PNAS
www.pnas.org

Supplementary Information for

Size and topology modulate the effects of frustration in protein folding

Kluber, Alex; Burt, Timothy; Clementi, Cecilia

Corresponding author: Cecilia Clementi
E-mail: cecilia@rice.edu

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.1801406115

Supporting Information Text

Simulation Model

We treat the effects of frustration by adding non-native interactions to a one-bead per residue (C_α) structure-based model. The potential energy of our model is composed by four terms:

$$E = E_{\text{back}} + E_{\text{nat}} + E_{\text{ex}} + E_{\text{non}}\,, \tag{1}$$

where E_{back}, E_{nat}, and E_{ex} are "structure-based" terms constructed to make the native conformation the global energy minimum and E_{non} incorporates frustration by including interactions between non-native pairs of residues (1). The backbone potential E_{back} biases towards native backbone geometry,

$$E_{\text{back}}(\vec{x}) = \sum_{\text{bonds}} \frac{k_b}{2}(r - r_0)^2 \; + \sum_{\text{angles}} \frac{k_\theta}{2}(\theta - \theta_0)^2 \tag{2}$$

$$+ \sum_{\text{dihedrals}} \frac{k_\phi}{2}\left[\cos\left(\phi - \phi_0\right) + 1\right] \; + \frac{k_\phi}{4}\left[\cos\left(3(\phi - \phi_0)\right) + 1\right]\,, \tag{3}$$

where subscript "0" indicates native state values of bond distances r_0, bond angles θ_0, and dihedral angles ϕ_0. The values of the parameters k_b, k_θ, and k_ϕ are as previously reported (2).

Long-range interactions are assigned to residue pairs depending on whether they are defined as "native", "non-native", or "neutral" pairs. Native pairs are residues in contact in the corresponding Protein Data Bank (PDB) structure, where "in contact" is defined as having at least one all-atom contact within a 6Å cutoff as assigned by the Shadow map algorithm (3).

The native contact potential E_{nat} places an attractive Gaussian well at the native distance:

$$E_{\text{nat}} = \sum_{ij} \left(\frac{r_{\text{ex}}}{r_{ij}}\right)^{12} \left[G_{ij}(r_{ij}) - 1\right] + \epsilon_{\text{nat}} G_{ij}(r_{ij})\,, \tag{4}$$

where G_{ij} is a Gaussian,

$$G_{ij}(r_{ij}) = \exp\left(\frac{-(r_{ij} - r_{ij}^0)^2}{2w_{\text{nat}}^2}\right)\,. \tag{5}$$

By this construction each native interaction is centered at its native contact distance $r_{ij} = r_{ij}^0$, has a depth of $\epsilon_{\text{nat}} = -1$, a width of $w_{\text{nat}} = 0.5$Å and an excluded volume radius $r_{\text{ex}} = 4$Å.

Neutral pairs are residues that have C_α's within 8Å in the native state but don't form native contacts (i.e. they are "near-native"). Neutral pairs interact only through an excluded volume interaction,

$$E_{\text{ex}} = \sum_{ij} \left(\frac{r_{\text{ex}}}{r_{ij}}\right)^{12}\,. \tag{6}$$

We define these residues as neutral, and not non-native, to prevent non-native interactions from changing the properties of the native state.

The remaining residue pairs are non-native interactions that can be attractive or repulsive:

$$E_{\text{non}} = \sum_{ij} E_{ij}^{\text{non}}(r_{ij})\,. \tag{7}$$

Each non-native pair is assigned an interaction strength $\epsilon_{ij}^{\text{non}}$ that is chosen from a zero-mean Gaussian distribution with standard deviation b: $\mathcal{N}(0, b)$. If $\epsilon_{ij}^{\text{non}} < 0$ then the interaction is attractive and takes the form:

$$E_{ij}^{\text{non}} = \left(\frac{r_{\text{ex}}}{r_{ij}}\right)^{12} \left[G_{ij}(r_{ij}) - 1\right] + \epsilon_{ij}^{\text{non}} G_{ij}(r_{ij})\,, \tag{8}$$

where non-native Gaussian have their minimum located at $r_{\text{non}} = 5$Å and width $w_{\text{non}} = 0.75$Å.

