



The role of stability in protein evolution

Citation

Razban, Rostam. 2020. The role of stability in protein evolution. Doctoral dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37369433>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. [Submit a story](#)

HARVARD UNIVERSITY
Graduate School of Arts and Sciences



DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Department of Chemistry & Chemical Biology
have examined a dissertation entitled:

The role of stability in protein evolution

presented by: Rostam Razban

candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

A handwritten signature in black ink, appearing to read 'E. Shakhnovich', written over a horizontal line.

Signature _____
Typed name: Professor Eugene Shakhnovich

A handwritten signature in black ink, appearing to read 'Alán Aspuru-Guzik', written over a horizontal line.

Signature _____
Typed name: Professor Alán Aspuru-Guzik

A handwritten signature in black ink, appearing to read 'Daniel Hartl', written over a horizontal line.

Signature _____
Typed name: Professor Daniel Hartl

Date: 18 November 2020

The role of stability in protein evolution

Rostam Razban

A dissertation presented in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the subject of Chemistry

Harvard University

November 2020

The role of stability in protein evolution

Abstract

Amino acids have been tuned throughout evolution such that their interactions lead the overall protein to predominantly fold into the same structure at physiological conditions. How often a protein finds its native structure at equilibrium is measured by its change in the Gibbs free energy of the folding process (ΔG).

This thesis explores biophysical constraints on and evolutionary consequences of proteins' optimal ΔG values. We explore whether stability is likely to influence long-term evolution through mutations that cause a loss of function or gain in toxicity (Chapter 1). In Chapter 2, we fully develop the framework to understand the proteomic relationship between abundance and evolutionary rate. We assess the differences between ΔG and another stability proxy called melting temperature to determine whether conclusions reached with available experimental melting temperature data are transferable (Chapter 3). Finally, in Chapter 4 we mathematically motivate an important assumption tying ΔG to evolution: the average change in ΔG from a single mutation.

Through bioinformatically-tested mathematical modeling, we elucidate the role protein stability plays in explaining the universal protein abundance – sequence

evolutionary rate correlation and the Gaussian form of single mutant ΔG effects seen across proteins.

Table of Contents

Acknowledgments.....	vi
Introduction.....	1
Chapter I. Distribution of folding free energies from protein orthologs.....	8
Evolutionary dynamics of folding free energy	12
Flux dynamics unlikely to explain folA orthologs folding free energies ..	21
Unfolding avoidance better explains rnhA and folA free energies.....	24
Chapter II. Mathematical framework of the misfolding avoidance hypothesis.....	31
Closed-form evolutionary rate expression as a function of abundance	34
Misfolding avoidance could underly proteome-wide trends	44
Chapter III. Melting temperature as a proxy for folding free energy	48
Mathematical relationship between stability metrics.....	50
Limits of using melting temperature to capture folding free energy	56
Chapter IV. Single mutants have Gaussian distributed folding free energies	64
Statistical mechanics of stability effects	65
Per residue position stability effects are Gaussian distributed.....	74
Per residue position to per protein stability effects.....	80
Conclusion	83
Data availability. ProteomeVis: a protein evolution web application	86
References.....	92

Acknowledgments

I decided to write this section last because any amount of time I spend is insufficient. There are so many people to thank that I will undoubtedly omit crucial people along my journey.

First and foremost, I would like to thank Eugene Shakhnovich. He provided a lot of freedom to pursue wherever my interests led me. Moreover, he provided patient guidance to strengthen my research by proposing thorough controls and helped extend my preliminary results beyond what I had initially thought possible by proposing new theoretical perspectives. This is most aptly captured by Chapter 1 of this thesis, although examples abound from each chapter. I had initially resisted to study the evolutionary dynamics of ΔG (Equation 5), however by carefully reasoning its importance in deriving the evolutionary rate expression in Chapter 2 (Equation 41), and mentioning its possibility to compare with the unique experimental dataset published by the lab a couple of years ago (Figure 5), Eugene won me over. Looking back, I am not sure how many other people would have been so patient with such a stubborn student when they had a brilliant idea. I appreciate Eugene allowing me to go astray before convincing me of an interesting way forward when I finally came back around.

I would also like to thank the two other members of my committee, Daniel Hartl and Alán Aspuru-Guzik. Despite both not currently being in the Harvard Chemistry department, I appreciate their willingness to be on my committee and offer their helpful comments year after year. I remember one meeting going over time and causing Dan to

be late in picking up a family member from the airport. And another, in which I poorly set up Skype and Alán had trouble hearing my presentation. However, I appreciate both sticking with me and helping me get to this point.

Shakhnovich group members have played a large role in me being where I am as I write this. Each member was willing to drop whatever they were doing to discuss a question that I had, and they were always willing to provide helpful feedback during group meetings to progress my projects. Special thanks goes out to Orit Peleg for getting me started in the group when I first joined; Eugene Serebryany, who taught me how protein biology is done *in vitro*; Michael Manhart, who instilled in me the fundamentals of population dynamics (Equation 24); Bharat Adkar introduced me to all things protein stability (Equation 45); Will Jacobs, who got the brunt of my interruptions by proximally being closest to my desk in the office, advised me on tractable models of protein folding and aggregation; Amy Gilson provided me the opportunity to work on ProteomeVis and introduced me to the curious abundance-evolutionary rate relationship that forms the backbone of this thesis; and Victor Zhao for being a great friend throughout the entirety of my Ph.D. and sharing his knowledge about computer programming.

In addition, I would like to thank Mobolaji Williams, who inspired my attempts at mathematical derivations, rather than be satisfied with just a numerical solution. The math described in the thesis is the result of incredibly helpful discussions with Mobolaji that led to constructing well-defined problems reflecting a biological question and analytically solving for the solution. For example, work in Chapter 2 began when Mobolaji found me at my desk with heaps of scratch paper lying around trying to rederive a result from a published paper. After working through it together on the

chalkboard and Mobolaji elucidating the method of steepest descent as the key approximation, I applied a similar approach to setup and solve Equation 28 to get a mathematical expression for evolutionary rate.

Saving the most important for last, I would like to acknowledge my family. Let me start with the newest member of my family, my fiancée, Xinyi Zhang. The most memorable event during my time at Harvard was spontaneously meeting Xinyi during my first week in my graduate studies at an ice cream social. Even if this whole Ph.D. thing does not pan out, I am content with the process because this is when I met the love of my life.

My sister, Golbarg, taught me confidence. From a young age, she was always fearless. Despite being one of the physically smallest students in her grade, no one dared to bully her (You might think that her tall big brother was the reason, but that was not the case.). Growing up, she took risks against my parents' wishes whenever she thought she was right. It is difficult to imagine that I would have the same level of confidence as I currently have in myself to tackle my hardest obstacles or even begin considering pursuing a graduate degree at the tail end of my undergraduate studies without my sister's influence.

The only reason you are reading these words are the sacrifices of my parents. It is difficult for me to fathom how much adversity my parents went through to raise my sister and I in the United States. They left their families and their homes; they mastered a new language and learned a new culture, so that their children can capitalize on amazing opportunities in the United States. Every decision they made, from the job they worked to

the location of our home, was a selfless action to ensure the well-being of my sister and I.

Maman and Baba, this thesis is a result of your tireless efforts. Khasteh nabashi

Introduction

The advent of protein sequencing in the 1950s (Edman 1950; Sanger and Tuppy 1951) led to the surprising discovery that the same protein in different organisms has different sequences (Margoliash 1963; Pauling and Zuckerkandl 1963; King and Jukes 1969). Even proteins essential to organism survival were found to have varying sequence evolutionary rates (ERs¹), some differing by more than 40% of their aligned amino acids (King and Jukes 1969)!

The neutral theory of molecular evolution was developed to understand sequence variation by positing that although the occurrence of a beneficial mutation is rare, consistent with Darwinian evolution, neutral mutations at the molecular level are common (Kimura 1968). To explain the finding that some proteins evolve faster than other proteins, Kimura and Ota (1974) proposed the protein functional importance hypothesis. More functionally important proteins are hypothesized to have fewer neutral mutations because attempted mutation that even slightly disrupt function could be catastrophic. At the time, proteome-wide experimental data did not exist to test the functional importance hypothesis. Then in the late 1990s, newly developed gene deletion experiments could measure on a continuous scale the relative importance of a protein to the well-being of the cell (Hurst and Smith 1999; Winzeler et al. 1999). Contrary to

¹ In this thesis, ER is defined as the relative number of amino acid differences between orthologous proteins (the same protein in different organisms). We do not divide ER by the divergence time because when comparing multiple orthologous proteins at once, we make sure that they all originate from the same two organisms. Thus, accounting for the divergence time only changes ERs by a scaling constant.

predictions of the functional importance hypothesis, gene dispensability was found to only weakly constrain ER in *S.cerevisiae* proteins (Hirsh and Fraser 2001).

More telling, mRNA abundance, another proteome-wide measurement that became possible around the same time (Holstege et al. 1998), was surprisingly shown to have a strong correlation with ER (Pál et al. 2001). No other proteome-wide measurement had a stronger correlation, and functional importance did not confound the relationship, i.e. more functionally important proteins were not more abundant (Pál et al. 2003; Drummond et al. 2005; Drummond et al. 2006). By the mid-2000s, proteome-wide experiments falsified the protein functional importance hypothesis, and the reason mRNA abundance constrained ER was a mystery. Drummond et al. (2005) filled the void by proposing the protein translation robustness hypothesis, which has since been expanded to become the protein misfolding avoidance hypothesis (MAH) (Yang et al. 2010). MAH posits that more abundant proteins² are more stable to avoid toxic misfolding (Drummond and Wilke 2008; Yang et al. 2010), and more stable proteins evolve slower because they have more destabilizing mutations (Serohijos et al. 2012). Stability is captured by protein folding free energy (ΔG), which can yield the probability that a protein is folded at a given temperature through equilibrium statistical mechanics (Equation 1). Stability effects from mutations are denoted by $\Delta\Delta G = \Delta G^{\text{mut}} - \Delta G^{\text{wt}}$. We adopt the convention that $\Delta G < 0$ refers to a stable protein, thus $\Delta\Delta G < 0$ refers to a stabilizing mutation.

Figure 1 summarizes expected MAH correlations and assumptions.

² When MAH was first proposed, protein abundances were hard to come by (Vogel and Marcotte 2012). Since then, studies have measured protein abundances on the proteome scale (Ghaemmaghami et al. 2003; Wang et al. 2015) and strong protein abundance-evolutionary rate correlations are seen as well (Drummond et al. 2006; Plata and Vitkup 2017; Razban et al. 2018).

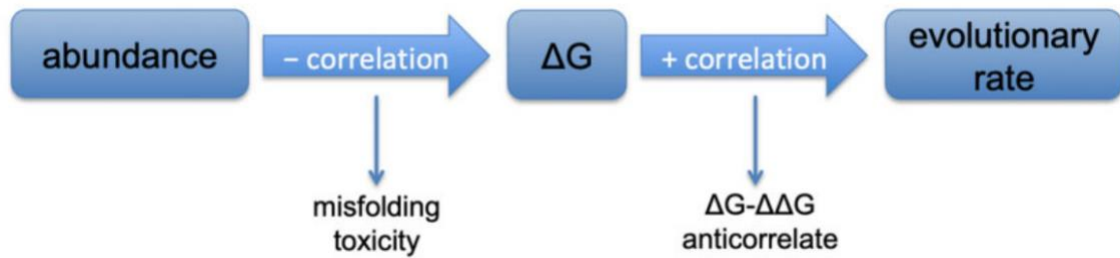


Figure 1: The protein misfolding avoidance hypothesis. Horizontal arrows denote expected correlations between protein properties. Vertical arrows denote the assumptions underlying their respective horizontal arrow. Reproduced with permission from *Molecular Biology and Evolution* (Razban 2019)..

Although the feasibility of the hypothesis has been verified by extensive computer simulations (Drummond and Wilke 2008; Yang et al. 2010; Serohijos et al. 2012), MAH lacks robust experimental testing. At the level of individual proteins, tens of proteins with deposited experimental ΔG s, and proteome-wide studies with ΔG proxies, the jury is still out as to whether MAH is true.

Previous experimental tests of the protein misfolding avoidance hypothesis

1) Individual proteins

Experimental studies of individual proteins have provided mixed results on the validity of MAH. Geiler-Samerotte et al. (2010) measured the growth rate of *S.cerevisiae* organisms expressing mutant yellow fluorescent proteins (YFPs) or mutant orotidine-5'-phosphate decarboxylases and found a linear relationship between insoluble protein and growth rate. In contrast to this result, a slightly earlier study found *E.coli* organisms to be indifferent to overexpressed wild-type or destabilized β -galactosidase mutants, even though mutants were more likely to be misfolded (Plata et al. 2010). In their paper, Geiler-Samerotte et al. (2010) responded to this inconsistency with MAH by arguing that overexpressed proteins activate the protein homeostasis machinery, making misfolding events less likely.

Although the expression levels of YFP are consistent with expression levels of *S.cerevisiae* endogenous proteins, Geiler-Samerotte et al. (2010)'s results on YFPs or orotidine-5'-phosphate decarboxylases in the presence of excess uracil may also not be biologically relevant given that their experiments ignore protein function a priori by focusing on a protein non-endogenous to *S.cerevisiae* or the environment is manipulated such that the protein's function is not required for the cell's survival, respectively. A study expressing mutant dihydrofolate reductases (folAs) via chromosomal incorporation in *E.coli*, found that the cumulative parameter called catalytic capacity ($A \times \text{catalytic activity}$) strongly determines growth rate (Bershtein, Choi, et al. 2015). This cumulative parameter fits more broadly into the flux dynamics hypothesis (Kacser and Burns 1981; Dykhuizen et al. 1987; Bershtein, Serohijos, et al. 2015), which can be thought of as a more quantitative form of the functional importance hypothesis (further discussed in Chapter 1). Only one of the four folA mutants studied exhibited protein misfolding toxicity, which the authors called a secondary effect (Bershtein, Choi, et al. 2015). In general, it is not clear whether MAH should hold robustly at the level of individual protein, or whether it can be only exhibited when pooling data from multiple proteins that are not necessary enzymes, where protein functional effects become irrelevant.

2) Proteins with experimentally measured ΔG s

Protein folding free energy can be experimentally measured by differential scanning calorimetry (Robertson and Murphy 1997) or denaturation curves (Pace and Shaw 2000). By mining the Protein Thermodynamic database (ProTherm) (Gromiha et al. 2016), 28 *E.coli* proteins (Plata et al. 2010; Usmanova et al. 2020), 5 *S.cerevisiae* proteins (Yang et al. 2010), and 47 *H.sapiens* proteins (Usmanova et al. 2020) were

identified with previously measured ΔG s. When correlating ΔG with protein abundance (A) or with evolutionary rate (ER), none of the organisms exhibited the MAH-expected negative and positive correlations, respectively. However, it is not clear that such a small sample size well-represents the proteome. With 28 experimental ΔG s, *E.coli* has the largest fraction of its proteome measured at less than 1%.

3) Proteome-wide studies with ΔG proxies

Making proteome-wide ΔG measurements has been a difficult endeavor, although progress is being made (Walker et al. 2018; Nisthal et al. 2019). Some studies employing ΔG proxies found indirect support for MAH. Highly expressed proteins in several proteomes, including *E.coli*, *S.cerevisiae* and *H.sapiens*, were shown to have sequences similar to those of thermophiles, whose proteins are assumed to be highly stable (Joshua L. Cherry 2010). Additionally, highly expressed proteins in *S.cerevisiae* also showed some enhanced features of proxies for ΔG , such as strength of hydrogen bonds and interatomic contacts, both calculated by the computational software Eris (Serohijos et al. 2013).

However, another study employing melting temperature (T_m) as a proxy for ΔG did not find support for MAH. Proteome-wide T_m measurements have recently become possible (Leuenberger et al. 2017) and provide the closest proxy for ΔG . Utilizing this dataset, Plata and Vitkup (2017) found only one MAH-consistent correlation among the four other organisms for which proteome-wide T_m was measured – *E.coli*, *S.cerevisiae*, *H. sapiens*, *T.thermophilus* – despite all four organisms demonstrated the universal negative correlation between A and ER. The one MAH-consistent correlation belonged to *E.coli*, where the A- T_m correlation was weakly significant. However, MAH as a whole is

still not supported by the *E.coli* proteome because the T_m -ER correlation was essentially nonexistent (Figure 1).

Thesis statement

We assess MAH by carefully deriving mathematical relationships among protein properties. As more proteome-wide stability data become available, it remains to be seen whether MAH is on the verge of being rejected, just like its predecessor, the protein functional importance hypothesis. Our approach in this thesis is to apply first-principles population dynamics and statistical mechanics to make sense of currently available data and reach a verdict.

In the first chapter, we reinvestigate the role of protein function in evolution by treating function more quantitatively with the flux dynamics hypothesis, for a unique dataset of *folA* orthologs. We develop a master equation approach to predict ΔG distributions of orthologs at steady state, finding the MAH fitness function to be more consistent with the data, rather than the flux dynamics fitness function. In the second chapter, we go on to derive quantitative expressions among A , ΔG and ER in a monoclonal evolution framework to fully quantify MAH. Using our expressions, we are able to indirectly test MAH with already available A and ER data. In the third chapter, we carefully consider whether proteome-wide T_m can replace ΔG to test MAH. We first derive an approximately accurate, proteome-evaluable relationship between the protein stability metrics. Then, we apply our quantitative expressions derived in Chapter 2 to obtain resulting A - T_m and T_m -ER correlations. Finally, in the fourth chapter, we focus on testing the assumption underlying the ΔG -ER correlation (Figure 1). The distribution of

single mutant ΔG s is posited to take a universal form, which we motivate through first-principles statistical mechanics.

Chapter I.

Distribution of folding free energies from protein orthologs

Previous studies have been able to quantitatively predict how biophysical properties of proteins, such as abundance, stability and catalytic activity, influence organisms' growth rates (Dykhuizen et al. 1987; Geiler-Samerotte et al. 2010; Jiang et al. 2013; Bershtein, Serohijos, et al. 2015; Adkar et al. 2017; Adkar et al. 2019) and their impact on the efficacy of protein-specific drugs (Rodrigues et al. 2016; Knies et al. 2017). Intriguingly, different studies on different proteins have found different forms of fitness functions that predominantly explain growth rates. Examples of two fitness functions were discussed in the Introduction: the misfolding avoidance hypothesis (MAH) and the flux dynamics hypothesis (FDH) fitness functions. Overall, MAH posits that misfolded proteins cause fitness decreases by gain of toxicity while FDH posits that misfolded proteins cause fitness decreases by loss of function.

Here, we study whether we can reconcile at longer timescales the two aforementioned fitness functions found to predominate at shorter timescales. Intuitively, one would expect the evolution of enzymes to be governed by its function, despite Geiler-Samerotte et al. (2010)'s contrary finding for one endogenous enzyme of *S.cerevisiae* in a particular environment making the protein nonessential. More comprehensively, protein function has been crudely measured proteome-wide in terms of growth rate effect upon gene knockout (Winzeler et al. 1999) and found not to dictate long-term evolutionary trends (Pál et al. 2003; Drummond et al. 2005; Wang and Zhang

2009). MAH on the other hand, is a leading hypothesis to capture long-term evolutionary trends because of its ability to explain *in silico* the bioinformatically observed result that protein abundance and protein sequence evolutionary rate correlate across proteomes of many organisms (Drummond and Wilke 2008; Yang et al. 2010; Serohijos et al. 2012; Zhang and Yang 2015).

To check that previously reported results are not skewed by a crude metric for protein function and to directly assess MAH plausibility with experimental data, we quantitatively apply the FDH and MAH fitness functions to previously published experimental data from our group on folA orthologs. FolA is an essential protein in *E.coli* (Baba et al. 2006), catalyzing the reaction converting dihydrofolate to tetrahydrofolate, a precursor in DNA synthesis. To determine whether FDH sculps evolutionary dynamics over long timescales beyond immediate organismal growth rate, we use first-principles population dynamics to derive the distributions of stabilities for protein orthologs. Our group previously measured experimental stabilities of 33 folAs belonging to mesophilic bacteria. By comparing this data to our theory with parameter values that are experimentally measured, we can assess the validity of FDH at longer timescales.

	Description	Value
$1/\beta$	Thermal energy	0.59 kcal/mol (temperature = 25°C)
s^{MAH}	Misfolding avoidance hypothesis selection coefficient	$\approx -cA^{\text{tot}}e^{\beta\Delta G}(e^{\beta\Delta\Delta G} - 1)$ (Equation 22)
c	Fitness cost to cell growth of a single misfolded protein	$8*10^{-7}$ (Geiler-Samerotte et al. 2010)
A^{tot}	Total abundance of the protein in the cell	folA: 4 copies (Krug et al. 2013) rnhA: 2 copies (Krug et al. 2013)
s^{FDH}	Flux dynamics hypothesis selection coefficient	$\approx -\frac{b}{A^{\text{tot}}(k^{\text{cat}}/K^{\text{M}})}e^{\beta\Delta G}(e^{\beta\Delta\Delta G} - 1)$
b	Protein effects in enzymatic chain	folA: $0.1 * A(E.coli)$ (Bershtein, Serohijos, et al. 2015)
$k^{\text{cat}}/K^{\text{M}}$	Catalytic activity	folA: $8.1 (s*\mu M)^{-1}$ (Bershtein, Serohijos, et al. 2015)

$p(\Delta\Delta G)$	Distribution of sampled $\Delta\Delta G$ s	$\frac{1}{\sqrt{2\pi}\sigma^2} \exp - \frac{(\Delta\Delta G - \langle\Delta\Delta G\rangle)^2}{2\sigma^2}$ (Equation 23)
$\langle\Delta\Delta G\rangle$	Average of $\Delta\Delta G$ distribution, $b_1 * \Delta G + b_2$	$-0.13 * \Delta G + 0.23$ kcal/mol (Serohijos et al. 2012)
σ	Standard deviation of $\Delta\Delta G$ distribution	1.7 kcal/mol (Serohijos et al. 2012)
P^{fix}	Probability mutation fixes in a monoclonal population	$\frac{1 - e^{-2s}}{1 - e^{-2N_e s}}$ (Equation 24)
N_e	Effective population size	<i>E.coli</i> : 10^8 , <i>S.cerevisiae</i> : 10^7 , <i>H.sapiens</i> : 10^4 (Lynch and Conery 2003)

Table 1: Biophysical parameters relevant to our monoclonal evolutionary framework under the misfolding avoidance or flux dynamics hypothesis. Throughout this thesis chapter and Chapter 2, protein variables and expressions involving variables are in italics.

Fitness functions and corresponding parameters

How often a protein finds its native structure at equilibrium (P^{nat}) is captured by its free energy of folding (ΔG). Assuming two-state folding (Privalov and Khechinashvili 1974),

Equation 1

$$P^{\text{nat}}(\Delta G) = \frac{1}{1 + e^{\beta\Delta G}}$$

The parameter $1/\beta$ is the thermal energy of the environment. We follow the convention throughout this text that stable proteins have $\Delta G < 0$.

The way P^{nat} can influence cell fitness and thus evolution is through abundance. At equilibrium, the soluble abundance of folded protein (A) is simply the total abundance of the protein (A^{tot}) multiplied by P^{nat} .

Equation 2

$$A(\Delta G) = A^{\text{tot}} * P^{\text{nat}}(\Delta G)$$

Experimental studies on individual proteins elucidated different fitness functions with different dependences on abundance. Geiler-Samerotte et al. (2010) found that the fitness function of the misfolding avoidance hypothesis (MAH) applies when measuring the growth rate of *S.cerevisiae* organisms expressing mutant yellow fluorescent proteins

(YFPs) and orotidine-5'-phosphate decarboxylases. The mathematical form of the MAH fitness function depends on the number of misfolded proteins (m) (Drummond and Wilke 2008).

Equation 3

$$f^{\text{MAH}}(\Delta G) = e^{-c*m(\Delta G)}$$

$$m(\Delta G) = A^{\text{tot}} - A(\Delta G)$$

The parameter c is the cellular cost of housing a misfolded protein. By finding an exponential relationship between insoluble YFPs and *S.cerevisiae* growth rates, c was fitted to be $8*10^{-7}$ (Geiler-Samerotte et al. 2010).

Another study expressing mutant dihydrofolate reductases (folAs) via chromosomal incorporation in *E.coli*, found that the cumulative parameter called catalytic capacity ($A*$ catalytic activity) to strongly determine growth rate (Bershtein, Choi, et al. 2015). The fitness function follows from the flux dynamics hypothesis (FDH) and takes the following form (Kacser and Burns 1981; Dykhuizen et al. 1987; Bershtein, Serohijos, et al. 2015; Adkar et al. 2017).

Equation 4

$$f^{\text{FDH}}(\Delta G) = \frac{A(\Delta G) * \left(k^{\text{cat}} / K^{\text{M}}\right)}{b + A(\Delta G) * \left(k^{\text{cat}} / K^{\text{M}}\right)}$$

Note that $k^{\text{cat}}/K^{\text{M}}$ is the parameter pertaining to the protein's catalytic activity. The parameter b is related to effects from other proteins in the enzymatic chain (Kacser and Burns 1981; Dykhuizen et al. 1987). In addition to single, double and triple mutants of folA (Bershtein, Choi, et al. 2015), the FDH fitness function was also found to dictate

growth rates of *E.coli* organisms expressing folA orthologs (Bershtein, Serohijos, et al. 2015).

Evolutionary dynamics of folding free energy

We consider an ensemble of organism populations undergoing evolutionary dynamics for a given protein. We assume that mutations appear on a protein according to a Poisson distribution with some rate μ (Hogg et al. 2005). The parameter μ is the mutation rate per amino acid of the protein in the population (Equation 18). The distribution of ΔG at $t+1$ ($P_{t+1}(\Delta G)$), or the master equation of ΔG , can be written as,

Equation 5

$$\begin{aligned}
 P_{t+1}(\Delta G) = & e^{-\mu} P_t(\Delta G) \\
 & + \mu e^{-\mu} \int_{-\infty}^{\infty} d(\Delta\Delta G) [1 - P^{\text{fix}}(s[\Delta G, \Delta\Delta G])] p(\Delta\Delta G[\Delta G]) P_t(\Delta G) \\
 & - \mu e^{-\mu} \int_{\Delta\Delta G=-\infty, \Delta\Delta G \neq 0}^{\infty} d(\Delta\Delta G) P^{\text{fix}}(s[\Delta G, \Delta\Delta G]) p(\Delta\Delta G[\Delta G]) P_t(\Delta G) \\
 & + \mu e^{-\mu} \int_{-\infty}^{\infty} d(\Delta G^*) P^{\text{fix}}(s[\Delta G, \Delta\Delta G]) p(\Delta G - \Delta G^*) P_t(\Delta G^*)
 \end{aligned}$$

Mutations only affect protein stability, where stability effects of a single mutant are denoted as $\Delta\Delta G = \Delta G^{\text{mut}} - \Delta G$. The probability that a single mutation fixes in the organism population is denoted by P^{fix} . P^{fix} is a function of the selection coefficient (s), which is a function of ΔG and $\Delta\Delta G$ (Equation 22). $p(\Delta\Delta G)$ is the probability distribution of attempted $\Delta\Delta G$ s, which is a function of ΔG to satisfy the assumption that more stable proteins have more destabilizing mutations (Equation 23).

The first line of Equation 5 corresponds to staying at ΔG when either no mutation occurs, or a mutation occurs but does not fix. The second line corresponds to a fixed

mutation causing efflux from the ΔG state. The second line excludes the value of $\Delta\Delta G=0$ in the integral because such a mutation stays in the ΔG state. The third line corresponds to a fixed mutation on a ΔG^* state, causing influx into the ΔG state. We ignore multiple mutations appearing at one time in our master equation because their occurrence is sufficiently rare that we can safely ignore them ($\mu \ll 1$). For our extensive folA dataset, this seems reasonable as positive selection is not detected along any sites in the phylogenetic tree according to the Phylogenetic Analysis of Maximum Likelihood software (Yang 2007).

Inserting MAH defined selection coefficient and $p(\Delta\Delta G)$ under a monoclonal evolution framework with experimentally measured parameters (Table 1), we numerically solve Equation 5 for a given initial condition across time. In Figure 2, we see that we reach a steady state solution regardless of initial conditions by $10^9/\mu$ timesteps for a population size of 10^8 .

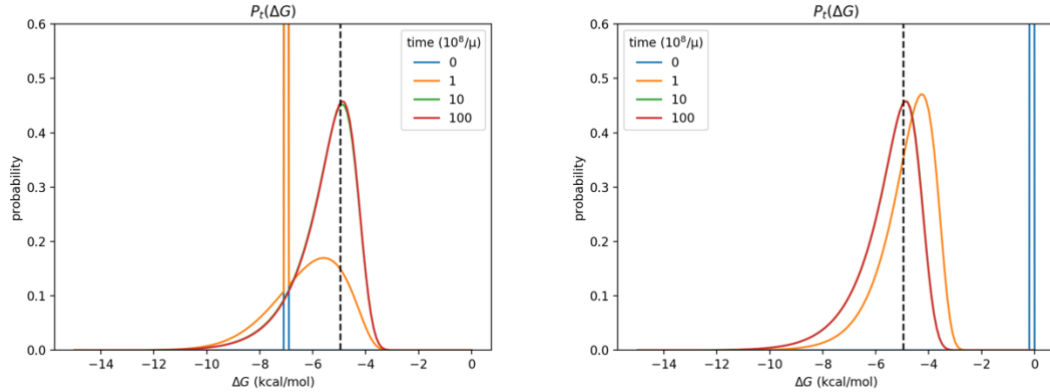


Figure 2: Solving the evolutionary dynamics equation (Equation 5) yields $P_t(\Delta G)$ distributions at steady state peaked at $\Delta G \sim -5$ kcal/mol (dashed line) under the misfolding avoidance fitness function at $A^{\text{tot}} = 10$. The left graph has an initial condition of $P_{t=0}(\Delta G = -7$ kcal/mol) the right graph, $P_{t=0}(\Delta G = 0$ kcal/mol). The green curve corresponding to timestep $10^9/\mu$, is hard to see because it overlaps with the red curve corresponding to timestep $10^{10}/\mu$. Parameter values are listed in Table 1, with $N_e = 10^8$. The initial condition is $P_{t=0}(\Delta G = -7$ or 0 kcal/mol) ~ 10 , such that $\int_{-\infty}^0 P_t(\Delta G) d(\Delta G) = 1$, and is outside the y-axis scale.

