Article

Dead-End Elimination with a Polarizable Force Field Repacks PCNA Structures

Stephen D. LuCore,[1] Jacob M. Litman,[2] Kyle T. Powers,[2] Shibo Gao,[2] Ava M. Lynn,[1] William T. A. Tollefson,[1] Timothy D. Fenn,[3] M. Todd Washington,[2] and Michael J. Schnieders[1,2,*]

[1]Department of Biomedical Engineering and [2]Department of Biochemistry, University of Iowa, Iowa City, Iowa; and [3]Boehringer Ingelheim, Ridgefield, Connecticut

ABSTRACT A balance of van der Waals, electrostatic, and hydrophobic forces drive the folding and packing of protein side chains. Although such interactions between residues are often approximated as being pairwise additive, in reality, higher-order many-body contributions that depend on environment drive hydrophobic collapse and cooperative electrostatics. Beginning from dead-end elimination, we derive the first algorithm, to our knowledge, capable of deterministic global repacking of side chains compatible with many-body energy functions. The approach is applied to seven PCNA x-ray crystallographic data sets with resolutions 2.5–3.8 Å (mean 3.0 Å) using an open-source software. While PDB_REDO models average an R_{free} value of 29.5% and MOLPROBITY score of 2.71 Å (77th percentile), dead-end elimination with the polarizable AMOEBA force field lowered R_{free} by 2.8–26.7% and improved mean MOLPROBITY score to atomic resolution at 1.25 Å (100th percentile). For structural biology applications that depend on side-chain repacking, including x-ray refinement, homology modeling, and protein design, the accuracy limitations of pairwise additivity can now be eliminated via polarizable or quantum mechanical potentials.

INTRODUCTION

The Protein Data Bank (PDB; http://www.rcsb.org/pdb/home/home.do) (1) now contains biomolecular structural models derived from >90,000 x-ray diffraction experiments conducted over the last half century. More than 80,000 of these structures have been deposited with their original diffraction data, which permits the experiments to be more fully interpreted as biomolecular refinement programs improve (2,3). Only a small fraction of PDB structures result from diffraction data to atomic resolution (i.e., ~1 Å). Mid- to low-resolution data sets, such as those for proliferating-cell nuclear antigen (PCNA) studied here, are much more common (Fig. 1). For these data, attainment of high-quality models relies heavily on the use of both systematic validation tools such as MOLPROBITY (4) and the prior chemical knowledge contained in molecular mechanics force fields (5). It is also possible to leverage previously solved structures to parameterize restraints based on elastic networks (6,7), although this level of coarse-graining is incapable of repacking side chains as networks deform or come together to form the interface of a complex.

To address the protein side-chain repacking problem, a brute force search over discrete conformations is computationally intractable for even small proteins due to a combinatorial explosion of conformational possibilities. However, by considering the relative energetics of discrete side-chain conformations (rotamers) for a single residue in the context of its interactions with the rest of the protein structure, unfavorable rotamers can be eliminated by proving that they cannot be part of the global minimum energy conformation (GMEC). The eliminated conformations are dead-ends in the search; therefore, the algorithm used to eliminate rotamers, rotamer pairs, and so on, is known as dead-end elimination (DEE).

The combination of low-energy side-chain rotamer libraries (8–10) with DEE (11,12) global optimization has been widely used for protein electrostatic network optimization and sequence design (13–17). However, rotamer elimination criteria have only been defined for pairwise-additive energy functions such as the OPLS-AA (18), AMBER (19), and CHARMM (20) families of fixed partial-charge force fields and pairwise decomposable continuum solvents (21–23). Explicit inclusion of many-body effects has been neglected such that the strength of the interaction between two residues must be independent of their mutual environment. Therefore, important molecular driving forces such as the hydrophobic effect (24) and electronic polarization (25), which are fundamentally many-body in nature, have been implicitly approximated or neglected entirely (Fig. 2). Here we overcome the restriction to pairwise energy functions by showing that both the DEE criteria (11) and more-stringent Goldstein criteria (12) can be derived in the context of many-body energy functions such as polarizable force fields (25,26)

Submitted March 16, 2015, and accepted for publication June 29, 2015.

*Correspondence: michael-schnieders@uiowa.edu

Editor: Nathan Baker.

0006-3495/15/08/0816/11

http://dx.doi.org/10.1016/j.bpj.2015.06.062

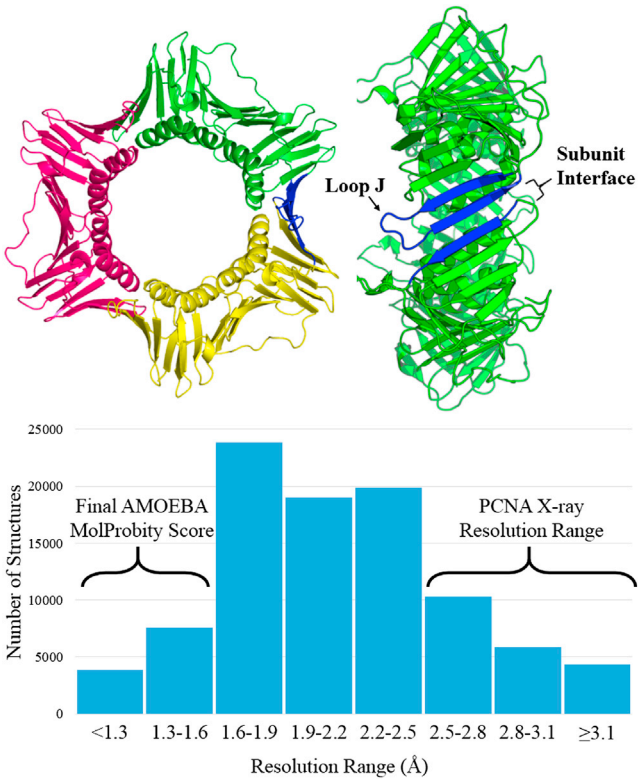


FIGURE 1 The biological unit of PCNA (*above*) is a trimeric ring with a central hole for binding double-strand DNA. Viewing PCNA from the side shows the subunit interface between two PCNA monomers (*blue*). Histogram (*below*) showing the number of x-ray diffraction data sets within resolution bins for structures deposited in the PDB as of January 20, 2015. The experimental resolution and deposited MOLPROBITY scores for the PCNA data sets used in this study averaged 2.96 and 2.86 Å, respectively. Structures determined by the algorithm described here yielded a mean MOLPROBITY score of 1.25 Å.

as well as quantum mechanical potentials and continuum solvents (27–30).

Rotamer and rotamer pair elimination criteria compatible with many-body energy functions are given below and their derivations supplied in the Supporting Material. For a pairwise decomposable energy function, the new expressions simplify to the established elimination criteria. The approach is used to refine a series of PCNA structures in the context of a many-body x-ray crystallographic target function $E_{\text{tot}} = E_{\text{chem}} + w_{A\text{ex-ray}}$. Here E_{chem} is a parallelized implementation of the polarizable AMOEBA force field that supports space group symmetry (31), $E_{\text{x-ray}}$ is a real-space electron density function (32,33), and w_A is used to weight the importance of the force field and x-ray terms (33). The resulting AMOEBA structures are compared to PDB_REDO (34) and pairwise DEE refinements based on the OPLS-AA/L(18) fixed charge force field.

Finally, functional insights into changes in PCNA stability due to single amino-acid mutations are discussed. PCNA plays an essential role in the maintenance of genome stability. It is a replication accessory factor that interacts with and regulates the activities of proteins involved in DNA replication, DNA repair, DNA recombination, chromatin modifications, sister chromatid cohesion, and cell-cycle control (35). Each PCNA subunit consists of two domains, which interact in a head-to-tail arrangement to form a ring-shaped homo-trimer possessing pseudo-sixfold symmetry (Fig. 1) (36). The PCNA trimer binds double-stranded DNA through the central pore of the ring. PCNA function is regulated in part by posttranslational modifications. For example, ubiquitylation of PCNA on Lys^{164} promotes translesion synthesis (TLS), which is the replicative and generally mutagenic bypass of damaged DNA (37). Several separation-of-function mutations in PCNA have been identified that inhibit various cellular processes including DNA mismatch repair as well as TLS (38,39). X-ray structures of wild-type PCNA, ubiquitin-modified PCNA, SUMO-modified PCNA, two separation-of-function mutant PCNA proteins that block mismatch repair, and two separation-of-function mutant proteins that block TLS have been determined (40–42).

MATERIALS AND METHODS

Theory

Side-chain repacking with a pairwise potential

For a potential energy function that approximates nonbonded interactions as being a pairwise sum over residues, the total energy of a protein $E(\mathbf{r})$ is given by

$$E(\mathbf{r}) = E_{\text{env}} + \sum_i E_{\text{self}}(r_i) + \sum_i \sum_{j>i} E_2(r_i, r_j), \quad (1)$$

where E_{env} is the energy of the environment (i.e., the protein backbone and residues that are not being optimized), $E_{\text{self}}(r_i)$ is the self-energy of residue i including its intramolecular bonded energy terms and nonbonded interactions with the backbone, and $E_2(r_i, r_j)$ is the two-body nonbonded interaction energy between residues i and j with other residues turned off. The self-energy and two-body terms, diagrammed in Fig. 2, are calculated as

$$E_{\text{self}}(r_i) = E_{BB/SC}(r_i) - E_{\text{env}}, \quad (2)$$

$$E_2(r_i, r_j) = E_{BB/SC}(r_i, r_j) - E_{\text{self}}(r_i) - E_{\text{self}}(r_j) - E_{\text{env}}, \quad (3)$$

where $E_{BB/SC}(r_i)$ is the energy of the protein backbone with only the side chain of residue i attached. Likewise, $E_{BB/SC}(r_i, r_j)$ is the energy of the backbone and only side chains i,j. E_{env} is subtracted from each self and two-body term to avoid double counting. The original elimination criteria for rotamers and rotamer pairs (11), respectively, under the approximation of a pairwise decomposable force field, are

$$E_{\text{self}}(r_i^\alpha) + \sum_{j'} \min_\gamma E_2(r_i^\alpha, r_j^\gamma) > E_{\text{self}}(r_i^\beta) + \sum_{j'} \max_\gamma E_2(r_i^\beta, r_j^\gamma), \quad (4)$$

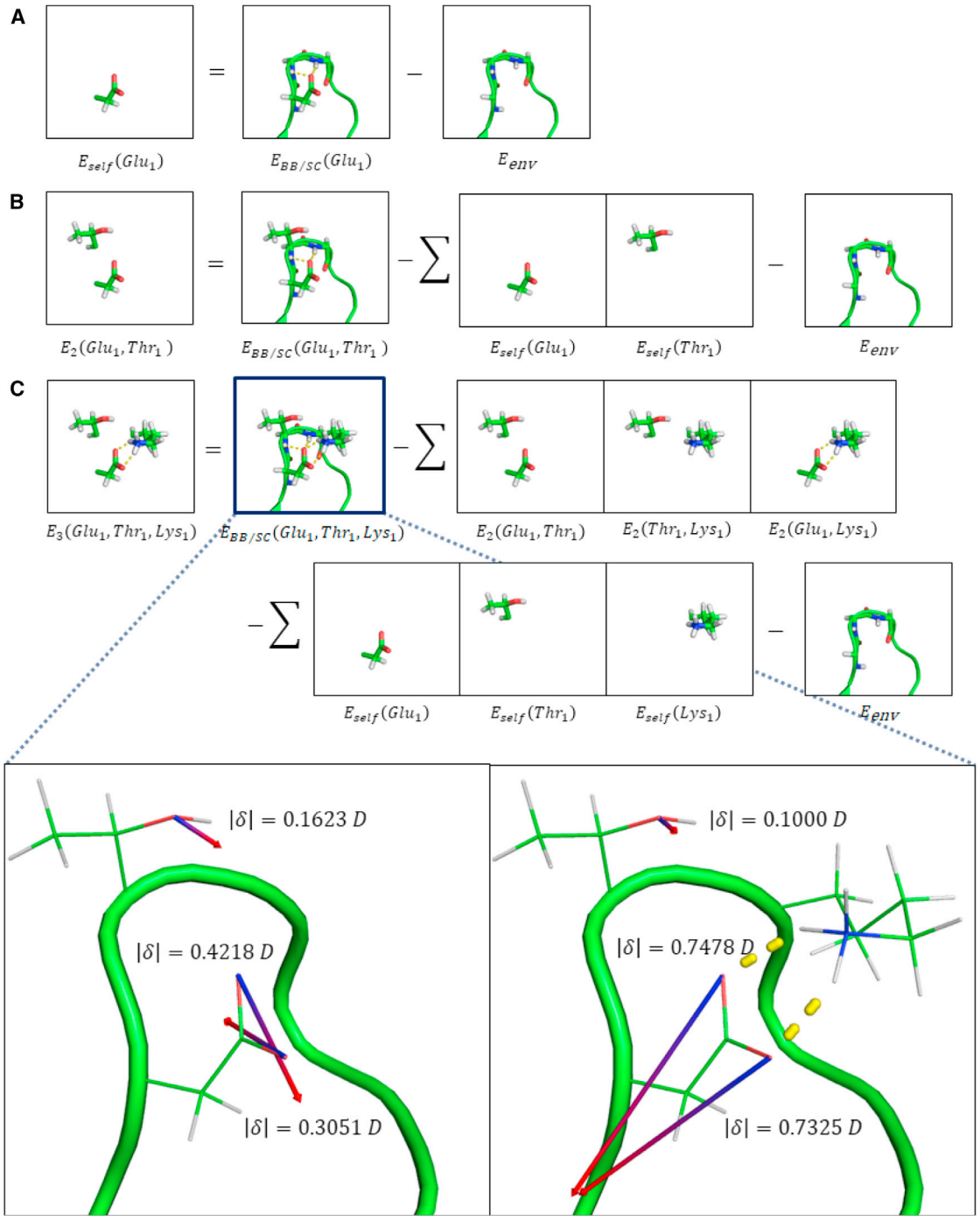


FIGURE 2 This diagram explains energy terms used in Eq. 1 including the (*A*) self-energy, (*B*) two-body energy, and (*C*) three-body energy. The induced dipole vectors for three side-chain oxygen atoms in wild-type PCNA are shown in the absence of Lys[107] (*left*) and in its presence (*right*). A vector length of 7 Å corresponds to 1 Debye. Changes in induced dipole direction and magnitude reflect the AMOEBA electronic polarization response to the total electrostatic field, which results in polarization energy that is not pairwise. For this example, the three-body energy is 1.55 kcal/mol.

