Supporting Information for "Nonaffine Displacements Encode Collective Conformational Fluctuations in Proteins"

Dube Dheeraj Prakashchand,[†] Navjeet Ahalawat,[‡] Satyabrata Bandyopadhyay,[†] Surajit Sengupta,[*,†] and Jagannath Mondal[*,†]

[†]*Tata Institute of Fundamental Research, Center for Interdisciplinary sciences, Hyderabad 500107, India*

[‡]*Tata Institute of Fundamental Research, Center for Interdisciplinary sciences, Hyderabad 500107, India and Department of Molecular Biology, Biotechnology and Bioinformatics, Chaudhary Charan Singh, Haryana Agricultural University, Hisar 125004, India*

E-mail: surajit@tifrh.res.in,+914020203089; jmondal@tifrh.res.in,+914020203091

NAP calculation:

The equation for NAP as reported in the main draft is as follows:

$$\mathrm{NAP}(i) = \min_{\mathcal{D}} \sum_{j=1}^{n_\Omega} [\mathbf{r}_j - \mathbf{r}_i - \mathcal{D}(\mathbf{R}_j - \mathbf{R}_i)]^2$$

Here the lowercase r_i and r_j are the reduced coordinates of the current time-step ($time = t$) in the MD simulation i.e. $r_j = r_j(t) - R_j$ and $r_i = r_i(t) - R_i$; while the uppercase R_i and R_j belong to the reference structure. A simple re-arrangement results in expression $r_j - r_i = [(r_j(t) - r_i(t)) - (R_j - R_i)]$, which accounts for a change that the length between jth and ith particle undergoes at ($time = t$) compared to that in the reference structure. The above expression can also be written as:

$\mathrm{NAP}(i) = \min_{\mathcal{D}} \sum_{j=1}^{n_\Omega} [\Delta_{ij} - \mathcal{D}(\mathbf{R}_j - \mathbf{R}_i)]^2$

where $\Delta_{ij} = r_j - r_i = [(r_j(t) - r_i(t)) - (R_j - R_i)]$

The minimization process of the above eq. can also be carried out by adopting the following projection-formalism.

$\boldsymbol{NAP(i) = (S(i)^T S(i))_{min}}$ where,

$$S(i) = \begin{pmatrix} \Delta_{1x} - [D_{xx}(R_1 - R_i)_x + D_{xy}(R_1 - R_i)_y + D_{xz}(R_1 - R_i)_z] \\ \Delta_{1y} - [D_{yx}(R_1 - R_i)_x + D_{yy}(R_1 - R_i)_y + D_{yz}(R_1 - R_i)_z] \\ \Delta_{1z} - [D_{zx}(R_1 - R_i)_x + D_{zy}(R_1 - R_i)_y + D_{zz}(R_1 - R_i)_z] \\ . \\ . \\ . \\ \Delta_{n_\Omega x} - [D_{xx}(R_{n_\Omega} - R_i)_x + D_{xy}(R_{n_\Omega} - R_i)_y + D_{xz}(R_{n_\Omega} - R_i)_z] \\ \Delta_{n_\Omega y} - [D_{yx}(R_{n_\Omega} - R_i)_x + D_{yy}(R_{n_\Omega} - R_i)_y + D_{yz}(R_{n_\Omega} - R_i)_z] \\ \Delta_{n_\Omega z} - [D_{zx}(R_{n_\Omega} - R_i)_x + D_{zy}(R_{n_\Omega} - R_i)_y + D_{zz}(R_{n_\Omega} - R_i)_z] \end{pmatrix} \tag{1}$$

