All-atom modeling of anisotropic atomic fluctuations in protein crystal structures

Jeffrey Hafner and Wenjun Zheng[a)]
Physics Department, University at Buffalo, Buffalo, New York 14260, USA

(Received 17 June 2011; accepted 14 September 2011; published online 13 October 2011)

The accurate modeling of protein dynamics in crystalline states is essential for the development of computational techniques for simulating protein dynamics under physiological conditions. Following a previous coarse-grained modeling study of atomic fluctuations in protein crystal structures, we have refined our modeling with all-atom representation and force field. We have calculated the anisotropic atomic fluctuations of a protein structure interacting with its crystalline environment either explicitly (by including neighboring proteins into modeling) or implicitly (by adding harmonic restraints to surface atoms involved in crystal contacts). The modeling results are assessed in comparison with the experimental anisotropic displacement parameters (ADP) determined by X-ray crystallography. For a list of 40 high-resolution protein crystal structures, we have found that the optimal modeling of ADPs is achieved when the protein-environment interactions are much weaker than the internal interactions within a protein structure. Therefore, the intrinsic dynamics of a protein structure is only weakly perturbed by crystal packing. We have also found no noticeable improvement in the accuracy of ADP modeling by using all-atom over coarse-grained representation and force field, which justifies the use of coarse-grained modeling to investigate protein dynamics with both efficiency and accuracy. © *2011 American Institute of Physics*. [doi:10.1063/1.3646312]

I. INTRODUCTION

Protein structural dynamics (such as atomic fluctuations of a protein in an equilibrium state), is widely recognized as a key player in protein structure-function relationship.[1] X-ray crystallography and nuclear magnetic resonance[2] are two primary sources of protein dynamics data at atomic or near-atomic resolution. In X-ray crystallography, atomic fluctuations of protein crystal structures are routinely quantified by the isotropic temperature factors (or B factors), which use an isotropic Gaussian distribution to characterize the spread of electron density of individual atoms.[3] The isotropic assumption is inadequate for large protein complexes which possess highly flexible structural components undergoing orientation-specific (anisotropic) motions. Anisotropic thermal parameters are needed to properly represent those motions in structural refinement. In recent years, a growing number of high-resolution protein crystal structures have been refined using anisotropic Gaussian distributions, which characterize atomic fluctuations by a symmetric tensor with six independent elements named anisotropic displacement parameters (ADPs).[4,5] Unlike the isotropic B factors which depict atoms as isotropic spheres, the ADPs describe atoms as anisotropic ellipsoids with the information of both magnitude and direction of atomic fluctuations included. So they offer much more details of protein structural dynamics in crystalline states, which is inherently anisotropic and anharmonic[6] (for an alternative way to improve structural refinement beyond isotropic Gaussian distributions, see Ref. 7).

To explore the fine details of protein structural dynamics, a range of computational methods from all-atom molecular dynamics simulation[8] to coarse-grained modeling techniques[9] have been developed and utilized. To quantify the dynamic contributions to crystallographic B factors and ADPs, low-frequency normal modes, which are known to capture large-amplitude collective motions of protein structures, have been used to fit B factors[10] and refine ADPs.[11] Initially, the normal modes were solved from all-atom force fields following energy minimization.[12–14] This procedure is, however, computationally expensive for large protein structures, and it is susceptible to structural distortions caused by energy minimization. More recently, elastic network models (ENM), including anisotropic network model (ANM) (Refs. 15–17) and its isotropic counterpart—Gaussian network model (GNM),[18,19] have been used instead of all-atom force fields to calculate normal modes. The ENM is typically constructed using a C_α-only structural representation, where neighboring C_α atoms are connected by Hookean springs with a uniform[15] or distance-dependent[20,21] force constant. Such simplification allows the coarse-grained normal modes to be calculated efficiently without the need for energy minimization. The collective motions described by low-frequency modes are largely unchanged by the use of coarse-grained representation and harmonic potential function.[22,23] The crystallographic B factors have been modeled by GNM,[24,25] ANM,[26] and their generalization[27] with notable success. ANM has also been used to model ADPs with reasonable success.[28–30] However, in these early studies, an isolated protein structure was modeled without considering its interactions with crystal environment. Therefore, the contributions

a)Author to whom correspondence should be addressed. Electronic mail: wjzheng@buffalo.edu.

of global rotations and translations (corresponding to six zero modes) could not be properly modeled.[31]

To deduce protein dynamics under physiological conditions from crystallographic data, it is imperative to relate the protein dynamics in crystalline states to the protein dynamics in solution. Early studies found significant effect of crystal packing on the atomic fluctuations of protein structures.[5,32–34] To accurately model the protein dynamics in crystalline states, recent efforts were made to simulate crystal packing effect in various ways, for example, by using GNM to model a protein structure together with its neighboring molecules,[35] or by using ENM coupled with rigorous treatments of boundary conditions and lattice vibrations.[36,37] In a recent study,[38] we adopted a different strategy to model crystal packing: a single protein molecule was selected as the main protein structure, while the rest of protein crystal was treated as its environment; the environment was then truncated based on the distance to the main protein structure or by keeping the nearest and next nearest neighbors of the main protein structure. The entire protein-environment system was modeled using three different ENM schemes, including ANM,[17] distance network model,[28] and a C_α-based ENM proposed by Hinsen *et al*.[20] Three different boundary conditions[38] (fixed, free, and buffered environment) were considered to account for the flexibility of environment to different extent. The dynamic effects of crystal packing were explored by varying the strength of protein-environment interactions relative to the intra-protein interactions. We performed ADP modeling for a list of 83 high-resolution crystal structures previously studied.[28,36,38] We found that the optimal modeling of ADPs was achieved when the protein-environment interactions are much weaker than the interactions within the main protein structure, which may be attributed to several causes (such as loose packing, large solvent screening for residues involved in crystal contacts[38]). As a result, the crystallographic ADPs and B factors are dominated by contributions from rigid-body motions of the main protein structure, and the internal protein dynamics is only weakly perturbed by crystal packing.

In this study, we will further refine our coarse-grained modeling of protein-environment systems with all-atom representation and force field. Our goal is to answer the following two key questions: (1) Does the above finding of weak protein-environment interactions still hold after the all-atom refinement? (2) Can we further improve ADP modeling by using all-atom representation and force field?

II. METHODS

A. All-atom modeling of protein-environment system

1. Explicit environment modeling (EEM)

To explicitly model the crystal packing effect while keeping a minimal system size, we construct an all-atom system consisting of a main protein structure and its environment comprised of the neighboring proteins that are within a minimal distance of 4.5 Å between heavy atoms (for an example, see Fig. 1). The atomic coordinates of the neighboring proteins are generated using the crystallographic symmetry transformations given by REMARK 290 and the SCALEN records of protein data bank (PDB) files.