$$E_{\text{pair}}\left(r_i^{\alpha}, r_j^{\beta}\right) + \sum_{k'} \min_{\varepsilon}\left[E_2\left(r_i^{\alpha}, r_k^{\varepsilon}\right) + E_2\left(r_j^{\beta}, r_k^{\varepsilon}\right)\right] > E_{\text{pair}}\left(r_i^{\gamma}, r_j^{\delta}\right) + \sum_{k'} \max_{\varepsilon}\left[E_2\left(r_i^{\gamma}, r_k^{\varepsilon}\right) + E_2\left(r_j^{\delta}, r_k^{\varepsilon}\right)\right], \tag{5}$$

$$E_{\text{pair}}\left(r_i^{\alpha}, r_j^{\beta}\right) = E_{\text{self}}\left(r_i^{\alpha}\right) + E_{\text{self}}\left(r_j^{\beta}\right) + E_2\left(r_i^{\alpha}, r_j^{\beta}\right), \tag{6}$$

where r_i^{α} and r_i^{β} are different rotamers of the same residue i. The prime notation indicates that the summation occurs over all residues $i \neq j$; similarly, k' implies $k \neq i$, $k \neq j$.

Side-chain repacking with a many-body potential

Under a many-body potential, the total energy of a protein $E(\mathbf{r})$ can be defined to arbitrary precision using the expansion

$$\begin{aligned} E(\mathbf{r}) =& E_{\text{env}} + \sum_i E_{\text{self}}(r_i) + \sum_i \sum_{j>i} E_2(r_i, r_j) \\ &+ \sum_i \sum_{j>i} \sum_{k>j} E_3(r_i, r_j, r_k) \\ &+ \sum_i \sum_{j>i} \sum_{k>j} \sum_{l>k} E_4(r_i, r_j, r_k, r_l) + \ldots, \end{aligned} \tag{7}$$

where the three- and four-body contributions, respectively, are given by

$$\begin{aligned} E_3(r_i, r_j, r_k) =& E_{BB/SC}(r_i, r_j, r_k) - E_{\text{self}}(r_i) - E_{\text{self}}(r_j) \\ &- E_{\text{self}}(r_k) - E_2(r_i, r_j) - E_2(r_i, r_k) \\ &- E_2(r_j, r_k) - E_{\text{env}}, \end{aligned} \tag{8}$$

$$\begin{aligned} E_4(r_i, r_j, r_k, r_l) =& E_{BB/SC}(r_i, r_j, r_k, r_l) - E_{\text{self}}(r_i) - E_{\text{self}}(r_j) \\ &- E_{\text{self}}(r_k) - E_{\text{self}}(r_l) - E_2(r_i, r_j) \\ &- E_2(r_i, r_k) - E_2(r_j, r_k) - E_2(r_i, r_l) \\ &- E_2(r_j, r_l) - E_2(r_k, r_l) - E_3(r_i, r_j, r_k) \\ &- E_3(r_i, r_j, r_l) - E_3(r_i, r_k, r_l) \\ &- E_3(r_j, r_k, r_l) - E_{\text{env}}. \end{aligned} \tag{9}$$

The DEE rotamer and rotamer pair elimination equations, respectively, can be extended to arbitrary order as follows:

$$\begin{aligned} &E_{\text{self}}(r_i^\alpha) + \sum_{j'} \min_\gamma \Big\{ E_2(r_i^\alpha, r_j^\gamma) \\ &\quad + \sum_{k'} \min_\delta \left[E_3(r_i^\alpha, r_j^\gamma, r_k^\delta) + \ldots \right] \Big\} > E_{\text{self}}(r_i^\beta) \\ &\quad + \sum_{j'} \max_\gamma \Big\{ E_2(r_i^\beta, r_j^\gamma) + \sum_{k'} \max_\delta \left[E_3(r_i^\beta, r_j^\gamma, r_k^\delta) + \ldots \right] \Big\}, \end{aligned} \tag{10}$$

$$\begin{aligned} &E_{\text{pair}}(r_i^\alpha, r_j^\beta) + \sum_{k'} \min_\varepsilon \Big\{ E_2(r_i^\alpha, r_k^\varepsilon) + E_2(r_j^\beta, r_k^\varepsilon) + E_3(r_i^\alpha, r_j^\beta, r_k^\varepsilon) + \sum_{l'} \min_\eta \left[E_3(r_i^\alpha, r_k^\varepsilon, r_l^\eta) + E_3(r_j^\beta, r_k^\varepsilon, r_l^\eta) \right. \\ &\quad \left. + E_4(r_i^\alpha, r_j^\beta, r_k^\varepsilon, r_l^\eta) + \ldots \right] \Big\} > E_{\text{pair}}(r_i^\gamma, r_j^\delta) + \sum_{k'} \max_\varepsilon \Big\{ E_2(r_i^\gamma, r_k^\varepsilon) + E_2(r_j^\delta, r_k^\varepsilon) + E_3(r_i^\gamma, r_j^\delta, r_k^\varepsilon) \\ &\quad + \sum_{l'} \max_\eta \left[E_3(r_i^\gamma, r_k^\varepsilon, r_l^\eta) + E_3(r_j^\delta, r_k^\varepsilon, r_l^\eta) + E_4(r_i^\gamma, r_j^\delta, r_k^\varepsilon, r_l^\eta) + \ldots \right] \Big\}, \end{aligned} \tag{11}$$

where the ellipses signify the presence of further higher-order terms up to n-body (see the Supporting Material for the derivation).

Although terms based on interactions between three or more residues are zero for pairwise decomposable energy functions such as OPLS-AA/L, for the polarizable AMOEBA force field the three-body term $E_3(r_i, r_j, r_k)$ captures energetic changes due to mutual polarization between the three residues and their environment (Fig. 2). In the context of continuum solvation, higher-order terms additionally capture energetic changes that result from modifications to the protein-solvent dielectric boundary.

Goldstein elimination

More stringent elimination criteria were introduced by Goldstein (12), which for a pairwise energy function are given by

$$E_{\text{self}}(r_i^\alpha) - E_{\text{self}}(r_i^\beta) + \sum_{j'} \min_\gamma \left[E_2(r_i^\alpha, r_j^\gamma) - E_2(r_i^\beta, r_j^\gamma) \right] > 0, \tag{12}$$

and

$$\begin{aligned} &E_{\text{pair}}(r_i^\alpha, r_j^\beta) - E_{\text{pair}}(r_i^\gamma, r_j^\delta) + \sum_{k'} \min_\varepsilon \left[E_2(r_i^\alpha, r_k^\varepsilon) \right. \\ &\quad \left. + E_2(r_j^\beta, r_k^\varepsilon) - E_2(r_i^\gamma, r_k^\varepsilon) - E_2(r_j^\delta, r_k^\varepsilon) \right] > 0, \end{aligned} \tag{13}$$

for rotamer and rotamer pair elimination, respectively. The Goldstein elimination criteria, extended to include higher-order energy components for rotamers and rotamer pairs, respectively, are given by

$$\begin{aligned} &E_{\text{self}}(r_i^\alpha) - E_{\text{self}}(r_i^\beta) + \sum_{j'} \min_\gamma \Big\{ E_2(r_i^\alpha, r_j^\gamma) - E_2(r_i^\beta, r_j^\gamma) \\ &\quad + \sum_{k'} \min_\delta \left[E_3(r_i^\alpha, r_j^\gamma, r_k^\delta) - E_3(r_i^\beta, r_j^\gamma, r_k^\delta) \ldots \right] \Big\} > 0, \end{aligned} \tag{14}$$

and

$$\begin{aligned} &E_{\text{pair}}(r_i^\alpha, r_j^\beta) - E_{\text{pair}}(r_i^\gamma, r_j^\delta) + \sum_{k'} \min_\varepsilon \Big\{ E_2(r_i^\alpha, r_k^\varepsilon) \\ &\quad + E_2(r_j^\beta, r_k^\varepsilon) + E_3(r_i^\alpha, r_j^\beta, r_k^\varepsilon) - E_2(r_i^\gamma, r_k^\varepsilon) - E_2(r_j^\delta, r_k^\varepsilon) \\ &\quad - E_3(r_i^\gamma, r_j^\delta, r_k^\varepsilon) + \sum_{l'} \min_\eta E_3(r_i^\alpha, r_k^\varepsilon, r_l^\eta) + E_3(r_j^\beta, r_k^\varepsilon, r_l^\eta) \\ &\quad + E_4(r_i^\alpha, r_j^\beta, r_k^\varepsilon, r_l^\eta) \left[- E_3(r_i^\gamma, r_k^\varepsilon, r_l^\eta) - E_3(r_i^\delta, r_k^\varepsilon, r_l^\eta) \right. \\ &\quad \left. - E_4(r_i^\gamma, r_j^\delta, r_k^\varepsilon, r_l^\eta) + \ldots \right] \Big\} > 0. \end{aligned} \tag{15}$$

Many-body x-ray refinement function

Pairwise molecular mechanics force fields have been used in tandem with experimental x-ray diffraction data to refine protein structural models for more than two decades (43,44). To quantify agreement between the

experimental and model electron densities, and avoid overfitting, both R and R_{free} values are monitored (45). To measure agreement between the structural model and prior chemical knowledge, the MOLPROBITY structure validation tool (4) compares van der Waals contacts, hydrogen-bond distances, side-chain rotamers, and peptide backbone conformation with tabulated values from high-resolution protein structures. The overall MOLPROBITY score was calibrated against the PDB to reflect the x-ray diffraction resolution that, on average, is needed to produce a structure of a given quality. For example, the average MOLPROBITY score for the original seven PCNA models indicates structure quality consistent with 2.86 Å diffraction data, which is near the actual 2.96 Å experimental resolution of the data. MOLPROBITY clash scores were corrected based on experimental evidence (46) and quantum mechanical calculations for the optimal CH…O hydrogen-bond distance (47). Although the ideal distance is reported to be 2.3 Å, MOLPROBITY incorrectly reports this separation as a clash (31). The corrected scores are denoted uwith a footnote as MOLPROBITY[a].

The optimization procedure used here operates on a hybrid target function based on maximum-likelihood principles (48). The target function (E_{Tot}) is composed of a weighted sum of force-field (25,49,50) (E_{chem}) and x-ray ($E_{x\text{-}ray}$) energy terms, where the latter is a measure of the agreement between a real-space map and the electron density of the model:

$$E_{Tot} = E_{chem} + w_A E_{x\text{-}ray}. \quad (16)$$

Calculation of real-space density maps followed the formalism of Read (51) and implementation of Cowtan (52) to compute σ_A and figure-of-merit coefficients for structure factors. Real-space density values at specific coordinates were computed using a Catmull-Rom spline ($\tau = 0.25$). OPLS-AA/L and AMOEBA electrostatics were evaluated using particle-mesh Ewald summation as described previously in Schnieders et al. (31).

Methods

The rotamer elimination criteria were implemented in the FORCE FIELD X (FFX) molecular biophysics software package (http://ffx.biochem.uiowa.edu) (31,33) and applied in an iterative fashion, such that rounds of rotamer and rotamer-pair elimination were performed until no new eliminations were produced. The target function for all remaining permutations was then evaluated to determine the GMEC. For all AMOEBA stages of this work, the electron density and potential energy terms were weighted equally (33). The electron density weight was doubled for OPLS-AA/L refinements ($w_A = 2$) as this was observed to yield output structures with a better balance between R and other quality metrics.