We can analytically solve for the steady state (ss) ΔG distribution ($P_{ss}(\Delta G)$) by first expressing Equation 5 as a differential equation and setting it to zero. Then, we apply the method of steepest descent to approximately obtain $P_{ss}(\Delta G)$.

To obtain a differential equation, we expand exponentials to first order in μ . This approximation is consistent with our assumption that populations evolve monoclally by assuming that at most only one mutation occurs at a given time on a protein, when sampling single mutant $\Delta\Delta G$ s ($\mu \ll 1$).

Equation 6

$$\begin{aligned}
P_{t+1}(\Delta G) &\approx (1 - \mu)P_t(\Delta G) + \mu(1 - \mu) \int_{-\infty}^{\infty} d(\Delta\Delta G)[1 - P^{\text{fix}}]p(\Delta\Delta G)P_t(\Delta G) \\
&\quad - \mu(1 - \mu) \int_{\Delta\Delta G=-\infty, \Delta\Delta G \neq 0}^{\infty} d(\Delta\Delta G)P^{\text{fix}}p(\Delta\Delta G)P_t(\Delta G) \\
&\quad + \mu(1 - \mu) \int_{-\infty}^{\infty} d(\Delta G^*)P^{\text{fix}}p(\Delta G - \Delta G^*)P_t(\Delta G^*) \\
P_{t+1}(\Delta G) &= P_t(\Delta G) \\
&\quad - \mu \left\{ \int_{-\infty}^{\infty} d(\Delta\Delta G)P^{\text{fix}}p(\Delta\Delta G)P_t(\Delta G) \right. \\
&\quad + \int_{\Delta\Delta G=-\infty, \Delta\Delta G \neq 0}^{\infty} d(\Delta\Delta G)P^{\text{fix}}p(\Delta\Delta G)P_t(\Delta G) \\
&\quad \left. - \int_{-\infty}^{\infty} d(\Delta G^*)P^{\text{fix}}p(\Delta G - \Delta G^*)P_t(\Delta G^*) \right\} \\
\frac{\partial P(\Delta G)}{\partial t} &\approx -\mu \left\{ \int_{-\infty}^{\infty} d(\Delta\Delta G)P^{\text{fix}}p(\Delta\Delta G)P(\Delta G) \right. \\
&\quad + \int_{\Delta\Delta G=-\infty, \Delta\Delta G \neq 0}^{\infty} d(\Delta\Delta G)P^{\text{fix}}p(\Delta\Delta G)P(\Delta G) \\
&\quad \left. - \int_{-\infty}^{\infty} d(\Delta G^*)P^{\text{fix}}p(\Delta G - \Delta G^*)P(\Delta G^*) \right\}
\end{aligned}$$

Setting the time derivative of $P(\Delta G)$ equal to zero, we obtain the following equation.

Equation 7

$$\begin{aligned}
P_{ss}(\Delta G) \int_{-\infty}^{\infty} d(\Delta\Delta G) P^{\text{fix}} p(\Delta\Delta G) + P_{ss}(\Delta G) \int_{\Delta\Delta G=-\infty, \Delta\Delta G \neq 0}^{\infty} d(\Delta\Delta G) P^{\text{fix}} p(\Delta\Delta G) \\
= \int_{-\infty}^{\infty} d(\Delta G^*) P^{\text{fix}} p(\Delta G - \Delta G^*) P_{ss}(\Delta G^*)
\end{aligned}$$

We solve for $P_{ss}(\Delta G)$ by applying the method of steepest descent to approximately solve each integral. Briefly, steepest descent is a useful method to approximate integrals of exponentials at their maximum (Erdelyi 1956). We evaluate the integrals at $\Delta\Delta G = 0$, or $\Delta G^* = \Delta G$. Figure 3 demonstrates that this is a good approximation for ΔG s near its optimal ΔG value of ~ -5 kcal/mol in this presented case.

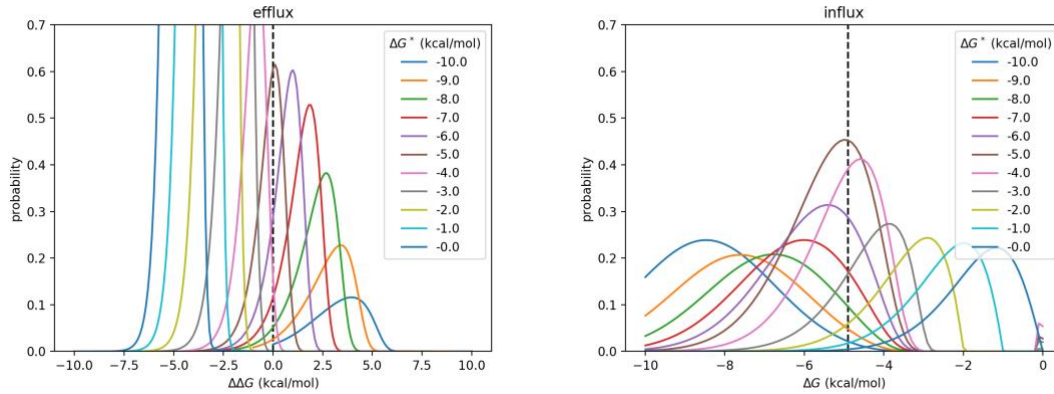


Figure 3: The method of steepest descent is a good approach to approximately solve individual integrals in Equation 7 for ΔG^* near its optimal value of -5 kcal/mol under the misfolding avoidance fitness function at $A^{\text{tot}}=10$ and $N_e = 10^8$. Efflux corresponds to the left hand side of Equation 7; influx, the right hand side of Equation 7. Both integrals are evaluated at time $= 10^9/\mu$, at which steady state has been shown to be reached (Figure 2). The efflux probabilities for $\Delta G^* > -5$ kcal/mol are outside the y-axis scale.

Applying the method of steepest descent, Equation 7 becomes,

Equation 8

$$P_{ss}(\Delta G) \sqrt{\frac{2\pi}{g''(\Delta\Delta G = 0)}} e^{g(\Delta\Delta G=0)} = \sqrt{\frac{2\pi}{h''(\Delta G^* = \Delta G)}} e^{h(\Delta G^*=\Delta G)}$$

Note that the second term of the left-hand side of Equation 7 makes no contribution because it is zero at $\Delta\Delta G = 0$. We define the functions $g(\Delta\Delta G)$ and $h(\Delta G)$ to

concisely capture integrands of the left hand and right hand side of Equation 7. g'' represents the second derivative of g with respect to $\Delta\Delta G$; h'' represents the second derivative of h with respect to ΔG^* .

Equation 9

$$g(\Delta\Delta G) = \ln P^{\text{fix}} + \ln p(\Delta\Delta G)$$

$$h(\Delta G^*) = \ln P^{\text{fix}} + \ln p(\Delta G - \Delta G^*) + \ln P_{\text{ss}}(\Delta G^*)$$

We can further simplify Equation 8, because $g(\Delta\Delta G=0) + \ln P_{\text{ss}}(\Delta G) = h(\Delta G^* = \Delta G)$.

Equation 10

$$g''(\Delta\Delta G = 0) = h''(\Delta G^* = \Delta G)$$

Evaluating second derivatives are tricky, and we show how to do it in detail for $g(\Delta\Delta G)$ in the next chapter (Equation 33). Upon evaluation we get,

Equation 11

$$\begin{aligned} & -\frac{1}{\sigma^2} + \left\{ -\frac{1}{3} \left(\frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta\Delta G} \right)^2 + \frac{\beta}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta\Delta G} \right\} \Bigg|_{\{\Delta\Delta G=0\}} \\ & = -\frac{(1 + b_1)^2}{\sigma^2} \\ & + \left\{ -\frac{1}{3} \left(\frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta G^*} \right)^2 \right. \\ & \quad \left. + \frac{\beta}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta G^*} \right\} \Bigg|_{\{\Delta G^*=\Delta G\}} + (\ln P_{\text{ss}}(\Delta G^*))''|_{\{\Delta G^*=\Delta G\}} \end{aligned}$$

Since $s'(\Delta\Delta G=0) = -s'(\Delta G^* = \Delta G)$, we can simplify Equation 11 to

Equation 12

$$\begin{aligned}
& -\frac{1}{\sigma^2} + \left\{ -\frac{1}{3} \left(\frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta G^*} \right)^2 - \frac{\beta}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta G^*} \right\} \Big|_{\{\Delta G^* = \Delta G\}} \\
& = -\frac{(1 + b_1)^2}{\sigma^2} \\
& + \left\{ -\frac{1}{3} \left(\frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta G^*} \right)^2 \right. \\
& \quad \left. - \frac{\beta}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta G^*} \right\} \Big|_{\{\Delta G^* = \Delta G\}} + (\ln P_{\text{ss}}(\Delta G^*))''|_{\{\Delta G^* = \Delta G\}} \\
0 & = -\frac{b_1^2 + 2b_1}{\sigma^2} + 2 \frac{\beta}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta G^*} \Big|_{\{\Delta G^* = \Delta G\}} + (\ln P_{\text{ss}}(\Delta G^*))''|_{\{\Delta G^* = \Delta G\}}
\end{aligned}$$

Evaluating $(\ln P^{\text{fix}}(\Delta G^* = \Delta G))' = c\beta N_e A^{\text{tot}} e^{\beta \Delta G}$ under MAH (Equation 37), we can solve the differential equation for $P_{\text{ss}}(\Delta G)$.

Equation 13

$$P_{\text{ss}}(\Delta G) = \frac{1}{Z} \exp \left[\frac{b_1(2 + b_1)}{2\sigma^2} \Delta G^2 + D\Delta G - 2cN_e A^{\text{tot}} e^{\beta \Delta G} \right]$$

The same procedure can be easily applied to find the corresponding expression under FDH because MAH and FDH fitness functions are proportional in the limit that $b/(A^{\text{tot}} k^{\text{cat}}/K^{\text{M}}) \ll 1$.

Equation 14

$$P_{\text{ss}}^{\text{FDH}}(\Delta G) = \frac{1}{Z} \exp \left[\frac{b_1(2 + b_1)}{2\sigma^2} \Delta G^2 + D\Delta G - 2 \frac{b}{A^{\text{tot}} \left(K^{\text{cat}}/K^{\text{M}} \right)} N_e e^{\beta \Delta G} \right]$$

Z is the normalization constant and D is an undetermined constant independent of ΔG . To pinpoint the value of D , we use the necessary condition to apply the method of

steepest descent that $h'(\Delta G^* = \Delta G) = 0$. In other words, the integrand is indeed maximized at ΔG .

Equation 15

$$0 = h'(\Delta G^* = \Delta G)$$

$$0 = -\frac{b_1 \Delta G + b_2}{\sigma^2} (1 + b_1) + \beta c N_e A^{\text{tot}} e^{\beta \Delta G} + (\ln P_{ss}(\Delta G^*))'|_{\{\Delta G^* = \Delta G\}}$$

$$0 = -\frac{b_1 \Delta G + b_2}{\sigma^2} (1 + b_1) + \beta c N_e A^{\text{tot}} e^{\beta \Delta G} + \frac{b_1(2 + b_1)}{\sigma^2} \Delta G + D - 2\beta c N_e A^{\text{tot}} e^{\beta \Delta G}$$

$$0 = \frac{b_1 \Delta G}{\sigma^2} - \frac{b_2 + b_1 b_2}{\sigma^2} - \beta c N_e A^{\text{tot}} e^{\beta \Delta G} + D$$

By also utilizing the condition when setting $g'(\Delta \Delta G=0) = 0$ that $(\ln P^{\text{fix}})'|_{\Delta \Delta G=0} = - (b_1 \Delta G^{\text{opt}} + b_2)/\sigma^2$ (Equation 38), we finally obtain D.

Equation 16

$$0 = \frac{\langle \Delta \Delta G \rangle}{\sigma^2} - \beta c N_e A^{\text{tot}} e^{\beta \Delta G} - \frac{2b_2 + b_1 b_2}{\sigma^2} + D$$

$$D \approx \frac{2b_2 + b_1 b_2}{\sigma^2}$$

Thus, our final $P_{ss}(\Delta G)$ expression reads,

Equation 17

$$P_{ss}(\Delta G) \propto \exp \left[\frac{b_1(2 + b_1)}{2\sigma^2} \Delta G^2 + \frac{b_2(2 + b_1)}{\sigma^2} \Delta G - 2c N_e A^{\text{tot}} e^{\beta \Delta G} \right]$$

We test the accuracy of our $P_{ss}(\Delta G)$ expression by comparing it to the numerical solution we find at steady state. Figure 4 demonstrates that we do a fairly good job of capturing numerical results. Around the optimal ΔG value of -5 kcal/mol, our distribution is a little depressed compared to numerical results. This can be explained by the inaccuracy of applying the method of steepest descent being greater at the large stability tail of the distribution. There is excess probability because our application of the method

of steepest descent starts to break down; $\Delta\Delta G$ is not very close to 0, or ΔG^* is not very close to ΔG (Figure 3). We could also fit D to take a larger value ($D \sim 0.5 \text{ (kcal/mol)}^{-1}$) such that $P_{ss}(\Delta G)$ becomes more sharply peaked. However, we find that the maximum of $P_{ss}(\Delta G)$ becomes slightly right-shifted to that of the numerical solution by $\sim 0.3 \text{ kcal/mol}$.

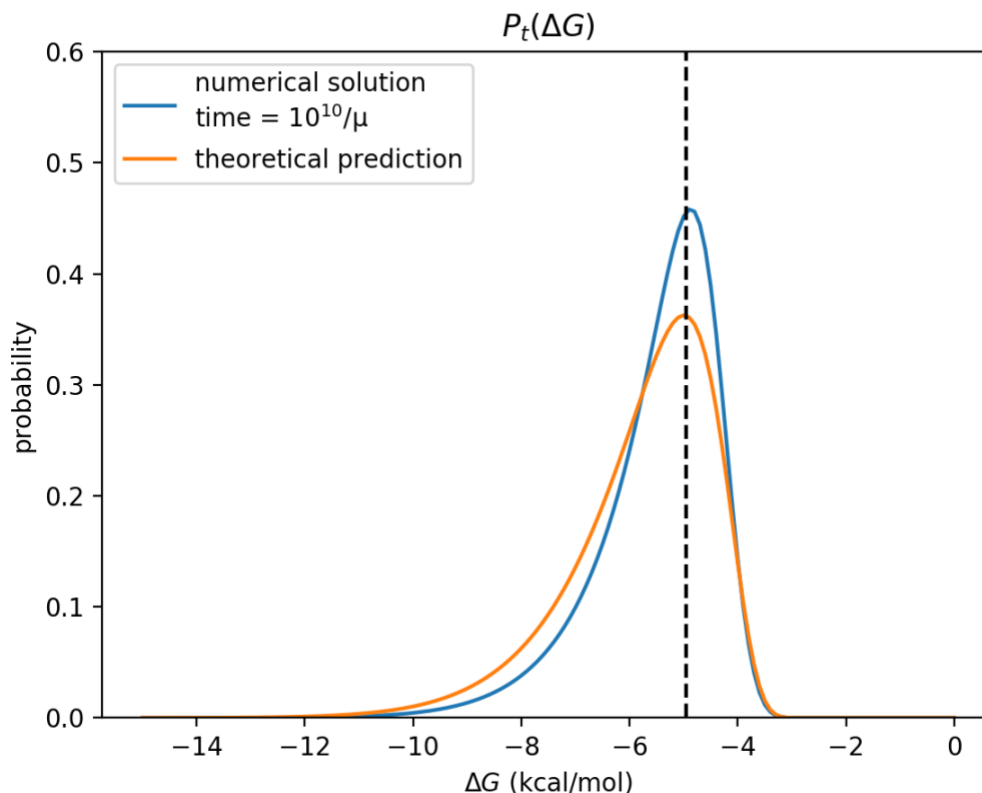


Figure 4: Our theoretical expression for $P_{ss}(\Delta G)$ fairly accurately captures numerical results. Parameter values are the same as those in Figure 2, $N_e = 10^8$ and $A = 10$ under the misfolding avoidance fitness function. The normalization constant Z was evaluated numerically.

Our analytical expression for $P_{ss}(\Delta G)$ will come in handy in Chapter 2 when we attempt to derive a closed form expression between abundance and evolutionary rate (Equation 28).

In the next section, we will compare our steady state numerical solution with experimental data to assay fitness function feasibility. A quick back of the envelope calculation demonstrates that for the *folA* orthologs we extensively study, the number of

reproductions needed for our evolutionary dynamics framework to reach steady state is orders of magnitudes less than the number of reproductions estimated to have transpired from the orthologs last common ancestor. In Figure 2, we demonstrate that steady state in $P(\Delta G)$ is reached by $10^9/\mu$ timesteps for a population size corresponding to *E.coli*. We can determine the number of reproduction events (n^r) this corresponds to by expanding μ in terms of the baseline mutation rate of an amino acid residue, the protein length (L) and the population size (N_e).

Equation 18

$$\begin{aligned} n^r &= \frac{10^9}{\mu_0 L N_e} \\ &= \frac{10^9}{10^{-10} * 100 * 10^8} = 10^9 \text{ reproductions} \end{aligned}$$

The baseline mutation rate of an amino acid residue is taken to be $\mu_0 \sim 10^{-10}$ mutations per residue (Drake et al. 1998) and the protein length ~ 100 residues (Zhang 2000). We divide by the effective population size because N_e organisms are evolving in parallel. Although population sizes of bacteria organisms are not all 10^8 , they have been shown to have values within a factor of 2 for those organisms with *folA* measurements that we study (Bobay and Ochman 2018).

According to the TimeTree database (Kumar et al. 2017), the last common ancestor of the *folA* orthologs studied here existed around 1 billion years ago. If we take the doubling time of *E.coli* in the wild to be 15 hours (584 reproductions per year) (Gibson et al. 2018) and representative across 1 billion years of evolution, we estimate that more than 10^{11} reproductions have transpired. This number is more than two orders of magnitude larger than the number of reproductions required by our evolutionary framework to reach steady state (Equation 18).

Flux dynamics unlikely to explain folA orthologs folding free energies

We focus on determining if MAH or FDH can capture experimental ΔG distributions. Previous work from our lab characterized melting temperatures for 33 mesophilic bacteria folA orthologs (Bershtein, Serohijos, et al. 2015). By fitting a linear relationship between ΔG and T_m (motivated by Equation 46 in Chapter 3), to *E.coli* folA mutants measured in Bershtein, Serohijos, et al. (2015), we transform folA orthologs T_m s to ΔG s.

When applying our evolutionary framework, we first assume that all orthologs have the same total abundance and that catalytic activity is static and does not evolve (We will analyze the validity of these assumptions later). Surprisingly we find that MAH is more likely to capture folA's ΔG distribution with a reasonable total abundance (Figure 5), despite FDH having been found to explain growth rates for *E.coli* organisms expressing these very same orthologs (Bershtein, Serohijos, et al. 2015).

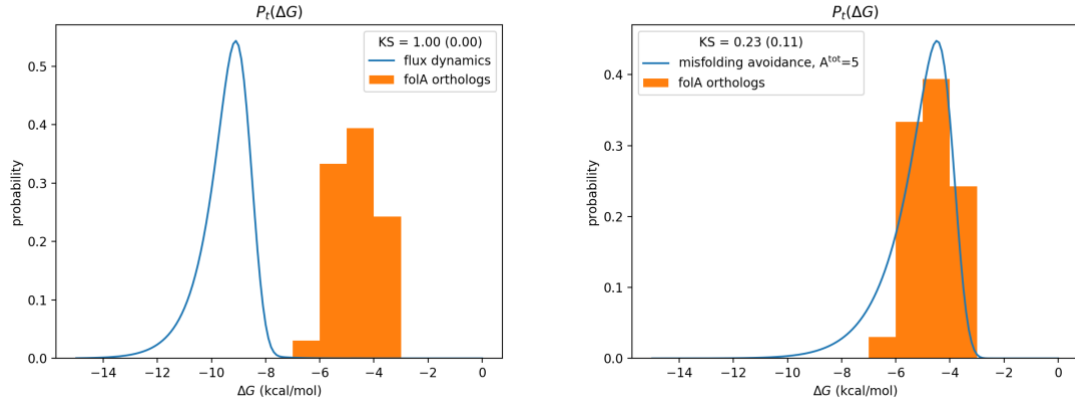


Figure 5: The misfolding avoidance fitness function (right plot) is more likely than flux dynamics (left plot) to capture folA orthologs ΔG distribution according to our evolutionary framework. KS corresponds to the Kolmogorov-Smirnov test and in parenthesis is the p-value. Higher KS and lower p-values means that the theory and experimental distributions are more likely to be distinct. Parameter values are from experimental measurements listed in Table 1.

We are aware of three studies that have measured *E.coli* folA abundance. Arike et al. (2012) found folA abundance to be 142 copies per cell; Taniguchi et al. (2011), 38;

and Krug et al. (2013), 4. Thus, the total abundance of 5 in Figure 5 is feasible, although on the low side. We further explore how well MAH captures the experimental ΔG distribution for different total abundances. There is a fairly narrow range of total abundances from 2 to 6 copies for which the theoretical ΔG distribution is indistinguishable from that of experiment according to the Kolmogorov-Smirnov test. In the next subsection, we will discuss that the range of permissible total abundances is larger when accounting for only the stability between the unfolded and intermediate state (Figure 10).

We check whether the assumption that catalytic parameters are fixed, is responsible for FDH's poor ability to match folA's ΔG distribution. We find that the prefactor of the selection coefficient, $b/(A^{\text{tot}} k^{\text{cat}}/K^M)$, needs to be 5,000 times smaller than its current value to obtain a matching ΔG distribution. In principle, catalytic activity (k^{cat}/K^M) can evolve to take larger values than the static value we used. However, none of the 30+ orthologs are seen to have catalytic activities 5,000 times larger (Figure 6, left). At most, catalytic activity of the orthologs studied are 10 times larger than that of *E.coli*, which we used as the parameter value in our evolutionary framework (Table 1). Thus, even if catalytic activity evolved, we would still be unable to qualitatively capture the experimental ΔG distribution.

We also consider whether A^{tot} can vary enough to explain the discrepancy. Taking relative abundances of folA across all bacteria organisms deposited in the Protein Abundance database (Wang et al. 2015), we find that the relative abundance to vary at most by a factor of 50. For those proteomes which have the same number of relative abundances, the variation is a lot less with all folA abundances closely equaling 50 parts

per million (Figure 6, right). This is consistent with previous experimental studies finding abundance to be an evolutionary conserved quantity (Liao and Zhang 2006; Arthur and Ruvinsky 2011; Vogel and Marcotte 2012; Wang et al. 2015; Yang et al. 2017). Relative abundance measurements are sensitive to the number of total protein measurements in the proteome (Wang et al. 2015), thus the approximately two orders of magnitude in variation could just be an artifact of the relative abundance metric. Nonetheless, this multiplicative factor of 50 variation in abundance is still not large enough, even when also accounting for $k^{\text{cat}}/K^{\text{M}}$ variation, to capture the 5,000 times difference necessary for our evolutionary framework to capture folA's ΔG distribution under FDH.

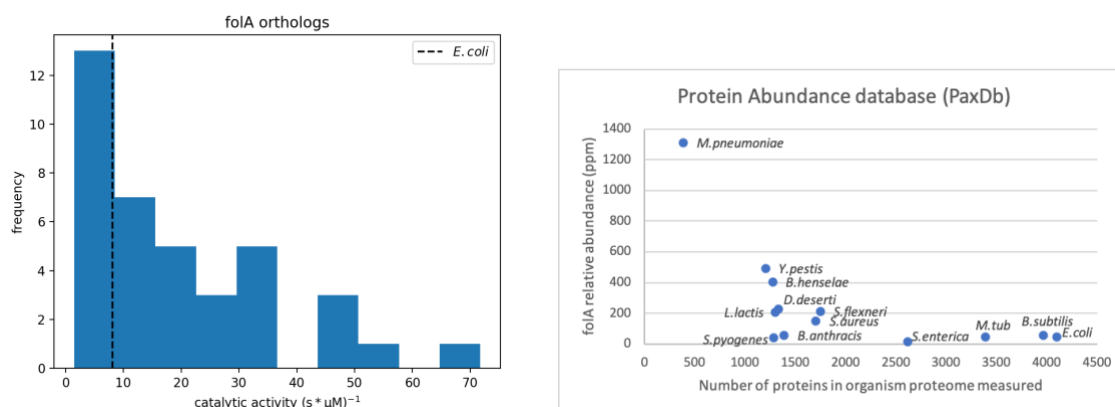


Figure 6: Model assumptions are not responsible for the poor agreement between flux dynamics fitness function and experimental stability data in Figure 5. Left plot: distribution of catalytic activities measured for the folA orthologs. Note that the *E. coli* catalytic activity value used in our evolution framework is 8.1 (s*μM)⁻¹ (dotted vertical line). Right plot: relative abundances of all bacterial folAs listed in the Protein Abundance database. Relative abundance is sensitive to the number of proteins measured in the respective organism's proteomes and is measured in parts per million (ppm) (Wang et al. 2015). *M.tub* is short for *M.tuberculosis*.

Finally, we consider whether a more relevant stability metric for FDH should account for enzyme binding to any cofactors. Ionescu et al. (2000) found that NADP⁺ stabilizes folA by around 1 kcal/mol at room temperature. Assuming that all orthologs are stabilized by the same amount of 1 kcal/mol, in such a scenario the prefactor of the selection coefficient needs to only be 500 times smaller than its current value to obtain a matching ΔG distribution. Taking into account the upper limits in A^{tot} and $k^{\text{cat}}/K^{\text{M}}$

discussed in the previous paragraphs, we can reach the threshold with a multiplicative factor of $50 \times 10 = 500$. However, as discussed before, it is dubious that A^{tot} varies by 50 times from species to species and seems to be an artifact of the relative protein abundance metric. Thus, even when accounting for ligand binding in the ΔG value, as well as variation in catalytic capacity ($A^{\text{tot}} * k^{\text{cat}}/K^{\text{M}}$), flux dynamics still seems unlikely to capture folA orthologs ΔG distribution under our evolution framework.

Unfolding avoidance better explains rnhA and folA free energies

To corroborate our findings that MAH seems to be a plausible fitness function to capture folA's long-term evolution, we apply MAH to another protein with measured stabilities. We employ ΔG measurements from Lim et al. (2016) for 5 reconstructed rnhAs along *E.coli*'s diverged branch from thermophilic organisms plus *E.coli*'s rnhA. Reconstructed rnhAs along the branch leading to the thermophiles are not included in our analysis because of variations in ambient temperature. Although not as extensive as the number of folA orthologs with available stability metrics, rnhA has enough ΔG measurements to make a statistical comparison with our theoretical distribution.

Interestingly, we find MAH to only capture rnhA ΔG measurements for a large total abundance of 10,000 copies per cell (Figure 7, left). This value for A^{tot} is unreasonable since rnhA is known to be less abundant than folA. We are only aware of one dataset that quantified the absolute number of rnhA proteins in *E.coli*. Krug et al. (2013) measured rnhA abundance at 2, compare to 4 for folA. Multiple relative abundance datasets deposited in the Protein Abundance database, corroborate the finding that folA is more abundant than rnhA (website assessed September 2020).

To try and explain the discrepancy between theory and experiment, we account for rnhA being a three-state folder across the orthologs considered (Lim et al. 2016). If MAH is valid in such a scenario, then the stability metric under selection is only that between the intermediate and native state (ΔG_{IN}), rather than that between the unfolded and native state (ΔG_{UN}) as considered before.



Just as we wrote down the probability to be in the native state (P^{nat}) for a two-state folder (Equation 1), we can write down the corresponding P^{nat} for a three-state folder and insert it into our expression for the number of misfolded proteins (Equation 21).

