1 A detailed description of the agglomerative clustering process

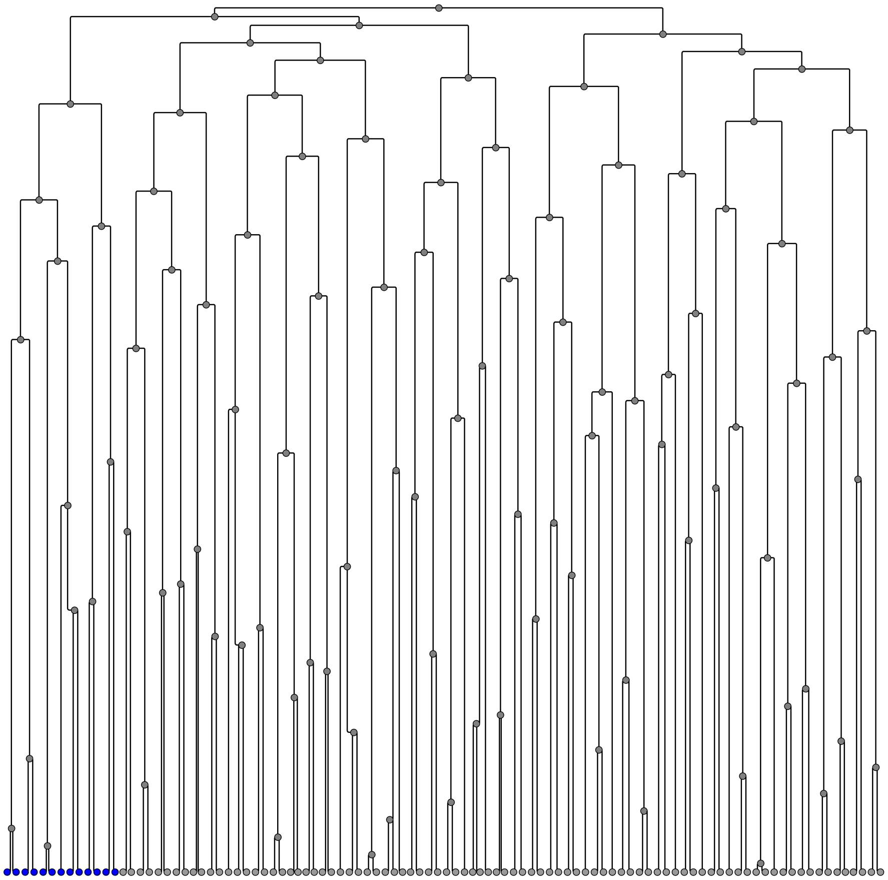
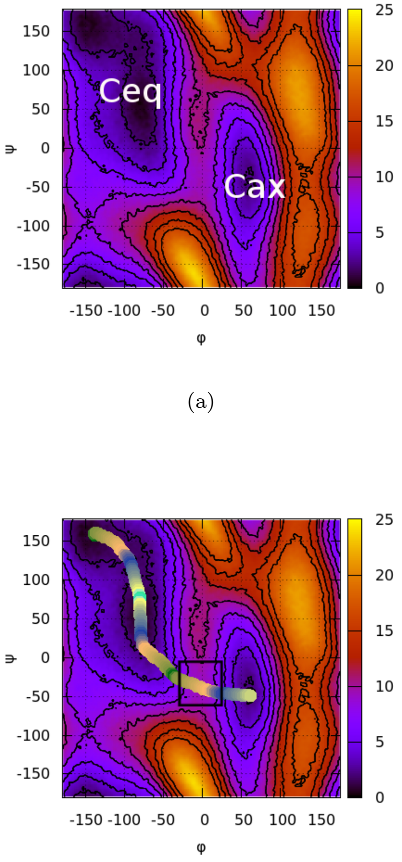


Figure 1: **A detailed description of the agglomerative clustering process.** During agglomerative clustering each of the objects are initially assigned to their own cluster and then pairs of clusters are repeatedly merged based on the same similarity function until all the initial objects are grouped in one cluster. The agglomeration can be represented as a tree or dendrogram. The leaves of the tree (terminal nodes, represented at the bottom of the figure) are the initial set of objects and newly created clusters are ordered on the vertical in the order of their discovery from bottom to top. Each object is connected to the objects it was created from. Here, we present the entire tree from which the one presented in Fig. 2 was extracted. The leaves of the tree represent the initial set of 100 partitional clusters. In blue we show the clusters that are spanned by the active conformation.

2 Using Focused Sampling on Networks (FSN) to study alanine dipeptide C_{eq} to C_{ax} transition

We show here how the FSN approach can be used to increase statistics along the transition from C_{eq} to C_{ax} sub-states of alanine dipeptide in vacuum. All simulations are run using NAMD 2.7 [3], at 300K, infinite cutoff, 0.5 fs integration time step and using Amber FF10 force field for proteins[1]. A potential of mean force profile along the φ, ψ torsion angles is shown in figure 2.



