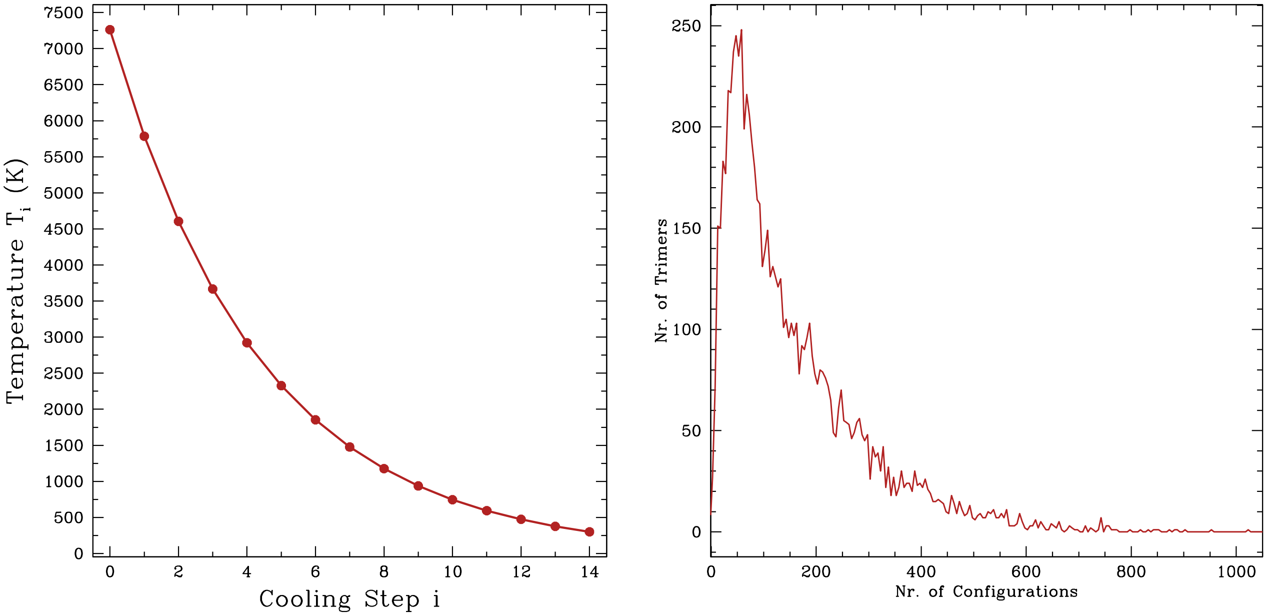
Supplemental Material

Below we provide further details on the two stages of the MuSE exploration. Additional results on higher-energy conformational ensembles obtained for the proteins in this study are also presented.

Updating Temperature in MC-SA

The initial temperature T_0 for the MC-SA scheme is determined from a desired acceptance probability associated with generated conformations. Initially, a coarse-grained conformation whose energy is $\delta E = 10$ kcal/mol higher than that of the previously generated conformation is accepted with a probability of 0.5. This acceptance probability corresponds to $e^{-\delta E/(k_{\mathrm{B}} T_0)} = 0.5$, which gives an initial temperature T_0 of ~ 7261 K (k_{B} denotes the Boltzmann constant). The final temperature T_f is set to 300 K. The temperature T_0 is progressively lowered k times according to a proportional cooling schedule that updates the MC temperature as in $T_{i+1} = T_i \cdot \frac{T_f}{T_0}^{\frac{1}{k+1}}$ until $T_k = T_f$. Temperatures for each $0 \le i \le k$ are shown in Figure 1(a).



(a)

(b)

Figure 1: (a) shows the temperatures during the MC-SA. (b) shows through a histogram the population of configurations for trimers in the local database.

Compiling a Local Database of Trimer Configurations

MuSE compiles and maintains a local fragment database during the coarse-grained exploration in the first stage. A PDB subset of nonredundant protein structures (as of July 2007) is obtained through the PISCES server (Wang and Dunbrack, 2003). Chosen proteins have $\le 40\%$ sequence similarity, ≤ 2.5 Å resolution if the structure is obtained through X-ray crystallography, or R-factor ≤ 0.2 if obtained through NMR. The $6,056$ protein chains in this subset are split into all possible overlapping fragments of three consecutive amino acids. For each 3-aa sequence (trimer), the local database maintains the list of configurations (6 backbone dihedral angles) populated by the trimer over all protein chains (a total of $10,072,004$ trimer configurations).