Equation 19

$$\begin{aligned} m(\Delta G) &= A^{\text{tot}}(1 - P^{\text{nat}}(\Delta G)) \\ &= A^{\text{tot}} \left(1 - \frac{e^{-\beta(\Delta G_{\text{UI}} + \Delta G_{\text{IN}})}}{1 + e^{-\beta\Delta G_{\text{UI}}} + e^{-\beta(\Delta G_{\text{UI}} + \Delta G_{\text{IN}})}} \right) \\ &= A^{\text{tot}} \left(\frac{1 + e^{-\beta\Delta G_{\text{UI}}}}{1 + e^{-\beta\Delta G_{\text{UI}}} + e^{-\beta(\Delta G_{\text{UI}} + \Delta G_{\text{IN}})}} \right) \\ &= A^{\text{tot}} \left(1 + \frac{e^{-\beta(\Delta G_{\text{UI}} + \Delta G_{\text{IN}})}}{1 + e^{-\beta\Delta G_{\text{UI}}}} \right)^{-1} \\ &= A^{\text{tot}} \left(1 + \frac{e^{-\beta\Delta G_{\text{IN}}}}{e^{\beta\Delta G_{\text{UI}}} + 1} \right)^{-1} \\ &\approx \frac{A^{\text{tot}}}{1 + e^{-\beta\Delta G_{\text{IN}}}} \end{aligned}$$

The approximation to yield the final line in Equation 19 requires the protein states relative to the unfolded state to be stable $\Delta G_s < 0$ kcal/mol. Note that m has the same form as when we assumed two-state folding, however instead of employing ΔG between

the unfolded and native states (Equation 21), we only need to employ the ΔG between the intermediate and native states. Swapping out this expression for m in Equation 22 yields the following selection coefficient,

Equation 20

$$s^{\text{MAH}} \approx -cA^{\text{tot}} e^{\beta \Delta G_{\text{IN}}^{\text{wt}}} (e^{\beta \Delta \Delta G_{\text{IN}}} - 1)$$

Employing direct ΔG_{IN} measurements from Lim et al. (2016), we still find that the A^{tot} needed to match the experimental ΔG distribution is unreasonable at 50 copies per cell (Figure 7, right), although much more feasible than the abundance found to capture the ΔG_{UN} distribution.

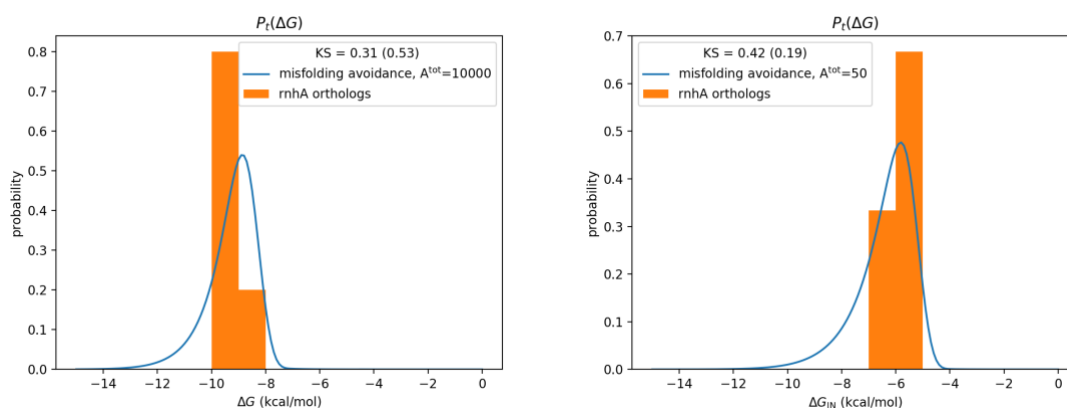


Figure 7: The misfolding avoidance fitness function under two state folding is unable to capture the ΔG distribution of rnhA orthologs with a reasonable total abundance under our evolutionary framework (left plot). In addition, accounting for rnhA being a three-state folder is still not enough to enable misfolding avoidance to capture the ΔG distribution of rnhA orthologs with a reasonable total abundance (right plot).

Finally, we consider whether the stability metric under selection is that between the unfolded and intermediate state (ΔG_{UI}). This situation corresponds to the unfolded state being toxic, while the intermediate state is harmless and also assumed to not be in equilibrium with the native state. Under the *unfolding* avoidance hypothesis (UAH), we are able to capture the experimental ΔG_{UI} distribution with the most experimentally consistent total abundance of 1 copy per cell (Figure 8).

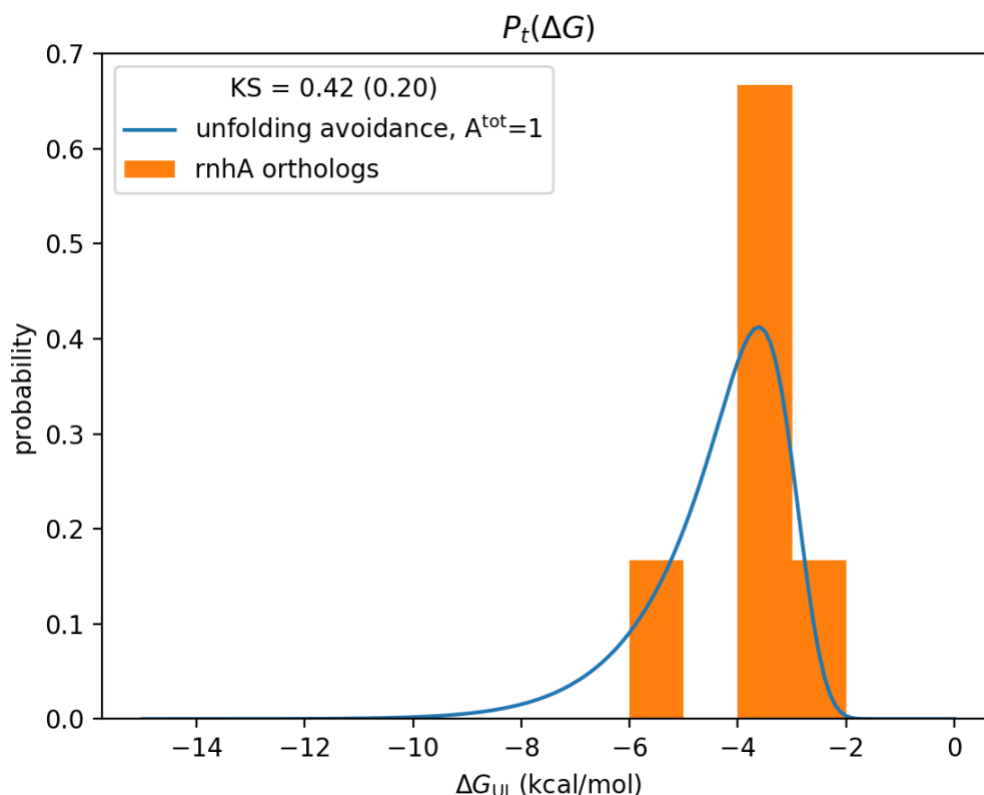


Figure 8: Under the unfolding avoidance fitness function, our evolutionary framework reproduces experimental ΔG measurements between the unfolded (U) and intermediate (I) state for *rnhA* orthologs with a reasonable total abundance. KS corresponds to the Kolmogorov-Smirnov test and in parenthesis is the *p*-value.

We return to the *folA* dataset to investigate whether the unfolded state is also the only toxic state for *folA*. *FolA*, just like *rnhA*, is a three-state folder (Ionescu et al. 2000). Previous work from our lab characterized the melting temperature of *folA*'s intermediate state through the fluorescent molecular probe, bis-1-anilinonaphthalene-8-sulfonate (ANS) (Bershtein et al. 2013). We can transform the ANS melting temperature to ΔG by applying the Gibbs-Helmholtz equation with previously measured stability parameters for the intermediate state (Ionescu et al. 2000). Applying our evolutionary framework, we find that the same $A^{\text{tot}} = 5$ for MAH in a two-state framework, is able to capture ΔG_{UI} under UAH (Figure 9). We reason that both MAH and UAH can be accurate for *folA* but not for *rnhA*, because ΔG_{IN} is smaller for *folA*. Also ΔG_{UN} for *folA* may not exactly

correspond to the unfolded to folded transition because the data we use to transform T_m to ΔG only found a two state transition under urea denaturation (Bershtein et al. 2012).

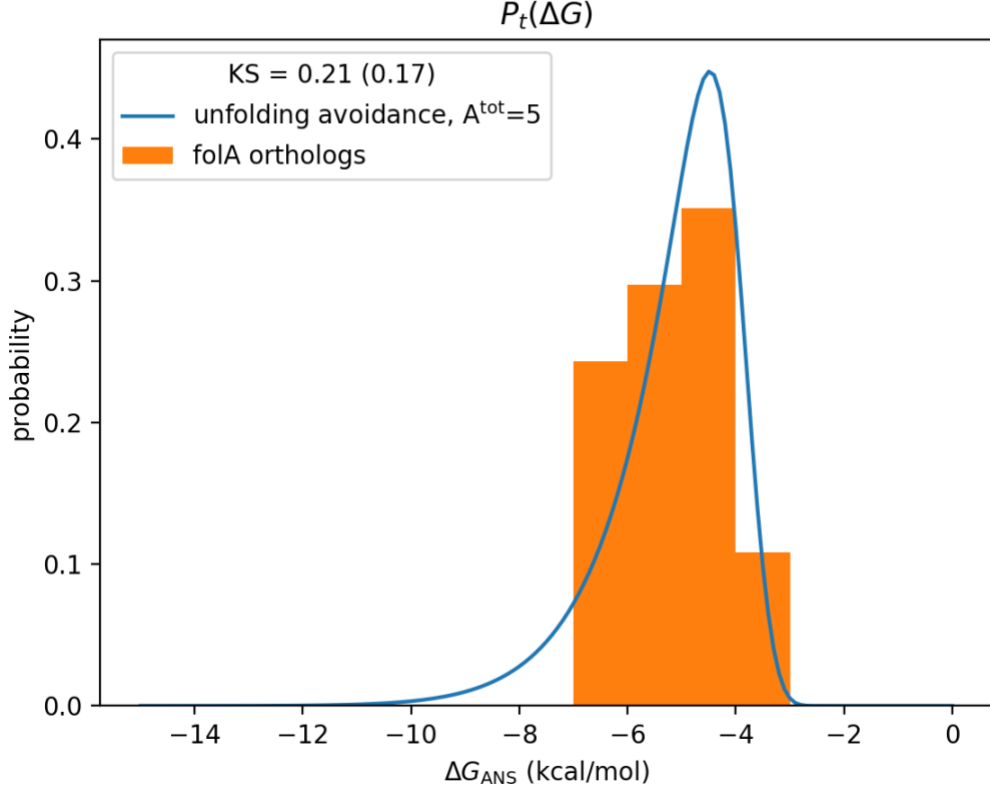


Figure 9: Our evolutionary framework reproduces experimental ΔG measurements between the unfolded and intermediate state for *folA* orthologs with a reasonable total abundance. *ANS* is a probe that detects the *folA* intermediate state (Bershtein et al. 2013). ΔG_{ANS} is a proxy for the ΔG between the unfolded and intermediate state. *KS* corresponds to the Kolmogorov-Smirnov test and in parenthesis is the *p*-value.

We investigate A^{tot} dependence on UAH's ability to match experimental ΔG_{UI} for *folA*. We find the range of permissible A^{tot} to be much wider in Figure 10 than the corresponding figure for MAH. From 5 to 15 copy numbers, A^{tot} can vary and still capture ΔG_{UIS} ; compared to the 2 to 6 copy numbers found for ΔG_{UNS} . The weaker dependence on A^{tot} for UAH compared to MAH in a two-state framework indicates that UAH is a slightly better fit to the experimental data. Protein expression in the cell is stochastic, thus it is unreasonable to assume that all orthologs have the exact same A^{tot} . Assuming protein expression to be normally distributed, *folA* was found to have a mean

expression of 38 copies and a standard deviation of 14 copies per cell (Taniguchi et al. 2011). If we rescale 38 to the 10 found to be the mean in our theory (Figure 10), the corresponding standard deviation would be 4 copies, which is well within the range of permissible A^{tot} that we found can still capture ΔG_{UIS} . It is reasonable that our A^{tot} does not perfectly match that measured in Taniguchi et al. (2011) given that biophysical parameter values used in our framework, such as the cost of a misfolded protein to the cell (c), are not fitted but are from experimental measurements on different proteins (Table 1).

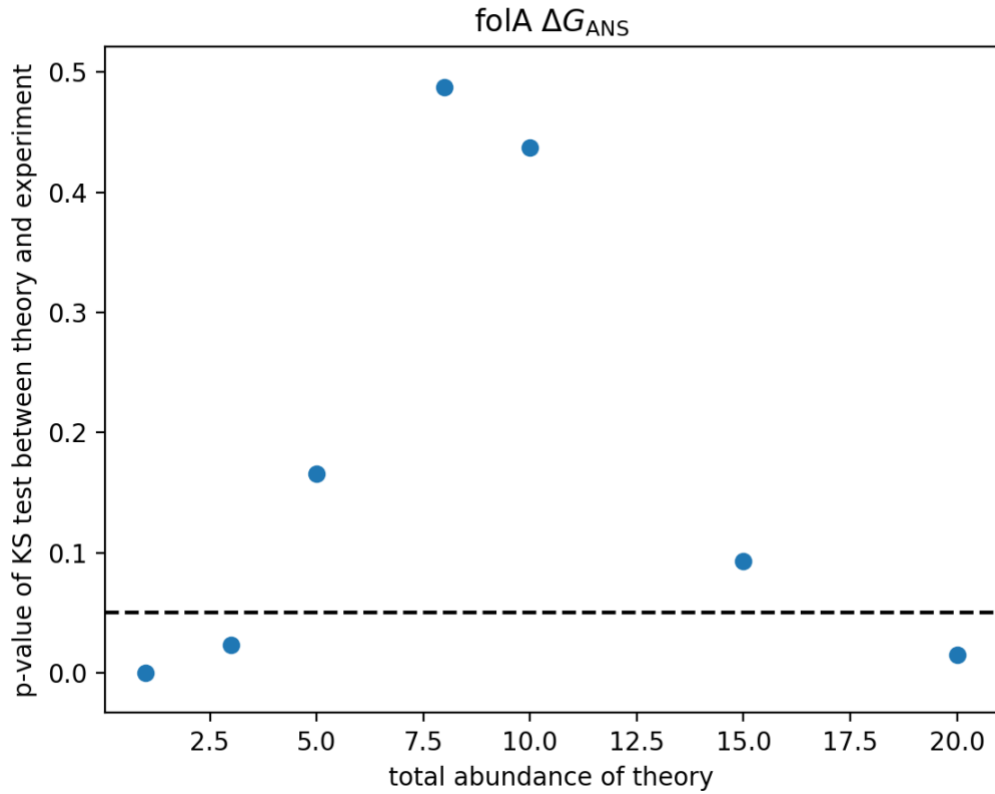


Figure 10: ΔG distributions from the unfolding avoidance hypothesis are not as sensitive to the value of the total abundance parameter compared to the misfolding avoidance hypothesis for *folA* (discussed in the text). KS corresponds to the Kolmogorov-Smirnov test between theory and experimental ΔG distribution and the corresponding p -value is plotted on the y -axis. The dashed horizontal line corresponds to a p -value threshold of 0.05. ANS is the fluorescent probe used to characterize the stability of the intermediate state in Bershtein et al. (2013).

To further corroborate that unfolding avoidance is a reasonable fitness function to explain long-term evolution, we need to apply it to other proteins besides folA and rnhA. Unfortunately, we are unaware of other studies *on multiple orthologs* that directly measured ΔG s besides rnhA (Hart et al. 2014; Lim et al. 2016). We are aware of several studies that measured melting temperatures for ancestral proteins (Hobbs et al. 2012; Akanuma et al. 2013; Risso et al. 2015; Trudeau et al. 2016), however we cannot find corresponding papers that measured relevant thermodynamic parameters to transform them into ΔG s, like we did with folA orthologs melting temperatures.

To extrapolate our evolutionary framework to other proteins such that it is experimentally testable, we derive a relationship between total abundance and evolutionary rate in the next chapter. Unlike protein stability data, abundance (A) and evolutionary rate (ER) measurements are widely available across proteins in their respective proteome (Wang et al. 2015; Zhang and Yang 2015). In addition, a robust A-ER correlation is seen among organisms' proteomes (Drummond and Wilke 2008; Zhang and Yang 2015; Razban et al. 2018). Thus, an accurate fitness function must be capable of quantitatively capturing this universal trend. Unfortunately, in this more global investigation, we cannot discriminate between misfolding avoidance or unfolding avoidance and simply assume that proteins fold in a two-state manner.

Chapter II.

Mathematical framework of the misfolding avoidance hypothesis

Protein misfolding avoidance hypothesis predictions are mostly limited to correlation signs: we expect $A-\Delta G$ to negatively correlate and $\Delta G-ER$ to positively correlate (Figure 1). Deriving an equation relating protein properties under MAH assumptions would provide a fully quantitative description. Instead of just knowing the correlation sign, we would be able pinpoint the optimal ΔG given an abundance value, and the optimal ER given the optimal ΔG . We could then obtain an overarching equation relating ER directly to A and test the accuracy of the equation experimentally.

Previous attempts have been made to fully quantify MAH (Wilke and Drummond 2006; Serohijos et al. 2013; Agozzino and Dill 2018), however, they all have shortcomings that make the final comparison with experimental data difficult to interpret. First, no study explicitly accounted for the assumption that more stable proteins have more destabilizing mutations. Second, two previous studies had additional assumptions unrelated to the two MAH assumptions, such as a priori classifying protein residues as either neutral or non-neutral (Wilke and Drummond 2006), or capturing protein folding by a hydrogen-polar model (Agozzino and Dill 2018). One study fitted biophysical parameters to match experimental data (Agozzino and Dill 2018). And only one study actually derived a closed form expression between ER and A (Wilke and Drummond 2006). These shortcomings make it difficult to assess the underlying physics when comparing with experimental data. Wilke and Drummond (2006) derived an expression

in disagreement with experimental data, while Agozzino and Dill (2018) numerically solved their expression with fitted parameters to match experimental evolutionary rates as a function of mRNA abundances. It is unclear whether one theory disproved MAH by showing that it does not correspond to reality, while another theory with different additional assumptions and fitted parameters proved its feasibility.

Here, we derive analytical expressions between all three protein properties – A , ΔG and ER – according to MAH’s two assumptions with experimentally measured parameters (Geiler-Samerotte et al. 2010; Serohijos et al. 2012). We make only one additional assumption: organisms evolve monoclonally, which has bioinformatics support (Lynch and Conery 2003) (Equation 24). We also use experimental measurements for parameters, allowing us to confidently claim that any agreement with experimental trends is the result of the underlying physics.

Mathematically, we can express MAH assumptions as follows.

Assumption 1) Misfolded proteins are toxic (underlies A - ΔG relationship)

We have already presented the definition of the MAH fitness function in the previous chapter (Equation 3). Here, we explicitly evaluate terms relevant to evaluating its selection coefficient. From equilibrium statistical mechanics and assuming two-state folding (Drummond and Wilke 2008), the number of misfolded proteins (Equation 3) can be evaluated as

Equation 21

$$\begin{aligned} m &= A^{\text{tot}} - A \\ &= A^{\text{tot}} - A^{\text{tot}} \left(\frac{1}{1 + e^{\beta \Delta G}} \right) = \frac{A^{\text{tot}}}{1 + e^{-\beta \Delta G}} \end{aligned}$$

The variable A refers to the soluble protein abundance and is the abundance variable that is typically compared with ER in bioinformatics studies. A^{tot} more generally encompasses soluble and insoluble protein abundance and is a constant independent of ΔG . Because almost all proteins have $\Delta G < -3$ kcal/mol at room temperature (Zeldovich et al. 2007; Ghosh and Dill 2009), P^{nat} is very close to 1 and thus A^{tot} and A are interchangeable in the final expression.

We can now write down the equation for the selection coefficient which is defined as $s = \ln f^{\text{mut}} - \ln f^{\text{wt}}$ (Sella and Hirsh 2005). The wildtype protein sequence is referred to as wt; mutant, mut.

Equation 22

$$s = -cA^{\text{tot}} \left(\frac{1}{1 + e^{-\beta(\Delta G^{\text{wt}} + \Delta\Delta G)}} - \frac{1}{1 + e^{-\beta\Delta G^{\text{wt}}}} \right)$$

$$\approx -cA^{\text{tot}} e^{\beta\Delta G^{\text{wt}}} (e^{\beta\Delta\Delta G} - 1)$$

To go from the first to the final expression on the second line, we use the fact that most proteins have $\Delta G < -3$ kcal/mol (Zeldovich et al. 2007; Ghosh and Dill 2009; Serohijos et al. 2012; Gromiha et al. 2016).

Assumption 2) Distribution of stability effects (underlies ΔG -ER relationship)

Theoretical studies originally approximated $p(\Delta\Delta G)$ by a Gaussian distribution based on deposited $\Delta\Delta G$ data for different proteins, with a mean of 1 kcal/mol and a standard deviation (σ) of 1.7 kcal/mol (Zeldovich et al. 2007; Chen and Shakhnovich 2009; Wylie and Shakhnovich 2011). However, a closer analysis with more data found that $\Delta\Delta G$ s are influenced by the wildtype protein's ΔG . Serohijos et al. (2012) found that $p(\Delta\Delta G)$'s mean is better described as $\langle\Delta\Delta G\rangle = b_1 * \Delta G + b_2$, where b_1 and b_2 were empirically determined to be -0.13 and 0.23 kcal/mol, respectively. This ΔG - $\Delta\Delta G$

anticorrelation is critical for recovering the MAH-expected ΔG -ER correlation, and thus the overall A-ER anticorrelation in simulations (Serohijos et al. 2012).

Equation 23

$$p(\Delta\Delta G[\Delta G]) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(\Delta\Delta G - [b_1\Delta G + b_2])^2}{2\sigma^2}\right]$$

We adopt the convention that more stable proteins have more negative ΔG values. Thus, $\Delta\Delta G < 0$ refers to a stabilizing mutation.

Probability of fixation (additional assumption of our theory)

We define P^{fix} based on Kimura's equation (Kimura 1962).

Equation 24

$$P^{\text{fix}}(s) = \frac{1 - e^{-2s}}{1 - e^{-2N_e s}}$$

Kimura's equation describes monoclonal populations evolving with an effective population size of N_e . The condition of monoclonality, $N_e\mu_0 \ll 1$, where μ_0 is the mutation rate per amino acid sequence, has been shown to be satisfied for many organisms (Lynch and Conery 2003).

Closed-form evolutionary rate expression as a function of abundance

We define ER for a given protein based on the relative number of amino acid changes between itself and an orthologous protein (Zhang and Yang 2015).³

³ We do not define ER with respect to DNA sequences to avoid complications surrounding synonymous mutations being neutral (Wall et al. 2005; Zhou et al. 2010; Jacobs and Shakhnovich 2017; Bhattacharyya et al. 2018).

Equation 25

$$ER = \frac{N^{\text{subs}}}{L}$$

N^{subs} is the number of amino acid substitutions and L is the alignment length of the two proteins.

The challenge in calculating ER is its inherent stochasticity. Beneficial mutations may be unlucky and not fix, while highly destabilizing mutations may beat the odds and fix by random chance. We first calculate the average ER at a given moment in time to smooth over the stochasticity. We express the average N^{subs} at a given moment in time t by the normalized sum over probabilities that a single mutation fixes in the population (P^{fix}) over all possible single mutations (N_t^{sample}). This averaging approach smooths over the stochasticity involving mutation fixation.

Equation 26

$$\begin{aligned} \langle ER_t \rangle &= \frac{\langle N_t^{\text{subs}} \rangle}{L} \\ &= \frac{1}{L} \frac{1}{N_t^{\text{sample}}} \sum_i^{N_t^{\text{sample}}} P^{\text{fix}}(s_i) \end{aligned}$$

Equation 26 is still general and could be used with any selection coefficient. We now narrow down Equation 26 by having s dependent only on protein stability. We refine the summation term to be over stability effects ($\Delta\Delta G$) or equivalently an integral over the $\Delta\Delta G$ probability distribution, assuming $\Delta\Delta G$ s are well sampled. Based on our investigation in the previous section, the time dependence comes into play by considering all possible ΔG states of the protein at time t .

Equation 27

$$\begin{aligned}\langle ER_t \rangle &= \frac{1}{L} \frac{1}{N_t^{\text{state}}} \sum_{\Delta G_j}^{N_t^{\text{state}}} \frac{1}{N^{\text{sample}}} \sum_{\Delta \Delta G_k}^{N^{\text{sample}}} P^{\text{fix}}(s[\Delta G_j, \Delta \Delta G_k]) \\ &= \frac{1}{L} \int_{-\infty}^0 d(\Delta G) P_t(\Delta G) \int_{-\infty}^{\infty} d(\Delta \Delta G) P^{\text{fix}}(s[\Delta G, \Delta \Delta G]) p(\Delta \Delta G[\Delta G])\end{aligned}$$

Normalization terms, N_t^{state} and N^{sample} , get subsumed by their respective probability distributions, $P_t(\Delta G)$ and $p(\Delta \Delta G)$.

We can take advantage of our analytical expression for $P(\Delta G)$ at steady state (ss) (Equation 17) to estimate $\langle ER \rangle$, the average evolutionary rate across time, by multiplying $\langle ER_{ss} \rangle$ by a scaling constant N^{atts} .

Equation 28

$$\begin{aligned}\langle ER \rangle &\approx N^{\text{atts}} \langle ER_{ss} \rangle \\ &= \frac{N^{\text{atts}}}{L} \int_{-\infty}^0 d(\Delta G) P_{ss}(\Delta G) \int_{-\infty}^{\infty} d(\Delta \Delta G) P^{\text{fix}}(s[\Delta G, \Delta \Delta G]) p(\Delta \Delta G[\Delta G])\end{aligned}$$

This approximation is key to getting an expression that we can compare to ER data. N^{atts} is the number of attempted N^{sample} s throughout evolution. It will prove to be the only parameter with no experimental measurement because it is difficult to go back in time to determine when steady state is reached for a given protein. However, it will be shown to be a multiplicative constant of our final ER expression, thus not altering correlation coefficients to assess its applicability to experimental A, ER data.

We can analytically solve Equation 28 by first using the method of steepest descent to approximately solve the $\Delta \Delta G$ integral.

Equation 29

$$\begin{aligned}\langle ER \rangle &= \frac{N_{\text{atts}}}{LZ} \int_{-\infty}^0 d(\Delta G) P_{\text{ss}}(\Delta G) \int_{-\infty}^{\infty} d(\Delta \Delta G) P^{\text{fix}}(s(\Delta G, \Delta \Delta G)) p(\Delta G, \Delta \Delta G) \\ &\approx \frac{N_{\text{atts}}}{LZ} \frac{1}{N_e} \int_{-\infty}^0 d(\Delta G) P_{\text{ss}}(\Delta G) \sqrt{\frac{2\pi}{g''(\Delta \Delta G = 0)}} e^{-\langle \Delta \Delta G \rangle^2 / (2\sigma^2)}\end{aligned}$$

To further progress, we treat the second derivative term as a constant and expand the exponent around $\Delta G = \Delta G^{\text{opt}}$ to first order in ΔG .

Equation 30

$$\begin{aligned}\langle ER \rangle &\approx \frac{N_{\text{atts}}}{LZ} \frac{1}{N_e} \sqrt{\frac{2\pi}{g''(\Delta \Delta G = 0)}} \int_{-\infty}^0 d(\Delta G) P_{\text{ss}}(\Delta G) \left(e^{-(b_1 \Delta G^{\text{opt}} + b_2)^2 / (2\sigma^2)} \right. \\ &\quad \left. - \frac{b_1^2}{\sigma^2} \Delta G e^{-(b_1 \Delta G^{\text{opt}} + b_2)^2 / (2\sigma^2)} \right) \\ &= \frac{N_{\text{atts}}}{LZ} \frac{1}{N_e} \sqrt{\frac{2\pi}{g''(\Delta \Delta G = 0)}} e^{-(b_1 \Delta G^{\text{opt}} + b_2)^2 / (2\sigma^2)} \left(Z - \int_{-\infty}^0 d(\Delta G) P_{\text{ss}}(\Delta G) \frac{b_1^2}{\sigma^2} \Delta G \right)\end{aligned}$$

We can solve the remaining integral by the fundamental theorem of calculus because $(\ln P_{\text{ss}}(\Delta G))' \sim 2b_1 \Delta G / \sigma^2$.

Equation 31

$$\begin{aligned}\langle ER \rangle &\approx \frac{N_{\text{atts}}}{LZ} \frac{1}{N_e} \sqrt{\frac{2\pi}{g''(\Delta \Delta G = 0)}} e^{-\frac{(b_1 \Delta G^{\text{opt}} + b_2)^2}{2\sigma^2}} \left(Z \right. \\ &\quad \left. - \int_{-\infty}^0 d(\Delta G) P_{\text{ss}}(\Delta G) \frac{b_1}{2} (\ln P_{\text{ss}}(\Delta G))' \right) \\ &= \frac{N_{\text{atts}}}{LZ} \frac{1}{N_e} \sqrt{\frac{2\pi}{g''(\Delta \Delta G = 0)}} e^{-(b_1 \Delta G^{\text{opt}} + b_2)^2 / (2\sigma^2)} \left(Z - \frac{b_1}{2} P_{\text{ss}}(\Delta G) \Big|_{-\infty}^0 \right)\end{aligned}$$

Finally, we note that the second derivative term times the exponential term corresponds to the second derivative term of $P_{\text{ss}}(\Delta G)$ if we were to apply the method of

steepest descent. Thus, we apply the ‘inverse’ method of steepest descent to obtain our final result and get rid of the normalization constant Z.

