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

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

Lincong Wang*

Abstract

Previous analyses of protein structures have focused primarily on their secondary structures, three-dimensional folds, 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 computed and 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 the average areas for three SES patch types each follows a power law with respect to the number of atoms n in a structure. Specifically, SAS area per atom μ_s decreases while probe area per atom μ_p increases with n. Most interestingly, the fitted power laws for μ_s intersect, respectively, with those for μ_p at, $n = 956, n = 934$ and $n = 1,024$ for $\mathbf{G}, \mathbf{S}$ and $\mathbf{M}$. They correspond approximately to $58, 60$ and 68 amino acid residues. Together they 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 predicted models for twenty seven proteomes. The analyses show that the μs for thirteen prokaryotic proteomes are similar to those for the experimentally-derived structures while the μs for fourteen eukaryotic proteomes differ largely from them. The differences may have significant implications 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 not adequate for, say, structure-based drug design.

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 secondary structure content, three-dimensional (3D)[1] folds, and the physical-chemical and geometric properties of ligand-protein interfaces. By comparison, protein surface has not been as well characterized. 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

[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; PDB, Protein Data Bank; 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 experimentally-derived structures the areas per atom for SAS* μ_s*, probe* μ_p *and toroidal patches* μ_t *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* $n = 956$ *for* $\mathbf{G}$*,* $n = 934$ *for* $\mathbf{S}$ *and* $n = 1,024$ *for* $\mathbf{M}$*. They correspond respectively to* $58, 60$ *and* 68 *residues. A SAS patch is convex while a probe one concave, thus the power laws for* μ_s *intersect with those 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 they 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* μ*s of the AlphaFold models for thirteen prokaryotic proteomes are similar to those for experimental structures. However, in stark contrast with the experimental structures and the models for prokaryotic proteomes, the* μ*s of the models for fourteen eukaryotic proteomes differ largely from them. The differences have significant implications for both theory 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 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. 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 $n = 956$ for $\mathbf{G}$, $n = 934$ for $\mathbf{S}$ and $n = 1,024$ for $\mathbf{M}$. They correspond, respectively, to $58, 60$ and 68 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 is equal to that with a positive curvature if one ignores the toroidal patches each of which includes regions of both 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 < 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

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

closely related to the structural stability of proteins and protein folding in the sense that $\gamma > 1$ is a necessary condition for protein folding.

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, TM score, electron density map [18], and other criteria [19, 20, 21]. 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 of experimentally-derived structures while the SESs for the latter differ largely from them. Specifically, the μs[3] of the models for the eukaryotic proteomes are broadly distributed, hardly change and only minimally follow a power law with n. The differences may have significant implications 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 not adequate for, say, structure-based drug design.

2 Materials and Methods

In this section we first describe the three sets of crystal structures and the AlphaFold models for twenty seven 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 fitting to power law of μs with n.

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 $4,780$ structures of monomeric protein with a resolution range of $0.64 - 3.50$Å and mean of 1.71Å, a R_{free} range of $0.075-0.350$ and mean of 0.216. $\mathbf{G}$ has $12,228$ structures with a resolution range of $0.68-3.5$Å and mean of 1.98Å, a R_{free} range of $0.078 - 0.350$ and mean of 0.228. Neither $\mathbf{S}$ nor $\mathbf{G}$ includes any membrane proteins. $\mathbf{M}$ has $4,504$ structures of membrane protein with a resolution range of $0.88 - 9.8$Å and mean of 3.01Å, a R_{free} range of $0.107 - 0.486$ and mean of 0.260.

From AlphaFold Protein Structure Database [22] we downloaded $350,974$ models for a total of 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 consisting of two yeasts, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*; four plants, *Arabidopsis thaliana, Glycine max, Oryza sativa* and *Zea mays*; and amoeba, *Dictyostelium discoideum*; fungus, *Candida albicans*; insect, *Drosophila melanogaster*; worm, *Caenorhabditis elegans*; fish, *Danio rerio*; and three mammals, *Mus musculus, Rattus norvegicus* and *Homo sapiens*.

The PDB files are preprocessed for area computation as described previously [23]. 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 selected. Protons are added using REDUCE [24] to any PDB structure that lacks them. Only protein atoms are used for area computation.

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

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

The program for analytic SES area computation, SESA[4], 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 algorithmic treatments of probe-probe intersections. 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 model. Specifically we analyze the SAS, probe and toroidal areas per (exterior) surface atom defined as follows.

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 SAS, toroidal and probe areas per atom, μ_s, μ_t and μ_p, are defined 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, A_s, A_t and A_p are, respectively, the total SES, SAS, toroidal and probe areas. The summations are over n_s. SAS area A_s, as a part of A, 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 the μ_s, μ_t and μ_p values for the exterior surface and n_s of a molecule.*

2.3 Fitting to power law

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

$$y = ax^b + c \tag{5}$$

The changes of μs with n, the fitted power laws and their intersections are the main topics of this paper.

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

3 Results

In this section we describe the changes of μs with n, the fitted power laws for $\mathbf{G}, \mathbf{S}$ and $\mathbf{M}$, and twenty seven AlphaFold predicted proteomes, and the intersections of the power laws for μ_s and μ_p. In addition, we briefly mention the intersections of the power laws for μ_t and μ_p.

3.1 The changes of μs with n and the fitted power laws and their intersections for $\mathbf{G}$

$\mathbf{G}$ has $12,228$ 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 having a larger slope while μ_p increases with n. The fitted power laws for μ_s and μ_p are respectively

$$\mu_s = 7.2159n^{-0.3722} + 1.2298 \tag{6}$$

$$\mu_p = -3.1320n^{-0.2410} + 2.3847 \tag{7}$$

They intersect at $n = 965$ (58 residues using the fitted linear equation (Fig. S1 of SI), $n_r = 0.0647n - 4.3341$, where n_r is the number of residues in a protein). In addition, the power law for μ_p intersects with that for μ_t at $n = 48,659$ ($3,144$ residues). However, the fitting to μ_t is worse (lower R_{square}) than that to either μ_s or μ_p (Fig. 1a).

3.2 The changes of μs with n and the fitted power laws and their intersections for $\mathbf{S}$

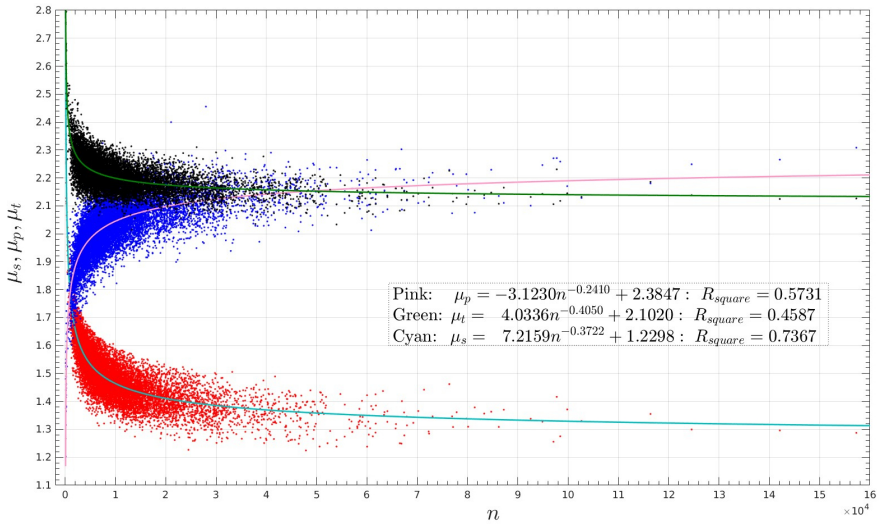
$\mathbf{S}$ has $4,780$ monomers. The smallest and largest ones are respectively 4A4J with 69 residues ($1,013$ atoms), and 7BY5 with $1,286$ residues ($20,734$ atoms). As shown in Fig. 1b and as with $\mathbf{S}$, both μ_s and μ_t decrease with n with the former having a larger slope while μ_p increases with n. The fitted power laws for μ_s and μ_p are, respectively, $\mu_s = 12.0335n^{-0.4716} + 1.3011$ and $\mu_p = -2.5619n^{-0.1299} + 2.8326$. They intersect at $n = 934$. Using the fitted linear equation between n_r and n, $n_r = 0.0643n + 0.4054$, it equals 60 residues, a purely theoretical number since all the monomers have $n_r \geq 69$ residues. Furthermore, the differences in μs between $\mathbf{S}$ and $\mathbf{G}$ are very small (Table 1).

Set	$\Delta\mu_s$ (mean / std)	$\Delta\mu_p$ (mean / std)	R_{square} (μ_s/μ_p)	$n_r : \mu_s = \mu_p$
G	0.0 / 0.0	0.0 / 0.0	*0.737 / 0.573*	*58*
S	-0.0170 / 0.0520	0.0005 / 0.0567	0.608 / 0.533	60
M	-0.0172 / 0.0719	0.0172 / 0.0882	0.628 / 0.440	68

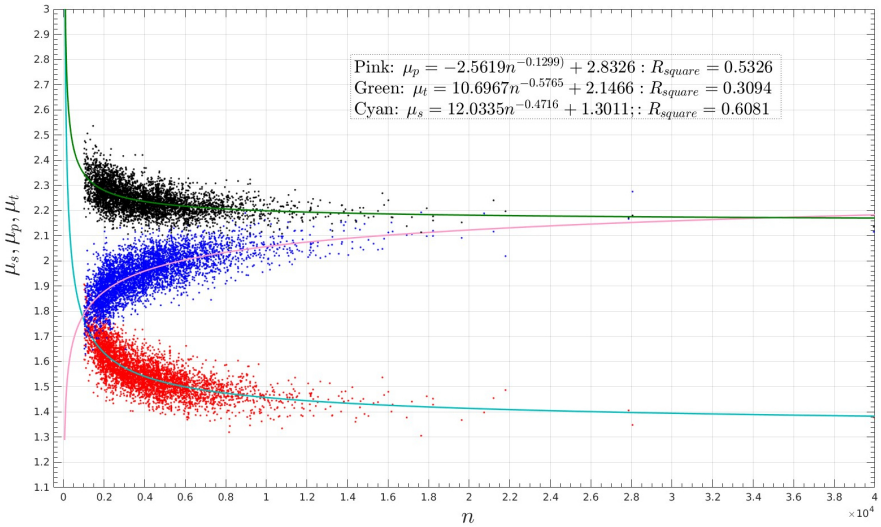
Table 1: **The differences in μs between $\mathbf{G}$ and $\mathbf{S}$, and between $\mathbf{G}$ and $\mathbf{M}$**. Column 1 lists set names. The differences in μ_s and μ_p of structure i of either $\mathbf{S}$ or $\mathbf{M}$ from the μ_ss and μ_ps for $\mathbf{G}$ are quantified as $\Delta\mu_s(i) = \mu_s(i) - \mu_s^g(i)$ and $\Delta\mu_p(i) = \mu_p(i) - \mu_p^g(i)$ where $\mu_s(i), \mu_p(i)$ are the μ_s, μ_p values for structure 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 in structure i. Column 2 lists the mean and standard deviation of the list of all the $\Delta\mu_s$s for a set, and column 3 lists the mean and standard deviation of the list of all the $\Delta\mu_p$s for a set. Column 4 lists the MATLAB R_{square}s for μ_s and μ_p. Column 5 lists the estimated numbers of residues n_r when $\mu_s(n) = \mu_p(n)$, i.e. when the power laws for μ_s and μ_p intersect.

3.3 The changes of μs with n and the fitted power laws and their intersections for $\mathbf{M}$

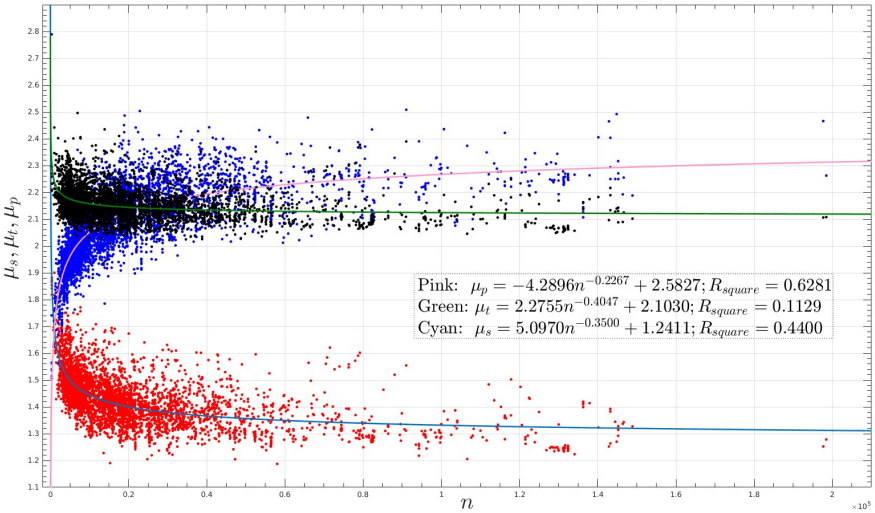
$\mathbf{M}$ has $4,504$ structures. The smallest and the largest ones are respectively 5N9I with 24 residues (328 atoms), and 3OAA with $13,028$ residues ($198,458$ atoms). As shown in Fig. 1c and as with $\mathbf{S}$ and $\mathbf{G}$, both μ_s and μ_t decrease with n with the former having a larger slope while μ_p increases with n. The fitted power laws for μ_s and μ_p are, respectively, $\mu_s = 5.0970n^{-0.3500} + 1.2411$ and $\mu_p = -4.2896n^{-0.2267} + 2.5827$. They intersect at $n = 1,024$ (68 residues assuming that there are 15 atoms per residue (Fig. S1) considering



(a) **G**



(b) **S**



(c) **M**

Figure 1: **The μ_ss, μ_ts and μ_ps of the experimentally-derived structures and their fitted power laws**. Figs. (**a, b**) and (**c**) depict respectively the μs and power laws for **G**, **S** and **M**. The μ_ss, μ_ts and μ_ps are colored respectively in red, black and blue with the fitted curves colored respectively in cyan, green and pink. The x-axes are n while the y-axes are μ_s, μ_t and μ_p values. The inserts list the fitted power laws and MATLAB fitting scores R_{square}s.

that many structures have missing atoms). Furthermore, the differences in μs between $\mathbf{M}$ and $\mathbf{G}$ are small (Table 1).

3.4 The changes of μs with n and the fitted power laws and their intersections for AlphaFold models

The changes of the μs with n 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 models for *E.coli* as a representative of the thirteen prokaryotic proteomes and the models for Homo Sapiens as a representative of the fourteen eukaryotic proteomes. Please refer to Table 2 for the points of intersection and R_{square}s, and Fig. 3 for the fitted power laws for the twenty seven proteomes.

AlphaFold predicted $4,363$ models for the *E.coli* proteome. The smallest and largest ones are, respectively, 257 atoms (16 residues), and $34,552$ atoms ($2,358$ residues). As shown in Figs. 2d, 2e and as with $\mathbf{S}$, $\mathbf{G}$ and $\mathbf{M}$ (Figs. 1, 2a and 2b), both the μ_ss and μ_ts decrease with n with the former having a larger slope (Figs. S3a and S3c) while the μ_ps increase with n. The fitted 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 $n = 1,393$ (87 residues). 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 number of residues in any *E.coli* protein.