$$\mathsf{S(i)} = \begin{pmatrix} \Delta_{1x} \\ \Delta_{1y} \\ \Delta_{1z} \\ . \\ . \\ . \\ \Delta_{n_\Omega x} \\ \Delta_{n_\Omega y} \\ \Delta_{n_\Omega z} \end{pmatrix} - \begin{pmatrix} (R_1 - R_i)_x & (R_1 - R_i)_y & (R_1 - R_i)_z & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & (R_1 - R_i)_x & (R_1 - R_i)_y & (R_1 - R_i)_z & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & (R_1 - R_i)_x & (R_1 - R_i)_y & (R_1 - R_i)_z \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ (R_{n_\Omega} - R_i)_x & (R_{n_\Omega} - R_i)_y & (R_{n_\Omega} - R_i)_z & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & (R_{n_\Omega} - R_i)_x & (R_{n_\Omega} - R_i)_y & (R_{n_\Omega} - R_i)_z & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & (R_{n_\Omega} - R_i)_x & (R_{n_\Omega} - R_i)_y & (R_{n_\Omega} - R_i)_z \end{pmatrix} \begin{pmatrix} D_{xx} \\ D_{xy} \\ D_{xz} \\ D_{yx} \\ D_{yy} \\ D_{yz} \\ D_{zx} \\ D_{zy} \\ D_{zz} \end{pmatrix} \tag{2}$$

The above matrix equation can written in a compact form as follows:

$$\boldsymbol{S} = \Delta - \boldsymbol{Re} \tag{3}$$

where the column format of deformation matrix $\boldsymbol{D}$ has been denoted by $\boldsymbol{e}$. **Note that the Matrix $\boldsymbol{R}$ is entirely made up of coordinates from the reference structure. $\boldsymbol{\Delta}$ is, by definition, made up of the coordinates from current time-step as well those from the reference structure while $\boldsymbol{D}$ matrix is the unknown.** Except for their format, technically, they are one and the same.

$$\boldsymbol{NAP} = min[(\Delta - \boldsymbol{Re})^T(\Delta - \boldsymbol{Re})] \tag{4}$$

$$\boldsymbol{NAP} = min[(\Delta^T - \boldsymbol{e}^T\boldsymbol{R}^T)(\Delta - \boldsymbol{Re})] \tag{5}$$

$$\boldsymbol{NAP} = min[\Delta^T\Delta - \Delta^T\boldsymbol{Re} - \boldsymbol{e}^T\boldsymbol{R}^T\Delta + \boldsymbol{e}^T\boldsymbol{R}^T\boldsymbol{Re}] \tag{6}$$

Let $[\Delta^T\Delta - \Delta^T\boldsymbol{Re} - \boldsymbol{e}^T\boldsymbol{R}^T\Delta + \boldsymbol{e}^T\boldsymbol{R}^T\boldsymbol{Re}]$ be $\boldsymbol{X}$

$$\frac{\boldsymbol{dX}}{\boldsymbol{dt}} = 0, \tag{7}$$

gives us,

$$\frac{\boldsymbol{dX}}{\boldsymbol{de}} = 0 - \frac{\boldsymbol{d}(\boldsymbol{\Delta^T Re})}{\boldsymbol{de}} - \frac{\boldsymbol{d}(\boldsymbol{e^T R^T \Delta})}{\boldsymbol{de}} + \frac{\boldsymbol{d}(\boldsymbol{e^T R^T Re})}{\boldsymbol{de}} \tag{8}$$

$$\frac{\boldsymbol{dX}}{\boldsymbol{de}} = -(\Delta^T\boldsymbol{R})^T - (\boldsymbol{R}^T\Delta) + 2(\boldsymbol{R}^T\boldsymbol{Re}) \tag{9}$$

Putting $\frac{\boldsymbol{dX}}{\boldsymbol{de}} = 0$ we get,

$$2(\boldsymbol{R}^T\boldsymbol{Re}) - 2(\boldsymbol{R}^T\Delta) = 0 \tag{10}$$

which gives us,

$$\boldsymbol{e} = \boldsymbol{Q}\Delta \tag{11}$$

where $\boldsymbol{Q} = (\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T$. Here R, and hence Q, is purely a function of the reference coordinates while the delta column matrix is made up of both the coordinates from $(time = t)$ as well as the reference structure.

