Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Structural and mechanistic insight into the substrate binding from the conformational dynamics in apo and substrate-bound DapE enzyme

Debodyuti Dutta, and Sabyashachi Mishra*

Department of Chemistry, Indian Institute of Technology Kharagpur, Kharagpur, India

E-mail: mishra@chem.iitkgp.ernet.in

*To whom correspondence should be addressed

Overall stabilization of the apo and enzyme-substrate complex

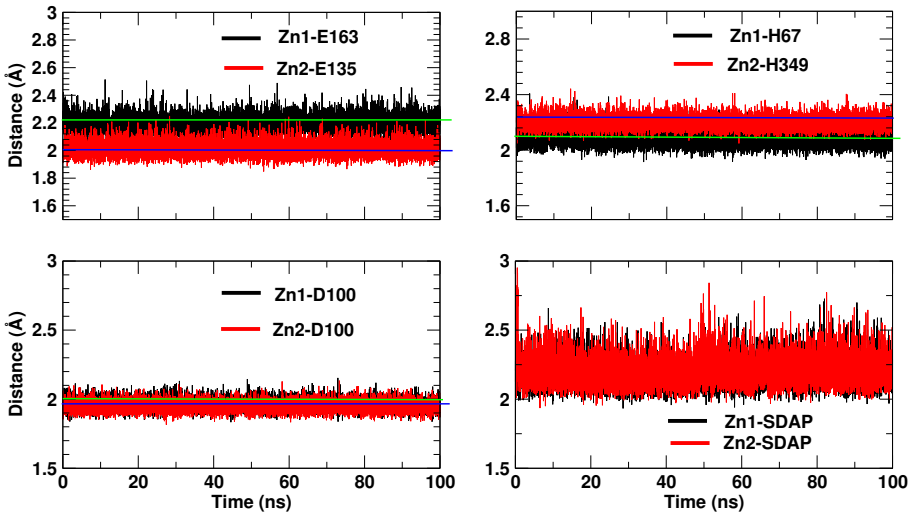

Figure S1: Distance of the active site residues from Zn atoms during the trajectory. Horizontal lines indicate the value of the same in the crystal structure of DapE.

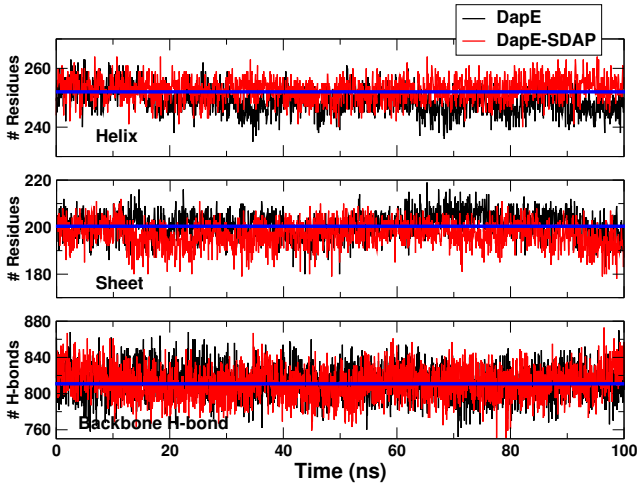

Figure S2: The stability of the enzyme native structure in the DapE and DapE-SDAP simulations shown by the number of residues that form helices (upper panel) and extended β-strands (middle panel) and number of backbone hydrogen bonds (lower panel) in both the systems. The horizontal lines indicate the corresponding values in the crystal structure.

Theoretical Background of Principal Component Analysis

The principal component analysis, also known as quasiharmonic analysis or essential dynamics method, is a well established technique to express the conformational dynamics in high dimensional complex systems in terms of a few principal modes or principal components of motion.[1,2] The first step of calculation of the principal components involves the calculation of the covariance matrix, **C**, of the positional deviations and the diagonalization of this matrix. The $3N$ dimensional covariance matrix **C** is calculated based on an ensemble of protein structures, and the elements of **C** are defined as[3]

$$C_{ij} = \langle (x_i - \langle x_i \rangle)(x_j - \langle x_j \rangle) \rangle, \tag{S1}$$

where $x_1, ..., x_{3N}$ are the Cartesian coordinates of an N-particle system and the brackets denote ensemble average. The diagonalization of the symmetric $3N \times 3N$ matrix **C** is equivalent to solving the eigenvalue problem[4,5]

$$\mathbf{R}^T \mathbf{C} \mathbf{R} = \lambda, \tag{S2}$$

where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_{3N}$ are the eigenvalues, and $\mathbf{R}^T$ is the transpose of **R**. The columns of **R** are the eigenvectors or the principal modes. The trajectory can be projected onto the eigenvectors to obtain the principal components[6] $q_i(i), i = 1, 2, ..., 3N$,

$$q = \mathbf{R}^T (x(t) - \langle x \rangle). \tag{S3}$$

The eigenvalues λ_i represent the mean-square fluctuation along the direction of the ith principal mode. The largest eigenvalue captures the largest fraction of the root-mean-square fluctuation (RMSF) of the overall dynamics followed by the smaller eigenvalues. The first few principal components typically describe the collective global motions of the system, while the principal components associated with smaller eigenvalues represent local motions largely dominated by random fluctuations.

To distinguish the actual protein motions from random fluctuations described by the dynamics of the first few principal components, the cosine contents (c_i) of these principal components were evaluated,[7]

$$c_i = \frac{2}{T} \frac{\left(\int_0^T \cos\left(\frac{i\pi t}{T}\right) q_i(t) \mathrm{d}t \right)^2}{\int_0^T {q_i}^2(t) \mathrm{d}t}. \tag{S4}$$

Based on the fact that the principal components of random diffusion are cosines with the number of periods equal to half the principal component index, the cosine content can take values between zero (no cosine) and 1 (a perfect cosine), representing the limits of relevant protein dynamics and random fluctuations, respectively.[8]

The eigenvectors corresponding to the larger eigenvalues, which can be used to describe the essential motions, tend to converge to a stable set in the nanosecond timescale, suggesting that MD simulations of nanosecond timescale can provide a reasonable definition of the essential subspace, valid well beyond the nanosecond range.[9–11] The convergence of the principal component analysis is further verified by determining the overlap of the sampling between the first and second half of the MD simulation of apo enzyme and enzyme-substrate complex. Since the elements of the covariance matrix are proportional to the square of the displacement, the extent of sampling can be accessed from the square root of the covariance matrix, which can be calculated from the eigenvalues λ_i and the eigenvectors (columns of the matrix $\mathbf{R}$).[6]

The overlap s between two parts of the trajectory is defined as,[6]

$$s(A,B) = 1 - \frac{d(A,B)}{\sqrt{\mathrm{tr}A + \mathrm{tr}B}}, \tag{S5}$$

where tr is the trace of a matrix and the difference d between covariance matrices A and B, associated with the two parts of the trajectory, are given by,[6]

$$d(A,B) = \sqrt{\mathrm{tr}\left(A^{1/2} - B^{1/2}\right)} \tag{S6}$$

$$= \left(\sum_i^N (\lambda_i^A + \lambda_i^B) - 2 \sum_i^N \sum_i^N \sqrt{\lambda_i^A \lambda_j^B} (R_i^A R_j^B)^2 \right)^{1/2}, \tag{S7}$$

with λ_i^A and λ_i^B as the ith eigenvalue of covariance matrices A and B, corresponding to the eigenvector columns R_i^A and R_j^B, respectively.