Figure 1(b) shows that the populations of different trimers are very heterogeneous. However, all possible 20^3 trimers are populated, and only $70/8,000$ have less than 10 configurations. Low populations for this small percentage of trimers are associated with bulky amino acids that are energetically penalized for being neighbors in a protein chain. It is worth pointing out that for all the protein sequences in this study, no trimers have less than 21 configurations in the local database.

Coarse-grained Energy Function

The coarse-grained energy function is a linear combination of non-local terms: $\mathrm{E} = \mathrm{E}_{\mathrm{Lennard-Jones}} + \mathrm{E}_{\mathrm{H-Bond}} + \mathrm{E}_{\mathrm{contact}} + \mathrm{E}_{\mathrm{water}} + \mathrm{E}_{\mathrm{burial}} + \mathrm{E}_{\mathrm{Rg}}$. $\mathrm{E}_{\mathrm{Lennard-Jones}}$ is a slightly modified version of the 12-6 Lennard-Jones potential employed in AMBER9 (Case *et al.*, 2006). The modification is introduced to allow for a soft penetration of vdw spheres. $\mathrm{E}_{\mathrm{H-Bond}}$ is a hydrogen-bonding term implemented as in (Gong *et al.*, 2005). E_{contact}, $\mathrm{E}_{\mathrm{water}}$, and $\mathrm{E}_{\mathrm{burial}}$ are implemented as in (Papoian *et al.*, 2004). These three terms, taken from the AMW function proposed by the Wolynes lab for structure prediction (Prentiss *et al.*, 2006), allow for water-mediated interactions in coarse-grained representations. The terms rely on C_β positions that are computed from the backbone of a conformation in MuSE as in (Milik *et al.*, 1997). The E_{Rg} term penalizes a conformation by $(\mathrm{Rg} - \mathrm{Rg}_{\mathrm{goal}})^2$ when its Rg is above $\mathrm{Rg}_{\mathrm{goal}}$.

Randomly Choosing Trimers in Each MC Move

Each of the $N-2$ moves in a cycle of an MC simulation in MuSE chooses a trimer randomly over the sequence of N amino acids. If instead, each move proceeds in order down the sequence, it is easy to get stuck trying to find acceptable configurations for the chosen trimer. Randomly picking trimers over the sequence allows getting out of such "local minima."

Length of a Monte Carlo Simulation

The duration of an MC simulation is determined by monitoring the convergence of averages in each of the terms of the coarse-grained energy values obtained during the simulation. If the averages converge to the same value in two successive windows of w cycles, the simulation terminates. For the proteins in this work, averages are measured every 500 cycles, and convergence is usually reached after $1,000$ cycles. An MC simulation is therefore carried for $N_{\mathrm{MC}} = 2,000$ cycles.

Determining Goal Radii of Gyration for Different Levels of Confinement

Inspection of native structures assumed by diverse proteins in the PDB reveals that one single radius of gyration cannot be imposed during the search. Traditionally, structure prediction methods bias towards native-like conformations with ideal radii of gyration $R_0 = 2.83 \times N^{0.34}$ (Gong *et al.*, 2005, Papoian *et al.*, 2004, Prentiss *et al.*, 2006). This value, close to that predicted by theory (Fleming and Rose, 2005), biases assembly towards collapsed conformations. MuSE instead aims to capture diverse functional states of proteins: non-collapsed conformations (assumed by proteins such as CaM) should not be discarded if they are energetically feasible. For this reason, each temperature in the MC-SA launches many MC simulations that employ different goal radii of gyration to allow for the possibility of non-collapsed conformations.