If $\epsilon_{ij}^{\text{non}} > 0$ then the residue pair is assigned a repulsive interaction:

$$E_{ij}^{\text{non}} = \left(\frac{r_{\text{ex}}}{r_{ij}}\right)^{12} + \frac{\epsilon_{ij}^{\text{non}}}{2}\left[\tanh\left(-\frac{(r_{ij} - r_{\text{non}})}{w_{\text{non}}}\right) + 1\right]\,. \tag{9}$$

Note that all non-native interactions are at the same distance $r_{\text{non}} = 5$Å; there is no geometric or steric information encoded in non-native interactions. We have assigned the non-native interaction width $w_{\text{non}} = 0.75$Å as non-native contacts are meant to be less specific. This choice for w_{non} also prevents artifacts for small $|\epsilon_{ij}^{\text{non}}|$ where a small oscillation in the potential is observed if smaller values of w_{non} are used.

Fraction of native contacts, radius of gyration, and degree of collapse

The reaction coordinate Q is defined as the fraction of native contacts:

$$Q = \frac{1}{N_{\text{nat}}} \sum_{ij}^{N_{\text{nat}}} \frac{1}{2} \left[\tanh \left(\frac{-(r_{ij} - r_{ij}^0 - r_c)}{w_{\text{nat}}} \right) + 1 \right] , \quad [10]$$

where N_{nat} is the total number of native contacts and w_{nat} is the width of the interaction energy well for native contacts, as defined above, and $r_c = 1$Å.

The radius of gyration R_g is calculated as the root-mean-squared deviation of bead positions $\vec{r}_i$ from the average position $\langle \vec{r} \rangle$, using MDTraj (4):

$$R_g = \left\langle \sum_i^N \frac{(\vec{r}_i - \langle \vec{r} \rangle)^2}{N} \right\rangle^{\frac{1}{2}} . \quad [11]$$

The degree of collapse η is calculated by normalizing the radius of gyration by minimum and maximum values:

$$\eta = \frac{R_g - R_g^{\max}}{R_g^{\min} - R_g^{\max}} . \quad [12]$$

We take $R_g^{\max}$ to be the maximum radius of gyration for the unfrustrated ($b = 0$) chain $R_g^{b=0}$. The smallest possible radius of gyration $R_g^{\min}$ corresponds to the chain being compacted into a tight ball. Consider the chain packed into some minimum volume $V_{\min} = \frac{4}{3}\pi R_{\min}^3$, if this was treated as a solid sphere its radius of gyration would be:

$$(R_g^{\min})^2 = \frac{\int_0^{R_{\min}} (R^2) R^2 dR}{\int_0^{R_{\min}} R^2 dR} = \frac{\frac{1}{5} R_{\min}^5}{\frac{1}{3} R_{\min}^3} = \frac{3}{5} R_{\min}^2 . \quad [13]$$

As a rough approximation, the minimum volume $V_{\min}$ is proportional to the volume of a monomer $v_0 = \frac{4\pi}{3} r_0^3$,

$$V_{\min} = \frac{4\pi}{3} r_0^3 N , \quad [14]$$

where r_0 is some effective radius of the monomers. Therefore the minimum radius of gyration is:

$$R_g^{\min} = \sqrt{\frac{3}{5}} r_0 N^{\frac{1}{3}} . \quad [15]$$

We take $r_0 = 3$Å as a rough estimate for the effective radius of the monomers, because it is about half of the non-native contact distance. Changing r_0 by $20 - 30\%$ does not change the qualitative interpretation of η. The important feature of Eq.15 is that it captures the proper scaling of $R_g^{\min}$ with N.

