Sub-Angstrom Accuracy in Protein Loop Reconstruction by Robotics-Inspired Conformational Sampling

Daniel J. Mandell, Evangelos A. Coutsias and Tanja Kortemme

Supplementary Materials

I. Supplementary Methods

Datasets. We use two independent benchmark datasets for loops in monomeric proteins: dataset 1, a set of 40 12-residue loops originally compiled by Fiser *et al.*[1], and later studied by Rohl *et al.*[2] and Wang *et al.*[3], to facilitate comparison to previous work using the Rosetta loop modeling methodology, and dataset 2, a set of 20 12-residue loops compiled by Zhu *et al.*[4] to allow direct comparison to studies by Jacobson *et al.* [5], Zhu *et al.*[4], and Sellers *et al.*[6] The latter dataset was selected from high quality structures (resolution ≤ 2.0Å, $R < 0.25$) for loops with diverse sequences (<40% sequence identity), low temperature factor (<35), lack of contacts to heteroatom groups (>4Å for neutral ligands, >6.5Å for metal ions), lack of secondary structure within the loop, lack of more than 4 loop residues adjacent to either loop endpoint, and pH 6.5 – 7.5. The monomer loop datasets are shown in Tables S1 and S2, respectively. Dataset 1 contained 15 loops with neutral ligands or charged ions within contact distance of the loop, using the criteria specified by dataset 2, so while these loops are included in Table S1 they are separated from the "filtered" dataset used for most subsequent analyses. Dataset 2 was simulated in two ways, first by the *de novo* method used on the Rosetta dataset, where KC is used to place the loop into a random starting conformation, and second, by the *perturbed* method, where the perturbed loops used in the simulations by Sellers *et al.*[6] were obtained from that group's website[7] and serve as starting conformations for the Rosetta simulations. The perturbed approach was used to enable direct comparison between the Rosetta and molecular mechanics methods, since the degree of initial backbone perturbation will influence the degree to which the side-chain environment is perturbed. The "de novo" and "perturbed" columns of Table S2 refer to this distinction.

A third independent dataset (dataset 3) was compiled to assess loop reconstruction of the same protein crystallized in complex with different partners (Table S3). This dataset contains 4 proteins (Rac, Ras, CDC42, Ubiquitin) crystallized with 18 different partners where the interface contains a loop that changes conformation across partners. For each of the four proteins, the reconstruction endpoints were defined by consecutive residues that contained any heavy atoms that were within 7Å of the binding partner in any crystal structure, and that lacked secondary structure in one or more of the crystal structures of that protein (7 residues minimum). Thus, the loop definitions were the same across complexes of the same protein, facilitating assessment of reconstruction accuracy. Nucleotides and metal co-factors were modeled explicitly, and GDP-aluminum fluoride was modeled as GTP.

Structure preparation. Structures were prepared by first discarding all native side-chain information (including side-chain bond lengths, bond angles, and chi angles) and replacing them with rotameric conformations from the Dunbrack backbone-dependent rotamer library[8] that are then simultaneously optimized by Metropolis Monte Carlo simulated annealing ("*repacking*") using Rosetta, as described in reference[9]. Each side-chain is then independently optimized by replacing it with the lowest energy conformation from the Dunbrack library and iterating through all positions until convergence is reached ("*rotamer trials*"). These procedures are followed by quasi-Newton all-atom energy minimization using the Davidon-Fletcher-Powell method[10] (*DFPmin*) on the loop backbone and side-chains within 10Å of the loop. The repacked, energy minimized structures serve as input to the loop modeling protocol, which is depicted in Figure S1.