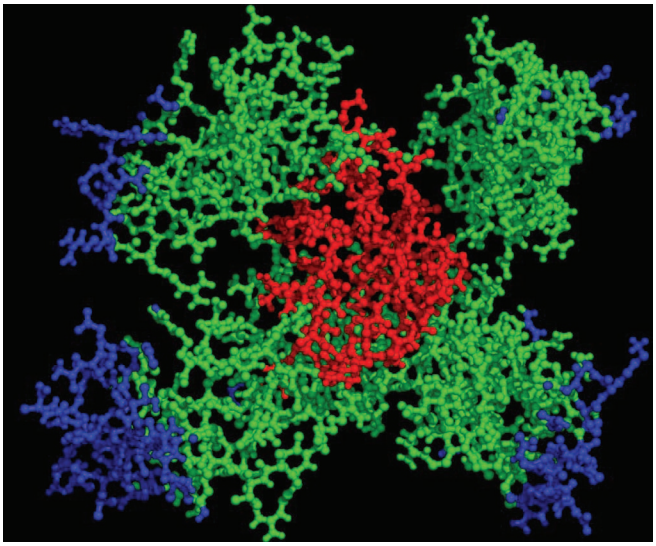

FIG. 1. Explicit modeling of crystalline environment for a crystal structure of pancreatic trypsin inhibitor (PDB code: 1g6x). The main protein structure, the residues in neighboring proteins within $r_c = 20$ Å from the main protein (E_1), and the rest of neighboring proteins (E_2) are colored red, green, and blue, respectively.[46]

The environment is further divided into two parts (E_1 and E_2, see Fig. 1). E_1 includes those residues of the neighboring proteins whose heavy atoms are within a cutoff distance r_c (varying from 0 to 20 Å) to the main protein's heavy atoms; the rest is called E_2. To properly model the flexibility of crystalline environment, we fix the atoms of E_2 while allowing the atoms of E_1 to move together with the main protein.

The total potential energy is minimized, initially by multiple 50-step steepest descent minimizations with first backbone atoms and then C_α atoms harmonically restrained, followed by 150 000 steps of unrestrained minimization using the adopted basis Newton-Raphson (ABNR) algorithm. Additional 10 000 steps of minimization by ABNR may be executed until the resulting Hessian matrix becomes positive semi-definite.

The Hessian matrix (comprised of second derivatives of potential energy) is calculated for the minimized system using the VIBRAN module of CHARMM program.

2. Implicit environment modeling (IEM)

To implicitly model the crystal packing effect in a protein crystal, we construct an all-atom system consisting of a main protein structure alone, and then apply harmonic positional restraints to its surface heavy atoms in contact with neighboring proteins (defined as the heavy atoms of the main protein within a minimal distance of 4.5 Å from the heavy atoms of neighboring proteins).

The main protein structure is minimized in the same manner as explicit environment modeling (EEM) but in the absence of neighboring proteins. Then the Hessian matrix is calculated for the minimized system using the VIBRAN module of CHARMM program.

The introduction of the harmonic restraints results in the addition of $K_E N_i$ to the diagonal elements of the ith 3×3 diagonal super block of the hessian matrix, where N_i is the number of heavy atoms of neighboring proteins within 4.5 Å from atom i of the main protein, and K_E is the force constant of a harmonic restraint. K_E is tuned to adjust the strength of protein-environment interactions and optimize the fitting of ADP data (see below).

The missing residues are modeled using the MODLOOP web server.[39] The CHARMM program[40] is used for energy minimizations and calculations of Hessian matrices. Energy minimization causes structural changes of the main protein with small root mean squared deviation (RMSD) (see Table I). For all-atom force field, we use an effective energy function combining the CHARMM 19 polar hydrogen energy function with an excluded volume implicit solvation model.[41] Our choice of the force field follows a previous study.[28] Its advantages are twofold: first, it gives smaller system size than the more standard CHARMM22 force filed because of the use of united hydrogen atoms; second, it gives reasonable account of solvent environment in protein crystals.

B. Coarse-grained modeling of protein-environment system

Following a previous study,[38] we construct a C_α-only ENM that consists of the following two components: first, the C_α atoms of a main protein structure (corresponding to the asymmetric unit of a protein crystal); second, the C_α atoms of neighboring protein molecules as environment. To reduce computing cost, the environment atoms are fixed in space. Other boundary conditions were also studied in our previous study.[38]

The potential energy of the two-component ENM is

$$E = \frac{1}{2}\sum_{i<j} C_{ij}(d_{ij} - d_{ij,0})^2 + \frac{1}{2} f_{env} \sum_{i,I} C_{iI}(d_{iI} - d_{iI,0})^2, \tag{1}$$

where i and j (I) are indices for C_α atoms in the main protein (environment). d_{ij} and d_{iI} are C_α–C_α atomic distances. $d_{ij,0}$ and $d_{iI,0}$ are the values of d_{ij} and d_{iI} given by the crystal structure. A new parameter f_{env} within the range [0, 1] is introduced to tune the strength of protein-environment interactions relative to intra-protein interactions. $f_{env} = 0$ corresponds to an isolated protein structure. $f_{env} = 1$ corresponds to equal strength between protein-environment and intra-protein interactions. C_{ij} or C_{iI} represents spring force constant: $C_{ij} = \begin{cases} 1, & \text{if } d_{ij,0} < R_c \\ 0, & \text{otherwise} \end{cases}$, where R_c is the cutoff distance. We use $R_c = 10$ Å because it was found to give optimal modeling of the ADPs.[38]

We then calculate a coarse-grained hessian matrix using the above ENM potential energy.

C. Calculation of ADP

The atomic mean-square fluctuations in a protein crystal are fitted using a trivariate Gaussian distribution described by the ADP. To obtain the theoretical values of ADP, we first calculate the $3N \times 3N$ covariance matrix for the following all-atom or coarse-grained models of protein-environment system (N is the number of atoms or residues).

TABLE I. Backbone RMSD of structural changes in main protein by energy minimization

PDB code	RMSD of EEM (Å)	RMSD of IEM (Å)
1g6x	1.14	1.20
2fdn	1.00	1.48
1rb9	0.68	0.74
1kth	1.05	0.95
1oai	0.72	1.00
1c75	1.02	1.59
1f94	1.24	1.12
1iqz	1.00	1.00
1vbw	0.95	0.94
1vyy	0.79	0.79
1tg0	0.94	1.11
1m1q	1.77	2.27
1ok0	1.07	0.81
1r6j	0.95	1.42
1xmk	0.97	0.95
1l9l	1.13	0.96
1x6z	1.14	1.14
1u2h	0.78	0.97
1iua	0.93	1.21
1lkk	1.03	1.03
1zzk	1.00	1.68
1ufy	0.97	1.43
2pvb	0.91	0.85
1j0p	1.82	2.69
1gqv	0.83	0.90
1nwz	0.78	0.82
1tqg	0.61	0.73
1r2m	1.17	2.24
1c7k	0.79	0.97
1g4i	1.29	1.26
1unq	1.06	2.16
1w0n	1.23	2.01
1mc2	1.26	1.19
1v6p	1.36	1.70
3lzt	0.81	0.97
1exr	1.27	2.23
1a6m	0.85	1.28
1f9y	0.77	0.93
1tt8	0.72	0.90
1eb6	0.99	1.01