Convergence Tests of the Principal Component Analysis

Due to the large size of the biomolecular system together with the large scale conformational dynamics observed during 100 ns of MD simulation, verification of the convergence of the principal component analysis is very important. With increasing length of MD simulation, the conformational dynamics of the system is expected to converge. For the present study, we have tested this by evaluating principal components for increasing length of simulations. Table S1 shows that the eigenvalues of the first three principal components converge as the simulation length approaches 100 ns. For further verification of the convergence of the results, the cosine contents (Equation S4) of the principal components, which help distinguish the essential protein motion from the random fluctuations, are evaluated for increasing length of simulation (Figure S4). While the principal component 2 requires only a few nanoseconds of simulations to acquire a cosine content close to zero, the principal component 1 requires much longer time to do so due to its higher rate of fluctuation. For biological systems as large as the present one, the limit of convergence of the cosine contents are around 0.5,[8] which is achieved for the principal component 1 by the end of 100 ns of MD simulations in both apo enzyme as well as enzyme-substrate complex.

Table S1: The relative contribution of the first three principal components to the overall protein dynamics for increasing length of MD trajectory of DapE and DapE-SDAP complex.

Time	DaPE			DaPE-SDAP		
	PC 1	PC 2	PC 3	PC 1	PC 2	PC 3
0-20 ns	0.41	0.20	0.12	0.62	0.13	0.08
0-40 ns	0.53	0.16	0.11	0.57	0.12	0.10
0-60 ns	0.49	0.20	0.13	0.46	0.19	0.10
0-80 ns	0.43	0.27	0.12	0.52	0.18	0.09
0-100 ns	0.40	0.29	0.11	0.54	0.18	0.08

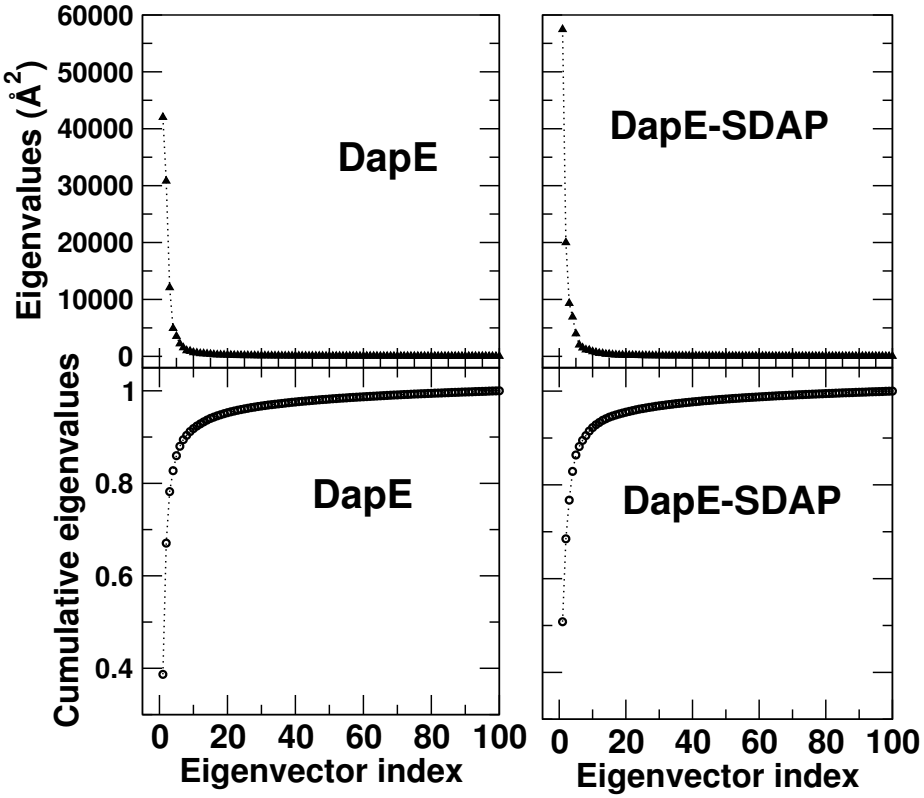

Figure S3: The eigenvalues of the covariance matrices obtained from the positional deviations of the enzyme backbone atoms in apo DapE and DapE-SDAP complex (upper panels) and the normalized cumulative eigenvalues of the two systems (lower panels). Only the first 100 eigenvalues are shown for clarity.

To evaluate the convergence of the phase-space sampling by the MD simulations, the overlap factors (Equation S5) are calculated from the two halves of the trajectory with gradually increasing the length of the MD trajectory. We see that when the first 20 ns of the trajectory is divided into two, the covariance matrices show least overlap (Figure S4c,d). The overlap gradually increases and is the highest (around 0.75) when the entire 100 ns MD trajectory is divided into two halves, indicating the degree of convergence of the phase space sampling during MD simulations.

The distribution of the displacement of the principal components gives a measure of randomness of the underlying dynamics.[12,13] Figure S5 shows the normalized distribution of the principal component 1 and principal component 2 and compares it with a relatively small amplitude component (principal component 50, chosen randomly). It is clear that the principal

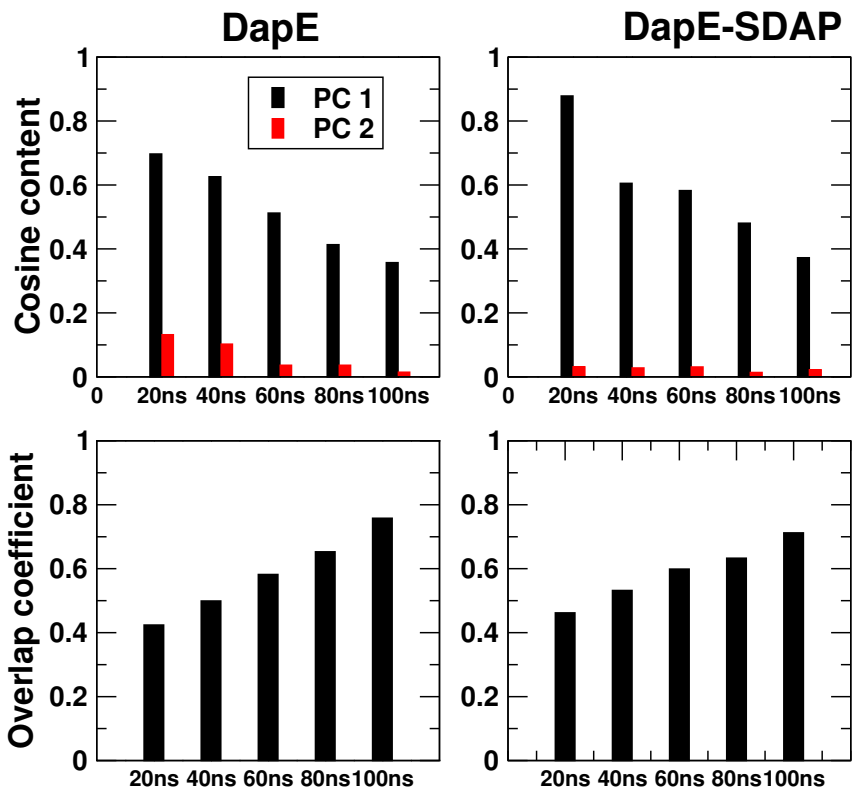

Figure S4: Cosine content analysis (Equation S4) of the principal components 1 and 2 in DapE and DapE-SDAP complex for increasing length of the MD trajectory (upper panels). The overlap coefficient (Equation S5) between two equal halves of the MD trajectory for increasing length of MD trajectory of DapE and DapE-SDAP complex (lower panels).