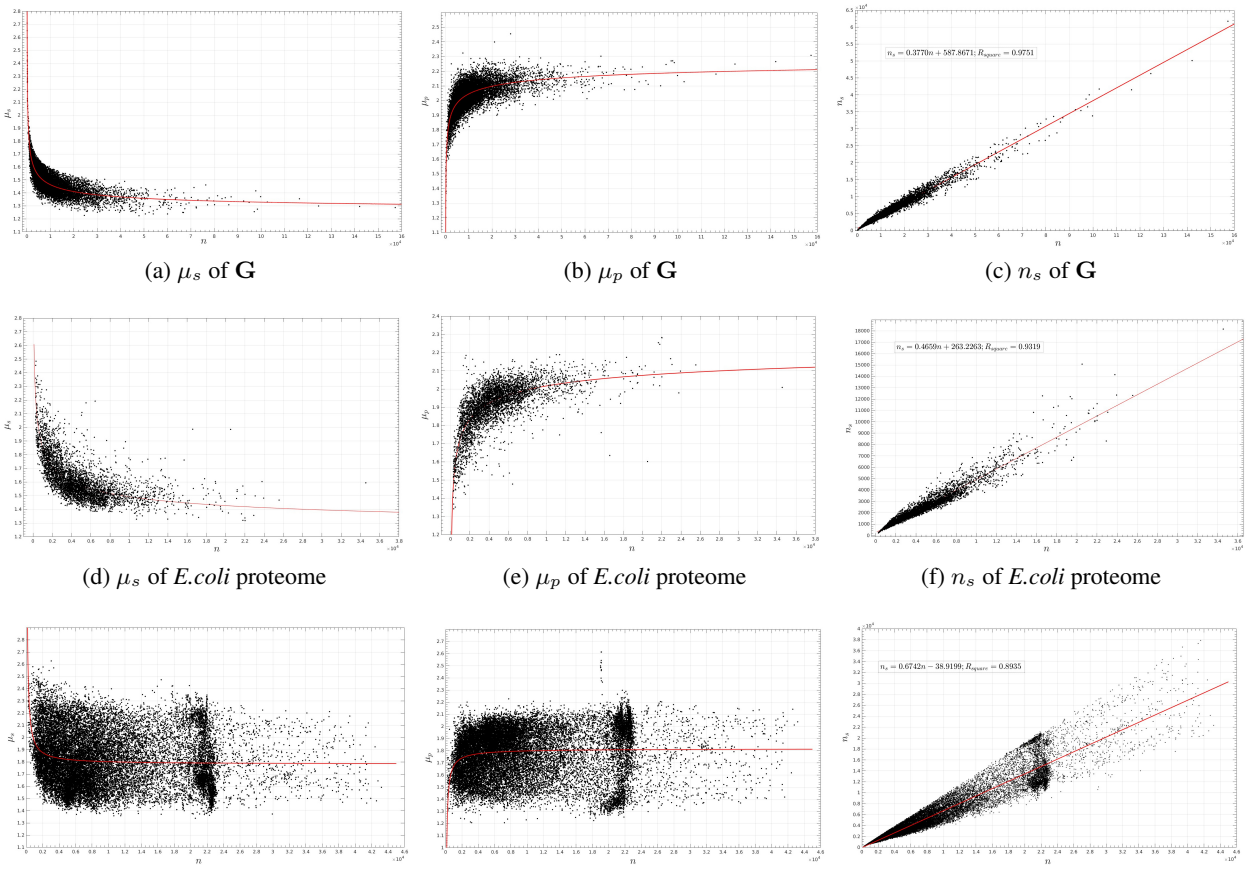
AlphaFold predicted $23,391$ models for the human proteome. The smallest and largest ones are, respectively, 238 atoms (16 residues) and $43,190$ atoms ($2,699$ residues). As shown in Figs. 2g and 2h, in stark contrast to the μs for both the experimental structures (Figs. 1, 2a and 2b) and the models for the thirteen prokaryotic proteomes (Figs. 2d and 2e and Figs. S3–S15), the μs for the fourteen eukaryotic proteomes (Figs. S16–S29) differ largely from both of them in the sense that when plotted with μ as y-axis and with n as x-axis, the μs are broadly distributed and hardly change with n. The changes of the μs with n only minimally follow a power law with a $R_{square} = 0.034$ for μ_s and $R_{square} = 0.025$ for μ_p. The power laws for μ_s and μ_p intersect at $n = 11,100$ (694 residues). However, the power law for μ_p *does not* intersect with that for μ_t. Furthermore, the slope of the fitted line (Fig. 2i), $n_s = 0.6742n - 38.9199$, is 0.6742, 1.8-fold larger than that for $\mathbf{G}$ (0.3770, Fig. 2c insert). The broad distributions in μs, the power laws and their intersections, the large slope for the fitted line between n_s and n, and low R_{square}s (Table 2) taken together show that there exist large differences in SES between the experimental structures and the predicted models for eukaryotic proteomes.

4 Discussion

We start with the previous analyses of experimentally-derived protein structures and their surfaces, and follow up with the molecular surfaces of both experimental structures and AlphaFold models focusing on the changes of their μs with molecular size, the fitted power laws and their intersections. Finally we list 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 secondary structures, folds, binding or active sites and structural qualities [25], 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



(a) μ_s of **G** (b) μ_p of **G** (c) n_s of **G**

(d) μ_s of *E.coli* proteome (e) μ_p of *E.coli* proteome (f) n_s of *E.coli* proteome

(g) μ_s of human proteome (h) μ_p of human proteome (i) n_s of human proteome

Figure 2: **The changes of μ_s, μ_p and n_s with n for G, and *E.coli* and human proteomes**. The fitted power laws (Figs. (**a, b, d, e, g, h**)) and lines (Figs. (**c, f, i**)) are depicted in red. The inserts in Figs. (**c, f, i**) list the fitted lines and their R_{square}s. The x-axes are n. The y-axes are, respectively, the μ_s (Figs. (**a, d, g**)), μ_p (Figs. (**b, e, h**)) and n_s (Figs. (**c, f, i**)) in a structure or model.

function [26], 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 types have been studied *separately and at atom-level* for their possible relevance to structural stability, folding and protein function.

4.2 The geometric properties of the molecular surfaces of experimental structures

We have analyzed the changes of total SES area, and areas for each of the three SES patch types with the numbers of surface atoms n_s, total atoms n and residues n_r, as well as secondary structure content and surface atom types. Among their geometric properties we found that the changes of μs, particularly μ_ss and μ_ps, with n are most unique (Figs. 1, 2, 3) and thus likely most relevant to protein structural stability, folding and function. The μs for experimental structures are fitted to n slightly better than to either n_s or n_r (Fig. S2), though there exist excellent linear relationships among them (Fig. 2c and Fig. S1). For example, the number of atoms per residue, i.e. 16 atoms per residue, could be determined accurately from the linear correlation between n_r and n for **G** (Fig. S1). The fitting to n is better than that to n_r likely because of the existence of missing atoms in many experimental structures, and to n_s likely because of the diversity in SES among the structures. Furthermore, the μs only change 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 of μs with n.

4.2.1 The characteristic changes of the μs with n and their possible meanings

The characteristic changes of the μs with n could be (1) a purely global geometrical feature, (2) the optimization of solvent-protein interaction, and (3) the minimization of misinteraction, and (4) the artifacts of crystal packing. We will not discuss the last possibility any further but will mention it later as a possible limitation of this study.

4.2.1.1 The change of the μs with n is unlikely a purely 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 intersections of their fitted power laws at ≤ 68 residues. The two power laws and their intersections mean 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 gradually larger than that of positive curvature. Thus the results show that both μ_s and μ_p are local geometric property of SES, and as a local property they should not change with molecular size.

4.2.1.2 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 water-soluble proteins or both the lipids and aqueous solvent for membrane proteins. In other words, the increase in μ_p suggests that the disruption to a solvent's intrinsic structures is minimized. For a water-soluble protein it means that the disruption to water hydrogen-bond network is minimized, the entropic increase of the whole system minimized, and the intermolecular hydrogen bonding and VDW interaction optimized. For a membrane protein, the disruption to the structural integrity of the membrane is minimized.

4.2.1.3 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 [27]. 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 for proteins to have a molecular surface with an area-weighted average curvature to be negative. 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 surface hydrophobic side chains, and that in turn reduces the chance for two proteins sticking together and for the formation of a protein inclusion body inside a cell.

4.2.2 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 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 [28] based on a purely 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 fitted power laws intersect at 68 residues. Furthermore, if the sequester of hydrophobic residues is the only driving force, then after the point of intersection, μ_s should stop decreasing and μ_p stop increasing. Thus the sequester of hydrophobic residues could only provide a partial explanation.

The power laws for μ_s intersect with those for μ_p at $n_r = 58$ for $\mathbf{G}$ and $n_r = 60$ for $\mathbf{S}$, close to the number of residues required for a polypeptide to adopt a dominant conformation in solution. The two values suggest 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 property of the SES for experimental structures provides a plausible explanation for the transition from peptide to protein. In addition, the points of intersection suggest that the ratio between the total probe area and total SAS area, $\gamma = \frac{\mu_p}{\mu_s}$, is an important parameter for describing 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 [29] and protein folding in the sense that $\gamma > 1$ is a necessary condition for protein folding.

4.3 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 the accuracy of mainly backbone atoms and all atoms, and overall structural qualities [20, 25]. To objectively judge AlphaFold's accuracy, some models have been compared directly with experimental crystallographic maps [18]. In addition, the topology of the models has been evaluated [30]. However, to our best knowledge, neither their SASs nor SESs have been analyzed.

4.3.1 The similarity in SES between the experimental structures and AlphaFold models for prokaryotic proteomes

The changes of the μs with n of the AlphaFold models for the thirteen prokaryotic proteomes resemble those for the crystal structures (Table 2, Figs. 2, 3 and S3–S15). Among them the μs for *E.coli* proteome are one

of the closest to those for the experimental ones (Figs. 2d–2f and S3). The sensitivity of neural-network to training data [31] suggests that the similarity likely results from (1) the abundance of sequenced prokaryotic genomes, (2) the similarity among prokaryotic genomes [32], (3) the high experimental coverage of at least one of the prokaryotic proteomes [32], (4) prokaryotic proteins are generally smaller in size and simpler in structure than their counterparts in eukaryotic proteomes, and (5) the relatively low content of unstructured regions in prokaryotic proteomes [33]. In other words, the abundance of high quality training data somehow compensates for the lack of surface features for AlphaFold model training.

Organism	$\Delta\mu_s$ (mean / std)	$\Delta\mu_p$ (mean / std)	R_{square} (μ_s/μ_p)	$n_r : \mu_s = \mu_p$
ecoli	*0.0455 / 0.1010*	*-0.0404 / 0.0825*	*0.582 / 0.571*	*87*
haeIn	0.0447 / 0.0994	-0.0366 / 0.0811	0.487 / 0.501	86
helPy	0.0552 / 0.1119	-0.0456 / 0.0894	0.499 / 0.523	89
klePh	0.0499 / 0.1047	-0.0501 / 0.0873	0.629 / 0.582	92
metJa	0.0252 / 0.0850	-0.0204 / 0.0724	0.487 / 0.488	75
mycLe	0.0903 / 0.1361	-0.0693 / 0.1090	0.372 / 0.331	110
mycTu	0.0802 / 0.1390	-0.0799 / 0.1099	0.369 / 0.379	113
neiG1	0.0703 / 0.1275	0.0559 / 0.1027	0.556 / 0.482	99
pseAr	0.0472 / 0.1117	-0.0568 / 0.0858	0.404 / 0.414	93
salTy	0.0511 / 0.1045	-0.0431 / 0.0846	0.504 / 0.457	91
shiDs	0.0545 / 0.1067	-0.0478 / 0.0883	0.509 / 0.503	93
staA8	0.0495 / 0.1121	-0.0284 / 0.0914	0.516 / 0.546	75
strR6	0.0491 / 0.0998	-0.0344 / 0.0837	0.505 / 0.503	81
yeast	0.2242 / 0.2065	-0.1410 / 0.1645	0.056 / 0.083	133
schPo	0.1989 / 0.1978	-0.1265 / 0.1600	0.056 / 0.086	125
araTh	0.2484 / 0.2087	-0.1557 / 0.1698	0.125 / 0.123	246
maize	0.2845 / 0.2141	-0.1890 / 0.1763	0.104 / 0.094	389
orySj	0.3017 / 0.2250	-0.1944 / 0.1854	0.206 / 0.172	391
soyBn	0.2499 / 0.2113	-0.1554 / 0.1754	0.093 / 0.081	250
celEg	0.2148 / 0.2068	-0.1285 / 0.1636	0.126 / 0.123	156
canAl	0.2307 / 0.2131	-0.1392 / 0.1654	0.036 / 0.072	145
dicDi	0.2503 / 0.2128	-0.1295 / 0.1805	0.092 / 0.085	116
danRe	0.2756 / 0.2155	-0.1837 / 0.1856	0.023 / 0.018	208
droMe	0.2830 / 0.2338	-0.1926 / 0.1944	0.037 / 0.040	327
mouse	0.2745 / 0.2292	-0.1961 / 0.1908	0.028 / 0.023	193
rat	0.2710 / 0.2274	-0.1940 / 0.1892	0.030 / 0.025	258
human	*0.3130 / 0.2289*	*-0.2261 / 0.2023*	*0.034 / 0.025*	*694*

Table 2: **The differences in SES between the** $350,974$ **AlphaFold models for twenty seven proteomes and experimental structures**. Column 1 lists organism names in abbreviation with their species names listed in Table S1. The differences in μ_s and μ_p of model i of a proteome from the μ_ss and μ_ps of **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 in a proteome, 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 in model i. Column 2 lists the mean and standard deviation of the list of all the $\Delta\mu_s$s for a proteome, and column 3 lists the mean and standard deviation of the list of all the $\Delta\mu_p$s for a proteome. Column 4 lists the R_{square}s for μ_s and μ_p. Column 5 lists the estimated numbers of residues n_r when $\mu_s(n) = \mu_p(n)$. The n_rs are estimated assuming that there are 16 atoms per residue considering that there are no missing atoms in an AlphaFold model (Fig. S1). Among the twenty seven proteomes, the human proteome has the largest point of intersection with an $n_r = 694$.

4.3.2 The large differences in SES between the experimental structures and AlphaFold models for eukaryotic proteomes

As illustrated in Figs. 2g–2i and shown in Table 2, Figs. 3 and S16–S29, the μs of the AlphaFold models for the fourteen eukaryotic proteomes are broadly distributed, and hardly change and only minimally follow a power law with n. Among them the μs of the models for human proteome (Figs. 2g–2i) are the furthest from those for experimental structures. However, the number of surface atoms n_s increases linearly with n though in an artificial manner in that the range of the n_ss also increases linearly with n (Figs. 2i and S16–S29). The large differences in SES between the AlphaFold models and experimental structures suggest that either the AlphaFold models have defects or there exist genuine differences between prokaryotic and eukaryotic proteomes. Such differences, if exist, are due possibly to the likelihood that a eukaryotic proteome has a higher percentage of unstructured regions than a prokaryotic one. It is estimated that on average a eukaryotic protein contains 32% disordered residues while a prokaryotic one only has 12% [33]. The SES of an unstructured region is similar to the SES of a typical peptide with a $\gamma < 1.0$. However, the 20% difference unlikely to be large enough to rationalize the large differences in SES as shown in Table 2 and Figs. S16–S29, even though it has been shown that the unstructured regions of several AlphaFold models agree well with experimental results [21]. Furthermore, since AlphaFold is trained using the currently available crystal structures and there remain linear relationships between n_s and n for all the twenty seven proteomes (Figs. S3–S29), out of the two possibilities, we believe that the differences are likely attributed to the lack of surface features in model training, which in turn leads to the defects in the molecular surfaces for the predicted models.

4.4 Possible limitations

The SES is computed by modeling both the probe and protein atoms as hard spheres. It means that SES itself is only an approximation to the real physical surface of a protein. Consequently both the power laws and their intersections are only an approximation to a likely complicated physical relationship between protein surface and molecular size. In addition, we could not completely exclude the possibility that both the power laws and their intersections originate largely from crystal packing.

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.