1. EEM

The atoms of main protein (M) and part of environment (E_1) are free to move, while the rest of environment (E_2) is fixed. So the atomic covariance matrix of the main protein is

$$\langle u_M u_M^T \rangle_{\text{EEM}} = k_B T \cdot \begin{bmatrix} H_{MM} & H_{ME_1} \\ H_{E_1 M} & H_{E_1 E_1} \end{bmatrix}^{-1}_{MM}, \tag{2}$$

where u_M represents the atomic displacement of the main protein, k_B is the Boltzmann constant, T is the temperature, H_{MM},

H_{ME_1}, H_{E_1M}, and $H_{E_1E_1}$ are four sub-matrices of the all-atom Hessian matrix that involve the main protein and E_1.

2. IEM

The atomic covariance matrix of the main protein is

$$\left\langle u_M u_M^T \right\rangle_{\text{IEM}} = k_B T (H_M + K_E N_E)^{-1}, \tag{3}$$

where H_M is the all-atom Hessian matrix of an isolated main protein, N_E is a diagonal matrix whose $(3i + j)$th diagonal element (j = 0, 1, 2) is given by the number of heavy atoms of neighboring proteins within 4.5 Å from atom i of the main protein, and K_E is the force constant of a harmonic restraint.

3. Two-component ENM with fixed environment

The covariance matrix of C_α atoms in the main protein is

$$\left\langle u_M u_M^T \right\rangle_{\text{ENM}} = k_B T (H_{MM})^{-1}, \tag{4}$$

where H_{MM} is the MM sub-matrix of the coarse-grained Hessian matrix.[38]

To exploit the sparseness of Hessian matrix, we use a sparse linear-equation solver CHOLMOD (Ref. 42) to calculate matrix inversion in Eqs. (2)–(4). It is computationally more efficient and accurate than the calculation of H^{-1} using a subset of low-frequency modes.[28]

Given the covariance matrix $\langle u_M u_M^T \rangle$, one can calculate the theoretical ADP tensor for atom i using the ith 3×3 diagonal block C_{ii} of $\langle u_M u_M^T \rangle$:

$$C_{ii} = \begin{bmatrix} \langle \delta x_i^2 \rangle & \langle \delta x_i \delta y_i \rangle & \langle \delta x_i \delta z_i \rangle \\ \langle \delta x_i \delta y_i \rangle & \langle \delta y_i^2 \rangle & \langle \delta y_i \delta z_i \rangle \\ \langle \delta x_i \delta z_i \rangle & \langle \delta y_i \delta z_i \rangle & \langle \delta z_i^2 \rangle \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ U_{12} & U_{22} & U_{23} \\ U_{13} & U_{23} & U_{33} \end{bmatrix}, \tag{5}$$

where the diagonal elements U_{11}, U_{22}, U_{33} give the mean-squared fluctuations of atom i along x, y, z direction, and the off-diagonal elements U_{12}, U_{13}, U_{23} describe the covariance among the displacements of the atom i along x, y, z direction. Together, the six ADP elements determine a three-dimensional Gaussian distribution function which describes both the direction and magnitude of the atomic fluctuations.[5] For fixed probability value, the distribution is ellipsoidal with a directional preference along the long axis, which is given by the eigenvector of ADP tensor with the largest eigenvalue. The anisotropy of the Gaussian distribution is defined as the ratio of the smallest to the largest eigenvalue of ADP tensor.

The isotropic B factor is related to the trace of ADP tensor as follows:

$$\text{B} = 8\pi^2 (U_{11} + U_{22} + U_{33})/3. \tag{6}$$

D. Comparison between theoretical and experimental ADP

We use the following metrics to assess the similarity between experimental and theoretical ADP tensors (represented as U and V).

1. Real-space correlation coefficient

The following real-space correlation coefficient is calculated to evaluate the overlap integral of two three-dimensional Gaussian distributions given by U and V:[5]

$$\text{cc}(U, V) = \frac{(\det U^{-1} \det V^{-1})^{1/4}}{[\det(U^{-1} + V^{-1})/8]^{1/2}}. \tag{7}$$

Based on the real-space correlation coefficient, the following two metrics have been introduced to evaluate the directional similarity of two ADPs.

a. Normalized correlation coefficient (ncc).

$$\text{ncc}(U, V) = \frac{\text{cc}(U, V)}{\text{cc}(U, U_{iso})\,\text{cc}(V, V_{iso})}, \tag{8}$$

where $U_{iso} = V_{iso} = I_3/3$, I_3 is a 3×3 identity matrix, U and V have been normalized by their trace. ncc measures the similarity between U and V relative to their similarities to an isotropic tensor.[5] Following a previous study,[30] we use a simple ratio of the number of ADPs with ncc > 1 and the total number of ADPs (named f_{ncc}) to measure the overall similarity between two sets of ADPs.

b. Modified correlation coefficient (cc_{mod}*)*

$$\text{cc}_{\text{mod}}(U, V) = \frac{\text{cc}(U, V) - \text{cc}(U, V^*)}{1 - \text{cc}(U, V^*)}, \tag{9}$$

where V^* is a 3×3 tensor generated by taking the eigenvectors of U and using the eigenvalues of V, with the largest and smallest switched to define the two ellipsoids with perfect misalignment.[28,36] cc_{mod} is 1.0 (0) if the two ellipsoids are perfectly aligned (misaligned).

2. Kullback-Leibler (KL) distance

The KL distance[43] evaluates the difference between the three-dimensional Gaussian distributions a and b as defined by U and V.[29] The KL distance can be expressed in terms of the eigenvalues (d_{ak} and d_{bk}, $k = 1, 2, 3$) and eigenvectors (v_{ak} and v_{bk}, $k = 1, 2, 3$) of U and V as follows:

$$D_{ab} = -\frac{3}{2} + \frac{1}{2} \sum_{k=1}^{3} ln \frac{d_{bk}}{d_{ak}} + \frac{1}{2} \sum_{k=1}^{3} \sum_{l=1}^{3} \frac{d_{ak}}{d_{bl}} \left| v_{ak}^T v_{bl} \right|^2. \tag{10}$$

Since the KL distance is asymmetric ($D_{ab} \neq D_{ba}$), the arithmetic average $(D_{ab} + D_{ba})/2$ was calculated previously.[29] We notice that D_{ab} diverges if the distribution b is highly anisotropic (with a near-zero eigenvalue). To avoid such

divergence, we use $min\{D_{ab}, D_{ba}\}$ instead of $(D_{ab} + D_{ba})/2$ as our KL distance metric.