component 1 and principal component 2 for both apo enzyme and the enzyme-substrate complex show non-Gaussian distribution, emphasizing the fact that the dynamics of the systems along these eigenmodes sample a large number of configurations, which is a characteristic of global motion (Figure S5). The higher eigenmodes corresponding to smaller eigenvalues, on the other hand, exhibit Gaussian distributions signifying independent local fluctuations as seen for principal component 50 (Figure S5).

In addition to the dominating first two principal components, the higher principal components have also been analyzed. We found the principal components 3 and 4 in both apo enzyme and the enzyme-substrate complex show essential motion exhibiting relative motion of the catalytic domains and dimerization domains. However, owing to their small amplitudes obtained from the diagonalization of the covariance matrix, their contributions to the overall enzyme

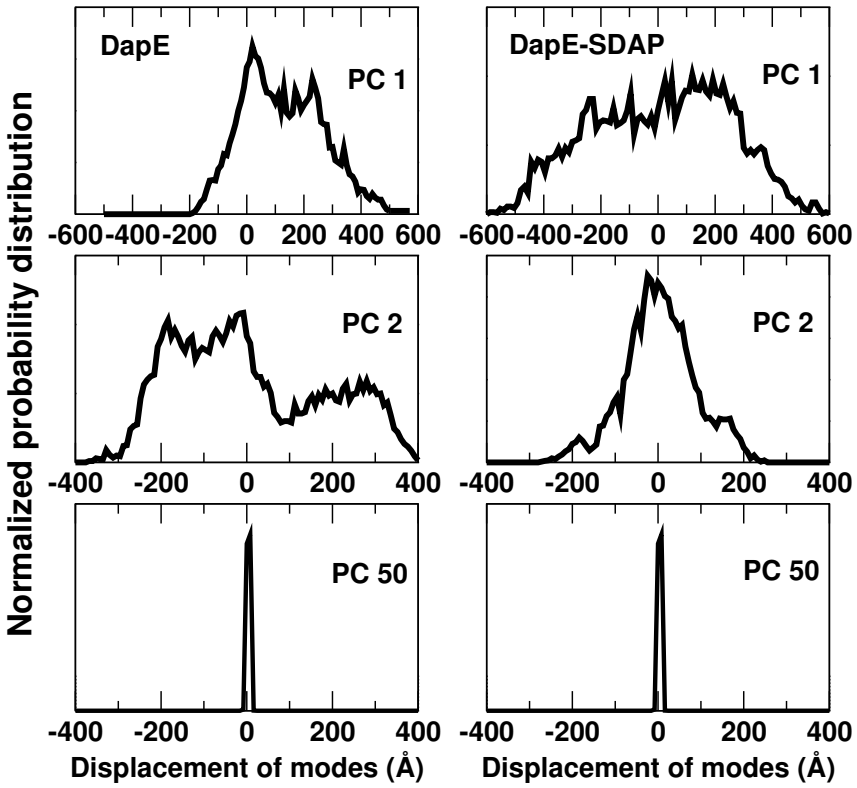

Figure S5: Normalized probability distribution of the principal components. While the distribution profiles for principal components 1 and 2 are non-Gaussian, the principal component 50 shows a Gaussian distribution signifying the random fluctuations.

conformational dynamics are expected to be marginal.

Characterization of the principal components

Characterization of the principal component 1

The time evolution of the distance between the centers of mass of the catalytic domain and dimerization domain in both chains of the apo enzyme and enzyme-substrate complex obtained from 100 ns of MD simulations are shown in the upper panels in Figure S6, while the middle panels show the distance between centers of mass of the catalytic domains and dimerization domains of the trajectory filtered along mode 1. The two data sets are strongly correlated, in particular, for the DapE-SDAP system with correlation coefficient of 0.73 for chain A and 0.66 for chain B, as compared to 0.56 and 0.43 for the chains A and B in the apo enzyme, respectively. The high values of correlations further underlines the predominant importance of mode 1 in the conformational dynamics of the enzyme-substrate complex. The striking similarities between the time evolution of the distance between the centers of mass of the catalytic domain and dimerization domain (middle panel in Figure S6) with the time series of the displacement of the principal component 1 (lower panel in Figure S6) reflects the fact that the choice of the internal coordinates (i.e., the distance between centers of mass of the catalytic domain and dimerization domain) indeed provides an accurate description of the dynamics along the principal component 1.

The time evolution of the distance between centers of mass of the catalytic domain and dimerization domain along mode 1, shows a decreasing distance between the catalytic and dimerization domains in the DapE-SDAP complex system, while the same is not true for the apo enzyme (middle panel in Figure S6). This differential correlation can be clearly seen from the cross-correlation matrix of C_α atomic fluctuations along principal component 1 (Figure S7). Highly positive regions (blue) are indicative of strong correlation in the movement of specific residues, whereas negative regions (red) are associated with strong anticorrelated motion of the residues. Along the diagonal, both the systems show highly correlated motions because of the correlation of a residue with itself. In Figure S7 only the extreme values of correlation or anti-correlation ($|correlation| > 0.9$) are shown for clarity. In the enzyme-substrate complex system,

