SUPPLEMENTARY INFORMATION: HTMD: High-throughput molecular dynamics for molecular discovery

S. Doerr,[1] M. J. Harvey,[2] Frank Noé,[3] and G. De Fabritiis[4, 1, *]

[1] *Computational Biophysics Laboratory (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), C/ Doctor Aiguader 88, 08003 Barcelona, Spain*

[2] *Acellera, Barcelona Biomedical Research Park (PRBB), C/ Doctor Aiguader 88, 08003 Barcelona, Spain*

[3] *Department of Mathematics, Computer Science and Bioinformatics, Free University of Berlin, Berlin, Germany*

[4] *Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, Barcelona 08010, Spain*

I. ADAPTIVE EQUATIONS

First, all currently available trajectories are projected onto a given metric space using any of the projection classes described in Table 1 of the main document. Then, the projected conformations are clustered using a clustering method into a set of N_k clusters. The number of clusters, N_k, is determined by the amount of total conformations using the curve show in Figure S1. After clustering, a Markov model is constructed at a lag-time of 1 step and N_m macrostates are obtained using the PCCA+[1] algorithm. Then, the number of conformations in each macrostate M_c^m is determined by summing up the conformations in the clusters corresponding to each macrostate.

$$M_c^m = \sum_{k=1}^{N_k} \mathbb{1}_m(k) C_k \tag{1}$$

where C_k indicates the number of confomations in cluster k and $\mathbb{1}_m(k)$ is the indicator function which is 1 if cluster k belongs to macrostate m and 0 otherwise. Lastly, the counts of each macrostate are inverted and normalized, giving a probability distribution $p(m)$.

$$z_m = \frac{1}{M_c^m} \tag{2}$$

$$p(m) = \frac{z_m}{\sum_{m=1}^{N_m} z_m} \tag{3}$$

We then define a multinomial distribution on the random variables X_m, with associated probabilities $p(m)$ with $nmax - nrun$ trials giving us the number of conformations we should restart from each macrostate.

---

* Electronic address: gianni.defabritiis@upf.edu

[1] P. Deuflhard and M. Weber, Linear Algebra Appl. **398**, 161 (2005), ISSN 0024-3795, URL `http://www.sciencedirect.com/science/article/pii/S0024379504004689`.

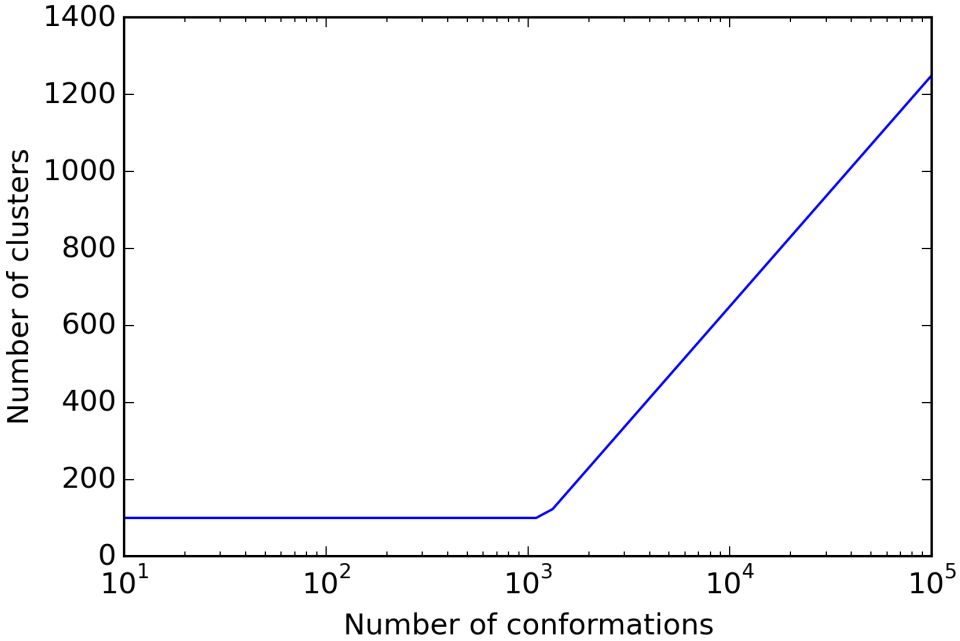

FIG. S1: Number of clusters used in the Markov models constructed during adaptive sampling. The number of clusters scales logarithmically with the number of conformations available.

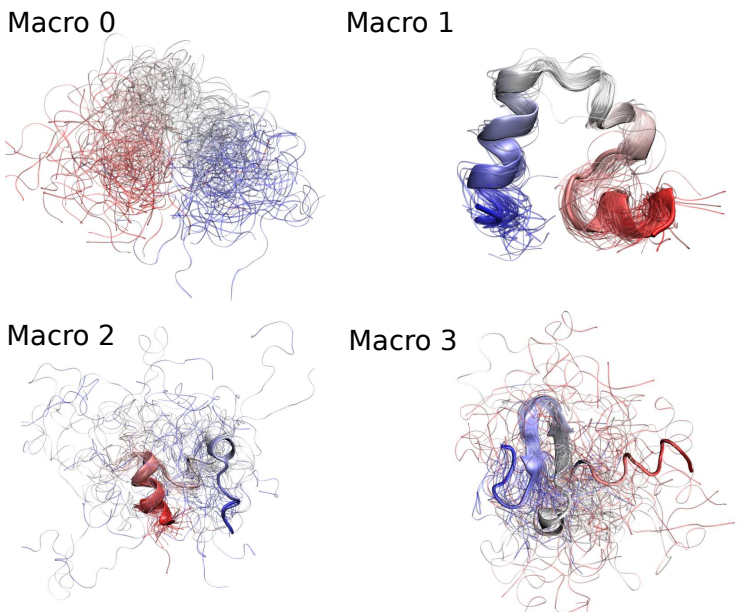

FIG. S2: The 4 macrostates produced by the Markov state model of Villin. Macrostate 0 is the unfolded state (containing various short-lived secondary structures), macrostate 1 corresponds to the folded structure, macrostate 2 corresponds to the red helix forming and macrostate 4 corresponds to a beta sheet formation.