Supporting information for: Performance and Free Energy Estimation for Solvated Polypeptides and Proteins Using Partial Infinite Swapping

Florent Hédin,[†] Nuria Plattner,[‡] J. D. Doll,[¶] and Markus Meuwly[*,†,¶]

[†]*Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland.*

[‡]*Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin.*

[¶]*Department of Chemistry, Brown University, Providence, Rhode Island 02912, USA.*

E-mail: m.meuwly@unibas.ch

1 RMSD analysis for $(Ala)_{10}$

Figure S1 illustrates the RMSD fluctuations during folding of Ala_{10}, for a 100 ns long MD simulation. The GENBORN implicit solvent model was used. The reference structure with RMSD = 0 is the folded $\alpha-$helix, cf. Figure 1 from the main text. After 15 ns one can observe a majority of quasi folded states, with a RMSD of ≈ 2 Å.

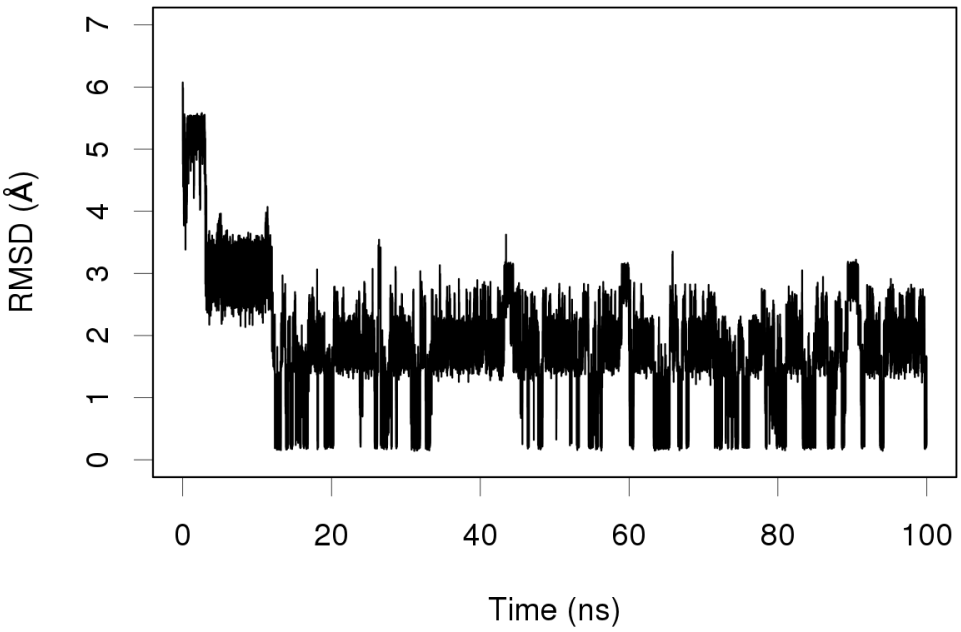


Figure S1: RMSD fluctuations observed for folding of (Ala_{10}), 100 ns of MD with GENBORN solvent. The reference structure is the α-helix from Figure 1 from the main text.

2 K-means clustering on the RMSD distributions

Table S1: RMSD clustering of the combined MD, PT, and PINS structures from Figure 2 from the main text, around 6 centres. After obtaining the cluster centres, each structure is assigned to the closest of the 6 centres.

centres (Å)	pop. MD (%)	pop. PT (%)	pop. PINS (%)
0.3	15.5	38.7	22.8
1.2	10.1	3.7	4.3
1.9	53.7	40.5	52.9
2.7	15.0	3.0	5.7
3.5	2.7	11.8	14.1
4.8	3.0	2.3	0.2

K-means clustering[S1–S3] is a straightforward method for characterizing the diversity of sampling based on a progression coordinate, which is the RMSD in the present case. Figures S2 – S3 illustrate the procedure used for choosing the number of clusters for performing the clustering.