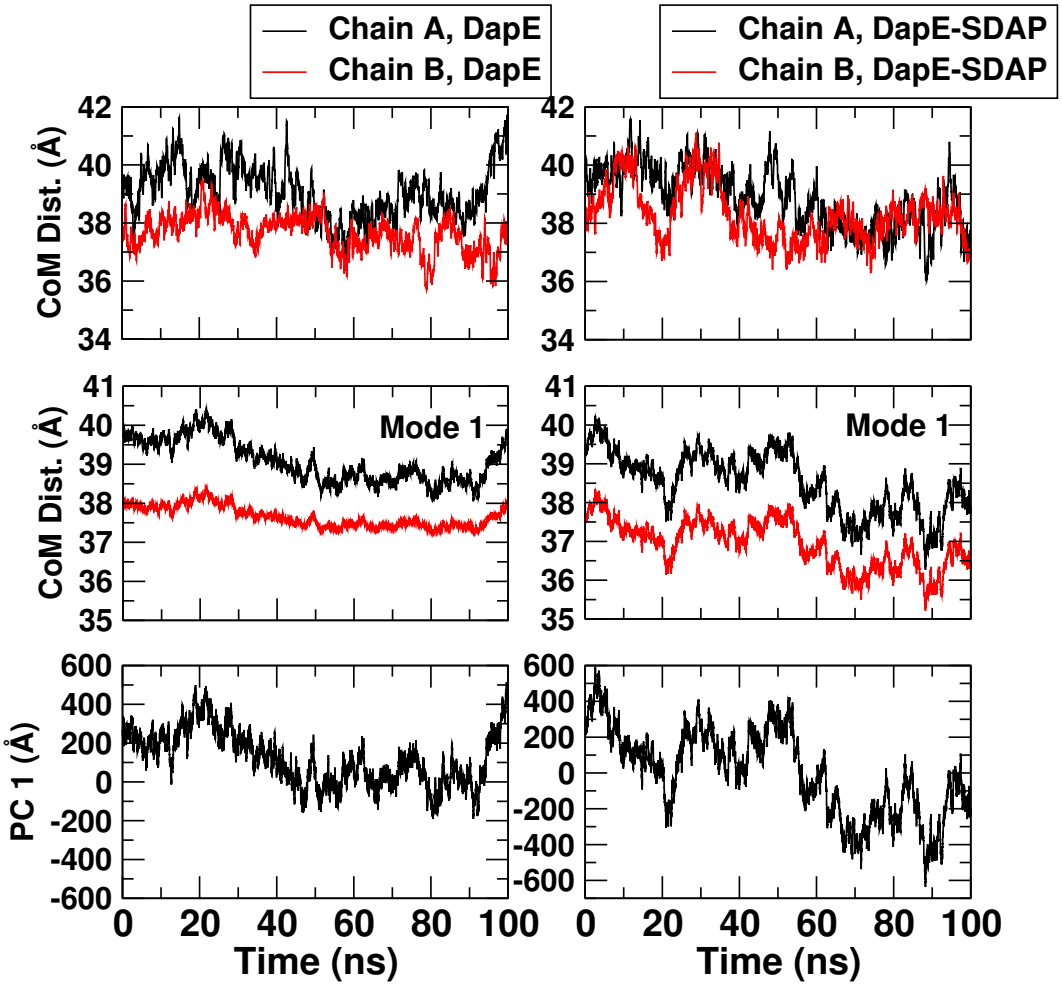

Figure S6: Distance between the centers of mass (CoM) of the catalytic domains and the dimerization domains in chain A and chain B of DapE and DapE-SDAP complex along the actual MD trajectory (upper panels), the trajectory filtered along principal component 1 (middle panels), and the displacement of the principal component 1 (lower panels).

the motion of the two catalytic domains exhibit more anticorrelation compared to that in the case of apo enzyme. This indicates the fact that in the DapE-SDAP systems, the two catalytic domains exhibit motions in opposite directions. Together with the observations from Figure 3 in the manuscript, the essential dynamics along the principal component 1 in the enzyme-substrate complex can be inferred to involve a folding of the catalytic domains onto the dimerization domains, which is further supported from the radius of gyration of the entire protein along the principal component 1 in of both the systems (Figure 5a of the manuscript). The decreasing radius of gyration in the DapE-SDAP complex along principal component 1 indicates the increase

of the compactness of the enzyme-substrate complex, compared to the apo enzyme.

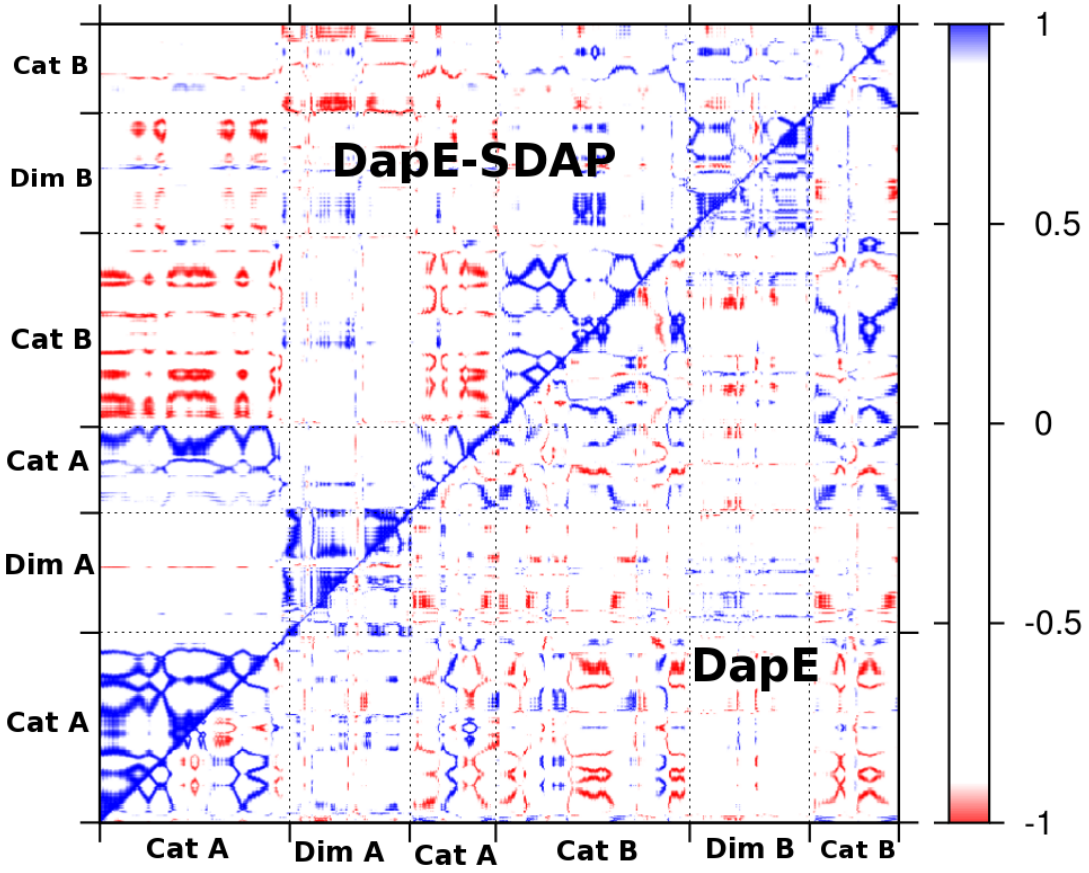

Figure S7: The C_α-C_α cross-correlation map of the trajectory filtered along principal component 1 for DapE (lower diagonal) and DapE-SDAP complex (upper diagonal).

Characterization of the principal component 2 The time evolution of the dihedral angle between the four centers of mass consisting of the catalytic domain of chain A, the dimerization domain of chain A, the dimerization domain of chain B and the catalytic domain of chain B of the apo enzyme and enzyme-substrate complex obtained from 100 ns of MD simulations are shown in the upper panels in Figure S8, while the middle panels show the same dihedral angle along principal component 2. The fluctuations in the dihedral angle are higher in case of the apo enzyme as compared to the enzyme-substrate complex, where principal component 1 is highly dominating in accounting for the enzyme conformational dynamics. Moreover, the correlation coefficients for the change in dihedral for the original trajectory and the trajectory filtered along mode 2 are 0.72 and 0.14 for DapE and DapE-SDAP, respectively. The displace-

ment along the principal component 2 in both apo enzyme and enzyme-substrate complex show excellent agreement with the time evolution of the dihedral angle along the principal component 2, suggesting the accurate description of the principal component 2 by the chosen internal coordinate.

