Effects of single mutations on protein stability are Gaussian distributed

R. M. Razban and E. I. Shakhnovich

Abstract: The distribution of protein stability effects is known to be well-approximated by a Gaussian distribution from previous empirical fits. Starting from first-principles statistical mechanics, we more rigorously motivate this empirical observation by deriving per residue protein stability effects to be Gaussian. Our derivation requires the number of amino acids to be large, which is satisfied by the standard set of 20 amino acids found in nature. No assumption is needed on the protein length or the number of residues in close proximity in space, in contrast to previous applications of the central limit theorem to protein energetics. We support our derivation results with computational and experimental data on mutant protein stabilities across all types of protein residues.

Statement of Significance

Defining the distribution of single mutant stability effects ($\Delta \Delta G$s) is the first step in modeling the role protein stability plays in evolution. Although empirical fits have been made to elucidate its form, a complete theoretical understanding of $\Delta \Delta G$ distributions is lacking. Here, we derive how a simple Gaussian form can arise, while still including the intricacies of protein sequence and structure. We backup our derivation with previously released computational and experimental $\Delta \Delta G$s.

Introduction

The overall stability of a protein is influenced by its residues and their interactions (1). Any change to the wildtype sequence, even a single mutation, has the potential to drastically affect stability because proteins are marginally stable (2, 3). Since mutations fix readily across evolution (4, 5), careful consideration must be made whether a certain mutation fixing in the population is the result of that mutation's stability effect. Unfolded proteins may no longer carry out their function in the cell (6) or may form toxic aggregates (7, 8). One of the first steps in modeling stability constraints requires defining the distribution of single mutant stability effects ($\Delta \Delta G$s)$^1$ (3, 9–13).

Zeldovich et al. (2007) took experimental $\Delta \Delta G$ data from different proteins deposited in the Protein Thermodynamic database (14) and found the distribution

$^1$ In this text, stating ‘$\Delta \Delta G$s’ always refers more broadly to all possible single mutant $\Delta \Delta G$s.
to be well-approximated by a Gaussian. Although limited by experimental data, Zeldovich et al. (2007)’s results seems to extend to individual proteins. Tokuriki et al. (2007) found a universal Gaussian form for all possible ΔΔGs calculated for a given protein using FoldX (16), a fast and relatively accurate computational software that requires a Protein Data Bank (PDB) (17) structure as an input. Faure and Koonin (2015) extended upon Tokuriki et al. (2007)’s study of 16 natural proteins by using FoldX to calculate ΔΔGs for all proteins with PDB structures in 5 organisms’ proteomes. They also found a universal Gaussian form for proteins of mesophilic proteomes (18).

In this paper, we apply first principles statistical mechanics to show that per residue ΔΔG distributions can by 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 energy (10, 19). Indeed, we show that our theory applies almost equally as well to those residues with fewer contacts for Tokuriki et al. (2007)’s computational data and for newly released experimental data from a study that quantified almost all ΔΔGs of a protein G domain (20).

**Derivation**

We solve for the distribution of energy effects for a given residue by first applying a continuous spin approximation to obtain the partition function of energy effects (21) and then employing an inverse Laplace transform over inverse temperature to proceed from the partition function to the density of states (22), also known as the distribution of energy effects.

**Protein energy and formalism**

The energy of the protein (E) is defined as the sum over individual contact energies,

Equation 1

\[
E = \frac{1}{2} \sum_{i,j}^N C_{ij} V_{s(i)s(j)}
\]

\[
= \frac{1}{2} \sigma^T U \sigma
\]

N is the number of amino acids in the protein. C is a N*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). V is a M*M interaction matrix, where M is the number of amino acid types (M = 20 standard amino acids), and an element of V is the pairwise energy of two specific amino acids when they are in contact. The variable s refers to the wildtype sequence; s(i) refers to the amino acid of residue (i).
σ is a NM dimensional vector, where M is the number of amino acid types. U = C ⊗ V and is a NM × NM dimensional matrix. Integrals will be simpler to perform with matrices, thus the rewrite in the second line of Equation 1. Each N block of σ 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.