Equation 32

$$\begin{aligned}\langle ER \rangle &\approx \frac{N_{\text{atts}}}{L} \frac{1}{N_e} \left(\sqrt{\frac{2\pi}{g''(\Delta\Delta G = 0)}} e^{-(b_1\Delta G^{\text{opt}} + b_2)^2/(2\sigma^2)} - \frac{b_1}{2Z} \int_{-\infty}^0 d(\Delta G) P_{\text{ss}}(\Delta G) \right) \\ &= \frac{N_{\text{atts}}}{L} \frac{1}{N_e} \left(\sqrt{\frac{2\pi}{g''(\Delta\Delta G = 0)}} e^{-(b_1\Delta G^{\text{opt}} + b_2)^2/(2\sigma^2)} - \frac{b_1}{2} \right)\end{aligned}$$

Another approach we could have taken to get an analytical expression for $\langle ER \rangle$ is to do the method of steepest descent in two dimensions: ΔG and $\Delta\Delta G$. The advantage of our current approach is that the normalization of $P_{\text{ss}}(\Delta G)$ is removed from our final expression. Z is tricky to work with because we cannot solve for it analytically and it depends on the abundance.

We now begin to evaluate $g''(\Delta\Delta G)$.

Equation 33

$$g''(\Delta\Delta G = 0) = -\frac{1}{\sigma^2} + \frac{\partial}{\partial\Delta\Delta G} \left\{ \frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial\Delta\Delta G} \right\} \Big|_{\{\Delta\Delta G=0\}}$$

The challenge of evaluating $g''(\Delta\Delta G = 0)$ lies the second term in Equation 33.

Equation 34

$$\begin{aligned}&\frac{\partial}{\partial\Delta\Delta G} \left\{ \frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial\Delta\Delta G} \right\} \\ &= -\left(\frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial\Delta\Delta G} \right)^2 + \frac{1}{P^{\text{fix}}} \frac{\partial^2 P^{\text{fix}}}{\partial s^2} \left(\frac{\partial s}{\partial\Delta\Delta G} \right)^2 + \frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial^2 s}{\partial\Delta\Delta G^2}\end{aligned}$$

Instead of explicitly evaluating Equation 34, we evaluate second derivatives at $\Delta\Delta G = 0$ in terms of first derivatives at $\Delta\Delta G=0$. This approach yields the following two identities,

Equation 35

$$\left. \frac{\partial^2 P^{\text{fix}}}{\partial s^2} \right|_{\{\Delta\Delta G=0\}} = \frac{2}{3} \left(-3 + \frac{2}{N_e} + N_e \right) \approx \frac{2}{3} \frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \left. \frac{\partial P^{\text{fix}}}{\partial s} \right|_{\{\Delta\Delta G=0\}}$$

$$\left. \frac{\partial^2 s}{\partial \Delta\Delta G^2} \right|_{\{\Delta\Delta G=0\}} = \beta \left. \frac{\partial s}{\partial \Delta\Delta G} \right|_{\{\Delta\Delta G=0\}}$$

Inserting these identities into Equation 34 evaluated at $\Delta\Delta G = 0$,

Equation 36

$$\left. \frac{\partial}{\partial \Delta\Delta G} \left\{ \frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta\Delta G} \right\} \right|_{\{\Delta\Delta G=0\}}$$

$$= \left\{ - \left(\frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta\Delta G} \right)^2 + \frac{2}{3} \left(\frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta\Delta G} \right)^2 \right.$$

$$\left. + \frac{\beta}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta\Delta G} \right\} \left. \right|_{\{\Delta\Delta G=0\}}$$

$$= \left\{ - \frac{1}{3} \left(\frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta\Delta G} \right)^2 + \frac{\beta}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta\Delta G} \right\} \left. \right|_{\{\Delta\Delta G=0\}}$$

To finally evaluate Equation 36 in terms of protein properties, we take advantage of the condition that the integrand is indeed maximized at $\Delta\Delta G=0$ ($g'(\Delta\Delta G=0)=0$), as is required by applying the method of steepest descent in this manner.

Equation 37

$$g'(\Delta\Delta G = 0) = -\frac{\Delta\Delta G - \langle\Delta\Delta G\rangle}{\sigma^2}\bigg|_{\{\Delta\Delta G=0\}} + \frac{1}{P^{\text{fix}}}\frac{\partial P^{\text{fix}}}{\partial s}\frac{\partial s}{\partial\Delta\Delta G}\bigg|_{\{\Delta\Delta G=0\}}$$

$$0 = \frac{b_1\Delta G^{\text{opt}} + b_2}{\sigma^2} + N_e\left(\frac{N_e - 1}{N_e}\right) (-c\beta A^{\text{tot}}e^{\beta\Delta G^{\text{opt}}})$$

We approximate N_e-1 as N_e since the effective population size of many organisms is very large (Table 1). Then, we expand $\langle\Delta\Delta G\rangle$ as $b_1\Delta G^{\text{opt}} + b_2$ and express Equation 37 in terms of ΔG^{opt} .

Equation 38

$$0 = \frac{b_1\Delta G^{\text{opt}} + b_2}{\sigma^2} - c\beta N_e A^{\text{tot}}e^{\beta\Delta G^{\text{opt}}}$$

$$\rightarrow \Delta G^{\text{opt}} = -\frac{b_2}{b_1} - \frac{1}{\beta} W\left[\frac{c\beta A^{\text{tot}}N_e\sigma^2 e^{-\beta b_2/b_1}}{-b_1}\right]$$

W is the Lambert W function, also known as the product logarithm. We can check the accuracy of Equation 38 by first numerically solving the integral involving $g(\Delta\Delta G)$ over $\Delta\Delta G$ in Equation 9 and then taking the minimum value of the integral along the three dimensional surface (blue points in Figure 11). Those points correspond to the smallest ERs that can fix for a protein with given A^{tot} s and ΔG^{opt} s, where $\Delta\Delta G = 0$. We find good agreement between our analytical expressions and numerical results (Figure 12).

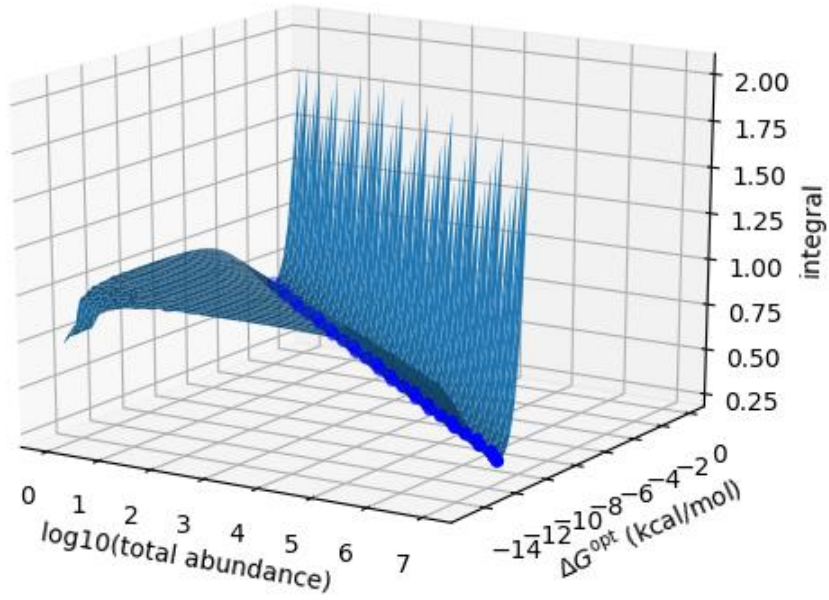


Figure 11: Dependence of wildtype folding free energy (ΔG^{opt}) and total abundance under the protein misfolding avoidance hypothesis. The three-dimensional surface is obtained by numerically solving the integral $= \int_{-\infty}^{\infty} d(\Delta\Delta G) P^{\text{fix}} p(\Delta\Delta G)$ with parameter values listed in Table 1 and $N_e = 10^7$. For clarity, integral values were multiplied by N_e and those greater than two are cut off. This figure is the same 'molecular clock surface' presented in Serohijos et al. (2013).

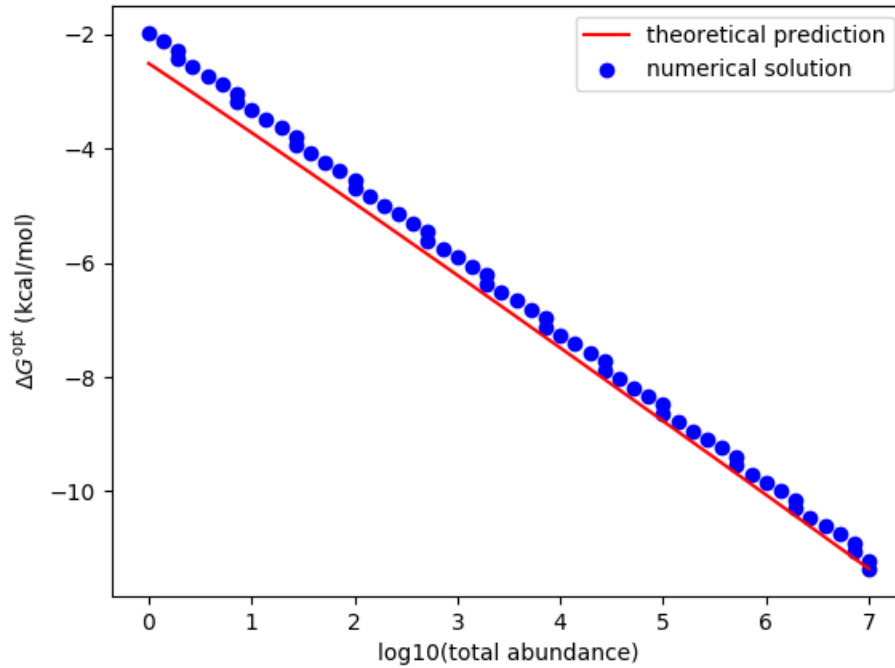


Figure 12: Our analytical expression between ΔG^{opt} and A^{tot} (red line) aptly capture numerical results at optimal integral values (blue points in Figure 11).

With Equation 38, we can now explicitly evaluate Equation 36 as

Equation 39

$$\left. \frac{\partial}{\partial \Delta \Delta G} \left\{ \frac{1}{P^{\text{fix}}} \frac{\partial P^{\text{fix}}}{\partial s} \frac{\partial s}{\partial \Delta \Delta G} \right\} \right|_{\{\Delta \Delta G=0\}} = -\frac{1}{3} \left(\frac{b_1 \Delta G^{\text{opt}} + b_2}{\sigma^2} \right)^2 + \beta \frac{b_1 \Delta G^{\text{opt}} + b_2}{\sigma^2}$$

This finally leads us to evaluating our $\langle ER \rangle$ expression (Equation 32) as,

Equation 40

$$\langle ER \rangle = \frac{N^{\text{atts}}}{L} \frac{1}{N_e} \left(-\frac{b_1}{2} + \frac{e^{-\langle \Delta \Delta G \rangle^2 / (2\sigma^2)}}{\sqrt{1 + \frac{\langle \Delta \Delta G \rangle^2}{3\sigma^2} + \beta \langle \Delta \Delta G \rangle}} (\Delta G = \Delta G^{\text{opt}}) \right)$$

By utilizing the result that $(\ln P^{\text{fix}})'|_{\Delta \Delta G=0} = -(b_1 \Delta G^{\text{opt}} + b_2)/\sigma^2$ (Equation 38), we can cast Equation 40 only in terms of A^{tot} .

Equation 41

$$\langle ER \rangle \approx \frac{N^{\text{atts}}}{L} \frac{1}{N_e} \left(-\frac{b_1}{2} + \frac{e^{-(A^{\text{tot}} Y)^2 / 2}}{\sqrt{1 + \frac{1}{3} (A^{\text{tot}} Y)^2 + A^{\text{tot}} Y \beta \sigma}} \right)$$

Y is a dimensionless parameter that indirectly depends on abundance through ΔG^{opt} .

Equation 42

$$Y = \frac{b_1 \Delta G^{\text{opt}} + b_2}{\sigma^2} \frac{\sigma}{A^{\text{tot}}} = c \beta N_e \sigma e^{\beta \Delta G^{\text{opt}}}$$

Figure 13 shows that our expression (Equation 41) is fairly accurate when compared with the exact numerical solution.

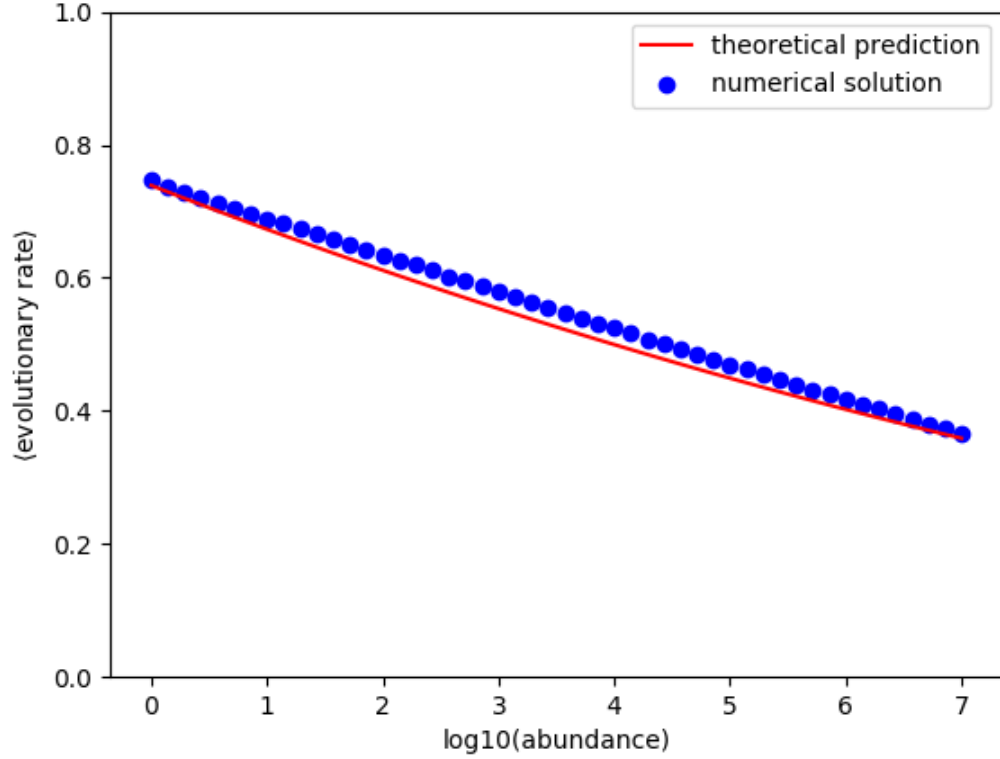


Figure 13: Our analytical expression (red line) aptly captures numerical results. Parameter values are listed in Table 1. N^{atts}/L was set to N_e , and $N_e = 10^7$.

Just as when finding an analytical expression for $P_{ss}(\Delta G)$ (Equation 14), we can find the corresponding analytical expression for $\langle ER \rangle$ as a function of A^{tot} under FDH because the selection coefficients are proportional in the $b/(A^{\text{tot}} \cdot k^{\text{cat}}/K^M) \ll 1$ limit. The overall expression is the same as that of Equation 41 except the expression for ΔG^{opt} changes, making the Y parameter different between FDH and MAH.

Equation 43

$$\Delta G^{\text{FDH,opt}} = -\frac{b_2}{b_1} - \frac{1}{\beta} W \left[\frac{c\beta b N_e \sigma^2 e^{-\beta b_2/b_1}}{-A^{\text{tot}} \left(k^{\text{cat}}/K^M \right) b_1} \right]$$

In the next section, we will not pursue whether experimental trends follow FDH predictions. Although k^{cat} has been measured for more than a 100 *E.coli* proteins (Davidi

et al. 2016), the b parameter means largely unknown. We are only aware of one other study besides our research group's work on *folA* (Bershtein, Serohijos, et al. 2015), that has determined the value of b experimentally (Jiang et al. 2013). If b was estimated to be roughly constant, FDH would predict a *positive* A-ER correlation, in complete contrast to what is experimentally observed. Thus, for FDH to have a chance at accurately capturing the A-ER anticorrelation observed across organisms' proteomes, b would have to vary between proteins. However, we currently have very limited experimental data to investigate this point.

Misfolding avoidance could underly proteome-wide trends

Despite the lack of ΔG data, with Equation 41 we can assess whether MAH is consistent with experimental data by analyzing the A-ER relationship. By taking advantage of the ΔG distribution being at steady state for timescales at which the A-ER correlation is usually studied (> 10 million years, Equation 18), Equation 41 calculates the number of mutations that fix while still maintaining steady-state on average (Equation 28). This is consistent with the neutral theory of molecular evolution (Kimura 1968; Koonin 2009; Zhang 2010), where we have quantitatively characterized neutral mutations based on biophysical parameters. It is worth emphasizing that our theory does not simply capture mutations that are only neutral. Beneficial mutations are accounted for within our framework but do not dictate trends at the steady state condition of mutation-selection balance because they are cancelled out by equally detrimental mutations that fix, making $\langle \Delta \Delta G \rangle = 0$ (Figure 14) (Sella and Hirsh 2005; Mustonen and Lässig 2009; Serohijos and Shakhnovich 2014).

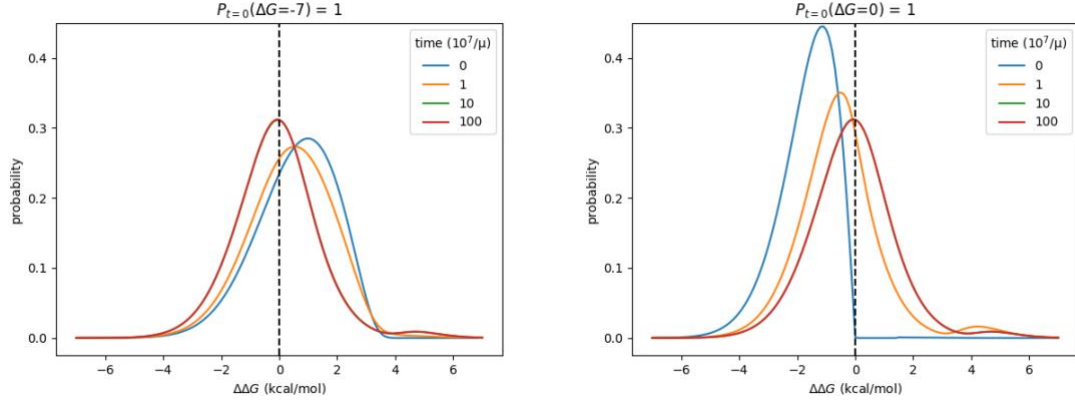


Figure 14: The $\Delta\Delta G$ distribution of fixing mutations ($\int_{-\infty}^{\infty} d(\Delta G) P^{\text{fix}}(\Delta G, \Delta\Delta G) p(\Delta\Delta G) P_t(\Delta G)$) reaches mutation-selection balance at $10^8/\mu$ timesteps and is centered at $\Delta\Delta G=0$. The left graph has an initial condition of $P_{t=0}(\Delta G=-7 \text{ kcal/mol})$; the right graph, $P_{t=0}(\Delta G=0 \text{ kcal/mol})$. Parameter values are the same as those in Table 1 for MAH with $N_e = 10^7$ and $A^{\text{tot}} = 100$.

We analyze A and ER data from *E.coli*, *S.cerevisiae* and *H.sapiens* deposited in the ProteomeVis database (Razban et al. 2018). Note that experiments typically measure soluble abundance (A), while our expression is for total abundance (A^{tot}). Because almost all proteins have $\Delta G < -3 \text{ kcal/mol}$ at room temperature (Zeldovich et al. 2007; Ghosh and Dill 2009), P^{nat} is very close to 1 and thus A^{tot} and A are interchangeable. All parameters in Equation 41 are experimentally known except for N^{atts} (Table 1). Although N^{atts} is protein dependent, N^{atts}/L is arguably protein independent because longer proteins have more attempted mutations - $N^{\text{atts}} \propto L$. Since correlations are unaffected by multiplicative constants, the exact value of N^{atts}/L is irrelevant in assessing MAH accuracy in this manner.

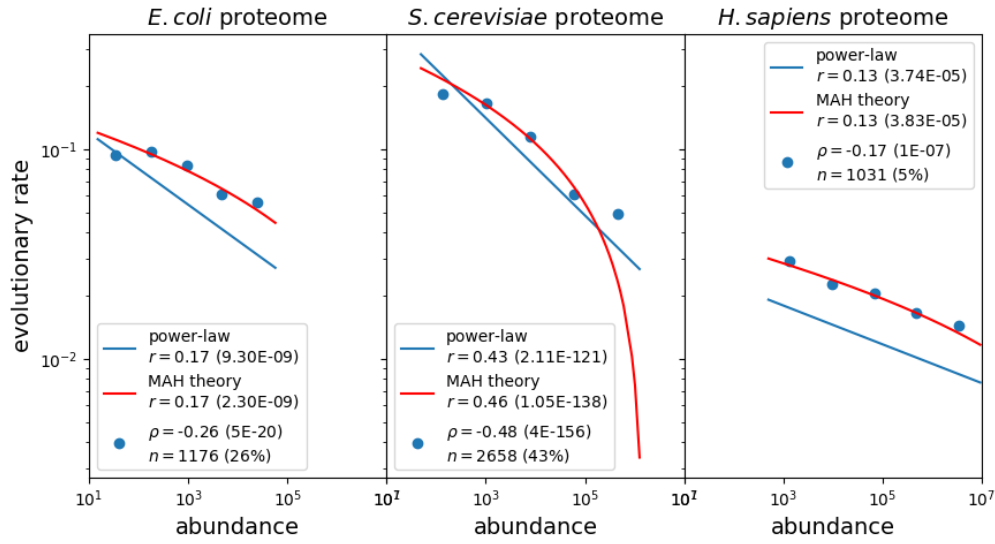


Figure 15: Our linearly fitted analytical expression, called MAH theory (red line), captures experimental trends for three organisms' proteomes at least just as well as an empirical power-law fit (blue line). r refers to the Pearson correlation between predicted and experimental evolutionary rate, in parenthesis is the p -value. $\rho_{A,ER}$ refers to the Spearman correlation between abundance and evolutionary rate, in parenthesis is the p -value. n refers to the number of proteins with abundance and evolutionary rate data, in parenthesis is the percentage of the proteome n captures. We only present the mean of binned datapoints for clarity; however, correlations are calculated by considering all datapoints.

From Figure 15, we see that MAH predicted evolutionary rates (Equation 41) capture overall proteome-wide A-ER trends. Pearson correlations between predicted evolutionary rates and experimental evolutionary rates are significant for all three organisms' proteomes tested (Figure 15). The correlation is strongest for *S.cerevisiae*, which is a result of its strong A-ER correlation. Likewise, the correlations for *E.coli* and *H.sapiens* are weaker because of the weaker A-ER correlation. The predictive power of our theory is limited by the A-ER correlation because A is the only variable on the right hand side Equation 41.

For reference, we compare MAH theory (red line) to a power-law fit (blue line in Figure 15), the best known empirical relationship between A and ER (Drummond et al. 2005). MAH theory has equivalent (*E.coli* and *H.sapiens*) or slightly better (*S.cerevisiae*) Pearson correlations than that of the power-law fit when including all datapoints in the correlation calculation. Matching correlation coefficients with empirical fits indicate that

the underlying physics of MAH could explain the universal A-ER correlation. Still, the corresponding curves are not identical and we are not aware of a way to mathematically reduce the MAH theory A-ER expression to a power-law form. Comparing best-fit curves, MAH theory always predicts larger ERs compared to a power-law fit for *E.coli* and *H.sapiens* proteins (red line above blue line in Figure 15). For *S.cerevisiae*, the trend is not as clear. For intermediate abundances, MAH theory predicts larger ERs; however, for small or large abundances, MAH theory predicts smaller ERs than those predicted by a power-law fit. When comparing with binned data, MAH theory is able to capture the non-linearity in mean ER across A bins in a log-log plot that a power-law fit cannot capture by definition (Figure 15).

Through abundance and ER data, we find that MAH could be true given that trends are consistent with an analytical relationship between the two variables under MAH. In the next chapter, we will further use predicted analytical relationships between protein properties to quantitatively determine whether a proxy for ΔG , called melting temperature, can take its place to assess MAH. Our analytical relationships enable us to test the control that if MAH is perfectly true, i.e. $r(A, \Delta G) \sim -1$ and $r(\Delta G, ER) \sim 1$ (Figure 1), what would be the corresponding correlations with melting temperature?

Chapter III.

Melting temperature as a proxy for folding free energy

Proteome-wide data has opened the door for testing complicated hypotheses that were once thought limited to *in silico* verification. This last decade witnessed a breakthrough in proteome-wide T_m measurements. Starting with studies laying the groundwork (Molina et al. 2013; Feng et al. 2014), large scale T_m measurements are now possible by coupling some form of proteolysis with mass spectrometry (Leuenberger et al. 2017; Becher et al. 2018; Mateus et al. 2018; Jarzab et al. 2020). This culminated most recently in proteome T_m measurements across 13 organisms (Jarzab et al. 2020)!

The advent of proteome-wide stability measurements provides an exciting opportunity to more conclusively test MAH. Leuenberger et al. (2017) demonstrated that MAH is validated when binning their proteome-wide data for *E.coli*. A more careful study treating each protein as a data point found a very weak correlation in support of the first-half of MAH: more abundant proteins are more stable for *E.coli* (Plata and Vitkup 2017). However, no T_m -ER correlation was exhibited. Furthermore, when they analyzed data from 3 other proteomes, none exhibited either of the two MAH-expected correlations (Plata and Vitkup 2017).

Before rushing to declaring MAH invalid with the data at hand, it is important to consider the difference between the two protein stability metrics. MAH assumes that ΔG underlies the A-ER correlation, not T_m . In general, just because x is correlated with y and y is correlated with z, does not mean that x is correlated with z. In other words, correlations are not transitive (Langford et al. 2001; Ellenberg 2014). Langford et al.

(2001) derived a relationship for the possible values of the Pearson correlations (r) between x and z (r_{xz}), given that r_{xy} and r_{yz} are known, to be as follows.

Equation 44

$$r_{xy}r_{yz} - \sqrt{(1 - r_{xy}^2)(1 - r_{yz}^2)} \leq r_{xz} \leq r_{xy}r_{yz} + \sqrt{(1 - r_{xy}^2)(1 - r_{yz}^2)}$$

Utilizing Equation 44 to determine ΔG correlations from corresponding T_m correlations, we find that ranges in possible ΔG correlations are very large (Figure 16). We conclude that we cannot make any statement on the accuracy of MAH just based on T_m correlations alone, since resulting ΔG correlations could just as well be positive or negative across the *E.coli*, *S.cerevisiae* and *H.sapiens* proteomes. However, Equation 44 does not account for any potential mathematical relationship between T_m and ΔG . For example, if T_m and ΔG were linearly related, then correlations involving T_m , have a direct one-to-one correspondence to correlations involving ΔG . We now commence to more carefully study the range in resulting ΔG correlations by considering the mathematical expression between ΔG and T_m according to statistical mechanics.

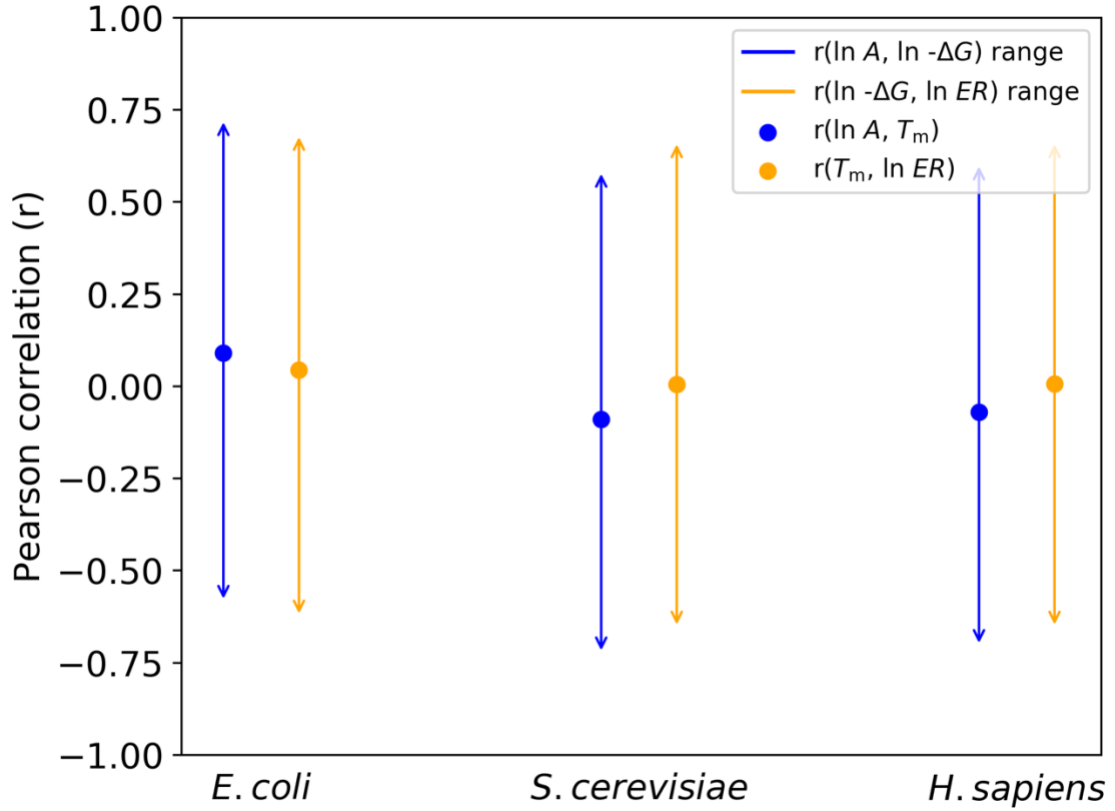


Figure 16: Wide ranges of Pearson correlation coefficients for $\ln -\Delta G$ with $\ln A$ and $\ln ER$ from corresponding T_m correlations with $\ln A$ and $\ln ER$ demonstrate that T_m measured by Leuenberger et al. (2017) cannot assess the protein misfolding avoidance hypothesis. Ranges are obtained by evaluating Equation 44 with $r(T_m, \ln -\Delta G) = 0.75$ from Supplementary Figure 1a, and $r(T_m, \ln A)$ and $r(T_m, \ln ER)$, which are represented as dots, taken from Figure 1 and 2, respectively, of Plata and Vitkup (2017). Although $r(T_m, \ln -\Delta G)$ is reported in Plata and Vitkup (2017), $r(T_m, -\Delta G) = 0.76$ is essentially identical to $r(T_m, \ln -\Delta G) = 0.75$. Therefore, ranges illustrated for $r(\ln -\Delta G, \ln A)$ and $r(\ln -\Delta G, \ln ER)$ are essentially identical to ranges for $r(\ln A, -\Delta G)$ and $r(-\Delta G, \ln ER)$ when $r(T_m, -\Delta G)$ is employed.

