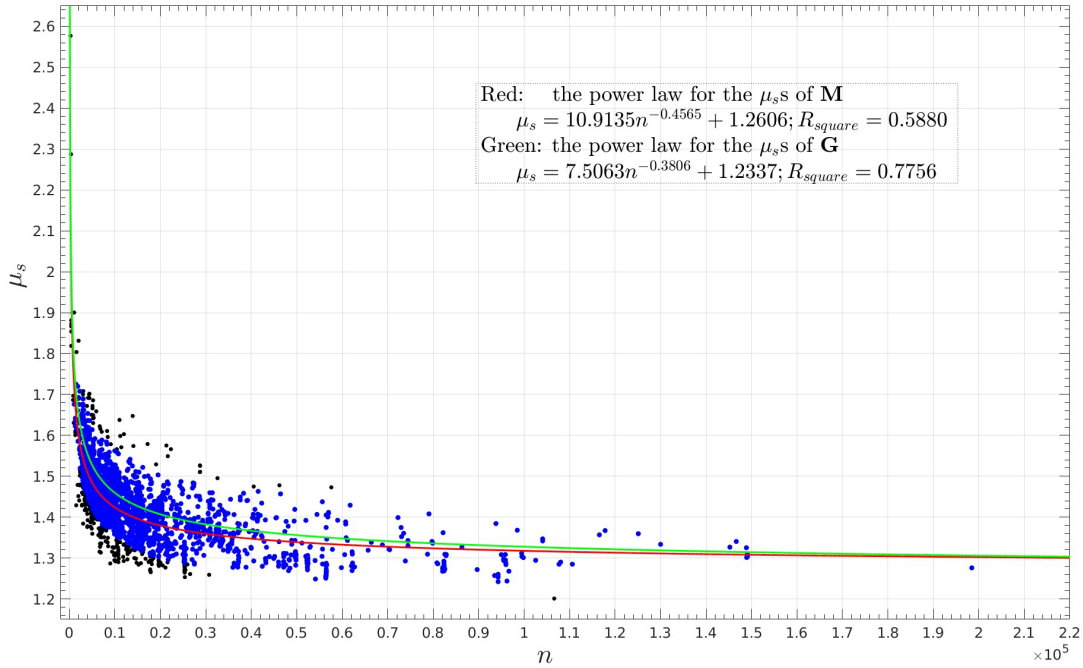
As shown in Fig. S3, the changes with n of the μs for $\mathbf{M}$ and the power laws are close to those for $\mathbf{G}$.



(a) The changes of the μ_ss with n



(b) The comparison of the μ_s of $\mathbf{M}$ with that of $\mathbf{G}$

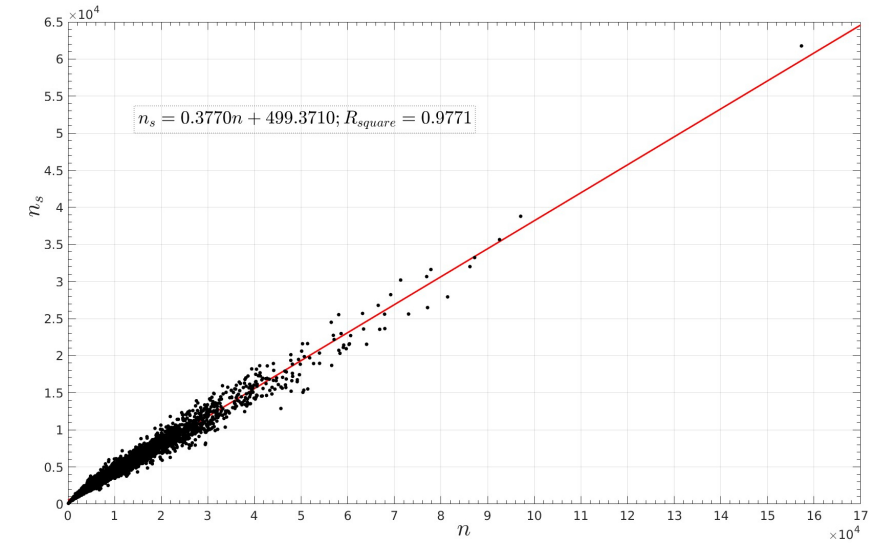


(c) The comparison of the μ_p of $\mathbf{M}$ with that of $\mathbf{G}$

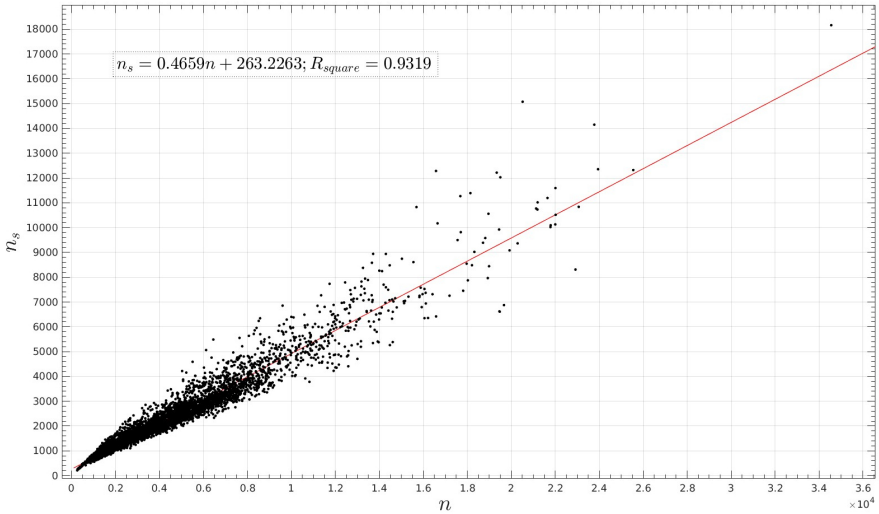
Figure S3: **The changes with n of the μs for $\mathbf{M}$ and the comparisons with those for $\mathbf{G}$**. Fig. (**a**) shows the μ_ss, μ_ts and μ_ps of $\mathbf{M}$ colored, respectively, in red, black and blue with the power law curves colored, respectively, in cyan, green and pink. Figs. (**b**) and (**c**) depict, respectively, the μ_ss and μ_ps of $\mathbf{M}$ and the power law curves. In Fig. (**b**) the μ_ss with $\Delta\mu_s < 2 \times \text{RMSE}_{G,s}$, where $\text{RMSE}_{G,s} = 0.0532$ is the RMSE of fitting the μ_ss for $\mathbf{G}$ to a power law, and in Fig. (**c**) the μ_ps with $\Delta\mu_p < 2 \times \text{RMSE}_{G,p}$, where $\text{RMSE}_{G,p} = 0.0631$ is the RMSE of fitting the μ_ps for $\mathbf{G}$ to a power law, are colored in blue and the rest in black. There are only 303 structures in $\mathbf{S}$ with a $\Delta\mu_s > 2 \times \text{RMSE}_{G,s}$, and 159 with a $\Delta\mu_p > 2 \times \text{RMSE}_{G,p}$ (Table 1 of the main text). The inserts list the power laws, R_{square}s and coloring schemes. The x-axes are n, and y-axes are μ_s, μ_t and μ_p (**a**), μ_s (**b**) and μ_p (**c**) in $\mathring{A}^2$.

S4: The increases with n of the n_ss for G, the *E.coli* and human proteomes

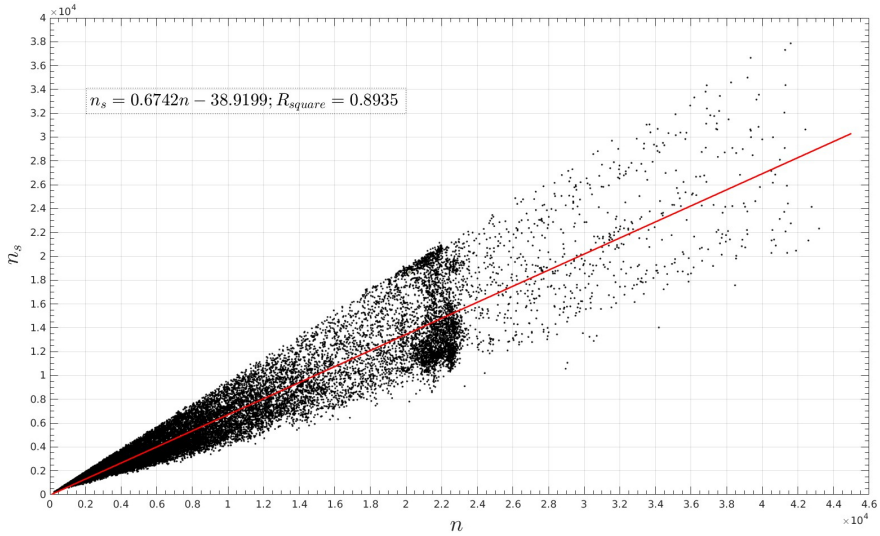
For **G**, the number of surface atoms n_s increases almost linearly with molecular size n with a $\tau = 0.3770$ (Fig. S4a). The τs for the *E.coli* proteome and human proteome are, respectively, 20.6% and 78.8% larger than that for **G** (Figs. S4b and S4c). Strangely, as shown in Fig. S4c and Figs. S22d–S35d, the changes with n of the n_ss for the fourteen eukaryotic proteomes look artificial with the range of n_s for approximately the same n also increases linearly with n.



(a) n_s of **G**



(b) n_s of the *E.coli* proteome

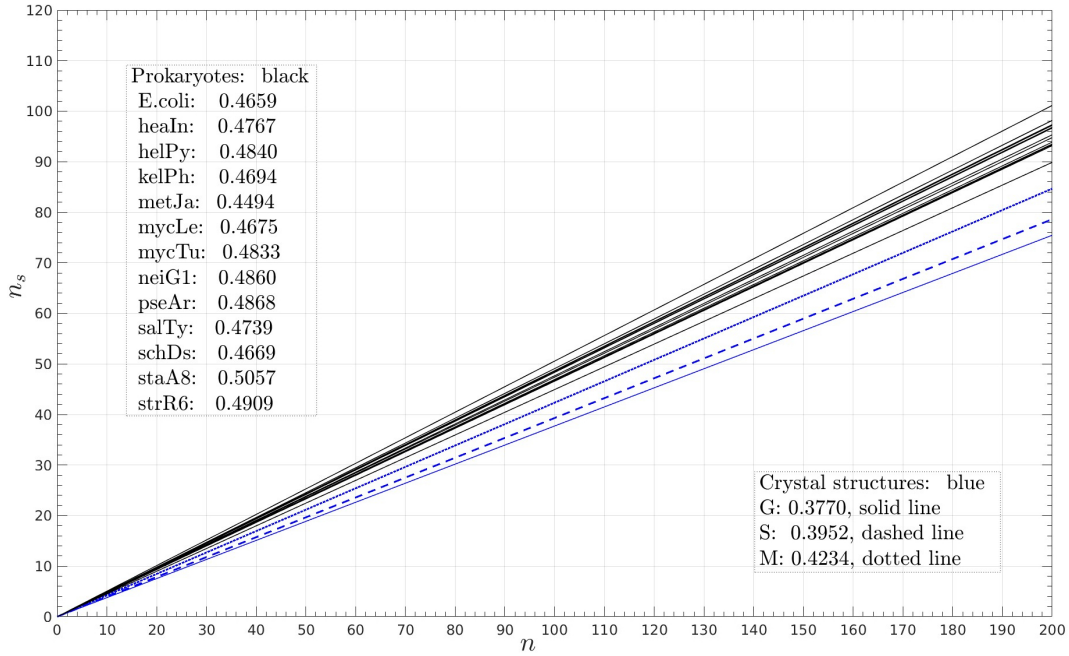


(c) n_s of the human proteome

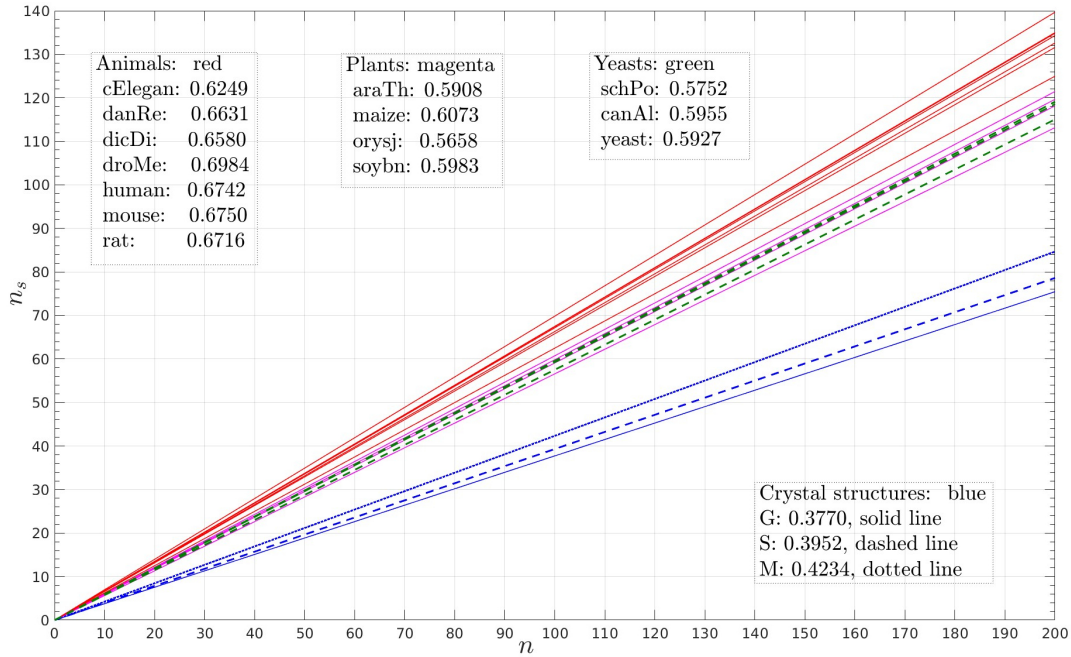
Figure S4: **The increases of the n_ss with n for G, and the *E.coli* and human proteomes**. The fitted lines are depicted in red. The inserts list the linear equations and their R_{square}s. As shown in (**c**) the increase of n_s with n looks artificial. The x-axes are n, and y-axes n_s.

S5: The τs for the crystal structures and AlphaFold proteomes

As shown in Fig. S5, the τs for the three sets of crystal structures are all smaller than any of the τs for the twenty seven proteomes.



(a) **G**



(b) **S**

Figure S5: **The τs for the crystal structures and AlphaFold proteomes**. Figs. (**a**) and (**b**) depict respectively the τs for the thirteen prokaryotic proteomes and fourteen eukaryotic ones. The τs for $\mathbf{G}$, $\mathbf{S}$ and $\mathbf{M}$ are shown in both (**a**) and b(**b**) for comparison. The inserts list the τ values for $\mathbf{G}$, $\mathbf{S}$ and $\mathbf{M}$ and the twenty seven proteomes. Please refers to Table S3 for their biological full names.

S6: The SES of the transmembrane region of a membrane protein

As shown in Fig. S6, the transmembrane region of a typical membrane protein is dominated by carbon and proton atoms; and the atoms with a partial charge differing largely from zero, if there, generally having a small SAS area.