Substituting this $\boldsymbol{e}$ in th eq for NAP , we get:

$$\boldsymbol{NAP} = \Delta^T\Delta - \Delta^T\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T\Delta - ((\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T\Delta)^T\boldsymbol{R}^T\Delta \tag{12}$$

$$+ ((\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T\Delta)^T(\boldsymbol{R}^T\boldsymbol{R})(\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T\Delta \tag{13}$$

$$\boldsymbol{NAP} = \Delta^T\Delta - \Delta^T\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T\Delta \tag{14}$$

We can write equation [27] as follows,

$$\boldsymbol{NAP} = \Delta^T\boldsymbol{I}\Delta - \Delta^T\boldsymbol{RQ}\Delta \tag{15}$$

$$\boldsymbol{NAP} = \Delta^T(\boldsymbol{I}\Delta - \boldsymbol{RQ}\Delta) \tag{16}$$

$$\boldsymbol{NAP} = \Delta^T(\boldsymbol{I} - \boldsymbol{RQ})\Delta \tag{17}$$

Let $(\boldsymbol{I} - \boldsymbol{RQ})$ be $\boldsymbol{P}$

$$\boldsymbol{NAP} = \Delta^T(\boldsymbol{P})\Delta \tag{18}$$

The above equation is the analytical form of NAP and this is what is implemented in our Plumed codes. The code is made available in following GitHub link `https://github.com/dheeraj08dube/GNAP_CALCULATION`

Here we see that NAP is actually a of projection of a 3N dimensional vector made out of fluctuations which has been obtained by ensuring that all the affine components of the displacements have been subtracted out i.e. we can say that the P-matrix projects Δ onto a space where there is no affine-ness remaining in the displacement vector at all. **Thus, NAP(i;t) i.e. NAP for particle i at** $(time = t)$ **is purely a function of** R_{i_Ω} **and** $r_{i_\Omega}(t)$ **i.e.** $\mathbf{NAP(i;t)} = \boldsymbol{NAP(R_{i_\Omega}, r_{i_\Omega}(t))}$ **,**

where $R_{i_\Omega} \Rightarrow$ coordinates of ith particle and the particles in its neighbourhood Ω in the reference structure

and $r_{i_\Omega}(t) \Rightarrow$ corresponding coordinates at $(time = t)$.

In order that P transformation matrix is indeed a Projection operation we should have the following properties:

1) P should be symmetric. $P^T = P$

$$\boldsymbol{P}^T = (\boldsymbol{I} - \boldsymbol{RQ})^T \tag{19}$$

$\boldsymbol{P}^T = \boldsymbol{I}^T - (\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T)^T$

$\boldsymbol{P}^T = \boldsymbol{I} - (\boldsymbol{R}^{T^T}((\boldsymbol{R}^T\boldsymbol{R})^{-1})^T\boldsymbol{R}^T)$

$\boldsymbol{P}^T = \boldsymbol{I} - (\boldsymbol{R}((\boldsymbol{R}^T\boldsymbol{R})^T)^{-1}\boldsymbol{R}^T)$

$\boldsymbol{P}^T = \boldsymbol{I} - (\boldsymbol{R}(\boldsymbol{R}^T\boldsymbol{R})^{-1}\boldsymbol{R}^T)$

Therefore, we end up with $P^T = P$.

2) Similarly it can be shown that $P^2 = P$. i.e. P should be idempotent.