3. Dot product

It is defined as the absolute value of the dot product between the two eigenvectors of U and V with the largest eigenvalue.[28] The dot product is 1 if the long axes of U and V are perfectly aligned, and 0 if their long axes are perpendicular to each other.

4. Pearson correlations

The above metrics only evaluate the directional similarity of two ADPs. To include the magnitude of ADPs into comparison, we compute the Pearson correlation (termed pc_{all}) between two sets of ADPs as two $6N'$-dimensional vectors $\tilde{U}$ and $\tilde{V}$ (Ref. 29) (N' is the number of ADPs):

$$pc = \frac{\sum_{j=1}^{6N'} (\tilde{U}_j - \langle\tilde{U}\rangle)(\tilde{V}_j - \langle\tilde{V}\rangle)}{\sqrt{\sum_{j=1}^{6N'} (\tilde{U}_j - \langle\tilde{U}\rangle)^2 \sum_{j=1}^{6N'} (\tilde{V}_j - \langle\tilde{V}\rangle)^2}}. \quad (11)$$

We also calculate the Pearson correlations for $3N'$ diagonal and $3N'$ off-diagonal ADP elements separately (termed $pc_{diagonal}$ and $pc_{offdiagonal}$, respectively), and the Pearson correlation between theoretical and experimental B factors (termed pc_{trace}).[29]

E. Crystallographic dataset for model evaluation

We evaluate our modeling of ADPs using a set of 40 high-resolution protein crystal structures (PDB codes: 1g6x, 2fdn, 1rb9, 1kth, 1oai, 1c75, 1f94, 1iqz, 1vbw, 1vyy, 1tg0, 1m1q, 1ok0, 1r6j, 1xmk, 1l9l, 1x6z, 1u2h, 1iua, 1lkk, 1zzk, 1ufy, 2pvb, 1j0p, 1gqv, 1nwz, 1tqg, 1r2m, 1c7k, 1g4i, 1unq, 1w0n, 1mc2, 1v6p, 3lzt, 1exr, 1a6m, 1f9y, 1tt8, 1eb6), which are the smallest 40 structures of an old list of 83 PDB structures previously studied.[28,36,38] From the PDB files of these structures, we collect 4178 usable ADPs for those C_α atoms with occupancy of 1.0.

Following earlier studies,[28,36,38] for the evaluation of Pearson correlations of all, diagonal, off-diagonal ADP elements, and B factors, we use all 4178 ADPs; for the evaluation of directional metrics (f_{ncc}, cc_{mod}, KL distance, and dot product), we use a subset of 1648 ADPs with anisotropy ≤ 0.5.

III. RESULTS

To explore how crystal packing affects the modeling of ADPs, we have performed all-atom modeling of a protein structure in a crystalline environment for a list of 40 high-resolution crystal structures taken from previous studies.[28,36,38] We have calculated theoretical ADP tensors and compared them with experimental ADPs.

A. Explicit vs. implicit modeling of protein-environment system

In several previous studies,[35–38] the effect of crystalline environment on the atomic fluctuations of a protein structure was modeled using a coarse-grained model (ENM) under various boundary conditions. It remains unknown how much the inherent inaccuracy of coarse-grained representation and elastic force field affects the modeling results. To address this issue, we have used all-atom representation and force field (see Sec. II) to refine the modeling of protein-environment system. Two complementary models of crystalline environment are considered—explicit environment modeling (EEM) and implicit environment modeling (IEM). The EEM explicitly models the atomic fluctuations of neighboring proteins (as environment) together with the main protein, while the IEM implicitly models the effect of crystal packing by applying harmonic positional restraints to the surface atoms of the main protein in contact with neighboring proteins (see Sec. II). The two approaches complement each other very well: the EEM describes protein-environment interactions more explicitly but is computationally more expensive, while the IEM allows the flexibility of tuning protein-environment interactions and is computationally cheaper. To evaluate the accuracy of both models, we have employed them to calculate the ADPs of C_α atoms in the main protein and then compared with the experimental ADPs determined by high-resolution X-ray crystallography (see Sec. II).

To offer a glimpse to the modeling results, we have shown the results for a crystal structure of pancreatic trypsin inhibitor (PDB code: 1g6x, see Fig. 1) using EEM and IEM. A better agreement between theoretical and experimental ADPs is found for IEM than EEM—the Pearson correlations for diagonal, off-diagonal, all ADP elements, and B factors (for definitions, see Sec. II) increase significantly from 0.17, −0.02, 0.48, 0.21 for EEM to 0.71, 0.55, 0.89, 0.80 for IEM. For the directional comparison of experimental and theoretical ADPs, we focus on 18 out of 58 experimental ADPs of 1g6x with anisotropy ≤ 0.5. We calculate four metrics for directional similarity between theoretical and experimental ADPs (f_{ncc}, cc_{mod}, KL distance, and dot product, see Sec. II for definitions), which all indicate improvement from EEM to IEM—f_{ncc} increases from 0.82 to 0.91 (see Fig. 2(b)), the average dot product increases from 0.45 to 0.68 (see Fig. 2(c)), the average cc_{mod} increases from 0.70 to 0.72 (see Fig. 2(d)), and the average KL distance decreases from 0.32 to 0.055 (see Fig. 2(e)).

The ADP modeling based on EEM and IEM has been performed for 4178 ADPs of C_α atoms collected from 40 high-resolution protein crystal structures, which is a subset of an old list used in previous studies[28,36,38] (see Sec. II). To deduce the overall performance of our ADP modeling, we average four Pearson correlations (for all, diagonal, off-diagonal ADP elements, and B factors) over 40 structures, and three directional metrics (cc_{mod}, dot product, and KL distance) over a subset of 1648 ADPs with anisotropy ≤ 0.5. Another directional metric f_{ncc} is also calculated over this subset (see Sec. II). For EEM, we explore how the ADP modeling quality depends on the level of environment flexibility by