Since we are interested in single mutant effects at a residue (i), we want ΔE(i) = E_mut,(i) − E_wt. First, we rewrite E in terms of the following matrix blocks,

**Equation 2**

\[
\sigma = \begin{bmatrix}
\sigma(a) \\
\sigma(i) \\
\sigma(z)
\end{bmatrix}
\]

\[
U = \begin{bmatrix}
U(a,a) & U(a,i) & U(a,z) \\
U(i,a) & 0(i,i) & U(i,z) \\
U(z,a) & U(z,i) & U(z,z)
\end{bmatrix}
\]

σ(a) corresponds to all σs numbered before residue (i); σ(z) after residue (i). U(i,i) = 0(i,i) is an 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, U is symmetric because C and V are symmetric by definition.

We commence to write the energy (Equation 1) in terms of the defined matrix blocks,

**Equation 3**

\[
E = \frac{1}{2} \left( \sigma_T(a) U(a,a) \sigma(a) + 2 \sigma_T(i) U(i,a) \sigma(i) + 2 \sigma_T(z) U(z,a) \sigma(a) + 2 \sigma_T(i) U(i,z) \sigma(z) + \sigma_T(z) U(z,z) \sigma(z) \right)
\]

Using Equation 3, ΔE(i) can be compactly expressed as,

**Equation 4**

\[
\Delta E(i) = \sigma_{\text{mut},(i)}^T U(i,a) \sigma(a) + \sigma_{\text{mut},(i)}^T U(i,z) \sigma(z) - E_{\text{wt},(i)}
\]

\[
= \delta_{\text{mut},(i)}^T U(i,a) \sigma(a) + \delta_{\text{mut},(i)}^T U(i,z) \sigma(z) - E_{\text{wt},(i)}
\]

Terms not involving (i) cancel because they are the same across mutant and wildtype. We also reduce σ(i) to only include amino acids that are non-wildtype in the second line of Equation 4; δ refers to an M-1 dimensional vector that does not include the wildtype amino acid. The same notation applies for U.

For future convenience in evaluating the partition function, let’s define δσ(i) = δ(i) - 1/(M-1). δσ(i) has the nice property of having a length of 0 when referring to an amino acid.
Equation 5

\[ \Delta E(i) = \delta \dot{\sigma}^T_{\text{mut},(i)} U_{(i,a)} \sigma(a) + \delta \dot{\sigma}^T_{\text{mut},(i)} U_{(i,z)} \sigma(z) + \frac{1}{M-1} \left( \hat{i} U_{(i,a)} \sigma(a) + \hat{i} U_{(i,z)} \sigma(z) \right) - E_{\text{wt},(i)} \]

Partition function

The partition function for one residue mutating can be approximately calculated by integrating each mutant term in \( \delta \dot{\sigma} \) over all space and picking out terms that correspond to a certain amino acid type with a Dirac delta function \( \delta(x) \).

Equation 6

\[ Z(i)(\beta) \approx \int_{-\infty}^{\infty} d^{M-1} \delta \dot{\sigma}_{\text{mut},(i)} \delta(|\dot{\sigma}_{\text{mut},(i)}|) e^{-\beta \Delta E_{\text{mut},(i)}} \]

This is the form of the continuous spin approximation that England and Shakhnovich (2003) also employed, except here we integrate over all mutant amino acids for residue \((i)\). Equation 6 is an approximation because the Dirac delta function as written does not generally correspond to the aforementioned condition that only one element of \( \dot{\sigma}(i) \) is equal to 1, while all other \( M - 2 \) elements are zero. Multiple terms can have values less than 1, such that the length of \( \dot{\sigma}(i) \) (denoted as \( |\dot{\sigma}(i)| \)) equals to 1, or equivalently \( |\delta \dot{\sigma}(i)| = 0 \). All amino acids and residues are equally affected; thus, we assume qualitative results are retained despite including those extra terms.

To make more progress, the delta function can be well-approximated by a Gaussian distribution when the variance approaches zero (23). This approximation is key to making the integral soluble and limiting terms to quadratic in \( \delta \dot{\sigma}(i) \). The variance can be expressed as,

Equation 7

\[ \text{Var} \delta \dot{\sigma}(i) = \langle \delta \dot{\sigma}^2(i) \rangle - \langle \delta \dot{\sigma}(i) \rangle^2 \]
\[ = \frac{1}{M-1} \sum_{x} \delta \dot{\sigma}^2_{(i),x} - 0 \]
\[ = \frac{1}{M-1} \left( 1 - \frac{1}{M-1} \right) \equiv \frac{1}{M'} \]

Thus, our Gaussian approximated delta function leads Equation 6 to look like,
**Equation 8**

\[ Z^{(i)}(\beta) \approx \int_{-\infty}^{\infty} d^{M-1}\delta\dot{\sigma}_{\text{mut},(i)} e^{-M'\delta^2\dot{\sigma}_{\text{mut},(i)}^2/2} e^{-\beta \Delta E_{\text{mut},(i)}} \]

We can now begin solving the integral. First, we insert the expression for \( \Delta E^{(i)} \) (Equation 5) into Equation 8 and pull out constants in front of the integral. Then we complete the square of the exponent in the second line.