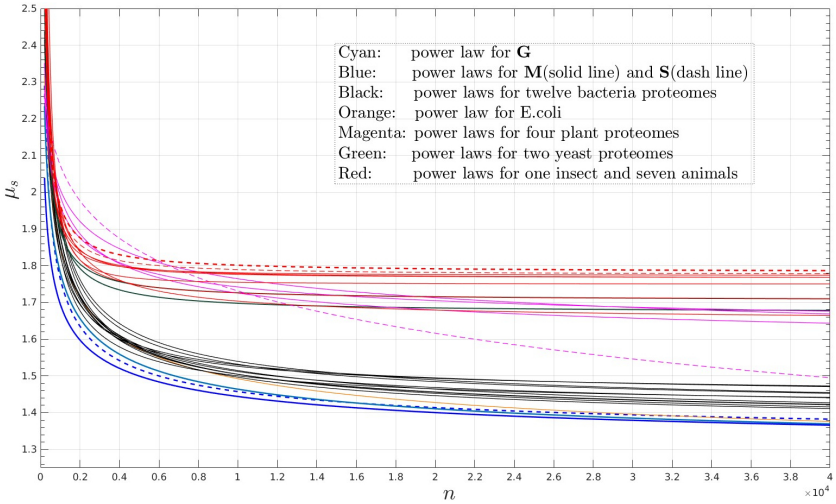
[4] D. Eisenberg and A. D. Andrew D. McLachlan. Solvation energy in protein folding and binding. *Nature*, 319:199–203, 1986.

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

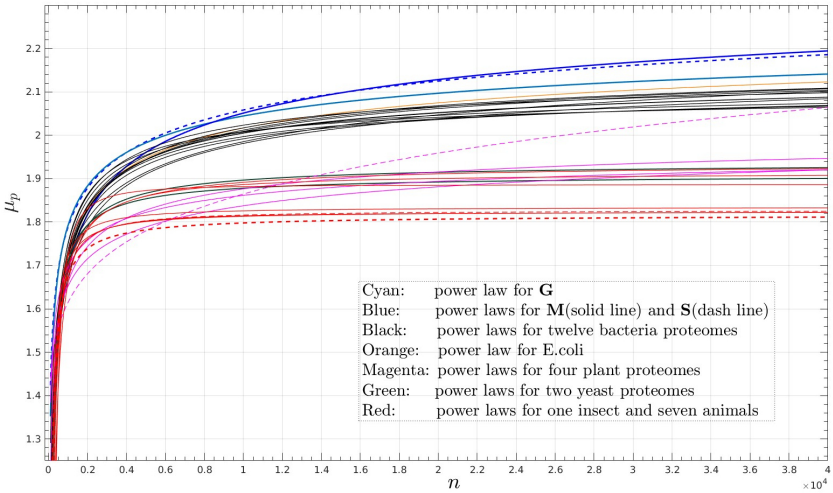
[6] T. Y. Lin and S. N. Timasheff. On the role of surface tension in the stabilization of globular proteins. *Protein science*, 5(2):372–381, 1996.

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

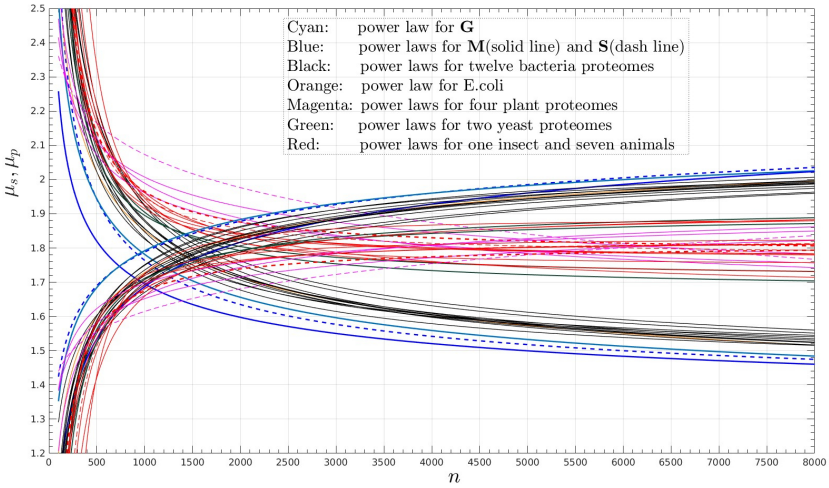
[8] M. Bellissent-Funel, A. Hassanali, M. Havenith, R. Henchman, P. Pohl, F. Sterpone, D. van der Spoel, Y. Xu, and A. E Garcia. Water determines the structure and dynamics of proteins. *Chemical Reviews*, 116(13):7673–7697, 2016. PMID: 27186992.



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



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



(c) The intersections of the fitted power laws for μ_s and μ_p

Figure 3: **The fitted power laws and their points of intersection for $\mathbf{G}, \mathbf{S}, \mathbf{M}$ and the AlphaFold models for twenty seven proteomes**. The x-axis is n while the y-axes are μ_s (**a**), μ_p (**b**) and μ_s, μ_p (**c**).

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

[20] Roberto Tejero, Yuanpeng Janet Huang, Theresa A. Ramelot, and Gaetano T. Montelione. Alphafold models of small proteins rival the accuracy of solution nmr structures. *Frontiers in Molecular Biosciences*, 9, 2022.

[21] 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.

[22] 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.

[23] 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.

[24] J.Michael Word, Simon C. Lovell, Jane S. Richardson, and David 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.

[25] Aneerban Bhattacharya, Roberto Tejero, and Gaetano T. Montelione. Evaluating protein structures determined by structural genomics consortia. *Proteins: Structure, Function, and Bioinformatics*, 66(4):778–795, 2007.

[26] 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.

[27] 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.

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

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

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

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

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

[33] Walter Basile, Marco Salvatore, Claudio Bassot, and Arne 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 change of the number of residues n_r with the number of atoms n for **G** , and the best-fitted linear equation between n_r and n. Section S2 describes the changes of μs with both n_r and the number of surface atoms n_s for **G**. Section S3 presents the changes of μs and n_s with n of the $350,974$ AlphaFold models for twenty seven proteomes.

S1: The change of n_r with n for G

As shown in Fig. S1, the change of n_r with n for **G** could be fitted to a linear equation with a MATLAB $R_{square} = 0.9981$. Because many structures in **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 μs with n_r (Fig. S2).

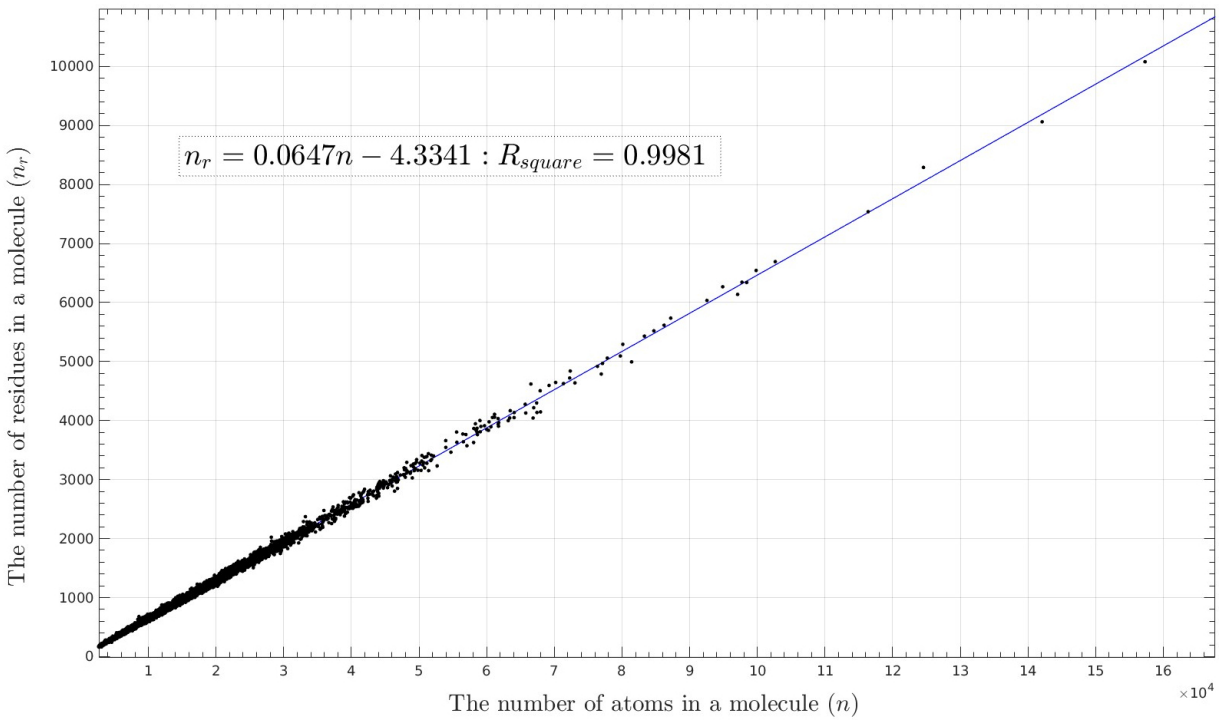
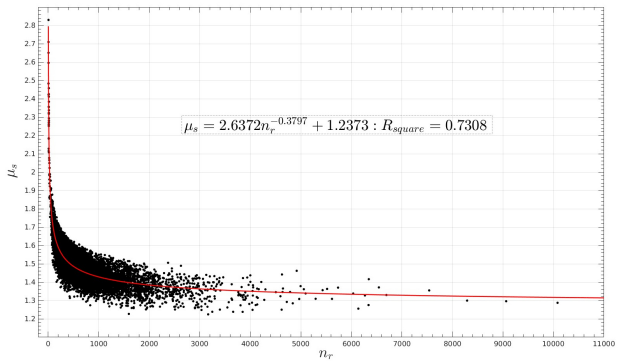


Figure S1: **The change of n_r with n for G**. From the slope of the best-fitted line we determine the number of atoms per residue as **16** considering that the y-intercept is a negative number. The insert is the MATLAB fitted linear equation, $n_r = 0.0647n - 4.3341$, with a $R_{square} = 0.9981$. As described in the main text, the equation is used to compute n_r from n for the points of intersection between two power laws for **G**. The number of atoms per residue, **16**, is used to estimate n_r from the point of intersection in n for the twenty seven AlphaFold predicted proteomes (Table 2 of the main text).

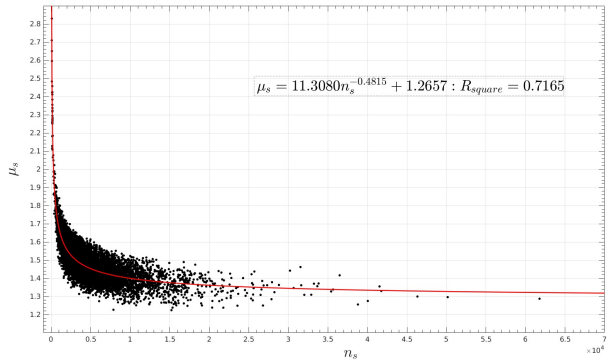
S2: The changes of μs with n_r and n_s for G

Fig. S2 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 with the changes with n, the R_{square}s for fitting the changes with n_r and n_s are all lower, except for the change of μ_t with n_s which has a slightly higher R_{square}. The low R_{square} scores for fitting

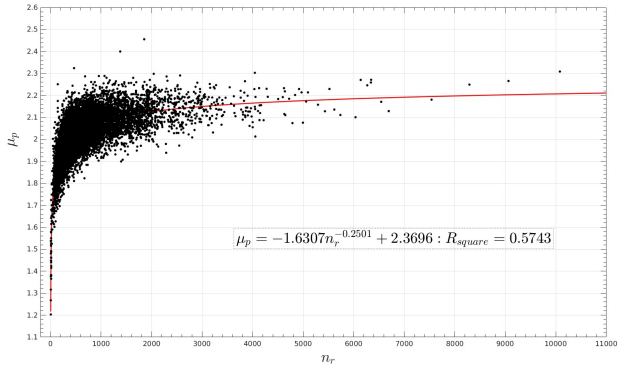
the changes with n_r and n_s are due likely to fact that many crystal structures in $\mathbf{G}$ have some missing atoms.



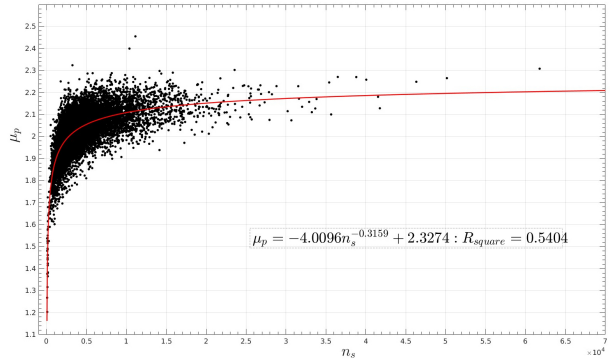
(a) μ_s vs n_r



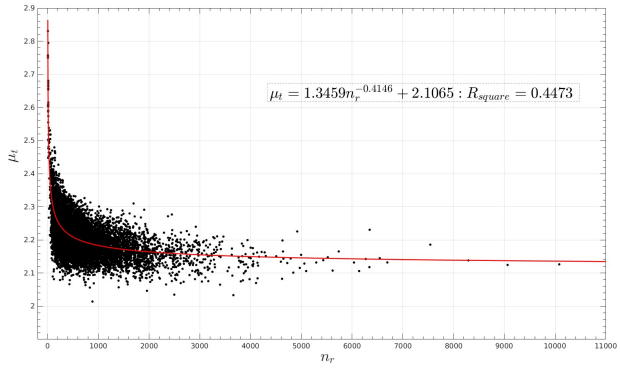
(b) μ_s vs n_s



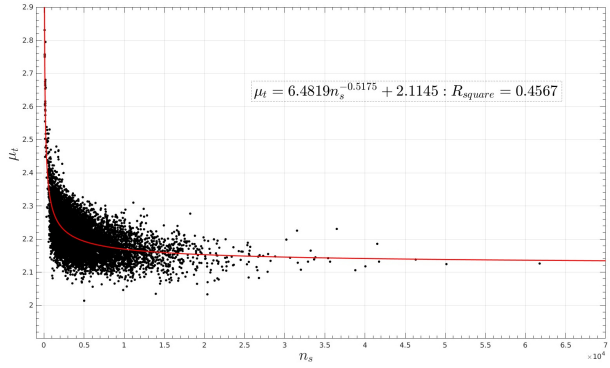
(c) μ_p vs n_r



(d) μ_p vs n_s



(e) μ_t vs n_r



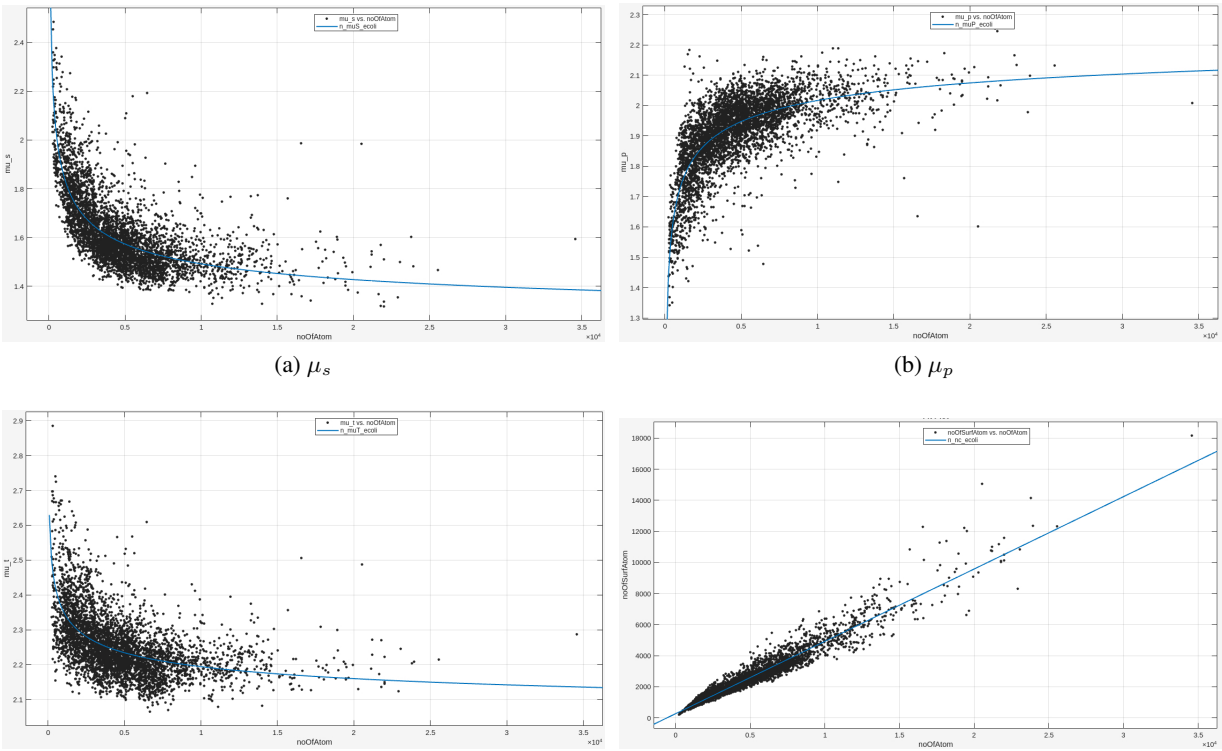
(f) μ_t vs n_s

Figure S2: **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 are the fitted power laws and fitting scores R_{square}s. The curves for the power laws 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 58 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 their upper limits.