Mathematical relationship between stability metrics

Under the experimentally validated assumption that the change in heat capacity at constant pressure (ΔC_p) is independent of temperature during protein folding, Becktel and Schellman (1987) derived a relationship between T_m and ΔG in terms of ΔC_p and the change in enthalpy (ΔH_m) at T_m , using the Gibbs-Helmholtz equation.

Equation 45

$$\Delta G = -\Delta H_m \left(1 - \frac{T}{T_m}\right) + \Delta C_p \left(T_m - T + T \ln \frac{T}{T_m}\right)$$

As seen in Equation 45, the dependence between ΔG and T_m involves two other protein-specific variables – ΔH_m and ΔC_p . In general, no simple monotonic relationship exists between ΔG and T_m . We find a weakly significant ΔG - T_m correlation $r = -0.36$ (p-value = 0.02) when taking the T_m , ΔH_m and ΔC_p reported for 43 proteins in Rees and Robertson (2001) Table 1A and calculating ΔG according to Equation 45 (Figure 18A).

As an attempt to obtain a simpler ΔG - T_m relationship that is physically motivated, we approximate Equation 45 by writing $T_m = T + \delta_m$ and Taylor expanding to first order around $\delta_m/T = 0$. The approximation is not drastic because temperatures in Equation 45 are in Kelvin ($T \sim 300$ K) and proteins have a mean stability of 333 K (Robertson and Murphy 1997), making $\delta_m/T \ll 1$.

Equation 46

$$\begin{aligned}\Delta G &= -\Delta H_m \left(1 - \frac{T}{T + \delta_m}\right) + \Delta C_p \left(\delta_m + T \ln \frac{T}{T + \delta_m}\right) \\ &= -\Delta H_m \left(1 - \left[1 + \frac{\delta_m}{T}\right]^{-1}\right) + \Delta C_p \left(\delta_m - T \ln \left[1 + \frac{\delta_m}{T}\right]\right) \\ &\approx -\Delta H_m \left(1 - \left[1 - \frac{\delta_m}{T}\right]\right) + \Delta C_p \left(\delta_m - T \left[\frac{\delta_m}{T}\right]\right) \\ &= -\Delta H_m \frac{\delta_m}{T} = -\Delta H_m \frac{T_m - T}{T}\end{aligned}$$

Equation 46 demonstrates that ΔC_p can be neglected to first order in δ_m/T . When fitting ΔG to the right hand side of Equation 46 (which we define as x_1) with the Rees and Robertson (2001) Table 1A dataset, we find $r = 0.91$ ($2E-17$) with a best-fit line of $\Delta G = 0.43 * x_1 - 5.9$ kJ/mol (Figure 18B). If Equation 46 is perfectly valid, $\Delta G = 1 * x_1 + 0$. We reason that the imperfection is a result of the $\delta_m/T \ll 1$ approximation. When only considering half the dataset with lower T_m , Equation 46 holds stronger with $r(x_1, \Delta G) =$

0.95 (5E-11) and a best-fit line of $\Delta G = 0.61 \cdot x_1 + 0.08$ kJ/mol. The slope in this case is closer to one and the intercept is nearly zero because the $\delta_m/T \ll 1$ approximation is less severe for proteins with lower T_m .

Although Equation 46 is simpler than Equation 45, Equation 46 still does not guarantee a monotonic relationship between T_m and ΔG because in the equation, ΔG still depends on one other protein property besides $T_m - \Delta H_m$. Moreover, ΔG cannot be evaluated explicitly because proteome-wide ΔH_m values are currently unknown.⁴ A previous study found ΔH and ΔC_p to scale with the number of residues (N) in a protein (Robertson and Murphy 1997) at 333 K, where $\Delta H(T=333K) = 2.92 \cdot N$ kJ/mol and $\Delta C_p = 0.058 \cdot N$ kJ/mol. Thus, $\Delta H_m = \Delta H(T=333K) + \Delta C_p (T_m - 333K) = 2.92 \cdot N + 0.058 \cdot N (T_m - 333K)$. For conciseness, we will use the symbols: $h = \Delta H(T=333K)/N$, $c_p = \Delta C_p/N$ and $T' = 333$ K.

Compact and cubic lattice proteins, the simplest model that captures contacts present in globular proteins that are important for assuming native conformations (Shakhnovich 1997), can motivate the N dependence. The energy of folded lattice proteins scales with the number of contacts and the number of contacts can be shown to scale as $2N$ as N approaches infinity for compact and cubic lattice proteins. Take for example Figure 17, a specific conformation of $3 \cdot 3 \cdot 3$ lattice protein.

⁴ ΔH_m are technically available by Leuenberger et al. (2017) for peptides, however, they are not consistent with Equation 45. In fitting experimental denaturant curves, Leuenberger et al. (2017) implicitly assumes that ΔH is temperature independent.

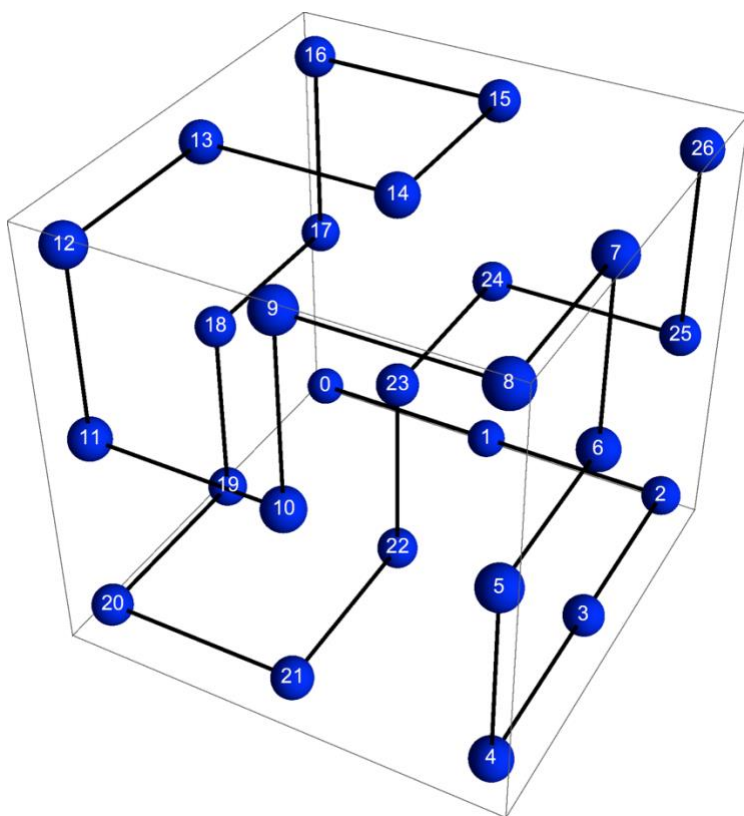


Figure 17: A compact 3*3*3 lattice protein.

A contact in a lattice protein is defined as adjacent residues not connected by the backbone chain. The number of contacts (n) for a cubic and compact lattice protein of any size can be determined by noting that residues in the lattice protein can be broken up into 4 types that each have a unique n (Table 2). For simplicity, we ignore the first and last residue of the chain in tabulating Table 2, which would have one more added contact.

	vertex	vertex-center	center-core	core
Corresponding residues in Figure 17	0, 2, 4, 8, 12, 16, 20, 26	1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 25	6, 10, 14, 18, 22, 24	23
n	1	2	3	4

Table 2: Summary of the four residue types and their number of contacts.

By deducing how the number of each residue type grows with the number of residues on the side (L) of a lattice protein, i.e. L=3 for Figure 17, we can write the total n for an arbitrary compact and cubic lattice protein as

Equation 47

$$\begin{aligned} n &= \frac{1}{2} [N_{vertex}n_{vertex} + N_{v-c}n_{v-c} + N_{c-c}n_{c-c} + N_{core}n_{core}] \\ &= \frac{1}{2} [8 * 1 + 12(L - 2) * 2 + 6(L - 2)^2 * 3 + (L - 2)^3 * 4] \\ &= 4 + 12(L - 2) + 9(L - 2)^2 + 2(L - 2)^3 \end{aligned}$$

N is the total number of residues in the protein and is equal to L^3 because every position in the lattice must be occupied by a residue, otherwise known as a Hamiltonian walk.

Thus, $n/N = 2$ as N approaches infinity.

We assume that $\Delta H \sim$ energy of the folded lattice protein (Gin et al. 2009; Best et al. 2013). The N scaling for ΔC_p immediately follows because ΔC_p is the temperature derivative of ΔH .

Inserting the $\Delta H_m(T_m, N)$ relationship into Equation 46, yields

Equation 48

$$\Delta G \approx -N \left(h + c_p (T_m - T') \right) \frac{T_m - T}{T}$$

When fitting ΔG to the right hand side of Equation 48 (which we define as x_2) with Rees and Robertson (2001) Table 1A, we find $r = 0.80$ (9E-11) with a best-fit line of $\Delta G = 0.38 * x_2 - 9.67$ kJ/mol (Figure 18C). $r(x_2, \Delta G)$ is not drastically smaller than the previous r corresponding to Equation 46 ($r = 0.91$). The slight loss in accuracy is acceptable given that ΔG no longer depends on ΔH_m and depends on N, a protein property that is known proteome-wide.

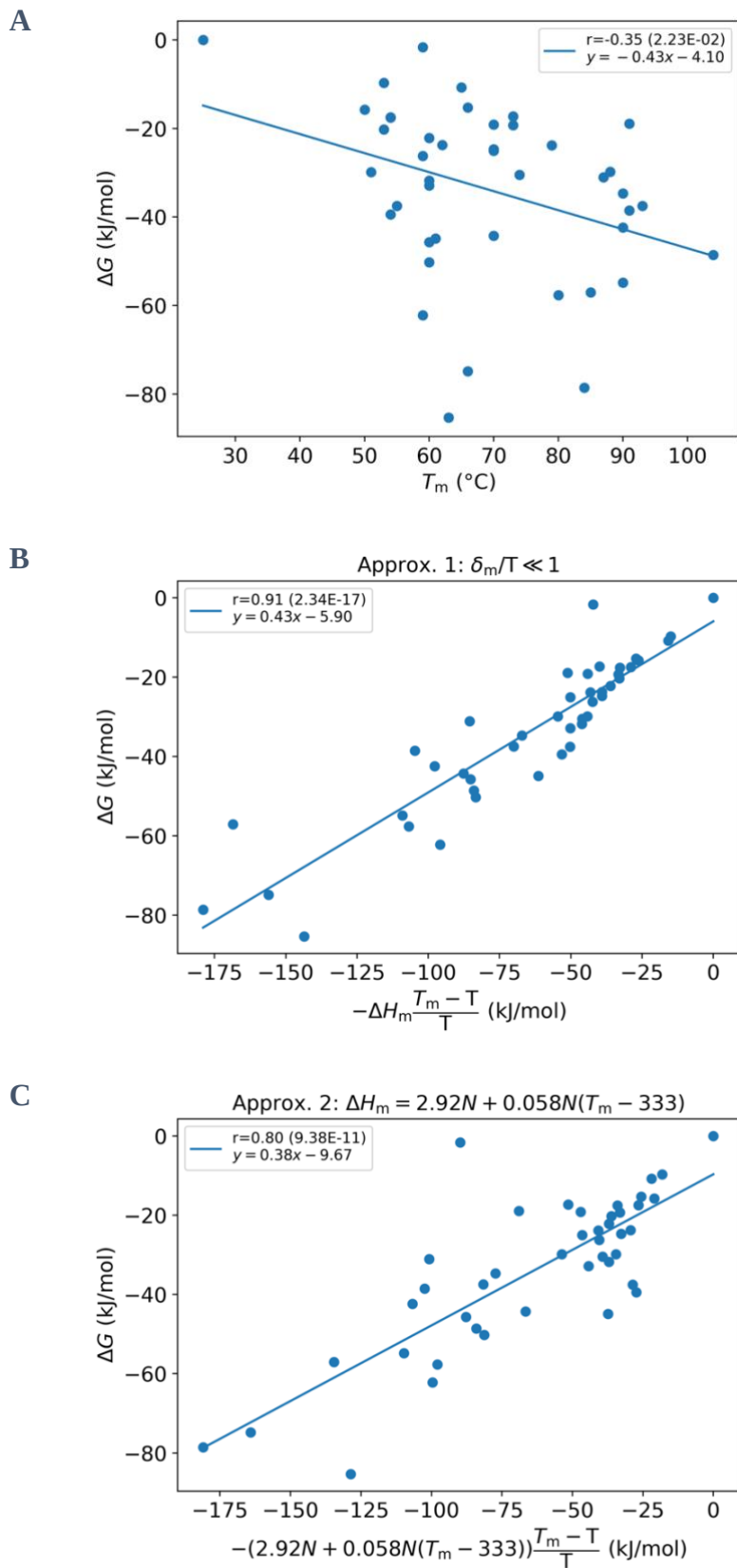


Figure 18: Workflow to find a simpler equation relating ΔG and T_m starting from Equation 45. After two approximations (B-C), an equation currently evaluable proteome-wide while still retaining accuracy, is obtained.

In principle, we could omit the first approximation altogether, and plug in N-dependent values directly into Equation 45. When doing so, the corresponding correlation with ΔG is similar to that seen for $r(x_2, \Delta G)$, $r = 0.82$ (3E-11). This makes sense because Equation 46 only has one protein-specific value that is approximated by N, while Equation 45 has two. The inaccuracy from employing N as a proxy for both ΔH_m and ΔC_p terms in Equation 45 is similar to employing N as a proxy for just ΔH_m in Equation 46, which approximates Equation 45. Besides depending less on the assumption that thermodynamic properties making up ΔG scale with N, we keep the first approximation because the resulting equation relating ΔG and T_m is simpler and has the added benefit of being invertible, i.e. we can write T_m as a function of ΔG (Equation 49).

Equation 49

$$T_m = \frac{1}{2c_p} \left(-h + c_p(T' + T) + \sqrt{c_p T^2 \left(c_p - \frac{4\Delta G}{N} \right) + 2c_p(h - c_p T')T + (h - c_p T')^2} \right)$$

Limits of using melting temperature to capture folding free energy

The scarcity of proteome-wide data on ΔG limits the ability to test key predictions stemming from MAH. This situation seems to have changed with the recent advance of proteome-wide measurements of melting temperature across four organisms (Leuenberger et al. 2017). Here, we assess the accuracy of T_m as a proxy for ΔG with respect to testing MAH. We first focus on the limit of T_m measurements perfectly consistent with MAH by employing our previous analytical expressions capturing the analytical relationship between abundance-evolutionary rate under MAH (Equation 41). We can write the ER expression to be a function of ΔG^{opt} , which we henceforth refer to ΔG^{MAH} , by taking advantage of the $g'(\Delta \Delta G) = 0$ identity (Equation 38).

It is possible to invert Equation 40 such that ΔG^{MAH} is a function of ER. Then we insert the ΔG^{MAH} expression into our T_m - ΔG expression, to obtain the corresponding T_m^{MAH} . In theory, there exists no T_m^{MAH} values that are in better agreement with MAH because the ΔG^{MAH} that T_m^{MAH} approximates is constructed to yield $r(\ln A, -\Delta G^{\text{MAH}}) \sim 1$. We analyze the $\ln A$ - T_m^{MAH} correlation and find a strong MAH consistent correlation across all three organisms (Figure 19). Even for *S.cerevisiae* and *H.sapiens* which exhibit negative $\ln A$ - T_m correlations with Leuenberger et al. (2017) data (dashed lines in bar plots Figure 19), MAH consistent positive $\ln A$ - T_m^{MAH} correlations are seen. Additionally, we find that all three organisms demonstrate significantly negative T_m^{MAH} - $\ln ER$, despite *S.cerevisiae* and *H.sapiens* insignificant correlations with Leuenberger et al. (2017) data. We conclude that perfectly accurate T_m could in principle yield MAH consistent correlations.

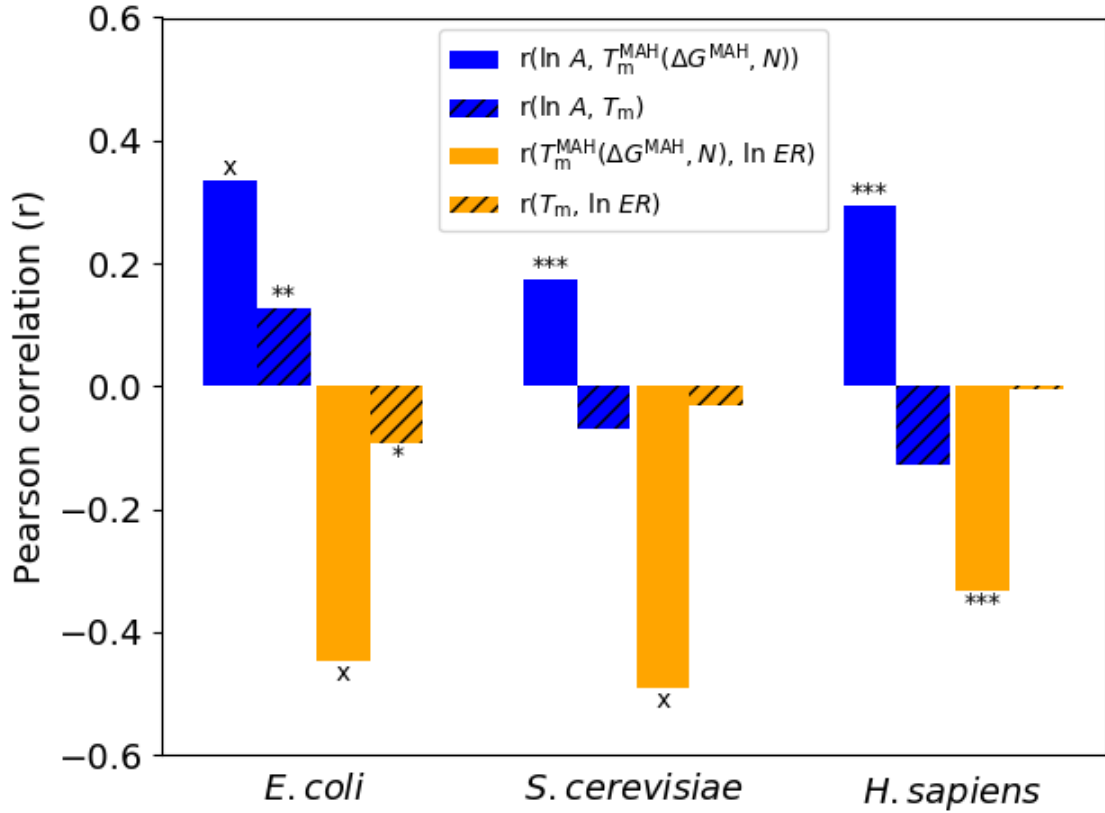


Figure 19: Utilizing our equation for ΔG^{MAH} , our derived T_m^{MAH} recovers MAH expected correlations. Corresponding T_m correlations have dashed lines and are derived from our curated datasets, with T_m from Leuenberger et al. (2017). X marks and asterisks above or below bar plots denote p-value ranges: $xx < 10^{-50}$, $x < 10^{-10}$, $*** < 0.001$, $** < 0.01$, $* < 0.05$.

In light of finding T_m^{MAH} correlations consistent with MAH, does the lack of MAH consistent correlations in the Leuenberger et al. (2017) T_m dataset indicate that MAH is wrong? Leuenberger et al. (2017) reported that T_m from their study had a Pearson correlation coefficient of 0.36 (p-value = $3\text{E-}4$) with experimental T_m measurements listed in ProTherm for *E.coli* proteins. T_m in ProTherm can be assumed to more accurately characterize a protein's T_m in isolation because differential scanning calorimetry, the standard protocol for T_m measurement (Robertson and Murphy 1997), is used. T_m in Leuenberger et al. (2017) employs a novel technique to measure T_m proteome-wide by employing limited proteolysis coupled with mass spectroscopy in cell extracts. The relatively low correlation between Leuenberger et al. (2017) T_m values and

ProTherm T_m values signifies that Equation 49 is not valid as written for Leuenberger et al. (2017) T_m because of measurement uncertainties. To make our previous analysis comparable to the Leuenberger et al. (2017) T_m dataset, we explicitly account for the inaccuracy in Leuenberger et al. (2017) T_m measurements by defining,

Equation 50

$$T_m^\alpha = T_m^{\text{MAH}} + \alpha \mathcal{N}(0,1)$$

T_m^α represents Leuenberger et al. (2017) T_m , T_m^{MAH} refers to ideally accurate T_m consistent with MAH as described previously, and $\alpha \mathcal{N}(0,1)$ captures the inaccuracy between the two variables. $\mathcal{N}(0,1)$ is the Normal distribution with a mean of 0 and variance of 1. The noise strength α modulates the variance. The added $\alpha \mathcal{N}(0,1)$ term is appropriate because the histogram of residuals from linearly fitting Leuenberger et al. (2017) T_m to ProTherm T_m resembles a Gaussian centered at 0 (Leuenberger et al. 2017), a general property of accurate linear relationships (Anscombe 1973).

We can analytically solve for α in terms of $r(T_m^{\text{MAH}}, T_m^\alpha)$ starting from the definition of the Pearson correlation coefficient and employing properties of the covariance (Cov) (Rice 2007).

Equation 51

$$\begin{aligned} r(T_m^{\text{MAH}}, T_m^\alpha) &= \frac{\text{Cov}(T_m^{\text{MAH}}, T_m^\alpha)}{\sqrt{\text{Var } T_m^{\text{MAH}}} \sqrt{\text{Var } T_m^\alpha}} = \frac{\text{Cov}(T_m^{\text{MAH}}, T_m^{\text{MAH}} + \alpha \mathcal{N}(0,1))}{\sqrt{\text{Var } T_m^{\text{MAH}}} \sqrt{\text{Var } [T_m^{\text{MAH}} + \alpha \mathcal{N}(0,1)]}} \\ &= \frac{\text{Cov}(T_m^{\text{MAH}}, T_m^{\text{MAH}})}{\sqrt{\text{Var } T_m^{\text{MAH}}} \sqrt{\text{Var } T_m^{\text{MAH}} + \alpha^2}} = \frac{1}{\sqrt{1 + \frac{\alpha^2}{\text{Var } T_m^{\text{MAH}}}}} \\ \rightarrow \alpha &= \sqrt{\text{Var } T_m^{\text{MAH}} \left(\frac{1}{r(T_m^{\text{MAH}}, T_m^\alpha)^2} - 1 \right)} \end{aligned}$$

We write down $r(\ln A, T_m^\alpha)$ in terms of $r(\ln A, T_m^{\text{MAH}})$ and insert our analytical solution for α that we solved for in Equation 51.

Equation 52

$$\begin{aligned}
r(\ln A, T_m^\alpha) &= \frac{\text{Cov}(\ln A, T_m^\alpha)}{\sqrt{\text{Var} \ln A} \sqrt{\text{Var} T_m^\alpha}} = \frac{\text{Cov}(\ln A, T_m^{\text{MAH}} + \alpha \mathcal{N}(0,1))}{\sqrt{\text{Var} \ln A} \sqrt{\text{Var} [T_m^{\text{MAH}} + \alpha \mathcal{N}(0,1)]}} \\
&= \frac{\text{Cov}(\ln A, T_m^{\text{MAH}})}{\sqrt{\text{Var} \ln A} \sqrt{\text{Var} T_m^{\text{MAH}} + \alpha^2}} = \frac{r(\ln A, T_m^{\text{MAH}})}{\sqrt{1 + \frac{\alpha^2}{\text{Var} T_m^{\text{MAH}}}}} \\
&= \frac{r(\ln A, T_m^{\text{MAH}})}{\sqrt{1 + \frac{\text{Var} T_m^{\text{MAH}} \left(\frac{1}{r(T_m^{\text{MAH}}, T_m^\alpha)^2} - 1 \right)}{\text{Var} T_m^{\text{MAH}}}}} \\
&= r(\ln A, T_m^{\text{MAH}}) r(T_m^{\text{MAH}}, T_m^\alpha)
\end{aligned}$$

Inserting $r(\ln A, T_m^{\text{MAH}}) = 0.33$ (Figure 19) and $r(T_m^{\text{MAH}}, T_m^\alpha) = 0.36$ (Leuenberger et al. 2017) for *E.coli* into Equation 52, yields $r(\ln A, T_m^\alpha) = 0.12$. Our $r(\ln A, T_m^\alpha)$ for *E.coli* is very similar to what we find for $r(\ln A, T_m) = 0.13$ using the Leuenberger et al. (2017) dataset (Figure 19). The same procedure can be applied to the other half of MAH to find that *E.coli* $r(T_m^\alpha, \ln ER) = -0.16$, compared to $r(\ln A, T_m) = -0.09$ for the Leuenberger et al. (2017) dataset. Our analysis demonstrates the possibility that MAH may be completely true, however the imperfect T_m – ΔG relationship and inaccurate Leuenberger et al. (2017) T_m values reduce the magnitude of MAH expected correlations.

It would be useful to expand results for the *E.coli* proteome to those of *S.cerevisiae* and *H.sapiens*, however it is difficult to repeat this analysis for these species because there are few reported *S.cerevisiae* and *H.sapiens* T_m measurements in

ProTherm. Regardless, our approach of adding noise to T_m^{MAH} (Equation 52) would not be able to recapitulate the opposite than MAH expected in A- T_m correlations observed for *S.cerevisiae* and *H.sapiens* using Leuenberger et al. (2017) T_m (Figure 19). The addition of noise to a perfect MAH correlate cannot explain this result because added noise only erases any correlation present; it cannot recreate a significant correlation in the opposite direction. An important assumption is that Leuenberger et al. (2017) T_m must actually relate to ΔG according to Equation 45. The negative ln A- T_m correlation for *S.cerevisiae* and *H.sapiens* could be the result of some variable confounding the T_m measurement, rather than disproving MAH. The only way to definitively determine if that is true is to have enough ProTherm T_m for the respective organism to assess the accuracy of Leuenberger et al. (2017)'s T_m .

We briefly investigate whether controlling for the number of protein-protein interaction (PPIs) partners could significantly change correlations between protein properties. Another study used a similar proteome-wide method as Leuenberger et al. (2017) to identify protein-protein interaction partners (Tan et al. 2018), indicating that T_m measurements could be influenced by PPIs. We find that the number of stable protein-protein interactions (PPIs) are right-shifted for *S.cerevisiae* and *H.sapiens* compared to that of *E.coli* (median PPI for *E.coli* = 8, *S.cerevisiae* = 16, *H.sapiens* = 25), consistent with previous reports (Reid et al. 2010; Schad et al. 2011). *E.coli* T_m from Leuenberger et al. (2017) would more likely correspond to T_m in ProTherm because cellular PPI effects on stability are less prevalent than in *S.cerevisiae* and *H.sapiens*. In Figure 20, we reanalyze the Leuenberger et al. (2017) T_m dataset by only including proteins with PPIs less than or equal to the median PPI found for the proteome of the respective organism,

hoping to recover the weakly consistent MAH correlations seen in *E.coli*, in *S.cerevisiae* and *H.sapiens* as well. However, MAH consistent correlations are not recovered for *S.cerevisiae* and *H.sapiens*. For *E.coli*, the MAH-consistent $\ln A$ – T_m correlation is retained, however the MAH-consistent T_m – $\ln ER$ is lost. This seems to be the result of *E.coli* proteins with more PPIs evolving slower (Razban et al. 2018).