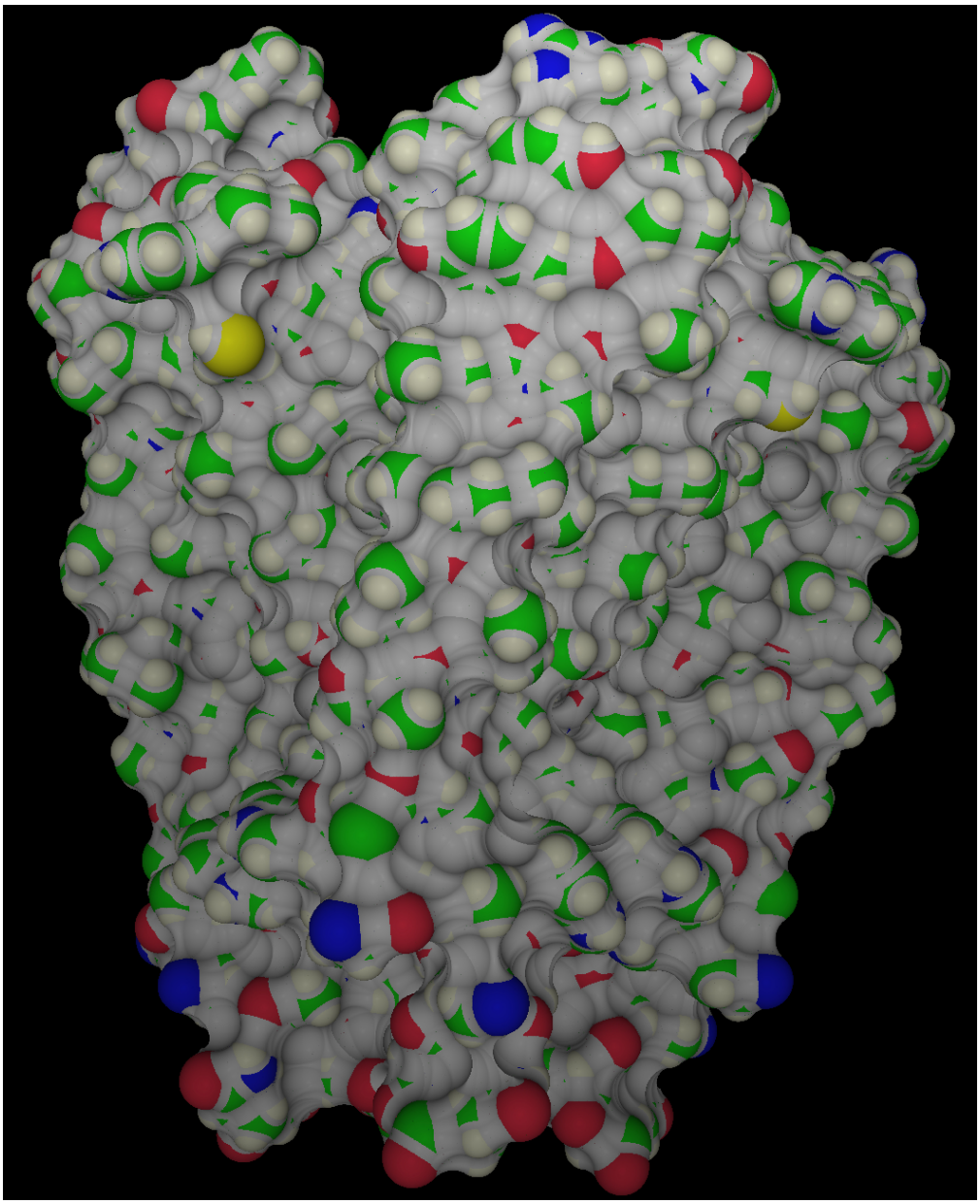
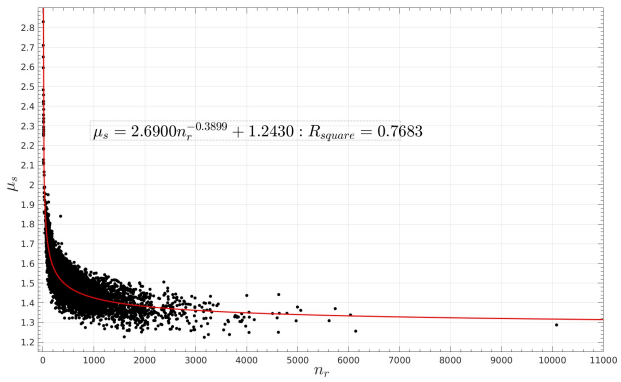


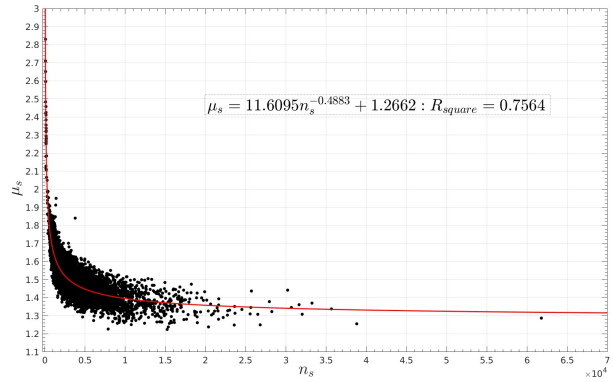
Figure S6: **The SES of the transmembrane region of a membrane protein (1BL8)**. The SAA polygons for the C, N, O and H atoms of a potassium channel are colored respectively in green, blue, red and misty rose while the rest of the SES in gray. The figure is prepared by our program SESA which is freely available at https://github.com/wlincong/SESexact.

S7: The changes with n_r and with n_s of the μs of G

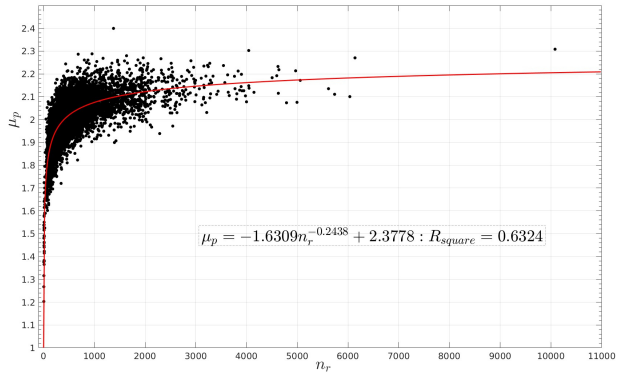
Fig. S7 depicts the changes of μ_s, μ_t and μ_p with n_r and n_s. As shown in the figure, compared with the R_{square}s for fitting to a power law, the changes of μs with n, the R_{square}s for fitting the changes of μs with n_r and n_s are all smaller, except for the change of μ_t with n_s, which has a R_{square} slightly larger than that for fitting to a power law the change of μ_t with n.



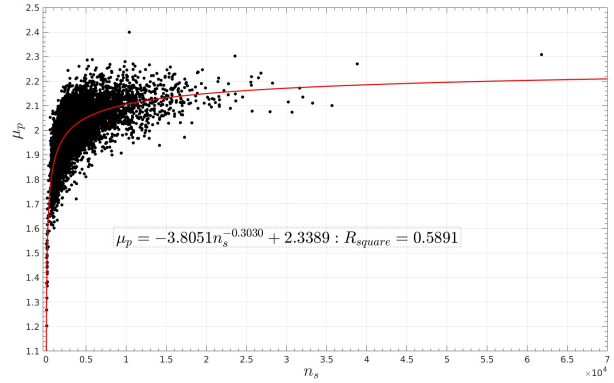
(a) μ_s vs n_r



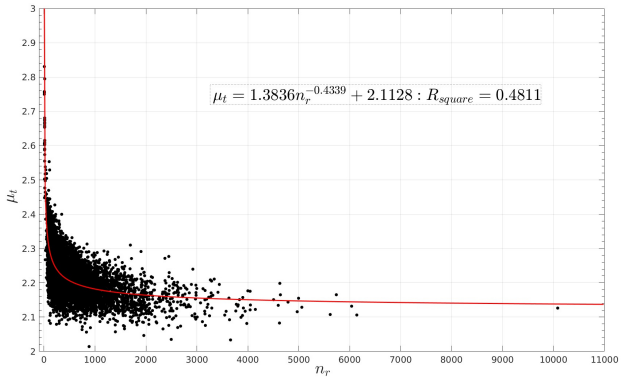
(b) μ_s vs n_s



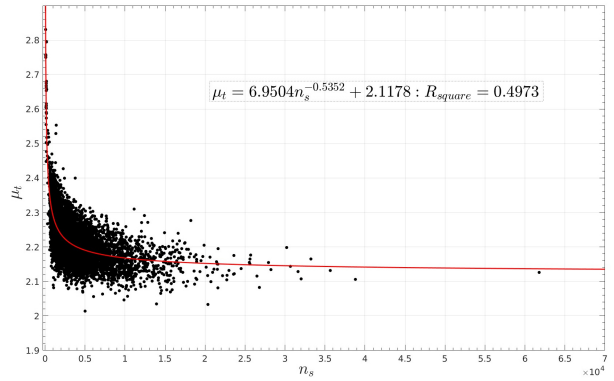
(c) μ_p vs n_r



(d) μ_p vs n_s



(e) μ_t vs n_r



(f) μ_t vs n_s

Figure S7: **The changes of μ_s, μ_p and μ_t with n_r and n_s for G**. Figs. (**a, c, e**) show the changes of μs with n_r while Figs. (**b, d, f**) the changes with n_s. The insets list the power laws and R_{square}s. The power law curves are depicted in red. The point of intersection of the power laws for μ_s and μ_p is $n_r = 62$, which is slightly larger than 60 residues computed using the linear equation in Fig. S1. The point of intersection of the power laws for μ_t and μ_p is $n_r = 3,241$. The existence of missing atoms in many structures means that these two n_r values are upper limits.

S8: The SESs of the individual chains of the multimers in G

We extract a set $\mathbf{G}_c$ of $20,050$ chains from all the multimers in $\mathbf{G}$ and compute their SESs individually. $\mathbf{G}_c$ includes no monomeric proteins. Among them the smallest chain is chain H of 4TY9 with only 3 residues and 60 atoms, and the largest one is chain A of 3EGW with $1,244$ residues and $19,403$ atoms. As shown in Fig. S8, the changes with n of the μs and the power laws for $\mathbf{G}_c$ are almost identical to those for $\mathbf{G}$ itself. In addition, both the μ_ss and μ_ps of $\mathbf{G}_c$ could be fitted to a power law better than the corresponding ones for $\mathbf{G}$. In addition, the closeness in SES of $\mathbf{G}_c$ to $\mathbf{G}$ and $\mathbf{S}$ to $\mathbf{G}$ shows that crystal packing, if exists, should have minimal impacts on the geometric properties studied in this paper.

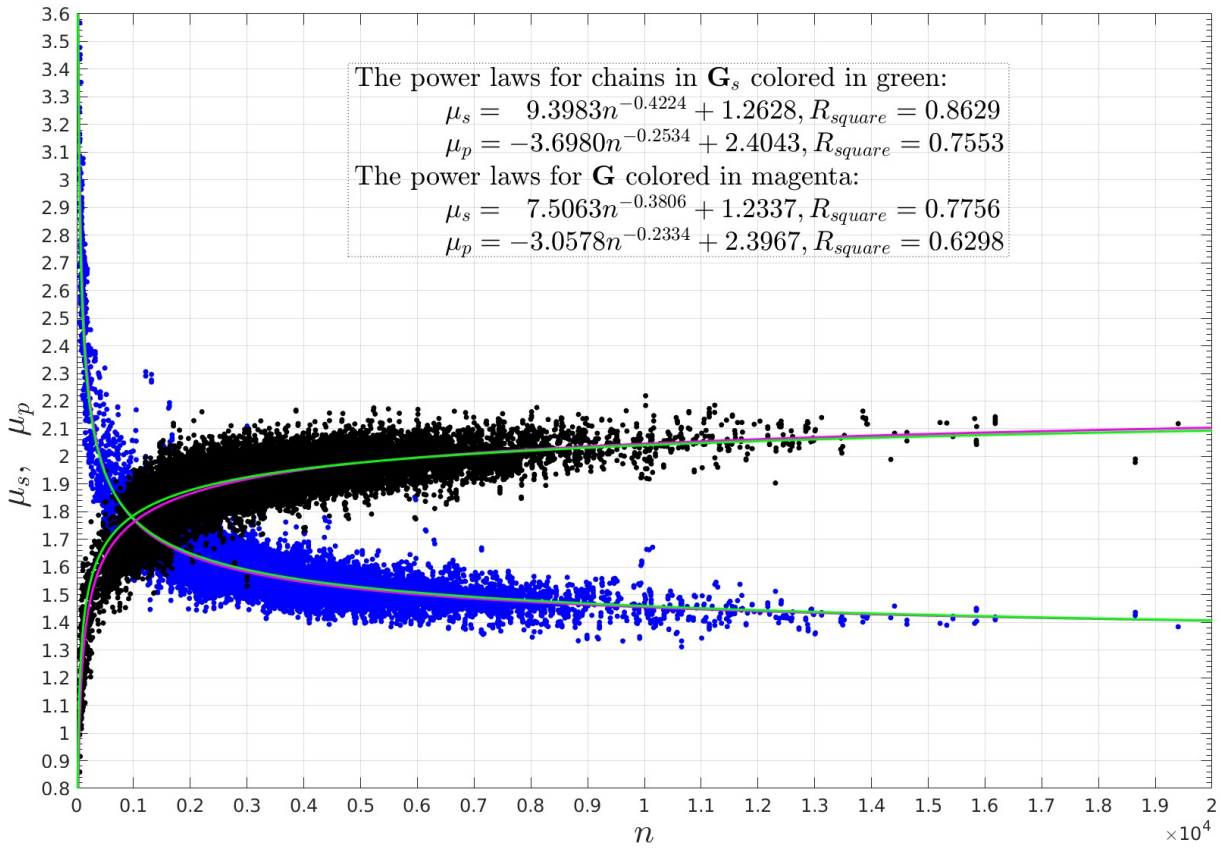


Figure S8: **The power laws for the μs of the individual chains in G**. The μ_ss and μ_ps are depicted, respectively, in blue and black. The insert lists the power laws for the μs of both $\mathbf{G}$ and $\mathbf{G}_c$, the set of $\mathbf{G}$'s individual chains. The power laws for $\mathbf{G}_c$ and n_{sp} are almost identical to those for $\mathbf{G}$ itself. The mean and standard deviation of the list of the $\Delta\mu_s$s (defined in the legend of Table 1 of the main text) are only -0.0054Å^2 and 0.0637Å^2, the absolute value of the mean is almost 10-fold smaller than $\text{RMSE}_{G,s} = 0.0532$. The mean and standard deviation of the list of the $\Delta\mu_p$s are only -0.0097Å^2 and 0.0673Å^2, the absolute value of the mean is almost 7-fold smaller than $\text{RMSE}_{G,p} = 0.0631$.

S9: The R_{square}s, N_{sp}s and τs for the crystal structures and AlphaFold models

Table S1 lists the R_{square}s, N_{sp}s (points of intersection) and τs for the three sets of crystal structures and the twenty seven AlphaFold proteomes.