S3: The changes of μs and n_s with n for the twenty seven AlphaFold proteomes

The figures S3 to S29 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 prokaryotes (Figs. (S3–S15)), two yeasts (Figs. S16 and S17), four plants (Figs. (S18–S21)) and eight animals (Figs. (S22–S29)).

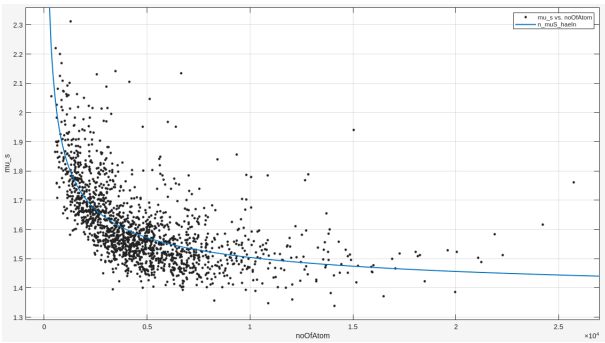
S3.1: The thirteen prokaryotic proteomes



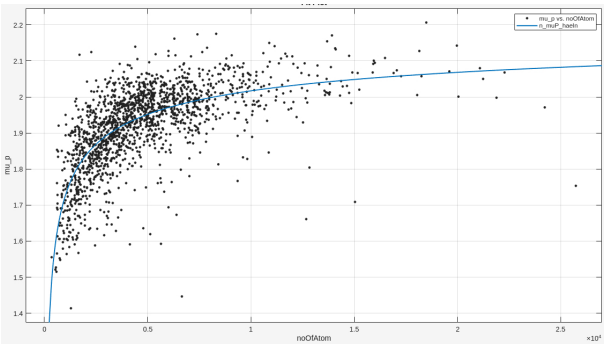
(a) μ_s (b) μ_p

(c) μ_t (d) n_s

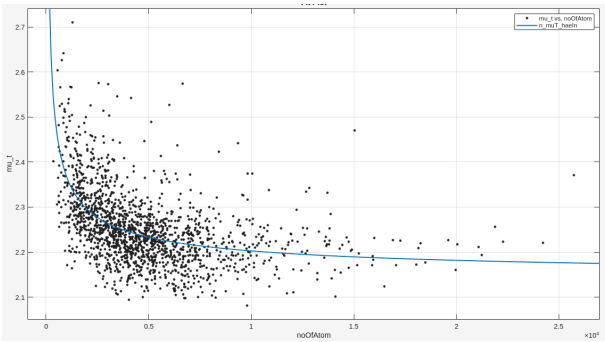
Figure S3: **The changes of μ_s, μ_p, μ_t and n_s with n for *Escherichia coli* proteome**.



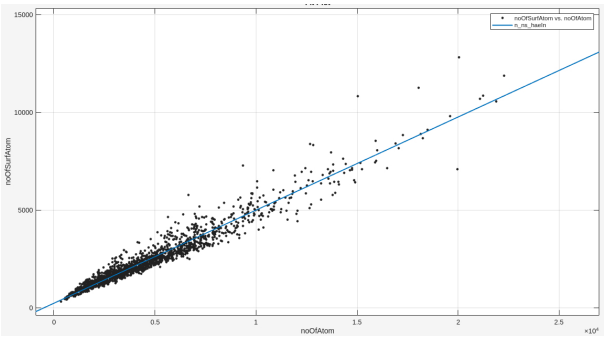
(a) μ_s



(b) μ_p

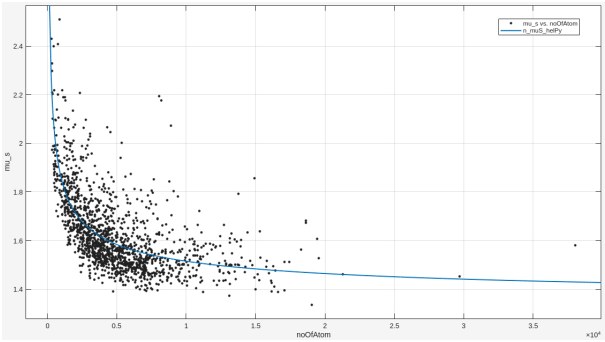


(c) μ_t

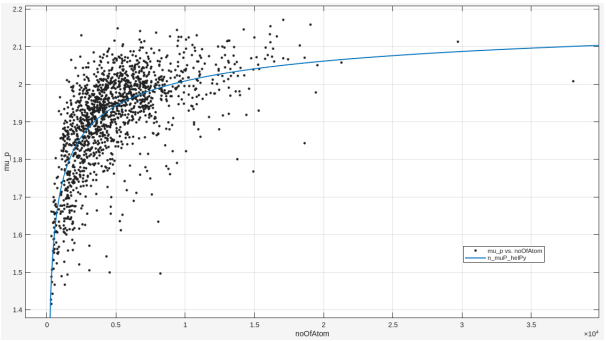


(d) n_s

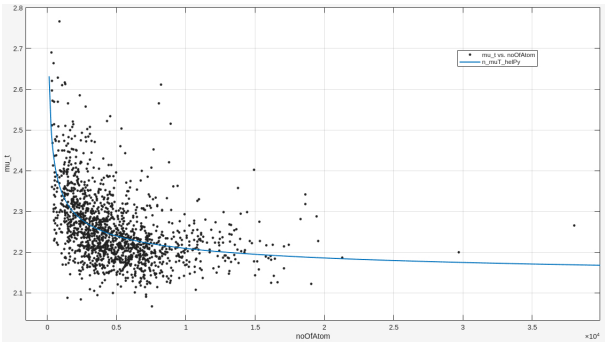
Figure S4: **The changes of μ_s, μ_p, μ_t and n_s with n for *Haemophilus influenzae* proteome**.



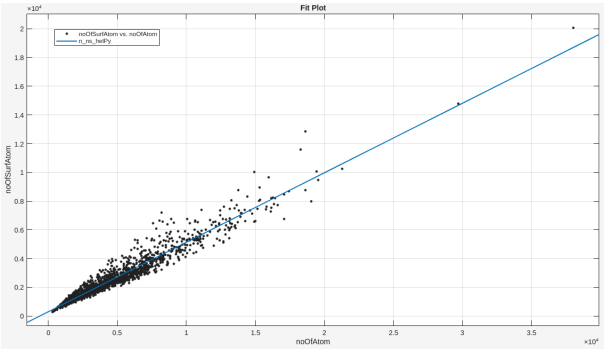
(a) μ_s



(b) μ_p

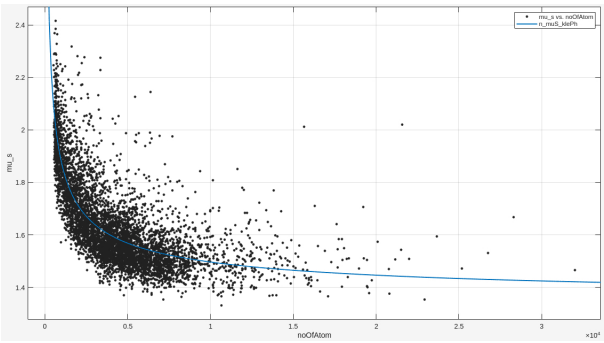


(c) μ_t

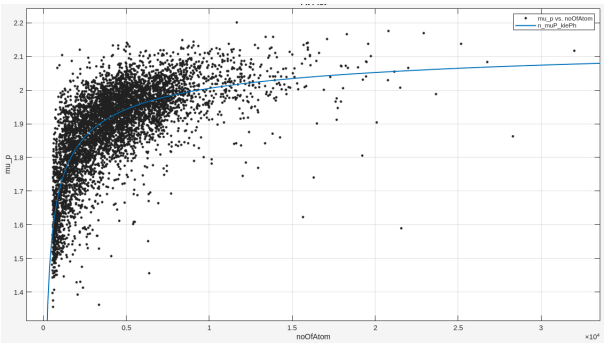


(d) n_s

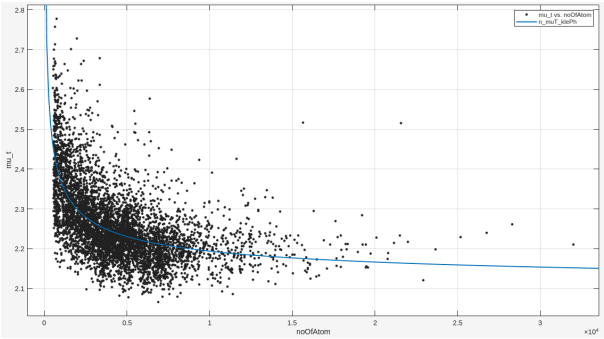
Figure S5: **The changes of μ_s, μ_p, μ_t and n_s with n for *Helicobacter pylori* proteome**.



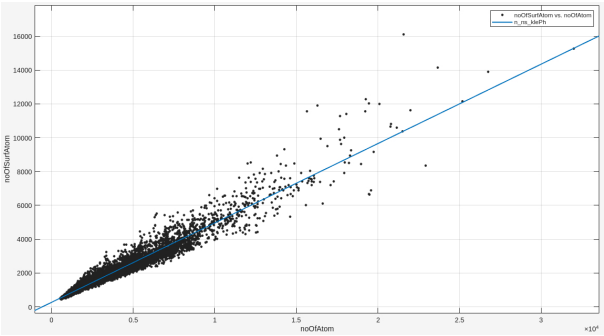
(a) μ_s



(b) μ_p

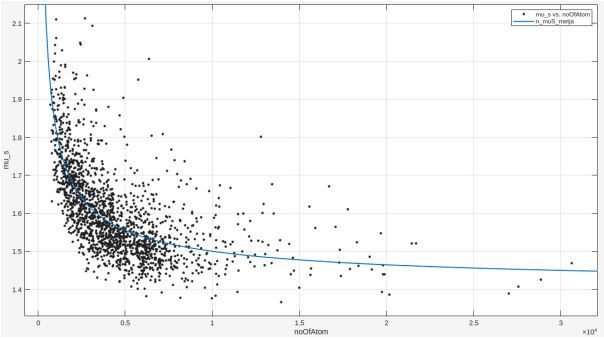


(c) μ_t

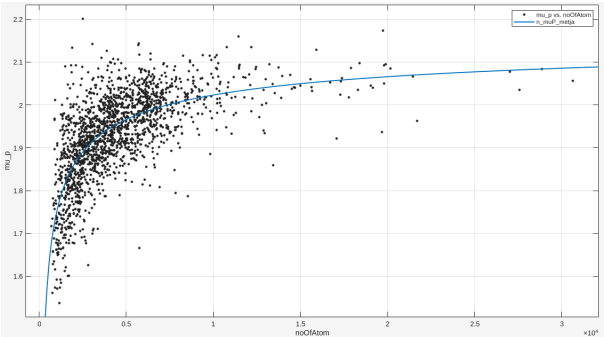


(d) n_s

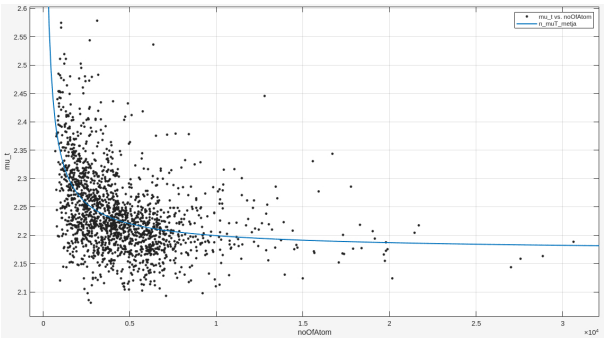
Figure S6: **The changes of μ_s, μ_p, μ_t and n_s with n for *Klebsiella pneumoniae* proteome**.



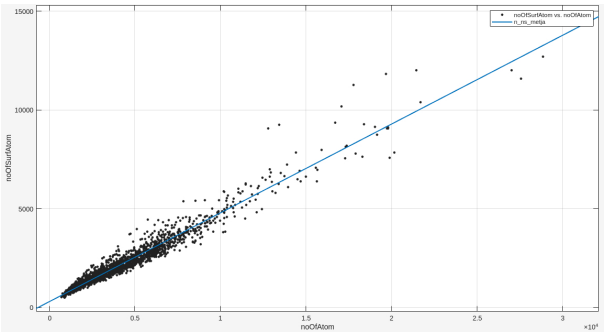
(a) μ_s



(b) μ_p

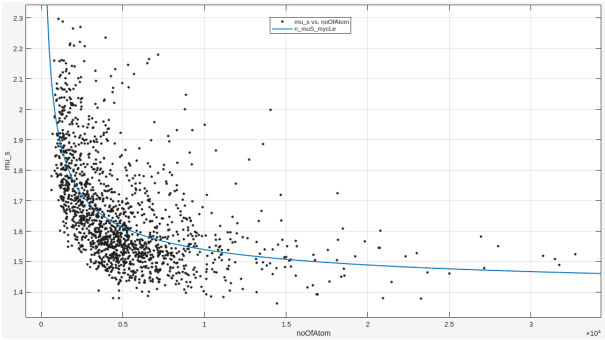


(c) μ_t

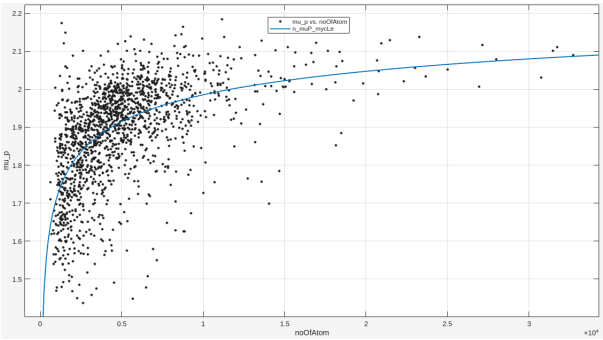


(d) n_s

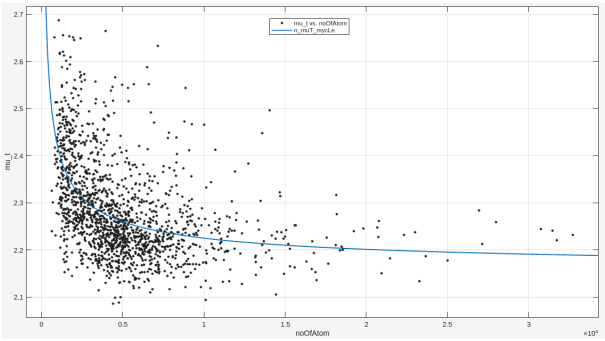
Figure S7: **The changes of μ_s, μ_p, μ_t and n_s with n for *Methanocaldococcus jannaschii* proteome**.