(a)

(b)

Figure 2: (a) PMF profile in the ϕ, ψ space for alanine dipeptide in vacuum at 300K, obtained using metadyanmics[2] and amber FF10 force field for proteins[1]. Contour lines are placed at 2 Kcal/mol intervals. (b) A graphical representation of the minimum free energy path between the two states. The transition state region is enclosed by a square.

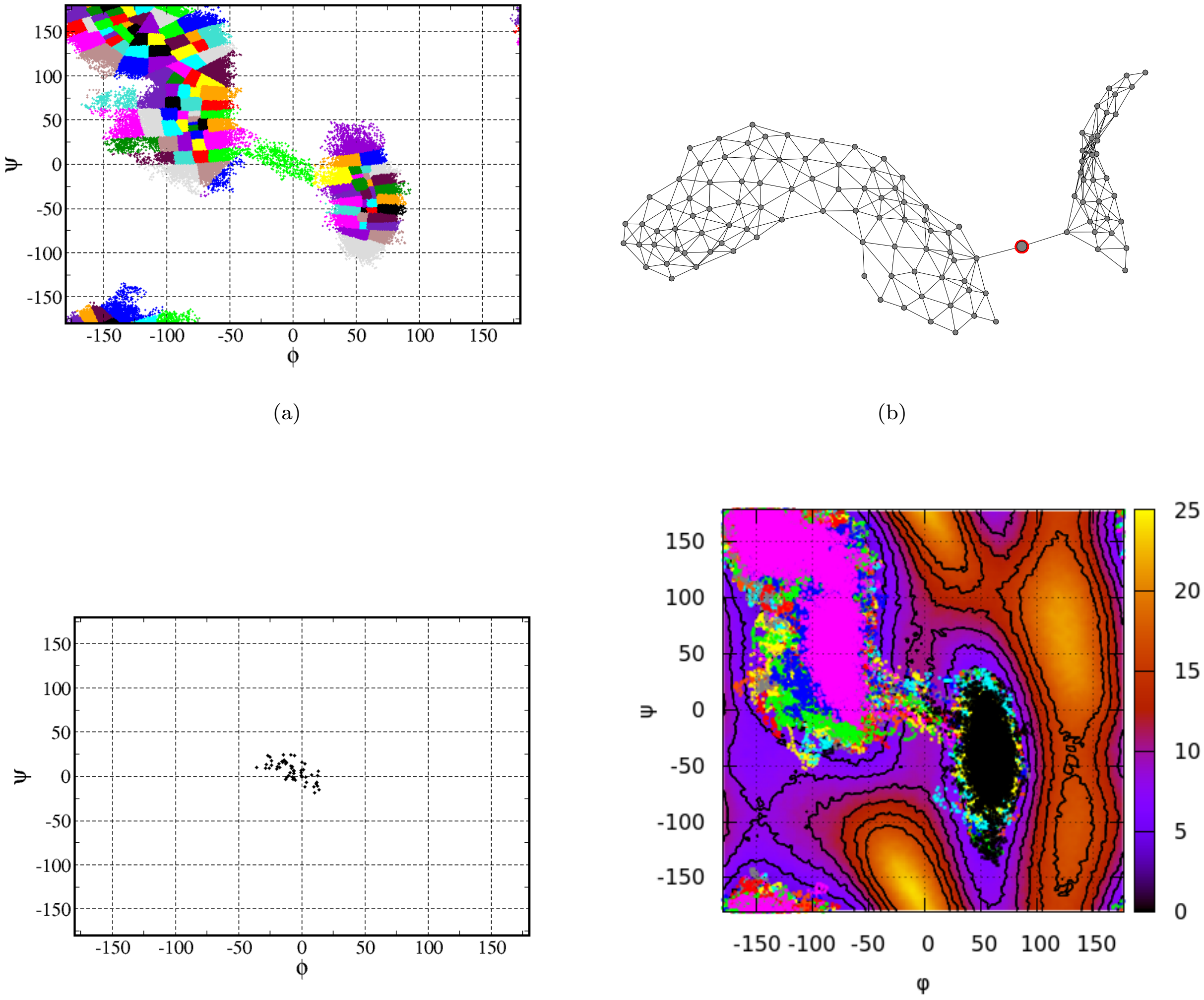
Focused sampling on networks (FSN)

Initial guess. The origins of initial swarms were chosen on the y=-x diagonal of the (ϕ, ψ) map with increments of 15 degrees. A set of 5 trajectories was initiated from each origin.

Focused sampling on networks was repeated for 40 times and had the following steps (see the main text for a further description of the method and a list of similar methods) :

- build a connected network from the existing sampled data. The network is built using 120 clusters obtained through a partitional approach of the ϕ, ψ angles represented as their corresponding pairs of sines and cosines. For efficiency, data from only the previous three iterations was used.
- evaluate sampling density. The sampling density was determined counting the number of samples within a cutoff distance, chosen to be the smallest of the clusters radii.
- evaluate traversal count using the all shortest paths between all the points.
- launch swarms of trajectories. The node of the network that ranked the highest in the traversal count and lowest in the sampling density was chosen as the origin of the swarm of trajectories. 60 trajectories were launched using randomly chosen structures belonging to the cluster associated with the node.