Seven structures of PCNA were optimized according to the following protocol: input structures were first minimized (in coordinates and temperature factors) to remove clashes, and then the coordinates of each side chain were recorded. Each unit cell was divided into subvolumes with axis lengths of 4 Å that were placed with 3 Å overlap between neighboring subvolumes. Residues were placed into any box containing their C_α atom. Side-chain optimization via DEE was performed on each box using the Richardson rotamer library augmented by the initial coordinates of each residue as an additional rotameric choice (Ponder and Richards (8)). Pairwise DEE was applied for OPLS-AA/L, while many-body DEE truncated after trimer interactions was used for AMOEBA. After another round of minimization in both coordinates and temperature factors, residues that remained in poor rotameric positions were optimized individually using the same criteria, but without using the initial coordinates as a rotameric choice. This final step was performed iteratively until no further improvement in structure quality metrics was achieved, which accounted for <5% of the final side-chain positions.

A conservative approximation was employed to significantly reduce the computational expense of applying the elimination criteria. All rotamers whose self-energy was 30 kcal/mol larger than that residue's self-energy minimum were pruned before continuing on to two- and three-body calculations. This approach is based on the observation that rotamers with self-energy disparities of this magnitude, which often arise from side-chain van der Waals clashes with the protein backbone, are inconsistent with a well-packed GMEC (11). Such prunable rotamers are also eliminated during application of the rotamer elimination criterion; however, removing them immediately after self-energy calculation drastically reduces the required number of two- and three-body energies. This pruning strategy produced even more benefit under a hybrid target function because in many cases the density map is well fit by only a handful of rotamers.

RESULTS AND DISCUSSION

Atomic-resolution quality from mid- to low-resolution diffraction

PCNA data sets (Table 1) are ideal for demonstrating the ability of the refinement approaches described above to achieve atomic-resolution structural quality from mid- to low-resolution diffraction data. Overall, many-body DEE with the AMOEBA polarizable force field yielded higher quality PCNA models than PDB_REDO, local minimization, or traditional two-body DEE using OPLS-AA/L (Table 2). Although each strategy was able to improve the original PDB models, many-body DEE displayed the most significant gains across all major quality metrics. Mean improvement in R_{free} was 3.0 for AMOEBA DEE versus 2.5 for pairwise OPLS-AA/L DEE and almost no reduction for PDB_REDO. The locally minimized structures were used as a baseline for comparing force-field energy, against which both OPLS-AA/L and AMOEBA DEE models

TABLE 1 The PDB ID, resolution, R/R_{free} values, and MOLPROBITY analyses for the deposited PCNA models

		Reported		FFX		MOLPROBITY[a]		Clash[a]		Ramachandran		Poor
Data Set	Resolution (Å)	R	R_{free}	R	R_{free}	Score	%	Score	%	Out (%)	Favorable (%)	Rotamer (%)
3F1W	2.90	22.8	25.5	23.5	25.9	2.81	81	35.3	65	0.4	95.2	3.9
3GPM	3.80	27.5	31.2	35.4	34.3	3.43	73	52.9	51	4.0	89.3	7.5
3GPN	2.50	23.6	27.3	23.8	27.3	2.19	91	11.8	92	0.0	98.0	6.2
3L0W	2.80	27.9	31.4	31.5	33.2	3.57	23	51.0	20	0.0	92.3	15.8
3L0X	3.00	24.4	26.7	24.3	25.7	2.79	86	15.0	97	0.0	94.4	9.2
3L10	2.80	27.9	31.4	31.8	34.4	3.56	23	51.8	20	0.0	92.3	15.1
WT	2.95	24.6	27.3	24.9	27.3	1.65	100	5.5	100	0.4	94.9	0.9
Mean	2.96	25.5	28.7	27.9	29.7	2.86	68	31.9	64	0.7	93.8	8.4

[a]See main text for explanation.

TABLE 2 The R/R_{free} values, change in force-field potential energy, and MOLPROBITY analyses for the PCNA data sets are given for PDB_REDO, OPLS-AA/L, and AMOEBA refinement methods

					MOLPROBITY[a]		Clash[a]		Ramachandran		Poor
Data Set	Model	R	R_{free}	E_{FF}	Score	%	Score	%	Out (%)	Favorable (%)	Rotamer (%)
3F1W	PDB_REDO	27.01	29.28		1.98	99	3.7	100	0.0	96.0	5.2
	OPLS-AA/L	23.78	26.86		1.78	100	1.0	100	0.8	93.3	5.2
	+ DEE	23.99	26.21	133	1.67	100	1.2	100	0.8	93.7	3.5
	AMOEBA	21.94	26.11		1.39	100	0.5	100	0.4	96.4	4.4
	+ DEE	22.12	26.25	−129	1.03	100	1.2	100	0.4	96.4	0.4
3GPM	PDB_REDO	30.14	32.01		2.68	96	8.5	97	4.8	86.9	6.1
	OPLS-AA/L	26.35	28.25		1.94	100	0.5	100	4.8	84.5	6.1
	+ DEE	23.88	26.32	40	1.51	100	0.5	100	3.2	85.3	1.8
	AMOEBA	29.53	30.43		2.13	99	1.3	100	5.6	84.9	6.6
	+ DEE	24.42	27.25	−110	1.33	100	0.0	100	4.4	85.7	1.8
3GPN	PDB_REDO	23.60	27.24		2.02	95	2.8	100	0.4	96.0	7.9
	OPLS-AA/L	21.26	25.60		1.86	97	1.8	100	0.4	95.6	6.6
	+ DEE	21.65	26.04	−242	1.41	100	1.3	100	0.8	95.6	2.2
	AMOEBA	20.75	25.19		2.07	94	2.3	100	0.0	96.4	12.3
	+ DEE	20.98	25.59	−351	1.28	100	0.5	100	0.0	96.4	3.1
3L0W	PDB_REDO	28.58	32.45		3.14	53	30.1	70	4.0	86.7	5.4
	OPLS-AA/L	30.07	31.50		1.99	99	1.0	100	1.2	93.2	10.1
	+ DEE	28.17	30.80	−316	1.86	99	0.8	100	1.9	92.3	7.1
	AMOEBA	27.09	29.71		2.29	96	2.9	100	1.2	92.3	9.4
	+ DEE	27.12	29.60	−260	1.17	100	1.0	100	0.6	94.7	1.7
3L0X	PDB_REDO	22.26	23.53		2.71	89	8.2	97	1.6	91.2	9.6
	OPLS-AA/L	21.33	24.27		2.05	99	1.3	100	1.6	93.6	10.9
	+ DEE	20.47	24.78	−334	1.51	100	0.8	100	1.6	92.8	2.6
	AMOEBA	20.92	24.10		2.22	98	3.0	100	0.8	92.4	7.4
	+ DEE	20.98	24.43	−125	1.25	100	1.8	100	1.6	92.8	0.4
3L10	PDB_REDO	28.01	32.04		3.35	37	43.8	33	5.0	84.2	5.4
	OPLS-AA/L	30.41	33.08		1.79	100	0.4	100	2.2	92.9	8.4
	+ DEE	27.04	30.65	−281	1.67	100	0.6	100	1.6	91.0	4.0
	AMOEBA	27.58	30.20		2.25	97	3.3	100	0.6	92.0	7.1
	+ DEE	26.82	29.56	−360	1.59	100	1.0	100	0.6	91.6	3.4
WT	PDB_REDO	28.71	30.15		3.06	68	14.6	96	0.8	92.1	15.6
	OPLS-AA/L	23.89	27.22		1.69	100	0.5	100	1.6	92.9	5.6
	+ DEE	23.21	25.57	−664	1.69	100	1.5	100	2.0	93.3	3.0
	AMOEBA	22.64	25.89		1.73	100	2.0	100	0.8	94.9	3.5
	+ DEE	21.63	24.24	−760	1.09	100	0.7	100	1.2	94.9	0.9
Mean	PDB_REDO	26.90	29.53		2.71	77	16.0	85	2.4	90.1	7.9
	OPLS-AA/L	25.30	28.11		1.87	99	0.9	100	1.8	92.3	7.6
	+ DEE	24.06	27.20	−238	1.62	100	0.9	100	1.7	92.0	3.5
	AMOEBA	24.35	27.38		2.01	98	2.2	100	1.3	92.8	7.2
	+ DEE	23.44	26.70	−299	1.25	100	0.9	100	1.3	93.2	1.7

All R/R_{free} values were calculated in FFX for consistency. Potential energy (E_{ff}) after DEE repacking is reported relative to the energy after local minimization (kcal/mol).

[a]See main text for explanation.

were favored by an average of >200 kcal/mol per structure. These large increases in stability may favorably affect downstream computational methods such as molecular-dynamics simulations, which generally begin from a crystal structure after local optimization using a chosen force field, but without side-chain repacking. A more targeted analysis on the effects of the three-body term is available in Table S3 in the Supporting Material, which compares structure quality for the two- versus three-body approximation under AMOEBA. Three-body optimization under AMOEBA is shown to yield additional improvements not obtained by any other combination.

The MOLPROBITY score was improved by ~1.0 using local minimization alone, by 1.24 using pairwise OPLS-AA/L rotamer optimization, and by 1.61 using many-body AMOEBA rotamer optimization. The latter placed all seven structures in the 100th MOLPROBITY percentile among structures of this resolution range (3.0 Å). Clash score was improved to the 100th percentile by all methods except PDB_REDO. The deposited structures averaged poor rotameric positions for 8.4% of side chains, which was not significantly improved by local minimization methods. DEE using both pairwise OPLS-AA/L and many-body AMOEBA algorithms reduced the percent of poor rotamers

by 4.9 and 6.7%, respectively. Although both DEE methods yielded marked improvements in most structure quality metrics, the many-body AMOEBA improvements were greatest. Relative to OPLS-AA/L pairwise DEE, AMOEBA DEE shows mean additional improvements of a lower R_{free} by 0.5, lower MOLPROBITY score by 0.37, 1.8% fewer poor rotamers, better clash score, and more favorable Ramachandran values. These additional improvements are driven by the inclusion of many-body polarization and atomic multipole electrostatics, which are critically important to capture the bifurcated hydrogen bonding that stabilizes both α-helical and β-sheet secondary structure (Fig. 3 and Fig. S1 in the Supporting Material) (46). As shown in Table 3, truncation of the many-body expansion at pairwise interactions neglects ~1 kcal/mol/residue of interaction energy under the AMOEBA polarizable force field.

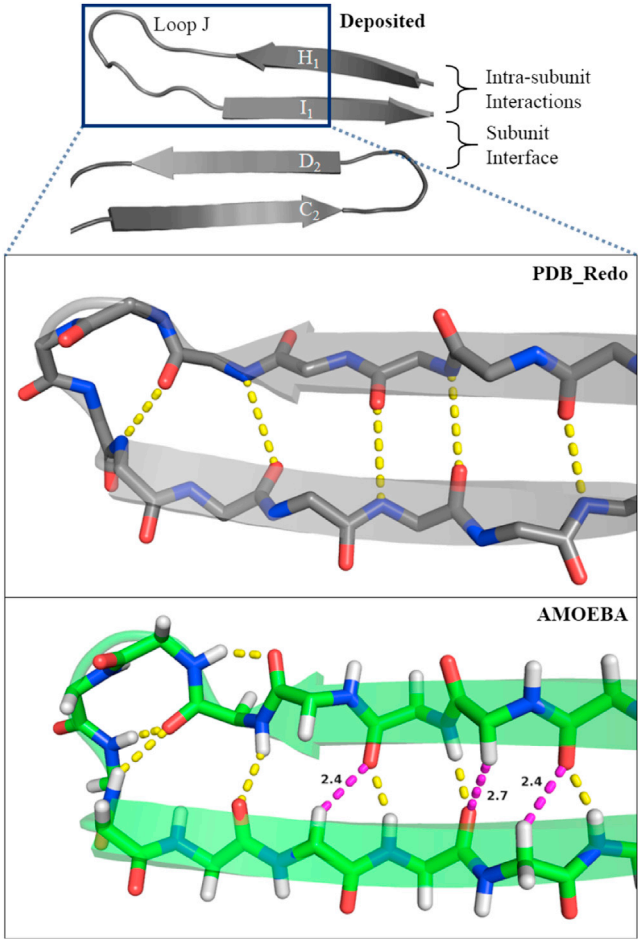


FIGURE 3 The PCNA H_1 and I_1 β-strands and loop *J* backbone electrostatic network for 3GPM from PDB_REDO and from the AMOEBA DEE refinement in FFX. The intrastrand interaction is stabilized by five backbone hydrogen bonds in the PDB_REDO structure (*above*), while the AMOEBA model (*below*) contains seven. The AMOEBA model has three additional low-energy $C_\alpha H\cdots O$ bonds (*purple*). The bifurcated hydrogen bonding is driven largely by the quadrupole moment of the carbonyl oxygen. See Fig. S1 for similar improvements in α-helical structure.