Figure S2 estimates which proportion $\frac{\sum_{i=1}^{K}\sigma^2(X_i)}{\sigma^2(X)}$ of the total variance of the RMSD dataset (denoted as X in the following equations) is reproduced when considering K clusters: for $K \to +\infty$ the total variance of the dataset is described. Here, the sum of the variance around each cluster is $\sigma^2(X_i)$ and the total variance of the original dataset is $\sigma^2(X)$. It is commonly observed that at some point increasing the number of clusters does not appreciably improve the variance description, and the value of K after this point is considered an acceptable value of k for the k-means clustering. The detection of such an inflection angle, is referred to as "The Elbow Method".[S4] Although this inflection point may be challenging to locate in some cases[S5] for the data analyzed here those points are easily found as $k = 6$ for MD and $k = 4$ for PT and PINS.

Figure S3 counts the sum of squares of the RMSD X within each group defined around

a cluster (WSS): this time for $K \to +\infty$ this WSS tends to 0. It is estimated according to

$$WSS = \sum_{n=1}^{K} \sum_{p=1}^{P} (X_p - X_n)^2 \quad (1)$$

where K is the number of clusters allowed for the k-means clustering, P is the total number of X points around a cluster k, X_p the RMSD value of point p and X_k is the RMSD value of the centre of the cluster k. The results from the previous Figure S2 are confirmed by Figure S3, i.e. values of $k = 6$ and $k = 4$ seem to be a reasonable choice for performing the k-means.

For those reasons it was decided to use $k = 6$ in all k-means analyses performed for the present study (see Table 4 from the main text and Table S1). Indeed this value of 6 appears to be required for describing well the RMSD distribution of the MD dataset, to which PT and PINS are compared, so it is practical to use the same k for the three methods.

But as the previous plots suggested $k = 4$ for PT and PINS, one could argue that providing $k = 6$ for those two methods adds an unnecessary number of clusters which may reduce the statistical significance of the results. Table S2 shows results of a k-means clustering with $k = 4$: when compared to Table 4 from the main text it is noticed that the 4 most populated centres are close in RMSD and then it could be concluded that using $k = 6$ instead of $k = 4$ for allowing a precise comparison with MD does not invalidate the discussion from the Applications section.

Table S2: Results of a k-means clustering of the RMSD data from Figure 2 from the main text, for PT and PINS with $k = 4$ as suggested by Figures S2 – S3. Results are similar to those with $k = 6$ in Table 4 from the main text.

centres (Å)	pop. (%)	centres (Å)	pop. (%)
0.3	40.0	0.3	22.8
1.9	43.7	1.2	4.4
3.4	10.7	1.9	55.6
4.2	5.6	3.4	17.2

