# Supporting Information for "Microscopic events in $\beta$-hairpin folding from alternative unfolded ensembles"

Robert B. Best,[†] and Jeetain Mittal[‡]

*Cambridge University, Department of Chemistry, Lensfield Road, Cambridge CB2 1EW, and Lehigh University, Department of Chemical Engineering, Bethlehem, PA 18015*

## Order parameters and folding times

We define the fraction of ordered contacts $Q_{\rm s}$ relative to a given structure "s" (not necessarily the native state) as

$$Q_{\rm s} = N_{\rm s}^{-1} \sum_{(i,j)} \frac{1}{1+\exp(\gamma(r_{ij} - \lambda r_{ij}^0))} \tag{1}$$

where the sum runs over the $N_{\rm s}$ pairs $(i,j)$ of native atomic contacts which are separated by distances $r_{ij}$ in the configuration of interest and by $r_{ij}^0$ in "s" ($\gamma = 5$ Å$^{-1}$; $\lambda = 1.5$). Only atom pairs closer than 4.5 Å in "s", belonging to residues separated by more than 2 in sequence are included. We then define the global order parameter $Q_{\rm n-nn} = Q_{\rm n} - Q_{\rm nn}$, where $Q_{\rm n}$ and $Q_{\rm nn}$ are defined as above using the native structure and misfolded intermediate illustrated in Fig. 1 in the main text. Note that while the symbol $Q$ is often used in the literature to refer to *native* contacts, we have generalized it here and in our previous work on the equilibrium properties of GB1[1] to refer to

*To whom correspondence should be addressed
[†]Cambridge University
[‡]Lehigh University

similarity to any structure; an analogous measure is the structural overlap function[2] (although the latter runs from 1 for unfolded to 0 for folded).

## Maximum Likelihood folding/unfolding times

Folding and unfolding times were estimated via maximum likelihood. In the case of a single exponential distribution of folding times, the maximum likelihood folding time is given by:

$$\tau_{\rm f} = N_{\rm f}^{-1}\Big\{\sum_{i=1}^{N_{\rm f}} t_i + \sum_{i=N_{\rm f}+1}^{N} T_i\Big\} \tag{2}$$

where the first sum runs over the first passage times $t_i$ for the $N_{\rm f}$ trajectories which fold, and the second over the simulation lengths $T_i$ for the trajectories which do not fold. The error was estimated as $\tau_{\rm f}/N_{\rm f}^{1/2}$. The corresponding maximum log-likelihood is

$$\ln L({\rm SE}) = -\tau_{\rm f}^{-1}\Big\{\sum_{i=1}^{N_{\rm f}} t_i + \sum_{i=N_{\rm f}+1}^{N} T_i\Big\} - N_{\rm f}\ln\tau_{\rm f} \tag{3}$$

For a double exponential distribution of folding times, the log-likelihood function is:

$$\begin{aligned}\ln L({\rm DE}) &= \sum_{i=1}^{N_{\rm f}} \ln[A_1\tau_1^{-1}e^{t_i/\tau_1} + (1-A_1)\tau_2^{-1}e^{t_i/\tau_2}] \\ &\quad + \sum_{i=N_{\rm f}+1}^{N} \ln[A_1 e^{T_i/\tau_1} + (1-A_1)e^{T_i/\tau_2}]\end{aligned} \tag{4}$$

Since there is no simple expression for the maximum likelihood parameters in this case, the parameter space was sampled by Metropolis Monte Carlo in which $-\ln L$ was used as the energy. The maximum likelihood parameters were determined by simulated annealing from a temperature of 5.0 to 0.0; the posterior distribution of parameters was obtained from a simulation at a temperature of 1.0 (uniform prior).