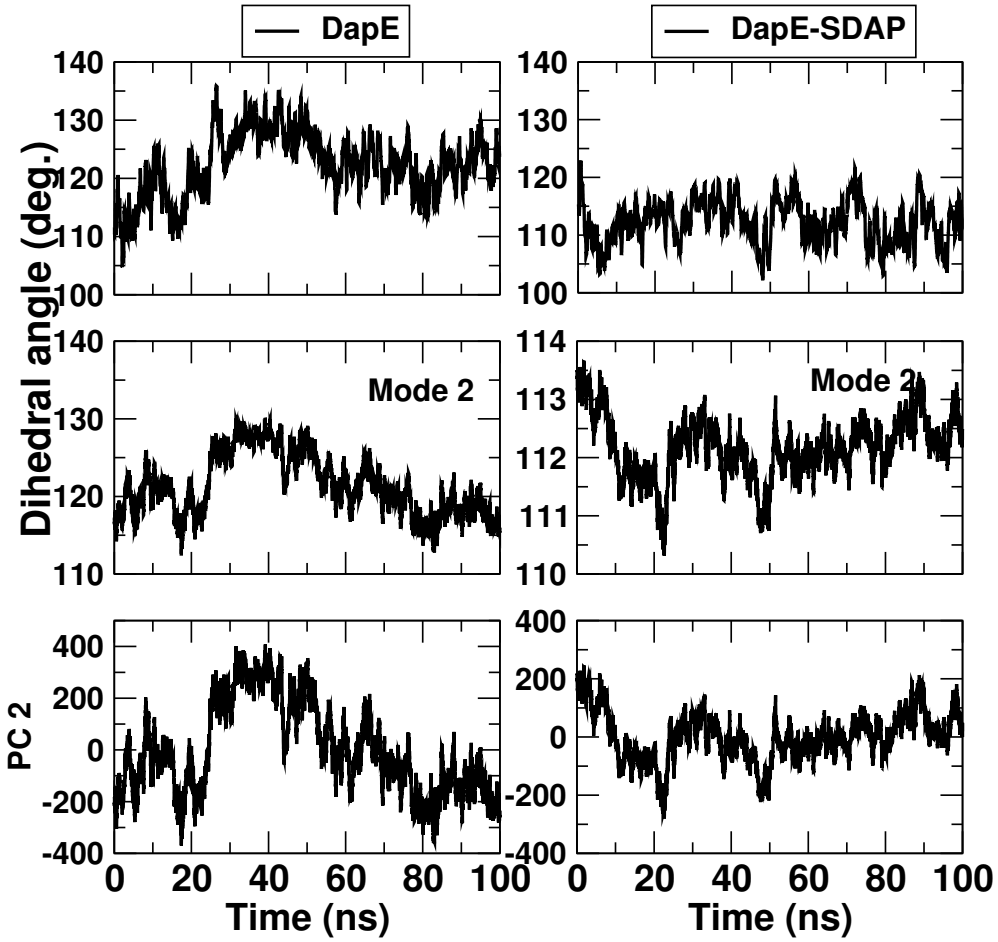

Figure S8: Dihedral angle between the centers of mass of catalytic domain (chain A), dimerization domain (chain A), dimerization domain (chain B) and catalytic domain (chain B) along the actual MD trajectory (upper panels), the trajectory filtered along the principal component 2 (middle panels), and the displacement of the principal component 2 (lower panels) for DapE and DapE-SDAP complex.

In addition to the dominating first two principal components, the higher principal components have also been analyzed. We found the principal components 3 and 4 in both apo enzyme and the enzyme-substrate complex show essential motion exhibiting relative motion of the catalytic domains and dimerization domains. However, owing to their small amplitudes obtained

from the diagonalization of the covariance matrix, their contributions to the overall enzyme conformational dynamics are expected to be marginal.

The distribution of the eigenvalues and the description of the first two principal components are found to be similar when the covariance matrix is constructed from the backbone atoms of the protein (data not shown). This underlines the fact that the covariance analysis generally gives an idea about the global motion of the system, where the backbone atoms play primary role. Although, the side-chains of some of the residues are important in substrate binding, many side chains undergo Gaussian type fluctuations with very small amplitude and therefore contribute marginally to the large time scale overall protein dynamics.

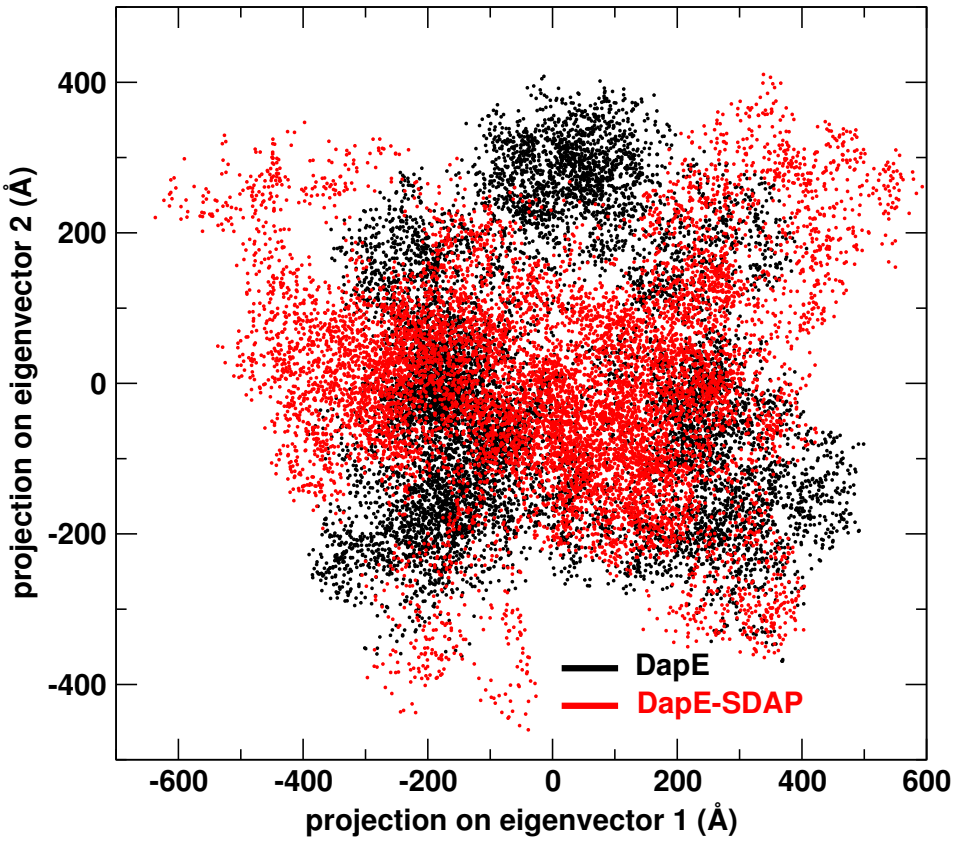

Figure S9: Distribution of the planes spanned by the first two modes for DapE and DapE-SDAP systems.