Organism	R_{square} (μ_s/μ_p)	τ	N_{sp}	μ_{sp}
Crystal structures				
G	**0.776 / 0.630**	**0.3770**	**60**	**1.783**
S	**0.560 / 0.521**	**0.3952**	**57**	**1.775**
M	**0.666 / 0.588**	**0.4234**	**64**	**1.714**
Prokaryotic proteomes				
E.coli	*0.582 / 0.571*	*0.4659*	*87*	*1.781*
haeIn	0.487 / 0.501	0.4767	86	1.780
helPy	0.499 / 0.523	0.4840	89	1.779
klePh	0.629 / 0.582	0.4694	92	1.779
metJa	0.487 / 0.488	0.4494	75	1.782
mycLe	0.372 / 0.331	0.4675	110	1.790
mycTu	0.369 / 0.379	0.4833	113	1.772
neiG1	0.556 / 0.482	0.4860	99	1.787
pseAr	0.404 / 0.414	0.4868	93	1.769
salTy	0.504 / 0.457	0.4739	91	1.779
shiDs	0.509 / 0.503	0.4669	93	1.787
staA8	0.516 / 0.546	0.5057	75	1.785
strR6	0.505 / 0.503	0.4909	81	1.783
Yeast proteomes				
yeast	0.056 / 0.083	0.5927	133	1.799
schPo	0.056 / 0.086	0.5752	125	1.793
canAl	0.036 / 0.072	0.5955	145	1.800
Plant proteomes				
araTh	0.125 / 0.123	0.5908	246	1.808
orySj	0.206 / 0.172	0.5658	391	1.806
soyBn	0.093 / 0.081	0.5983	250	1.808
maize	0.104 / 0.094	0.6073	389	1.805
Eukaryotic proteomes				
celEg	0.126 / 0.123	0.6249	156	1.809
dicDi	0.023 / 0.018	0.6631	208	1.824
danRe	0.092 / 0.085	0.6580	116	1.805
droMe	0.037 / 0.040	0.6984	327	1.804
mouse	0.028 / 0.023	0.6750	193	1.798
rat	0.030 / 0.025	0.6716	258	1.797
human	***0.034 / 0.025***	***0.6742***	***694***	***1.800***

Table S1: The R_{square}s, τs and N_{sp}s for the three sets of crystal structures and twenty seven AlphaFold proteomes. Column 1 lists species names (Table S3). Column 2 lists the R_{square}s for μ_s and μ_p. Column 3 lists the slopes (τs) of the fitted linear equations between n and n_s. Column 4 lists the N_{sp}s (computed using, $N_{sp} = \frac{n_{sp}}{16}$, Fig. S1). Among the twenty seven proteomes, the human proteome has the largest $N_{sp} = 694$ and the third largest $\tau = 0.6742$.

S10: The differences in SES between the *E.coli* proteome and the other twenty six proteomes

Table S2 lists the $\Delta\mu_s$s, $\Delta\mu_p$s and P_ss and P_ps for the twenty six proteomes computed with the power laws for the *E.coli* proteome as references.

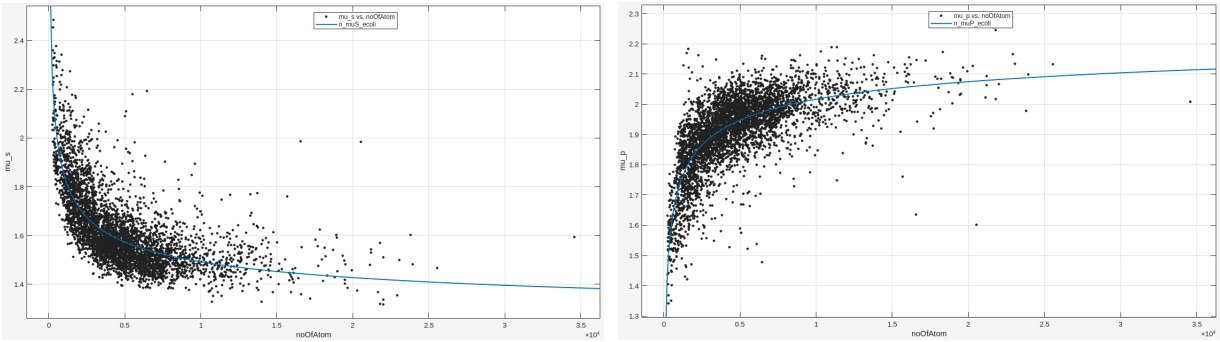
Organism	$\Delta\mu_s$ (mean / std) ($\mathring{A}^2$)	$\Delta\mu_p$ (mean / std) ($\mathring{A}^2$)	$P_s(\%)$	$P_p(\%)$
Prokaryotic proteomes				
haeIn	-0.0001 / 0.0992	0.0030 / 0.0804	5.1	4.9
helPy	0.0105 / 0.1117	-0.0057 / 0.0885	7.1	6.1
klePh	0.0029 / 0.1027	-0.0082 / 0.0857	4.9	6.1
metja	***-0.0207 / 0.0859***	***0.0199 / 0.0724***	***2.4***	***4.3***
mycLe	0.0458 / 0.1352	-0.0300 / 0.1089	12.9	12.7
mycTu	0.0355 / 0.1382	-0.0404 / 0.1089	12.6	14.5
neiG1	0.0219 / 0.1252	-0.0130 / 0.1015	9.5	10.7
pseAr	0.0036 / 0.1112	-0.0184 / 0.0854	6.0	5.8
salTy	0.0064 / 0.1036	-0.0036 / 0.0843	5.7	5.8
shiDs	0.0076 / 0.1057	-0.0064 / 0.0872	5.6	6.5
staA8	0.0014 / 0.1135	0.0147 / 0.0911	6.0	5.8
strR6	0.0025 / 0.1008	0.0069 / 0.0836	5.2	5.5
Yeast proteomes				
schPo	0.1603 / 0.2009	-0.0914 / 0.1617	34.8	28.7
yeast	0.1859 / 0.2104	-0.1058 / 0.1669	39.7	32.9
canAl	0.1936 / 0.2159	-0.1053 / 0.1668	40.9	31.5
Plant proteomes				
araTh	0.2074 / 0.2100	-0.1188 / 0.1709	43.1	36.3
soyBn	0.2076 / 0.2136	-0.1174 / 0.1775	43.3	36.6
maize	0.2423 / 0.2155	-0.1512 / 0.1778	51.6	43.7
orysj	0.2548 / 0.2245	-0.1527 / 0.1857	53.1	47.9
Eukaryotic proteomes				
celEg	0.1740 / 0.2089	-0.0916 / 0.1652	37.7	29.9
dicDi	0.2119 / 0.2182	-0.0938 / 0.1841	46.0	33.6
danRe	0.2386 / 0.2196	-0.1497 / 0.1883	50.2	40.2
droMe	0.2445 / 0.2380	-0.1573 / 0.1973	49.8	41.9
mouse	0.2370 / 0.2330	-0.1617 / 0.1934	50.1	43.1
rat	0.2329 / 0.2312	-0.1593 / 0.1918	49.5	42.7
human	***0.2786 / 0.2330***	***-0.1935 / 0.2050***	***57.7***	***48.5***

Table S2: The differences in SES between the *E.coli* proteome and other twenty six proteomes. Column 1 lists organism names (Table S3). The differences in μ_s and μ_p of model i of a proteome from the μ_ss and μ_ps of the *E.coli* proteome are quantified as $\Delta\mu_s(i) = \mu_s^\alpha(i) - \mu_s^e(i)$ and $\Delta\mu_p(i) = \mu_p^\alpha(i) - \mu_p^e(i))$, where $\mu_s^\alpha(i), \mu_p^\alpha(i)$ are the μ_s, μ_p values of model i, and $\mu_s^e(i), \mu_p^e(i)$ are computed using, respectively, the power laws for the μ_ss and μ_ps of the *E.coli* proteome, with n in Eq. (5) being the number of atoms of model i. Column 2 lists the mean and standard deviation of the list of the $\Delta\mu_s$s for a proteome. Column 3 lists the mean and standard deviation of the list of the $\Delta\mu_p$s for a proteome. Columns 4 and 5 list, respectively, the P_ss and P_ps for the twenty-six proteomes, and Fig. 6 of the main text visualizes the P_s and P_p for the human proteome. The P_s and P_p have the same meanings as those in Table 1 of the main text except that they are calculated with the $\text{RMSE}_{E,s} = 0.1002$ for fitting to a power law the μ_ss of the *E.coli* proteome, and $\text{RMSE}_{E,p} = 0.0816$ for fitting to a power law the μ_ps of the *E.coli* proteome. As shown in the last row, the μs, P_s and P_e for the human proteome are the furthest from those for the *E.coli* proteome.

S11: The changes with n of the μs and the power laws, and the n_ss and the linear equations for the twenty seven AlphaFold proteomes

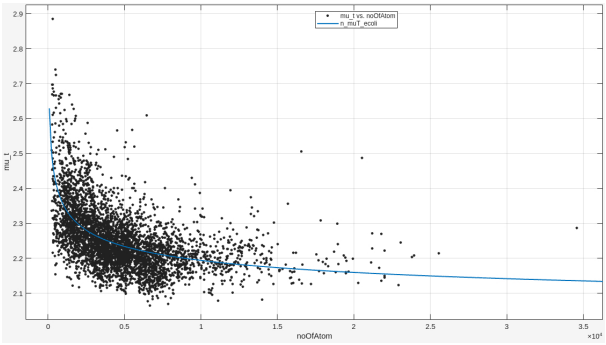
The Figs. S9 to S35 depict the changes of μ_s, μ_p, μ_t and n_s with respect to n for the twenty seven proteomes in the order of thirteen prokaryotic (Figs. S9–S21), three yeast (Figs. S22–S24), four plant (Figs. (S25–S28)) and seven animal (Figs. S29–S35) proteomes.

S11.1: The thirteen prokaryotic proteomes

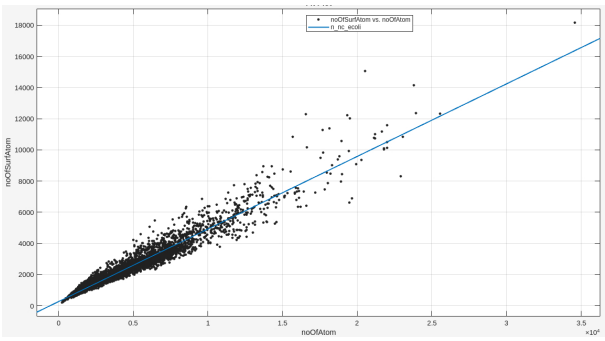


(a) μ_s

(b) μ_p

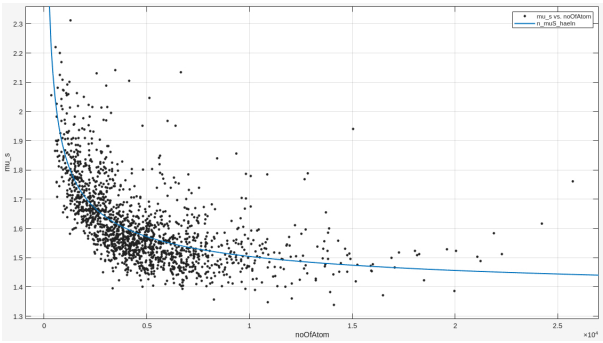


(c) μ_t

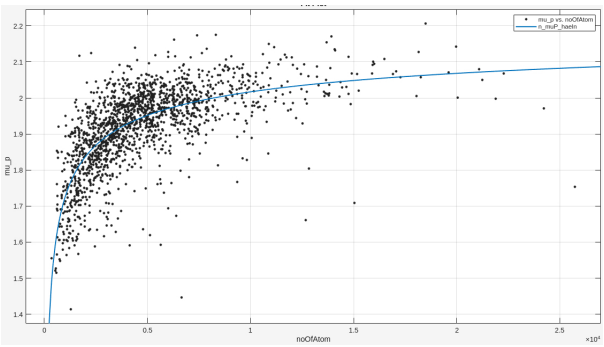


(d) n_s

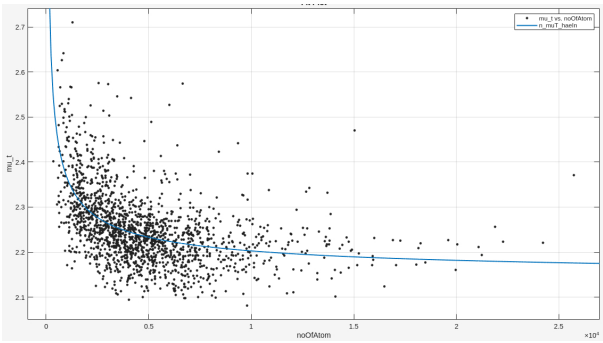
Figure S9: **The changes with n of the μ_s, μ_p, μ_t and n_s for the *Escherichia coli* proteome**.



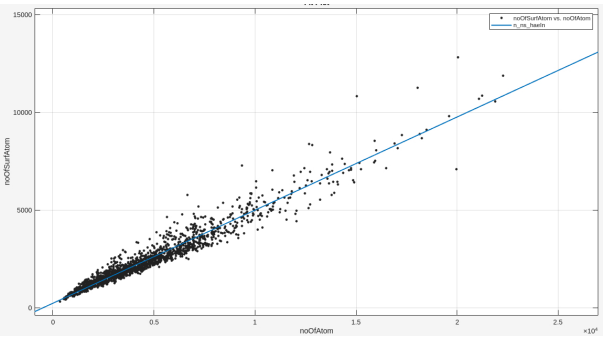
(a) μ_s



(b) μ_p

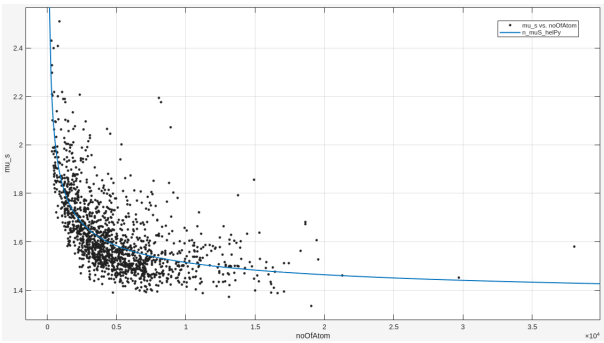


(c) μ_t

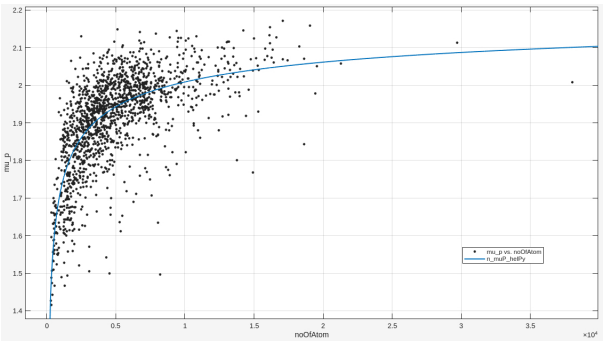


(d) n_s

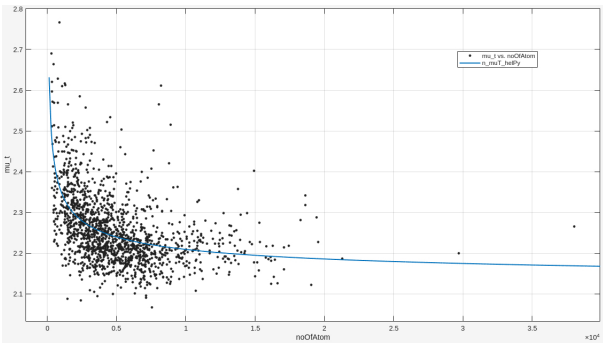
Figure S10: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Haemophilus influenzae* proteome**.