Inherent structure analysis

The global topography of the energy landscape in protein folding resembles a funnel towards the native state. The native energy E_{nat} measures the progress down the funnel and each stratum of the funnel has roughness coming from fluctuations in the non-native interactions ΔE_{non}. We use "inherent structure" (IS) analysis (5) to estimate the roughness of the energy landscape ΔE_{non}, by inspecting the distribution of energy minima in the unfolded state.

The idea of IS analysis is that the energy landscape can be partitioned into basins that surround each energy minimum, each minimum corresponding to an "inherent structure" on the landscape. A trajectory sampled at temperature T can be thought of as quickly fluctuating within and between energy basins.

We apply gradient descent energy minimization in GROMACS (6) to structures originally sampled at the folding temperature $T = T_f$. This maps each sampled structure to the nearest energy minimum (or "inherent structure")

on the landscape, eliminating thermal fluctuations. In general, the probability of observing an energy minimum E during a trajectory at temperature T is (5):

$$P(E) = \frac{\Omega(E)e^{-\beta E - \beta f_{\text{vib}}(E)}}{Z} = \frac{e^{-\beta(E - TS + f_{\text{vib}}(E))}}{Z}\,, \quad [16]$$

where $\beta = \frac{1}{k_B T}$, k_B is the Boltzmann constant, $\Omega(E)$ is the density of states, $S(E) = k_B \ln \Omega(E)$ is the microcanonical entropy, Z is the partition function, and $f_{\text{vib}}(E)$ is a vibrational free energy that depends on the shape of the basin around the minimum. Here we have made the assumption that the vibrational free energy is primarily a function of energy $f_{\text{vib}} = f_{\text{vib}}(E)$, which has been found to be the case for structure-based models (7, 8).

The funneled shape of the energy landscape is reflected in the competition between E and $S(E)$ which both take their smallest value in the native state and increase when moving up the funnel to the unfolded state. We are primarily interested in what the landscape looks like in the unfolded state: we want to know the fluctuations in energy ΔE_{non}.

In order to identify the unfolded state we split the total energy into native and non-native terms, $E = E_{\text{nat}} + E_{\text{non}}$. We define the native (N) and unfolded (U) states as bins around the peaks in $P(E_{\text{nat}})$, located at E^N_{nat} and E^U_{nat}, respectively. The width of each state goes until half the maximum of the peak. Non-native interactions are negligible in the native state, by design of our model, but play a very important role in the unfolded state ($E_{\text{nat}} = E^U_{\text{nat}}$). In particular, energy landscape theory posits that non-native interactions decrease the microcanonical entropy of the unfolded state by (9),

$$S(E^U_{\text{nat}}, E_{\text{non}}) = S_0(E^U_{\text{nat}}) - \frac{(E_{\text{non}} - \overline{E}(E^U_{\text{nat}}))^2}{2\Delta E^2_{\text{non}}}\,, \quad [17]$$

where S_0 and $\overline{E}$ are the entropy and average energy of the chain without non-native interactions, respectively. Crucially, the frustration ΔE_{non} sets the slope of the parabola expressed by Eq.17.

Estimating ΔE_{non} requires estimating S. Unfortunately, the probability of observing a minimum on the landscape,

$$P(E_{\text{nat}}, E_{\text{non}}) = \frac{1}{Z}\exp(-\beta(E_{\text{nat}} + E_{\text{non}}) - TS(E_{\text{nat}}, E_{\text{non}}) + f_{\text{vib}}(E_{\text{nat}}, E_{\text{non}}))\,, \quad [18]$$

cannot be inverted for S directly, because we don't know Z or f_{vib}.