The combined displacements along the first two principal components provide information about the various conformational states the systems visit. The corresponding two-dimensional scatter plot distribution along the first two principal components (Figure S9) represents the tra-

jectories projected on two planes, defined by two eigenvectors for apo and complex systems. While it may be difficult to isolate the representative structures, it can be readily seen that during the 100 ns simulation, both the DapE and DapE-SDAP systems visit a wide range of conformations. Along eigenvector 1, DapE-SDAP spans a wider range of conformations compared to DapE, whereas along eigenvector 2, the conformational dynamics in DapE is wider than the DapE-SDAP complex.

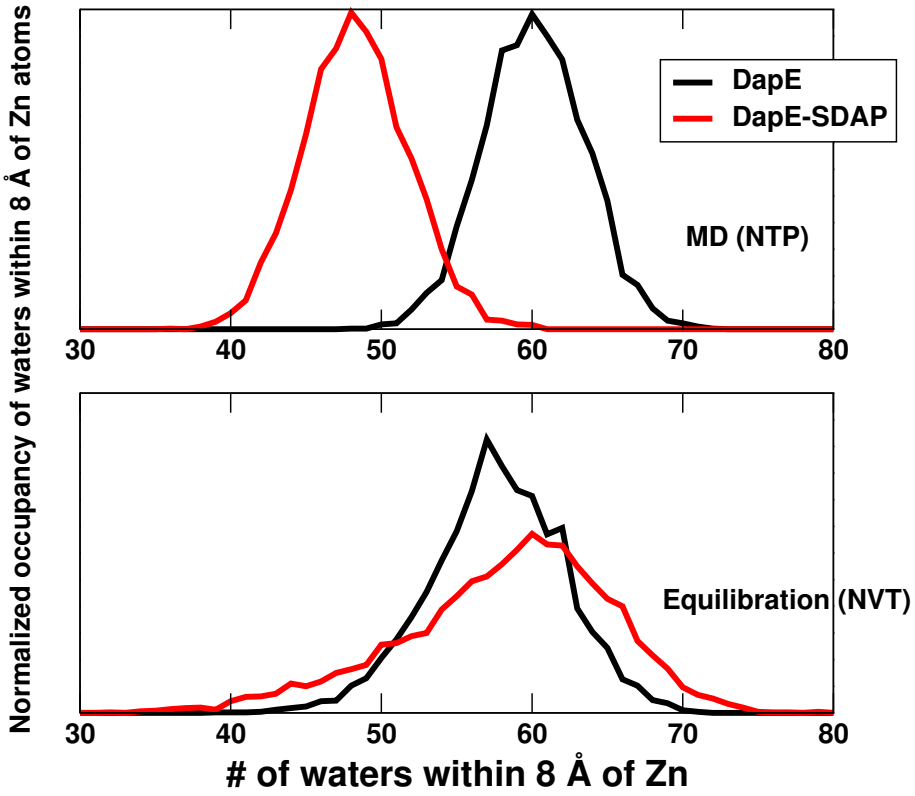

Figure S10: The number of water molecules within 8 Å of Zn atoms in apo enzyme and enzyme-substrate complex during the MD trajectory (upper panel) and equilibration (lower panel). The space occupied by the substrate is not considered in this figure. In this figure, 8 Å radius around the metal ions in both DapE and DapE-SDAP systems is considered to calculate the number of water molecules near the active site. In the complex, the substrate SDAP falls within this radius. However, the volume occupied by the substrate can account for at the most a total of six water molecules (even if we assume perfect packing of water molecules). Additionally, since SDAP is present within the protein matrix, it is unlikely that the region vacated by the substrate will be entirely occupied by water molecules. Hence the difference in the number of water molecules must be the direct consequence of the types of motion of the catalytic domains that are described in the manuscript. This is further verified, when one compares the results during the equilibration (lower panel), where the number of water molecules with 8 Å of the metal center remains nearly same for the apo and complex system.

Table S2: Residue-wise decomposition of the binding free energy (kcal/mol) when the ligand is SDAP, or Zn1, or Zn2. During SDAP binding, most of the active site residues (connected to metal centers) exhibit positive ΔG, while all the hydrogen-bonding partners of SDAP and the few dimeric-domain residues that have been mentioned in the manuscript, exhibit negative ΔG. The positive ΔG of the active site residues results from the electrostatic repulsion of the negatively charged ends of the active site residues (which are connected to Zn) and that of SDAP, which arises due to the fact that in MM-PBSA an infinite cut-off of electrostatics has been used. The active site residues, on the other hand, strongly stabilize the binding of metal centers (columns 3 and 4) which in turn stabilize the binding of substrate SDAP (column 2).

Residue	**ligand SDAP**	**ligand Zn1**	**ligand Zn2**
HSD 67	0.82	-38.03	-4.93
ASP 100	4.48	-94.26	-89.43
GLU 135	3.15	-29.57	-103.19
GLU 163	2.45	-115.73	-21.29
HSD 349	-0.03	-3.51	-33.41
GLU 134	-0.86	-13.12	-5.99
LYS 175	-0.74	0.59	0.40
ARG 178	-0.89	0.63	0.48
GLY 322	-0.09	0.06	0.02
GLY 323	-0.31	0.54	0.20
GLY 324	-1.26	0.83	0.38
THR 325	-3.30	1.00	0.35
ASP 327	-0.45	-1.38	-0.93
ARG 329	-1.32	2.70	1.23
ASN 345	-0.16	0.34	0.15
LYS 350	-0.42	-0.18	-0.30
ARG 258	-0.52	0.28	0.31
THR 261	-0.06	-0.00	-0.00
SER 290	-0.76	-0.01	-0.01
SER 181	-0.07	0.01	0.01
PRO 293	-0.08	0.04	0.02
TYR 259	0.08	-0.02	-0.02
Zn1	-4.47	33.80	86.83
Zn2	-5.15	83.43	36.63
SDAP	-58.988	-108.66	-101.82
ΔG_{total}	-66	-282	-238

Table S3: Hydrogen bonds summary between SDAP and DapE in chain A and chain B during 100 ns MD. The numbers in the parentheses represent the hydrogen bonds summary along the principal component 1.

H-bond summary of LL-SDAP			
Residues	%Occupancy	⟨lifetime⟩ (ps)	#events
Glu134	38.7 (33.6)	15.28 (14.86)	2532 (2240)
Lys175	11.2 (8.9)	3.95 (3.90)	2837 (2282)
Arg178	73.6 (61.3)	91.4 (90.1)	805 (680)
Thr325	78.6 (71.9)	23.46 (22.91)	3350 (3139)
Asp327	74.6 (48.3)	18.36(17.02)	4030 (2837)
Arg329	19.8 (15.8)	22.76 (21.94)	870 (720)
Asn345	42.5 (38.0)	106.25 (104.1)	400 (365)
Lys350	5.8 (0)	6.10	950