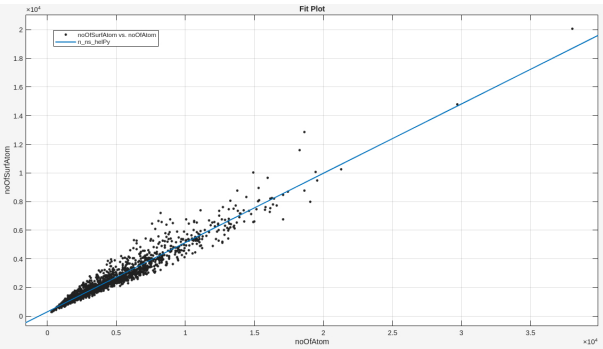
(a) μ_s



(b) μ_p

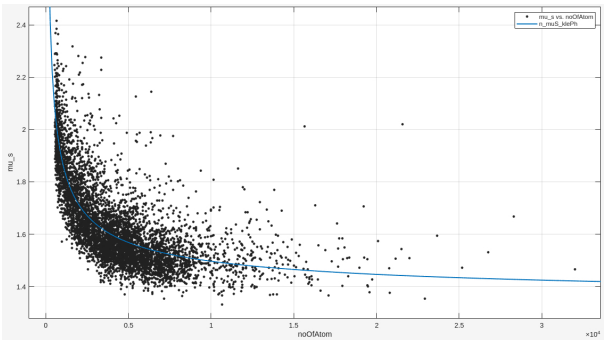


(c) μ_t

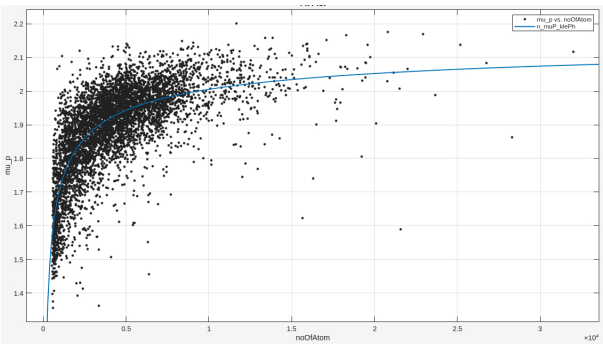


(d) n_s

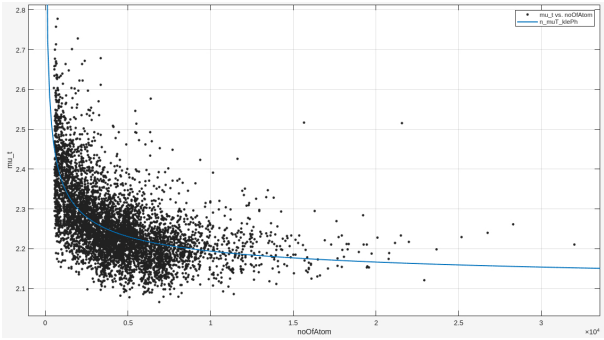
Figure S11: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Helicobacter pylori* proteome**.



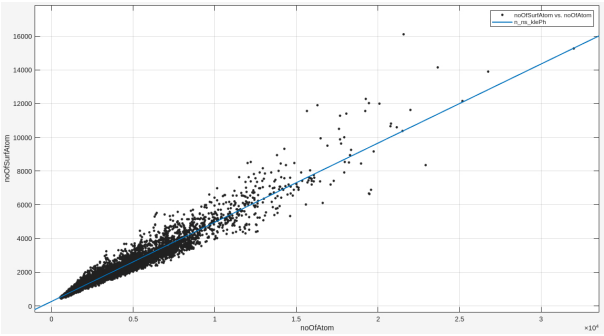
(a) μ_s



(b) μ_p

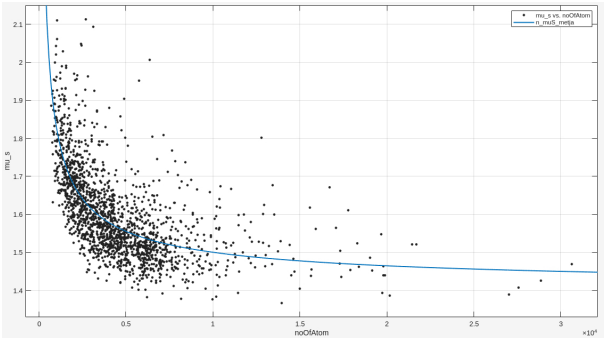


(c) μ_t

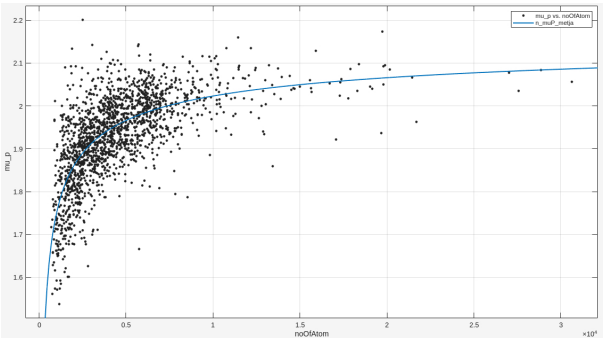


(d) n_s

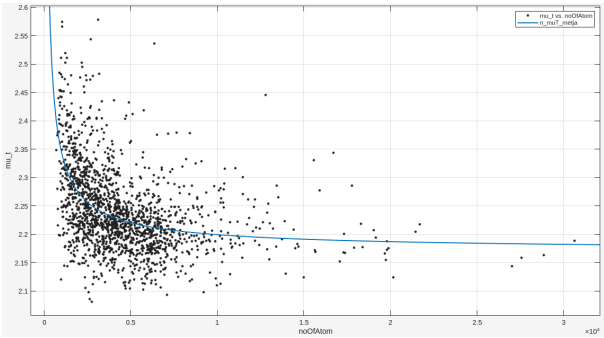
Figure S12: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Klebsiella pneumoniae* proteome**.



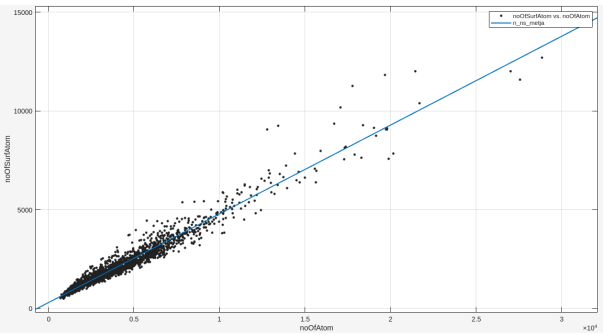
(a) μ_s



(b) μ_p

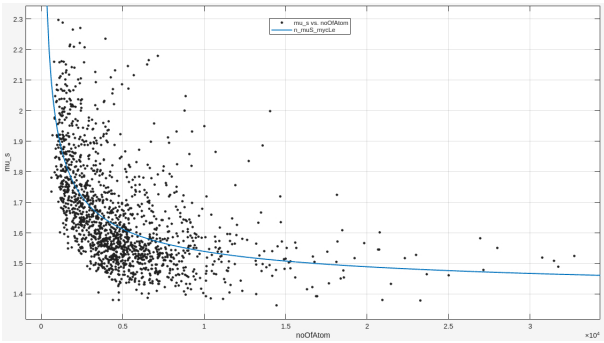


(c) μ_t

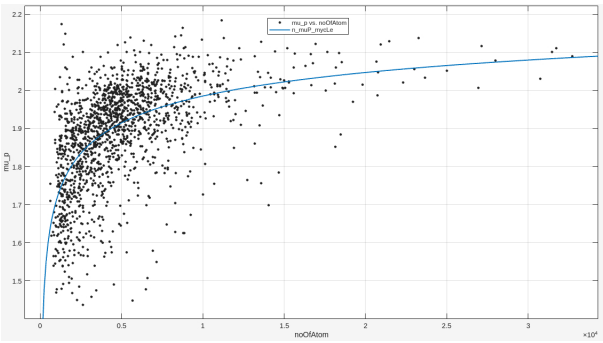


(d) n_s

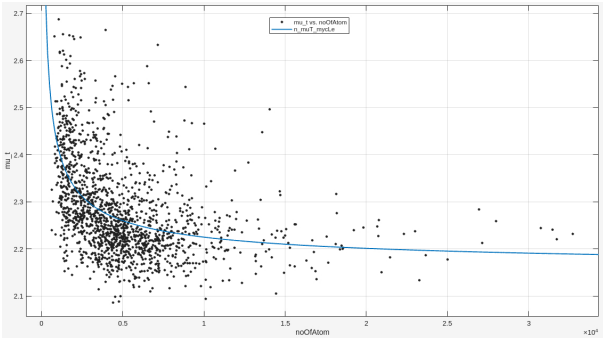
Figure S13: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Methanocaldococcus jannaschii* proteome**.



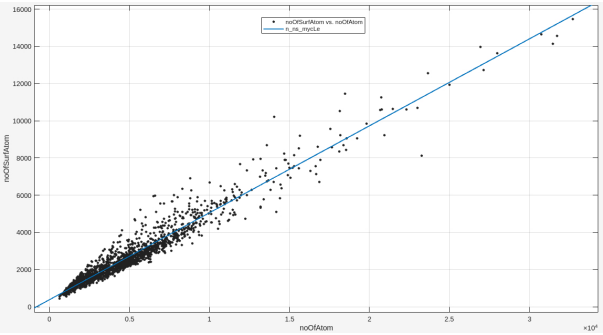
(a) μ_s



(b) μ_p

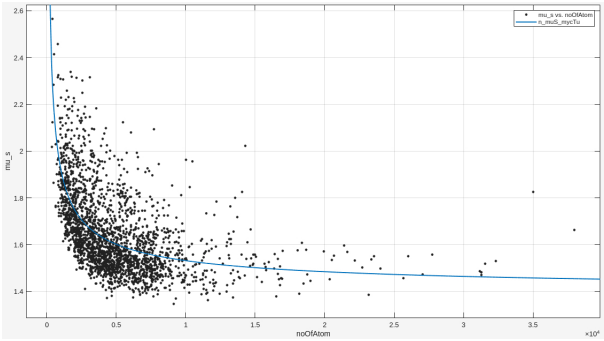


(c) μ_t

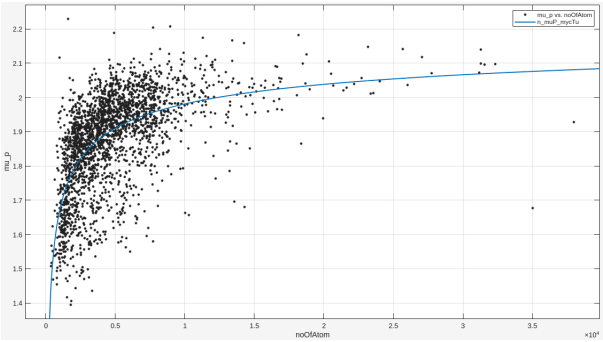


(d) n_s

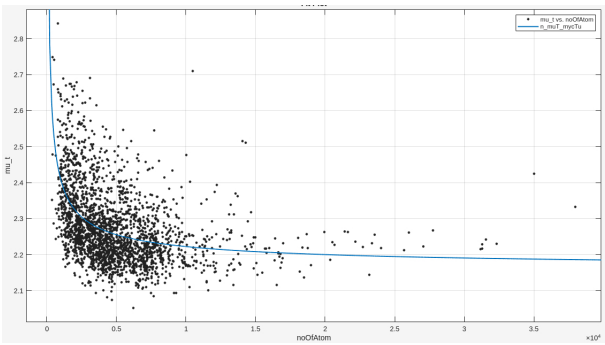
Figure S14: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Mycobacterium leprae* proteome**.



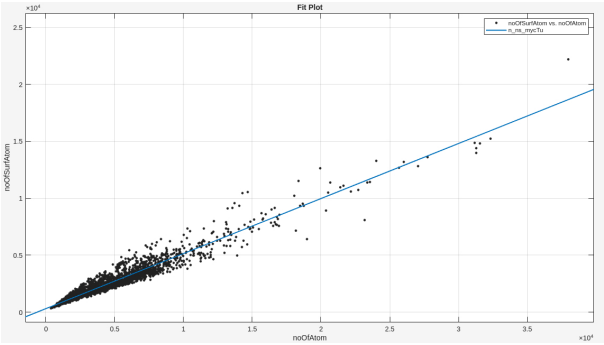
(a) μ_s



(b) μ_p

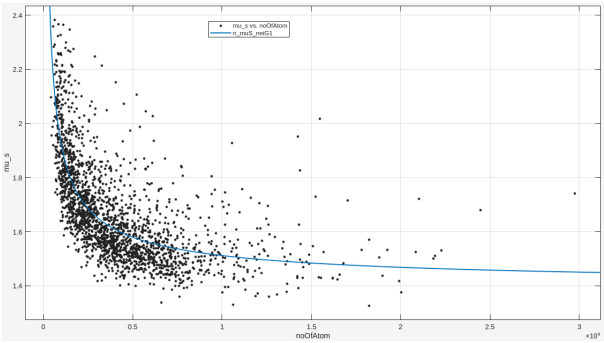


(c) μ_t

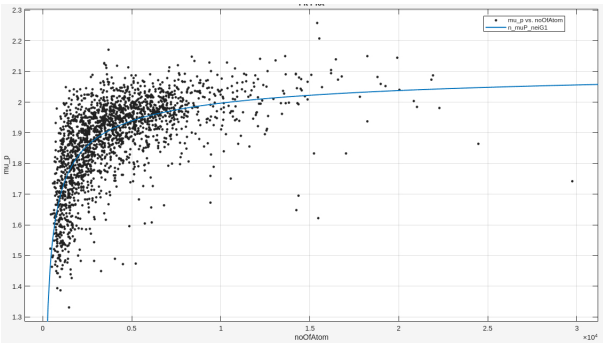


(d) n_s

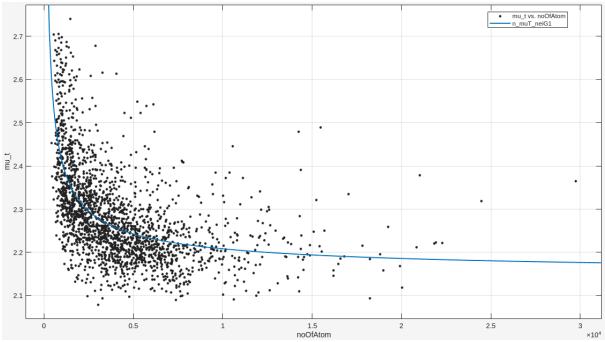
Figure S15: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Mycobacterium tuberculosis* proteome**.