Fortunately, truncation at three-body interactions neglects <0.1 kcal/mol/residue, which is a reasonable compromise between efficiency and residual error due to higher-order neglected interactions (i.e., four-body and higher). Distributions of self, pair, and three-body energies for the wild-type structure, as well as distributions of slack (i.e., the amount of energy by which the elimination criterion was exceeded), are available in Fig. S4. Ninety-percent of three-body energies by absolute magnitude are greater than only 0.04% of self-elimination slacks and none of the pair slacks. The largest individual three-body energy, however, is >10 kcal/mol. We thus expect that there exist individual fourth- and higher-order energies (at short distances) with significant impact on elimination, but calculation of fourth-order energies represents an infeasible computational cost for structures of PCNA's size. Comparison of run times for the methods tested herein is available in Table S4.

Structural insights into the relative stability of PCNA mutants

The newly refined AMOEBA models provide structural and mechanistic insights that are supported by the x-ray diffraction data, but were not achieved in the original models due to limitations in available refinement algorithms. To demonstrate this, we now focus on E113G and G178S TLS-deficient separation-of-function PCNA mutants (38,53). These substitutions, E113G in β-strand I_1 and G178S in β-strand D_2, are at the subunit interface of PCNA, where antiparallel strand interactions between I_1 and D_2 stabilize the PCNA trimer (36). The original structural models demonstrated partial separation of these β-strands in both mutant proteins relative to the wild-type protein (40). In addition, biochemical studies showed that both mutant proteins have significantly reduced trimer stability relative to the wild-type protein, which is responsible for their inability to support TLS (54).

The structural basis for the separation of β I_1 and β D_2 was indicated by the original model of the G178S mutant protein. The side-chain hydroxyl group on substituted Ser^{178} (on β D_2) forms a new hydrogen bond with the backbone carbonyl of Glu^{113} (on β I_1), and this interaction alters the trajectory of β I_1 in the mutant protein (40). By contrast, the structural basis for the strand separation was not clear from the original model of the E113G mutant protein. The newly refined models, however, have provided what are, to our knowledge, novel insights into how the E113G substitution alters the structure of the subunit interface.

In the AMOEBA side-chain optimized model of the E113G mutant, the interaction between β I_1 and β H_1 is stabilized by increased hydrogen bonding. Comparing the structures of the wild-type and E113G mutant protein, we see that β H_1 is extended by one residue (position 105) and that β I_1 is extended by two residues (positions 109

TABLE 3 Neglected higher-order energy when truncating at two- or three-body interactions when using the AMOEBA force field

						$E_{neglected}$		$E_{neglected}$/Residue	
Data Set	E_{total}	$E_{backbone}$	ΣE_{self}	ΣE_{pair}	ΣE_{trimer}	Two-Body	Three-Body	Two-Body	Three-Body
3F1W	−7303.7	−2392.7	−2737.2	−2387.8	230.0	213.9	−16.0	0.8	−0.1
3GPM	−7777.1	−2459.3	−2886.1	−2687.5	283.0	255.7	−27.3	1.0	−0.1
3GPN	−7618.3	−2838.6	−2368.5	−2628.5	232.7	217.3	−15.4	0.9	−0.1
Wild-type	−7591.7	−2385.2	−2685.5	−2765.7	269.6	244.7	−24.9	1.0	−0.1
Mean						232.9	−20.9	0.9	−0.1

Total energy neglected when truncating the expansion after pairwise interactions is ~1 kcal/mol/residue. By contrast, truncation after three-body interactions reduces the neglected energy by an order of magnitude to <0.1 kcal/mol/residue.

and 110) in the mutant protein. In addition, the loop between β H_1 and β I_1 (loop *J*) appears to be in a more energetically favorable conformation in the mutant proteins. There are three intrastrand backbone hydrogen bonds in the mutant protein that are not present in AMOEBA side-chain optimized wild-type protein. A similar mechanism was proposed for a loss-of-flexibility S115P mutant (36) that caused trimer instability due to loss of interstrand hydrogen bonds. In wild-type PCNA, the trimeric form is more stable than the monomeric form by 1667 kcal/mol of AMOEBA energy; this stabilization drops to 1424 kcal/mol in the E113G mutant. Our results suggest a similar mechanism for the gain-of-flexibility E113G mutant based on an energetic tradeoff between the intermolecular interactions of β D_2 and β I_1 at the subunit interface and intramolecular interactions between β H_1 and β I_1 and within the backbone of loop *J*. The greater flexibility of β I_1 due to introduction of glycine at position 113 has shifted the balance in favor of stronger intramolecular β H_1-β I_1 interactions and loop-*J* hydrogen bonds (see Fig. 3 and Table S1). This is a possible explanation for the observed separation of the subunit interface and is consistent with reduced trimer stability. The AMOEBA PCNA electrostatic networks at the subunit interface are supported not only by lower MOLPROBITY score and lower R/R_{free} values, but also by dramatically cleaner σ_A-weighted F_o-F_c electron density maps (Fig. 4).

CONCLUSIONS

Biomolecular x-ray refinement strategies that place side chains, such as PDB_REDO (34) and RINGER (55), have achieved some success in improving the quality and interpretation of x-ray diffraction experiments. However, protein structure refinement methods have been limited by their assumption of side-chain independence and/or the absence of rigorous electrostatic interactions. For example, PDB_REDO is based on choosing a rotameric state for one residue at a time (56), which is reflected by a mean poor rotamer percentage of 6.6% for the PCNA structures examined here. On the other hand, many-body DEE using AMOEBA reduced the percentage of poor rotamers to 1.7% while simultaneously improving overall MOLPROBITY score and lowering both R_{free} and AMOEBA potential energy.

Model bias is an important consideration for any refinement procedure that optimizes atomic coordinates to a target function that depends on calculated phases. Neither systematic removal of backbone model bias nor optimization of backbone conformation beyond what is achieved by local minimization was considered in this work. However, several methods have been proposed for considering limited backbone flexibility during repacking, which could be coupled to many-body DEE in the future. For example, generation of a discrete set of backbone conformations to include during DEE has been described in Su and Mayo (57). Alternatively, deterministic DEE has been extended to find a flexible-backbone rigid-rotamer GMEC by calculating bounds on rotameric interaction energies given a limited range of backbone dihedral movements imposed by per-residue restraining boxes (58).

The side-chain repacking algorithm presented here, to our knowledge, is the first deterministic method compatible with many-body potential energy functions. This opens the door to using polarizable force fields, Poisson-Boltzmann electrostatics, and quantum mechanical potential energy functions alone or in combination with experimental data to improve protein structural models. In this work, a hybrid target function has shown success in improving MOLPROBITY score and lowering both R_{free} and AMOEBA potential energy based on a series of mid- to low-resolution PCNA x-ray diffraction data sets. In this case, electrostatic networks from coupled side-chain reorientations, which are difficult or impossible to refine by hand, revealed intramolecular stabilization of PNCA monomers at the expense of intermolecular hydrogen-bonding and destabilization of the active PNCA trimer.

In addition to x-ray structure determination, this work sets the foundation for application of many-body potential energy functions to computational protein design, homology modeling, and design of protein-ligand interactions. The advantage of many-body over pairwise DEE is of greatest importance for driving molecular forces that are inherently many-body in nature, including polarizable electrostatic interactions and the hydrophobic effect. For example, it has been suggested that the inherent many-body nature of the

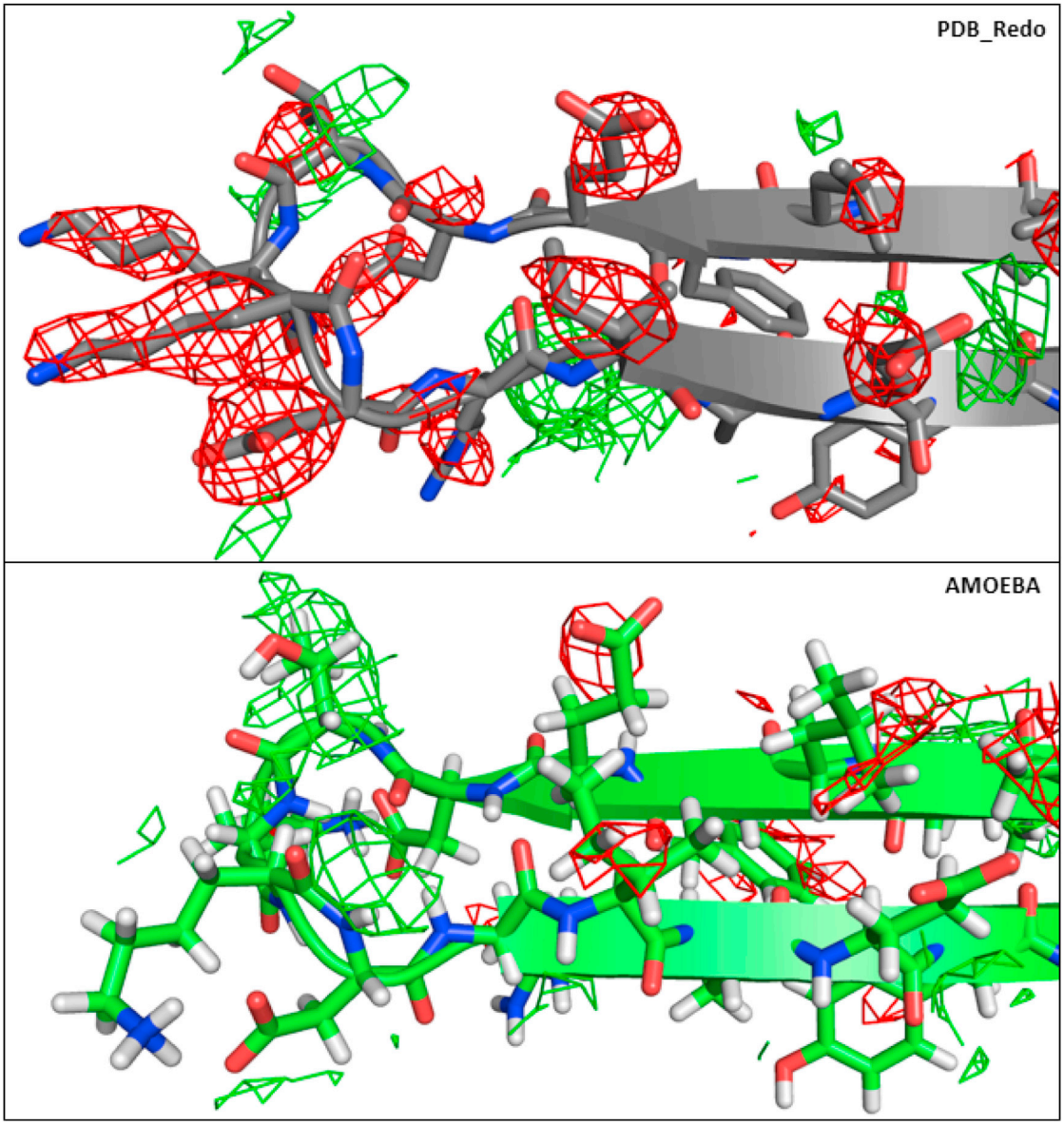


FIGURE 4 The β-strands H_1 and I_1 of PCNA G178S mutant (3F1W) are shown with F_o-F_c maps contoured at 2σ (*green*) and -2σ (*red*) for PDB_REDO (*above*) and AMOEBA (*below*).

hydrophobic effect has made computational protein design a challenge for implicit solvents (59). Future applications of many-body DEE may help determine whether the use of polarizable force fields (26) and self-consistent reaction-field implicit solvents (28–30) can overcome the limitations of previous generation pairwise force fields (14) and pairwise implicit solvents (21–23) for computational protein design (15–17).

SUPPORTING MATERIAL

Supporting Materials and Methods, three figures, four tables, and derivations of many-body dead-end elimination criteria are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(15)00676-1.

AUTHOR CONTRIBUTIONS

S.D.L., S.G., W.T.A.T., and M.J.S. conceived the theory; S.D.L. performed the experiments; S.D.L., J.M.L., K.T.P., M.T.W., and M.J.S. analyzed the data; S.D.L., J.M.L., K.T.P., A.M.L., W.T.A.T., T.D.F., M.T.W., and M.J.S. contributed code/tools/structures; and S.D.L., J.M.L., K.T.P., M.T.W., and M.J.S. wrote the article.

ACKNOWLEDGMENTS

The authors thank Lokesh Gakhar and Lynne Dieckman for helpful discussions. All computations were performed on The University of Iowa NEON cluster with support and guidance from Glenn Johnson and Ben Rogers.

M.J.S. was supported by National Science Foundation award No. CHE-1404147 and National Institutes of Health award No. R01 DC002842

from The National Institute on Deafness and Other Communication Disorders. M.T.W. was supported by National Institutes of Health award No. 01-GM081433 from the National Institute of General Medical Sciences. S.D.L. acknowledges a National Institutes of Health fellowship from award No. T32-GM067795. J.M.L. acknowledges a National Institutes of Health fellowship from award No. T32-GM008365 and a University of Iowa Presidential Fellowship. S.G. acknowledges support from a University of Iowa Biochemistry Summer Undergraduate Research Fellowship. A.M.L. and W.T.A.T. were partially supported by fellowships from The University of Iowa Center for Research by Undergraduates.