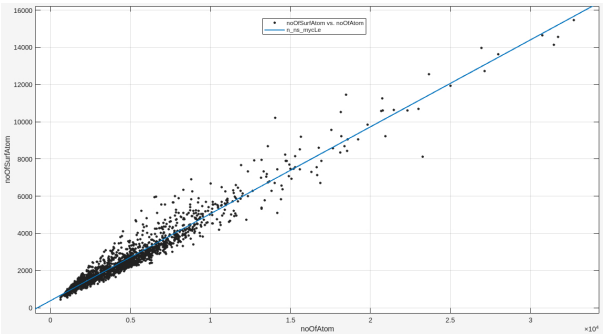
(a) μ_s



(b) μ_p

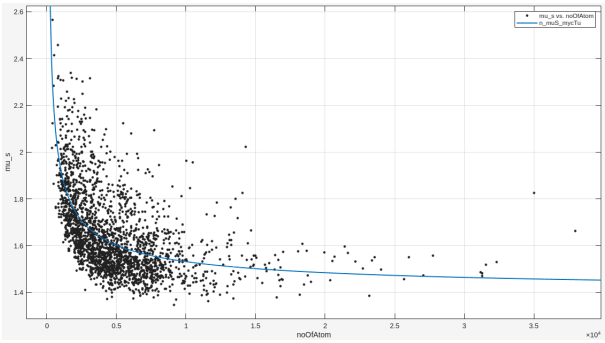


(c) μ_t

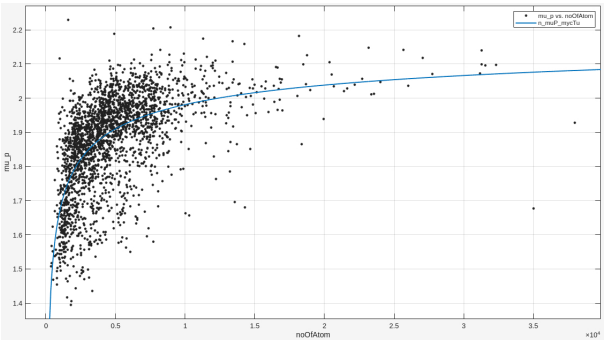


(d) n_s

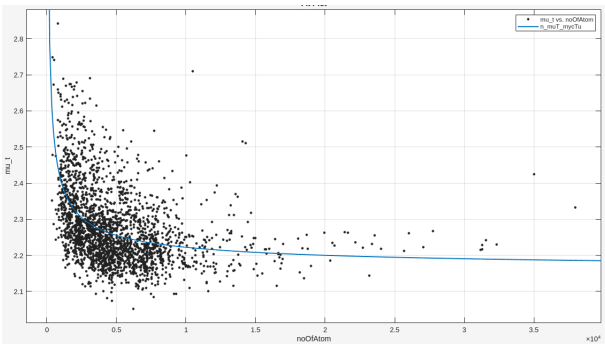
Figure S8: **The changes of μ_s, μ_p, μ_t and n_s with n for *Mycobacterium leprae* proteome**.



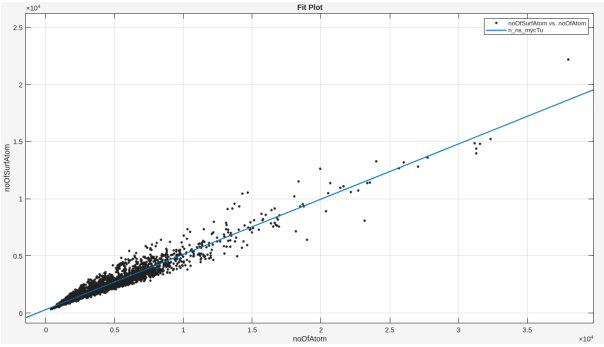
(a) μ_s



(b) μ_p

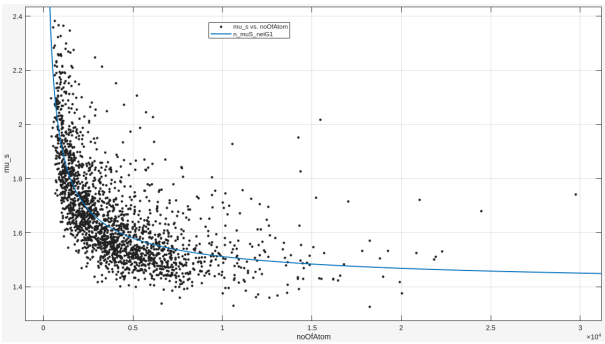


(c) μ_t

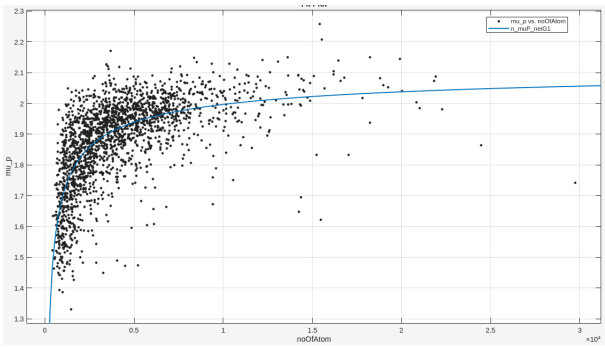


(d) n_s

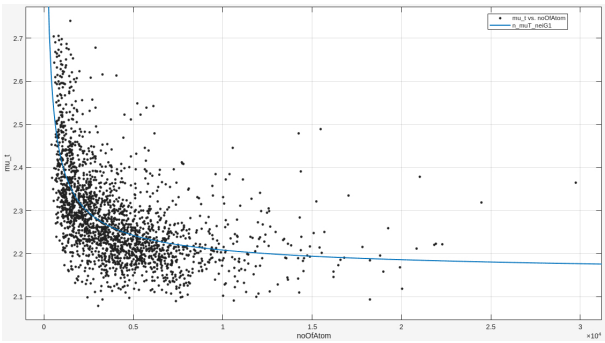
Figure S9: **The changes of μ_s, μ_p, μ_t and n_s with n for *Mycobacterium tuberculosis* proteome**.



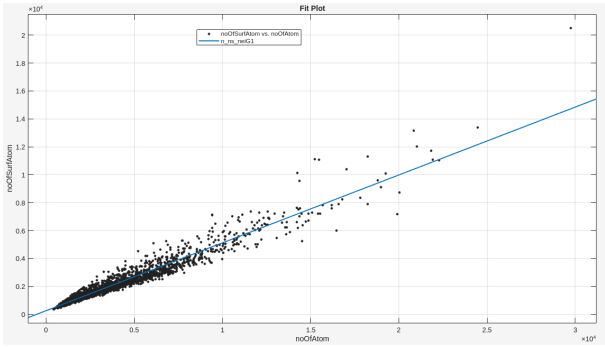
(a) μ_s



(b) μ_p

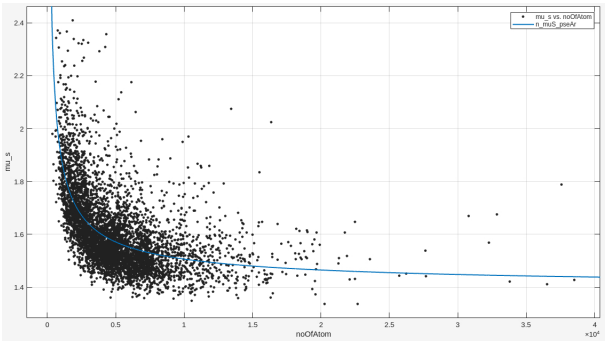


(c) μ_t

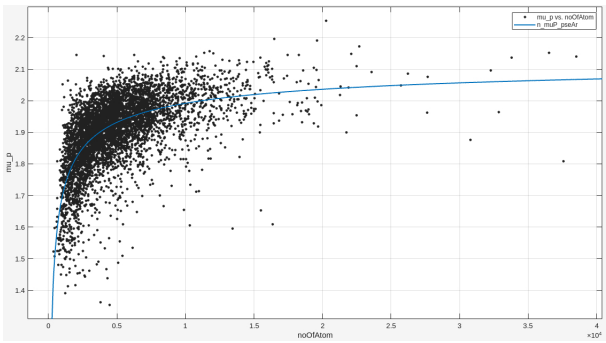


(d) n_s

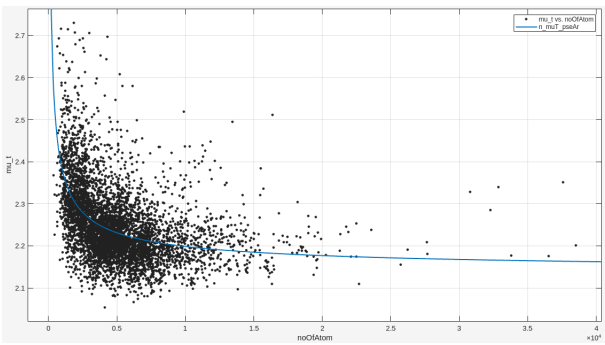
Figure S10: **The changes of μ_s, μ_p, μ_t and n_s with n for *Neisseria gonorrhoeae* proteome**.



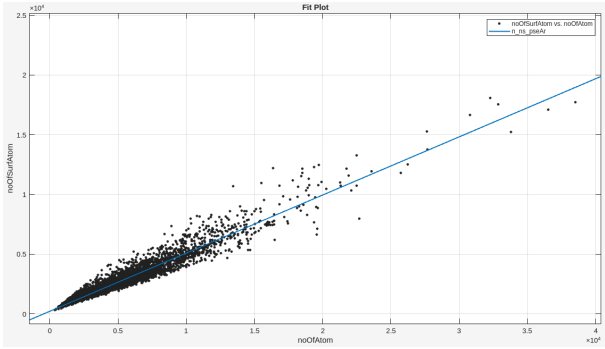
(a) μ_s



(b) μ_p

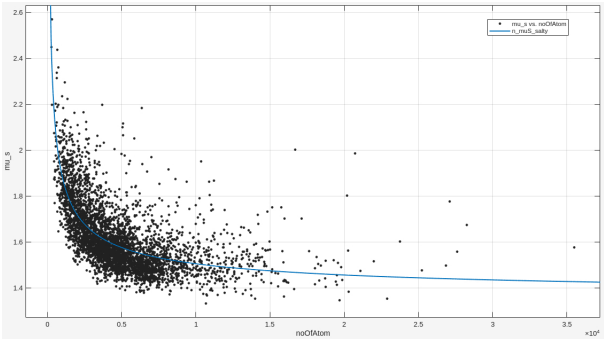


(c) μ_t

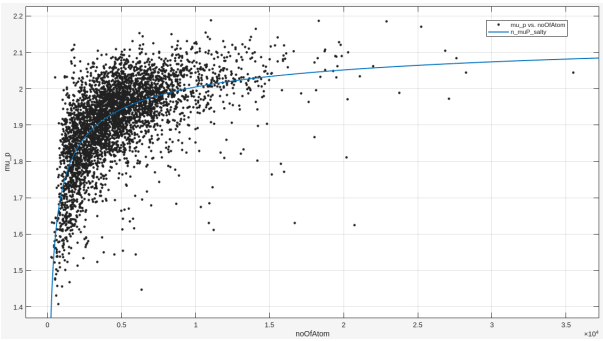


(d) n_s

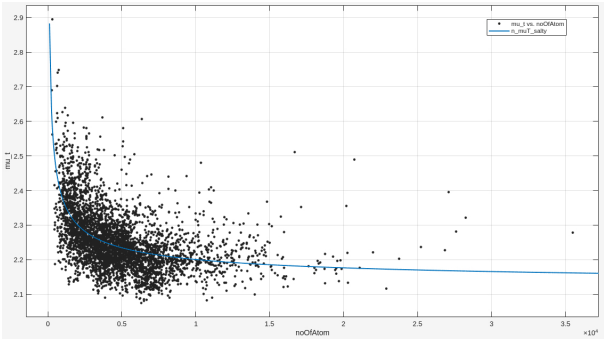
Figure S11: **The changes of μ_s, μ_p, μ_t and n_s with n for *Pseudomonas aeruginosa* proteome**.



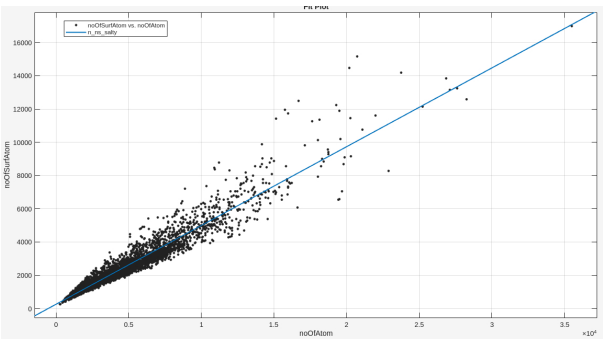
(a) μ_s



(b) μ_p

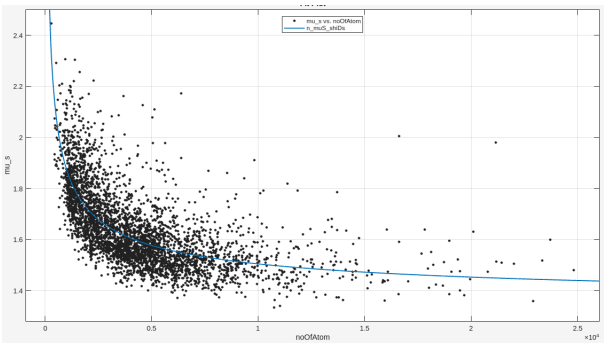


(c) μ_t

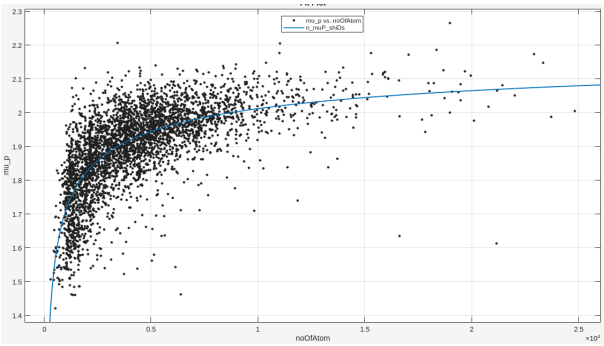


(d) n_s

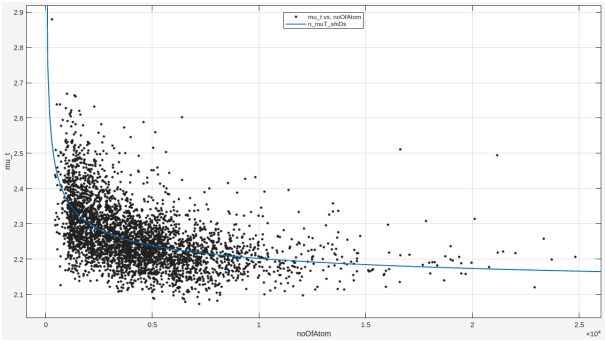
Figure S12: **The changes of μ_s, μ_p, μ_t and n_s with n for *Salmonella typhimurium* proteome**.



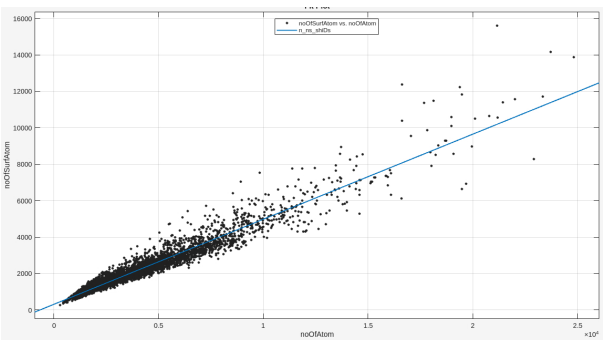
(a) μ_s



(b) μ_p

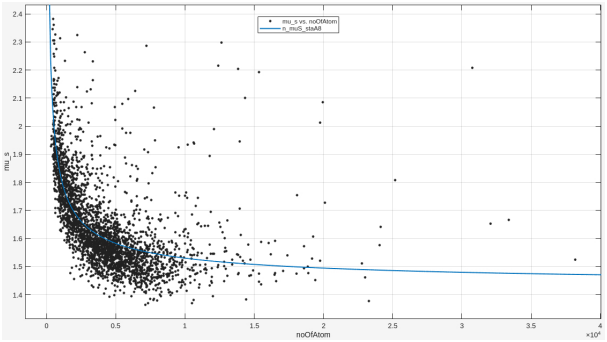


(c) μ_t

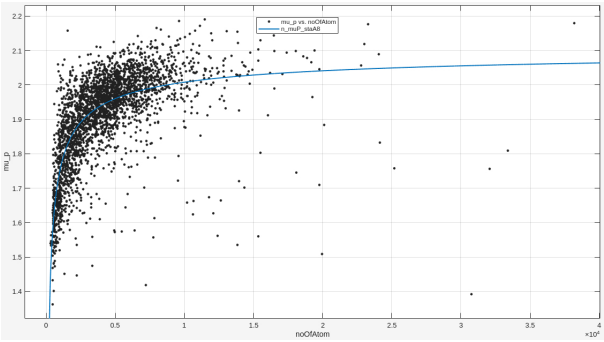


(d) n_s

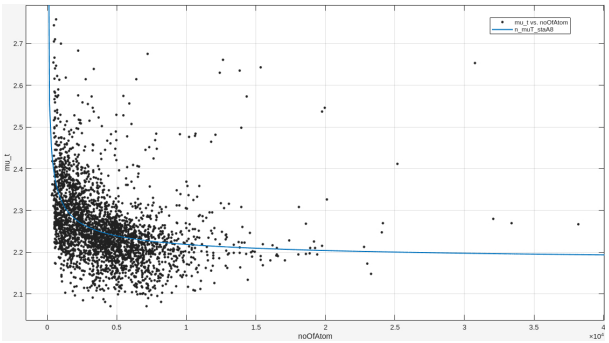
Figure S13: **The changes of μ_s, μ_p, μ_t and n_s with n for *Shigella dysenteriae* proteome**.