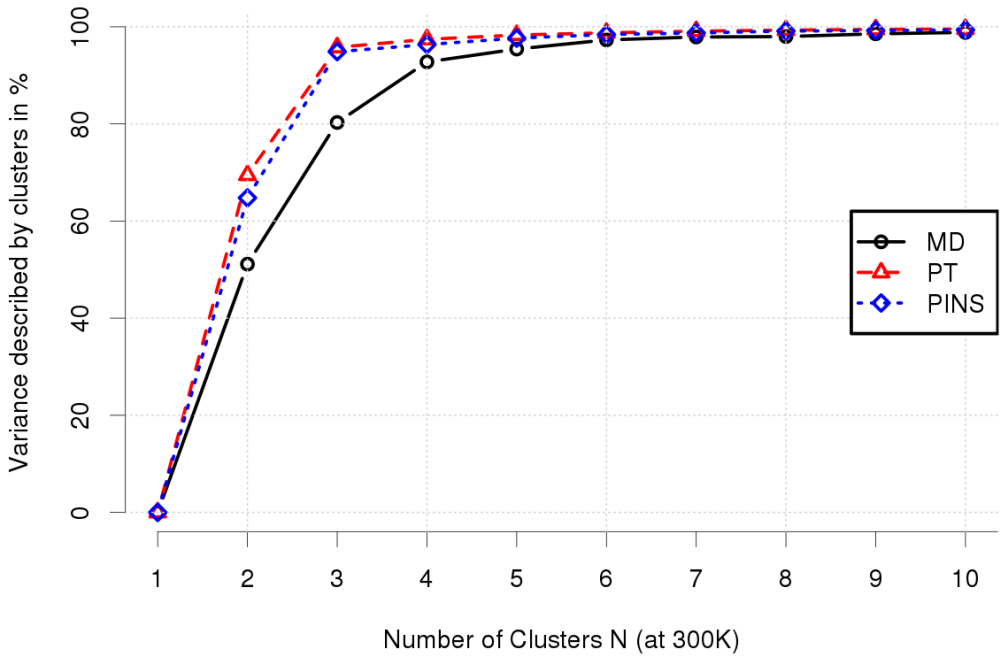


Figure S2: Proportion of the variance of the RMSD dataset described by N clusters, for the 20 ns long simulations from Figure 2 from the main text. The asymptotic behaviour for $K \geq k$ indicates that k clusters are apparently enough for describing accurately the RMSD, with $k = 6$ clusters for MD, and $k = 4$ clusters for PT/PINS.

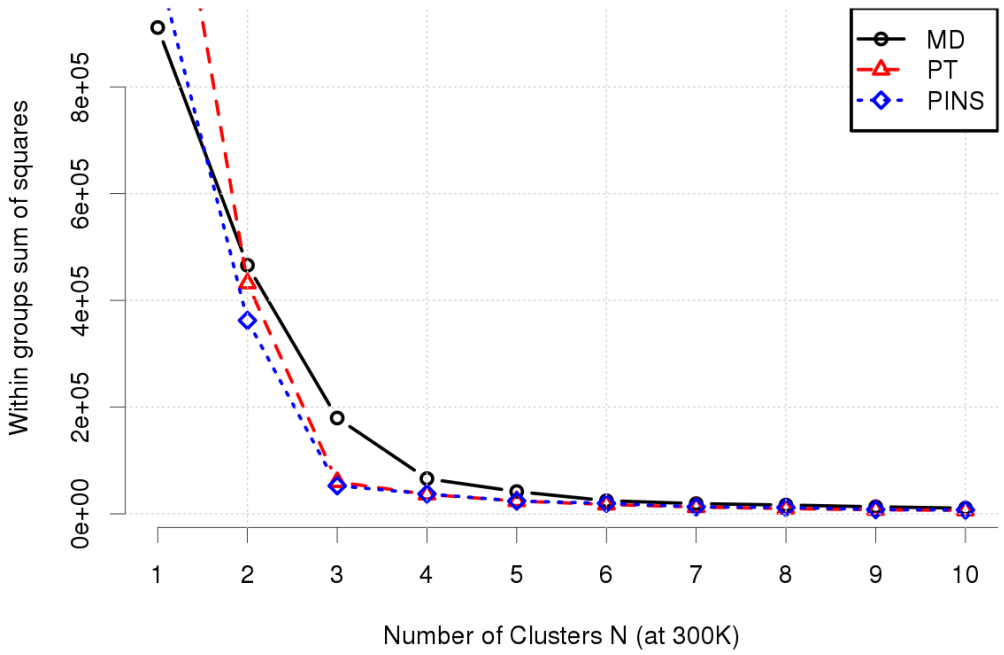


Figure S3: Evolution of the total WSS (Within groups Sum of Squares) when increasing the number of clusters k for the k-means clustering, applied to the 100 ns long implicit solvent simulations from Figure 2 from the main text. Values of $k = 6$ and $k = 4$, for respectively MD and PT/PINS, look reasonable, as adding more clusters does not reduce the overall WSS.