REFERENCES

1. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
2. Adams, P. D., P. V. Afonine, …, P. H. Zwart. 2010. PHENIX: a comprehensive PYTHON-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* 66:213–221.
3. Winn, M. D., C. C. Ballard, …, K. S. Wilson. 2011. Overview of the CCP4 suite and current developments. *Acta Crystallogr. D Biol. Crystallogr.* 67:235–242.
4. Chen, V. B., W. B. Arendall, 3rd, …, D. C. Richardson. 2010. MOLPROBITY: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* 66:12–21.
5. Ponder, J. W., and D. A. Case. 2003. Force fields for protein simulations. *In* Advances in Protein Chemistry, *Vol. 66*. Academic Press, London, UK, pp. 27–85.
6. Tama, F., O. Miyashita, and C. L. Brooks, 3rd. 2004. Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J. Mol. Biol.* 337: 985–999.
7. Schröder, G. F., M. Levitt, and A. T. Brunger. 2010. Super-resolution biomolecular crystallography with low-resolution data. *Nature.* 464:1218–1222.
8. Ponder, J. W., and F. M. Richards. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J. Mol. Biol.* 193:775–791.
9. Lovell, S. C., J. M. Word, …, D. C. Richardson. 2000. The penultimate rotamer library. *Proteins.* 40:389–408.
10. Shapovalov, M. V., and R. L. Dunbrack, Jr. 2011. A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. *Structure.* 19:844–858.
11. Desmet, J., M. De Maeyer, …, I. Lasters. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature.* 356:539–542.
12. Goldstein, R. F. 1994. Efficient rotamer elimination applied to protein side-chains and related spin glasses. *Biophys. J.* 66:1335–1340.
13. Dahiyat, B. I., and S. L. Mayo. 1997. De novo protein design: fully automated sequence selection. *Science.* 278:82–87.
14. Boas, F. E., and P. B. Harbury. 2007. Potential energy functions for protein design. *Curr. Opin. Struct. Biol.* 17:199–204.
15. Das, R., and D. Baker. 2008. Macromolecular modeling with ROSETTA. *Annu. Rev. Biochem.* 77:363–382.
16. Gainza, P., K. E. Roberts, …, B. R. Donald. 2013. OSPREY: protein design with ensembles, flexibility, and provable algorithms. *Methods Enzymol.* 523:87–107.
17. Simonson, T., T. Gaillard, …, G. Archontis. 2013. Computational protein design: the PROTEUS software and selected applications. *J. Comput. Chem.* 34:2472–2484.
18. Kaminski, G. A., R. A. Friesner, …, W. L. Jorgensen. 2001. Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys. Chem. B.* 105:6474–6487.
19. Cornell, W. D., P. Cieplak, …, P. A. Kollman. 1996. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 118:2309.
20. MacKerell, A. D., D. Bashford, …, M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616.
21. Lazaridis, T., and M. Karplus. 1999. Effective energy function for proteins in solution. *Proteins.* 35:133–152.
22. Marshall, S. A., C. L. Vizcarra, and S. L. Mayo. 2005. One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. *Protein Sci.* 14:1293–1304.
23. Gaillard, T., and T. Simonson. 2014. Pairwise decomposition of an MMGBSA energy function for computational protein design. *J. Comput. Chem.* 35:1371–1387.
24. Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *In* Advances in Protein Chemistry, *Vol. 14* Anfinsen, Jr., C. B., M. L. Anson, K. Bailey, and J. T. Edsall, editors. Academic Press, London, UK.
25. Ponder, J. W., C. Wu, …, T. Head-Gordon. 2010. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B.* 114:2549–2564.
26. Lopes, P. E. M., B. Roux, and A. D. MacKerell, Jr. 2009. Molecular modeling and dynamics studies with explicit inclusion of electronic polarizability. Theory and applications. *Theor. Chem. Acc.* 124:11–28.
27. Tomasi, J., B. Mennucci, and R. Cammi. 2005. Quantum mechanical continuum solvation models. *Chem. Rev.* 105:2999–3093.
28. Maple, J. R., Y. X. Cao, …, R. A. Friesner. 2005. A polarizable force field and continuum solvation methodology for modeling of protein-ligand interactions. *J. Chem. Theory Comput.* 1:694–715.
29. Schnieders, M. J., and J. W. Ponder. 2007. Polarizable atomic multipole solutes in a generalized Kirkwood continuum. *J. Chem. Theory Comput.* 3:2083–2097.
30. Schnieders, M. J., N. A. Baker, …, J. W. Ponder. 2007. Polarizable atomic multipole solutes in a Poisson-Boltzmann continuum. *J. Chem. Phys.* 126:124114.
31. Schnieders, M. J., T. D. Fenn, and V. S. Pande. 2011. Polarizable atomic multipole x-ray refinement: particle mesh Ewald electrostatics for macromolecular crystals. *J. Chem. Theory Comput.* 7:1141–1156.
32. Fenn, T. D., M. J. Schnieders, and A. T. Brunger. 2010. A smooth and differentiable bulk-solvent model for macromolecular diffraction. *Acta Crystallogr. D Biol. Crystallogr.* 66:1024–1031.
33. Fenn, T. D., and M. J. Schnieders. 2011. Polarizable atomic multipole x-ray refinement: weighting schemes for macromolecular diffraction. *Acta Crystallogr. D Biol. Crystallogr.* 67:957–965.
34. Joosten, R. P., J. Salzemann, …, G. Vriend. 2009. PDB_REDO: automated re-refinement of x-ray structure models in the PDB. *J. Appl. Cryst.* 42:376–384.
35. Moldovan, G.-L., B. Pfander, and S. Jentsch. 2007. PCNA, the maestro of the replication fork. *Cell.* 129:665–679.
36. Krishna, T. S. R., X.-P. Kong, …, J. Kuriyan. 1994. Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA. *Cell.* 79:1233–1243.
37. Stelter, P., and H. D. Ulrich. 2003. Control of spontaneous and damage-induced mutagenesis by SUMO and ubiquitin conjugation. *Nature.* 425:188–191.
38. Amin, N. S., and C. Holm. 1996. In vivo analysis reveals that the interdomain region of the yeast proliferating cell nuclear antigen is important for DNA replication and DNA repair. *Genetics.* 144:479–493.
39. Lau, P. J., H. Flores-Rozas, and R. D. Kolodner. 2002. Isolation and characterization of new proliferating cell nuclear antigen (POL30) mutator mutants that are defective in DNA mismatch repair. *Mol. Cell. Biol.* 22:6669–6680.
40. Freudenthal, B. D., S. Ramaswamy, …, M. T. Washington. 2008. Structure of a mutant form of proliferating cell nuclear antigen that blocks translesion DNA synthesis. *Biochemistry.* 47:13354–13361.

41. Freudenthal, B. D., L. Gakhar, …, M. T. Washington. 2010. Structure of monoubiquitinated PCNA and implications for translesion synthesis and DNA polymerase exchange. *Nat. Struct. Mol. Biol.* 17:479–484.

42. Dieckman, L. M., E. M. Boehm, …, M. T. Washington. 2013. Distinct structural alterations in proliferating cell nuclear antigen block DNA mismatch repair. *Biochemistry.* 52:5611–5619.

43. Brünger, A. T., J. Kuriyan, and M. Karplus. 1987. Crystallographic *R* factor refinement by molecular dynamics. *Science.* 235:458–460.

44. Moulinier, L., D. A. Case, and T. Simonson. 2003. Reintroducing electrostatics into protein x-ray structure refinement: bulk solvent treated as a dielectric continuum. *Acta Crystallogr. D Biol. Crystallogr.* 59:2094–2103.

45. Brünger, A. T. 1992. FREE *R* VALUE: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature.* 355:472–475.

46. Ho, B. K., and P. M. G. Curmi. 2002. Twist and shear in β-sheets and β-ribbons. *J. Mol. Biol.* 317:291–308.

47. Vargas, R., J. Garza, …, B. P. Hay. 2000. How strong is the C_α–H$\cdots$OC hydrogen bond? *J. Am. Chem. Soc.* 122:4750–4755.

48. Murshudov, G. N., A. A. Vagin, and E. J. Dodson. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D Biol. Crystallogr.* 53:240–255.

49. Ren, P., C. Wu, and J. W. Ponder. 2011. Polarizable atomic multipole-based molecular mechanics for organic molecules. *J. Chem. Theory Comput.* 7:3143–3161.

50. Shi, Y., Z. Xia, …, P. Ren. 2013. The polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory Comput.* 9:4046–4063.

51. Read, R. 1986. Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. A.* 42:140–149.

52. Cowtan, K. 2005. Likelihood weighting of partial structure factors using spline coefficients. *J. Appl. Cryst.* 38:193–198.

53. Zhang, H., P. E. M. Gibbs, and C. W. Lawrence. 2006. The *Saccharomyces cerevisiae* rev6-1 mutation, which inhibits both the lesion bypass and the recombination mode of DNA damage tolerance, is an allele of POL30, encoding proliferating cell nuclear antigen. *Genetics.* 173:1983–1989.

54. Dieckman, L. M., and M. T. Washington. 2013. PCNA trimer instability inhibits translesion synthesis by DNA polymerase η and by DNA polymerase δ. *DNA Repair (Amst.).* 12:367–376.

55. Fraser, J. S., M. W. Clarkson, …, T. Alber. 2009. Hidden alternative structures of proline isomerase essential for catalysis. *Nature.* 462:669–673.

56. Joosten, R. P., K. Joosten, …, A. Perrakis. 2011. Automatic rebuilding and optimization of crystallographic structures in the Protein Data Bank. *Bioinformatics.* 27:3392–3398.

57. Su, A., and S. L. Mayo. 1997. Coupling backbone flexibility and amino acid sequence selection in protein design. *Protein Sci.* 6:1701–1707.

58. Georgiev, I., D. Keedy, …, B. R. Donald. 2008. Algorithm for backrub motions in protein design. *Bioinformatics.* 24:i196–i204.

59. Jaramillo, A., and S. J. Wodak. 2005. Computational protein design is a challenge for implicit solvation models. *Biophys. J.* 88:156–171.

Biophysical Journal

Supporting Material

Dead-End Elimination with a Polarizable Force Field Repacks PCNA Structures

Stephen D. LuCore,[1] Jacob M. Litman,[2] Kyle T. Powers,[2] Shibo Gao,[2] Ava M. Lynn,[1] William T. A. Tollefson,[1] Timothy D. Fenn,[3] M. Todd Washington,[2] and Michael J. Schnieders[1,2,*]

[1]Department of Biomedical Engineering and [2]Department of Biochemistry, University of Iowa, Iowa City, Iowa; and [3]Boehringer Ingelheim, Ridgefield, Connecticut

