Supporting Information

Ceriotti et al. 10.1073/pnas.1108486108

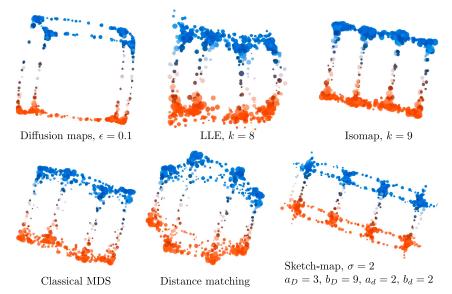


Fig. 51. 2D projections of the landmark points selected from the dataset depicted in Fig. 2 of the main text. As for sketch-map distances were calculated in a way that takes the periodicity of the space into account. Projections obtained using a number of commonly used manifold learning algorithms are shown along with the parameters used (for all the algorithms a number of different parameters were tried and similar results were obtained). Projected points are colored, using the key shown in Fig. 1C of the text, in accordance with the value of one of the three underlying variables. This figure makes clear that sketch-map gives a far better representation of the topology of the 3D distribution than any of the other algorithms.

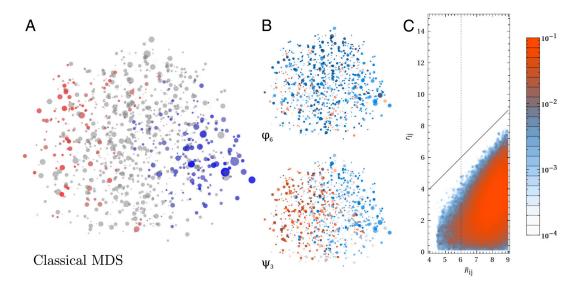


Fig. S2. Diagnostic information on the embedding of the ala12 landmark points using classical multidimensional scaling (MDS). Distances here are calculated using the differences between torsional angles in the landmark configurations taking into account the periodicity of the space. A should be compared with Fig. 4 of the main text as here points are colored according to the protein secondary structure. *B* should be compared with Fig. 5C of the main text as here points are colored according to the protein secondary structure. *B* should be compared with Fig. 5C of the main text as here points are colored according to the value of a particular dihedral angle. Lastly, *C* shows the joint probability distribution of the distances between two frames in high dimension (R_{ij}) and the distance between the corresponding low-dimensional projections (r_{ij}).

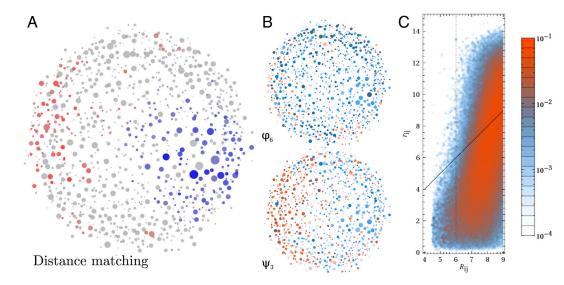


Fig. S3. Diagnostic information on the embedding of the ala12 landmark points using distance matching. Distances here are calculated using the differences between torsional angles in the landmark configurations taking into account the periodicity of the space. A should be compared with Fig. 4 of the main text as here points are colored according to the protein secondary structure. *B* should be compared with Fig. 5C of the main text as here points are colored according to the value of a particular dihedral angle. Lastly, C shows the joint probability distribution of the distances between two frames in high dimension (R_{ij}) and the distance between the corresponding low-dimensional projections (r_{ij}).

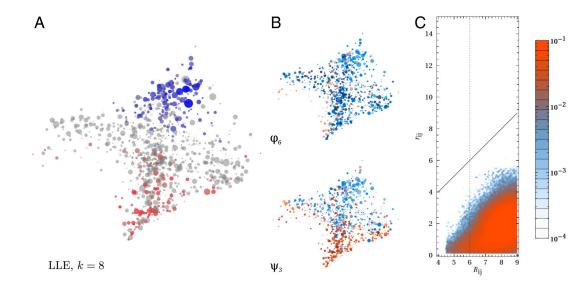


Fig. 54. Diagnostic information on the embedding of the ala12 landmark points using locally linear embedding (LLE). Each landmark configuration was described as a 24-dimensional vector of torsional angles and in calculating the vectors connecting configurations we took the periodicity of the space into account. A should be compared with Fig. 4 of the main text as here points are colored according to the protein secondary structure. *B* should be compared with Fig. 5C of the main text as here points are colored according to the value of a particular dihedral angle. Lastly, C shows the joint probability distribution of the distances between two frames in high dimension (R_{ij}) and the distance between the corresponding low-dimensional projections (r_{ij}).