3 2D density estimation using KDE, and MEPs finding method

The R[S6] package **gdistance**[S7] provides classes and functions to calculate various distance measures and routes in heterogeneous geographic spaces represented as grids, but it is possible to apply the algorithm to any surface. The `shortestPath()` function was used for finding the Minimum Energy Path (MEP), based on the Dijkstra[S8] algorithm.

The Dijkstra algorithm expects no discontinuity on the grid when searching for a path: when building a surface using a standard 2D Histogram ($\Delta F(\xi, \alpha) = -RT \ln(\rho(\xi, \alpha))$, see Figure S4 in red for an example with deca-alanine) the transition areas are sometimes sampled poorly, and the application of the path finding algorithm may be challenging. For this reason, Kernel Density Estimation methods[S9,S10] were used for providing a trustful interpolation of the ΔF values at poorly sampled grid areas (see Figure S4 in black). Figure 8 A to F from the main text are examples of such interpolated KDE surfaces.

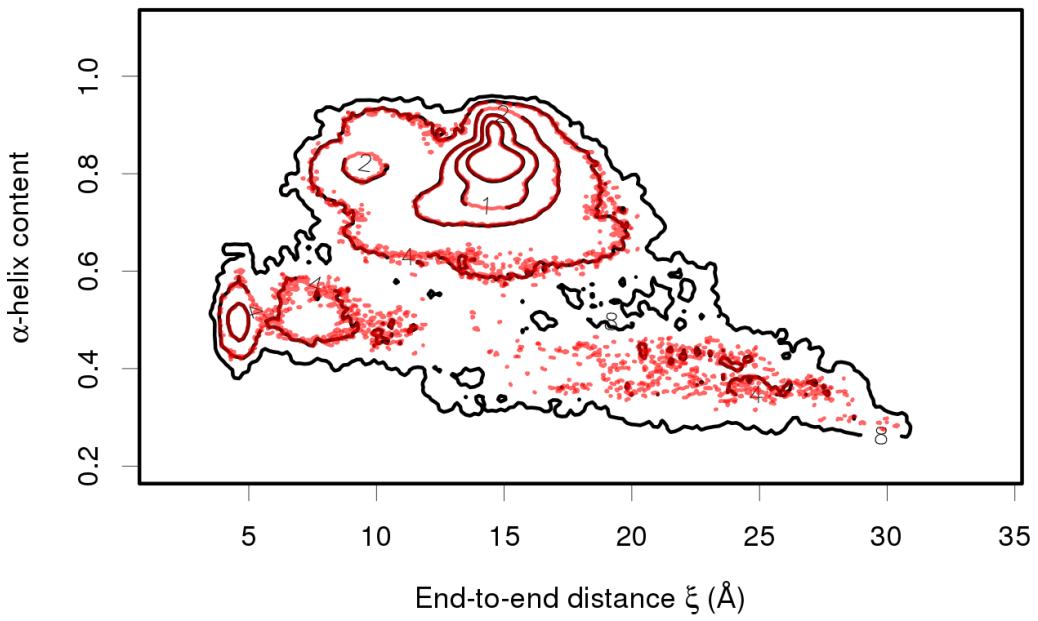


Figure S4: Free Energy contour plots built using a 2D Histogram (red) or on a 2Dim Kernel Density Estimation (black), using the end-to-end distance ($x-$axis) and the α-helix content ($y-$axis), for Ala_{10} MD simulations at 300K in implicit GENBORN solvent. The sparsity of the contour when using standard histograms (red) justifies the use of the KDE method for interpolating results (black), as one can see on Figure 8 A from the main text.

4 Calculation of the α-helical content

In order to build meaningful 2D free energy surfaces for deca-alanine, it is required to use as coordinates two properties which are easy to map to real numbers. It was decided to use the end-to-end distance ξ between carbonyls' carbons from the first and last residue (see Figure 1 from the main text), and a helicity score α detailed below. Those two coordinates were already successfully used for investigating the folding of the deca-alanine by Hénin et al.[S11] and implemented in the **colvars** package.