The likelihood ratio $D = -2(\ln L({\rm SE}) - \ln L({\rm DE}))$ was used to test whether the double exponential (three parameters) was significantly better than single exponential (one parameter). For the

folding simulations initiated from the U-A ensemble, the log-likelihood was essentially identical for single and double exponential solutions, therefore the single exponential was more appropriate. For the simulations initiated from the U-B ensemble at 350 K, $D = 13.1$, i.e. the double exponential provides a better fit at better than 1% significance. For the simulations initiated from the U-B ensemble at 300 K, $D = 3.15$, indicating that the double exponential is probably a better fit, but only at an $\sim 20\%$ significance level.

## Theoretical estimate of transition-path durations

An estimate of the transition-path time can be made by assuming diffusive dynamics over a parabolic barrier, using an expression due to Szabo:[3]

$$\langle \tau_{\mathrm{TP}} \rangle \approx \frac{\ln[2e^{\gamma}\beta\Delta G^{\ddagger}]}{D\beta(\omega^{\ddagger})^2} \tag{5}$$

In the above, $\gamma = 0.577..$ is the Euler-Mascheroni constant, $\Delta G^{\ddagger}$ is the height of the barrier, $D$ is the diffusion coefficient for movement along the coordinate, and $(\omega^{\ddagger})^2$ gives the curvature of the barrier. If in addition it is assumed that curvature of the unfolded well $(\omega)^2$ is similar to that of the barrier on the given coordinate (see Fig. 1 of the main text), then the expression can be rewritten in terms of the Kramers preexponential factor $k_0 = D\omega\omega^{\ddagger}\beta/2\pi$, giving

$$\langle \tau_{\mathrm{TP}} \rangle \approx \frac{\ln[2e^{\gamma}\beta\Delta G^{\ddagger}]}{2\pi k_0} = \frac{\ln[2e^{\gamma}\beta\Delta G^{\ddagger}]}{2\pi k \exp[\beta\Delta G^{\ddagger}]} \tag{6}$$

Using the barrier heights for folding on $Q_{\mathrm{n-nn}}$ of 6.0, 4.4 and 4.3 $k_{\mathrm{B}}T$ at 300, 325 and 350 K respectively, and the mean first passage times from U-A in Table 1 of the main text, we estimate transition-path times of 71, 66 and 31 ns at 300, 325 and 350 K respectively, similar to those obtained from simulation.

## Explanation for observed folding amplitudes

The observed amplitudes in the simulations initiated from $U_A$ and $U_B$ can be explicitly rationalized using a simplified chemical kinetics model, shown below. Note that this model is a minimal explanation for the observed kinetics, but a more complex model would be needed to accurately reflect the complete microscopic folding dynamics. We plan to build such a model in our future work.

$$\mathsf{U_S} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathsf{U_B} \xrightarrow{k_2} \mathsf{F}$$

In this scheme, the unfolded state is considered to consist of an unstructured subensemble $U_B$ as defined in the main text and a structured subensemble $U_S$, defined as the remainder of the unfolded state. The motivation for the chosen 3-state scheme comes from the fact that folding events always appear from the unstructured subensemble $U_B$ (several examples are given in Fig. 3 in the main text). The equilibrium unfolded state, $U_A$, comprises both $U_S$ and $U_B$ with the correct relative weights. We note that the fraction folded $P_F(\tau)$ obtained from this scheme would be identical to the cumulative distribution of first passage times $P(t < \tau)$ for folding discussed in the main text. The two non-zero eigenvalues of the scheme are given by

$$\lambda_{\pm} = [k_1 + k_{-1} + k_2 \pm \sqrt{(k_1 + k_{-1} + k_2)^2 - 4k_1k_2}]/2 \tag{7}$$

The evolution of the folded fraction is determined by the initial populations $P_S^0$, $P_B^0$, where $P_S^0 + P_B^0 = 1$ as:

$$\begin{aligned} P_F(\tau) = P_S^0 \Big[1 + \frac{\lambda_-}{\lambda_+ - \lambda_-} e^{-\lambda_+ \tau} - \frac{\lambda_+}{\lambda_+ - \lambda_-} e^{-\lambda_- \tau}\Big] \\ + P_B^0 \Big[1 + \frac{\lambda_- - k_2}{\lambda_+ - \lambda_-} e^{-\lambda_+ \tau} - \frac{\lambda_+ - k_2}{\lambda_+ - \lambda_-} e^{-\lambda_- \tau}\Big] \end{aligned} \tag{8}$$