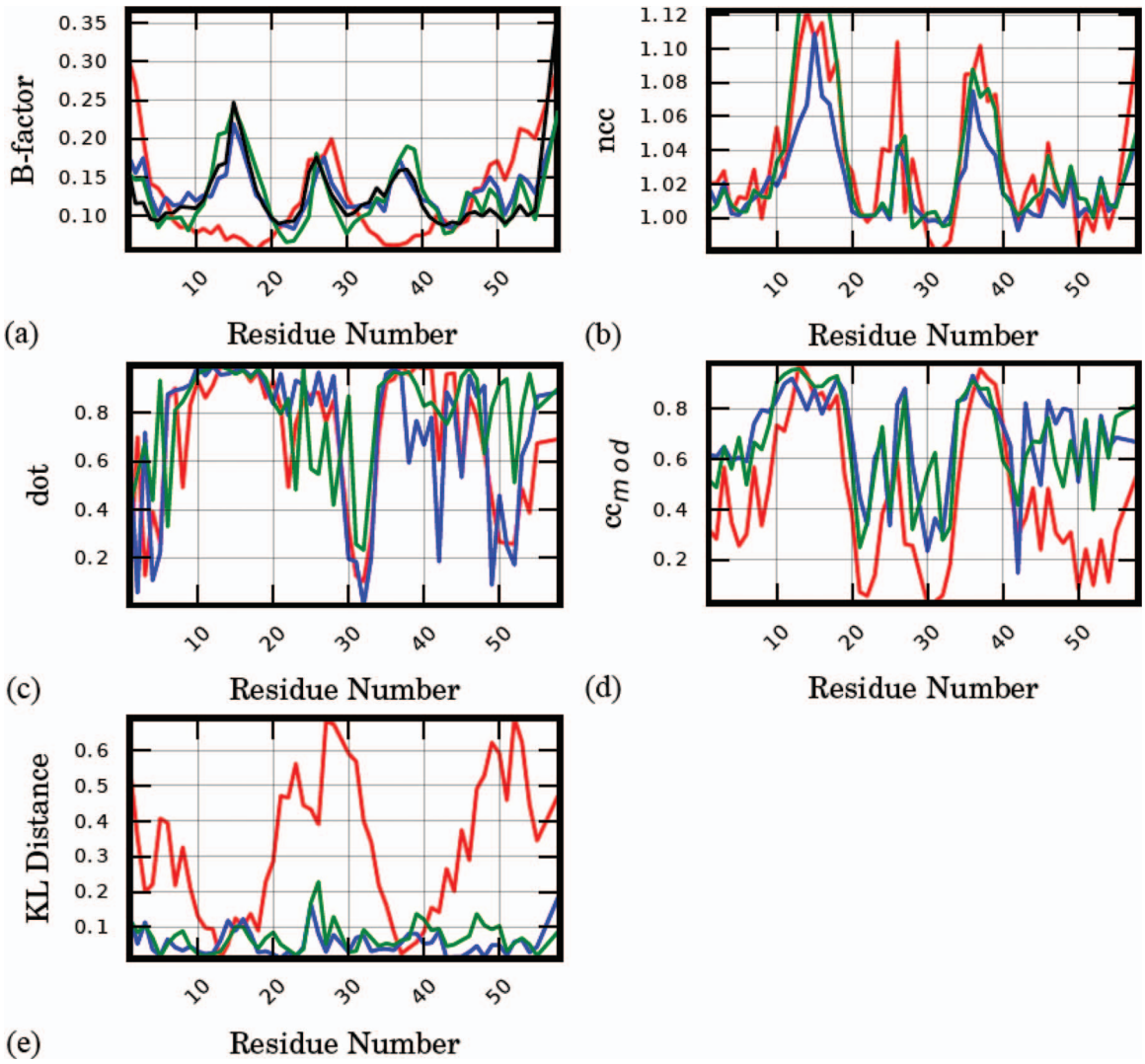

FIG. 2. The results of ADP modeling for a crystal structure of pancreatic trypsin inhibitor (PDB code: 1g6x): panels (a)–(e) show the B factors and four directional metrics (ncc, dot product, cc_{mod}, KL distance) as a function of residue number for EEM ($r_c = 20$ Å, colored red), IEM ($K_E = 0.01$, colored blue), and ENM ($f_{env} = 0.02$, colored green). In panel (a), the experimental B factors (rescaled by $3/8\pi^2$) are also shown in black.

varying the cutoff distance r_c that determines the partition of environment to moving and fixed parts (see Sec. II)—a higher r_c implies higher environment flexibility. For IEM, we explore how the ADP modeling quality depends on the strength of protein-environment interactions by varying the force constant K_E for harmonic restraints (see Sec. II)—a higher K_E implies stronger protein-environment interactions.

For EEM, it is found that all metrics gradually improve as r_c increases (see Fig. 3(a)). Meanwhile, all metrics seem to saturate as r_c approaches 20 Å (see Fig. 3(a)). Therefore, the atomic fluctuations of the main protein depend more on the flexibility of the environment atoms near the main protein than those environment atoms far away from the main protein. This observation justifies our truncation of environment based on distances from the main protein. Consequently, we can use the EEM results at $r_c = 20$ Å to assess the accuracy of ADP modeling based on explicit and equal treatment of intra-protein and protein-environment interactions (i.e., no tuning of protein-environment interactions).

For IEM, it is found that the minimum of average KL distance and the maxima of other metrics are roughly aligned near $K_E \sim 0.02$ (see Fig. 3(b)), where the optimal ADP modeling is attained. Furthermore, the optimal ADP modeling by IEM (at $K_E \sim 0.02$) is significantly better than that of EEM (at $r_c = 20$ Å)—the average Pearson correlations for all, diagonal, off-diagonal ADP elements, and B factors increase from 0.78, 0.40, 0.32, and 0.47 for EEM (see Fig. 3(a)) to 0.84, 0.55, 0.42, and 0.63 for IEM (see Fig. 3(b)). We then compare the directional metrics between EEM and IEM (see Figs. 3(a) and 3(b)): f_{ncc} increases from 0.87 to 0.91, the average cc_{mod} increases from 0.65 to 0.68, the average dot product increases from 0.69 to 0.71, and the average KL distance decreases from 0.14 to 0.10. This finding supports the importance of tuning protein-environment interactions (as in IEM) in optimizing the modeling of ADPs, which agrees with our earlier study.[38] In particular, the optimal modeling of ADPs is not achieved by treating intra-protein and protein-environment interactions equally (as in