However, we show these issues can be overcome by using relative probabilities. In particular, we follow refs (7, 8), by considering probabilities relative to the native state bin. Then Eq.18 becomes,

$$\ln\left(\frac{P(E^U_{\text{nat}}, E_{\text{non}})}{P(E^N_{\text{nat}}, E^N_{\text{non}})}\right) = \frac{\Delta S}{k_B} - \beta\Delta E - \beta\Delta f_{\text{vib}}\,, \quad [19]$$

where Δ indicates the difference between unfolded and native state, that is $\Delta X = X(E^U_{\text{nat}}, E_{\text{non}}) - X(E^N_{\text{nat}}, E^N_{\text{non}})$. In Eq.19, the non-native energy is evaluated in the reference probability $P(E^N_{\text{nat}}, E^N_{\text{non}})$, but allowed to vary in the unfolded state $P(E^U_{\text{nat}}, E_{\text{non}})$. Thus Eq.19 is a function of only one variable: the non-native energy E_{non} in the unfolded state. The native state is a natural choice for the reference probability in our case because it is unique in our model ($S(E^N_{\text{nat}}, E^N_{\text{non}}) = 0$) and well sampled in our simulation.

Therefore, dropping the Δ for S we get a relationship for microcanonical entropy as a function of E_{non} only (analogous to Eq.17):

$$\frac{S}{k_B} = \ln\left(\frac{P(E^U_{\text{nat}}, E_{\text{non}})}{P(E^N_{\text{nat}}, E^N_{\text{non}})}\right) + \beta\Delta E + \beta\Delta f_{\text{vib}}\,. \quad [20]$$

If we represent the right-hand side as a second-order function,

$$\frac{S}{k_B} = aE^2_{\text{non}} + bE_{\text{non}} + c\,, \quad [21]$$

we can make a correspondence with the coefficients in Eq.17 to solve for ΔE_{non}, S_0, and $\overline{E}$. We can then calculate the "glass temperature" T_g from these parameters as:

$$T_g = \frac{\Delta E_{\text{non}}}{\sqrt{2k_B S_0}}\,. \quad [22]$$

In our analysis, we neglect Δf_{vib}. Assuming Δf_{vib} does not depend on E_{non}, then it would only affect the constant term in 21, and consequently S_0, but not our determination of ΔE_{non}.

Correlations between crossover heterogeneity and structural metrics

The size and topology effects we observed are reflected in the correlation of absolute contact order (ACO) with important levels of non-native heterogeneity we have indicated with b^*. Absolute contact order (ACO) is defined as the average sequence separation between native contacts,

$$\mathrm{ACO} = \frac{1}{N_{\mathrm{nat}}} \sum_{ij}^{N_{\mathrm{nat}}} l_{ij} \qquad [23]$$

where $l_{ij} = |i - j|$ is the sequence separation between residues that make a native contact and N_{nat} is the number of native contacts. Fig.S3 shows that b^* increases with ACO, appearing to level off when ACO > 25. Since ACO correlates with size this indicates the crossover into a large-size limit where b^* no longer depends on size. Notably, b^* is less correlated with size (N) or relative contact order ($RCO = ACO/N$) as shown in Table S1 and Fig.S3.

Fluctuations in folding time across parameter sets

In this work, we have presented average structural and kinetic quantities from many realizations of non-native parameters. Thus we are able to discern trends that do not depend on any particular parameter set, but only on the statistical properties of the non-native parameters: their mean $\overline{\epsilon_{ij}^{\mathrm{non}}} = 0$ and standard deviation $\sigma_{\epsilon_{ij}^{\mathrm{non}}} = b$. Fig.S4 shows that as non-native heterogeneity is increased the fluctuations of the folding time τ_f and free energy profile $G(Q)$ tend to increase. Fig.S4B shows that the fluctuations in folding time relative to the parameter-mean $\sigma_{\tau_f}/\overline{\tau_f}$ increase sharply when $b > 1$. Large variations between parameter sets indicate that an observable depends on more than just the overall statistical properties of the parameters and also depends on how the non-native interaction parameters are assigned. This means, for example, that particularly attractive or repulsive non-native interactions, or clusters of such interactions, have a large influence on the free energy barrier height (see Fig.S4A). As a result, the variations in folding time across parameter sets increases with the heterogeneity of non-native interactions (see Fig.S4B).