These solutions can be used to explain the observed amplitudes for the different initial conditions in the main text. For example, at 350 K, the eigenvalues from our fit are $\lambda_+ = 1/\tau_2 = 100\ \mu s^{-1}$

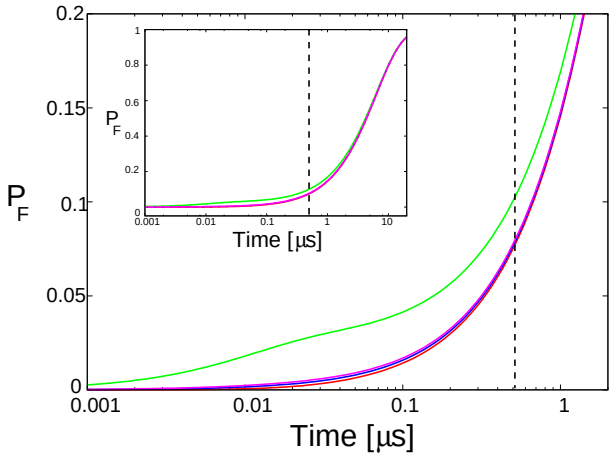


Figure S5: Folded fraction $P_{\mathrm{F}}(t)$ obtained from kinetic models. Blue curve is single exponential with time scale $\lambda_{-}$. Green, red, magenta curves are obtained from Eq. 9 with respective initial conditions $(P_{\mathrm{S}}^0, P_{\mathrm{B}}^0)$ of (0,1), (1,0), (0.9,0.1). Inset: expanded time scale showing more of the complete transients. Broken vertical line indicates the time at which trajectories at 350 K were truncated (0.5 $\mu$s).

and $\lambda_{-} = 1/\tau_1 = 0.159\ \mu\mathrm{s}^{-1}$. Note that the amplitudes for the relaxation of the $U_{\mathrm{S}}$ population are determined entirely by the kinetic eigenvalues, and the amplitudes for simulations initiated from $U_{\mathrm{B}}$ are known from the fit in the main text (0.026 and 0.974 for the fast and slow phases respectively). We can therefore write:

$$\begin{aligned} P_{\mathrm{F}}(\tau) \approx P_{\mathrm{S}}^0 \left[1 + 0.0016 e^{-\lambda_+ \tau} - 1.0016 e^{-\lambda_- \tau}\right] \\ + P_{\mathrm{B}}^0 \left[1 - 0.026 e^{-\lambda_+ \tau} - 0.974 e^{-\lambda_- \tau}\right] \end{aligned} \tag{9}$$

Although not needed here, these amplitudes correspond to the microscopic rates $k_1 = 5.77\ \mu\mathrm{s}^{-1}$, $k_{-1} = 91.6\ \mu\mathrm{s}^{-1}$ and $k_2 = 2.76\ \mu\mathrm{s}^{-1}$ in the above scheme. The reason the fast mode is not observed for simulations initiated from $\mathrm{U_A}$ is that in the equilibrium unfolded state, $P_{\mathrm{S}}^0 \gg P_{\mathrm{B}}^0$, and the contribution to the relaxation of $P_{\mathrm{S}}^0$ from the fast mode $\lambda_+$ is very small. Only when the simulations are initiated from $\mathrm{U_B}$ ($P_{\mathrm{B}}^0 = 1$), is there an appreciable contribution from the fast mode. This can be seen from the curves plotted in Figure S5: of the given initial conditions, only $P_{\mathrm{B}}^0 = 1$ results in an noticeable deviation from a slow single-exponential relaxation.