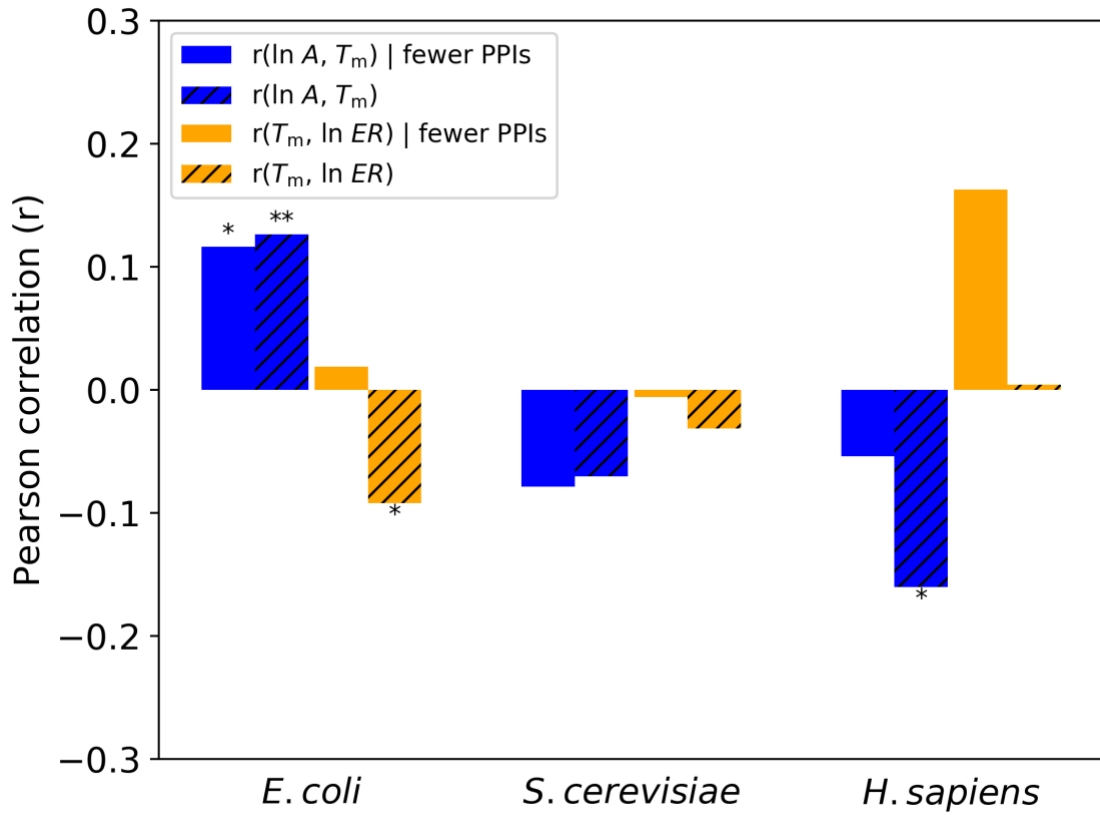


Figure 20: MAH-consistent correlations are not recovered for *S.cerevisiae* and *H.sapiens* when only considering proteins with PPIs less than or equal to the median value of the given organism. Asterisks above or below bar plots denote p-value ranges: *** <0.001, ** <0.01, * <0.05.

Overall, our results indicate the care required to interpretate correlations involving proxies of variables, despite T_m and ΔG both being metrics of protein stability. A common misconception is that Pearson correlation coefficients are transitive: if X and Y positively correlate, and Y and Z positively correlate, then X and Z must also positively correlate (Castro Sotos et al. 2009). If we want to do better than the general equation

providing bounds for a variable's correlations based on those of its proxies (Equation 44), we can consider their mathematical relationship. In Equation 48, we demonstrate that despite several reasonable approximations, we cannot derive an accurate relationship that has ΔG and T_m linearly related. Instead, there is one other additional protein variable confounding the relationship between the two stability metrics.⁵ Inserting the values into the corresponding equations and accounting for current inaccuracies in measuring T_m proteome-wide (Equation 52), we conclude that proteome-wide T_m measurements are not accurate enough to assess MAH.⁶

⁵ This additional variable becomes approximately constant when comparing stabilities only among protein orthologs because protein structures of orthologs are highly similar. Thus, we are justified in linearly transforming T_m into ΔG for folA orthologs in Chapter 1.

⁶ This chapter is a slightly modified version of Razban (2019) published in *Molecular Biology and Evolution* and has been reproduced here with the permission of the copyright holder.

Chapter IV.

Single mutants have Gaussian distributed folding free energies

When calculating all possible single mutant stability effects ($\Delta\Delta G$ ⁷) using FoldX (Guerois et al. 2002) – a fast and relatively accurate computational software that requires a Protein Data Bank (PDB) structure (Berman et al. 2000) – Tokuriki et al. (2007) made a curious observation. Across the 16 natural proteins studied, they found each $\Delta\Delta G$ distribution to take a universal bi-Gaussian form. They also pooled available experimental data and verified that this result was not an artifact of computational software and seemed to reflect reality (Tokuriki et al. 2007). A further study extended this initial observation across 5 organisms' proteomes, finding that the same universal bi-Gaussian form to hold generally for mesophilic proteomes (Faure and Koonin 2015).

In light of the Introduction, the universality of the $\Delta\Delta G$ distribution is critical to the makeup of the protein misfolding avoidance hypothesis and understanding protein evolution across the proteome. As summarized in Figure 1 and mathematically modelled in Equation 23, $\Delta\Delta G$ s sampled throughout evolution are assumed to follow a Gaussian form with a mean dependent on its current protein folding free energy. Computational results (Tokuriki et al. 2007; Faure and Koonin 2015) and limited experimental data (Tokuriki et al. 2007; Zeldovich et al. 2007) seem to support a Gaussian form for the $\Delta\Delta G$ distribution. However, a mechanistic understanding for the universality is lacking

⁷ In this thesis chapter, stating ' $\Delta\Delta G$ ' always refers more broadly to all possible single mutant $\Delta\Delta G$ s

In this chapter, we apply first principles statistical mechanics to understand how a universal Gaussian form can arise to capture $\Delta\Delta G$ distributions despite the unique sequence and structure of individual proteins. We derive that per residue position $\Delta\Delta G$ distributions are accurately approximated by Gaussians with means and variances shaped by their local (in space) sequence and structure. Our derivation elucidates that the large number of amino acid types is responsible for the Gaussian being a good approximation, unlike previous arguments in the field which relied on the large number of pairwise contact energies making up protein folding free energy (Shakhnovich and Gutin 1990; Tiana et al. 2001). Indeed, we show that our derivation applies almost equally as well to those residues with fewer contacts for previously studied computational data on 16 proteins (Tokuriki et al. 2007) and for newly released experimental data on a protein G domain (Nisthal et al. 2019).

Statistical mechanics of stability effects

We solve for the distribution of stability effects for a given residue position by first applying a continuous spin approximation to obtain the partition function of stability effects (England and Shakhnovich 2003) and then employing an inverse Laplace transform over inverse temperature to proceed from the partition function to the density of states (Horwitz 1983).

Protein folding free energy expression

The free energy of the protein (ΔG) is approximated as the sum over individual contact energies (Gō 1983). It assumes that two-body interactions in the native state make dominant contributions to protein stability.

Equation 53

$$\begin{aligned}\Delta G &\approx \frac{1}{2} \sum_{i,j}^N C_{ij} V_{s(i)s(j)} \\ &= \frac{1}{2} \boldsymbol{\sigma}^T \mathbf{U} \boldsymbol{\sigma}\end{aligned}$$

N is the number of amino acids in the protein. $\mathbf{C}$ is a $N \times N$ contact matrix whose elements can only take values of 1 (two residues spatially close in proximity, or in contact) or 0 (not in contact). $\mathbf{V}$ is a $M \times M$ interaction matrix, where M is the number of amino acid types ($M = 20$ standard amino acids), and an element of $\mathbf{V}$ is the pairwise free energy of two specific amino acids when they are in contact. The symbol s refers to the wildtype sequence; $s(i)$ refers to the amino acid of position (i).

Integrals will be simpler to perform with matrices, thus the rewrite of the summation by the equivalent matrix multiplication in the second line of Equation 53. $\boldsymbol{\sigma}$ is a NM dimensional vector, where M is the number of amino acid types. Each N block of $\boldsymbol{\sigma}$ has M elements, in which one element is equal to one while the other $M - 1$ elements are equal to zero. The index of the non-zero element sets the amino acid type. $\mathbf{U}$ is the Kronecker product (Laub 2005) between $\mathbf{C}$ and $\mathbf{V}$ ($\mathbf{C} \otimes \mathbf{V}$), making $\mathbf{U}$ a $NM \times NM$ dimensional matrix.

In principle, $\mathbf{V}$ can capture residue-residue interactions, as well as energetic and entropic solvent effects (Miyazawa and Jernigan 1996). Equation 53 is still an approximation because terms such as the unfolded state and conformational entropy are not taken into account. We point out at the end of this section that the final form of the distribution is robust to the exact definition of ΔG as long interactions are one- or two-body in $\boldsymbol{\sigma}$.

Since we are interested in single mutant effects at a position (i), we want $\Delta\Delta G_{(i)} = \Delta G_{\text{mut},(i)} - \Delta G_{\text{wt}}$. We split up σ and $\mathbf{U}$ into the following blocks,

Equation 54

$$\sigma = \begin{bmatrix} \sigma_{(a)} \\ \sigma_{(i)} \\ \sigma_{(z)} \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} \mathbf{U}_{(a,a)} & \mathbf{U}_{(a,i)} & \mathbf{U}_{(a,z)} \\ \mathbf{U}_{(i,a)} & \mathbf{0}_{(i,i)} & \mathbf{U}_{(i,z)} \\ \mathbf{U}_{(z,a)} & \mathbf{U}_{(z,i)} & \mathbf{U}_{(z,z)} \end{bmatrix}$$

$\sigma_{(a)}$ corresponds to all σ s numbered before position (i); $\sigma_{(z)}$, after position (i). $\mathbf{U}_{(i,i)} = \mathbf{0}_{(i,i)}$ is a $M * M$ dimensional zero matrix because the contact matrix has zeros along the diagonal. In other words, residues cannot be in contact with themselves. In addition, $\mathbf{U}$ is symmetric because $\mathbf{C}$ and $\mathbf{V}$ are symmetric by definition.

We commence to write the free energy (Equation 53) in terms of the defined matrix blocks,

Equation 55

$$\Delta G = \frac{1}{2} \left(\sigma_{(a)}^T \mathbf{U}_{(a,a)} \sigma_{(a)} + 2\sigma_{(i)}^T \mathbf{U}_{(i,a)} \sigma_{(a)} + 2\sigma_{(z)}^T \mathbf{U}_{(z,a)} \sigma_{(a)} + 2\sigma_{(i)}^T \mathbf{U}_{(i,z)} \sigma_{(z)} + \sigma_{(z)}^T \mathbf{U}_{(z,z)} \sigma_{(z)} \right)$$

Using Equation 55, $\Delta\Delta G_{(i)}$ can be compactly expressed as,

Equation 56

$$\begin{aligned} \Delta\Delta G_{(i)} &= \sigma_{\text{mut},(i)}^T \mathbf{U}_{(i,a)} \sigma_{(a)} + \sigma_{\text{mut},(i)}^T \mathbf{U}_{(i,z)} \sigma_{(z)} - \Delta G_{\text{wt},(i)} \\ &= \dot{\sigma}_{\text{mut},(i)}^T \dot{\mathbf{U}}_{(i,a)} \sigma_{(a)} + \dot{\sigma}_{\text{mut},(i)}^T \dot{\mathbf{U}}_{(i,z)} \sigma_{(z)} - \Delta G_{\text{wt},(i)} \end{aligned}$$

Terms not involving (i) cancel because they are the same across mutant and wildtype. We also reduce $\sigma_{(i)}$ to only include amino acids that are non-wildtype in the second line of

Equation 56; $\dot{\sigma}_{(i)}$ refers to a $M-1$ dimensional vector that does not include the wildtype amino acid. The same notation applies for $\dot{\mathbf{U}}$.

Partition function

We first transform the summation into an integral over $\dot{\sigma}_{\text{mut},(i)}$ space, picking out terms that correspond to amino acid types with Dirac delta functions.

Equation 57

$$\begin{aligned} Z_{(i)}(\beta) &= \sum_m^{M-1} e^{-\beta \Delta \Delta G_{(i),m}} \\ &= \int_{-\infty}^{\infty} d^{M-1} \dot{\sigma}_{\text{mut},(i)} \left\{ \sum_m^{M-1} \delta(\dot{\sigma}_{\text{mut},(i),m} - 1) \prod_{n \neq m}^{M-2} \delta(\dot{\sigma}_{\text{mut},(i),n}) \right\} e^{-\beta \Delta \Delta G_{(i)}} \end{aligned}$$

The integral in the second line exactly corresponds to the discrete sum. That integral, however, is not analytically solvable, so we approximate each delta function with a corresponding Gaussian distribution with variance v (Piela 2013). We call this the continuous spin approximation.

Equation 58

$$\begin{aligned} Z_{(i)}(\beta) &\approx \int_{-\infty}^{\infty} d^{M-1} \dot{\sigma}_{\text{mut},(i)} \left\{ \sum_m^{M-1} \exp \left(-\frac{1}{2v} [\dot{\sigma}_{\text{mut},(i),m} \right. \right. \\ &\quad \left. \left. - 1]^2 \right) \prod_{n \neq m}^{M-2} \exp \left(-\frac{1}{2v} \dot{\sigma}_{\text{mut},(i),n}^2 \right) \right\} e^{-\beta \Delta \Delta G_{(i)}} \\ &= \int_{-\infty}^{\infty} d^{M-1} \dot{\sigma}_{\text{mut},(i)} \left\{ \exp \left[-\frac{\dot{\sigma}_{\text{mut},(i)}^T \dot{\sigma}_{\text{mut},(i)}}{2v} \right] \exp \left[-\frac{1}{2v} \sum_m^{M-1} \exp \left(\frac{\dot{\sigma}_{\text{mut},(i),m}}{v} \right) \right] \right\} e^{-\beta \Delta \Delta G_{(i)}} \end{aligned}$$

If the variance approaches zero, then approximating the delta functions as Gaussians is accurate (Piela 2013). To test whether this approximation holds here, we write out the variance when referring to a single amino acid,

Equation 59

$$\begin{aligned}
 v &= \langle \dot{\sigma}_{\text{mut},(i)}^2 \rangle - \langle \dot{\sigma}_{\text{mut},(i)} \rangle^2 \\
 &= \frac{1}{M-1} \sum_m^{M-1} \dot{\sigma}_{\text{mut},(i),m}^2 - \left(\frac{1}{M-1} \sum_m^{M-1} \dot{\sigma}_{\text{mut},(i),m} \right)^2 \\
 &= \frac{1}{M-1} - \frac{1}{(M-1)^2}
 \end{aligned}$$

For the standard set of amino acids, $M = 20$ is large enough such that the variance approaches zero. Thus, applying the continuous spin approximation is validated.

To numerically substantiate the claim that qualitative accuracy is not lost by applying the continuous spin approximation, we calculate $Z_{(i)}$ exactly (Equation 57) and under the continuous spin approximation (Equation 58) for a designed 64mer lattice protein (Venev and Zeldovich 2015). We find that the exact and approximate $Z_{(i)}$ are significantly correlated across its 64 residues (Figure 21).

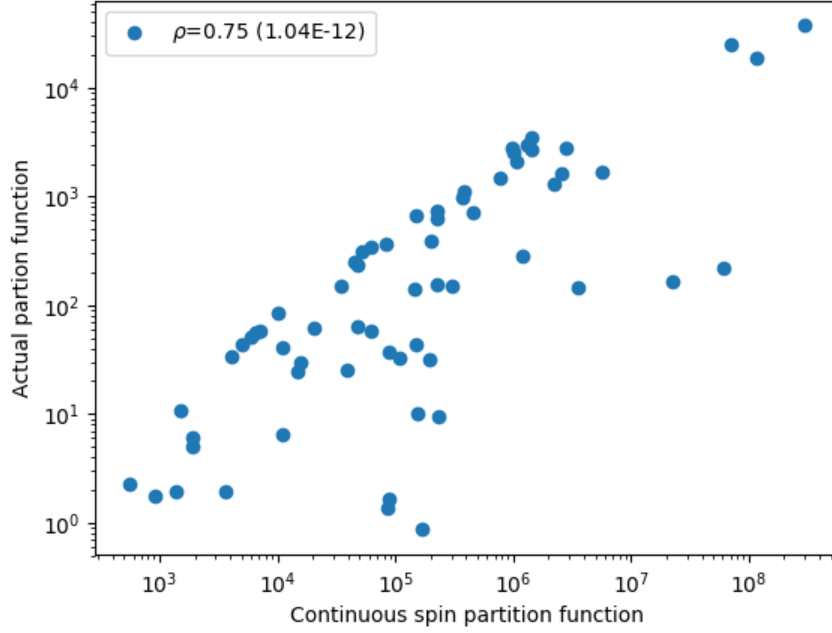


Figure 21: The continuous spin approximated and actual per residue position partition function has a significant Spearman correlation coefficient (ρ) with a p-value of $1E-12$, for a 64mer lattice protein under the Miyazawa-Jernigan potential (Miyazawa and Jernigan 1996). We sample each summation term in Equation 58 100 times to estimate the continuous spin partition function. The wildtype sequence was chosen by first evolving a random sequence for a given lattice protein structure under the Miyazawa-Jernigan potential to have a folding probability > 0.99 .

By approximating each delta function as a Gaussian, we end up with a global constraint on $\dot{\sigma}_{mut,(i)}^T \dot{\sigma}_{mut,(i)}$, or the length of $\dot{\sigma}_{mut,(i)}$. Still, having the term involving the summation of exponentials makes evaluating the integral over the vector $\dot{\sigma}_{mut,(i)}$ tricky, so we employ Jensen's inequality (Chandler 1987).

Equation 60

$$\frac{1}{M-1} \sum_m^{M-1} \exp\left(\frac{\dot{\sigma}_{mut,(i),m}}{v}\right) \geq \exp\left[\frac{\dot{\sigma}_{mut,(i)}^T \hat{\mathbf{1}}}{v(M-1)}\right]$$

$\hat{\mathbf{1}}$ is a $M-1$ dimensional ones vector. Inserting Equation 60 into Equation 58, we finally obtain a solvable integral.

Equation 61

$$\begin{aligned}
Z_{(i)}(\beta) &\geq \int_{-\infty}^{\infty} d^{M-1} \dot{\sigma}_{\text{mut},(i)} \left\{ \exp \left[-\frac{\dot{\sigma}_{\text{mut},(i)}^T \dot{\sigma}_{\text{mut},(i)}}{2v} \right] \exp \left[-\frac{1}{2v} \right] (M \right. \\
&\quad \left. - 1) \exp \left[\frac{\dot{\sigma}_{\text{mut},(i)}^T \mathbf{1}}{v(M-1)} \right] \right\} e^{-\beta \Delta \Delta G_{(i)}} \\
&= (M-1) \int_{-\infty}^{\infty} d^{M-1} \dot{\sigma}_{\text{mut},(i)} \exp \left[-\frac{1}{2v} \left(\dot{\sigma}_{\text{mut},(i)} - \frac{\mathbf{1}}{M-1} \right)^2 \right] e^{-\beta \Delta \Delta G_{(i)}}
\end{aligned}$$

We begin solving the integral by completing the square of the exponent and pulling out constants in front.

Equation 62

$$\begin{aligned}
Z_{(i)}(\beta) &= (M-1) e^{\beta \Delta G_{\text{wt},(i)}} \exp \left[-\frac{\beta}{M-1} \mathbf{1} \{ \dot{\mathbf{U}}_{(i,a)} \boldsymbol{\sigma}_{(a)} + \dot{\mathbf{U}}_{(i,z)} \boldsymbol{\sigma}_{(z)} \} \right. \\
&\quad \left. + \frac{1}{2v} (\beta v \{ \dot{\mathbf{U}}_{(i,a)} \boldsymbol{\sigma}_{(a)} + \dot{\mathbf{U}}_{(i,z)} \boldsymbol{\sigma}_{(z)} \})^2 \right] \int_{-\infty}^{\infty} d^{M-1} \dot{\sigma}_{\text{mut},(i)} \exp \left[-\frac{1}{2v} \left(\dot{\sigma}_{\text{mut},(i)} \right. \right. \\
&\quad \left. \left. - \frac{\mathbf{1}}{M-1} + \beta v \{ \dot{\mathbf{U}}_{(i,a)} \boldsymbol{\sigma}_{(a)} + \dot{\mathbf{U}}_{(i,z)} \boldsymbol{\sigma}_{(z)} \} \right)^2 \right]
\end{aligned}$$

The multidimensional integral represents $M-1$ identical Gaussian integrals and can be easily solved to be independent of β .

Before moving on and obtaining the density of states, let's do two things. We define a reduced partition function $z(\beta) = Z(\beta)/Z(0)$, since constants will eventually cancel out when calculating the density of states and ensuring that it is normalized. Let's also revert matrices back to the more intuitive summation notation.

Equation 63

$$z_{(i)}(\beta) = \frac{Z_{(i)}(\beta)}{Z_{(i)}(0)}$$

$$\begin{aligned}
&= \exp \left[\beta \Delta G_{\text{wt},(i)} - \frac{\beta}{M-1} \mathbf{1} \{ \dot{\mathbf{U}}_{(i,a)} \boldsymbol{\sigma}_{(a)} + \dot{\mathbf{U}}_{(i,z)} \boldsymbol{\sigma}_{(z)} \} \right. \\
&\quad \left. + \frac{\beta^2 v}{2} (\dot{\mathbf{U}}_{(i,a)} \boldsymbol{\sigma}_{(a)} + \dot{\mathbf{U}}_{(i,z)} \boldsymbol{\sigma}_{(z)})^2 \right] \\
&= \exp \left[\beta \left(-\frac{1}{M-1} \sum_j^N \sum_{m \neq s(i)}^M C_{ij} V_{m,s(j)} \right. \right. \\
&\quad \left. \left. + \Delta G_{\text{wt},(i)} \right) \right] \exp \left[\frac{\beta^2 v}{2} \sum_{j,k}^N \sum_{m \neq s(i)}^M C_{ij} V_{m,s(j)} C_{ik} V_{m,s(k)} \right]
\end{aligned}$$

Density of states

The density of states (n), which in our case is equivalent to the distribution of stability effects, is related to the partition function by an inverse Laplace transform (Horwitz 1983).

Equation 64

$$\begin{aligned}
n_{(i)}(\Delta \Delta G) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\beta e^{i\beta \Delta \Delta G} z_{(i)}(i\beta) \\
&= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\beta \exp \left[-\frac{\beta^2 v}{2} \sum_{j,k}^N \sum_{m \neq s(i)}^M C_{ij} V_{m,s(j)} C_{ik} V_{m,s(k)} \right] \exp \left[i\beta \left(\Delta \Delta G \right. \right. \\
&\quad \left. \left. - \frac{1}{M-1} \sum_j^N \sum_{m \neq s(i)}^M C_{ij} V_{m,s(j)} + \Delta G_{\text{wt},(i)} \right) \right]
\end{aligned}$$

The integral can be solved by completing the square. Skipping steps, it can be shown that the following Gaussian distribution ($\mathcal{N}(\text{mean}, \text{variance})$) is exactly obtained (Smith 2011).

Equation 65

$$n_{(i)}(\Delta\Delta G) = \mathcal{N} \left(\frac{1}{M-1} \sum_j^N \sum_{m \neq s(i)}^M C_{ij} V_{ms(j)} - \sum_j^N C_{ij} V_{s(i)s(j)}, v \sum_{j,k}^N \sum_{m \neq s(i)}^M C_{ij} V_{ms(j)} C_{ik} V_{ms(k)} \right)$$

The expressions for the mean and variance of the Gaussian in Equation 65 are intuitive; they are just what one would expect if fitting a Gaussian to data whose free energy is determined by Equation 53. The mean is $\langle \Delta\Delta G_{(i)} \rangle$ and the variance is $\langle \Delta\Delta G_{(i)}^2 \rangle - \langle \Delta\Delta G_{(i)} \rangle^2$ given that $v = 1/(M-1) * [1-1/(M-1)]$ (Equation 59) and approximating $\langle \Delta\Delta G_{(i)} \rangle^2$ as $\langle \Delta\Delta G_{(i)}^2 \rangle / (M-1)$ (Equation 66).

Equation 66

$$\begin{aligned} \langle \Delta\Delta G_{(i)} \rangle^2 &= \left(\frac{1}{M-1} \sum_m^{M-1} \Delta\Delta G_{(i),m} \right)^2 \\ &= \frac{1}{(M-1)^2} \sum_m^{M-1} \sum_n^{M-1} \Delta\Delta G_{(i),m} \Delta\Delta G_{(i),n} \\ &= \frac{1}{(M-1)^2} \left(\sum_m^{M-1} \Delta\Delta G_{(i),m}^2 + \sum_m^{M-1} \sum_{n \neq m}^{M-2} \Delta\Delta G_{(i),m} \Delta\Delta G_{(i),n} \right) \\ &\approx \frac{1}{(M-1)^2} \sum_m^{M-1} \Delta\Delta G_{(i),m}^2 = \frac{1}{M-1} \langle \Delta\Delta G_{(i)}^2 \rangle \end{aligned}$$

In principle, the form of the distribution yielded by our framework is fairly robust to the precise definition of ΔG in Equation 53. As long as terms making up ΔG are no higher than quadratic in σ and per residue position terms cannot self-interact ($\mathbf{U}_{(i,i)} = \mathbf{0}_{(i,i)}$), a Gaussian distribution can be exactly derived within our theoretical framework.

In other words, after applying the continuous spin approximation and taking $M \gg 1$, the inverse Laplace transform of the partition function exactly yields a Gaussian (no high temperature approximation is needed). However, omitted terms from our definition of ΔG , like conformational entropy, probably cannot be approximated by just one-body or two-body interactions (D'Aquino et al. 1996; Jacobs and Shakhnovich 2016).

In the next section, we assess whether our underlying assumptions are reasonable by testing predictions with computational and experimental $\Delta\Delta G$ data. Computational $\Delta\Delta G$ data comes from FoldX, a freely available software whose force field is much more complicated than the two-body interactions considered here. Among other terms, FoldX accounts for the unfolded state and conformational entropy to yield a final $\Delta\Delta G$ value (Schymkowitz et al. 2005). We posit that residue positions found to not have $\Delta\Delta G$ s that are Gaussian distributed represent a breakdown of the theory because higher-order interactions are non-negligible.

Per residue position stability effects are Gaussian distributed

Protein name	acronym	PDB	length	organism
Recombinant serum paraoxonase	PON	1v04	332	<i>Oryctolagus cuniculus</i>
Lipase	Lip	1ex9	285	<i>Pseudomonas aeruginosa</i>
β -Lactamase	TEM1	1btl	263	<i>Escherichia coli</i>
Human carbonic anhydrase II	CAII	1lug	259	<i>Homo sapiens</i>
Dihydrofolate reductase	DHFR	1rx2	159	<i>Escherichia coli</i>
Ribonuclease H	RNaH	2rn2	155	<i>Escherichia coli</i>
Myoglobin	Myo	1a6k	151	<i>Physeter macrocephalus</i>
Staphylococcus nuclease	SNase	1stn	136	<i>Staphylococcus aureus</i>
Human lysozyme	huLys	1rex	130	<i>Homo sapiens</i>
Hen lysozyme	henLys	1dpx	129	<i>Gallus gallus</i>
Ribonuclease A	RNaA	1fs3	124	<i>Bos taurus</i>
Barnase	Barn	1a2p	108	<i>Bacillus amyloliquefaciens</i>
Acylphosphatase	AcP	2acy	98	<i>Bos taurus</i>
Ubiquitin	Ubiqu	1ubq	76	<i>Homo sapiens</i>
Protein G, IgG-binding domain	proG	2igd	61	<i>Streptococcus sp. group G</i>
Cro repressor	Cro	1orc	59	<i>Escherichia phage lambda</i>

Protein G, β 1 domain	G β 1	1pga	56	<i>Streptococcus sp. group G</i>
-----------------------------	-------------	------	----	----------------------------------

Table 3: Proteins whose $\Delta\Delta$ Gs are studied. All protein $\Delta\Delta$ Gs, except G β 1, was computationally calculated via FoldX by Tokuriki et al. (2007). G β 1 $\Delta\Delta$ Gs were experimentally measured by Nisthal et al. (2019). Not all of G β 1 $\Delta\Delta$ Gs were measured because of experimental constraints.

To test whether per residue position $\Delta\Delta$ G distributions are indeed Gaussian as predicted by Equation 65, we first organize $\Delta\Delta$ Gs into groups of 19 $\Delta\Delta$ Gs belonging to individual residue positions to obtain per position $\Delta\Delta$ G distributions. We then run the Shapiro-Wilk test (Shapiro and Wilk 1965) to analyze how well a Gaussian distribution fits the 19 $\Delta\Delta$ Gs, with a mean and variance calculated from the data. Our null hypothesis is that the distribution is a Gaussian. As the Shapiro-Wilk test statistic decreases and the p-value subsequently decreases, we can be more confident that the distribution is not Gaussian, and our null hypothesis is wrong. For example, in Figure 22 we run the Shapiro-Wilk test on the first position of dihydrofolate reductase (DHFR) and find that the test statistic is small enough that we can reject the null hypothesis with a p-value threshold of 0.05. In other words, the 19 $\Delta\Delta$ Gs pertaining to the first position of DHFR is not Gaussian distributed, at odds with our prediction (further discussed below).