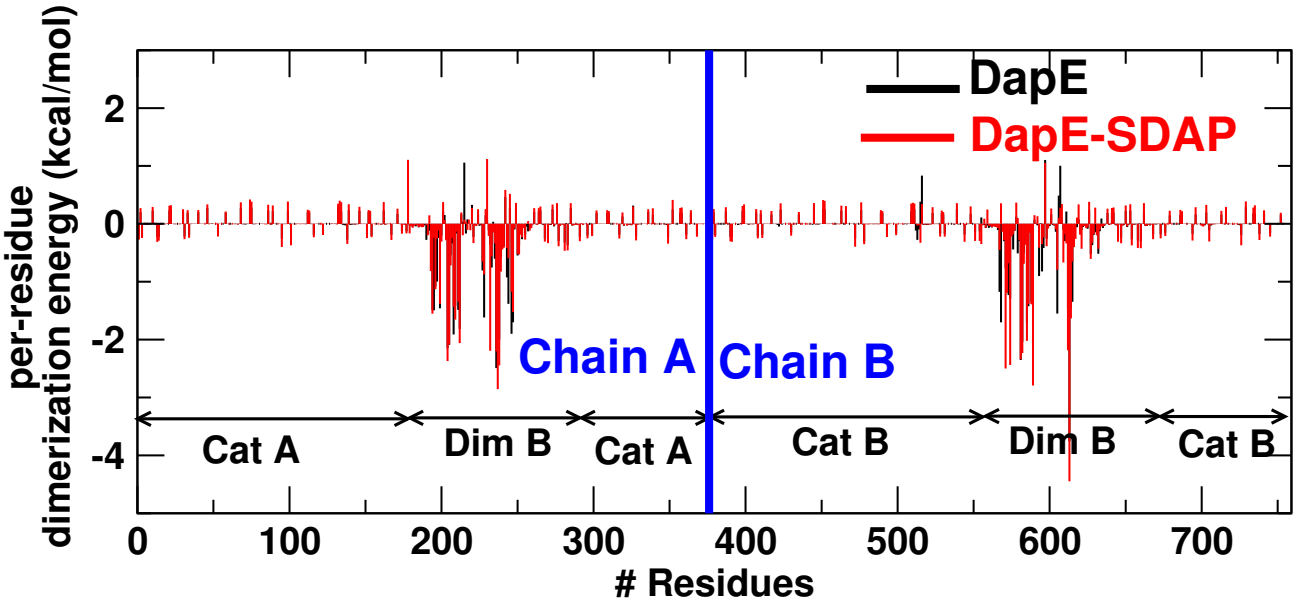

Figure S11: Residue-wise decomposition of the total dimerization free energy (the free energy of formation of the dimer DapE) obtained from MM-PBSA analysis of apo-DapE MD trajectory (black) and DapE-SDAP MD trajectory (red). The two chains (monomers) are separated by the vertical blue line and the domains within a monomer are shown as horizontal arrows.

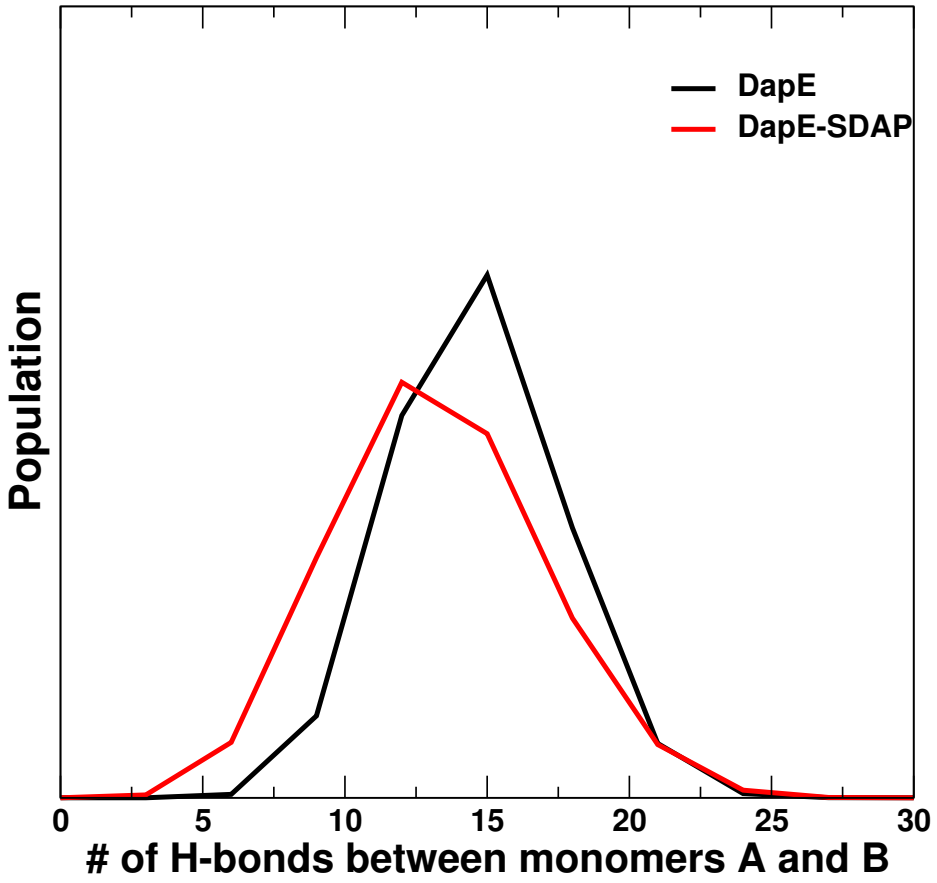

Figure S12: Distribution of the hydrogen bonds between the two chains of DapE during apo-DapE and DapE-SDAP MD simulations, shown in black and red, respectively.