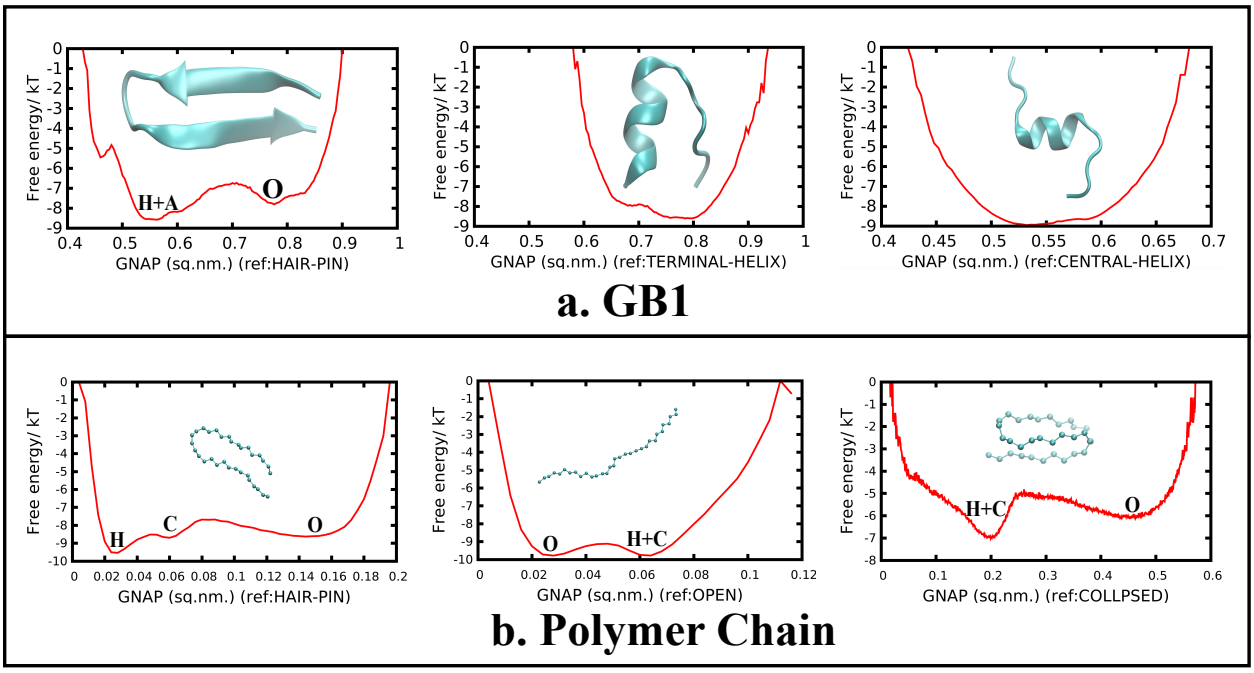

Figure S1: Illustration of different free energy curve arising due to choice of different reference structure for computing GNAP.

Details of the Protocol

Step-1: The first step is to generate the adaptive sampling trajectories for the system we wish to study. Here we ensure that the various short trajectories are adaptively and efficiently sampled for long times so that there is a thorough mixing of all the trajectories. In our analysis, for all the systems, we chose RMSD relative to the respective native structures as the only parameter which after clustering gave us the representative reference structures. The mixing of the RMSD of the 200 trajectories generated is demonstrated in the figure S2.

Step-2: As we have identified in the manuscript, the calculation of NAP (and hence GNAP) would depend on a reference structure and , the free energy profile captured by GNAP, would be sensitive to the choice of reference structures. This has been demonstrated in figure S1, which showed that free energy profile along GNAP would vary substantially as a function of reference structure. We note that this issue is not specifically limited to GNAP only. Rather, this is a generic scenario for any collective variable that depends on reference