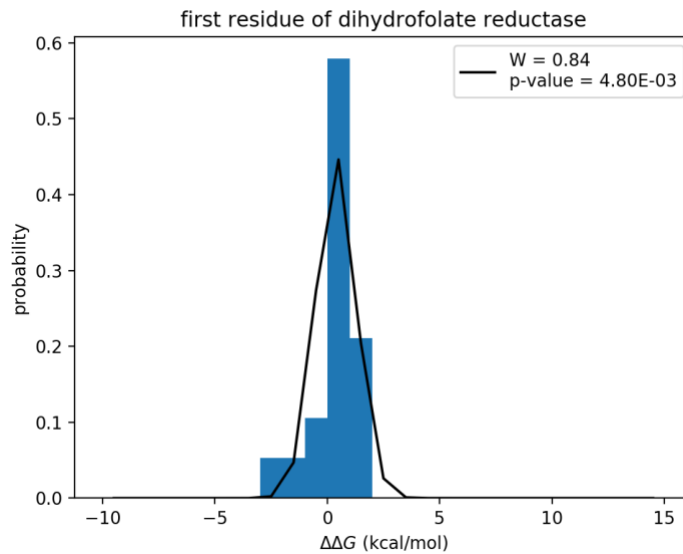


Figure 22: Example which demonstrates how we employ the Shapiro-Wilk test to determine whether the $\Delta\Delta$ G distribution across all mutants for a given residue position is Gaussian. We find that the first position of dihydrofolate

reductase has a $\Delta\Delta G$ distribution (blue bar plots) significantly different from that of a Gaussian (black line) according to the Shapiro-Wilk test. W is the Shapiro-Wilk test statistic.

Running the Shapiro-Wilk test for all residue positions from 16 proteins for which FoldX data has been previously tabulated (Tokuriki et al. 2007) (Table 3), we find that a majority of positions have $\Delta\Delta G$ distributions indistinguishable from a Gaussian (Figure 23). With a more stringent p-value threshold of 0.001, no protein rejects the null hypothesis for more than 4% of its residue positions (hatched bar plots in Figure 23). We can argue that a more stringent p-value threshold is necessary to address the multiple testing problem (Yoav and Hochberg 1995).

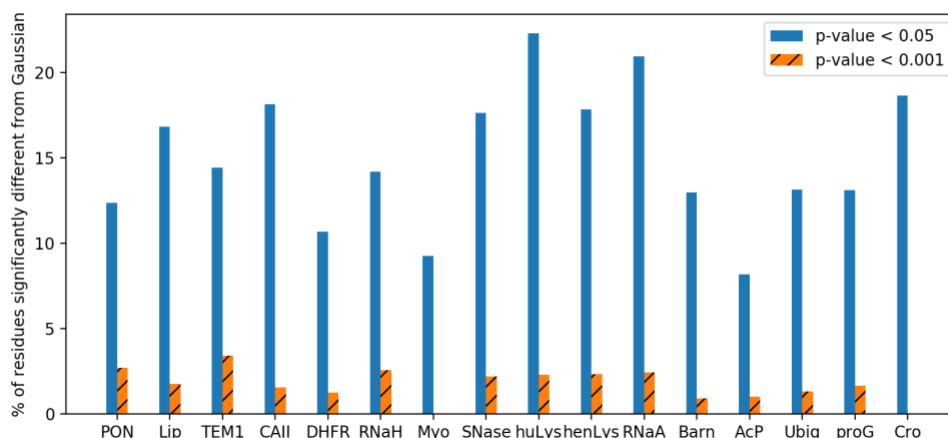


Figure 23: Relatively few residue positions per protein fail to match our theoretical prediction that per position $\Delta\Delta G$ s are Gaussian distributed according to the Shapiro-Wilk test. Solid bar plots correspond to residues that reject the null at a p-value threshold of 0.05; hatched, reject the null at a p-value of 0.001. For protein acronyms please see Table 3.

We further investigate the 10-25% of positions that do reject the null hypothesis at a p-value threshold of 0.05, like the first position of DHFR mentioned above. In Figure 24, we show box plots of contacts for positions that reject the null and for all positions. For 11 proteins, the median number of contacts for positions that reject the null is less than the median number of contacts for all positions. For five proteins, the medians are equivalent; and for one protein, Ribonuclease A (RNaseA), the median number of contacts for positions that reject the null is greater than the median number of contacts for all

positions. However, a more detailed analysis comparing the distributions of the number of contacts using the Kolmogorov-Smirnov test (Kolmogorov 1933), yields only human carbonic anhydrase II (CAII) to have significantly different contact distributions between those positions that reject the null and all positions (p -value = 0.037). The lack of a structural signature of those positions that reject the null makes sense in light of our derivation, which requires no assumption on the number of contacts.

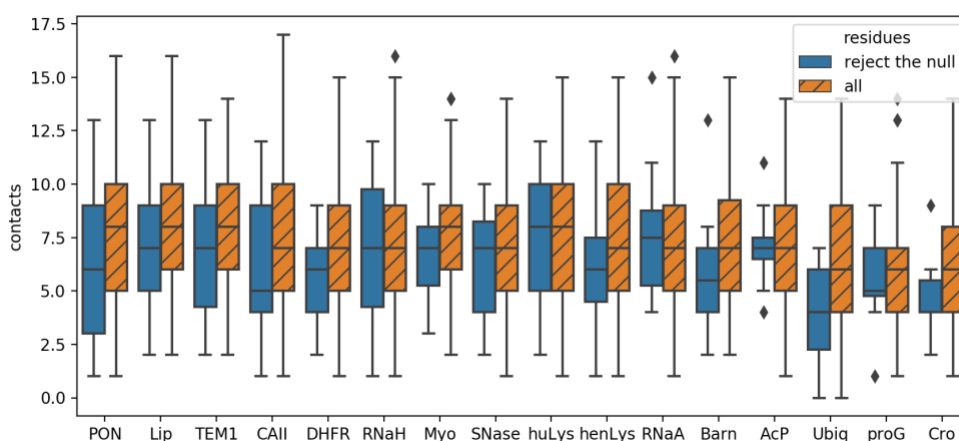


Figure 24: The distribution of contacts for residue positions found to reject the null hypothesis in Figure 23 at a p -value threshold of 0.05 (solid box plots) is similar to that of all protein residue positions (hatched box plots). Horizontal black lines correspond to medians; diamonds correspond to outliers.

When analyzing the wildtype amino acids of those residue positions that reject the null at a p -value threshold of 0.05 according to the binomial test, glycine is the most prominent among all 16 proteins studied (Figure 25). Glycine was found to be enriched at values significantly higher than its composition in the respective protein for five different proteins. Our model sometimes fails for glycine because it is the smallest amino acid and conformational entropy decreases for its mutants (D'Aquino et al. 1996). Positions with charged wildtype residues are also sometimes found to be outliers. For example, four proteins have lysine residues that reject the null at enriched values. Neglecting the unfolded state in our model may be the culprit, as long-range charge interactions may

greatly alter the unfolding for a non-charged amino acid mutant. Lysine is also known to entropically stabilize the native state (Berezovsky et al. 2005). However, it is worth noting that every major amino acid class: hydrophobic, polar and charged, has at least one amino acid not enriched for any of the 16 proteins studied (Figure 25). This indicates that our derivation does not systematically fail and lends credence to the form of our ΔG expression in Equation 53.

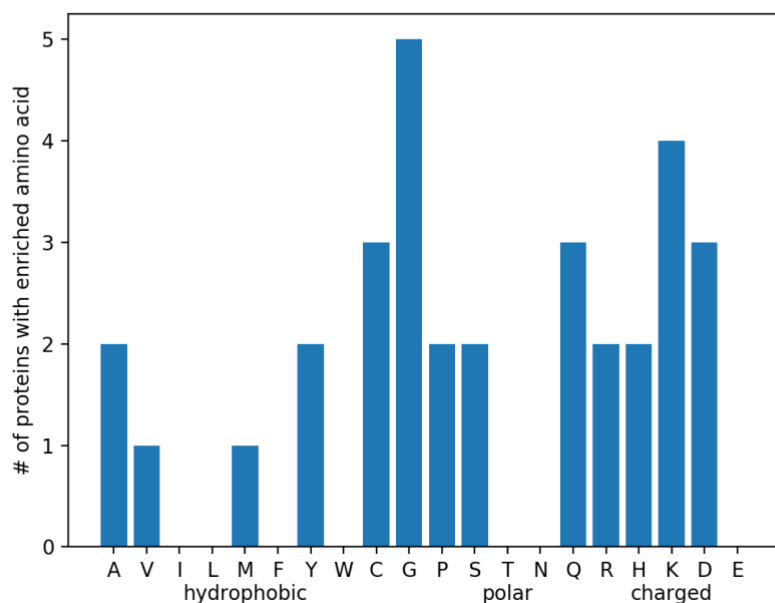


Figure 25: Amino acids that are enriched among residue positions found to reject the null hypothesis in Figure 23 at a p -value threshold of 0.05 for all 16 proteins studied. An amino acid is considered enriched for that protein if it is rejected at a higher fraction than its respective protein composition at a p -value less than 0.05, according to the binomial test. Results are not biased by any one of the 16 proteins; the maximal number of different enriched amino acids for one protein is found to be 4 and each protein has at least one enriched amino acid.

Experimental $\Delta\Delta G$ s

A major caveat of the previous subsection is that $\Delta\Delta G$ s are generated by FoldX. Exhaustive studies of protein $\Delta\Delta G$ distributions have been limited to computational calculations because experiments are costly and difficult to perform in a reasonable timeframe. Recent developments have made protein-wide $\Delta\Delta G$ measurements feasible by developing an experimental protocol utilizing automated robotics for protein purification

and chemical denaturant curve measurements (Nisthal et al. 2019). With this protocol, Nisthal et al. (2019) quantitatively measured the $\Delta\Delta G$ s of almost all single mutants (78%) of the $\beta 1$ domain of Streptococcal protein G ($G\beta 1$).

We find that per residue position $\Delta\Delta G$ measurements are mostly Gaussian distributed, but at a ratio on the high end of those found in Figure 23. Including all $G\beta 1$ residue positions with at least three measured $\Delta\Delta G$ s (54 out of 56 total residues), 13 (24%) positions reject the null hypothesis at a p-value threshold of 0.05; 6 (11%), at a p-value of 0.001 (Figure 26). We test whether having few $\Delta\Delta G$ s measured for some positions explains the relatively high percentages of null hypotheses rejected, however, this was not the case. When only analyzing positions with the maximal number of 17 $\Delta\Delta G$ s measured (29 out of 56 total residues), 11 positions still reject the null at a p-value threshold of 0.05; 5, at a p-value of 0.001 (Figure 26). The contact distribution of those 11 positions that reject the null does not matter either; its contact distribution is indistinguishable with those of all positions according to the Kolmogorov-Smirnov test (p-value = 0.40).

When analyzing the amino acid types of the 11 residue positions with $\Delta\Delta G$ s that reject the null according to the binomial test, we find that alanine (p-value = 0.0037) and isoleucine (p-value = 0.016) reject the null at enriched fractions with respect to $G\beta 1$'s amino acid composition. Given that alanine and isoleucine are small amino acids, we posit that our derivation is failing because configurational entropy is not accounted, as also seen for some of the 16 proteins computationally studied in the previous subsection (Figure 25).

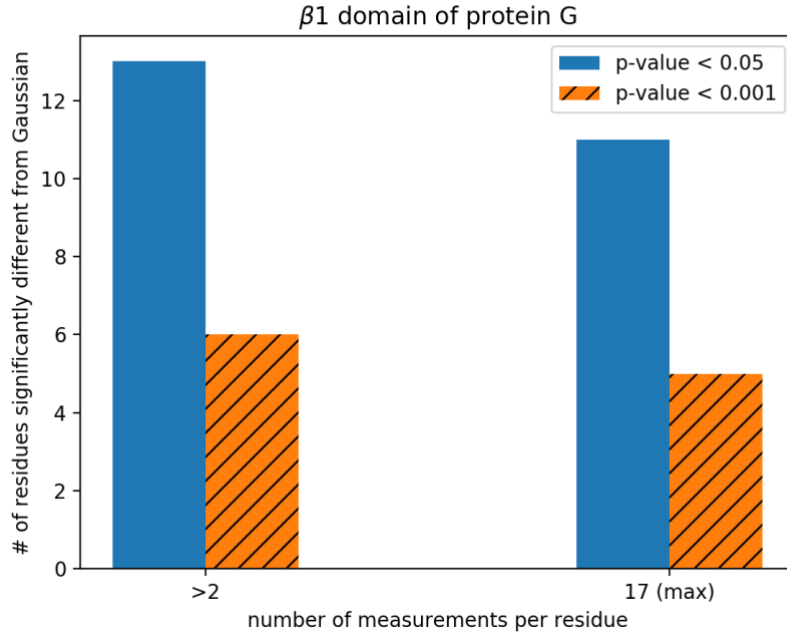


Figure 26: Relatively few residue positions of $\beta 1$ domain of Streptococcal protein G (56 total residues) fail to match our theoretical prediction that per residue position $\Delta\Delta G$ s are Gaussian distributed according to the Shapiro-Wilk test. Solid bar plots correspond to residue positions that reject the null at a p-value threshold of 0.05; hatched, reject the null at a p-value of 0.001. Positions that reject the null mostly have the maximal number of mutants measured in the study (right two-most bar plots).

Per residue position to per protein stability effects

Our derivation falls short from generally showing that the overall $\Delta\Delta G$ distribution of all single mutants across all residue positions ($n(\Delta\Delta G)$) is Gaussian. By definition, $n(\Delta\Delta G)$ is the weighted sum of all N Gaussians from Equation 65 because it corresponds to sequentially picking a single mutant out of N residue positions, i.e. a statistical OR operation.

Equation 67

$$n(\Delta\Delta G) = \frac{1}{N} \sum_i^N n_{(i)}(\Delta\Delta G)$$

We are not aware of an approach using first principles to reduce the mixture of N Gaussians in Equation 67. Although the sum of two normally distributed random variables is normally distributed, it does not generally hold that the sum of two normal

distributions is another normal distribution (Lemons and Langevin 2002). However, we bioinformatically demonstrate that the N Gaussian model reduces to a bi-Gaussian (Figure 27 and Figure 28). Therefore, a more tractable Gaussian mixture model with fewer Gaussians can be employed with little loss in accuracy.

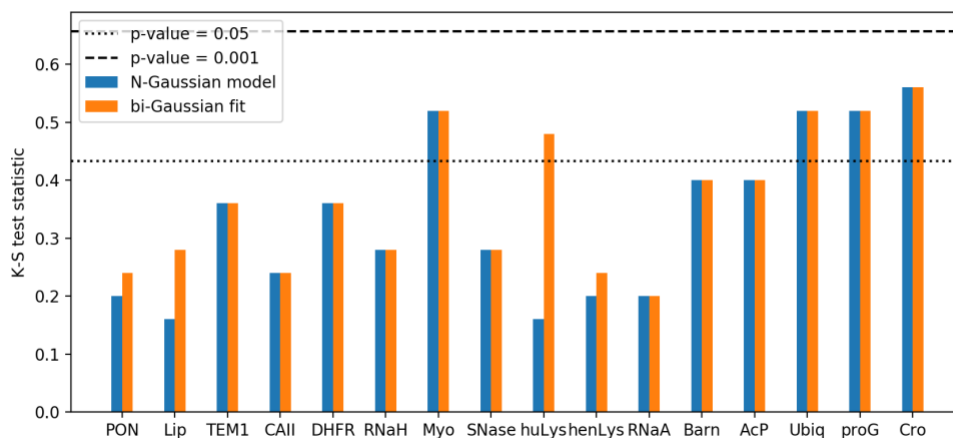


Figure 27: Our derived N-Gaussian model (blue bar plots) fits the computational $\Delta\Delta G$ data just as well, or even better than a bi-Gaussian fit (orange bar plots). The Kolmogorov-Smirnov (K-S) test was employed to determine goodness-of-fit.

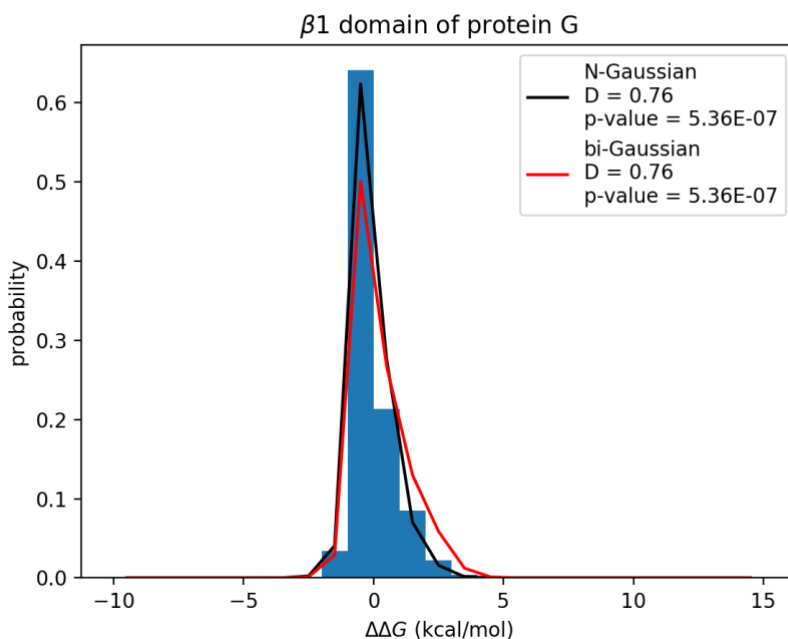


Figure 28: Our N-Gaussian model fits the experimental $\Delta\Delta G$ data of the $\beta 1$ domain of Streptococcal protein G just as well as a bi-Gaussian fit. However, both curves poorly fit the $\Delta\Delta G$ data according to the Kolmogorov-Smirnov test, where D is the Kolmogorov-Smirnov test statistic. One possible reason for the poor fit may be that experimental $\Delta\Delta G$ measurements are bound by the wildtype stability of the protein. Because $G\beta 1$'s $\Delta G = -4.0$ kcal/mol, it is not possible to purify a mutant protein with $\Delta\Delta G > 4.0$ kcal/mol and subsequently measure its $\Delta\Delta G$. Assuming protein mutants

without quantitative measurements in Nisthal et al. (2019) fall into this category, they could account for as much as 22% of $G\beta I$'s $\Delta\Delta G$ s. FoldX, on the other hand, is not bound by wildtype stability and can in principle yield large enough $\Delta\Delta G$ s that the mutant $\Delta G > 0$ kcal/mol. For the 16 natural proteins, non-negligible density in the distribution is seen for $\Delta\Delta G$ s as large as 6 kcal/mol (Tokuriki et al. 2007). Experiments that measure $\Delta\Delta G$ s for other proteins with greater wildtype stability such that a greater proportion of destabilizing mutants can be measured, will more conclusively determine whether our N-Gaussian model not only matches FoldX $\Delta\Delta G$ s (Figure 27), but also matches experimentally measured $\Delta\Delta G$ s.

Our results substantiate why a universal Gaussian distribution is empirically found for experimental and computational $\Delta\Delta G$ data (Tokuriki et al. 2007; Zeldovich et al. 2007; Faure and Koonin 2015). Regardless of protein structure, sequence, and to a certain extent the explicit $\Delta\Delta G$ expression (as long as it is dominated by pairwise interactions), we find that a Gaussian distribution is to be expected at the per residue position level based on first-principles statistical mechanics because of the large number of amino acid types. This helps substantiates the second assumption of the protein misfolding avoidance hypothesis (Figure 1 and Equation 23), which must apply to capture the abundance-evolutionary rate seen across proteins of all organisms.⁸

⁸ This chapter is a slightly modified version of Razban and Shakhnovich (2020) published in *Biophysical Journal* and has been reproduced here with the permission of the copyright holder.

Conclusion

Evolution explains everything in biology (Dobzhansky 1973). By identifying the biophysical factors influencing organismal fitness, we can begin to understand the curious observations from data collected on extant organisms. The biological factor we focus on in this thesis is protein stability. And the curious observations we attempt to explain is the stability distribution of orthologs (Chapter 1), the universal abundance-evolutionary rate correlation seen across all organisms' proteomes (Chapter 2, 3) and the universal Gaussian form of protein mutant stability effects seen across all single mutants of proteins (Chapter 4).

The role of stability in all three observations is more precisely defined by the framework of the protein misfolding avoidance hypothesis (MAH). MAH is a leading hypothesis positing that misfolding toxicity is the universal force influencing protein evolution. Our mathematical modeling, based on first principles population dynamics and statistical mechanics, helps make sense of currently available data. We conclude that MAH may be true based on stability distribution of orthologs (Chapter 1) and abundance-evolutionary rate trends (Chapter 2), despite proteome-wide melting temperature data not correlating as expected with abundance or evolutionary rate (Chapter 3). In Chapter 4, we more rigorously motivate MAH's assumption that single mutant protein stability effects are universally distributed across proteins, regardless of structure or sequence.

Our results do not eliminate the need for proteome-wide protein free energies in order to conclusively test MAH and determine once and for all whether protein stability is a universal force in protein evolution. However, our results make progress in understanding currently available data, demonstrating that they fail to falsify MAH. If future data does conflict with MAH, there are numerous alternative hypotheses that may take its place (Zhang and Yang 2015). Here, we only tested one other alternative hypothesis, called flux dynamics theory, which quantitatively describes protein function. We determined that flux dynamics theory is unlikely to govern long term evolution for *folA* orthologs (Chapter 1).

Other hypotheses can be broadly separated into two classes: protein expression cost (J. L. Cherry 2010; Plata et al. 2010; Kepp and Dasmeh 2014) and cell toxicity due to some biophysical property (Tartaglia et al. 2007; Yang et al. 2012; Park et al. 2013). An outstanding question is whether the validity of multiple hypotheses is the reason for a robust A-ER correlation, as the different properties constructively work together to cause abundance to constrain evolution. Or whether one hypothesis dominates evolution, and all others are negligible. The challenge of assessing other hypotheses besides MAH is that key experimental parameters have yet to be measured. As shown in Chapter 1, fitness functions of different hypothesis (MAH and FDH) can be proportional. Thus, parameter fitting will fail to discern whether one hypothesis more accurately describes the data because in such a scenario both will yield equally adequate results. Throughout this thesis, we attempt to always employ experimentally measured parameters in our mathematical expression to make a full quantitative assessment. Although this limits the

number of options considered, the conclusions reached are stronger because they more directly result from the underlying physics.

Data availability.

ProteomeVis: a protein evolution web application

Proteomic data employed in Chapter 2 and 3 are deposited in ProteomeVis (proteomevis.chem.harvard.edu). We developed this web application to facilitate protein evolutionary studies by bringing disparate data onto one user-friendly web application (Figure 29). Besides protein abundances and sequence evolutionary rates, there is data on protein-protein interaction partners, dosage tolerance and different protein structure metrics, obtained through analyzing PDB files. ProteomeVis can also provide comparison of proteins based on user-defined sequence identity and structure similarity scores, bridging the long timescale of protein structure evolution (~1 billion years) to the shorter timescales of protein sequence evolution.

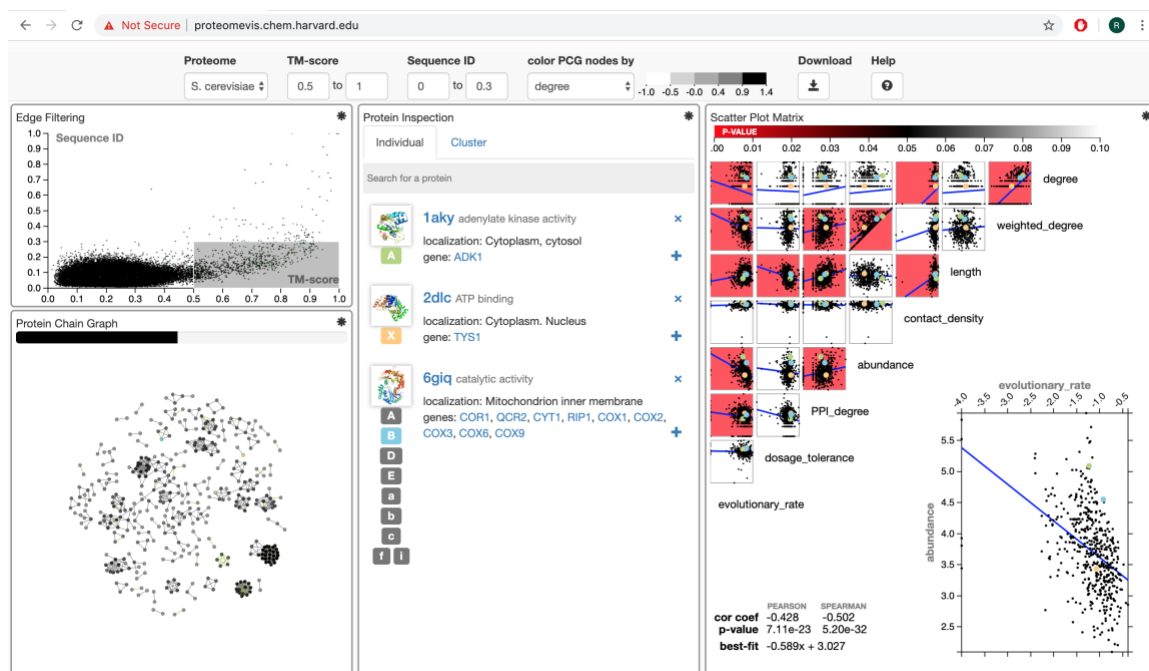


Figure 29: The ProteomeVis web app after exploration has occurred. Hovering over asterisks located in the top-right corner of each panel opens an informational window. Pressing the help button (question mark symbol) located to the right of the Control Strip at the top of the webpage leads to a quick tour. Pressing the download button, next to the

help button, allows users to download the underlying data. For a much more detailed introduction of the web app, please refer to Razban et al. (2018).

To validate the data making up ProteomeVis, we attempted to reproduce well-established results in the protein evolution field: protein structure similarities have a scale-free distribution and abundance is the strongest sequence evolutionary rate correlate. An interesting result in the field of protein structure evolution is that despite relatively high sequence divergence between two proteins, their structures can be the same (Chothia and Lesk 1986; Rost 1999; Gilson et al. 2017). To study this behavior quantitatively, people have plotted the distribution of such proteins and found the characteristic scale-free distribution of $k^{-\gamma}$, where k is the degree of the node and γ is the scaling exponent (Dokholyan et al. 2002; Deeds et al. 2004). The scaling exponent for *E.coli* proteins was measured to be 1.56, which we are able to reproduce with data in ProteomeVis (Figure 30). Note that graphs are composed of protein pairs from the respective organism with template-modeling (TM) scores greater than 0.5 and sequence identities less than 0.25 (Xu and Zhang 2010).

Moreover, ProteomeVis allows us to extend previous studies done only on one bacteria proteomes by examining protein structures of the *S.cerevisiae* proteome in the same manner as we did for *E.coli* proteins. Both organisms exhibit scaling constants that fall within the 1.4–1.9 range of scaling constants output by Big Bang model simulations (Dokholyan et al. 2002). This model assumes the protein structure diversity results from divergence, rather than convergence. Although *S.cerevisiae* and *E.coli* belong to different kingdoms of life, their scaling constants are captured by the Big Bang model, indicating that the mechanism behind protein structure evolution may be inherent to the last universal common ancestor.

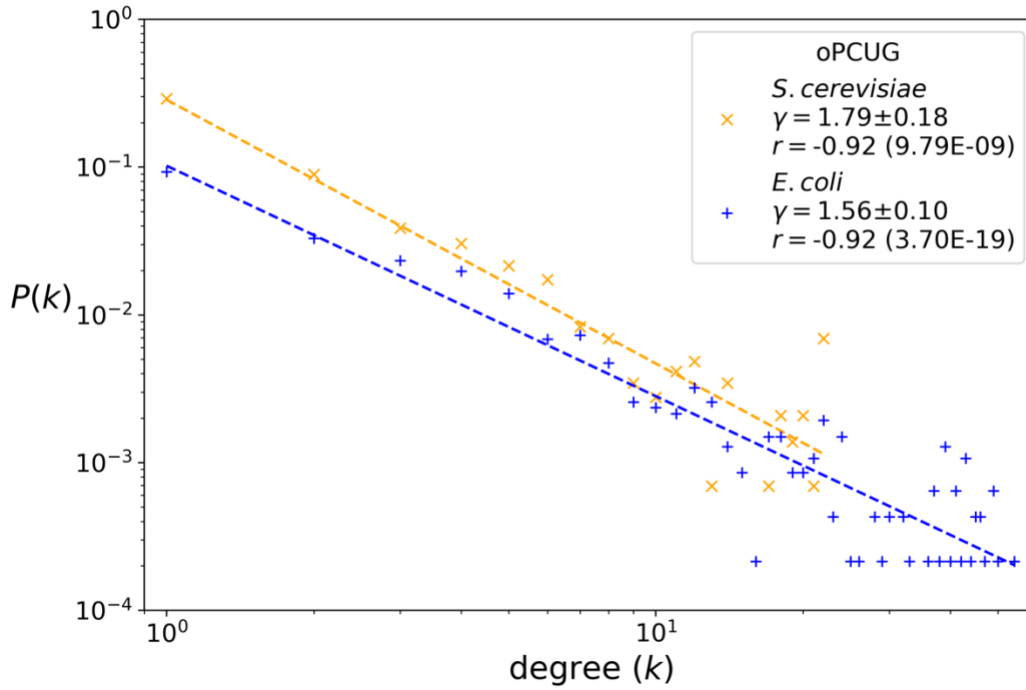


Figure 30: Degree distribution of the organismal protein chain universe graph (oPCUG) with its best-fit line on a log-log plot. The slope of the best-fit line (scaling constant, γ) with its standard error and the Pearson correlation coefficient (r) with its P-value in parenthesis are calculated by Python SciPy's linear regression function. Less *S.cerevisiae* datapoints are seen because its maximal degree is less than *E.coli*'s maximal degree. All degrees are horizontally shifted by 1 unit to allow viewing of orphans (degree=0) and be consistent with Deeds et al. (2004) in evaluating the scaling constant.