References

1. Clementi C, Plotkin SS (2004) The effects of nonnative interactions on protein folding rates: theory and simulation. *Protein Sci.* 13(7):1750–1766.
2. Clementi C, Nymeyer H, Onuchic JN (2000) Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298(5):937–53.
3. Noel JK, Whitford PC, Onuchic JN (2012) The shadow map: a general contact definition for capturing the dynamics of biomolecular folding and function. *J. Phys. Chem. B* 116(29):8692–702.
4. McGibbon RT, et al. (2015) MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* 109(8):1528–1532.
5. Stillinger FH, Weber TA (1982) Hidden structure in liquids. *Phys. Rev. A* 25(2):978–989.
6. Van Der Spoel D, et al. (2005) GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26(16):1701–1718.
7. Nakagawa N, Peyrard M (2006) The inherent structure landscape of a protein. *Proc Natl Acad Sci* 103(14):5279–84.
8. Ming D, Anghel M, Wall ME (2008) Hidden structure in protein energy landscapes. *Phys. Rev. E* 77(2):021902.
9. Bryngelson JD, Wolynes PG (1987) Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci* 84:7524–7528.

ρ	b^*_D	b^*_η	b^*_E
N	0.6156	0.5717	0.5815
ACO	0.8750	0.7319	0.7340
RCO	0.3688	0.2371	0.2903
ρ_{rk}	rank(b^*_D)	rank(b^*_η)	rank(b^*_E)
rank(N)	0.6727	0.5636	0.4857
rank(ACO)	0.9636	0.8061	0.9429
rank(RCO)	0.4182	0.3091	0.1429

Table S1. Pearson (ρ) and Spearman rank (ρ_{rk}) correlation coefficients between the crossover heterogeneity $b*$ and the following structural metrics: size N, absolute contact order (ACO), and relative contact order (RCO).

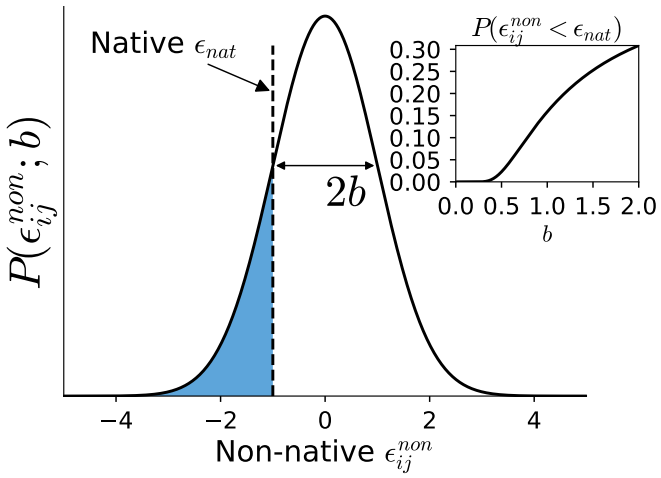

Fig. S1. Distribution of non-native parameters $\epsilon_{ij}^{\text{non}}$ is Gaussian with standard deviation b. The shaded region indicates the probability of a non-native interaction being more energetically stabilizing than a native contact which has strength ϵ. (inset) As b increases, the fraction of non-native interactions that are more attractive than native contacts increases.