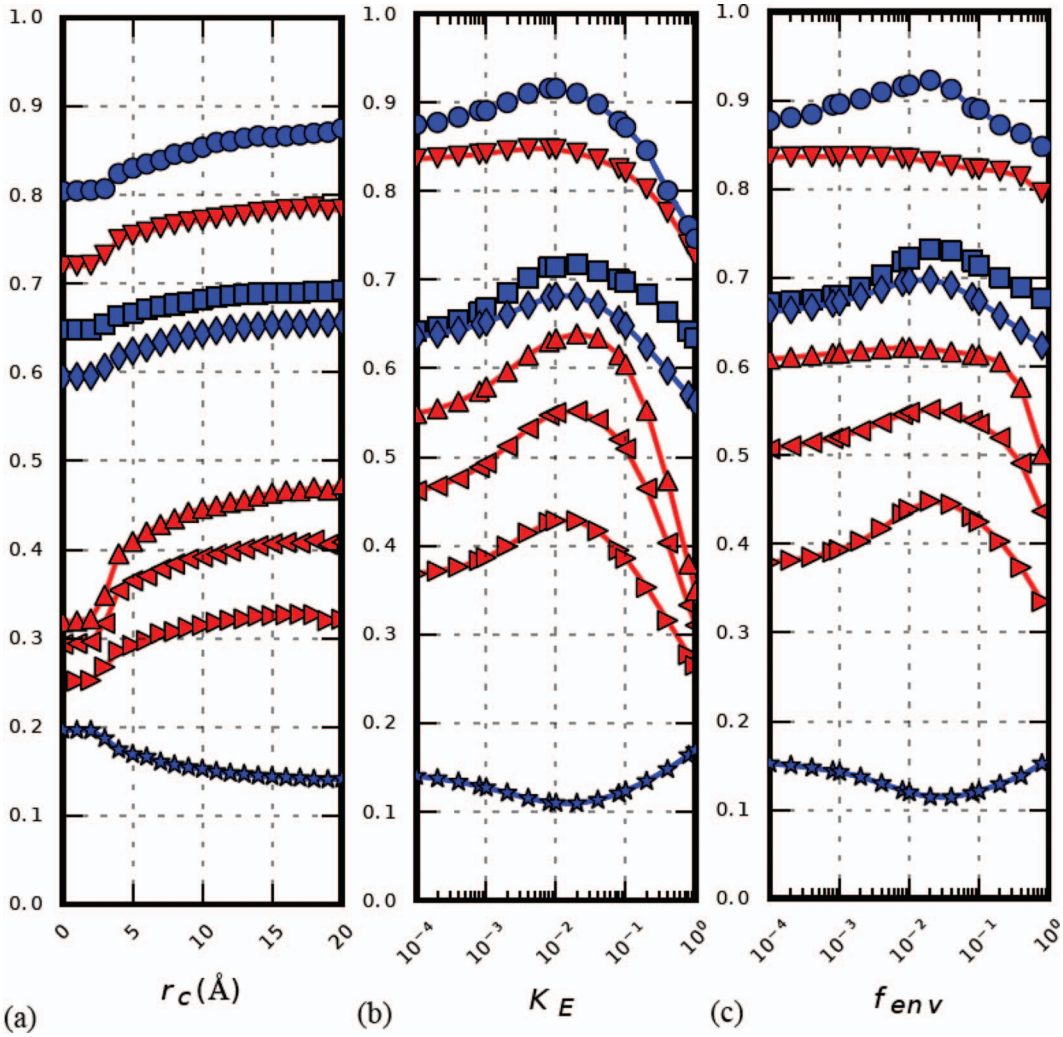

FIG. 3. The results of ADP modeling averaged over 40 protein crystal structures for: (a) EEM; (b) IEM; (c) ENM. Shown here are Pearson correlations of diagonal (◄), off-diagonal (►), all elements (▼) of ADPs and B factors (▲), and directional metrics including f_{ncc} (●), dot product (■), cc_{mod} (♦), KL distance (★).

EEM). We will discuss whether the optimal modeling of ADPs by IEM is achieved at weak protein-environment interactions in Subsection III C.

To assess the statistical significance of our finding of better ADP modeling by IEM than EEM, we have compared the ADP modeling quality of IEM and EEM for each of the 40 crystal structures. For both IEM and EEM, we have calculated all metrics for each structure and then averaged the three directional metrics (cc_{mod}, dot product, and KL distance) over the ADPs of each structure with anisotropy $\leq$0.5. It is found that better ADP modeling from EEM to IEM is achieved for 80%, 90%, 75%, 90%, 60%, 68%, 68%, and 85% of all structures as assessed by the Pearson correlations for all, diagonal, off-diagonal ADP elements, B factors, f_{ncc}, cc_{mod}, dot product, and KL distance, respectively (see Fig. 4). So IEM outperforms EEM according to all eight metrics, which strongly supports its statistical significance.

B. All-atom vs. coarse-grained modeling of protein-environment system

Next, we ask if the all-atom modeling improves the fitting of experimental ADPs over our previous coarse-grained modeling.[38] To answer this question, we have performed ENM-based modeling of a protein structure in a crystalline environment for the same list of 40 high-resolution crystal structures.[38] We have then calculated theoretical ADP tensors and compared them with experimental ADPs. The similarity metrics between theoretical and experimental ADPs are calculated and plotted as a function of f_{env} (see Fig. 3(c)). Here f_{env} tunes the strength of protein-environment interactions relative to intra-protein interactions (see Sec, II).[38]

Same as our previous study,[38] it is found that the minimum of average KL distance and the maxima of other metrics are roughly aligned near $f_{env} \sim 0.02$ (see Fig. 3(c)). So the optimal ADP modeling is attained at weak protein-environment interactions relative to intra-protein interactions.[38] The optimal values of all metrics are very similar between ENM and IEM (see Figs. 3(b) and 3(c)): for ENM (IEM), the average Pearson correlations for all, diagonal, off-diagonal ADP elements, and B factors are 0.84, 0.55, 0.42, and 0.63 (0.83, 0.55, 0.44, and 0.61), f_{ncc} is 0.91 (0.92), the average cc_{mod} is 0.68 (0.69), the average dot product is 0.71 (0.73), and the average KL distance is 0.10 (0.11). Therefore, all-atom refinement