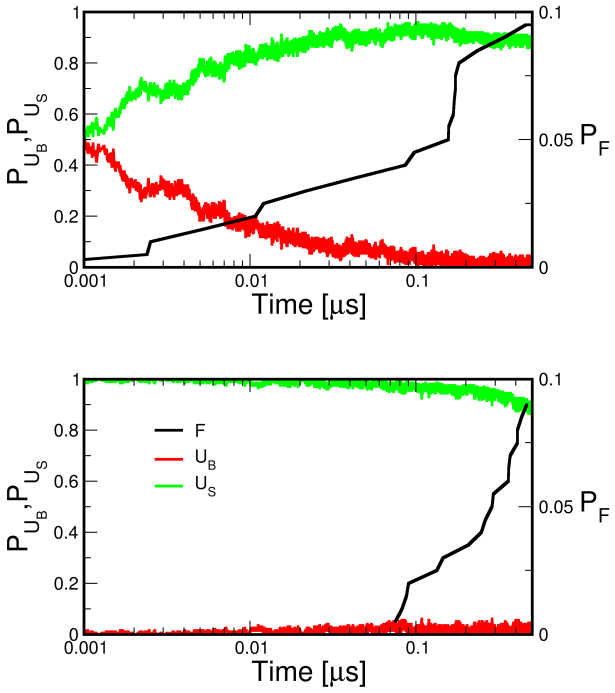


Figure S6: The population in each state, folded (F), structured unfolded ($U_B$), and unstructured unfolded ($U_S$) is shown as a function of time for simulations at 350 K. The top and bottom panels are the data from simulations starting from U-B and U-A ensembles, respectively.

The reason the first passage times are biased toward the fast phase is because the simulations are terminated after a relatively short time, where the major contribution to the increase in $P_{\mathrm{F}}(\tau)$, or $P(t < \tau)$ comes from the fast mode. Comparing the $P_{\mathrm{F}}(\tau)$ on short and long time scales (inset of Figure S5) makes the origin of this bias clear.

Figure S6 shows the population data in each state from starting ensembles U-A (bottom panel) and U-B (top panel) as a function of time, averaged over simulations at 350 K. It is clear that when simulations are started from U-B, the fast mode is associated with folding as well as non-native structure formation (an increase in $U_S$).

## Potentials of mean force for folded/unfolded state

We have calculated the potential of mean force (PMF) as a function of all the reaction coordinates separately for folded ($Q_{\mathrm{n-nn}} > 0.7$) and unfolded ($Q_{\mathrm{n-nn}} < 0.1$) states over all the trajectories at 300 K, shown in Figure S7. Whereas most of the order parameters ($Q_{\mathrm{turn}}$, $\phi$, $Q_{\mathrm{hb}}$) assume only

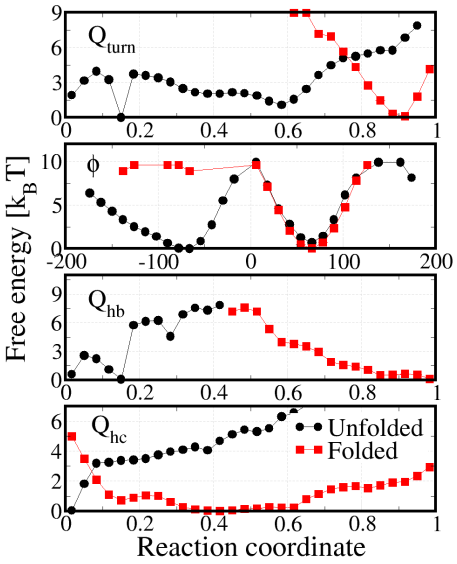


Figure S7: Separate free energy surfaces for the folded state ($Q_{n-nn} > 0.7$; red) and unfolded state ($Q_{n-nn} < 0.2$; black) for four of the reaction coordinates are shown at 300 K.

native-like values in the native state, the distribution of hydrophobic contacts is found to be rather broad. Fig. 4 (A)-(D) in the main text show that, although folding clearly does favour formation of hydrophobic interactions, these fluctuate much more than the other contacts, and are less obviously correlated with overall folding. This clearly suggests that the formation of hydrophobic contacts, while correlated with folding, is not by itself a reliable indicator of folding, i.e. is a poor reaction coordinate.