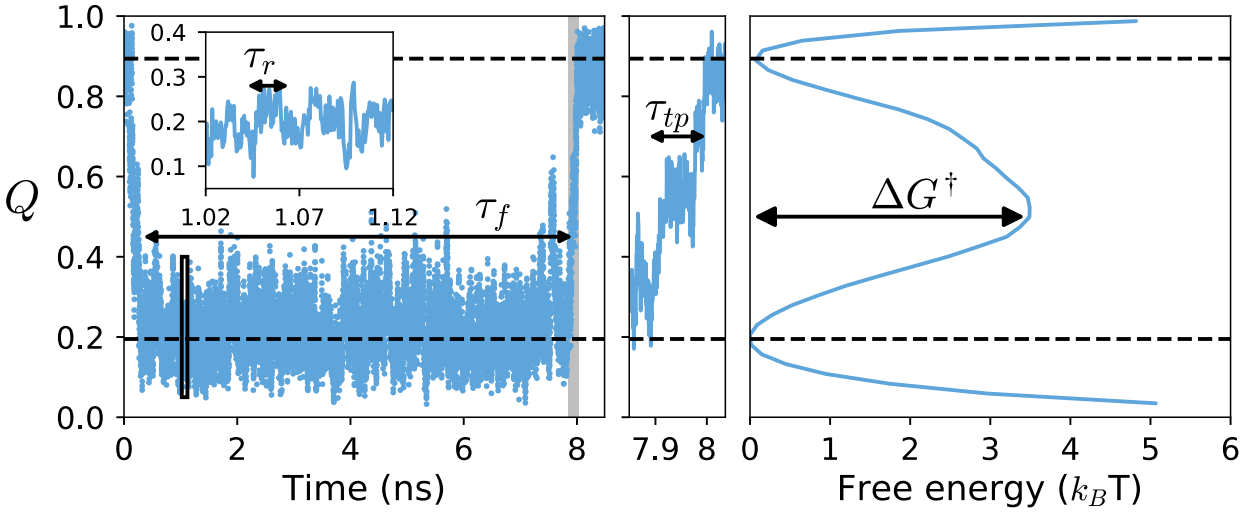

Fig. S2. Folding time τ_f, transition path time τ_{tp}, and reconfiguration time τ_r, are shown for illustration on a trajectory segment (left panel). Dashed lines indicate the (un)folded states along the free energy profile on the right panel. Note that coarse-grain time units do not correspond exactly to real time units, so the different timescales should be considered relative to each other and not for their absolute values.

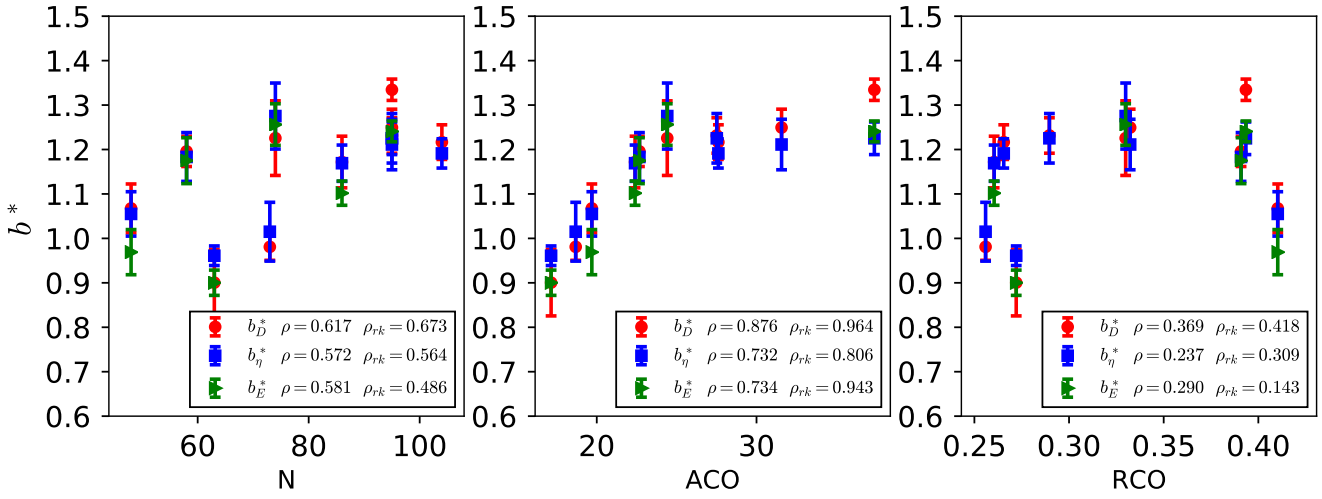

Fig. S3. Correlation between the crossover heterogeneity b^* and different protein-dependent quantities: size (left), Absolute Contact Order (middle), and Relative Contact Order (right). Legend shows Pearson correlation coefficient ρ and Spearman rank correlation coefficient ρ_{rk}.