Very long unconfined MC simulations are carried out at the highest temperature T_0 from slightly perturbed extended conformations. The distribution of Rg values among generated conformations is analyzed to determine the average radius $\langle\mathrm{Rg}\rangle$. $\langle\mathrm{Rg}\rangle$ is 23.4 Å for calbindin $\mathrm{D}_{9\mathrm{k}}$ and 33.4 Å for both CaM and ADK. Then, $m = 11$ different values are specified for $\mathrm{Rg}_{\mathrm{goal}}$ as follows: $\mathrm{Rg}_{\mathrm{goal}} = \infty$ (meaning no confinement is imposed), $\mathrm{Rg}_{\mathrm{goal}} = \langle\mathrm{Rg}\rangle$, and 9 more values are selected for $\mathrm{Rg}_{\mathrm{goal}}$ at consecutive $\Delta\mathrm{Rg} = 1.4$ Å decrements from $\langle\mathrm{Rg}\rangle$. Using this scheme, in the most confined MC simulation for calbindin $\mathrm{D}_{9\mathrm{k}}$, $\mathrm{Rg}_{\mathrm{goal}} = 10.8$Å, which matches the value predicted from theory for a chain of 76 amino acids (Fleming and Rose, 2005).

Seeding Monte Carlo Simulations

Conformations obtained at temperature T_i are collected in the ensemble Ω_{T_i}. Conformations with energies no higher than the average energy $\langle E_{T_i}\rangle$ over Ω_{T_i} are retained in the ensemble $\Omega^*_{T_i}$. This energetic criterion ensures that conformations selected as seeds for the next MC simulations will come from low-energy regions in the coarse-grained energy landscape.

A structural analysis is conducted over $\Omega^*_{T_i}$ to identify "basins" in the coarse-grained energy landscape. Conformations are binned by radii of gyration to yield $m = 11$ sub-ensembles $\Omega^*_{T_i,R_{\mathrm{goal}}}$ for each of the selected values of $\mathrm{Rg}_{\mathrm{goal}}$. A conformation C with Rg(C) is placed in a specific bin $\mathrm{Rg}_{\mathrm{goal}}$ if $|\mathrm{Rg(C)} - \mathrm{Rg}_{\mathrm{goal}}| \leq \delta\mathrm{Rg}$, where $\delta\mathrm{Rg} = \min\{\Delta\mathrm{Rg}, \sqrt{\mathrm{Rg}_{\mathrm{goal}}\mathrm{T}_{\mathrm{i+1}}}\}$ (ΔRg is the difference between two consecutive $\mathrm{Rg}_{\mathrm{goal}}$ values, as discussed above). This binning limits the expected increase in energy by the confinement penalty at the next temperature T_{i+1} to 1.0 kcal/mol.

Conformations to seed MC simulation at the next temperature T_{i+1} and with a confinement radius R_{goal} are now chosen from the ensemble $\Omega^*_{T_i,R_{\mathrm{goal}}}$. Conformations in $\Omega^*_{T_i,R_{\mathrm{goal}}}$ are clustered according to lRMSD with the Leader clustering algorithm (Jain *et al.*, 1987). Conformations are binned in a specific cluster if they are within an lRMSD radius c_{rad} from the cluster centroid. In this work, $c_{\mathrm{rad}} = 2.0$ Å. The obtained clusters with more than $n_{\mathrm{pop}} = 5$ conformations are ordered according to their populations and the centroids for the $n_c = 100$ most populated clusters are chosen*. If less than n_c clusters meet this cutoff, the rest of the conformations are chosen from $\Omega^*_{T_i,R_{\mathrm{goal}}}$ to maximize their lRMSD from those already selected. This strategy identifies geometrically distinct conformations in the absence of highly-populated ones.