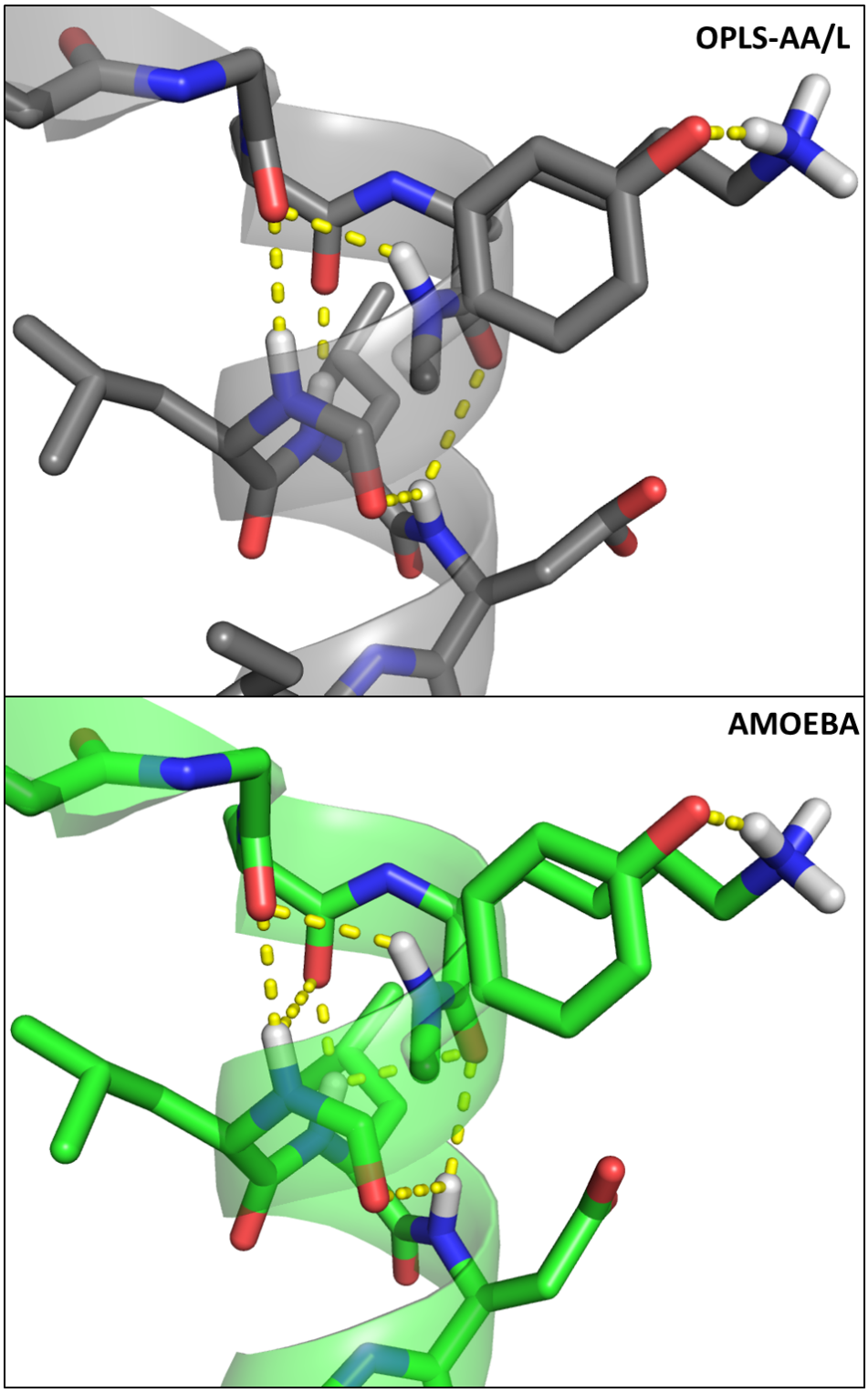
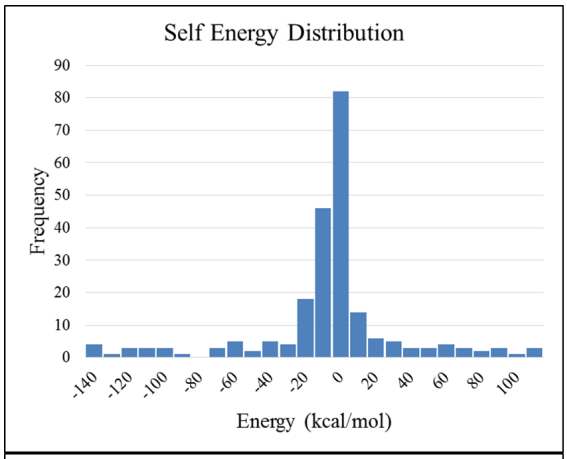
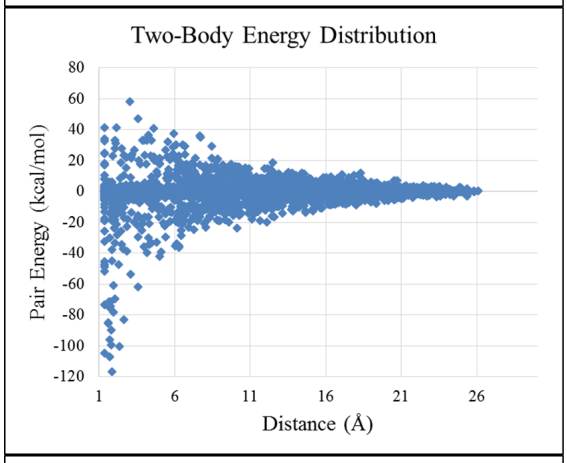
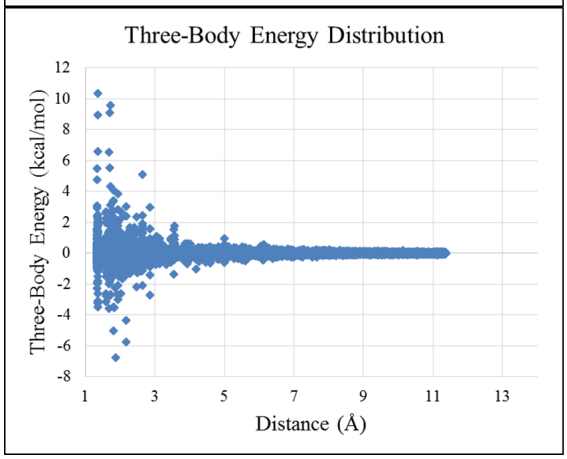
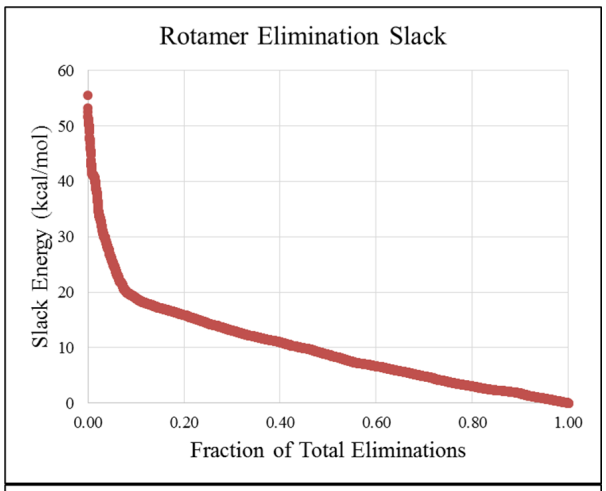


Figure S1. Hydrogen bonding for an α-helix from wild type PCNA is shown after application of OPLS-AA/L pairwise DEE and AMOEBA many-body DEE. Hydrogen atoms not involved in hydrogen bonding have been hidden for clarity. The AMOEBA model (lower panel) shows two additional hydrogen bonds not present in the OPLS-AA/L model (upper panel) due to backbone nitrogen atoms of residue *i* donating to both the residue *i*+3 and *i*+4 carbonyl oxygen atoms.









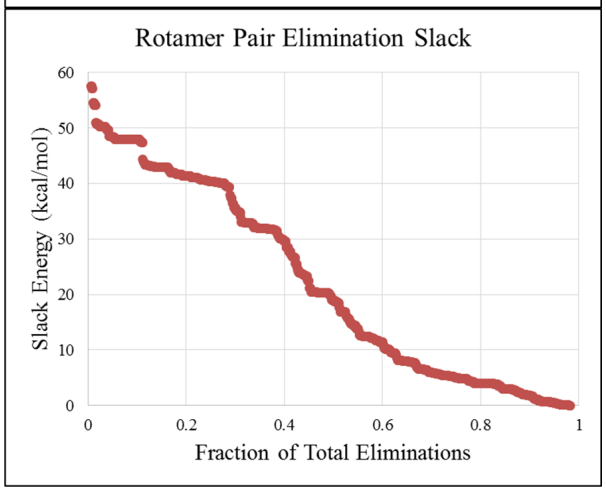


Figure S2. Shown to the left are distributions of self, 2-body, and 3-body energies from the wild type PCNA data as a function of distance. Shown to the right are sorted distributions of slack for rotamer and rotamer pair eliminations under AMOEBA many-body DEE for wild type PCNA (2 values for rotamer slack and 28 values for rotamer pair slack are greater than 60 kcal/mol and were not included). The maximum absolute 3-body energy is 10.4 kcal/mol, which is greater than 57% of the rotamer slack energies and 37% of the rotamer pair slack energies. Although it is infeasible to compute the distribution of all 4-body energies (there are more than 100 million), a subset of 1.9 million establishes their maximum absolute value to be approximately 1 kcal/mol (see Figure S3 below). We note that only 0.06% of rotamer slack energies and 0.06% of rotamer pair slack energies are less than 1 kcal/mol.

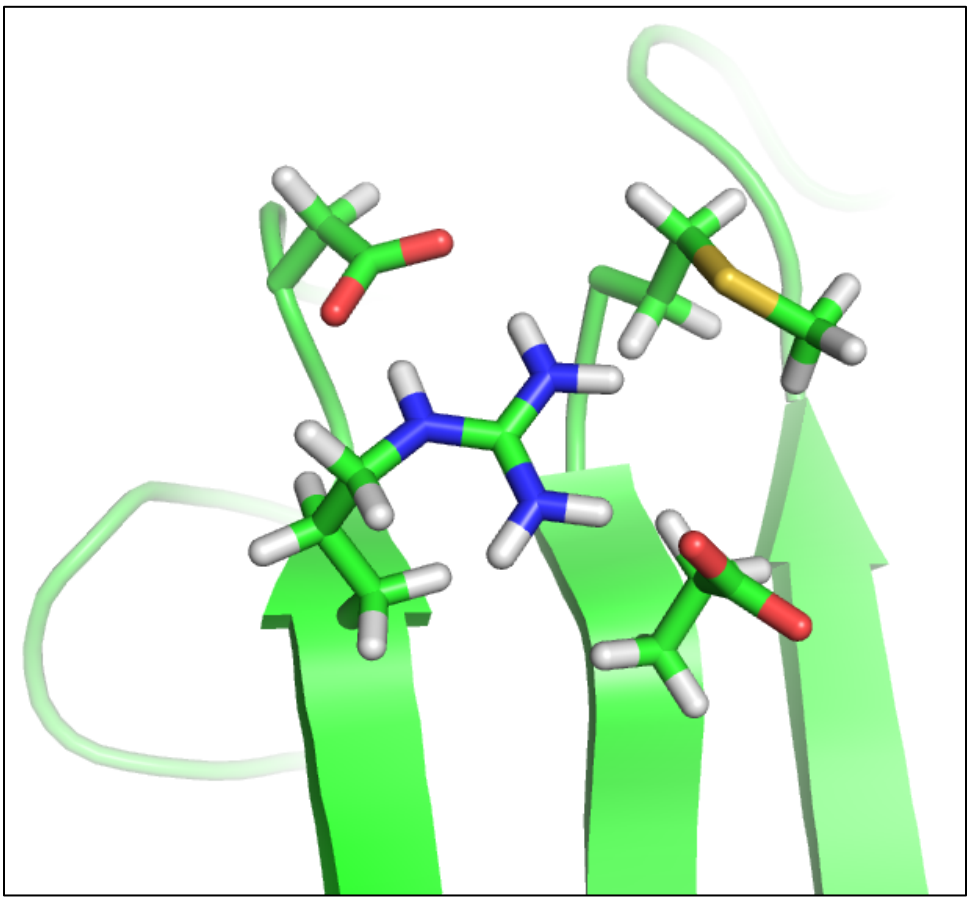


Figure S3. The shown collection of 4 residues from wild type PCNA produced the largest absolute magnitude 4-body energy among the subset of 1.9 million evaluated that contained the N-terminal methionine (-0.98 kcal/mol). The second largest 4-body energy identified was 0.39 kcal/mol and the average 4-body energy was -9e-07 kcal/mol.

Table S1. Shown is quantification of hydrogen bonding at the PCNA subunit interface. Intra-molecular hydrogen bonds are those spanning either βH_1-βI_1 or βC_1-βD_1. Inter-molecular hydrogen bonds are those spanning the subunit interface βI_1-βD_1. The original structural models do not show a clear trend in hydrogen bonding for the E113G and G178S mutants relative to wild type. However, the AMOEBA DEE repacked models reveal an increase in intra-subunit hydrogen bonding in the mutant structures that is consistent with reduced trimer stability.

Data		**Intra-molecular**		**Inter-molecular**		**Delta**
Set	**Model**	**Total**	**Change**	**Total**	**Change**	**Change**
WT	Original	22		5		
	PDB_Redo	20		5		
	OPLS-AA/L	29		5		
	AMOEBA	28		5		
3F1W	Original	22	0	4	-1	+1
(E133G)	PDB_Redo	22	2	4	-1	+3
	OPLS-AA/L	32	3	5	0	+3
	AMOEBA	31	3	6	+1	+2
3GPM	Original	22	0	4	-1	+1
(G178S)	PDB_Redo	19	-1	5	0	-1
	OPLS-AA/L	33	4	6	+1	+3
	AMOEBA	34	6	5	0	+6

Table S2. Shown are coordinate RMSDs relative to deposited coordinates and mean deviations in bonds lengths and angle bends.