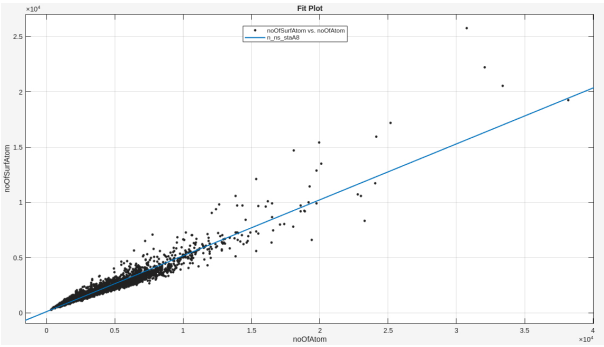
(a) μ_s



(b) μ_p

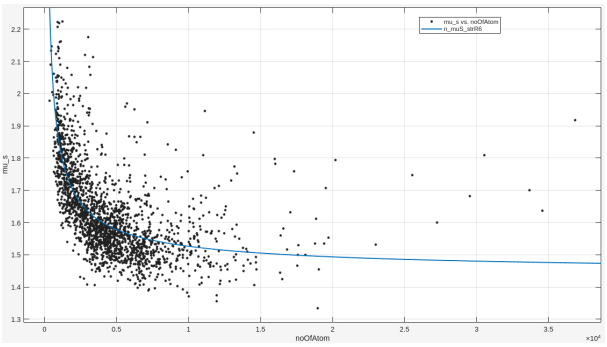


(c) μ_t

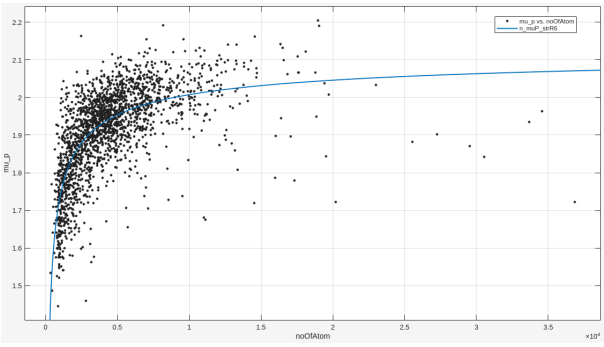


(d) n_s

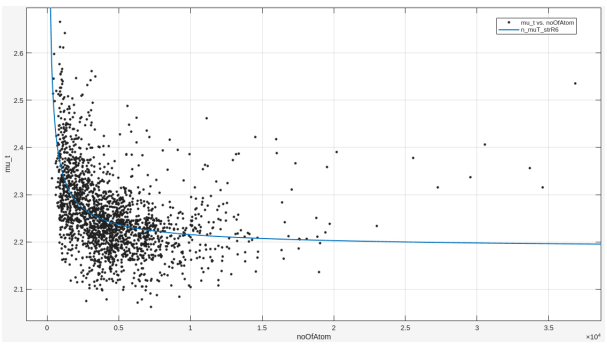
Figure S14: **The changes of μ_s, μ_p, μ_t and n_s with n for *Staphylococcus aureus* proteome**.



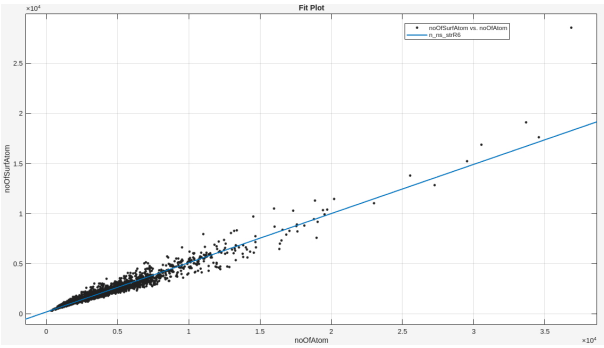
(a) μ_s



(b) μ_p



(c) μ_t

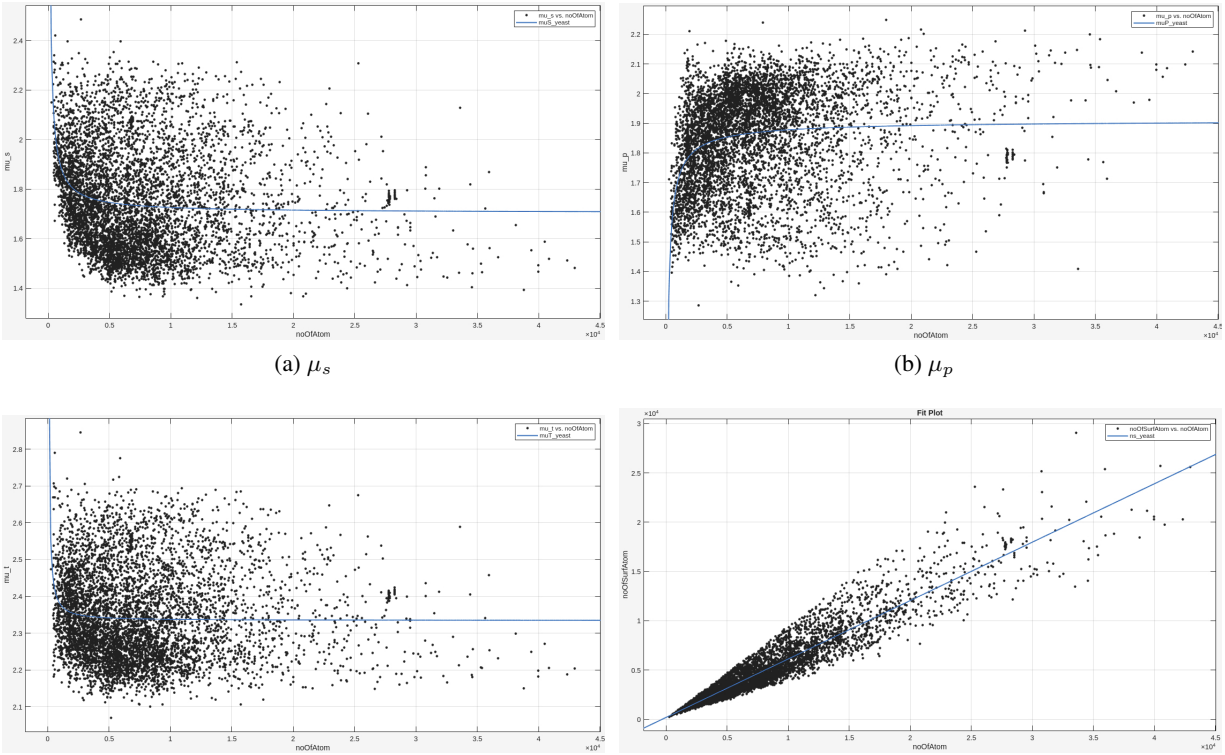


(d) n_s

Figure S15: **The changes of μ_s, μ_p, μ_t and n_s with n for *Streptococcus pneumoniae* proteome**.

S3.2: The fourteen eukaryotic proteomes

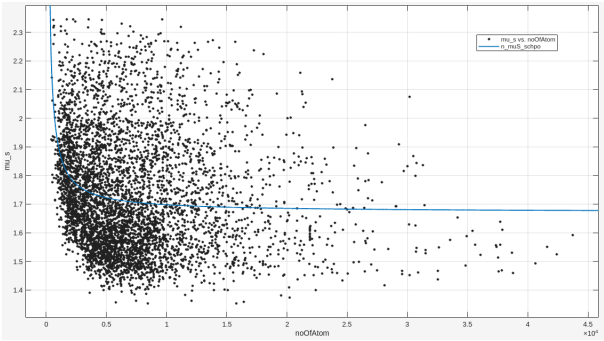
S3.2.1: Two yeast proteomes



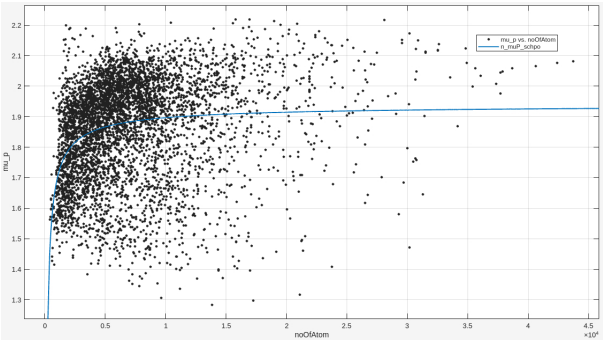
(a) μ_s (b) μ_p

(c) μ_t (d) n_s

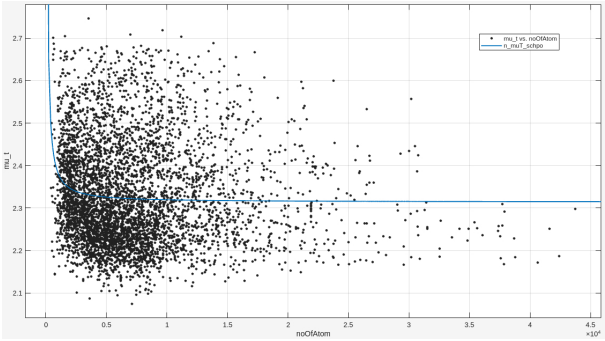
Figure S16: **The changes of μ_s, μ_p, μ_t and n_s with n for *Saccharomyces cerevisiae* proteome**.



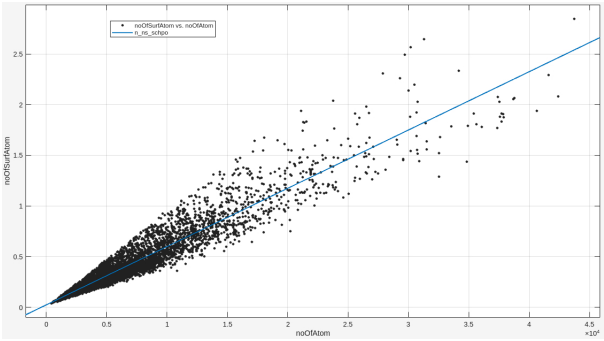
(a) μ_s



(b) μ_p



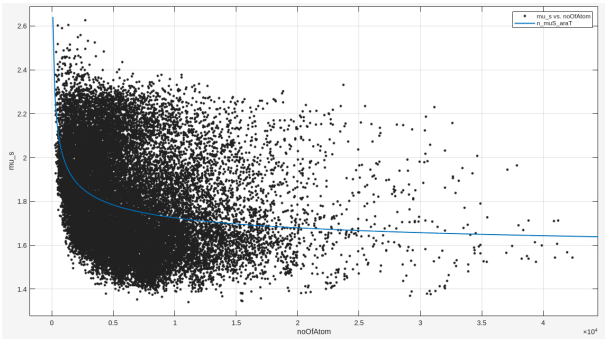
(c) μ_t



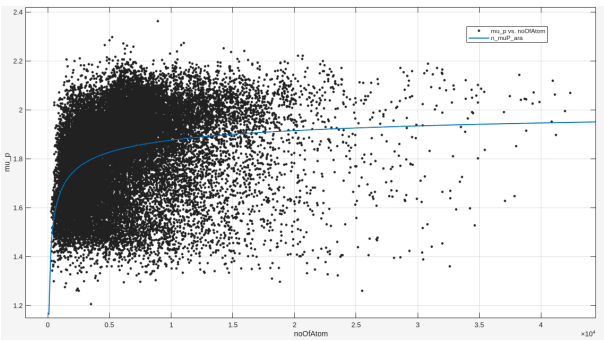
(d) n_s

Figure S17: **The changes of μ_s, μ_p, μ_t and n_s with n for *Schizosaccharomyces pombe* proteome**.

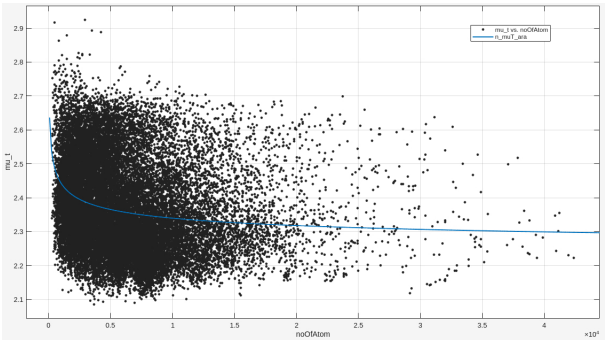
S3.2.2: Four plant proteomes



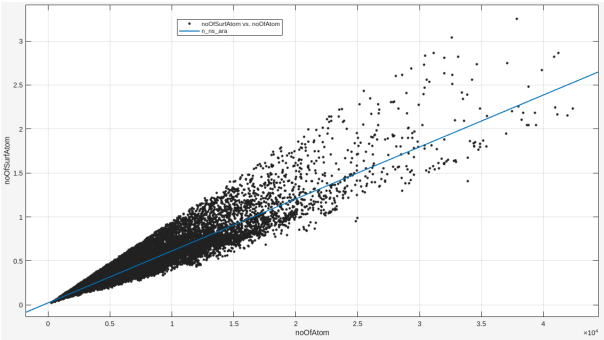
(a) μ_s



(b) μ_p

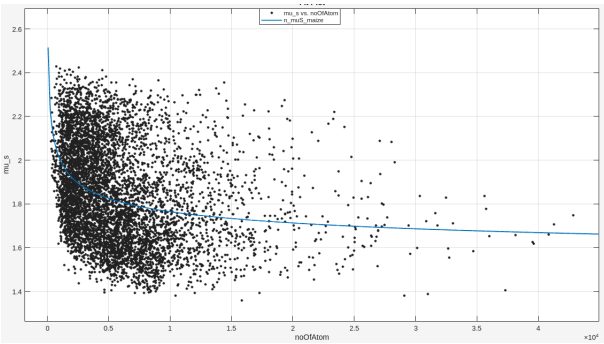


(c) μ_t

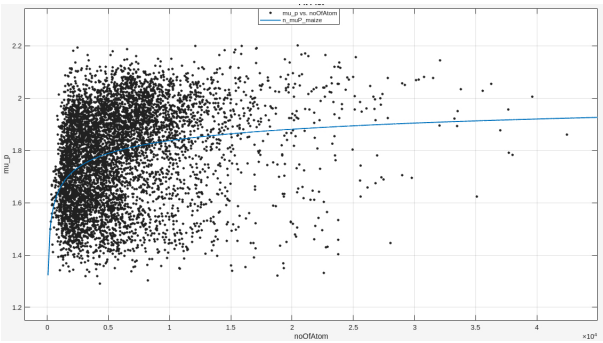


(d) n_s

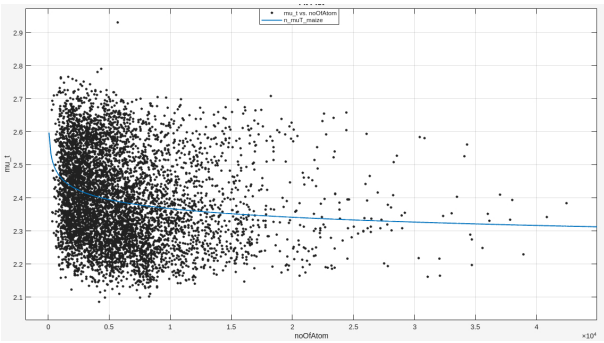
Figure S18: **The changes of μ_s, μ_p, μ_t and n_s with n for *Arabidopsis thaliana* proteome**.