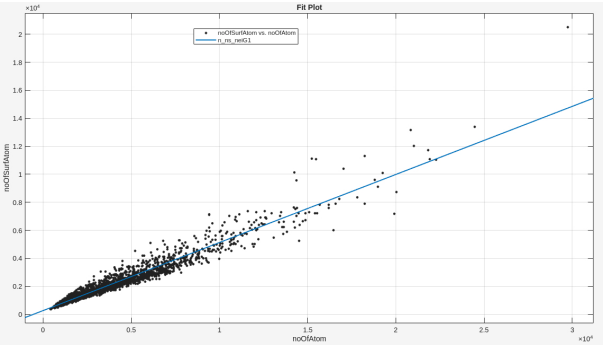
(a) μ_s



(b) μ_p

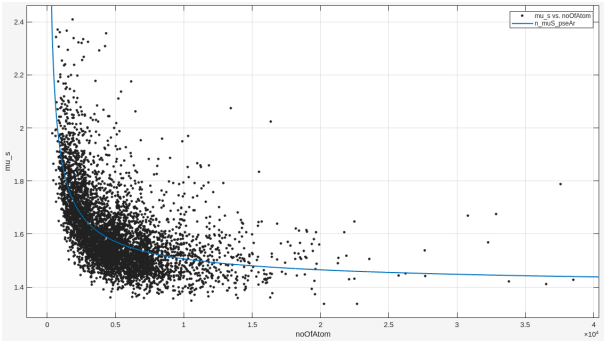


(c) μ_t

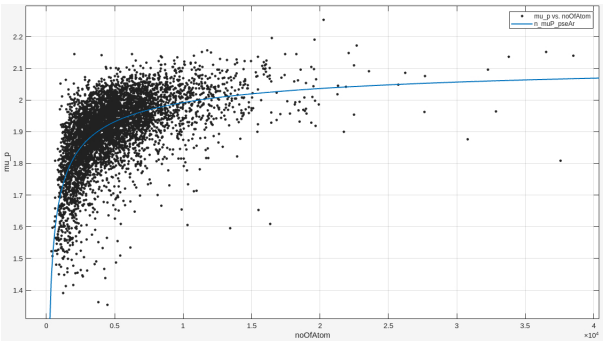


(d) n_s

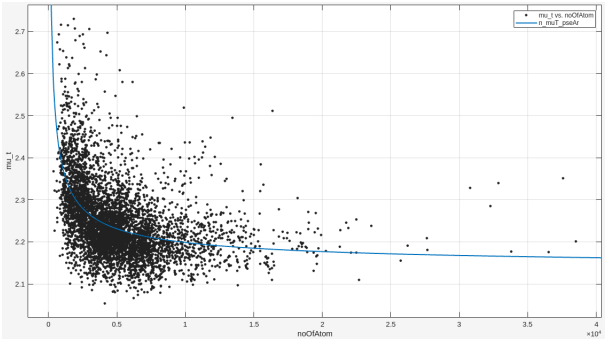
Figure S16: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Neisseria gonorrhoeae* proteome**.



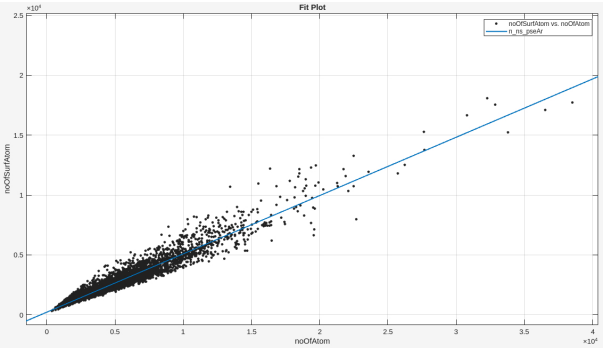
(a) μ_s



(b) μ_p

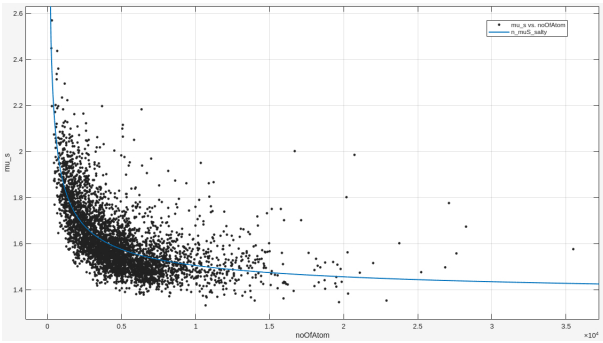


(c) μ_t

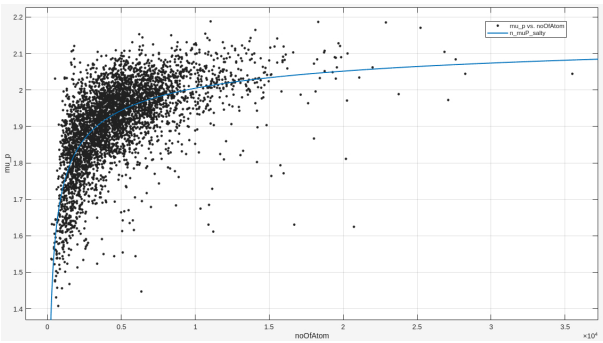


(d) n_s

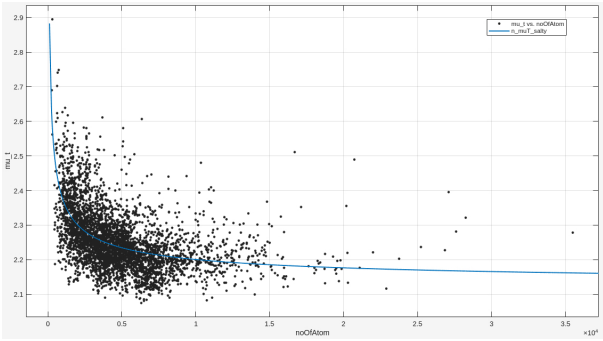
Figure S17: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Pseudomonas aeruginosa* proteome**.



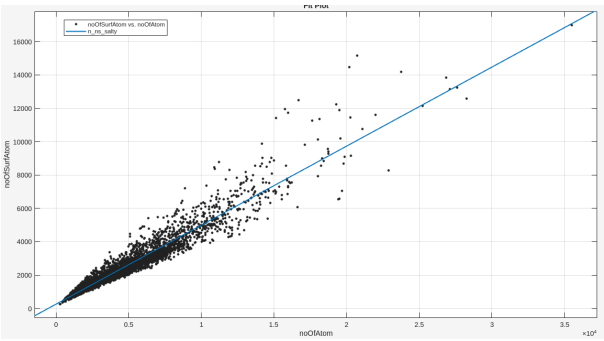
(a) μ_s



(b) μ_p

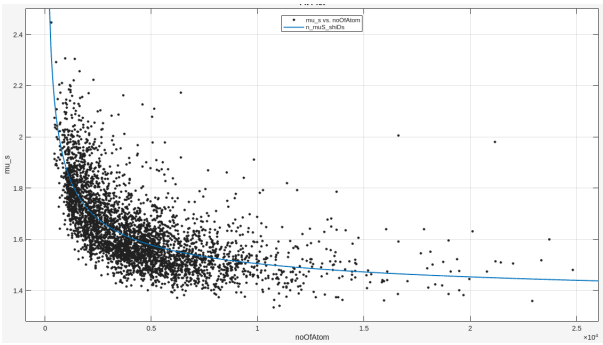


(c) μ_t

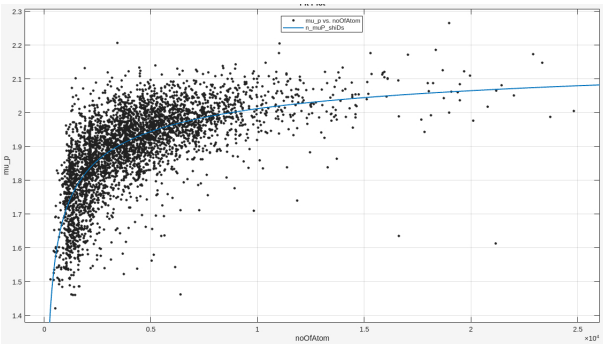


(d) n_s

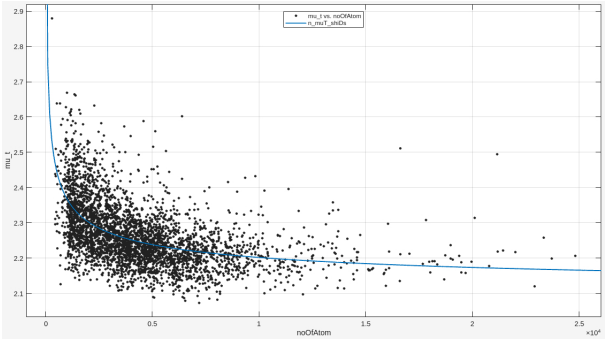
Figure S18: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Salmonella typhimurium* proteome**.



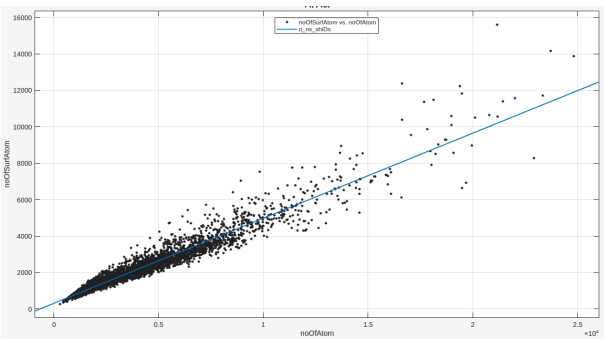
(a) μ_s



(b) μ_p

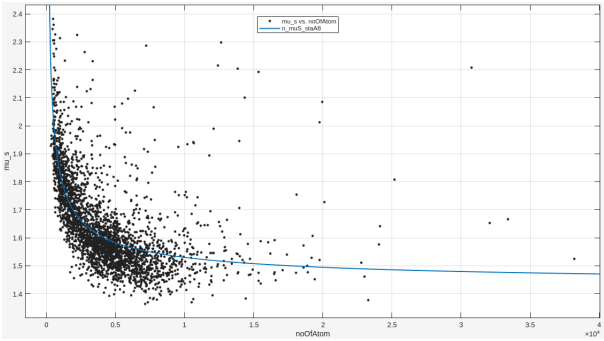


(c) μ_t

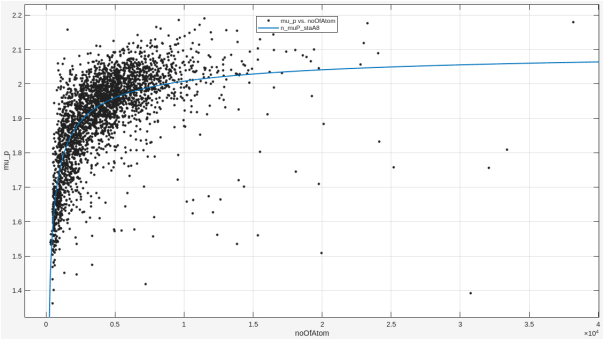


(d) n_s

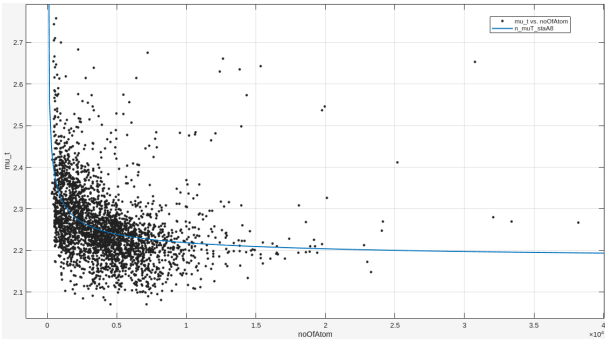
Figure S19: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Shigella dysenteriae* proteome**.



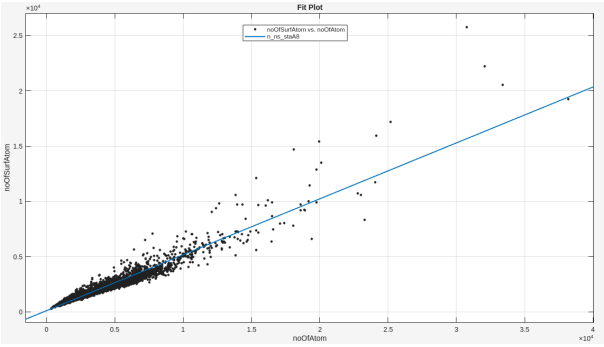
(a) μ_s



(b) μ_p

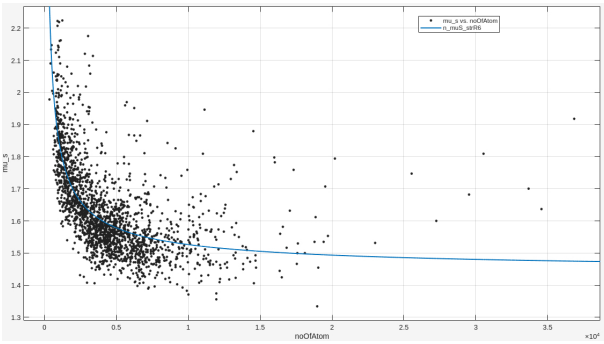


(c) μ_t

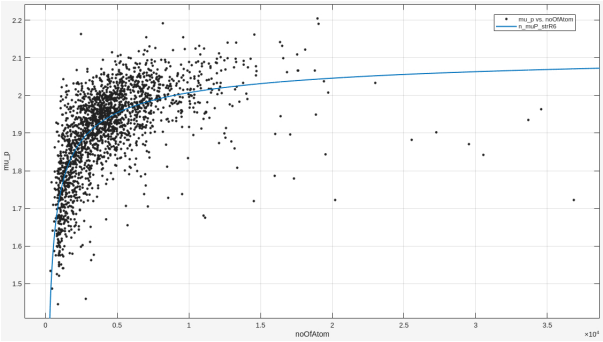


(d) n_s

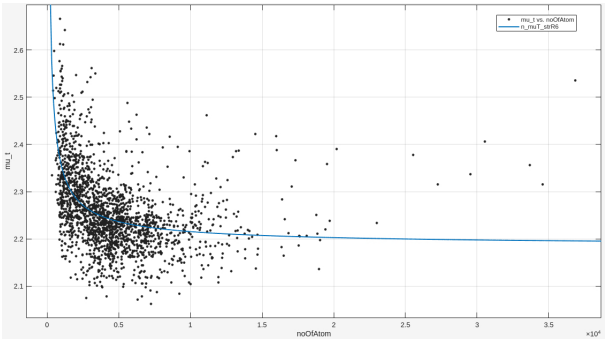
Figure S20: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Staphylococcus aureus* proteome**.



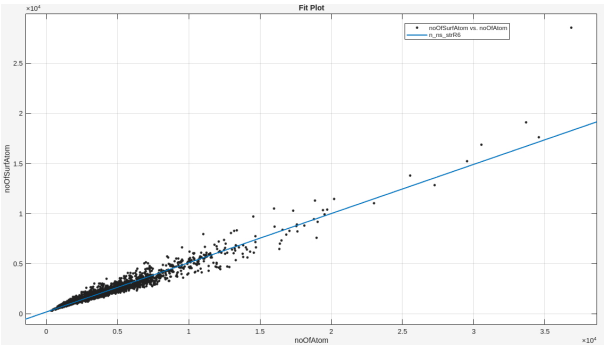
(a) μ_s



(b) μ_p



(c) μ_t

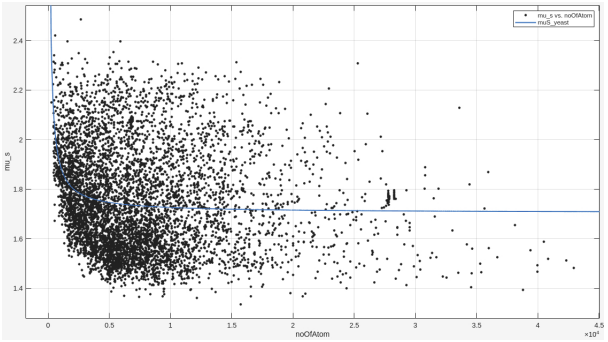


(d) n_s

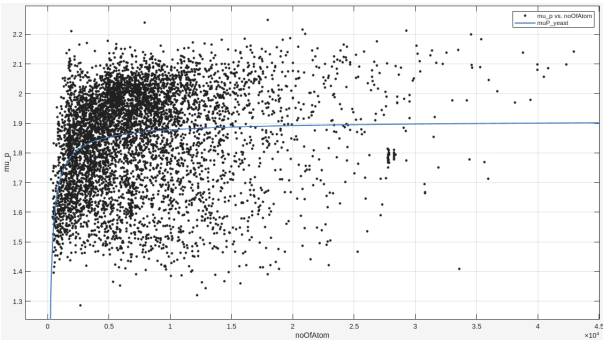
Figure S21: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Streptococcus pneumoniae* proteome**.