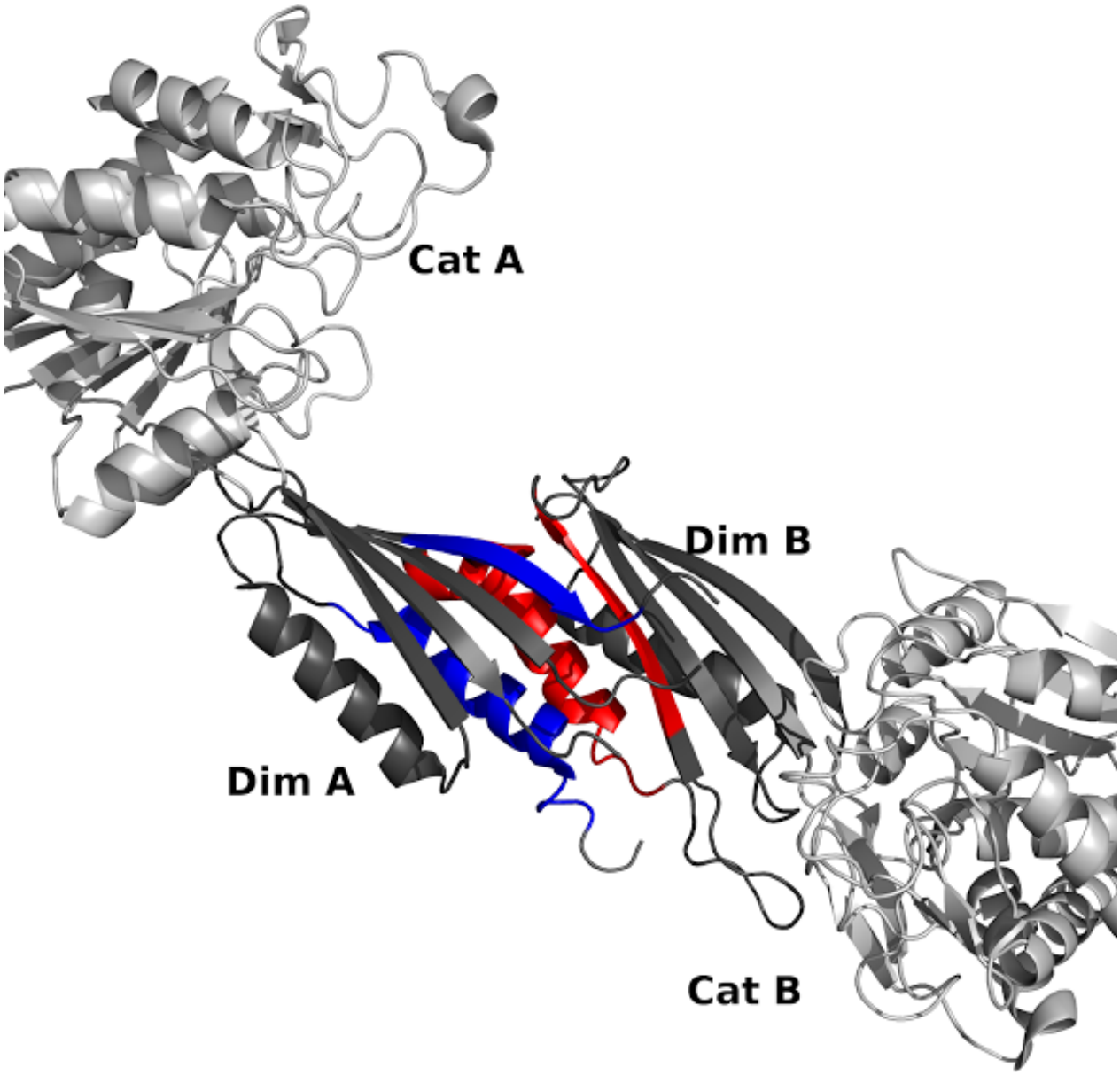

Figure S13: The contact between the two monomers of the dimeric DapE is primarily between the $\alpha 5$ helices and $\beta 8$ strands of both the monomers. The catalytic domains are shown in gray, the dimerization domains are shown in dark gray, and the $\alpha 5$ helices and *beta*8 strands are shown in blue and red for the chain A and chain B, respectively.

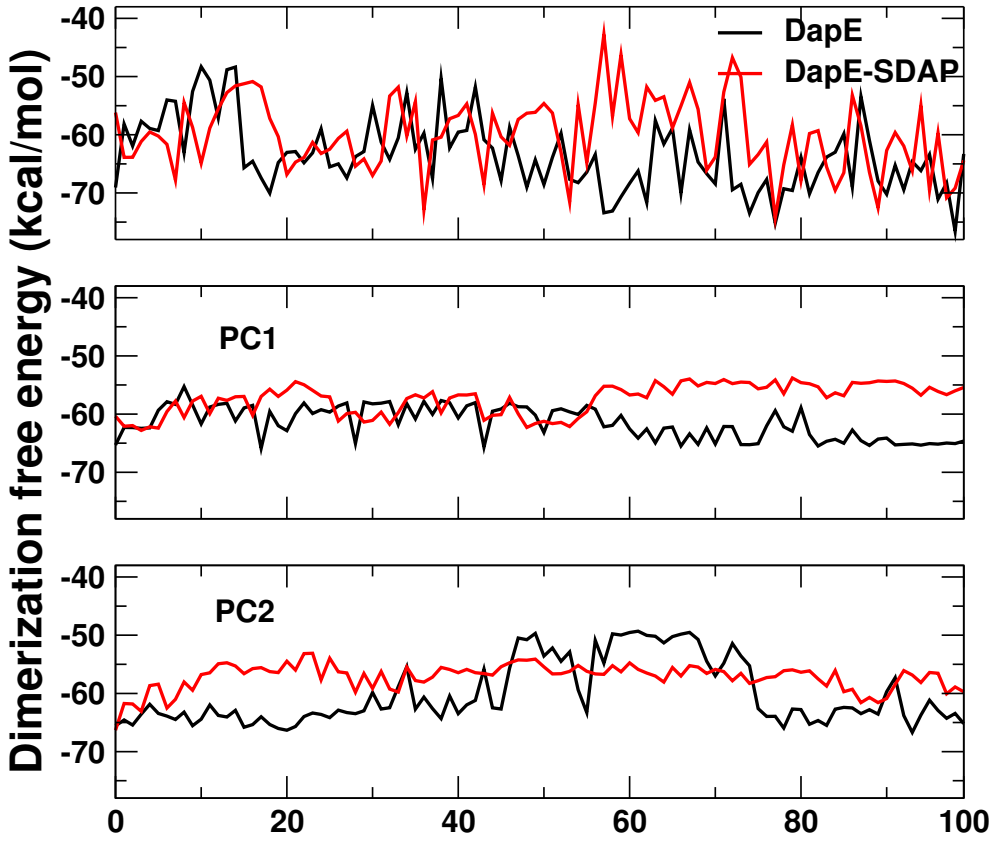

Figure S14: The free energy of formation of the dimer DapE (the dimerization free energy) during the 100 ns MD simulation (upper panel). The dimerization free energy from the trajectories filtered along the principal components 1 and 2, middle and lower panels, respectively. The average dimerization energy for the MD trajectory, PC 1, and PC 2, for the apo-DapE are estimated as -63, -61, -61 kcal/mol, respectively and the same for the SDAP-DapE complex are obtained as -60, -57, -57 kcal/mol, respectively.

References

(1) Kitao, A.; Hirata, F.; Go, N. *Chem. Phys.* **1991**, *158*, 447–472.

(2) Garcia, A. *Phys. Rev. Lett.* **1992**, *68*, 2696–2699.

(3) Levy, R. M.; Srinivasan, A.; Olson, W.; McCammon, J. *Biopolymers* **1984**, *23*, 1099–1122.

(4) Hayward, S.; Go, N. *Annu. Rev. Phys. Chem.* **1995**, *46*, 223–250.

(5) Tournier, A. L.; Smith, J. C. *Phys. Rev. Lett.* **2003**, *91*, 2081061–2081064.

(6) Hess, B. *Phys. Rev. E* **2002**, *65*, 319101 – 319110.

(7) Hess, B. *Phys. Rev. E* **2000**, *62*, 4838–8448.

(8) Maisuradze, G. G.; Liwo, A.; Scherga, H. A. *J. Mol. Biol.* **2009**, *385(1)*, 312–329.

(9) Van Aalten, D. M. F.; Groot, B. L. D.; Findlay, J. B. C.; Berendsen, H. J. C.; Amadei, A. *J. Comput. Chem.* **1997**, *18*, 169–181.

(10) Arcangeli, C.; Bizzarri, A. R.; Cannistraro, S. *Biophys. Chem.* **2001**, *90*, 45–56.

(11) Kubitzki, M. B.; de Groot., B. L. *Biophys. J.* **2007**, *92*, 4262–4270.

(12) Amadei, A.; Linssen, A. B. M.; Berendsen, H. J. C. *Proteins: Structure, Function, and Bioinformatics* **1993**, *17*, 412.

(13) Laberge, M.; Yonetani, T. *Biophys. J.* **2008**, *94*, 2737–2751.