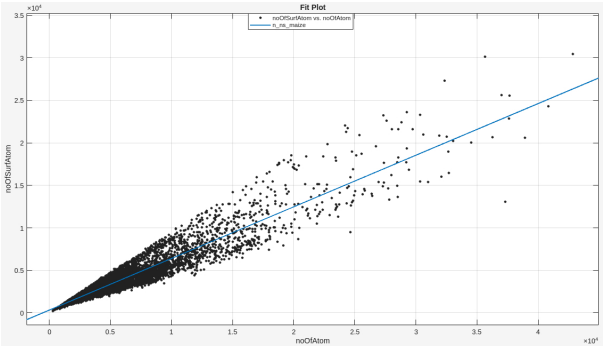
(a) μ_s



(b) μ_p

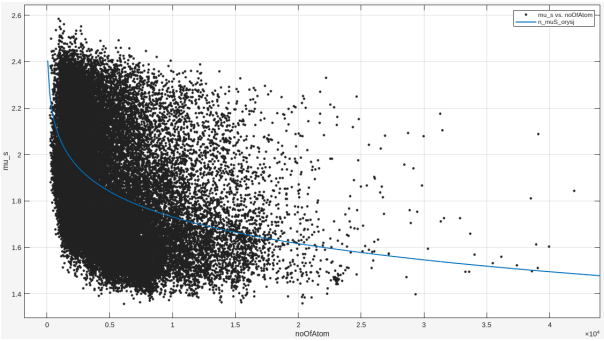


(c) μ_t

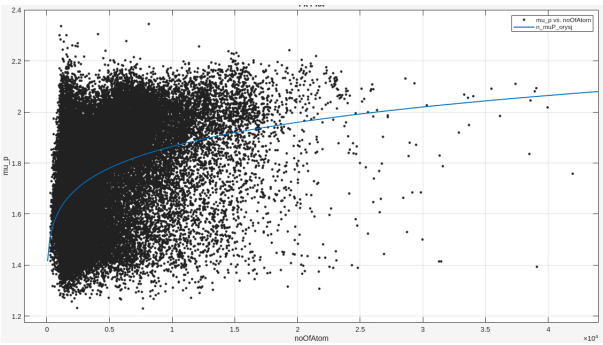


(d) n_s

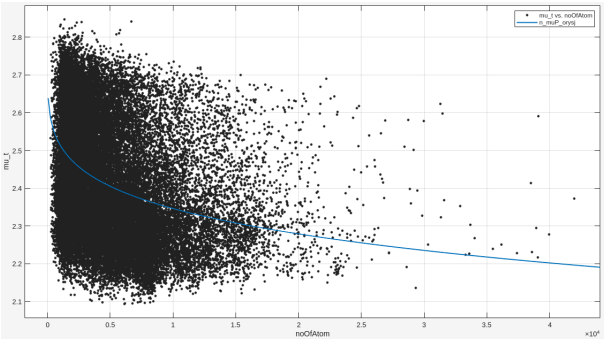
Figure S19: **The changes of μ_s, μ_p, μ_t and n_s with n for *Zea mays* proteome**.



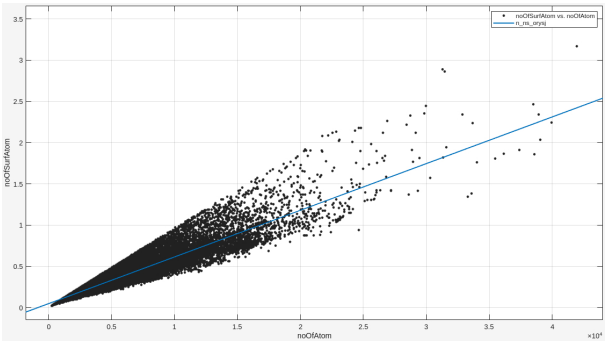
(a) μ_s



(b) μ_p

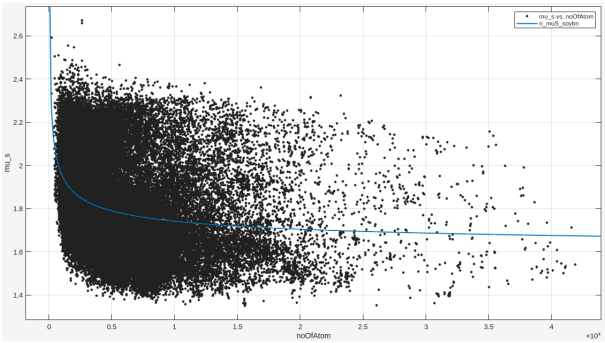


(c) μ_t

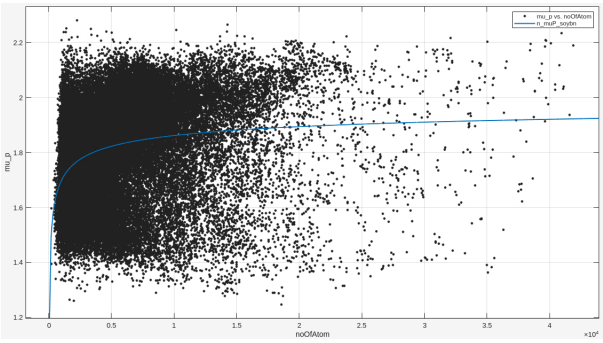


(d) n_s

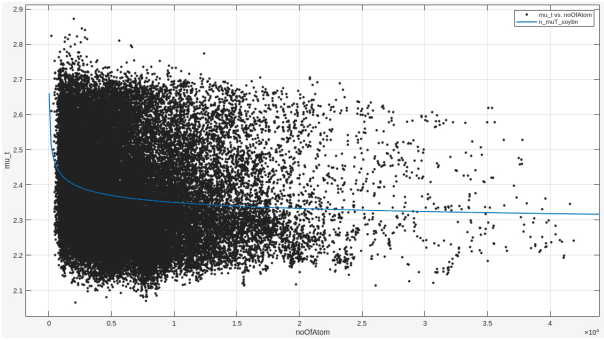
Figure S20: **The changes of μ_s, μ_p, μ_t and n_s with n for *Oryza sativa* proteome**.



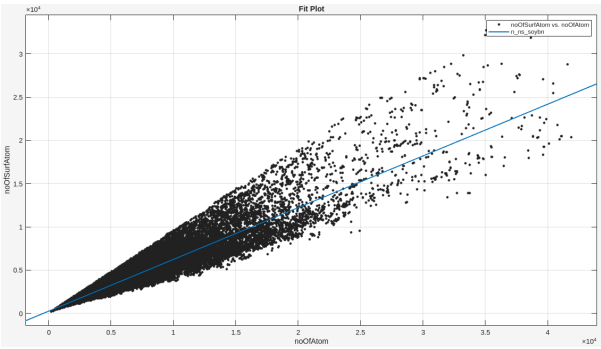
(a) μ_s



(b) μ_p



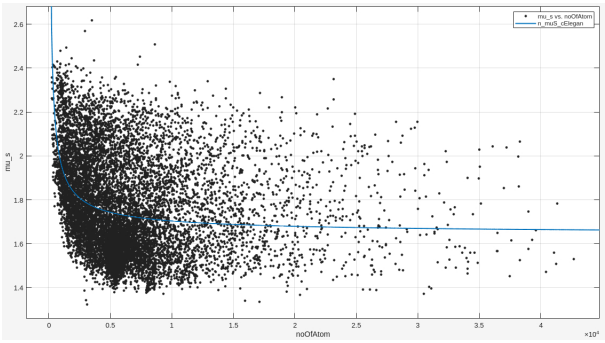
(c) μ_t



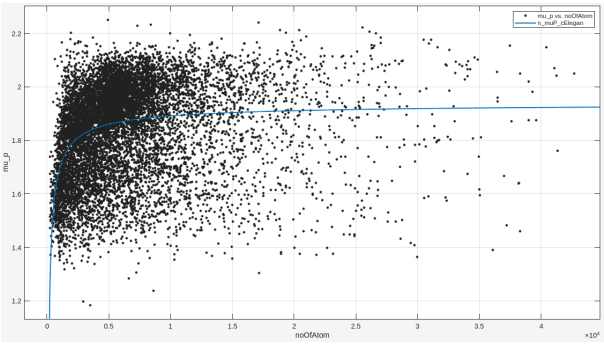
(d) n_s

Figure S21: **The changes of μ_s, μ_p, μ_t and n_s with n for *Glycine max* proteome**.

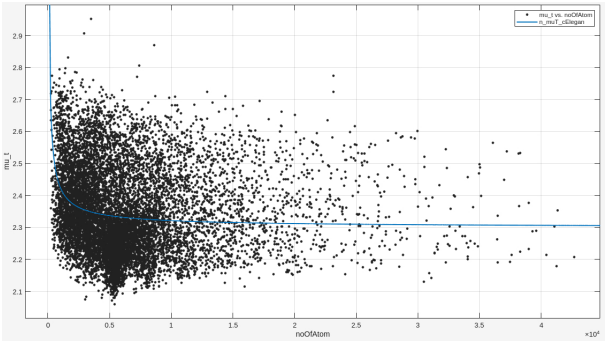
S3.2.3: Eight animal proteomes



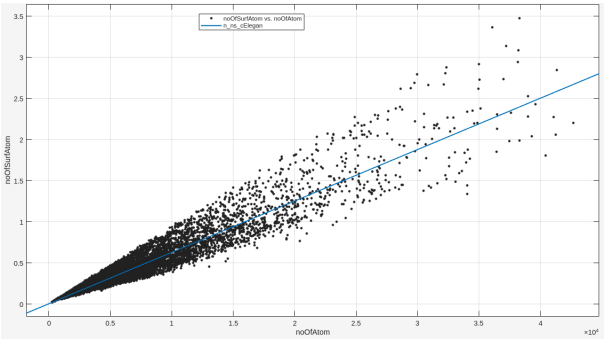
(a) μ_s



(b) μ_p

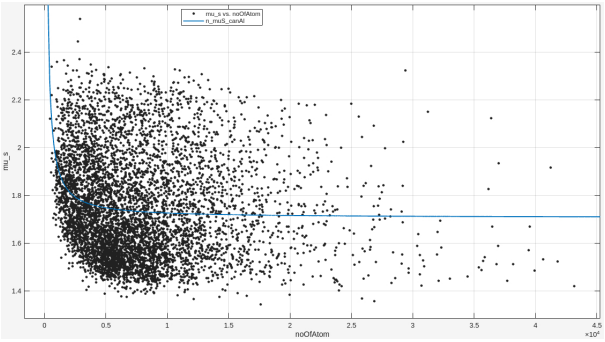


(c) μ_t

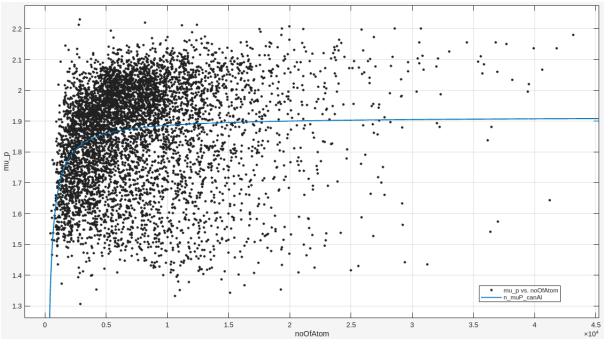


(d) n_s

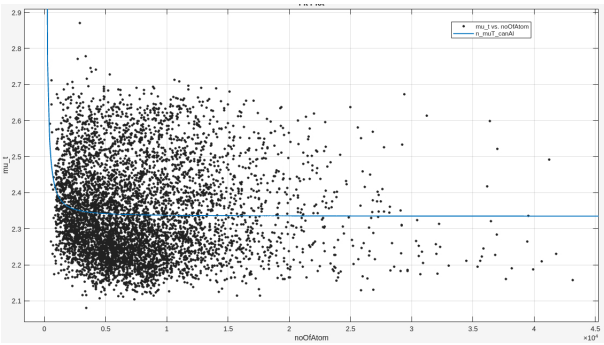
Figure S22: **The changes of μ_s, μ_p, μ_t and n_s with n for *Caenorhabditis elegans* proteome**.



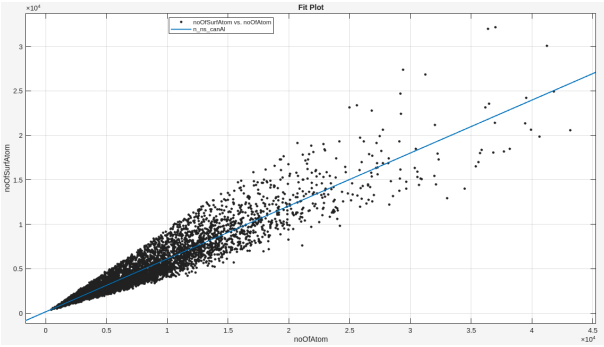
(a) μ_s



(b) μ_p

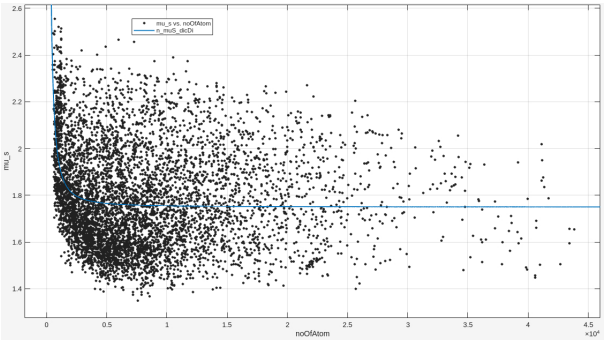


(c) μ_t

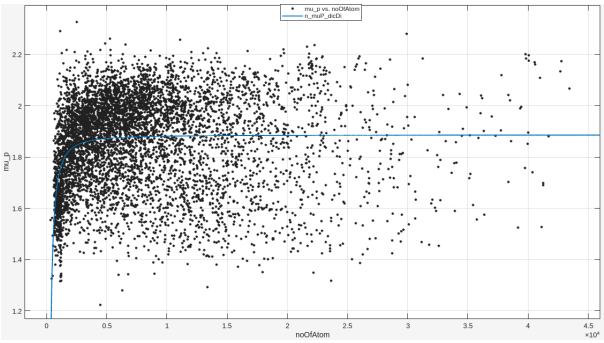


(d) n_s

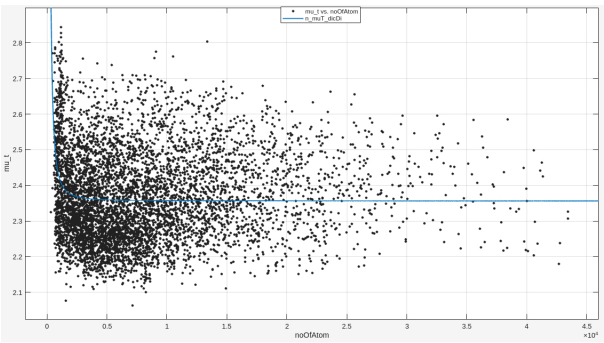
Figure S23: **The changes of μ_s, μ_p, μ_t and n_s with n for *Candida albicans* proteome**.