S11.2: The fourteen eukaryotic proteomes

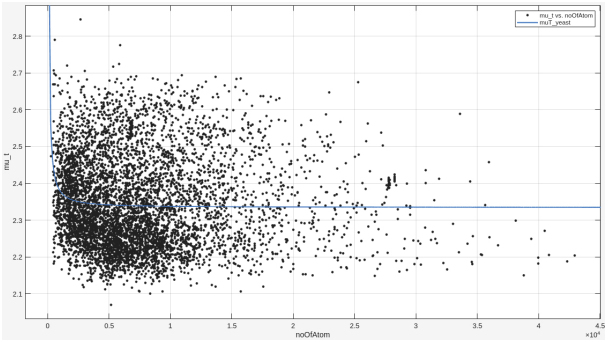
S11.2.1: Three yeast proteomes



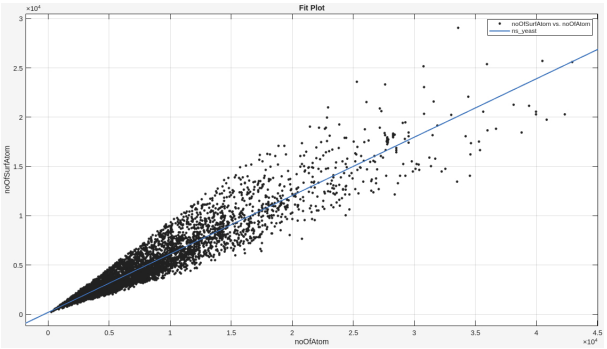
(a) μ_s



(b) μ_p

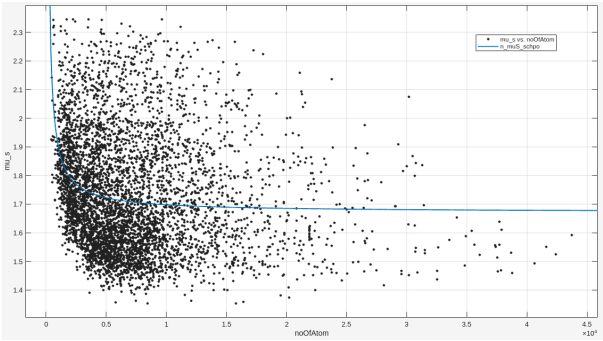


(c) μ_t

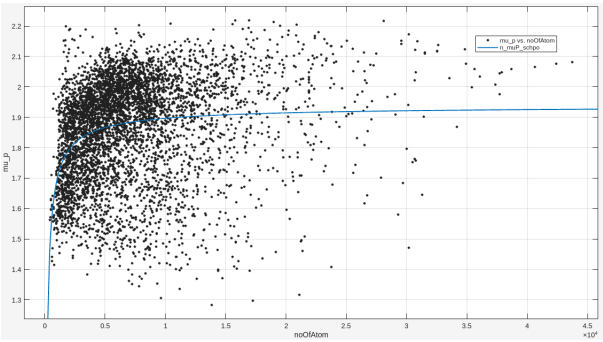


(d) n_s

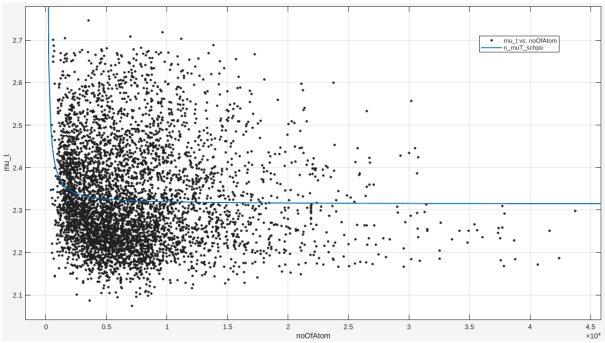
Figure S22: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Saccharomyces cerevisiae* proteome**.



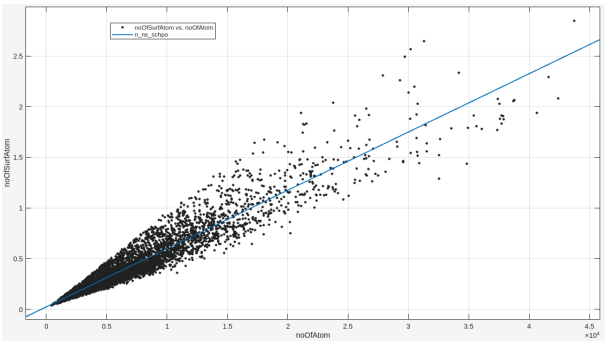
(a) μ_s



(b) μ_p

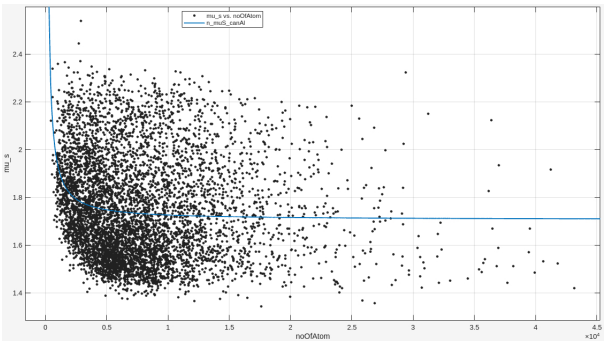


(c) μ_t

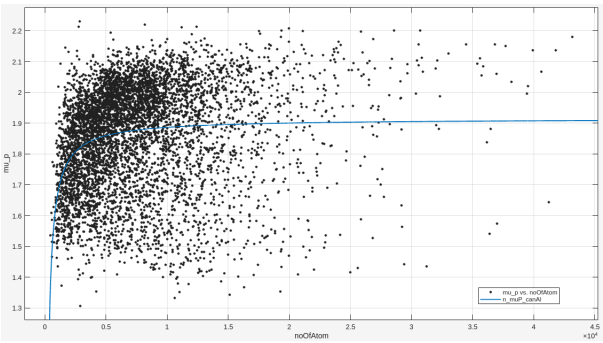


(d) n_s

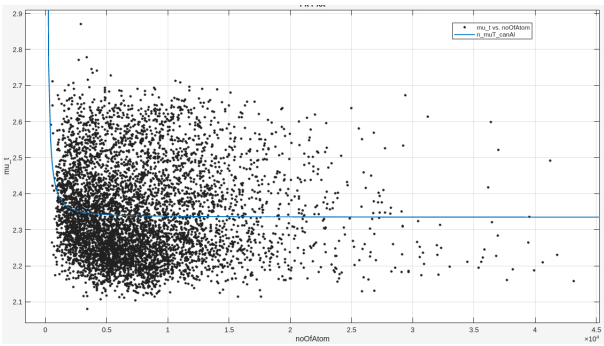
Figure S23: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Schizosaccharomyces pombe* proteome**.



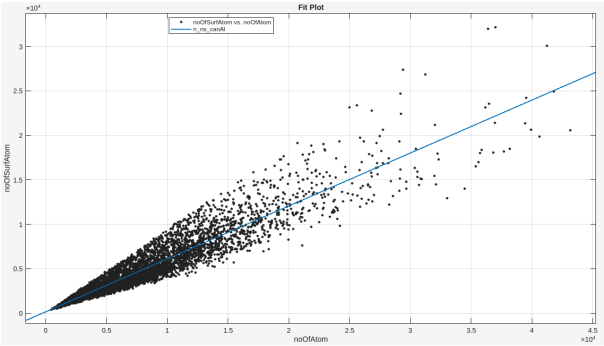
(a) μ_s



(b) μ_p



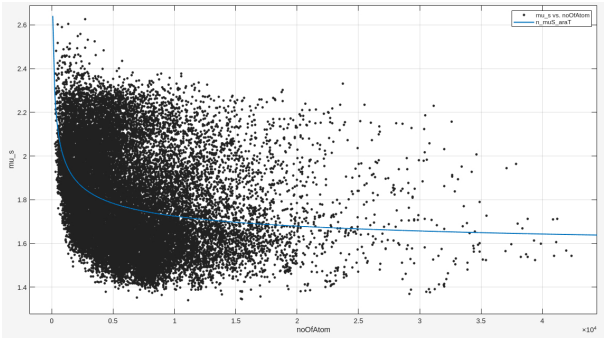
(c) μ_t



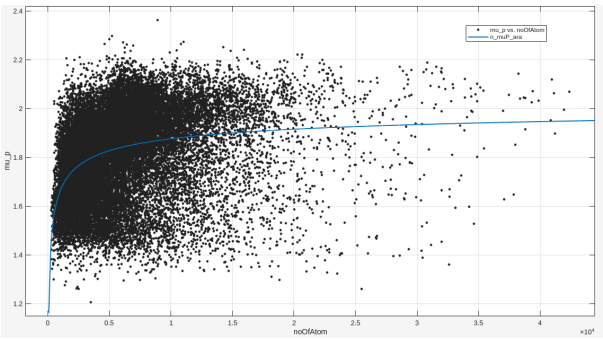
(d) n_s

Figure S24: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Candida albicans* proteome**.

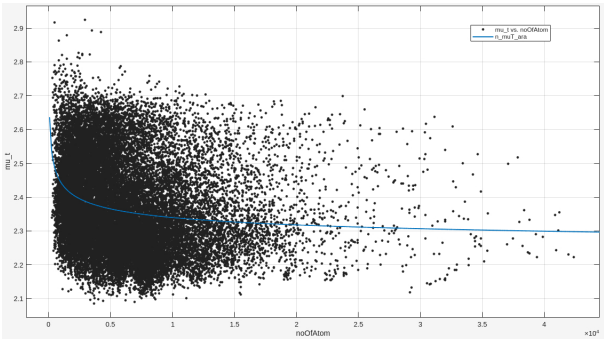
S11.2.2: Four plant proteomes



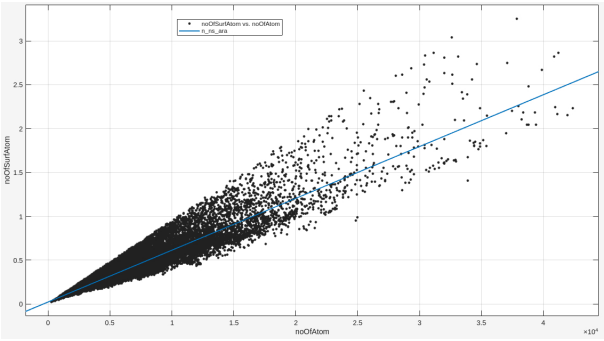
(a) μ_s



(b) μ_p

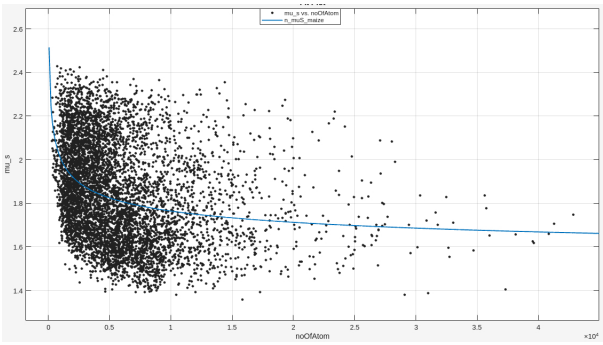


(c) μ_t

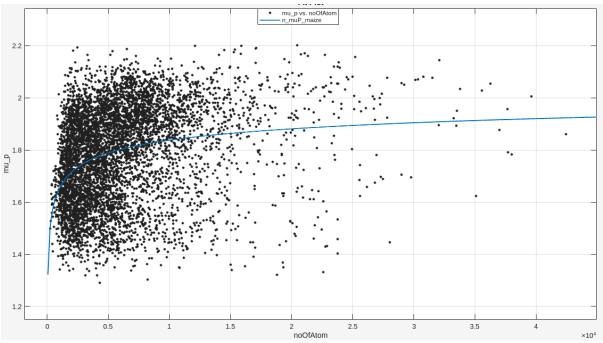


(d) n_s

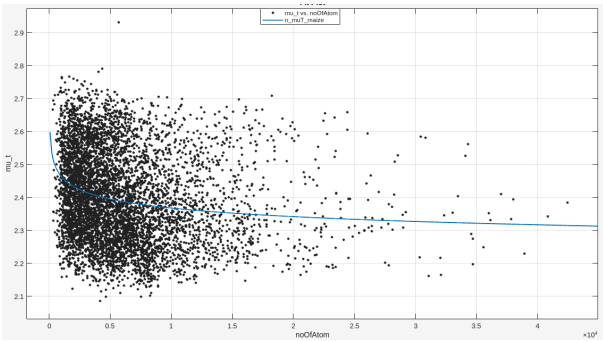
Figure S25: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Arabidopsis thaliana* proteome**.



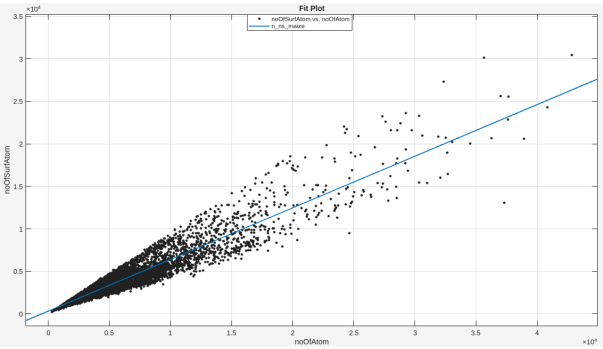
(a) μ_s



(b) μ_p

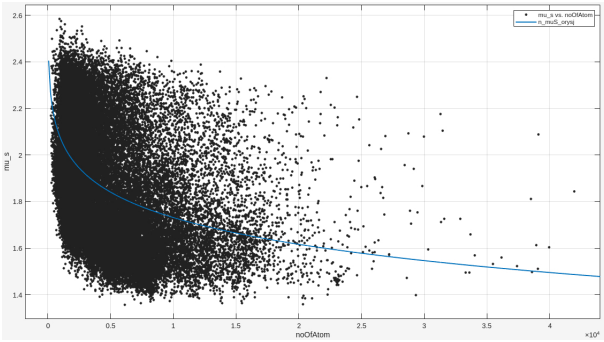


(c) μ_t

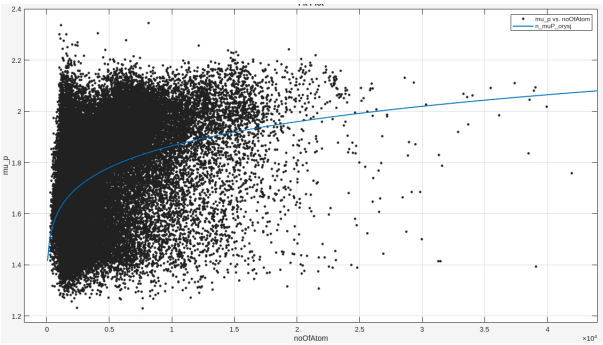


(d) n_s

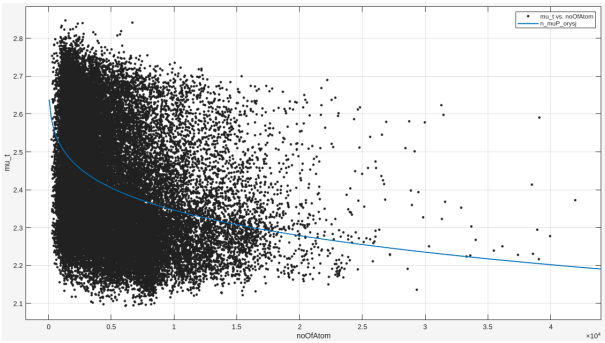
Figure S26: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Zea mays* proteome**.