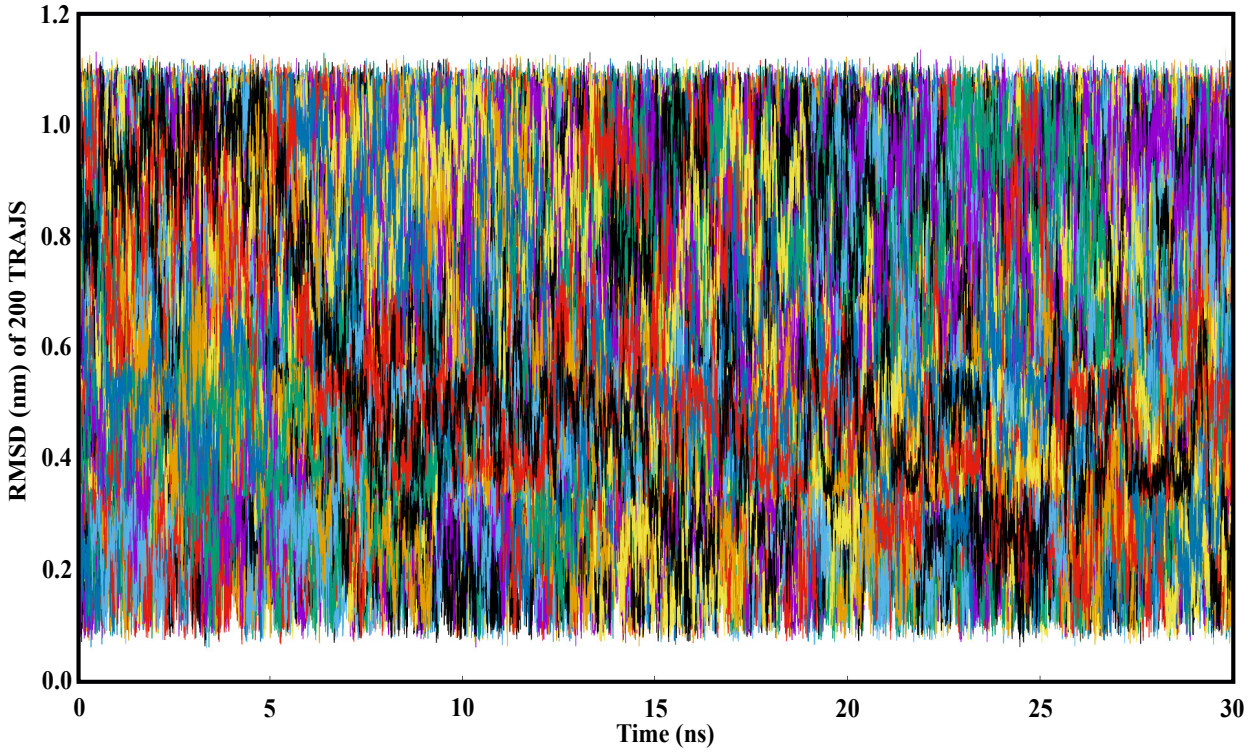

Figure S2: Mixing of rmsd for the 200 trajectories in case of the 32-bead polymer chain.

structure. Hence to circumvent this issue, one needs many different GNAP dimensions each based on the different reference structures. Each of these various GNAP dimensions would help in capturing some important characteristic features of the various energy basins in the free-energy landscape and, hence, ultimately help in reconstructing the entire energy landscape with a great accuracy. In principle, the greater the number of GNAP dimensions the greater is the accuracy with which the energy landscape is reconstructed. Effectively, this would need identifying different reference structures present in the conformational ensemble of the macromolecules. In this work, this identification of different reference structures has been done by exploring the RMSD of the structures and subsequently clustering the structures by RMSD. That in other words, in our work, RMSD appears only for deciding on a set of non-overlapping reference structures, as this is one of the most conventional ways to identify the key structures. It does not have any other role. In principle, one does not have to be limited in choice of RMSD for clustering metric.

Step-3: The multiple GNAP values relative to each of the representative structures are then calculated using a code that has been implemented as a Collective Variable in PLUMED a gromacs plugin software which helps in performing enhanced samplings. The code is now uploaded in GitHub.

Step-4: After obtaining the various GNAP dimensions, we go ahead to perform TICA over these raw GNAP dimensions. The idea here is to combine the features carried by each of the GNAP dimensions and project it on the each of the slowest TIC dimensions.

Step-5: Once we get the TIC dimensions, we take into account only those TIC dimensions (in slowest-to-fastest order) that, in totality, account for the 90% of kinetic variance. Thereafter, using these TIC dimensions we generate Markov State Model. We also studying the 2D free energy surface built out of the first two slowest relaxing TICs.

Demonstrating the Reproducibility of the Reference structures

Out of the 200 trajectories, we formed two sets of randomly chosen 100 trajectories. Performing a K-mean clustering over these two sets of trajectories based on RMSD values, we get two sets of 12 structures. Calculating the RMSD values of these two sets of 12 structures and arranging them in increasing order of their RMSD we get figure S3. This plot highlights a striking similarity of the two obtained sets.

Description of Supplemental Movies:

- Movie S1 demonstrates the non-affine modes responsible for transition of GB1 from beta hair-pin structure to the central helix structure.
- Movie S2 demonstrates the PCA models responsible for transition of GB1 from beta hair-pin structure to the central helix structure.