Data			RMSD (Å)		Mean Deviations	
Set	Model	Algorithm	All	Side-Chain	Bond (Å)	Angle (^{o})
3F1W	PDB_Redo		0.80	1.12	0.011	2.5
	OPLS-AA/L	Minimize	0.55	0.72	0.010	2.4
		+ DEE	0.81	1.12	0.011	2.4
	AMOEBA	Minimize	0.47	0.61	0.014	2.7
		+ DEE	0.65	0.89	0.013	2.7
3GPM	PDB_Redo		0.47	0.58	0.012	2.5
	OPLS-AA/L	Minimize	1.10	1.40	0.010	2.4
		+ DEE	1.22	1.58	0.010	2.4
	AMOEBA	Minimize	0.98	1.23	0.014	2.7
		+ DEE	1.07	1.36	0.014	2.7
3GPN	PDB_Redo		0.23	0.28	0.012	2.5
	OPLS-AA/L	Minimize	0.63	0.87	0.011	2.3
		+ DEE	0.85	1.19	0.011	2.3
	AMOEBA	Minimize	0.56	0.77	0.014	2.6
		+ DEE	0.74	1.03	0.014	2.6
3L0W	PDB_Redo		0.42	0.46	0.011	2.4
	OPLS-AA/L	Minimize	1.03	1.28	0.009	2.3
		+ DEE	1.20	1.51	0.010	2.3
	AMOEBA	Minimize	0.87	1.05	0.013	2.7
		+ DEE	1.00	1.24	0.012	2.6
3L0X	PDB_Redo		0.29	0.36	0.011	2.5
	OPLS-AA/L	Minimize	0.58	0.73	0.011	2.4
		+ DEE	0.82	1.09	0.011	2.3
	AMOEBA	Minimize	0.48	0.60	0.014	2.7
		+ DEE	0.79	1.07	0.013	2.7
3L10	PDB_Redo		0.45	0.52	0.011	2.4
	OPLS-AA/L	Minimize	1.05	1.30	0.009	2.3
		+ DEE	1.24	1.58	0.010	2.4
	AMOEBA	Minimize	0.85	1.03	0.013	2.7
		+ DEE	1.06	1.31	0.013	2.7
WT	PDB_Redo		1.04	1.42	0.015	3.3
	OPLS-AA/L	Minimize	0.60	0.63	0.011	2.5
		+ DEE	0.95	1.26	0.011	2.5
	AMOEBA	Minimize	0.52	0.63	0.014	2.7
		+ DEE	0.91	1.22	0.014	2.7
Mean	PDB_Redo		0.53	0.68	0.012	2.6
	OPLS-AA/L	Minimize	0.79	0.99	0.010	2.4
		+ DEE	1.03	1.36	0.010	2.4
	AMOEBA	Minimize	0.67	0.84	0.014	2.7
		+ DEE	0.89	1.16	0.013	2.7

Table S3. A comparison of structure quality metrics after AMOEBA DEE refinement using 2-body and 3-body approximations is shown. The AMOEBA 2-body optimization provides higher quality structures than OPLS-AA/L (see Table 2), however, 3-body optimization yields additional improvements beyond those of all other strategies.

PDB					MolProbity		Clash		Ramachandran		Poor
Res.	Refinement	R	R_{free}	E_{FF}	Score	%	Score	%	Out. %	Fav. %	Rotamers %
3f1w	Original	23.46	25.87		2.81	81	35.3	65	0.4	95.2	3.9
2.9 Å	2-body	22.74	27.02	-257	1.56	100	1.0	100	0.8	94.1	3.0
	3-body	22.12	26.25	-129	1.03	100	1.2	100	0.4	96.4	0.4
3gpm	Original	35.42	34.31		3.43	73	52.9	51	4.0	89.3	7.5
3.8 Å	2-body	24.13	27.20	-172	1.41	100	0.0	100	4.8	85.3	2.2
	3-body	24.42	27.25	-110	1.33	100	0.0	100	4.4	85.7	1.8
3gpn	Original	23.81	27.29		2.19	91	11.8	92	0.0	98.0	6.2
2.5 Å	2-body	20.97	25.55	-307	1.58	100	1.8	100	0.0	96.0	3.1
	3-body	20.98	25.59	-351	1.28	100	0.5	100	0.0	96.4	3.1
3l0w	Original	31.45	33.17		3.57	23	51.0	20	0.0	92.3	15.8
2.8 Å	2-body	27.10	29.66	-140	1.34	100	0.4	100	0.6	94.4	2.7
	3-body	27.12	29.60	-260	1.17	100	1.0	100	0.6	94.7	1.7
3l0x	Original	24.27	25.65		2.79	86	15.0	97	0.0	94.4	9.2
3.0 Å	2-body	20.70	24.47	-117	1.37	100	1.7	100	2.0	93.2	0.4
	3-body	20.98	24.43	-125	1.25	100	1.8	100	1.6	92.8	0.4
3l10	Original	31.83	34.36		3.56	23	51.8	20	0.0	92.3	15.1
2.8 Å	2-body	26.88	29.89	-319	1.69	100	0.8	100	0.6	92.6	4.4
	3-body	26.82	29.56	-360	1.59	100	1.0	100	0.6	91.6	3.4
WT	Original	24.89	27.25		1.65	100	5.5	100	0.4	94.9	0.9
3.0 Å	2-body	22.29	24.34	-665	1.15	100	1.0	100	1.2	94.9	0.9
	3-body	21.63	24.24	-760	1.09	100	0.7	100	1.2	94.9	0.9
Mean	Original	27.88	29.70		2.86	68	31.9	64	0.7	93.8	8.4
3.0 Å	2-body	23.54	26.88	-282	1.44	100	0.94	100	1.4	92.9	2.4
	3-body	23.44	26.70	-299	1.25	100	0.88	100	1.3	93.2	1.7
	Δ 3- vs. 2-body	-0.10	-0.18	-17	-0.19	0	-0.06	0	-0.2	0.3	-0.7

Table S4. Run times by structure and method on a single 16-core compute node at 2.6GHz. OPLS-AA/L and AMOEBA are nearly equivalent for minimization due to the X-ray scattering term being the limiting factor. The cost of AMOEBA 3-body DEE is approximately 15x greater than OPLS-AA/L 2-body DEE due to 1) the increased cost of each energy evaluation and 2) computation of 3-body terms.

PDB ID	Force Field	Algorithm	Run Time
3f1w	OPLS-AA/L	Minimize	35 sec
		+ DEE	25 hours
	AMOEBA	Minimize	39 sec
		+ DEE	17 days
3gpm	OPLS-AA/L	Minimize	29 sec
		+ DEE	18 hours
	AMOEBA	Minimize	30 sec
		+ DEE	18 days
3gpn	OPLS-AA/L	Minimize	23 sec
		+ DEE	11 hours
	AMOEBA	Minimize	28 sec
		+ DEE	12 days
3l0w	OPLS-AA/L	Minimize	43 sec
		+ DEE	28 hours
	AMOEBA	Minimize	51 sec
		+ DEE	15 days
3l0x	OPLS-AA/L	Minimize	30 sec
		+ DEE	18 hours
	AMOEBA	Minimize	39 sec
		+ DEE	11 days
3l10	OPLS-AA/L	Minimize	80 sec
		+ DEE	29 hours
	AMOEBA	Minimize	53 sec
		+ DEE	16 days
WT	OPLS-AA/L	Minimize	30 sec
		+ DEE	20 hours
	AMOEBA	Minimize	36 sec
		+ DEE	20 days
Mean	OPLS-AA/L	Minimize	39 sec
		+ DEE	21 hours
	AMOEBA	Minimize	39 sec
		+ DEE	15 days

Supplemental Derivations

I. Many-Body Inclusive Singles Elimination Criterion

We start from the knowledge that any given point in the global rotamer space has a minimum energy equal to the global minimum energy conformation.

$$E_{\text{global}} \geq E_{\text{GMEC}}$$

Equation 1

We denote a superscript g as being the rotamer of a particular residue as it exists in the global minimum energy conformation.

$$E_{\text{global}} = E_{BB} + E(r_i^{\alpha}) + \sum_{j\prime}^{n}\{E(r_j^{g}) + \sum_{k\prime}^{n}[E(r_j^{g}, r_k^{g}) + \sum_{l\prime}^{n}(E(r_j^{g}, r_k^{g}, r_l^{g}) + \cdots)]\} + \sum_{j\prime}^{n}\{E(r_i^{\alpha}, r_j^{g}) + \sum_{k\prime}^{n}[E(r_i^{\alpha}, r_j^{g}, r_k^{g}) + \sum_{l\prime}^{n}(E(r_i^{\alpha}, r_j^{g}, r_k^{g}, r_l^{g}) + \cdots)]\}$$

Equation 2

$$E_{\text{GMEC}} = E_{BB} + E(r_i^{g}) + \sum_{j\prime}^{n}\{E(r_j^{g}) + \sum_{k\prime}^{n}[E(r_j^{g}, r_k^{g}) + \sum_{l\prime}^{n}(E(r_j^{g}, r_k^{g}, r_l^{g}) + \cdots)]\} + \sum_{j\prime}^{n}\{E(r_i^{g}, r_j^{g}) + \sum_{k\prime}^{n}[E(r_i^{g}, r_j^{g}, r_k^{g}) + \sum_{l\prime}^{n}(E(r_i^{g}, r_j^{g}, r_k^{g}, r_l^{g}) + \cdots)]\}$$

Equation 3

Herein E_{BB} is the backbone energy, $E(r_i^{\alpha})$ is the self-energy of residue i in rotamer α, $E\left(r_i^{\alpha}, r_j^{\beta}\right)$ is the two-body energy of residues i, j in rotamers α, β and so on. Self, two-body, and many-body energies are as defined in the main text. Ellipses signify the presence of higher-order terms out to n-body, where n is the number of residues in the system. After explicitly enumerating all energy components including many-body energy, we substitute Eqs. 2 and 3 into Equation 1. Terms without dependence on r_i cancel out. We then find an expression for the remaining portion that doesn't require knowledge of the GMEC conformation.

$$\begin{cases} \max\limits_{s}\left[E(r_i^{\alpha}, r_k^{s}) + \sum\limits_{l\prime}^{n} E(r_i^{\alpha}, r_k^{s}, r_l^{g}) + \cdots\right] \geq E(r_i^{\alpha}, r_k^{g}) + \sum\limits_{l'}^{n} E(r_i^{\alpha}, r_k^{g}, r_l^{g}) + \cdots \\ \min\limits_{s}\left[E(r_i^{g}, r_k^{s}) + \sum\limits_{l\prime}^{n} E(r_i^{g}, r_k^{s}, r_l^{g}) + \cdots\right] \leq E(r_i^{g}, r_k^{g}) + \sum\limits_{l'}^{n} E(r_i^{g}, r_k^{g}, r_l^{g}) + \cdots \end{cases}$$

Equation 4

$$\begin{cases} \max\limits_{t}[E(r_i^{\alpha}, r_k^{s}, r_l^{t}) + \cdots] \geq E\left(r_i^{\alpha}, r_k^{s}, r_l^{g}\right) + \cdots \\ \min\limits_{t}\left[E\left(r_i^{g}, r_k^{s}, r_l^{t}\right) + \cdots\right] \leq E\left(r_i^{g}, r_k^{s}, r_l^{g}\right) + \cdots \end{cases}$$

Equation 5

$$\begin{cases} \max\limits_{s}\left[E(r_i^{\alpha}, r_k^{s}) + \sum\limits_{l\prime}^{n} \max\limits_{t}\left(E(r_i^{\alpha}, r_k^{s}, r_l^{t})\right) + \cdots\right] \geq E\left(r_i^{\alpha}, r_k^{g}\right) + \sum\limits_{l'}^{n} E\left(r_i^{\alpha}, r_k^{g}, r_l^{g}\right) + \cdots \\ \min\limits_{s}\left[E\left(r_i^{g}, r_k^{s}\right) + \sum\limits_{l\prime}^{n} \min\limits_{t}\left(E\left(r_i^{g}, r_k^{s}, r_l^{t}\right)\right) + \cdots\right] \leq E\left(r_i^{g}, r_k^{g}\right) + \sum\limits_{l'}^{n} E\left(r_i^{g}, r_k^{g}, r_l^{g}\right) + \cdots \end{cases}$$

Equation 6

Expressing the substituted Eq. 1 using the left-hand side of Eq. 6 yields the final singles elimination criterion.