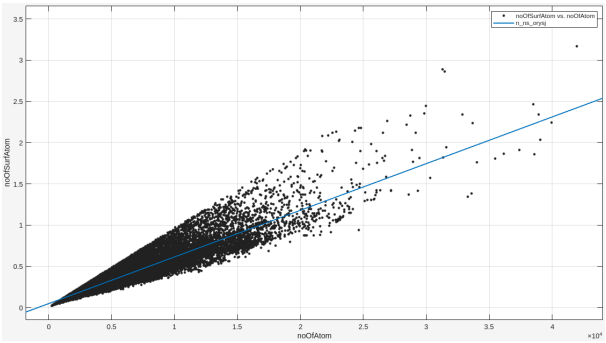
(a) μ_s



(b) μ_p

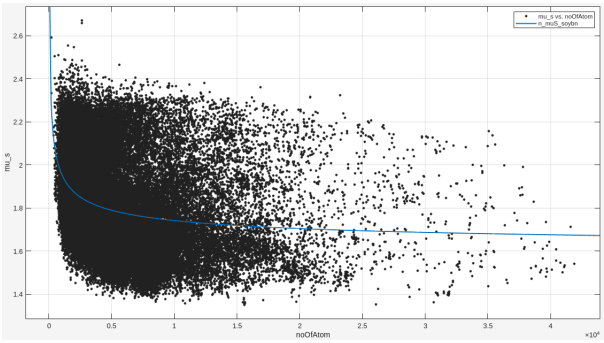


(c) μ_t

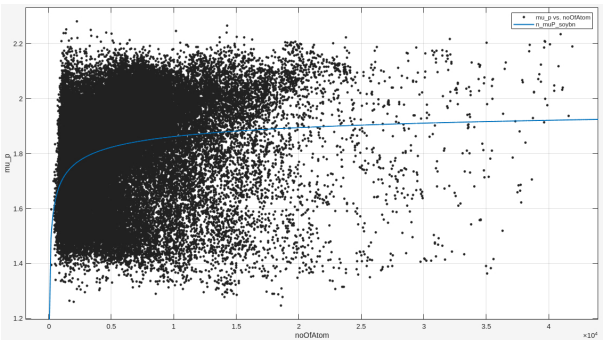


(d) n_s

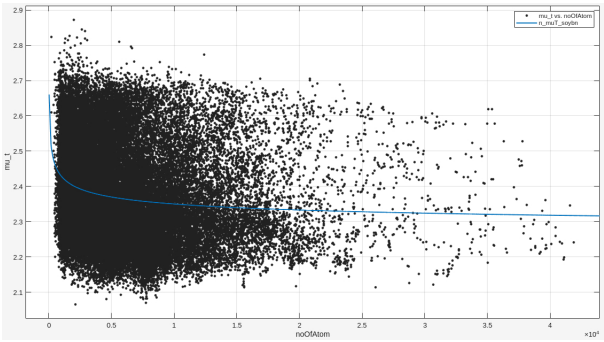
Figure S27: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Oryza sativa* proteome**.



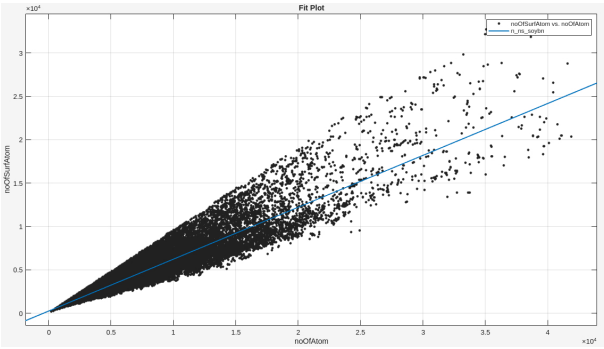
(a) μ_s



(b) μ_p



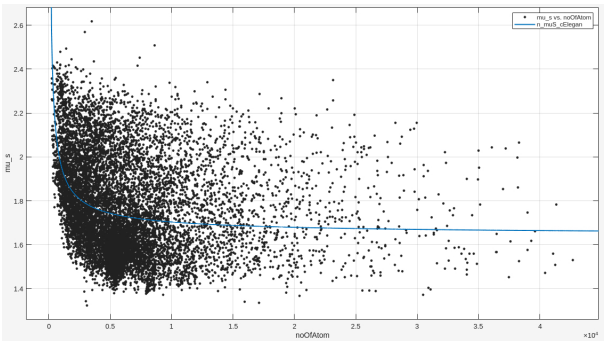
(c) μ_t



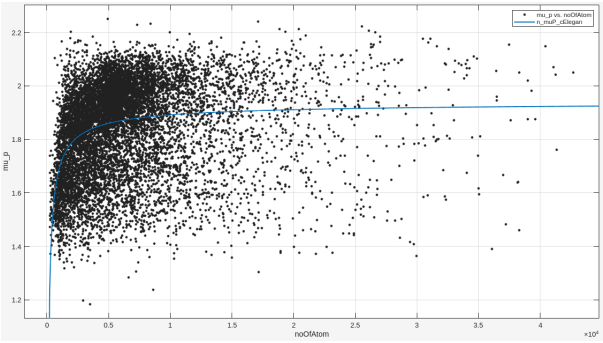
(d) n_s

Figure S28: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Glycine max* proteome**.

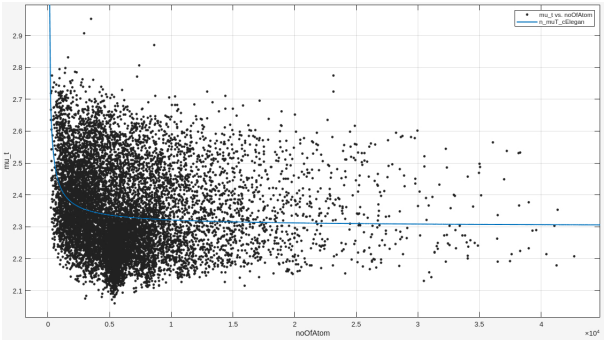
S11.2.3: Seven animal proteomes



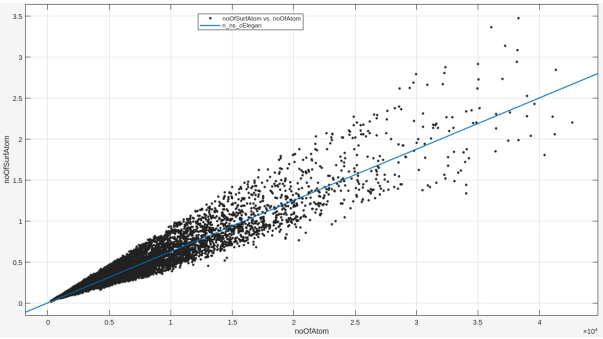
(a) μ_s



(b) μ_p

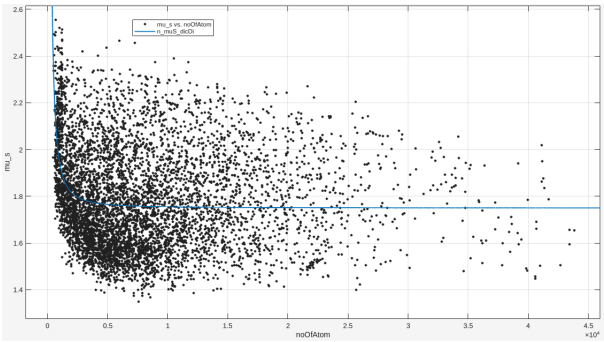


(c) μ_t

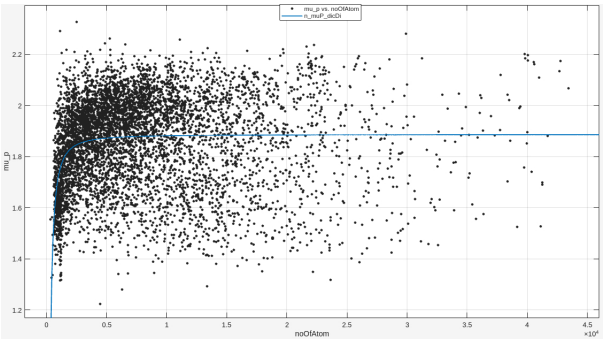


(d) n_s

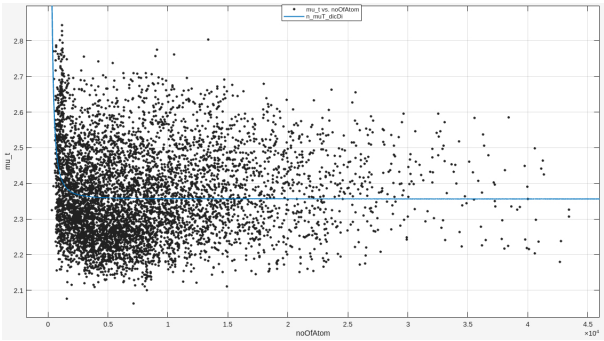
Figure S29: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Caenorhabditis elegans* proteome**.



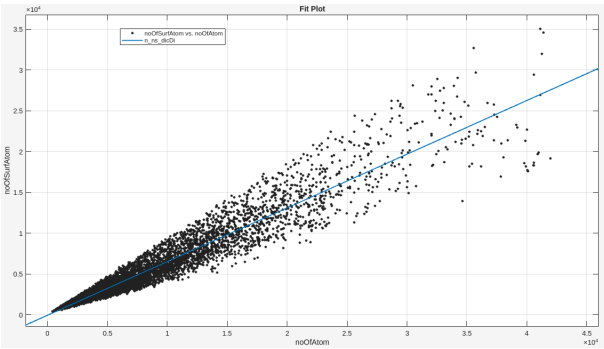
(a) μ_s



(b) μ_p

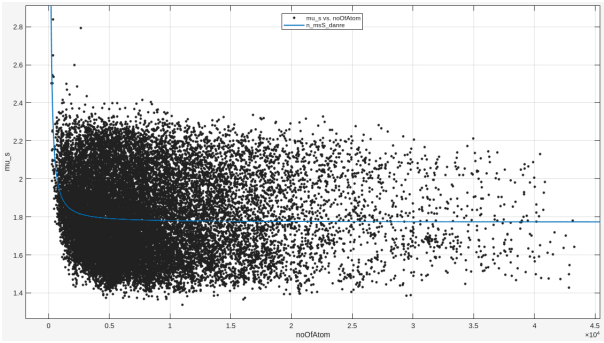


(c) μ_t

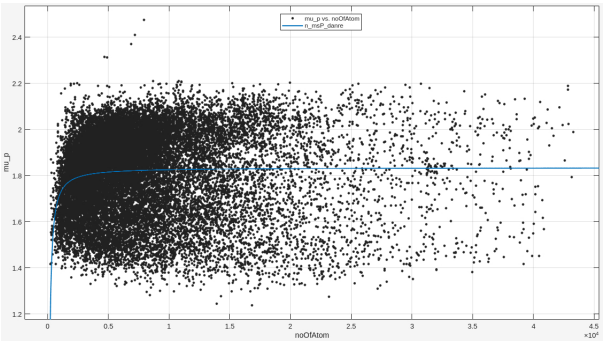


(d) n_s

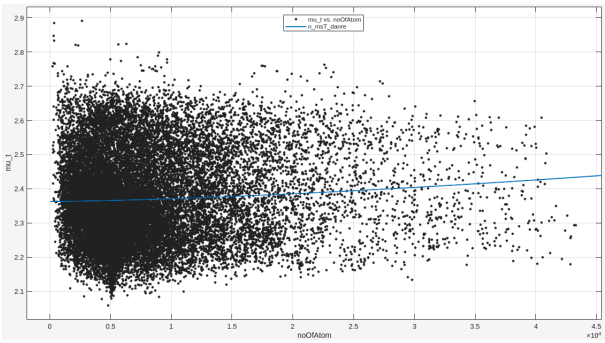
Figure S30: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Dictyostelium discoideum* proteome**.



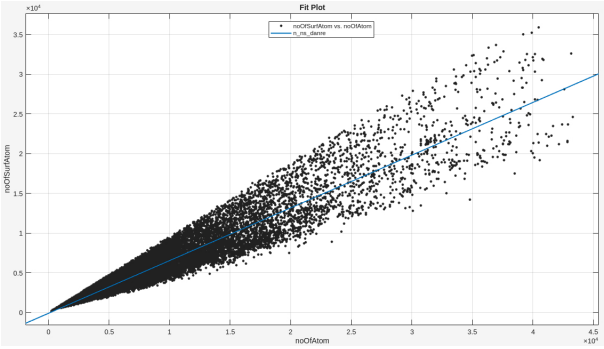
(a) μ_s



(b) μ_p

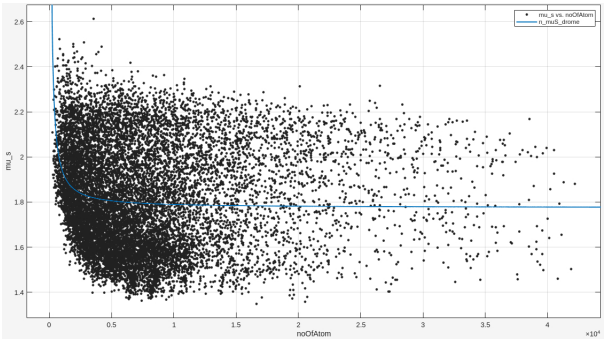


(c) μ_t

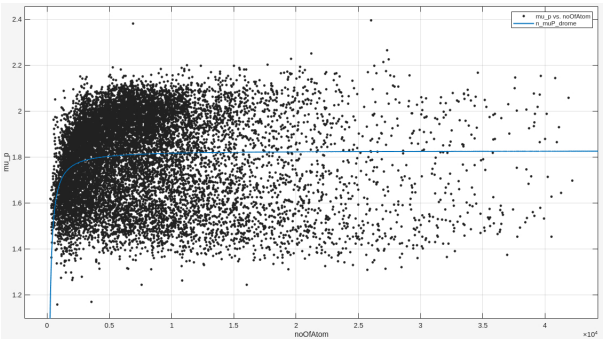


(d) n_s

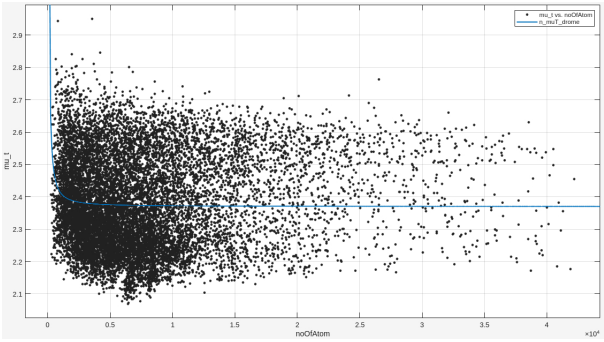
Figure S31: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Danio rerio* proteome**.