The α-helical content for the $N+1$ residues N_0 to N_0+N is calculated using the formula:

$$\alpha = \frac{1}{2(N-2)} \sum_{n=N_0}^{N_0+N-2} \text{angf}\left(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}\right) + \frac{1}{2(N-4)} \sum_{n=N_0}^{N_0+N-4} \text{hbf}\left(\text{O}^{(n)}, \text{N}^{(n+4)}\right) \tag{2}$$

where the scoring function angf(...) for the $\text{C}_\alpha - \text{C}_\alpha - \text{C}_\alpha$ angle is defined as:

$$\text{angf}\left(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}\right) = \frac{1 - \left(\theta(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}) - \theta_0\right)^2 / \left(\Delta\theta_{\text{tol}}\right)^2}{1 - \left(\theta(\text{C}_\alpha^{(n)}, \text{C}_\alpha^{(n+1)}, \text{C}_\alpha^{(n+2)}) - \theta_0\right)^4 / \left(\Delta\theta_{\text{tol}}\right)^4} \tag{3}$$

and the scoring function for the hydrogen bonding, hbf(...), is defined using:

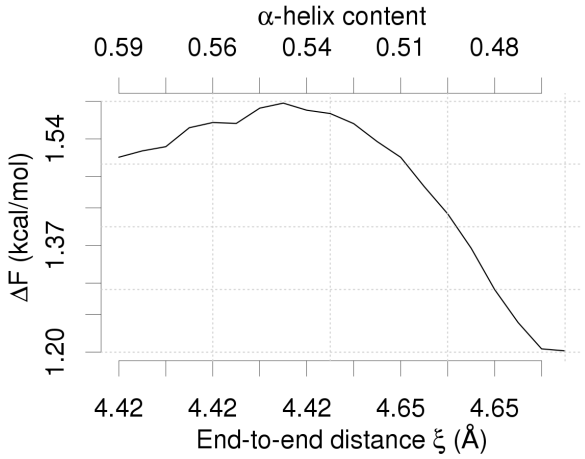
$$\text{hbf}\left(\text{O}^{(n)}, \text{N}^{(n+4)}\right) = \sum_{i \in \text{O}^{(n)}} \sum_{j \in \text{N}^{(n+4)}} \frac{1 - \left(|\mathbf{x}_i - \mathbf{x}_j| \ / \ hb_{cut}\right)^6}{1 - \left(|\mathbf{x}_i - \mathbf{x}_j| \ / \ hb_{cut}\right)^8} \tag{4}$$

where $\theta_0 = 88°$ and $\Delta\theta_{\text{tol}} = 15°$ are respectively reference and tolerance values of the $\text{C}_\alpha - \text{C}_\alpha - \text{C}_\alpha$ angle ; and $hb_{cut} = 3.3$ Å is the cutoff value under which a hydrogen bond is defined.

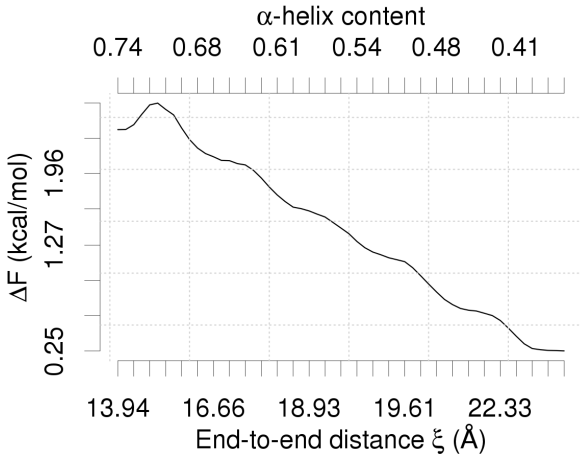
The final value of α maps to a real number between 0 and 1. When combined to the ξ end-to-end distance, one can build meaningful 2D surfaces, as seen in Figure 8 from the main text.