**Equation 9**

\[
\begin{align*}
Z^{(i)}(\beta) &= \exp \left[ \beta \left( \frac{1}{M - 1} \left( i \mathbf{U}_{(i,a)} \mathbf{\sigma}(a) + i \mathbf{U}_{(i,z)} \mathbf{\sigma}(z) \right) \right) \\
&\quad + E_{\text{wt},(i)} \right] \int_{-\infty}^{\infty} d^{M-1}\delta\dot{\sigma}_{\text{mut},(i)} e^{-\frac{M'\delta^2\dot{\sigma}_{\text{mut},(i)}^2}{2}} e^{-\beta \left( \delta\dot{\sigma}_{\text{mut},(i)}^T \mathbf{U}_{(i,a)} \mathbf{\sigma}(a) + \delta\dot{\sigma}_{\text{mut},(i)}^T \mathbf{U}_{(i,z)} \mathbf{\sigma}(z) \right)} \\
&= \exp \left[ \beta \left( \frac{1}{M - 1} \left( i \mathbf{U}_{(i,a)} \mathbf{\sigma}(a) + i \mathbf{U}_{(i,z)} \mathbf{\sigma}(z) \right) \right) \\
&\quad + E_{\text{wt},(i)} \right] \exp \left[ \frac{1}{2M'} \left( \beta \mathbf{U}_{(i,a)} \mathbf{\sigma}(a) + \mathbf{U}_{(i,z)} \mathbf{\sigma}(z) \right)^2 \right] \int_{-\infty}^{\infty} d^{M-1}\delta\dot{\sigma}_{\text{mut},(i)} e^{-\frac{M'\delta^2\dot{\sigma}_{\text{mut},(i)}^2}{2} \left( \frac{\beta}{M'} \mathbf{U}_{(i,a)} \mathbf{\sigma}(a) + \mathbf{U}_{(i,z)} \mathbf{\sigma}(z) \right)^2} \\
\end{align*}
\]

The multidimensional integral represents \( M - 1 \) identical Gaussian integrals and can be easily solved to be a constant independent of \( \beta \).

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 10**

\[
\begin{align*}
z^{(i)}(\beta) &= \frac{Z^{(i)}(\beta)}{Z^{(i)}(0)} \\
&= \exp \left[ \beta \left( \frac{1}{M - 1} \left( i \mathbf{U}_{(i,a)} \mathbf{\sigma}(a) + i \mathbf{U}_{(i,z)} \mathbf{\sigma}(z) \right) \right) \\
&\quad + E_{\text{wt},(i)} \right] \exp \left[ \frac{1}{2M'} \left( \beta \left( \mathbf{U}_{(i,a)} \mathbf{\sigma}(a) + \mathbf{U}_{(i,z)} \mathbf{\sigma}(z) \right) \right)^2 \right]
\end{align*}
\]


\[
\exp \left( -\frac{1}{M-1} \sum_{i} \sum_{j \neq s(i)} C_{ij} V_{m,s(j)} + E_{wt,(i)} \right) \exp \left( \frac{\beta^2}{2M'} \sum_{j,k} \sum_{m \neq s(i)} C_{ij} V_{m,s(j)} C_{ik} V_{m,s(k)} \right)
\]

**Density of states**

The density of states \(n\) is related to the partition function by an inverse Laplace transform (22).

**Equation 11**

\[
n_{(i)}(\Delta E) = \frac{1}{2\pi} \int_{-\infty}^{\infty} d\beta \, e^{i\beta \Delta E} Z_{(i)}(i\beta)
\]

\[
= \frac{1}{2\pi} \int_{-\infty}^{\infty} d\beta \exp \left[ \frac{\beta^2}{2M'} \sum_{j,k} \sum_{m \neq s(i)} C_{ij} V_{m,s(j)} C_{ik} V_{m,s(k)} \right] \exp \left[ \beta \left( \Delta E - \sum_{m \neq s(i)} C_{ij} V_{m,s(j)} + E_{wt,(i)} \right) \right]
\]

The integral can be solved by completing the square. Skipping steps, it can be shown that the following Gaussian distribution (G(mean, variance)) is exactly obtained.