Many protein properties have been proposed to constrain sequence evolution. In Figure 31, we reproduce various correlations involving sequence evolutionary rate (ER) with our extended dataset for *E.coli* and *S.cerevisiae*. For the most part, dashed bar plots (literature values) agree with solid bar plots (ProteomeVis values). One correlation that disagrees with previous values is that between length and ER for *E.coli*. Although ProteomeVis agrees with Zhou et al. (2008) regarding correlations between contact density and ER, they vary regarding the correlation between protein length and ER for *E.coli*. Methodological differences do not explain the discrepancy. Zhou et al. (2008) applied the same filter based on length to identify structures in the respective organism's proteome as in ProteomeVis. As expected, the length versus ER correlation for

S.cerevisiae in ProteomeVis and Zhou et al. (2008) are consistent. Further studies are required to elucidate why inconsistent length-ER correlation are seen in *E.coli* but not *S.cerevisiae*.

Our results go beyond replicating results by demonstrating that the mRNA abundance-ER correlation previously seen in the literature is roughly equivalent to the A-ER correlation. This is nontrivial because the correlation between mRNA and protein abundance is far from perfect (Greenbaum et al. 2003; Taniguchi et al. 2011; Lahtvee et al. 2017). With respect to the protein misfolding avoidance hypothesis, this is an important result because protein abundance is the property of interest, not mRNA abundance (Figure 1).

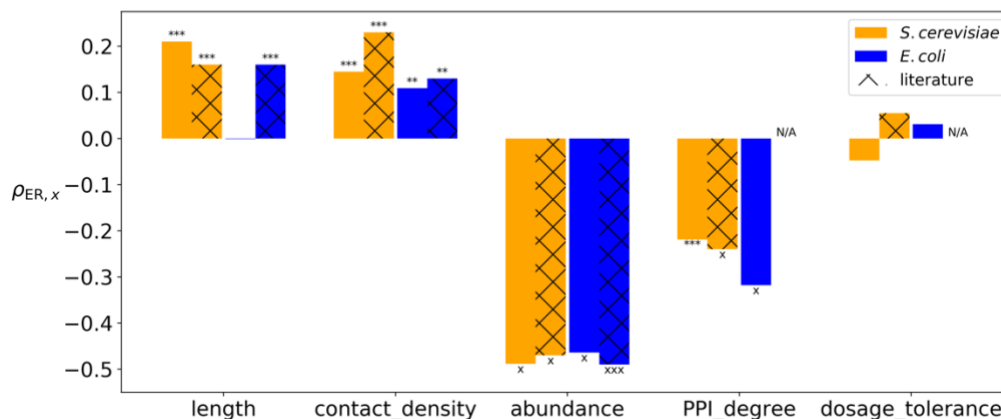


Figure 31: Spearman correlation coefficients (ρ) between evolutionary rate (ER) and five protein properties (x) in ProteomeVis and the literature. X marks and asterisks above or below bar plots denote p-value ranges: xxx = $P\text{-value} < 10^{100}$, xx < 10^{50} , x < 10^{10} , *** < 0.001, ** < 0.01, * < 0.05. Length (Zhou et al. 2008), contact density (Zhou et al. 2008) and (mRNA) abundance (Zhang and Yang 2015) correlations with evolutionary rate in the literature for both organisms are presented. PPI degree and dosage tolerance correlations with evolutionary rate in the literature for *S.cerevisiae* are from (Fraser et al. 2003) and (Vavouri et al. 2009), respectively; for *E.coli*, not available (N/A).

By integrating data from three published papers, programmatically accessed from four databases and generated by running three software packages, ProteomeVis enables users to make quick insights with a disparate set of data. Once more detailed hypotheses are formulated throughout their web sessions, users can freely download the data to

further probe on their own (Figure 29). We recommend rigorous hypothesis testing rather than superficially scanning for significant p-values to avoid the multiple testing problem (Yoav and Hochberg 1995).⁹

Data and code are available on GitHub

Source code and scripts employed to collect the underlying data in ProteomeVis can be found at github.com/rrazban. The Django web framework encoding the ProteomeVis web app layout can also be assessed at github.com/rrazban/proteomevis.

In general, data and data analysis scripts employed to reach conclusions reported in this thesis can also be accessed online through GitHub. An overview of repositories relevant to this thesis are summarized in Table 4. For example, to check the numerical expressions with our analytical expressions for $P_{ss}(\Delta G)$ and $ER(A^{tot})$ (Figure 4 and Figure 13, respectively), please take a look at the `dG_in_evolution` repository. To further scrutinize computational or experimental $\Delta\Delta G$ data in Chapter 4, please check out ‘`ddG_distribution`.’ Abundance and evolutionary rate experimental data used in Chapters 2 and 3 can be found in the `proteomevis_scripts` repository. Raw data relevant to Chapter 3 in assessing the accuracy of the approximate ΔG mathematical relationship as a function of melting temperature (Equation 46 and Equation 48) can be found in the Supplementary Material in Razban (2019).

Repository name	Description	Thesis chapter
<code>dG_in_evolution</code>	Solves ΔG master equation and compares analytical expressions with numerical solutions	1, 2
<code>ddG_distribution</code>	Performs statistical tests to determine how well Gaussian distribution fits $\Delta\Delta G$ s	4

⁹ This section is a slightly modified version of Razban et al. (2018) published in *Bioinformatics* and has been reproduced here with the permission of the copyright holder.

proteomevis_scripts	Protein abundance, melting temperature and evolutionary rate data	2, 3
proteomevis	Source code for web application that presents data collected in proteomevis_scripts	

Table 4: Repository names that can be found at github.com/rrazban. To access a repository online, simply insert the repository name after my GitHub directory, e.g. github.com/rrazban/dG_in_evolution.

References

- Adkar B V., Bhattacharyya S, Gilson AI, Zhang W, Shakhnovich EI. 2019. Substrate inhibition imposes fitness penalty at high protein stability. *Proc. Natl. Acad. Sci. U. S. A.*:201821447.
- Adkar B V., Manhart M, Bhattacharyya S, Tian J, Musharbash M, Shakhnovich EI. 2017. Optimization of lag phase shapes the evolution of a bacterial enzyme. *Nat. Ecol. Evol.* 1:1–6.
- Agozzino L, Dill KA. 2018. Protein evolution speed depends on its stability and abundance and on chaperone concentrations. *Proc. Natl. Acad. Sci.*:201810194.
- Akanuma S, Nakajima Y, Yokobori SI, Kimura M, Nemoto N, Mase T, Miyazono KI, Tanokura M, Yamagishi A. 2013. Experimental evidence for the thermophilicity of ancestral life. *Proc. Natl. Acad. Sci. U. S. A.* 110:11067–11072.
- Anscombe FJ. 1973. Graphs in Statistical Analysis. *Am. Stat.* 27:17–21.
- Arike L, Valgepea K, Peil L, Nahku R, Adamberg K, Vilu R. 2012. Comparison and applications of label-free absolute proteome quantification methods on *Escherichia coli*. *J. Proteomics* 75:5437–5448.
- Arthur RK, Ruvinsky I. 2011. Evidence that purifying selection acts on promoter sequences. *Genetics* 189:1121–1126.
- Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: The Keio collection. *Mol. Syst. Biol.* 2.
- Becher I, Andre A, Romanov N, Bork P, Beck M, Savitski MM, Romanov N, Stein F, Schramm M, Baudin F, et al. 2018. Pervasive Protein Thermal Stability Variation during the Cell Cycle Pervasive Protein Thermal Stability Variation during the Cell Cycle. *Cell* 173:1495–1507.
- Becktel WJ, Schellman JA. 1987. Protein stability curves. *Biopolymers* 26:1859–1877.
- Berezovsky IN, Chen WW, Choi PJ, Shakhnovich EI. 2005. Entropic stabilization of proteins and its proteomic consequences. *PLoS Comput. Biol.* 1:0322–0332.
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res.* 28:235–242.
- Bershtein S, Choi JM, Bhattacharyya S, Budnik B, Shakhnovich E. 2015. Systems-Level

- Response to Point Mutations in a Core Metabolic Enzyme Modulates Genotype-Phenotype Relationship. *Cell Rep.* 11:645–656.
- Bershtein S, Mu W, Serohijos AWR, Zhou J, Shakhnovich EI. 2013. Protein Quality Control Acts on Folding Intermediates to Shape the Effects of Mutations on Organismal Fitness. *Mol. Cell* 49:133–144.
- Bershtein S, Mu W, Shakhnovich EI. 2012. Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. *Proc. Natl. Acad. Sci.* 109:7126–7126.
- Bershtein S, Serohijos AWR, Bhattacharyya S, Manhart M, Choi J-M, Mu W, Zhou J, Shakhnovich EI. 2015. Protein Homeostasis Imposes a Barrier on Functional Integration of Horizontally Transferred Genes in Bacteria. *PLoS Genet.* 11.
- Best RB, Hummer G, Eaton WA. 2013. Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci.* 110:17874–17879.
- Bhattacharyya S, Jacobs WM, Adkar B V., Yan J, Zhang W, Shakhnovich EI. 2018. Accessibility of the Shine-Dalgarno Sequence Dictates N-Terminal Codon Bias in *E. coli*. *Mol. Cell* 70:894-905.e5.
- Bobay L-M, Ochman H. 2018. Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evol. Biol.* 18:1–12.
- Castro Sotos AE, Vanhoof S, Van den Noortgate W, Onghena P. 2009. The Transitivity Misconception of Pearson's Correlation Coefficient. *Stat. Educ. Res. J.* 8:33–55.
- Chandler D. 1987. Introduction to modern statistical mechanics. Oxford University Press
- Chen P, Shakhnovich EI. 2009. Lethal mutagenesis in viruses and bacteria. *Genetics* 183:639–650.
- Cherry Joshua L. 2010. Highly expressed and slowly evolving proteins share compositional properties with thermophilic proteins. *Mol. Biol. Evol.* 27:735–741.
- Cherry J. L. 2010. Expression level, evolutionary rate, and the cost of expression. *Genome Biol. Evol.* 2:757–769.
- Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–826.
- D'Aquino JA, Gómez J, Hilser VJ, Lee KH, Amzel LM, Freire E. 1996. The magnitude of the backbone conformational entropy change in protein folding. *Proteins Struct. Funct. Genet.* 25:143–156.
- Davidi D, Noor E, Liebermeister W, Bar-Even A, Flamholz A, Tummeler K, Barenholz U, Goldenfeld M, Shlomi T, Milo R. 2016. Global characterization of in vivo enzyme catalytic rates and their correspondence to in vitro k_{cat} measurements. *Proc. Natl.*

- Acad. Sci. 113:3401–3406.
- Deeds EJ, Shakhnovich B, Shakhnovich EI. 2004. Proteomic Traces of Speciation. *J. Mol. Biol.* 336:695–706.
- Dobzhansky T. 1973. Nothing in Biology Makes Sense Except in the Light of Evolution. *Am. Biol. Teach.* 35:125–129.
- Dokholyan N V, Shakhnovich B, Shakhnovich EI. 2002. Expanding protein universe and its origin from the biological Big Bang. *Proc. Natl. Acad. Sci. U. S. A.* 99:14132–14136.
- Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of Spontaneous Mutation. *Genetics* 148:1667–1686.
- Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U. S. A.* 102:14338–14343.
- Drummond DA, Raval A, Wilke CO. 2006. A single determinant dominates the rate of yeast protein evolution. *Mol. Biol. Evol.* 23:327–337.
- Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.
- Dykhuizen DE, Dean AM, Hartl DL. 1987. Metabolic flux and fitness. *Genetics* 115:25–31.
- Edman P. 1950. Method for Determination of the Amino Acid Sequence in Peptides. *Acta Chem. Scand.* 4:283–293.
- Ellenberg J. 2014. How not to be wrong. New York: The Penguin Press
- England JL, Shakhnovich EI. 2003. Structural determinant of protein designability. *Phys. Rev. Lett.* 90:218101.
- Erdelyi A. 1956. Asymptotic Expansions. New York: Dover Publications, Inc.
- Faure G, Koonin E V. 2015. Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. *Phys. Biol.* 12:035001.
- Feng Y, De Franceschi G, Kahraman A, Soste M, Melnik A, Boersema PJ, De Laureto PP, Nikolaev Y, Oliveira AP, Picotti P. 2014. Global analysis of protein structural changes in complex proteomes. *Nat. Biotechnol.* 32:1036–1044.
- Fraser HB, Wall DP, Hirsh AE. 2003. A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol. Biol.* 3:1–6.
- Geiler-Samerotte KA, Dion MF, Budnik BA, Wang SM, Hartl DL. 2010. Misfolded

- proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc. Natl. Acad. Sci. U. S. A.* 108:680–685.
- Ghaemmaghami S, Huh W-K, Bower K, Howson RW, Belle A, Dephoure N, O’Shea EK, Weissman JS. 2003. Global analysis of protein expression in yeast. *Nature* 425:737–741.
- Ghosh K, Dill KA. 2009. Computing protein stabilities from their chain lengths. *Proc. Natl. Acad. Sci. U. S. A.* 106:10649–10654.
- Gibson B, Wilson DJ, Feil E, Eyre-Walker A. 2018. The distribution of bacterial doubling times in the wild. *Proc. R. Soc. B Biol. Sci.* 285.
- Gilson AI, Marshall-Christensen A, Choi JM, Shakhnovich EI. 2017. The role of evolutionary selection in the dynamics of protein structure evolution. *Biophys. J.* 112:1350–1365.
- Gin BC, Garrahan JP, Geissler PL. 2009. The Limited Role of Nonnative Contacts in the Folding Pathways of a Lattice Protein. *J. Mol. Biol.* 392:1303–1314.
- Gō N. 1983. Theoretical Studies of Protein Folding. *Ann. Rev. Biophys. Bioeng.* 12:183–210.
- Greenbaum D, Colangelo C, Williams K, Gerstein M. 2003. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.* 4:117.
- Gromiha MM, Anoosha P, Huang L. 2016. Applications of Protein Thermodynamic Database for Understanding Protein Mutant Stability and Designing Stable Mutants. In: Carugo O, Walker JM, editors. *Data Mining Techniques for the Life Sciences*. Second. New York: Humana Press. p. 71–89.
- Guerois R, Nielsen JE, Serrano L. 2002. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J. Mol. Biol.* 320:369–387.
- Hart KM, Harms MJ, Schmidt BH, Elya C, Thornton JW, Marqusee S. 2014. Thermodynamic System Drift in Protein Evolution. *PLoS Biol.* 12.
- Hirsh AE, Fraser HB. 2001. Protein dispensability and rate of evolution. *Nature* 411:1046–1049.
- Hobbs JK, Shepherd C, Saul DJ, Demetras NJ, Haaning S, Monk CR, Daniel RM, Arcus VL. 2012. On the origin and evolution of thermophily: Reconstruction of functional precambrian enzymes from ancestors of *Bacillus*. *Mol. Biol. Evol.* 29:825–835.
- Hogg R V, McKean J, Craig AT. 2005. *Introduction to mathematical statistics*. Pearson Education

- Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA. 1998. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95:717–728.
- Horwitz G. 1983. Steepest descent path for the microcanonical ensemble-resolution of an ambiguity. *Commun. Math. Phys.* 89:117–129.
- Hurst LD, Smith NGC. 1999. Do essential genes evolve slowly? *Curr. Biol.* 9:747–750.
- Ionescu RM, Smith VF, O’Neill JC, Matthews CR. 2000. Multistate equilibrium unfolding of *Escherichia coli* dihydrofolate reductase: Thermodynamic and spectroscopic description of the native, intermediate, and unfolded ensembles. *Biochemistry* 39:9540–9550.
- Jacobs WM, Shakhnovich EI. 2016. Structure-Based Prediction of Protein-Folding Transition Paths. *Biophys. J.* 111:925–936.
- Jacobs WM, Shakhnovich EI. 2017. Evidence of evolutionary selection for co-translational folding. *Proc. Natl. Acad. Sci. U. S. A.* 114:11434–11439.
- Jarzab A, Kurzawa N, Hopf T, Moerch M, Zecha J, Leijten N, Bian Y, Musiol E, Maschberger M, Stoehr G, et al. 2020. Meltome atlas—thermal proteome stability across the tree of life. *Nat. Methods*.
- Jiang L, Mishra P, Hietpas RT, Zeldovich KB, Bolon DNA. 2013. Latent Effects of Hsp90 Mutants Revealed at Reduced Expression Levels. *PLoS Genet.* 9.
- Kacser H, Burns JA. 1981. The molecular basis of dominance. *Genetics* 97:639–666.
- Kepp KP, Dasmeh P. 2014. A model of proteostatic energy cost and its use in analysis of proteome trends and sequence evolution. *PLoS One* 9.
- Kimura M. 1962. On the probability of fixation of mutant genes in a population. *Genetics* 47:713–719.
- Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–626.
- Kimura M, Ota T. 1974. On some principles governing molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 71:2848–2852.
- King JL, Jukes TH. 1969. Non-Darwinian Evolution. *Science* 164:788–798.
- Knies JL, Cai F, Weinreich DM. 2017. Enzyme Efficiency but Not Thermostability Drives Cefotaxime Resistance Evolution in TEM-1 β -Lactamase. *Mol. Biol. Evol.* 34:1040–1054.
- Kolmogorov A. 1933. Sulla determinazione empirica di una legge di distribuzione. *Inst. Ital. Attuari, Giorn.* 4:83–91.

- Koonin E V. 2009. The Origin at 150: is a new evolutionary synthesis in sight? *Trends Genet.* 25:473–475.
- Krug K, Carpy A, Behrends G, Matic K, Soares NC, Macek B. 2013. Deep coverage of the *Escherichia coli* proteome enables the assessment of false discovery rates in simple proteogenomic experiments. *Mol. Cell. Proteomics* 12:3420–3430.
- Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* 34:1812–1819.
- Lahtvee PJ, Sánchez BJ, Smialowska A, Kasvandik S, Elsemman IE, Gatto F, Nielsen J. 2017. Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast. *Cell Syst.* 4:495–504.e5.
- Langford E, Schwertman N, Owens M. 2001. Is the Property of Being Positively Correlated Transitive? *Am. Stat.* 55:322–325.
- Laub AJ. 2005. Kronecker Products. In: *Matrix Analysis for Scientists and Engineers*. Philadelphia: Society for Industrial and Applied Mathematics. p. 147–158.
- Lemons DS, Langevin P. 2002. *An introduction to stochastic processes in physics*. JHU Press
- Leuenberger P, Gansch S, Kahraman A, Cappelletti V, Boersema PJ, von Mering C, Claassen M, Picotti P. 2017. Cell-wide analysis of protein thermal unfolding reveals determinants of thermostability. *Science* 355:eaai7825.
- Liao BY, Zhang J. 2006. Evolutionary conservation of expression profiles between human and mouse orthologous genes. *Mol. Biol. Evol.* 23:530–540.
- Lim SA, Hart KM, Harms MJ, Marqusee S. 2016. Evolutionary trend toward kinetic stability in the folding trajectory of RNases H. *Proc. Natl. Acad. Sci. U. S. A.* 113:13045–13050.
- Lynch M, Conery JS. 2003. The Origins of Genome Complexity. *Science* 302:1401–1404.
- Margoliash E. 1963. Primary Structure and Evolution of Cytochrome C. *Proc. Natl. Acad. Sci. U. S. A.* 50:672–679.
- Mateus A, Bobonis J, Kurzawa N, Stein F, Helm D, Hevler J, Typas A, Savitski MM. 2018. Thermal proteome profiling in bacteria: probing protein state in vivo. *Mol. Syst. Biol.* 14:e8242.
- Miyazawa S, Jernigan RL. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256:623–644.

- Molina DM, Jafari R, Ignatushchenko M, Seki T, Larsson EA, Dan C, Sreekumar L, Cao Y, Nordlund P. 2013. Monitoring Drug Target Engagement in Cells and Tissues Using the Cellular Thermal Shift Assay. *Science* 341:84–88.
- Mustonen V, Lässig M. 2009. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet.* 25:111–119.
- Nisthal A, Wang CY, Ary ML, Mayo SL. 2019. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc. Natl. Acad. Sci. U. S. A.* 116:16367–16377.
- Pace CN, Shaw KL. 2000. Linear extrapolation method of analyzing solvent denaturation curves. *Proteins Suppl* 4:1–7.
- Pál C, Papp B, Hurst LD. 2001. Highly expressed genes in yeast evolve slowly. *Genet. Soc. Am.* 158:927–931.
- Pál C, Papp B, Hurst LD. 2003. Rate of evolution and gene dispensability. *Nat. Br. Commun.* 421:496–497.
- Park C, Chen X, Yang J-R, Zhang J. 2013. Differential requirements for mRNA folding partially explain why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. U. S. A.* 110:E678–E686.
- Pauling L, Zuckerkandl E. 1963. Chemical Paleogenetics Molecular “Resurrection Studies” of Extinct Forms of Life. *Acta Chem. Scand.* 17:9–16.
- Piela L. 2013. Dirac Delta Function. In: *Ideas of Quantum Chemistry*. Second. Elsevier. p. 69–72.
- Plata G, Gottesman ME, Vitkup D. 2010. The rate of the molecular clock and the cost of gratuitous protein synthesis. *Genome Biol.* 11:R98.
- Plata G, Vitkup D. 2017. Protein stability and avoidance of toxic misfolding do not explain the sequence constraints of highly expressed proteins. *Mol. Biol. Evol.* 35:700–703.
- Privalov PL, Khechinashvili NN. 1974. A thermodynamic approach to the problem of stabilization of globular protein structure: A calorimetric study. *J. Mol. Biol.* 86:665–684.
- Razban RM. 2019. Protein Melting Temperature Cannot Fully Assess Whether Protein Folding Free Energy Underlies the Universal Abundance–Evolutionary Rate Correlation Seen in Proteins. *Mol. Biol. Evol.* 36:1955–1963.
- Razban RM, Gilson AI, Durfee N, Strobelt H, Dinkla K, Choi JM, Pfister H, Shakhnovich EI. 2018. ProteomeVis: A web app for exploration of protein properties from structure to sequence evolution across organisms’ proteomes.

- Bioinformatics 34:3557–3565.
- Razban RM, Shakhnovich EI. 2020. Effects of Single Mutations on Protein Stability Are Gaussian Distributed. *Biophys. J.* 118:2872–2878.
- Rees DC, Robertson AD. 2001. Some thermodynamic implications for the thermostability of proteins. *Protein Sci.* 10:1187–1194.
- Reid AJ, Ranea JAG, Orengo CA. 2010. Comparative evolutionary analysis of protein complexes in *E. coli* and yeast. *BMC Genomics* 11:1–16.
- Rice J. 2007. Expected Values. In: *Mathematical Statistics and Data Analysis*. Third. Belmont, CA: Brooks/Cole Cengage Learning.
- Risso VA, Manssour-Triedo F, Delgado-Delgado A, Arco R, Barroso-delJesus A, Ingles-Prieto A, Godoy-Ruiz R, Gavira JA, Gaucher EA, Ibarra-Molero B, et al. 2015. Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history. *Mol. Biol. Evol.* 32:440–455.
- Robertson AD, Murphy KP. 1997. Protein Structure and the Energetics of Protein Stability. *Chem. Rev.* 97:1251–1268.
- Rodrigues J V, Bershtein S, Li A, Lozovsky ER, Hartl DL, Shakhnovich EI. 2016. Biophysical principles predict fitness landscapes of drug resistance. *Proc. Natl. Acad. Sci.* 113:E1964–E1964.
- Rost B. 1999. Twilight zone of protein sequence alignments. *Protein Eng. Des. Sel.* 12:85–94.
- Sanger F, Tuppy H. 1951. The Amino-acid Sequence in the Phenylalanyl Chain of Insulin 1. The identification of lower peptides from partial hydrolysates. *Biochem. J.* 49:463–481.
- Schad E, Tompa P, Hegyi H. 2011. The relationship between proteome size, proteome complexity and disorder and organism complexity. *Genome Biol.* 12:1–13.
- Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. 2005. The FoldX web server: an online force field. *Nucleic Acids Res.* 33:382–388.
- Sella G, Hirsh AE. 2005. The application of statistical physics to evolutionary biology. *Proc. Natl. Acad. Sci. U. S. A.* 102:9541–9546.
- Serohijos AWR, Lee SYR, Shakhnovich EI. 2013. Highly abundant proteins favor more stable 3D structures in yeast. *Biophys. J.* 104:L1–L3.
- Serohijos AWR, Rimas Z, Shakhnovich EI. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep.* 2:249–256.

- Serohijos AWR, Shakhnovich EI. 2014. Merging molecular mechanism and evolution: Theory and computation at the interface of biophysics and evolutionary population genetics. *Curr. Opin. Struct. Biol.* 26:84–91.
- Shakhnovich EI. 1997. Theoretical studies of protein-folding thermodynamics and kinetics. *Curr. Op. Struct. Biol.* 7:3–9.
- Shakhnovich EI, Gutin AM. 1990. Implications of thermodynamics of protein folding for evolution of primary sequences. *Nature* 346:773–775.
- Shapiro SS, Wilk MB. 1965. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52:591–611.
- Smith JO. 2011. Gaussian Integral with Complex Offset. In: *Spectral Audio Signal Processing*. W3K.
- Tan CSH, Go KD, Bisteau X, Dai L, Yong CH, Prabhu N, Ozturk MB, Lim YT, Sreekumar L, Lengqvist J, et al. 2018. Thermal proximity coaggregation for system-wide profiling of protein complex dynamics in cells. *Science* 1177:1–11.
- Taniguchi Y, Choi PJ, Li G, Chen H, Babu M, Hearn J, Emili A, Xie XS. 2011. Quantifying E.coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells. *Science* 329:533–539.
- Tartaglia GG, Pechmann S, Dobson CM, Vendruscolo M. 2007. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem. Sci.* 32:199–204.
- Tiana G, Broglia R, Provasi D. 2001. Designability of lattice model heteropolymers. *Phys. Rev. E* 64:011904.
- Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. 2007. The Stability Effects of Protein Mutations Appear to be Universally Distributed. *J. Mol. Biol.* 369:1318–1332.
- Trudeau DL, Kaltenbach M, Tawfik DS. 2016. On the Potential Origins of the High Stability of Reconstructed Ancestral Proteins. *Mol. Biol. Evol.* 33:2633–2641.
- Usmanova DR, Plata G, Vitkup D. 2020. Ability of the misfolding avoidance hypothesis to explain protein evolutionary rates in the light of empirical evidence. *bioRxiv*:1–4.
- Vavouri T, Semple JI, Garcia-Verdugo R, Lehner B. 2009. Intrinsic Protein Disorder and Interaction Promiscuity Are Widely Associated with Dosage Sensitivity. *Cell* 138:198–208.
- Venev S V., Zeldovich KB. 2015. Massively parallel sampling of lattice proteins reveals foundations of thermal adaptation. *J. Chem. Phys.* 143:055101.

- Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* 13:227–232.
- Walker EJ, Bettinger JQ, Welle KA, Hryhorenko JR, Ghaemmaghami S. 2018. Global analysis of methionine oxidation provides a census of folding stabilities for the human proteome. *Proc. Natl. Acad. Sci. U. S. A.*
- Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB, Feldman MW. 2005. Functional genomic analysis of the rates of protein evolution. *Proc. Natl. Acad. Sci. U. S. A.* 102:5483–5488.
- Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. 2015. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 15:3163–3168.
- Wang Z, Zhang J. 2009. Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS Genet.* 5.
- Wilke CO, Drummond DA. 2006. Population genetics of translational robustness. *Genetics* 173:473–481.
- Winzeler EA, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, et al. 1999. Functional Characterization of the *S. cerevisiae* Genome by Gene deletion and Parallel Analysis. *Science* 285:901–907.
- Wylie CS, Shakhnovich EI. 2011. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc. Natl. Acad. Sci. U. S. A.* 108:9916–9921.
- Xu J, Zhang Y. 2010. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* 26:889–895.
- Yang J-R, Liao B-Y, Zhuang S-M, Zhang J. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl. Acad. Sci. U. S. A.* 109:E831–E840.
- Yang JR, Maclean CJ, Park C, Zhao H, Zhang J. 2017. Intra and interspecific variations of gene expression levels in yeast are largely neutral: (Nei Lecture, SMBE 2016, Gold Coast). *Mol. Biol. Evol.* 34:2125–2139.
- Yang JR, Zhuang SM, Zhang J. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. *Mol. Syst. Biol.* 6:1–13.
- Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–1591.
- Yoav B, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57:289–300.

- Zeldovich KB, Chen P, Shakhnovich EI. 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc. Natl. Acad. Sci. U. S. A.* 104:16152–16157.
- Zhang J. 2000. Protein-length distributions for the three domains of life. *Trends Genet.* 16:107–109.
- Zhang J. 2010. Evolutionary Genetics: Progress and Challenges. In: *Evolution Since Darwin The First 150 Years*. p. 87–118.
- Zhang J, Yang J-R. 2015. Determinants of the rate of protein sequence evolution. *Nat. Rev. Genet.* 16:409–420.
- Zhou T, Drummond DA, Wilke CO. 2008. Contact density affects protein evolutionary rate from bacteria to animals. *J. Mol. Evol.* 66:395–404.
- Zhou T, Gu W, Wilke CO. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. *Mol. Biol. Evol.* 27:1912–1922.