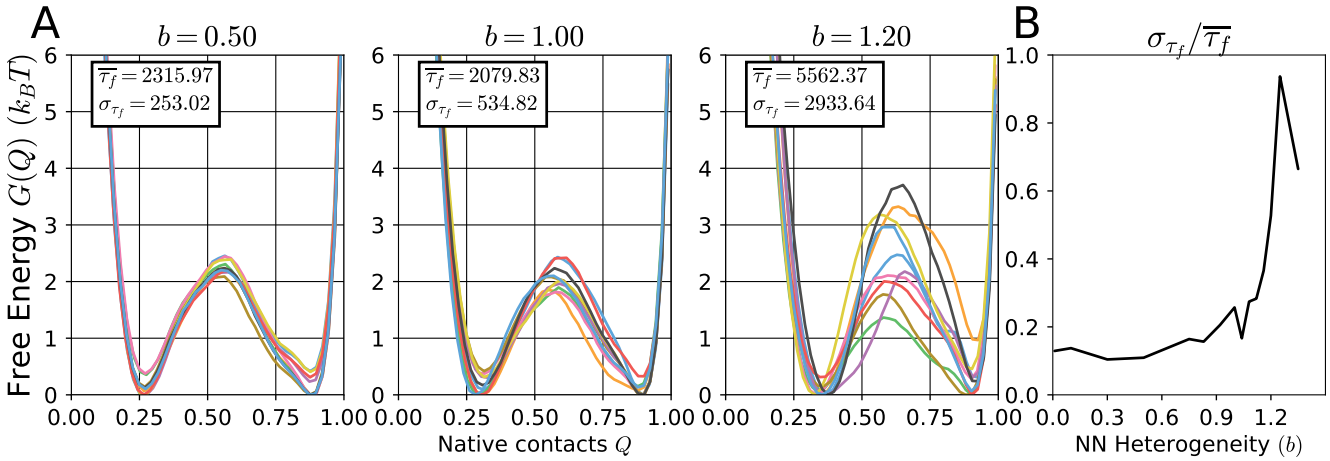

Fig. S4. (A) Free energy profiles for protein 1imq at three different values of non-native heterogeneity b. Colors indicate different sets of non-native interaction parameters. The average and standard-deviation of the folding time over all parameter sets at a given b are indicated as $\overline{\tau_f}$ and σ_{τ_f}, respectively, on each panel. (B) The relative size of the fluctuations in folding time of parameter sets versus non-native heterogeneity.

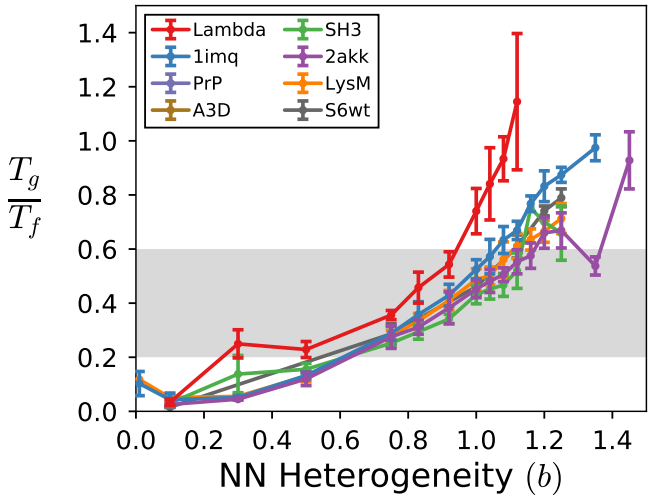

Fig. S5. The glass temperature compared to the folding temperature $\frac{T_g}{T_f}$ as a function of b. Grey rectangle indicates the theoretical range from other studies: $T_g/T_f = 0.2 - 0.6$