II. Many-Body Inclusive Pairwise Elimination Criterion

$$E_{\text{global}} \geq E_{\text{GMEC}}$$

Equation 7

$$\begin{aligned} &E(r_i^{\alpha}) + E\left(r_j^{\beta}\right) + E\left(r_i^{\alpha}, r_j^{\beta}\right) + \textstyle\sum_{k\prime}^{n}\left\{E\left(r_i^{\alpha}, r_k^{g}\right) + E\left(r_j^{\beta}, r_k^{g}\right) + E\left(r_i^{\alpha}, r_j^{\beta}, r_k^{g}\right) + \right. \\ &\left.\textstyle\sum_{l\prime}^{n}\left[E\left(r_i^{\alpha}, r_k^{g}, r_l^{g}\right) + E\left(r_j^{\beta}, r_k^{g}, r_l^{g}\right) + E\left(r_i^{\alpha}, r_j^{\beta}, r_k^{g}, r_l^{g}\right) \ldots\right]\right\} \geq \\ &E\left(r_i^{g}\right) + E\left(r_j^{g}\right) + E\left(r_i^{g}, r_j^{g}\right) + \textstyle\sum_{k\prime}^{n}\left\{E\left(r_i^{g}, r_k^{g}\right) + E\left(r_j^{g}, r_k^{g}\right) + E\left(r_i^{g}, r_j^{g}, r_k^{g}\right) + \right. \\ &\left.\textstyle\sum_{l\prime}^{n}\left[E\left(r_i^{g}, r_k^{g}, r_l^{g}\right) + E\left(r_j^{g}, r_k^{g}, r_l^{g}\right) + E\left(r_i^{g}, r_j^{g}, r_k^{g}, r_l^{g}\right) \ldots\right]\right\} \end{aligned}$$

Equation 8

We begin again from Equation 7. After explicitly enumerating all energy components of E_{global}, E_{GMEC} and substituting, all terms not involving r_i or r_j cancel out. We then find expressions for the remaining terms that do not require knowledge of the GMEC configuration.

$$\max_s \Bigg[E(r_i^\alpha, r_k^s) + E\left(r_j^\beta, r_k^s\right) + E\left(r_i^\alpha, r_j^\beta, r_k^s\right) + \sum_{l'}^{n} \left(E\left(r_i^\alpha, r_k^s, r_l^g\right) + E\left(r_j^\beta, r_k^s, r_l^g\right) + E\left(r_i^\alpha, r_j^\beta, r_k^s, r_l^g\right) + \cdots \right) \Bigg] \geq E\left(r_i^\alpha, r_k^g\right) + E\left(r_j^\beta, r_k^g\right) + E\left(r_i^\alpha, r_j^\beta, r_k^g\right) + \sum_{l'}^{n} \left(E\left(r_i^\alpha, r_k^g, r_l^g\right) + E\left(r_j^\beta, r_k^g, r_l^g\right) + E\left(r_i^\alpha, r_j^\beta, r_k^g, r_l^g\right) + \cdots \right)$$

Equation 9

$$\min_s \Bigg[E\left(r_i^g, r_k^s\right) + E\left(r_j^g, r_k^s\right) + E\left(r_i^g, r_j^g, r_k^s\right) + \sum_{l'}^{n} \left(E\left(r_i^g, r_k^s, r_l^g\right) + E\left(r_j^g, r_k^s, r_l^g\right) + E\left(r_i^g, r_j^g, r_k^s, r_l^g\right) + \cdots \right) \Bigg] \leq E\left(r_i^g, r_k^g\right) + E\left(r_j^g, r_k^g\right) + E\left(r_i^g, r_j^g, r_k^g\right) + \sum_{l'}^{n} \left(E\left(r_i^g, r_k^g, r_l^g\right) + E\left(r_j^g, r_k^g, r_l^g\right) + E\left(r_i^g, r_j^g, r_k^g, r_l^g\right) + \cdots \right)$$

Equation 10

$$\max_t \left[E(r_i^\alpha, r_k^s, r_l^t) + E\left(r_j^\beta, r_k^s, r_l^t\right) + E\left(r_i^\alpha, r_j^\beta, r_k^s, r_l^t\right) + \cdots \right] \geq E\left(r_i^\alpha, r_k^s, r_l^g\right) + E\left(r_j^\beta, r_k^s, r_l^g\right) + E\left(r_i^\alpha, r_j^\beta, r_k^s, r_l^g\right) + \cdots$$

$$\min_t \left[E\left(r_i^g, r_k^s, r_l^t\right) + E\left(r_j^g, r_k^s, r_l^t\right) + E\left(r_i^g, r_j^g, r_k^s, r_l^t\right) + \cdots \right] \leq E\left(r_i^g, r_k^s, r_l^g\right) + E\left(r_j^g, r_k^s, r_l^g\right) + E\left(r_i^g, r_j^g, r_k^s, r_l^g\right) + \cdots$$

Equation 11

Substituting Eq. 11 into Eqs. 9 and 10, we get:

$$\max_s \Bigg[E(r_i^\alpha, r_k^s) + E\left(r_j^\beta, r_k^s\right) + E\left(r_i^\alpha, r_j^\beta, r_k^s\right) + \sum_{l'}^{n} \max_t \left(E(r_i^\alpha, r_k^s, r_l^t) + E\left(r_j^\beta, r_k^s, r_l^t\right) + E\left(r_i^\alpha, r_j^\beta, r_k^s, r_l^t\right) + \cdots \right) \Bigg] \geq E\left(r_i^\alpha, r_k^g\right) + E\left(r_j^\beta, r_k^g\right) + E\left(r_i^\alpha, r_j^\beta, r_k^g\right) + \sum_{l'}^{n} E\left(r_i^\alpha, r_k^g, r_l^g\right) + E\left(r_j^\beta, r_k^g, r_l^g\right) + E\left(r_i^\alpha, r_j^\beta, r_k^g, r_l^g\right) + \cdots$$

Equation 12

$$\min_s \Bigg[E\left(r_i^g, r_k^s\right) + E\left(r_j^g, r_k^s\right) + E\left(r_i^g, r_j^g, r_k^s\right) + \sum_{l'}^{n} \min_t \left(E\left(r_i^g, r_k^s, r_l^t\right) + E\left(r_j^g, r_k^s, r_l^t\right) + E\left(r_i^g, r_j^g, r_k^s, r_l^t\right) + \cdots \right) \Bigg]$$
$$\leq E\left(r_i^g, r_k^g\right) + E\left(r_j^g, r_k^g\right) + E\left(r_i^g, r_j^g, r_k^g\right) + \sum_{l'}^{n} E\left(r_i^g, r_k^g, r_l^g\right) + E\left(r_j^g, r_k^g, r_l^g\right) + E\left(r_i^g, r_j^g, r_k^g, r_l^g\right) + \cdots$$

Equation 13

Expressing Equation 7 using the left-hand side of Eqs. 12 and 13 (for E_{global} and E_{GMEC} respectively) yields the final pairwise elimination criterion.

III. Many-Body Generalized Goldstein Singles Elimination Criterion

We begin from the substituted Eqs 1 through 3.

$$E(r_i^\alpha) + \sum_{j'}^{n} \left\{ E\left(r_i^\alpha, r_j^g\right) + \sum_{k'}^{n} \left[E\left(r_i^\alpha, r_j^g, r_k^g\right) + \sum_{l'}^{n} \left(E\left(r_i^\alpha, r_j^g, r_k^g, r_l^g\right) + \cdots \right) \right] \right\} \geq$$
$$E\left(r_i^g\right) + \sum_{j'}^{n} \left\{ E\left(r_i^g, r_j^g\right) + \sum_{k'}^{n} \left[E\left(r_i^g, r_j^g, r_k^g\right) + \sum_{l'}^{n} \left(E\left(r_i^g, r_j^g, r_k^g, r_l^g\right) + \cdots \right) \right] \right\}$$

Equation 14

In contrast to the original singles elimination derivation, we first subtract the right-hand side before applying the min operator.

$$E(r_i^\alpha) - E\left(r_i^g\right) + \sum_{j'}^{n} \left\{ E\left(r_i^\alpha, r_j^g\right) - E\left(r_i^g, r_j^g\right) + \sum_{k'}^{n} \left[E\left(r_i^\alpha, r_j^g, r_k^g\right) - E\left(r_i^g, r_j^g, r_k^g\right) + \right.\right.$$
$$\left.\left. \sum_{l'}^{n} \left(E\left(r_i^\alpha, r_j^g, r_k^g, r_l^g\right) - E\left(r_i^g, r_j^g, r_k^g, r_l^g\right) \ldots \right) \right] \right\} \geq 0$$

Equation 15

$$\min_s \Bigg\{ E\left(r_i^{\alpha}, r_j^{s}\right) - E\left(r_i^{g}, r_j^{s}\right) + \sum_{k'}^{n} \Bigg[E\left(r_i^{\alpha}, r_j^{s}, r_k^{g}\right) - E\left(r_i^{g}, r_j^{s}, r_k^{g}\right) + \sum_{l'}^{n} \left(E\left(r_i^{\alpha}, r_j^{s}, r_k^{g}, r_l^{g}\right) - E\left(r_i^{g}, r_j^{s}, r_k^{g}, r_l^{g}\right) \dots \right) \Bigg] \Bigg\} \leq E\left(r_i^{\alpha}, r_j^{g}\right) - E\left(r_i^{g}, r_j^{g}\right) + \sum_{k'}^{n} \Bigg[E\left(r_i^{\alpha}, r_j^{g}, r_k^{g}\right) - E\left(r_i^{g}, r_j^{g}, r_k^{g}\right) + \sum_{l'}^{n} \left(E\left(r_i^{\alpha}, r_j^{g}, r_k^{g}, r_l^{g}\right) - E\left(r_i^{g}, r_j^{g}, r_k^{g}, r_l^{g}\right) \dots \right) \Bigg]$$

Equation 16

$$\min_t \left[E\left(r_i^{\alpha}, r_j^{s}, r_k^{t}\right) - E\left(r_i^{g}, r_j^{s}, r_k^{t}\right) + \sum_{l'}^{n} \left(E\left(r_i^{\alpha}, r_j^{s}, r_k^{t}, r_l^{g}\right) - E\left(r_i^{g}, r_j^{s}, r_k^{t}, r_l^{g}\right) \dots \right) \right] \leq E\left(r_i^{\alpha}, r_j^{s}, r_k^{g}\right) - E\left(r_i^{g}, r_j^{s}, r_k^{g}\right) + \sum_{l'}^{n} \left(E\left(r_i^{\alpha}, r_j^{s}, r_k^{g}, r_l^{g}\right) - E\left(r_i^{g}, r_j^{s}, r_k^{g}, r_l^{g}\right) \dots \right)$$

Equation 17

$$\min_u \left(E\left(r_i^{\alpha}, r_j^{s}, r_k^{t}, r_l^{u}\right) - E\left(r_i^{g}, r_j^{s}, r_k^{t}, r_l^{u}\right) \dots \right) \leq \left(E\left(r_i^{\alpha}, r_j^{s}, r_k^{t}, r_l^{g}\right) - E\left(r_i^{g}, r_j^{s}, r_k^{t}, r_l^{g}\right) \dots \right)$$

Equation 18

As before, we then identify max and min inequalities that relieve us of reliance on knowing g. Substituting Eqs. 16-18 into Eq. 15, we arrive at the general Goldstein criterion.

IV. Many-Body Generalized Goldstein Pairwise Elimination Criterion

This derivation follows from the many-body Goldstein singles elimination in the same fashion that the original pairwise elimination followed from the original singles elimination.

$$E(r_i^{\alpha}) + E\left(r_j^{\beta}\right) + E\left(r_i^{\alpha}, r_j^{\beta}\right) + \sum_{k'}^{n} \left\{ E\left(r_i^{\alpha}, r_k^{g}\right) + E\left(r_j^{\beta}, r_k^{g}\right) + E\left(r_i^{\alpha}, r_j^{\beta}, r_k^{g}\right) + \sum_{l'}^{n} \left[E\left(r_i^{\alpha}, r_k^{g}, r_l^{g}\right) + E\left(r_j^{\beta}, r_k^{g}, r_l^{g}\right) + E\left(r_i^{\alpha}, r_j^{\beta}, r_k^{g}, r_l^{g}\right) + \cdots \right] \right\} \geq$$

$$E\left(r_i^g\right) + E\left(r_j^g\right) + E\left(r_i^g, r_j^g\right) + \sum_{k\prime}^{n}\{E\left(r_i^g, r_k^g\right) + E\left(r_j^g, r_k^g\right) + E\left(r_i^g, r_j^g, r_k^g\right) + \sum_{l\prime}^{n}\left[E\left(r_i^g, r_k^g, r_l^g\right) + E\left(r_j^g, r_k^g, r_l^g\right) + E\left(r_i^g, r_j^g, r_k^g, r_l^g\right) + \cdots\right]\}$$

Equation 19

$$E(r_i^\alpha) - E\left(r_i^g\right) + E\left(r_j^\beta\right) - E\left(r_j^g\right) + E\left(r_i^\alpha, r_j^\beta\right) - E\left(r_i^g, r_j^g\right) + \sum_{k\prime}^{n}\left\{E\left(r_i^\alpha, r_k^g\right) - E\left(r_i^g, r_k^g\right) + E\left(r_j^\beta, r_k^g\right) - E\left(r_j^g, r_k^g\right) + E\left(r_i^\alpha, r_j^\beta, r_k^g\right) - E\left(r_i^g, r_j^g, r_k^g\right) + \sum_{l\prime}^{n}\left[E\left(r_i^\alpha, r_k^g, r_l^g\right) - E\left(r_i^g, r_k^g, r_l^g\right) + E\left(r_j^\beta, r_k^g, r_l^g\right) - E\left(r_j^g, r_k^g, r_l^g\right) + E\left(r_i^\alpha, r_j^\beta, r_k^g, r_l^g\right) - E\left(r_i^g, r_j^g, r_k^g, r_l^g\right) + \cdots\right]\right\} \geq 0$$

Equation 20

$$\min_s\left[E(r_i^\alpha, r_k^s) - E\left(r_i^g, r_k^s\right) + E\left(r_j^\beta, r_k^s\right) - E\left(r_j^g, r_k^s\right) + E\left(r_i^\alpha, r_j^\beta, r_k^s\right) + E\left(r_i^g, r_j^g, r_k^s\right) + \cdots\right] \leq E\left(r_i^\alpha, r_k^g\right) - E\left(r_i^g, r_k^g\right) + E\left(r_j^\beta, r_k^g\right) - E\left(r_j^g, r_k^g\right) + E\left(r_i^\alpha, r_j^\beta, r_k^g\right) + E\left(r_i^g, r_j^g, r_k^g\right) + \cdots$$

Equation 21

Downstream min and max operators are applied just as before and are substituted into Eq. 20 to yield the many-body generalized pairwise Goldstein criterion.