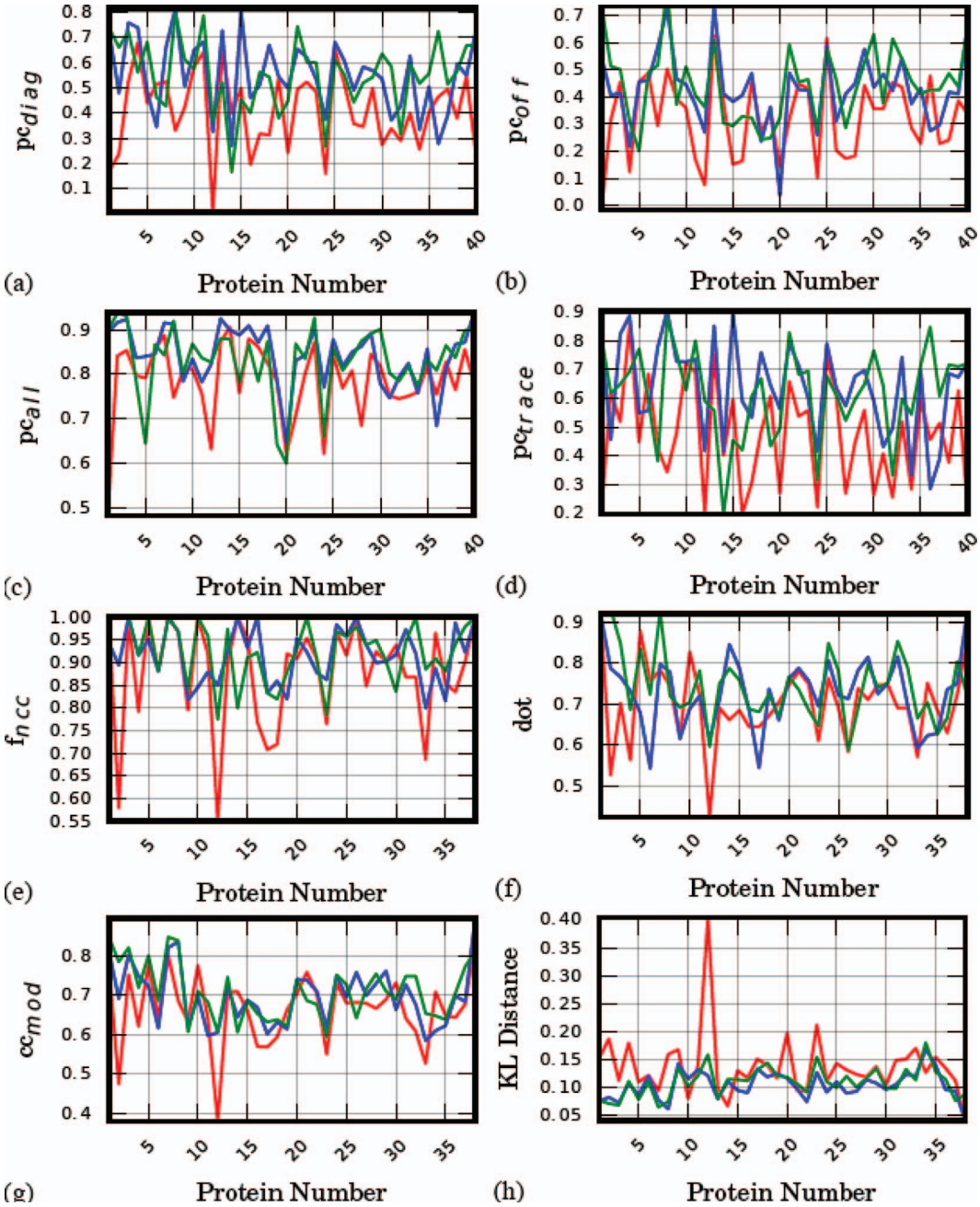

FIG. 4. The variation of ADP modeling results among 40 protein crystal structures calculated using EEM (in red), IEM (in blue), and ENM (green): panels (a)–(h) show Pearson correlations of diagonal, off-diagonal, all elements of ADPs and B factors, and directional metrics including $f_{ncc,}$ dot product, cc_{mod}, KL distance.

does not lead to noticeable improvement of ADP modeling over coarse-grained modeling.[38]

To further assess the statistical significance of our finding that ADP modeling by IEM and ENM attains similar accuracy, we have compared the ADP modeling quality of IEM and ENM for each of the 40 crystal structures. For both IEM and ENM, we have calculated all metrics for each structure and then averaged the three directional metrics (cc_{mod}, dot product, and KL distance) over the ADPs of each structure with anisotropy $\leq$0.5. It is found that the improvement of ADP modeling from ENM to IEM is achieved for 58%, 58%, 40%, 58%, 43%, 35%, 48%, and 73% of all structures as assessed by the Pearson correlations for all, diagonal, off-diagonal ADP elements, B factors, f_{ncc}, cc_{mod}, dot product, and KL distance, respectively (see Fig. 4). So IEM performs better according to four metrics (Pearson correlations for all, diagonal ADP elements, B factors, KL distance), while ENM performs better according to the other four metrics (Pearson correlation for off-diagonal ADP elements, f_{ncc}, cc_{mod}, dot product). Overall, their performance is comparable to each other.

C. All-atom modeling supports weak protein-environment interactions

In our previous study,[38] we found that ENM-based ADP modeling is optimal when the protein-environment interactions are much weaker than intra-protein interactions. Does

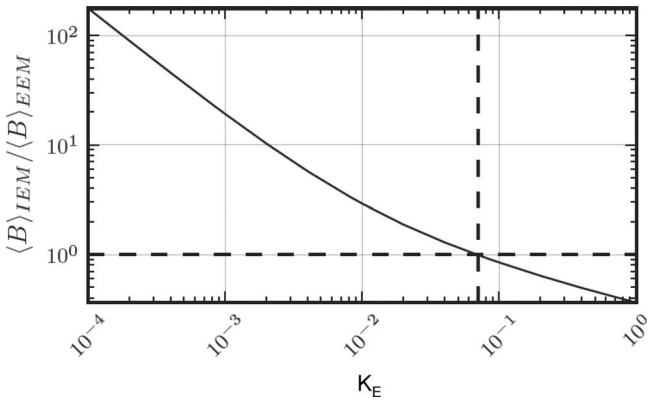

FIG. 5. The ratio between the average of B factors for 40 protein crystal structures calculated from IEM and that calculated from EEM as a function of the force constant of harmonic restraints (K_E).

the same conclusion hold for all-atom modeling based on IEM? If so, the weak protein-environment interactions can account for the better performance of IEM than EEM (see Figs. 3(a) and 3(b)). To answer this question, we need to determine the force constant K_E in IEM that corresponds to EEM (because EEM assumes equal strength of intra-protein and protein-environment interactions). To this end, we have computed the ratio between the average of B factors for 40 structures calculated by IEM and that calculated by EEM. As expected, this ratio decreases as K_E increases, and it crosses value 1 when $K_E \sim 0.07$ (see Fig. 5), which corresponds to equal strength of intra-protein and protein-environment interactions. Since IEM attains optimal ADP modeling at $K_E \sim 0.02 \ll 0.07$, we infer that the optimal ADP modeling by IEM is indeed achieved at weak protein-environment interactions relative to intra-protein interactions.

Although qualitatively similar, the findings by coarse-grained and all-atom modeling are quantitatively distinct. The optimal ADP modeling by ENM is attained when protein-environment interactions are much weaker than intra-protein interactions (between residues) by a factor of ~50.[38] In contrast, the optimal ADP modeling by IEM requires a much smaller ratio of ~4 between protein-environment and intra-protein interactions (between atoms). Such reduction is attributed to the use of more realistic all-atom force field (with atom-specific van der waals and electrostatic forces replacing uniform elastic forces between residues or atoms[38]).

IV. CONCLUSION

In conclusion, we have refined the modeling of atomic fluctuations in a protein crystal structure using all-atom representation and force field. The crystalline environment is modeled either explicitly (assuming equal intra-protein and protein-environment interactions) or implicitly (as harmonic restraints with tunable force constant). We have then evaluated the modeling by comparing theoretical ADPs with experimental ADPs for 40 high-resolution protein crystal structures. Our findings are summarized as follows:

1. The implicit treatment of crystalline environment by IEM outperforms the explicit treatment of crystalline environment by EEM, which implies a difference in strength between protein-environment and intra-protein interactions.
2. ADP modeling is optimal when the protein-environment interactions are much weaker than the intra-protein interactions, which is in qualitative agreement with our previous study based on coarse-grained modeling.[38] Compared with the coarse-grained modeling,[38] the all-atom modeling has found much smaller difference between protein-environment and intra-protein interactions, thanks to the use of more realistic all-atom force field over simple elastic force field of ENM. Future improvement in the explicit modeling of protein-environment interactions with better account of electrostatic screening and solvation effect may eventually remove the need to adjust the strength of protein-environment interactions to attain optimal fitting of experimental ADPs.
3. All-atom modeling does not lead to noticeable improvement over coarse-grained modeling, which supports the usefulness of coarse-grained modeling, and also suggests that the accuracy of ADP modeling is limited by factors other than force field accuracy. Future studies should go beyond harmonic approximation and improve the modeling of other aspects of protein dynamics including anharmonicity[44] and solvent-damped diffusive motions.[45]

ACKNOWLEDGMENTS

We acknowledge funding support from The American Heart Association (Grant No. 0835292N) and the computing resources from The Center for Computational Research at the University at Buffalo.