# Definition of ACF matrices and distances

We quantify the order of contact formation using an "average contact formation" (ACF) matrix **A**. We define a fraction of native contacts for each residue $i$, $Q_i$, as the fraction of all the native contacts in which $i$ is a member of the contacting pair. We then define the elements of **A**, $a_{ij}$ as the average degree of contact formation $Q_i$ of residue $i$ when $Q_{\mathrm{n-nn}}$ lies in interval $j$. We discretized $Q_{\mathrm{n-nn}}$ into intervals of width 0.1 between 0 and 0.8. The average was calculated for each transition path, defined as a trajectory segment spanning $Q_{\mathrm{n-nn}} = 0.1$ and $Q_{\mathrm{n-nn}} = 0.8$. To provide context, 1000 steps either side of each transition path were also included in the average.

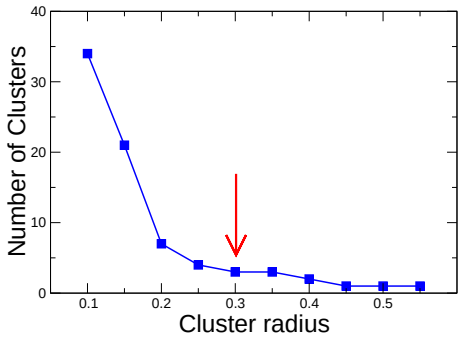


Figure S8: Dependence of number of clusters on the cluster radius, for clustering of ACF matrices with the leader algorithm. The red arrow indicates the chosen cluster radius for analysis.

The "distance" between two pathways is given by the Euclidean distance between their ACF matrices. We clustered the pathways using a simple "leader" algorithm with a cluster radius of 0.3. The leader algorithm assigns the first data point to the first cluster. Subsequent data points are compared to the first member of each cluster. If the distance between a new data point and the first member of at least one of the current clusters is less than the cutoff, the datum is assigned to the cluster whose first member it is closest to. Otherwise the data point becomes the first member of a new cluster. The process is continued until all data are assigned to a cluster.

We found that over a range of cutoff distances, we obtained 3 clusters, one of which was an outlier (Figure S8). The outlier corresponds to a very short-lived unfolding event where the protein only visits $Q_{\mathrm{n-nn}} < 0.1$ briefly, before refolding: this therefore probably does not correspond to a true unfolding event. Below $\sim 0.3$, we found a rapid increase in clusters, while above $\sim 0.45$, all data fell into a single cluster. We used a cutoff of 0.3 for subsequent analysis.

The ACF matrices for all transition paths are summarized in Figure S9.

## References

(1) Best, RB, Mittal, J (2010) Balance between $\alpha$ and $\beta$ structures in *ab initio* protein folding. *J. Phys. Chem. B* 114:8790–8798.

(2) Guo, Z, Thirumalai, D (1996) Kinetics and thermodynamics of folding of a de Novo designed four-helix bundle protein. *J. Mol. Biol.* 263:323–343.

(3) Chung, HS, Louis, JM, Eaton, WA (2009) Experimental determination of upper bound for transition path times in protein folding from single-molecule photon-by-photon trajectories. *Proc. Natl. Acad. Sci. USA* 106:11837–11844.