The resulting n_c conformations capture distinct "basins" in the coarse-grained landscape. However, due to inherent approximations in the coarse-grained energy function, it is not guaranteed that these basins remain low-energy minima in all-atom detail. An energetic analysis is then performed, which first adds all-atom detail to the n_c conformations as in (Heath *et al.*, 2007). The resulting all-atom n_c conformations are then energetically minimized with the AMBER ff03 force field (Duan *et al.*, 2003) using the GB implicit solvation model (Still *et al.*, 1990). The minimization uses a conjugate gradient descent that checks for convergence in energy. The correlation between coarse-grained energies and all-atom energies of conformations after the minimization is computed to associate the following score to each conformation: $(1 - R) \cdot \epsilon + R \cdot dE_{\mathrm{all-atom}}$, where R denotes the Pearson correlation coefficient, ϵ denotes the residual errors in the least squares fit between the coarse-grained and all-atom energies, and $dE_{\mathrm{all-atom}}$ denotes the difference between all-atom energies and the minimum all-atom energy obtained. Picking conformations with the lowest score ensures that, when the correlation is high, conformations whose all-atom energies best match their coarse-grained energies are chosen as seeds. Otherwise, conformations with lowest all-atom energies become more probable to be selected. A total of $n_s = 5$ conformations are considered as seeds for each value of $\mathrm{Rg}_{\mathrm{goal}}$ at each temperature.

This selection strategy takes into account the uncertainty in the coarse-grained energy function that may affect the determination of basins. Since the coarse-grained energy function integrates out all DOFs besides backbone heavy atoms, it is important to regularly estimate whether basins

*The centroid of a cluster is the lowest-energy conformation populating the cluster.

in the coarse-grained landscape remain low-energy regions when adding all-atom detail. Once the n_s seed conformations are selected, the extra DOFs (side chains) are removed and the resulting coarse-grained conformations are used to start the MC simulations at the next temperature.

Exploring Conformational Space Around an Energy Minimum

The second stage in MuSE employs the Protein Ensemble Method (PEM) to explore the all-atom conformational space around an energy minimum. PEM allows MuSE to generate more low-energy all-atom conformations around a minimum. PEM has been proposed in (Shehu *et al.*, 2006) and applied to various proteins in (Shehu *et al.*, 2007) to explore equilibrium fluctuations around a representative equilibrium conformation. The lowest-energy conformation in each pseudo free-energy minimum is used as the representative conformation for PEM in the second stage.

PEM defines consecutive fragments of length $l = 30$ and overlap of $\delta l = 5$ amino acids over the backbone of the representative conformation. The length l ensures that large fluctuations will be explored around the conformation. The overlap δl ensures that fluctuations will be consistent for neighboring fragments. An ensemble of low-energy all-atom conformations is obtained for each fragment, maintaining the rest of the backbone fixed as in the reference conformation. This ensemble is obtained by first using a coarse-grained level of detail to address geometric constraints imposed by a fixed backbone and then employing all-atom detail for energetic considerations. The resulting ensemble of all-atom low-energy conformations obtained with PEM allows MuSE to add more structural detail to an energy minimum.

Nonlinear Dimensionality Reduction of All-atom Conformational Space

MuSE employs `Sc`alable `Isomap` (ScIMAP) to project the high-dimensional space populated by low-energy all-atom conformations onto a low-dimensional space. Proposed in (Das *et al.*, 2006) and tested in (Plaku *et al.*, 2007, Shehu *et al.*, 2008), ScIMAP analyzes nonlinear surfaces associated with protein simulation data. ScIMAP first computes a nearest-neighbors graph G where each conformation is connected to its nearest neighbors. The distance between two conformations is then measured as the length of the shortest path that connects the conformations in G. The shortest-path distances between L conformations selected as landmarks and the remaining conformations are stored in a matrix M. The top eigenvectors of M provide an orthogonal basis that MuSE uses to project the computed all-atom conformations onto few global coordinates that capture the structural variability among conformations. Different number of landmarks (3000-5000) and nearest neighbors (30-50) have been tested to ensure accuracy and robustness of the obtained projections.