In figure 3 we show the details of the first iteration of FSN and in figure 4 we show the swarms of trajectories launched for the subsequent iterations. If for the first iteration the swarm of trajectories do not cover the transition state region, it can be observed that after 10 iterations the swarms of trajectories start covering the transition state region.



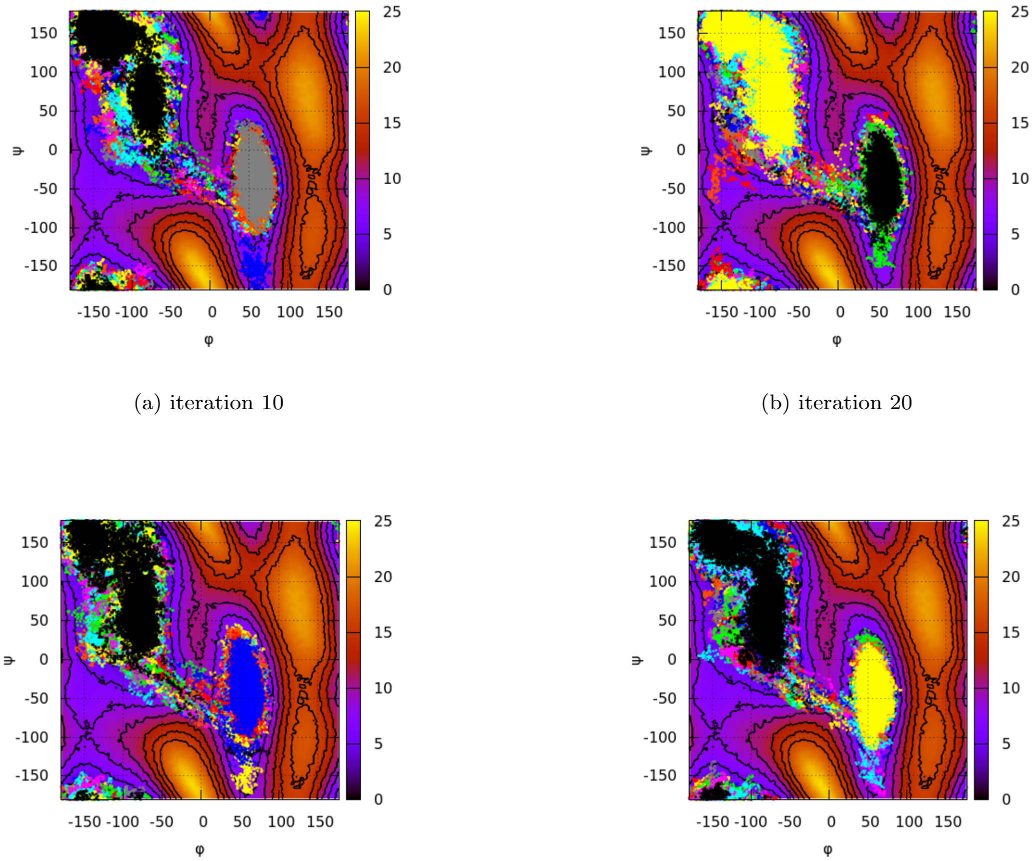
(a)

(b)

(c)

(d)

Figure 3: Inital iteration of FSN. (a) Data obtained from the initial set of swarms is clustered into 120 clusters. (b) A CSN is built assigning each cluster to a node in the network. The node that ranked highest in the traversal count and lowest in the sampling density (highlighted in red) is chosen as the next launching point of the swarm of trajectories (c) The 60 launching points are shown on the ϕ, ψ map. (d) Overalp of the obtained trajectories and the PMF profile. Each of the 60 trajectories is assigned a different color.



(a) iteration 10

(b) iteration 20

(c) iteration 30

(d) iteration 40

Figure 4: Subsequent iterations of FSN. Superimposition of the swarm of trajectories initiated for each iteration with the PMF profile. Each trajectory of the swarm is represented with a different color. It can be observed that, contrary to the first iteration shown in figure 3, the trajectories cover the transition state region.

References

[1] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of Multiple Amber Force Fields and Development of Improved Protein Backbone Parameters. *Proteins*, 65:712–725, 2006.

[2] Alessandro Laio and Michele Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA*, 99:12562–12566, 2002.

[3] James C. Phillips, Rosemary Braun, Wei Wang, James Gumbart, Emad Tajkhorshid, Elizabeth Villa, Christophe Chipot, Robert D. Skeel, Laxmikant Kaleé, and Klaus Schulten. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.*, 26:1781–1802, 2005.