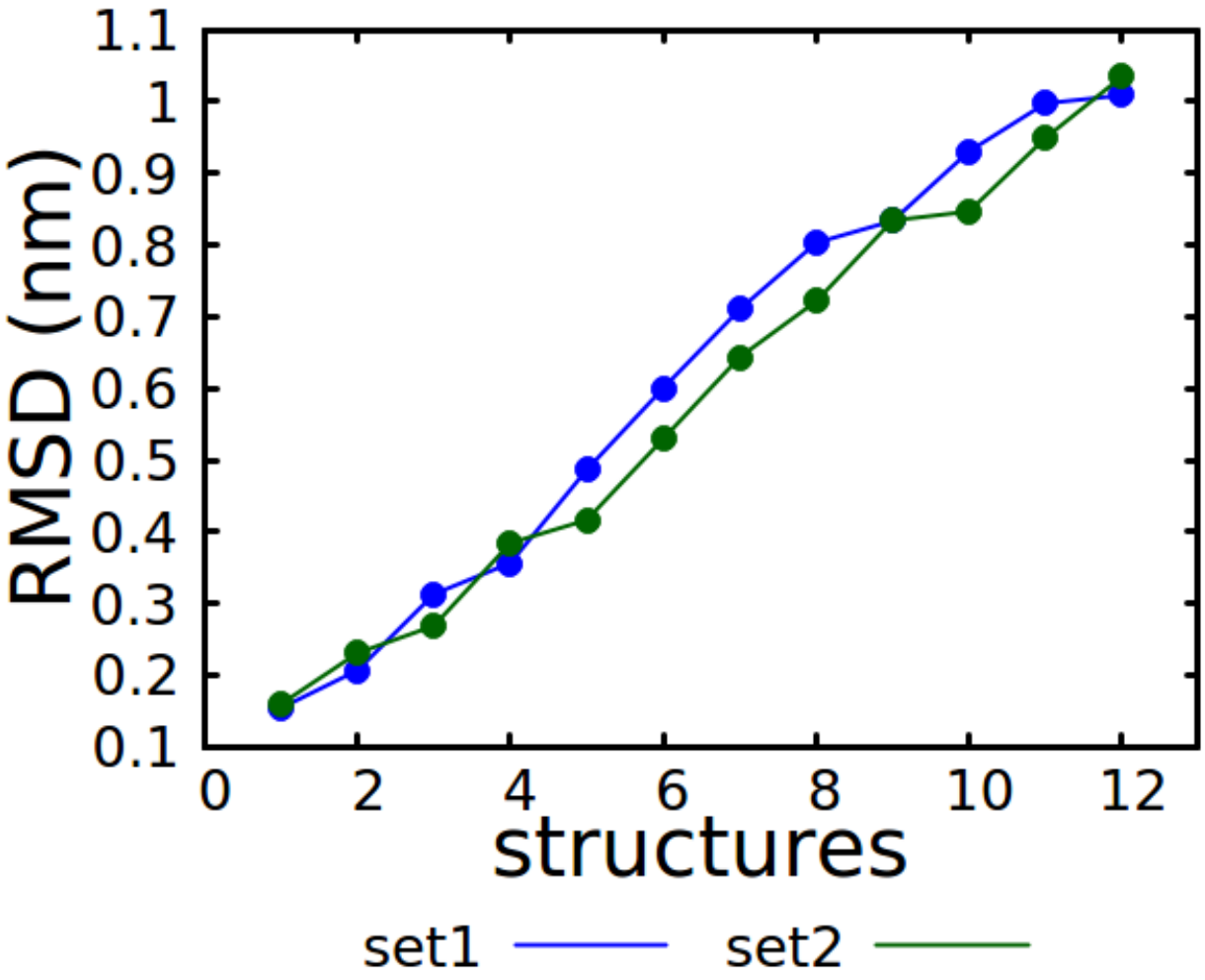

Figure S3: Reproducibility of the Reference structures