Calculation of Pseudo Free-Energy Values on Low-dimensional Space

Pseudo free-energy values are calculated on the low-dimensional conformational space obtained by ScIMAP. The calculations are carried out using a modified version of the weighted histogram method (WHAM) (Ferrenberg and Swendsen, 1988, 1989). The modification takes into account that the resulting low-energy all-atom conformations in MuSE do not come from a single constant temperature simulation at a given resolution; that is, the conformations do not define a canonical ensemble. First, a grid with uniform cell size is imposed on the low-dimensional space that spans conformations. The potential energy associated with conformations whose projections fall on a particular cell is averaged in order to smooth out noise originating from the different resolutions, the force field, or the solvation model used. The population of each grid cell is used to associate an entropy value to the cell and so obtain a pseudo free-energy value for each cell. Energy values calculated this way are then used to color-code the grid for visualization purposes and to reveal

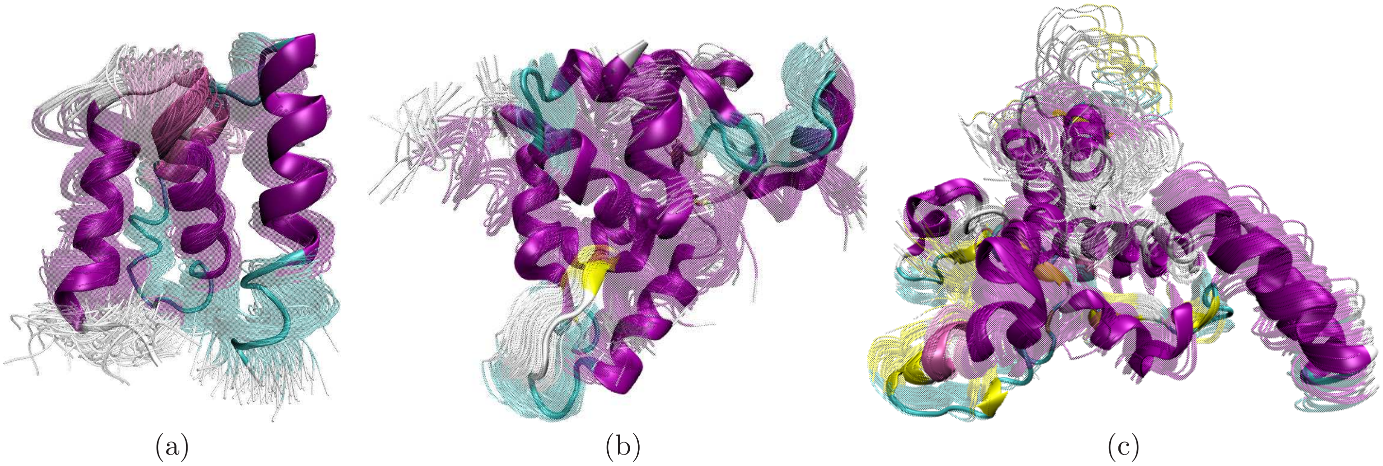


Figure 2: Figure shows higher-energy conformational ensemble obtained for calbindin D_{9k} in (a), CaM in (b) and ADK in (c). The lowest-energy conformations within each ensemble are drawn in opaque, superimposing the remaining conformations in an ensemble in transparent.

regions of the grid with lowest pseudo free-energy values. Different grid cell sizes have been tried in the calculations to ensure accuracy and consistency of the results.

Higher-energy Conformational Ensembles

Figure 2 shows some of the higher-energy conformational ensembles obtained by MuSE for the proteins in this work. The ensemble shown for calbindin D_{9k} in Figure 2(a) corresponds to the region $\{10 \leq x \leq 20, 2 \leq y \leq 8\}$ in the 2D pseudo free-energy landscape obtained for the protein (x and y refer to the two axes) and shown in the *Results* section of the paper. This ensemble displays an orientation of the EF-hand helices that is very different from that associated with the lowest energy minima obtained for calbindin D_{9k}. Figure 2(b) shows a higher-energy ensemble obtained for CaM that corresponds to the region $\{5 \leq x \leq 10, 5 \leq y \leq 10\}$ in the 2D pseudo free-energy landscape obtained for CaM. The ensemble contains collapsed conformation where the central linker is bent as in the collapsed state of CaM. The ensemble shown for ADK in Figure 2(c) contains higher-energy conformations corresponding to the region $\{-35 \leq x \leq -25, -15 \leq y \leq 0\}$ in the 2D pseudo free-energy landscape obtained for ADK. This ensemble also contains collapsed conformations. The issue of whether these ensembles are yet to be observed in experiment or are spurious consequences of approximations of the method is discussed in the *Discussion and Conclusion* section of the paper.