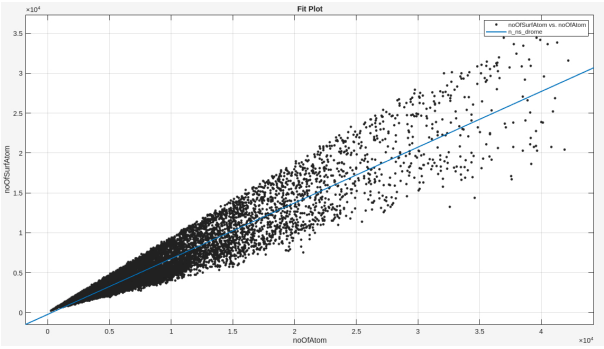
(a) μ_s



(b) μ_p

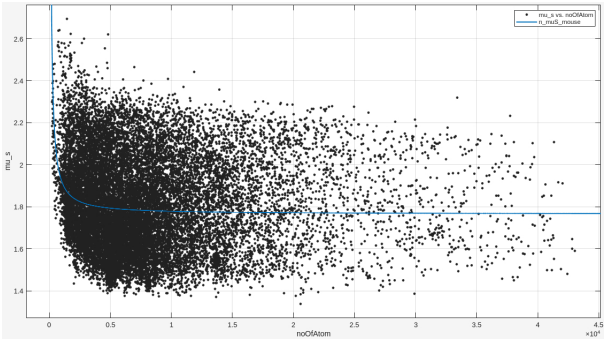


(c) μ_t

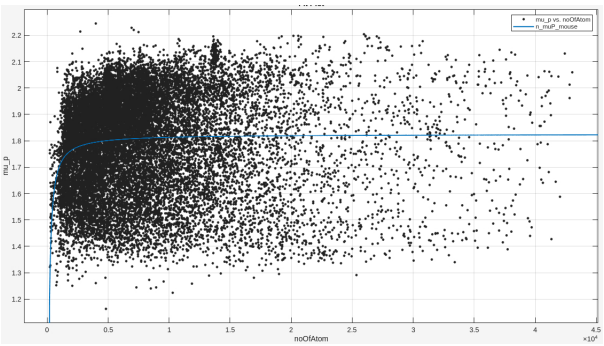


(d) n_s

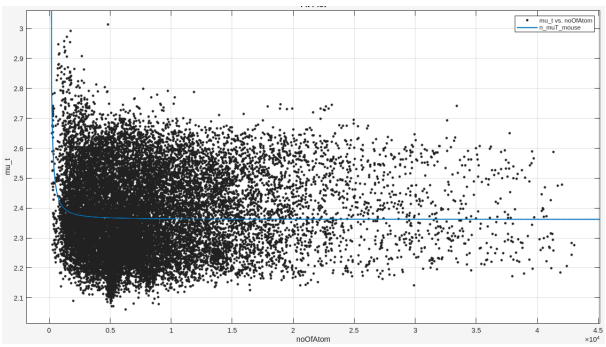
Figure S32: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Drosophila melanogaster* proteome**.



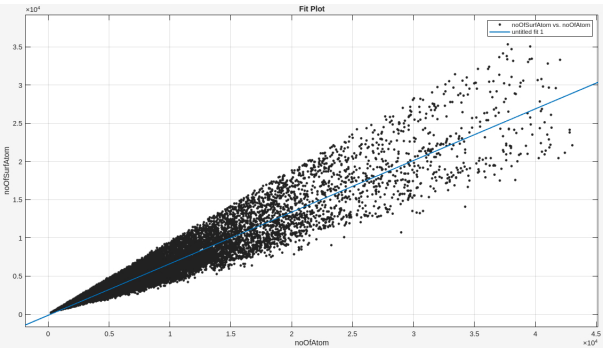
(a) μ_s



(b) μ_p

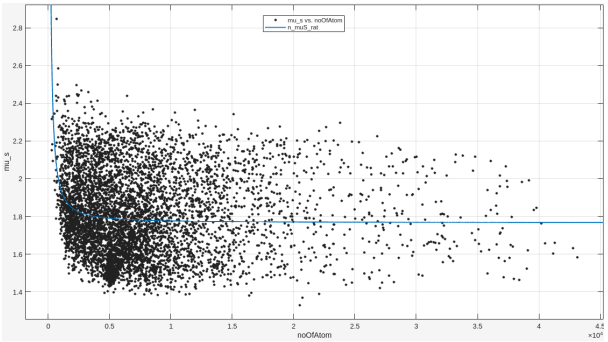


(c) μ_t

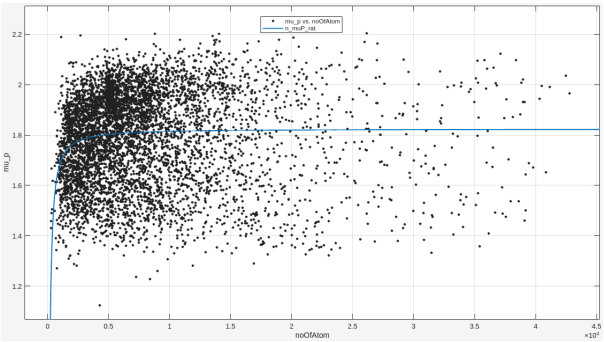


(d) n_s

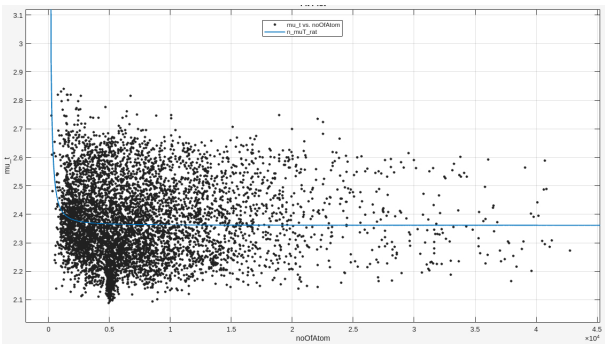
Figure S33: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Mus musculus* proteome**.



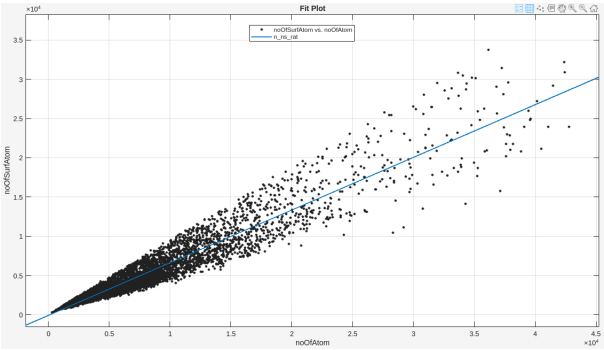
(a) μ_s



(b) μ_p

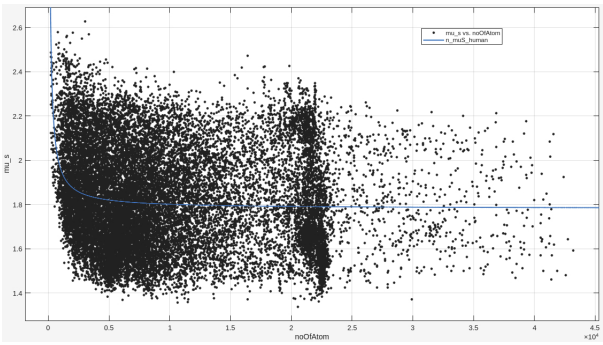


(c) μ_t

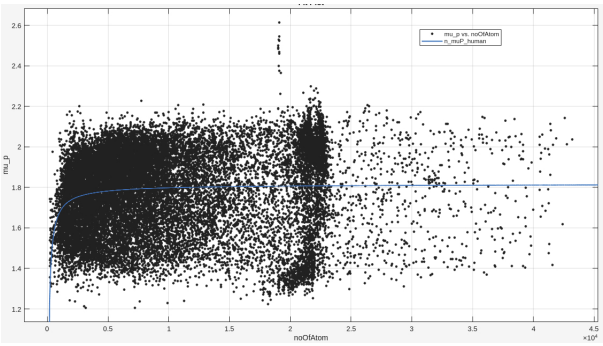


(d) n_s

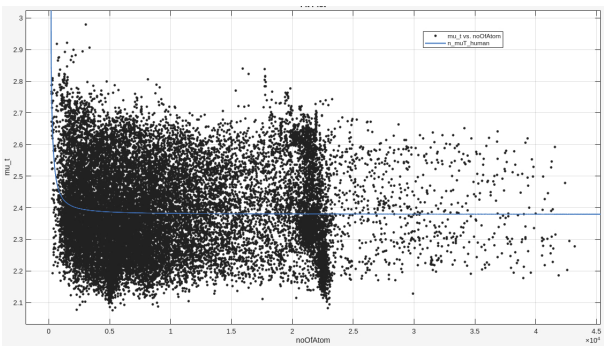
Figure S34: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Rattus norvegicus* proteome**.



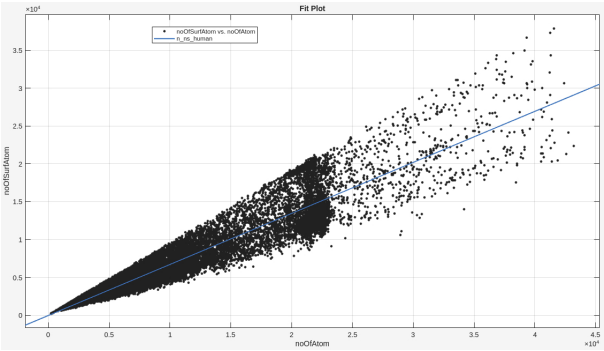
(a) μ_s



(b) μ_p



(c) μ_t



(d) n_s

Figure S35: **The changes with n of the μ_s, μ_p, μ_ts and n_ss for the *Homo sapiens* proteome**.

S12: The biological names of the twenty seven species and their abbreviations

Table S3 lists the biological names of the twenty seven species and their abbreviations used in Table 2 of the main text and Tables S1 and S2, and Fig. S5.

Species	Abbreviation
Escherichia coli	ecoli
Haemophilus influenzae	haeIn
Helicobacter pylori	helPy
Klebsiella pneumoniae	klePh
Methanocaldococcus jannaschii	metJa
Mycobacterium leprae	mycLe
Mycobacterium tuberculosis	mycTu
Neisseria gonorrhoeae	neiG1
Pseudomonas aeruginosa	pseAr
Salmonella typhimurium	salTy
Shigella dysenteriae	shiDs
Staphylococcus aureus	staA8
Streptococcus pneumoniae	strR6
Saccharomyces cerevisiae	yeast
Schizosaccharomyces pombe	schPo
Candida albicans	canals
Arabidopsis thaliana	araTh
Zea mays	maize
Oryza sativa	orySj
Glycine max	soyBn
Caenorhabditis elegans	celEg
Dictyostelium discoideum	dicDi
Danio rerio	danRe
Drosophila melanogaster	droMe
Mus musculus	mouse
Rattus norvegicus	rat
Homo sapiens	human

Table S3: **The name of the twenty seven species and their abbreviations used in Table 2 of the main text and Tables S1 and S2 and Fig. S5**.

S13: An example of the interface formed by crystal packing

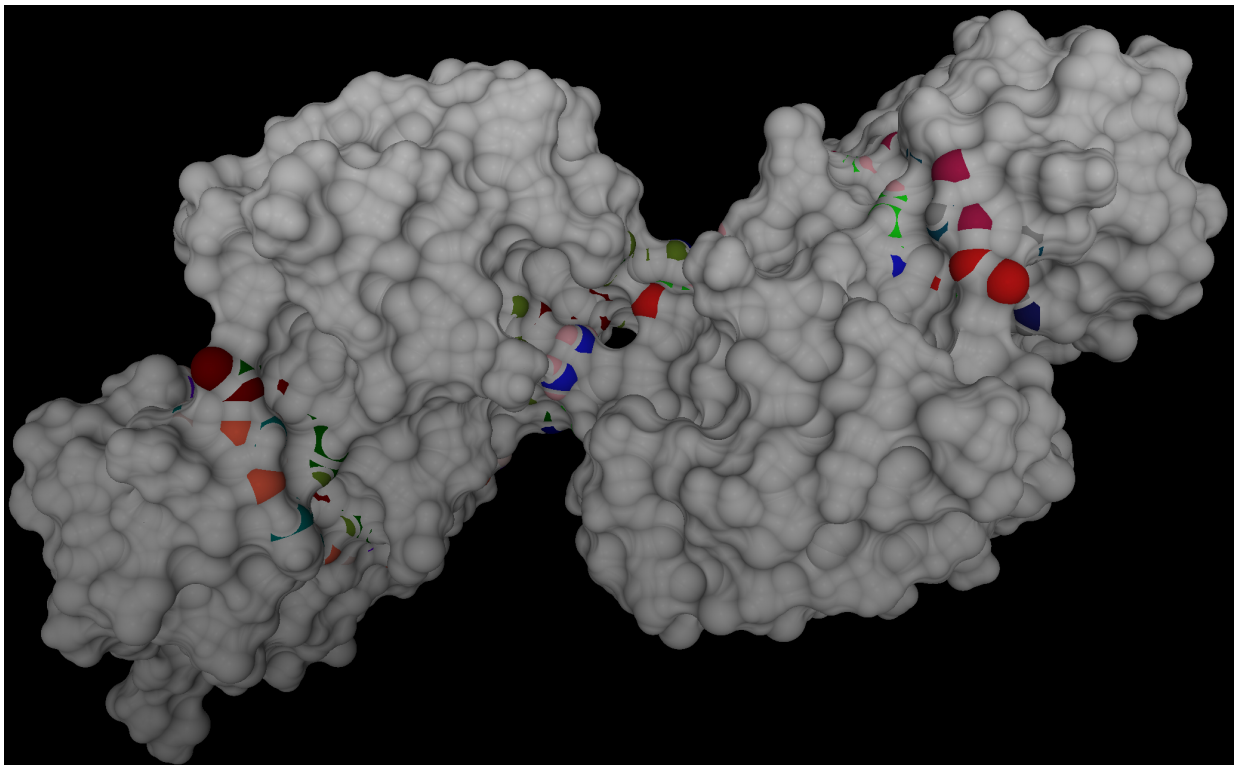


Figure S36: **The crystal packing between barnase-barstar complex**. The two PPI interfaces of barnase barstar complex are in the right and the left while the interface formed by crystal packing is in the middle. The SAS polygons of the interface atoms are depicted in bright colors while the rest of the SES in gray. As illustrated here, the interface of crystal packing is relatively small and thus its contribution to the geometric properties studied in this paper is minor. The figure is prepared by our program SESA which is freely available at https://github.com/wlincong/SESexact.

S14: The pdbids of the three sets of crystal structures

The file "pdbidOfGSM.tar.gz" includes three files that list respectively the pdbids for $\mathbf{G}$, $\mathbf{S}$ and $\mathbf{M}$.