5 ΔF along the MEPs in explicit solvent Ala_{10} simulations

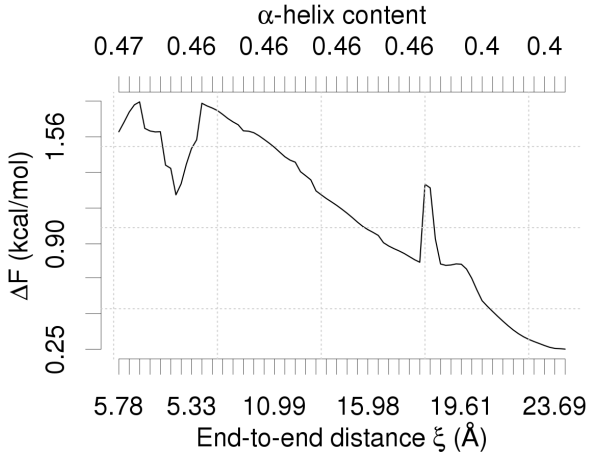
Figure S5 shows the free energy extracted along the four MEPs represented as colored lines in Figure 8 B from the main text. The barriers between points 2–3 and 4–5 are approximately of 0.4 – 0.5 kcal/mol, making transitions between those points highly probable. The free energy profile for paths 4–1 and 4–3, respectively connecting extended states to the β-hairpin and α-helix conformations, are shown on Figures S5b and S5c. The free energy change ($\Delta\Delta F$) is respectively of 2 and 1.25 kcal/mol emphasizing again the easy conformational changes during the simulation.



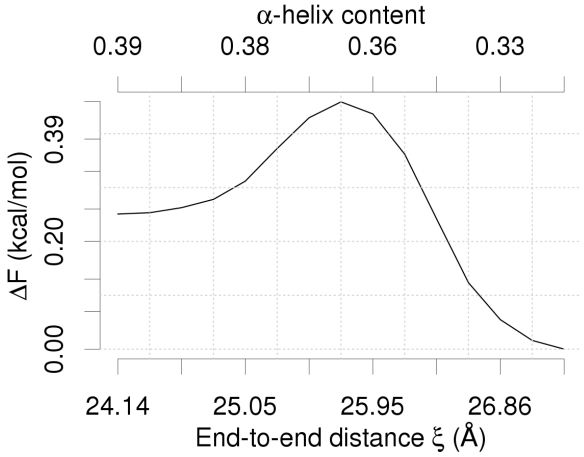
(a) ΔF between points 2 and 3 from Figure 8 B from the main text.



(b) ΔF between points 4 and 1 from Figure 8 B from the main text.



(c) ΔF between points 4 and 3 from Figure 8 B from the main text.



(d) ΔF between points 4 and 5 from Figure 8 B from the main text.

Figure S5: Free energy of the paths (ΔF in kcal/mol) displayed on Figure 8 B from the main text. The two dramatic changes in panel (c) are most probably errors either from the KDE smoothing of the MEP finding algorithm and should not be considered during analysis.

6 Effect of post-processing

The effect of the post-processing procedure is to increase convergence of simulations by enriching a given analysis state by bringing information from other states (for example higher temperatures). The procedure is introduced in the article Section "Computational Methods".

Figure S6 illustrates the effect of post-processing on a 2-dimensional Ala_{10} FES, built using the previously introduced α-helical content and the end-to-end distance ξ. Fig.S6 (Left) uses data from replica at 300K without use of the post-processing procedure, while Fig.S6 (Right) uses the same data but this time the post-processing procedure was applied.

The information brought to the 300 K replica from higher T replicas allows a more accurate sampling of the high energy configurations characterised by extended and poorly helical configurations.