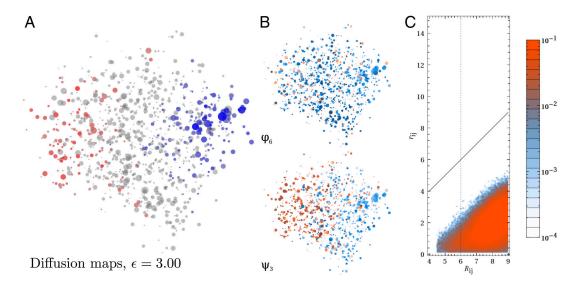


Fig. S5. Diagnostic information on the embedding of the ala12 landmark points using diffusion maps. Distances here are calculated using the differences between torsional angles in the landmark configurations taking into account the periodicity of the space. A should be compared with Fig. 4 of the main text as here points are colored according to the protein secondary structure. *B* should be compared with Fig. 5C of the main text as here points are colored according to the value of a particular dihedral angle. Lastly, C shows the joint probability distribution of the distances between two frames in high dimension (R_{ij}) and the distance between the corresponding low-dimensional projections (r_{ij}).

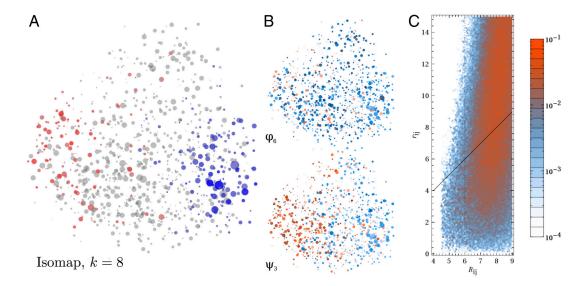


Fig. S6. Diagnostic information on the embedding of the ala12 landmark points using isomap. To calculate the *k*-nearest neighbors we computed the distances between configurations as the square root of the sum of the differences (taking periodicity into account) between the torsional angles in the reference configurations. *A* should be compared with Fig. 4 of the main text as here points are colored according to the protein secondary structure. *B* should be compared with Fig. 5C of the main text as here points are colored according to the value of a particular dihedral angle. Lastly, *C* shows the joint probability distribution of the distances between two frames in high dimension (R_{ij}) and the distance between the corresponding low-dimensional projections, (r_{ij}). Isomap is better than all the methods, other than sketch-map, in that the fraction of points which are projected close together when they are in actually far apart is noticeably smaller.

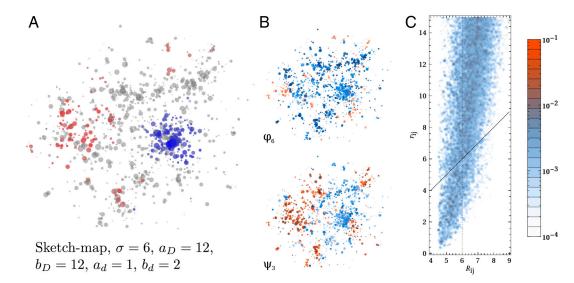


Fig. 57. Diagnostic information for a sketch-map projection of only the landmark points. In *A* points are colored according to the protein secondary structure. In *B* points are colored according to the value of a particular dihedral angle. C shows the joint probability distribution of the distances between two frames in high dimension (R_{ij}) and the distance between the corresponding low-dimensional projections, (r_{ij}). This figure is similar to the figures on this projection in the main text, but here we do not show the embeddings of nonlandmark frames. This figure is provided so a more direct comparison can be made between the results of sketch-map and the results of the other dimensionality reduction algorithms presented in the supplementary information.

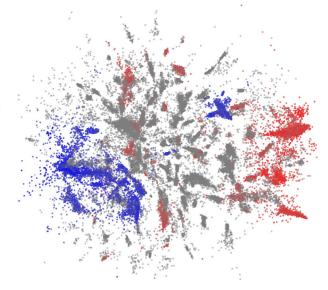


Fig. S8. Embedding of configurations from the reconnaissance metadynamcis simulations on ala12 in which the protein configurations were described in terms of the matrix of 66 $C_a - C_a$ distances rather than the torsions. In the sketch-map fitting 1,000 landmarks points were selected using the farthest point strategy. The parameters were set equal to $\sigma = 12$, $a_D = 9$, $b_D = 3$, $a_d = 1$, and $b_d = 3$ based on an analysis of the histogram of distances between configurations much like as was done for the dihedral based fitting. The embedded positions of the nonlandmark points are also shown and everything is colored according to the protein secondary structure like in Fig. 4 of the main text.