**Equation 12**

\[
n_{(i)}(\Delta E) = G \left( \frac{1}{M-1} \sum_{j} \sum_{m \neq s(i)} C_{ij} V_{m,s(j)} - \sum_{j} C_{ij} V_{s(i)s(j)}, \frac{1}{M'} \sum_{j,k} \sum_{m \neq s(i)} C_{ij} V_{ms(j)} C_{ik} V_{ms(k)} \right)
\]

The expressions for the mean and variance of the Gaussian in Equation 12 are intuitive. The mean is \(\langle \Delta E_{(i)} \rangle\) and the variance is \(\langle \Delta E_{(i)}^2 \rangle\) if \(M' = M - 1\). This is what we would expect if we were to fit a Gaussian to data whose energy is determined by Equation 1, except that the variance would be \(\langle \Delta E_{(i)}^2 \rangle - \langle \Delta E_{(i)} \rangle^2\). Some inaccuracy is reasonable because we employed a continuous spin approximation (Equation 6) and approximated the delta function as a Gaussian (Equation 8).

Our derivation found a distribution for \(\Delta E\), not \(\Delta G\). We assume that the unfolded state and entropic effects are less important and can be accounted for by properly fitting the interaction matrix, thus not changing the final form of the distribution. In other words, we assume \(n_{(i)}(\Delta E) \sim n_{(i)}(\Delta G)\).
Bioinformatics

Computational ΔΔGs

To test whether per residue ΔΔG distributions are Gaussian (Equation 12), we first organize ΔΔGs from Tokuriki et al. (2007) into groups of 19 ΔΔGs belonging to individual residues to obtain per residue ΔΔG distributions (Materials and Methods). We then run the Shapiro-Wilk test (24) to analyze how well a Gaussian distribution fits the 19 ΔΔGs, 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 (Materials and Methods). For example, in Figure S1 we run the Shapiro-Wilk test on the first residue 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 ΔΔGs pertaining to the first residue of DHFR is not Gaussian distributed, at odds with our prediction (further discussed below).

Running the Shapiro-Wilk test for all residues from the 16 natural proteins in Tokuriki et al. (2007), we find that a majority of residues have ΔΔG distributions indistinguishable from a Gaussian (Figure 1). We further investigate the 10-25% of residues that do reject the null hypothesis at a p-value threshold of 0.05, like the first residue of DHFR discussed above. In Figure 2, we show box plots of contacts for residues that reject the null and for all residues. For 11 proteins, the median number of contacts for residues that reject the null is less than the median number of contacts for all residues. For five proteins, the medians are equivalent; and for one protein, Ribonuclease A (RNAA), the median number of contacts for residues that reject the null is greater than the median number of contacts for all residues. However, a more detailed analysis comparing the distributions of the number of contacts using the Kolmogorov-Smirnov test (25), yields only human carbonic anhydrase II (CAII) to have significantly different contact distributions between those residues that reject the null and all residues (p-value = 0.037). The lack of a structural signature of those residues that reject the null makes sense in light of our theory, which does not require an assumptions on the number of contacts.

With a more stringent p-value threshold of 0.001, no protein rejects the null hypothesis for more than 4% of its residues (hatched bar plots in Figure 1). We can argue that a more stringent p-value threshold is necessary to address the multiple testing problem (26).
Figure 1: Relatively few residues per protein fail to match our theoretical prediction that per residue \( \Delta \Delta G \)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 1 in Materials and Methods.

Figure 2: The distribution of contacts for residues found to reject the null hypothesis in Figure 1 at a p-value threshold of 0.05 (solid box plots) is similar to that of all protein residues (hatched box plots). Horizontal black lines correspond to medians; diamonds correspond to outliers.

Experimental \( \Delta \Delta G \)s

A major caveat of the previous subsection is that \( \Delta \Delta G \)s are generated from FoldX. As noted in the Introduction, exhaustive studies of protein \( \Delta \Delta G \) distributions have been limited to computational calculations because the number of experimental measurements is too large to complete in a reasonable timeframe. Recently, Nisthal (2012) made exhaustive \( \Delta \Delta G \) measurements feasible by developing an experimental
protocol utilizing automated robotics for protein purification and chemical denaturant curve measurements. With this protocol, Nisthal et al. (2019) quantitatively measured the ΔΔGs of almost all single mutants (78%) of the β1 domain of Streptococcal protein G (Gβ1).