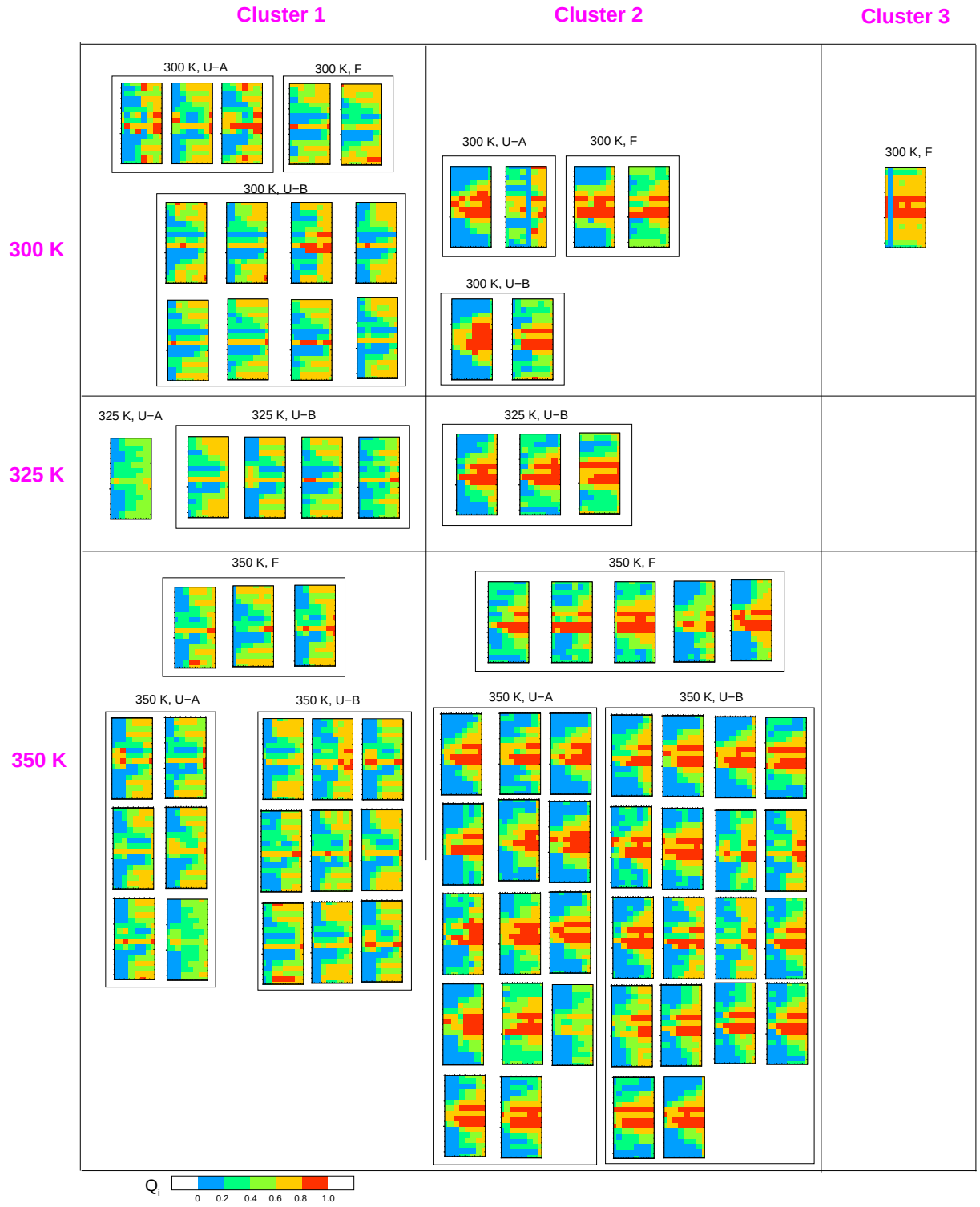


Figure S9: ACF matrices for all folding/unfolding trajectories. Matrices are grouped by cluster number (cluster 1 = "termini first"; cluster 2 = "zipper"; cluster 3 is an outlier) and by temperature. The temperature and initial conditions used for each trajectory are shown above the matrices. Note that a line of zeros in the centre of a matrix is due to a rapid crossing for which no data were saved in that interval of $Q_{\mathrm{n-nn}}$.