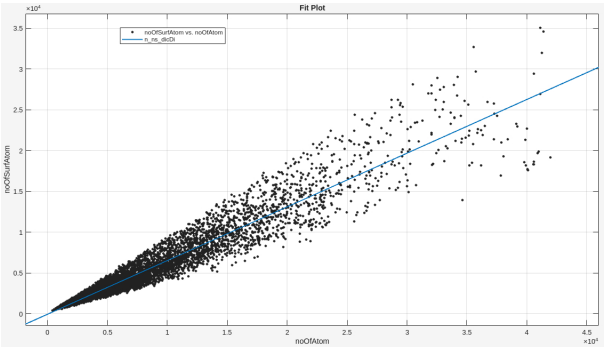
(a) μ_s



(b) μ_p

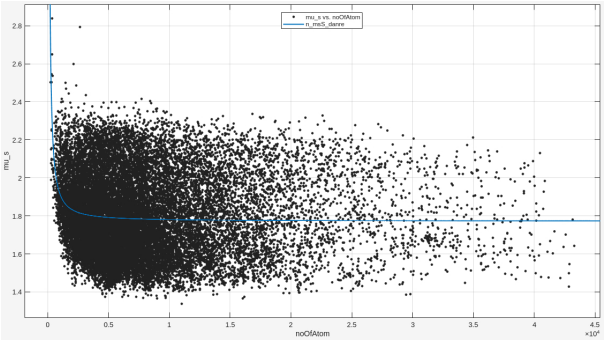


(c) μ_t

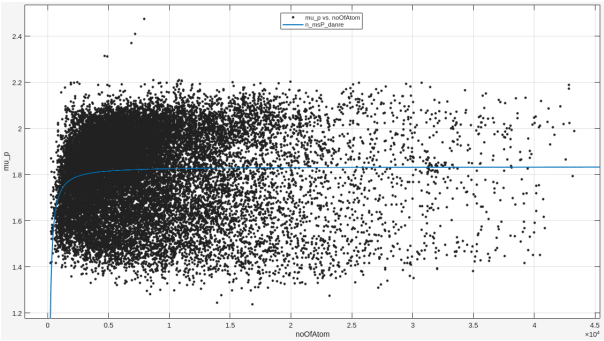


(d) n_s

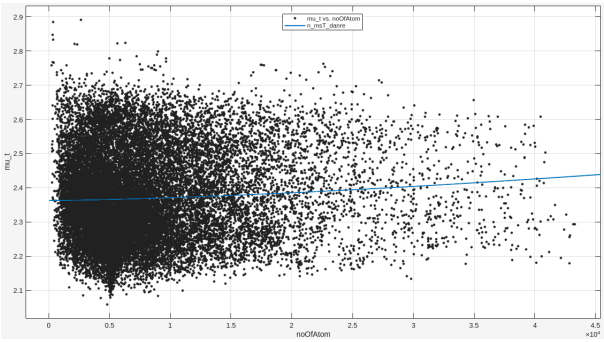
Figure S24: **The changes of μ_s, μ_p, μ_t and n_s with n for *Dictyostelium discoideum* proteome**.



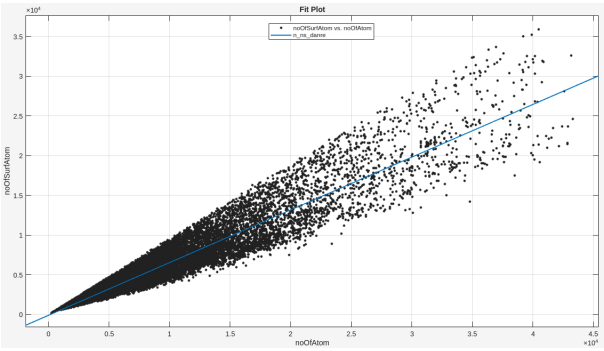
(a) μ_s



(b) μ_p

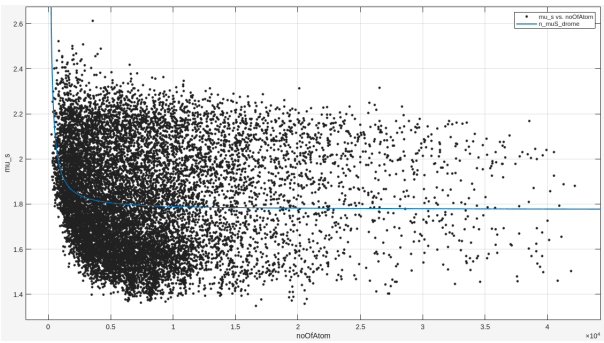


(c) μ_t

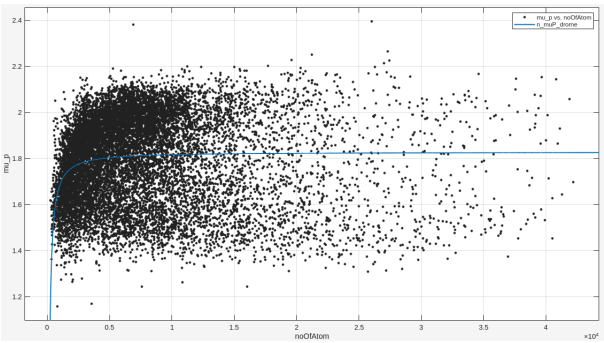


(d) n_s

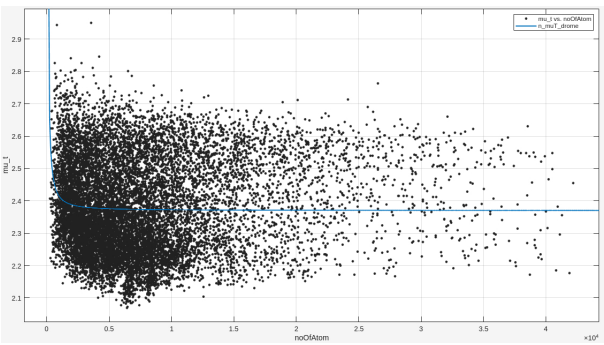
Figure S25: **The changes of μ_s, μ_p, μ_t and n_s with n for *Danio rerio* proteome**.



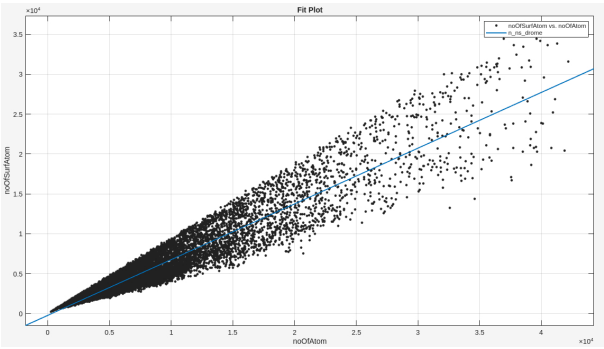
(a) μ_s



(b) μ_p

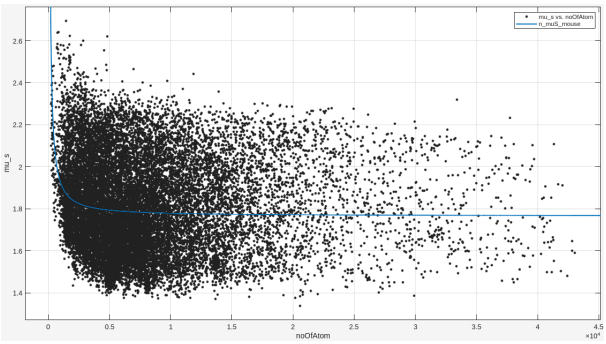


(c) μ_t

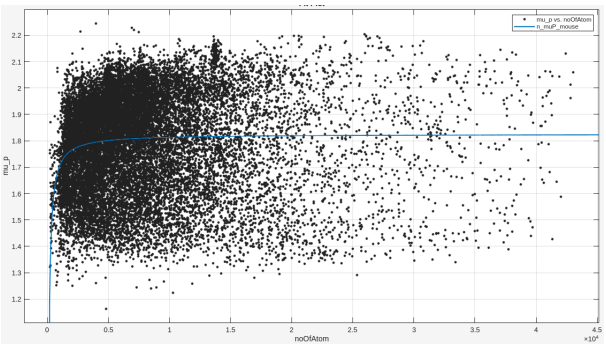


(d) n_s

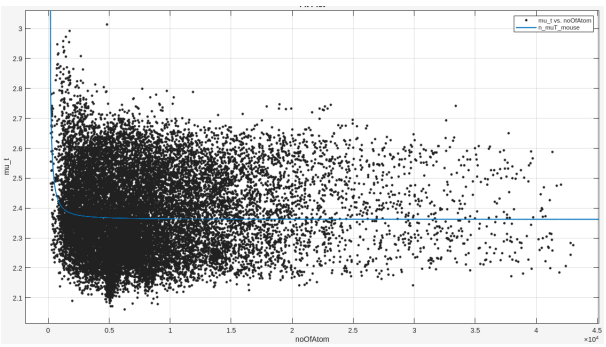
Figure S26: **The changes of μ_s, μ_p, μ_t and n_s with n for *Drosophila melanogaster* proteome**.



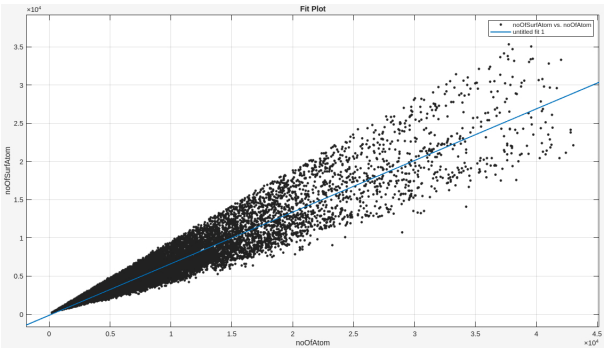
(a) μ_s



(b) μ_p

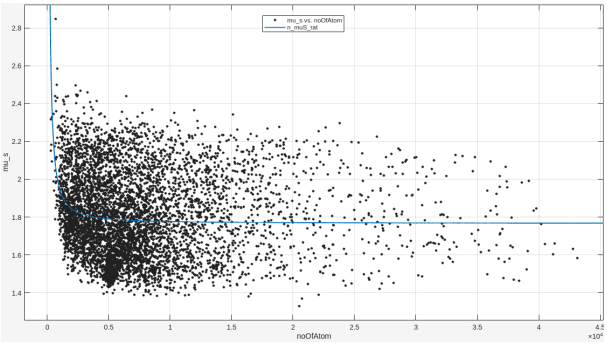


(c) μ_t

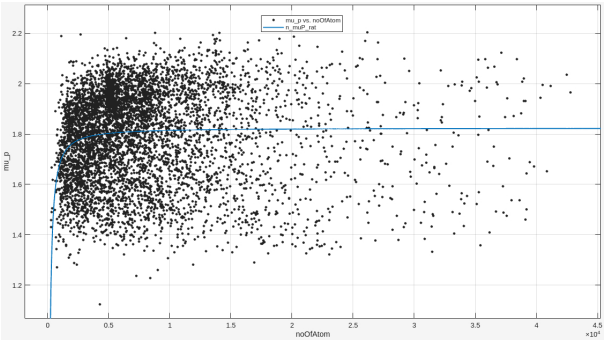


(d) n_s

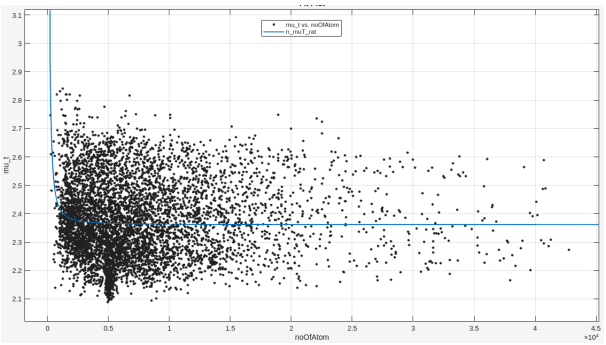
Figure S27: **The changes of μ_s, μ_p, μ_t and n_s with n for *Mus musculus* proteome**.



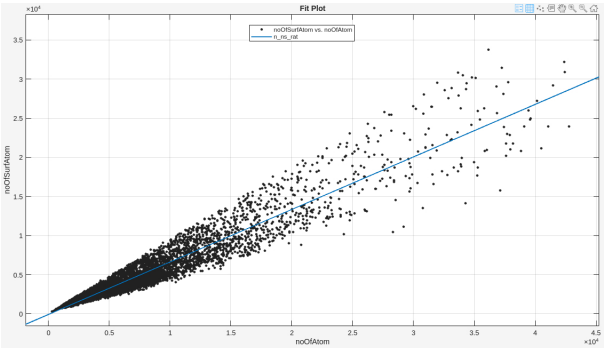
(a) μ_s



(b) μ_p

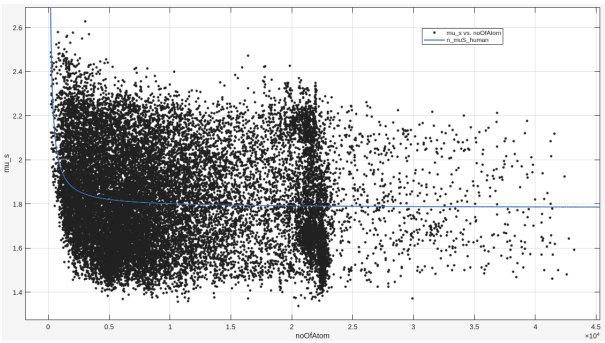


(c) μ_t

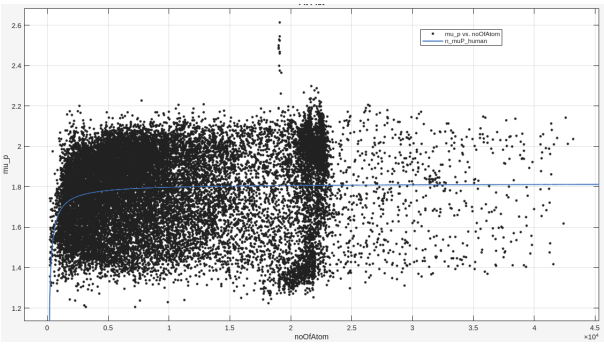


(d) n_s

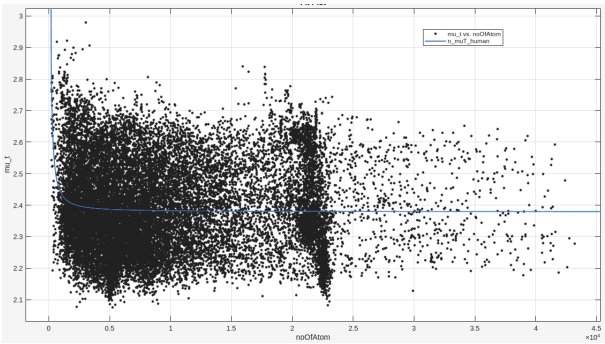
Figure S28: **The changes of μ_s, μ_p, μ_t and n_s with n for *Rattus norvegicus* proteome**.



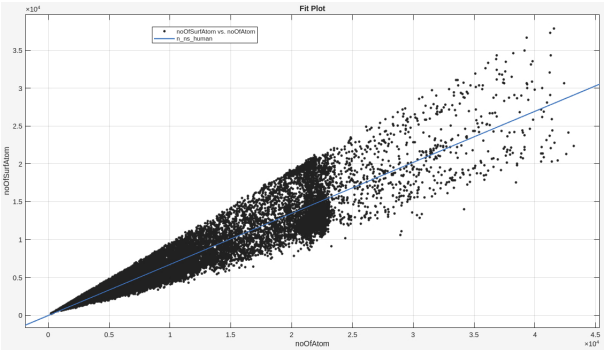
(a) μ_s



(b) μ_p



(c) μ_t



(d) n_s

Figure S29: **The changes of μ_s, μ_p, μ_t and n_s with n for *Homo sapiens* proteome**.

S4: The biological names of the twenty seven species and their abbreviations

Table S1 lists the biological names of the twenty seven species and their abbreviations used in Table 2 of the main text.

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
Arabidopsis thaliana	araTh
Zea mays	maize
Oryza sativa	orySj
Glycine max	soyBn
Caenorhabditis elegans	celEg
Candida albicans	canals
Dictyostelium discoideum	dicDi
Danio rerio	danRe
Drosophila melanogaster	droMe
Mus musculus	mouse
Rattus norvegicus	rat
Homo sapiens	human

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

S4: The pdbids of the three sets of experimentally-derived crystal structures

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