We find that per residue ΔΔG measurements are mostly Gaussian distributed, but at a ratio on the high end of those found in Figure 1. Including all Gβ1 residues with at least three measured ΔΔGs (54 out of 56 total residues), 13 (24%) residues reject the null hypothesis at a p-value threshold of 0.05; 6 (11%), at a p-value of 0.001 (Figure 3). We tested whether having few ΔΔGs measured for some residues explained the relatively high percentages of null hypotheses rejected, however, this was not the case. When only analyzing residues with the maximal number of 17 ΔΔGs measured (29 out of 56 total residues), 11 residues still reject the null at a p-value threshold of 0.05; 5, at a p-value of 0.001 (Figure 3). The contact distribution of those 11 residues that reject the null does not matter either. We find that the contact distribution is indistinguishable with those of all residues according to the Kolmogorov-Smirnov test (p-value = 0.40).

Figure 3: Relatively few residues of β1 domain of Streptococcal protein G (56 total residues) fail to match our theoretical prediction that per residue ΔΔGs 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. Residues that reject the null mostly have the maximal number of mutants measured in the study (right two-most bar plots).

**Discussion**

Employing a previously established theoretical framework (21), we derive from first principles statistical mechanics that per residue $\Delta \Delta G$s are Gaussian distributed. Our theory yields a Gaussian while still considering the intricacies of the protein structure and unique interactions of amino acid types. We verify with computational and experimental data that per residue $\Delta \Delta G$s are Gaussian distributed for most residues (Figure 1 and Figure 3).

The finding that per residue $\Delta \Delta G$s are Gaussian distributed may not seem surprising in light of the central limit theorem. However, the central limit theorem applies not because individual contact energies can be treated as random variables, but because each residue has 19 possible mutant amino acid identities (Equation 8). This yields Gaussian distributions for all residues, regardless of its number of native contacts (Figure 2).

The mathematical framework we employed closely follows that of England and Shakhnovich (2003). They derived from first principles an expression for the number of sequences that folds into a given structure, called protein designability. Unlike here where we integrate over all mutant amino acid types for a given residue (Equation 6), England and Shakhnovich (2003) integrated over all amino acid types across all residues to obtain the partition function. Interestingly, when applying an inverse Laplace transform to obtain the density of states, England and Shakhnovich (2003) cannot exactly obtain a Gaussian. Within the theory, a Gaussian distribution is only exactly obtained for per residue mutants. However, when considering all possible mutants at once (not just single mutants), a Gaussian distribution can only be approximated.

Our derivation falls short from generally showing that the overall $\Delta E$ distribution of all single mutants ($n(\Delta E)$) is Gaussian. By definition, $n(\Delta E)$ is the weighted sum of all $N$ Gaussians from Equation 12 because it corresponds to sequentially picking a single mutant out of $N$ residues, i.e. a statistical OR operation.

\[ n(\Delta E) = \frac{1}{N} \sum_{i}^{N} n(i)(\Delta E) \]
We are not aware of an approach using first principles to reduce the mixture of N Gaussians in Equation 13. However, we can bioinformatically demonstrate that our N Gaussian model reduces to a bi-Gaussian (Figure S2 and Figure S3). Therefore, a more tractable Gaussian mixture model with fewer Gaussians than the exact N Gaussian model elucidated here can be employed with little loss in accuracy.

Future studies can use Equation 12 to fit a 20*20 interaction matrix to yield accurate means and variances of per residue $\Delta \Delta G$ distributions for experimental data. A 20*20 interaction matrix is too simple to generally capture protein folding (28), however for single mutant $\Delta \Delta G$s that level of detail may suffice.

**Acknowledgements**

We are grateful to Dan Tawfik and Alex Nisthal for fruitful email exchanges. We would also like to thank Bharat Adkar and Mobolaji Williams for useful comments on this manuscript.

This work is supported by the National Institutes of Health (grant number GM067680).

**Author Contributions**

RMR and EIS designed research, analyzed data, wrote the manuscript. RMR performed the research.

**References**


---

Notice that we are summing normal distributions and not random variables that are normally distributed. The sum of two normally distributed random variables is normally distributed; however, it does not generally hold that the sum of two normal distributions is another normal distribution (30).