[1] K. Henzler-Wildman and D. Kern, Nature (London) **450**(7172), 964 (2007).
[2] D. D. Boehr, H. J. Dyson, and P. E. Wright, Chem. Rev. **106**(8), 3055 (2006).
[3] B. T. M. Willis, *Thermal Vibrations in Crystallography* (Cambridge University Press, London, 1975).
[4] C. Scheringer, Acta Crystallogr., Sect. A: Cryst. Phys., Diffr., Theor. Gen. Crystallogr. **33**, 879 (1977).
[5] E. A. Merritt, Acta Crystallogr., Sect. D: Biol. Crystallogr. **55**(Pt 12), 1997 (1999).
[6] A. E. Garcia, J. A. Krumhansl, and H. Frauenfelder, Proteins **29**(2), 153 (1997).
[7] M. Pellegrini, N. Gronbech-Jensen, J. A. Kelly, G. M. Pfluegl, and T. O. Yeates, Proteins **29**(4), 426 (1997).
[8] M. Karplus and J. A. McCammon, Nat. Struct. Biol. **9**(9), 646 (2002).
[9] V. Tozzini, Curr. Opin. Struct. Biol. **15**(2), 144 (2005).
[10] R. Diamond, Acta Crystallogr. A **46**(Pt 6), 425 (1990); A. Kidera and N. Go, Proc. Natl. Acad. Sci. U.S.A. **87**(10), 3718 (1990).
[11] B. K. Poon, X. Chen, M. Lu, N. K. Vyas, F. A. Quiocho, Q. Wang, and J. Ma, Proc. Natl. Acad. Sci. U.S.A. **104**(19), 7869 (2007); X. Chen, B. K. Poon, A. Dousis, Q. Wang, and J. Ma, Structure **15**(8), 955 (2007).
[12] M. Levitt, C. Sander, and P. S. Stern, J. Mol. Biol. **181**(3), 423 (1985).
[13] N. Go, T. Noguti, and T. Nishikawa, Proc. Natl. Acad. Sci. U.S.A. **80**(12), 3696 (1983).
[14] B. Brooks and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **80**(21), 6571 (1983).
[15] M. M. Tirion, Phys. Rev. Lett. **77**(9), 1905 (1996).
[16] F. Tama and Y. H. Sanejouand, Protein Eng. **14**(1), 1 (2001).
[17] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Biophys. J. **80**(1), 505 (2001).
[18] I. Bahar, A. R. Atilgan, and B. Erman, Folding Des. **2**(3), 173 (1997).

[19]T. Haliloglu, I. Bahar, and B. Erman, Phys. Rev. Lett. **79**, 3090 (1997).
[20]K. Hinsen, A. Petrescu, S. Dellerue, M. Bellissent-Funel, and G. R. Kneller, Chem. Phys. **261**, 25 (2000).
[21]L. Yang, G. Song, and R. L. Jernigan, Proc. Natl. Acad. Sci. U.S.A. **106**(30), 12347 (2009).
[22]I. Bahar and A. J. Rader, Curr. Opin. Struct. Biol. **15**(5), 586 (2005).
[23]F. Tama and C. L. Brooks, Annu. Rev. Biophys. Biomol. Struct. **35**, 115 (2006).
[24]T. Z. Sen, Y. Feng, J. V. Garcia, A. Kloczkowski, and R. L. Jernigan, J. Chem. Theory. Comput. **2**(3), 696 (2006).
[25]D. A. Kondrashov, Q. Cui, and G. N. Phillips, Jr., Biophys. J. **91**(8), 2760 (2006).
[26]E. Eyal, L.-W. Yang, and I. Bahar, Bioinformatics **22**, 2619 (2006).
[27]W. Zheng, Biophys. J. **94**(10), 3853 (2008).
[28]D. A. Kondrashov, A. W. Van Wynsberghe, R. M. Bannen, Q. Cui, and G. N. Phillips, Jr., Structure **15**(2), 169 (2007).
[29]E. Eyal, C. Chennubhotla, L. W. Yang, and I. Bahar, Bioinformatics **23**(13), i175 (2007).
[30]L. Yang, G. Song, and R. L. Jernigan, Proteins **76**(1), 164 (2009).
[31]R. Soheilifard, D. E. Makarov, and G. J. Rodin, Phys. Biol. **5**(2), 26008 (2008).
[32]G. N. Phillips, Jr., Biophys. J. **57**(2), 381 (1990).
[33]L. W. Yang, E. Eyal, C. Chennubhotla, J. Jee, A. M. Gronenborn, and I. Bahar, Structure **15**(6), 741 (2007).
[34]L. Yang, G. Song, A. Carriquiry, and R. L. Jernigan, Structure **16**(2), 321 (2008).
[35]S. Kundu, J. S. Melton, D. C. Sorensen, and G. N. Phillips, Jr., Biophys. J. **83**(2), 723 (2002).
[36]D. Riccardi, Q. Cui, and G. N. Phillips, Jr., Biophys. J. **96**(2), 464 (2009).
[37]K. Hinsen, Bioinformatics **24**(4), 521 (2008).
[38]J. Hafner and W. Zheng, J. Chem. Phys. **132**(1), 014111 (2010).
[39]A. Fiser, R. K. Do, and A. Sali, Protein. Sci. **9**(9), 1753 (2000).
[40]B. R. Brooks, C. L. Brooks, 3rd, A. D. Mackerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus, J. Comput. Chem. **30**(10), 1545 (2009).
[41]T. Lazaridis and M. Karplus, Proteins **35**(2), 133 (1999).
[42]Y. Chen, T. A. Davis, W. W. Hagner, and S. Rajamanickam, ACM Trans. Math. Softw. **35**, 1 (2008).
[43]S. Kullback and R. A. Leibler, Ann. Math. Stat. **22**, 79 (1951).
[44]W. Zheng, Biophys. J. **98**(12), 3025 (2010).
[45]B. T. Miller, W. Zheng, R. M. Venable, R. W. Pastor, and B. R. Brooks, J. Phys. Chem. B **112**(19), 6274 (2008).
[46]See supplementary material at http://dx.doi.org/10.1063/1.3646312 for additional results of ADP calculations using EEM and direct adjusting protein-environment interactions.