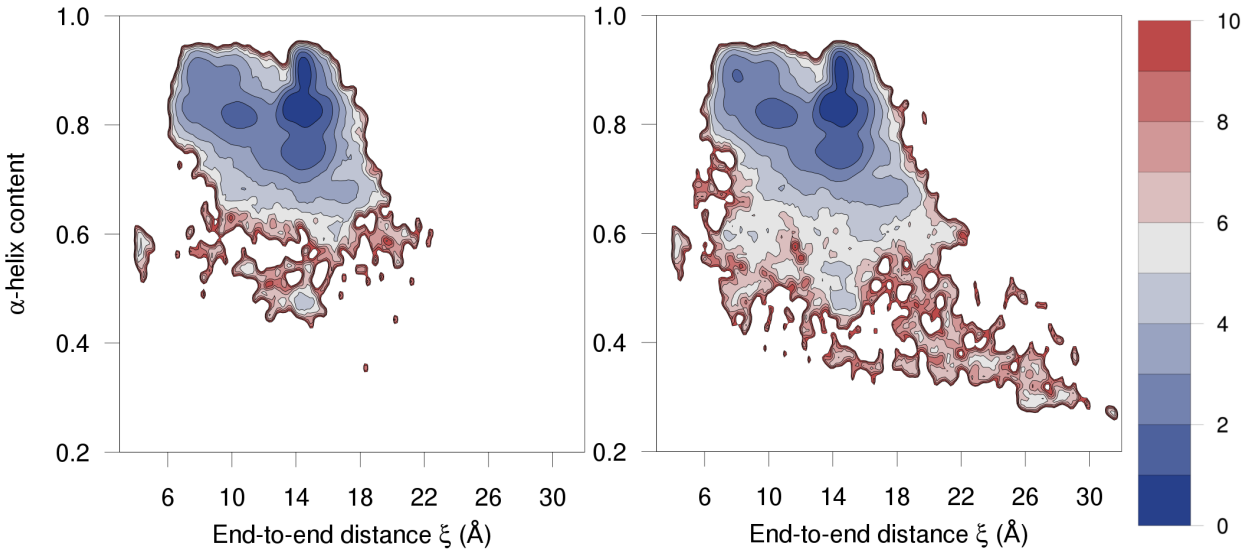


Figure S6: Effect of PINS post-processing for the Ala_{10} FES: the left Figure uses data from the replica at 300K without post-processing, while the right Figure is based on post-processing. Colour legend corresponds to the free energy given in kcal/mol.

References

(S1) MacQueen, J. Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. Berkeley, Calif., 1967; pp 281–297.

(S2) Lloyd, S. Least squares quantization in PCM. *IEEE T. Inform. Theory* **1982**, *28*, 129–137.

(S3) J. A. Hartigan, M. A. W. Algorithm AS 136: A K-Means Clustering Algorithm. *J. Roy. Stat. Soc. C-App.* **1979**, *28*, 100–108.

(S4) Thorndike, R. Who belongs in the family? *Psychometrika* **1953**, *18*, 267–276.

(S5) Ketchen, D. J.; Shook, C. L. The Application Of Cluster Analysis In Strategic Management Research: An Analysis And Critique. *Strateg. Manag. J.* **1996**, *17*, 441–458.

(S6) R Core Team, *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.

(S7) van Etten, J. *gdistance: distances and routes on geographical grids. R package version 1.1-9*; 2015.

(S8) Dijkstra, E. A note on two problems in connexion with graphs. *Numer. Math.* **1959**, *1*, 269–271.

(S9) Rosenblatt, M. Remarks on Some Nonparametric Estimates of a Density Function. *Ann. Math. Statist.* **1956**, *27*, 832–837.

(S10) Parzen, E. On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist.* **1962**, *33*, 1065–1076.

(S11) Fiorin, G.; Klein, M. L.; Hénin, J. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.* **2013**, *111*, 3345–3362.