References

Case, D. A., Darden, T. A., Cheatham, T. E. I., Simmerling, C. L., Wang, J., Duke, R. E., Luo, R., Merz, K. M., Pearlman, D. A., Crowley, M., Walker, R. C., Zhang, W., Wang, B., Hayik, S., Roitberg, A., Seabra, G., Wong, K. F., Paesani, F., Wu, X., Brozell, S., Tsui, V., Gohlke, H., Yang, L., Tan, C., Mongan, J., Hornak, V., Cui, G., Beroza, P., Mathews, D. H., Schafmeister, C., Ross, W. S., and Kollman, P. A. (2006). Amber 9.

Das, P., Moll, M., Stamati, H., Kavraki, L. E., and Clementi, C. (2006). Low-dimensional free energy landscapes of protein folding reactions by nonlinear dimensionality reduction. *Proc. Natl. Acad. Sci. USA*, **103**(26), 9885–9890.

Duan, Y., Wu, C., Chowdhury, S., Lee, M. C., Xiong, G. M., Zhang, W., Yang, R., Cieplak, P., Luo, R., Lee, T., Caldwell, J., Wang, J. M., and Kollman, P. (2003). A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, **24**(16), 1999–2012.

Ferrenberg, A. M. and Swendsen, R. H. (1988). New monte carlo technique for studying phase transitions. *Phys. Rev. Lett.*, **61**(23), 2635–2638.

Ferrenberg, A. M. and Swendsen, R. H. (1989). Optimized monte carlo data analysis. *Phys. Rev. Lett.*, **63**(12), 1185–1198.

Fleming, P. J. and Rose, G. D. (2005). *Protein Folding Handbook*. Wiley-VCH, Weinheim, Germany, 1 edition.

Gong, H., Fleming, P. J., and Rose, G. D. (2005). Building native protein conformation from highly approximate backbone torsion angles. *Proc. Natl. Acad. Sci. USA*, **102**(45), 16227–16232.

Heath, A. P., Kavraki, L. E., and Clementi, C. (2007). From coarse-grain to all-atom: Towards multiscale analysis of protein landscapes. *Proteins: Struct. Funct. Bioinf.*, **68**(3), 646–661.

Jain, A. K., Dubes, R. C., and Chen, C. C. (1987). Bootstrap techniques for error estimation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **9**(55), 628–633.

Milik, M., Kolinski, A., and Skolnick, J. (1997). Algorithm for rapid reconstruction of protein backbone from alpha carbon coordinates. *J. Comput. Chem.*, **18**(1), 80–85.

Papoian, G. A., Ulander, J., Eastwood, M. P., Luthey-Schulten, Z., and Wolynes, P. G. (2004). Water in protein structure prediction. *Proc. Natl. Acad. Sci. USA*, **101**(10), 3352–3357.

Plaku, E., Stamati, H., Clementi, C., and Kavraki, L. E. (2007). Fast and reliable analysis of molecular motions using proximity relations and dimensionality reduction. *Proteins: Struct. Funct. Bioinf.*, **67**(4), 897–907.

Prentiss, M. C., Hardin, C., Eastwood, M. P., Zong, C., and Wolynes, P. G. (2006). Protein structure prediction: The next generation. *J. Chem. Theory Comput.*, **2**(3), 705–716.

Shehu, A., Clementi, C., and Kavraki, L. E. (2006). Modeling protein conformational ensembles: From missing loops to equilibrium fluctuations. *Proteins: Struct. Funct. Bioinf.*, **65**(1), 164–179.

Shehu, A., Kavraki, L. E., and Clementi, C. (2007). On the characterization of protein native state ensembles. *Biophys. J.*, **92**(5), 1503–1511.

Shehu, A., Kavraki, L. E., and Clementi, C. (2008). Unfolding the fold of cyclic cysteine-rich peptides. *Protein Sci.*, **17**(3), 482–493.

Still, W. C., Tempczyk, A., Hawley, R. C., and Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.*, **112**(16), 6127–6129.

Wang, G. and Dunbrack, R. L. (2003). Pisces: a protein sequence culling server. *Bioinformatics*, **19**(12), 1589–1591.