Materials and Methods

The computational data for the 16 natural proteins studied in Tokuriki et al. (2007) was kindly provided by Dan Tawfik. The full names of the 16 proteins and the
acronyms used in this text, along with their lengths and endogenous organisms can be found in Table 1. The lone protein presented with experimental $\Delta \Delta G$ data, the $\beta_1$ domain of protein G, is also included at the end of Table 1. The experimental data was kindly provided by Alex Nisthal. We adopt the convention that a more negative $\Delta G$ equates to a more stable protein, thus $\Delta \Delta G = \Delta G^{\text{mut}} - \Delta G^{\text{wt}} < 0$ is a stabilizing mutation.

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Acronym</th>
<th>PDB</th>
<th>Length</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recombinant serum paraoxonase</td>
<td>PON</td>
<td>1v04</td>
<td>332</td>
<td>Oryctolagus cuniculus</td>
</tr>
<tr>
<td>Lipase</td>
<td>Lip</td>
<td>1ex9</td>
<td>285</td>
<td>Pseudomonas aeruginosa</td>
</tr>
<tr>
<td>$\beta$-Lactamase</td>
<td>TEM1</td>
<td>1btl</td>
<td>263</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>Human carbonic anhydrase II</td>
<td>CAII</td>
<td>1lug</td>
<td>259</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Dihydrofolate reductase</td>
<td>DHFR</td>
<td>1rx2</td>
<td>159</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>Ribonuclease H</td>
<td>RNaH</td>
<td>2rn2</td>
<td>155</td>
<td>Escherichia coli</td>
</tr>
<tr>
<td>Myoglobin</td>
<td>Myo</td>
<td>1a6k</td>
<td>151</td>
<td>Physeter macrocephalus</td>
</tr>
<tr>
<td>Staphylococcus nuclease</td>
<td>SNase</td>
<td>1stn</td>
<td>136</td>
<td>Staphylococcus aureus</td>
</tr>
<tr>
<td>Human lysozyme</td>
<td>huLys</td>
<td>1rex</td>
<td>130</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Hen lysozyme</td>
<td>henLys</td>
<td>1dpx</td>
<td>129</td>
<td>Gallus gallus</td>
</tr>
<tr>
<td>Ribonuclease A</td>
<td>RNaA</td>
<td>1fs3</td>
<td>124</td>
<td>Bos taurus</td>
</tr>
<tr>
<td>Barnase</td>
<td>Barn</td>
<td>1a2p</td>
<td>108</td>
<td>Bacillus amyloliquefaciens</td>
</tr>
<tr>
<td>Acylphosphatase</td>
<td>AcP</td>
<td>2acy</td>
<td>98</td>
<td>Bos taurus</td>
</tr>
<tr>
<td>Ubiquitin</td>
<td>Ubiq</td>
<td>1ubq</td>
<td>76</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>Protein G, IgG-binding domain</td>
<td>proG</td>
<td>2igd</td>
<td>61</td>
<td>Streptococcus sp. group G</td>
</tr>
</tbody>
</table>
Table 1: Proteins whose ΔΔGs are studied. All protein ΔΔGs, except Gβ1, originate from Tokuriki et al. (2007). Gβ1 ΔΔGs were experimentally measured by Nisthal et al. (2019). Not all of Gβ1 ΔΔGs were measured because of experimental constraints.

We employ the Shapiro-Wilk (S-W) test to determine whether per residue ΔΔGs are Gaussian distributed. S-W can only be applied for normality testing and has been shown to be the most powerful in this regards (29). If the test-statistic, which ranges from 0 to 1, approaches zero and has an appropriately minuscule p-value, then we can reject the null that our data fits a Gaussian distribution.

We calculated the number of contacts for each residue from in-house scripts analyzing corresponding PDB structures (Table 1). A two-sample Kolmogorov-Smirnov (K-S) test is employed to determine whether our contact distributions between residues that reject the null and all residues differ.

For Figure S2 and Figure S3, the K-S test is employed to determine whether our N-Gaussian model or a bi-Gaussian fit, matches the distribution from the ΔΔG data. We carry out a two-sample K-S test to compare binned data to each other. All binning is performed exactly as done in Tokuriki et al. (2007); ΔΔGs are parsed in 1 kcal/mol bins from -10 kcal/mol to 15 kcal/mol. Calculated FoldX ΔΔGs beyond those limits are set to the corresponding binning limit.

Scripts used to analyze data from Tokuriki et al. (2007) and Nisthal et al. (2019) can be found on GitHub (github.com/rrazban/ddG_distribution).