Loop modeling protocol. Loop endpoints for protein monomers are defined as in refs[3,6] and shown in Tables S1 and S2, and loop endpoints for the complexes set are defined as above (see Datasets) and shown in Table S3. The simulation proceeds through two stages of Monte Carlo simulated annealing, as shown in Figure S1. In the first, low-resolution stage, all side-chains are represented as centroids for coarse-grained conformational sampling. An initial kinematic closure (KC) is performed on the entire loop to place it into a non-native starting conformation with randomly chosen phi and psi torsion angles at non-pivot residues and phi/psi torsion angles at pivot residues determined by the kinematic closure algorithm (see Kinematic Closure section, below). During this step, native phi and psi torsions in the loop region are discarded, and bond lengths, bond angles, and omega torsions are set to ideal values. The 720 simulated annealing MC steps consist of applying KC to a random subsegment of the loop region of length 3 to *N* (for an *N* residue loop). KC moves are followed by line minimization of backbone torsions. The new conformation is scored and accepted or rejected by the Metropolis criterion. In the centroid stage the temperature decays exponentially from 2.0 *k*T to 1.0 *k*T, where *k* is Boltzmann's constant. The lowest energy conformation proceeds to the high-resolution all-atom stage. The repacked, minimized side-chains from the input conformation (see Structure preparation) are restored and those in the loop and on the surrounding scaffold with any heavy atoms within 10Å of the new loop conformation are then repacked and subject to rotamer trials. If the loop is part of an interface (e.g., on dataset 3), side-chains from the binding partner within 10Å of the loop are optimized as well. Relaxing the neighboring side-chains around a non-native loop conformation has the effect of starting the full-atom stage in a perturbed side-chain environment. This step makes loop reconstruction considerably more difficult, since neighboring side-chain conformations must be sampled and evaluated in addition to the loop side-chains and backbone conformations. The utility increases, however, because in many applications (e.g., homology modeling, interface redesign) it cannot be assumed that the neighboring side-chain conformations are known *a priori*. We compare our results to the method presented by Sellers *et al*.[6] that also reconstructs loops in a perturbed side-chain environment.

The 720 MC steps of the high-resolution stage consist of kinematic closure on random subsegments of the loop region, with one exponential simulated annealing cycle from 1.5 *k*T to 0.5 *k*T. In this high-resolution stage, KC is followed by side-chain repacking (every 20 steps) and rotamer trials within 10Å of the new loop conformation, and DFPmin on the loop backbone and side-chains within 10Å of the new loop conformation. The lowest energy conformation explored during the high-resolution stage is recorded. The protocol is then iterated, and may be run over multiple processes in parallel. Reported loop reconstructions represent the lowest energy structure out of 1000 separate simulations (see Figure S1), costing an average of ~320 CPU-hours per protein on a single 2.2 GHz Opteron processor. Each simulation trajectory is independent from the others, so they may be parallelized to dramatically speed up the protocol (up to one CPU-core per trajectory requiring less than 20 minutes per protein on average).

Kinematic closure. The atomic coordinates of the backbone atoms (N, Cα, C) of a random loop sub-segment of length 3 to *N* (for a loop of *N* residues) are supplied to the kinematic solver. The Cα atoms of the first, middle, and last residues are designated as pivots, and the remaining *N*-3 Cα atoms are designated as non-pivots. Torsions for each non-pivot Cα are assigned according to the Ramachandran probabilities for the residue type, and N-Cα-C bond angles are set to random values within one-half the standard deviation ($\sigma = 2.48^\circ$) above and below the mean (110.86°) observed in ultra-high-resolution crystal structures (<1.0Å resolution) in the Protein Data Bank (pdb). This step effectively *opens* the loop segment at the pivots, breaking the continuity of the peptide chain. To close the loop, the kinematic solver finds values of the six pivot torsions for which the perturbed segments may be rejoined to form a new closed loop. As discussed in the next section (Polynomial resultants), there may be up to sixteen sets of such solutions, or none. Solutions are randomly applied to the loop segment until two filters are passed. The first filter computes the Rosetta Ramachandran score, which is a statistical potential derived from a smoothed, highly flattened version of the residue- and secondary structure-specific frequency with which a given (phi/psi) pair occurs in a set of high-resolution crystal structures[11], and accepts or rejects the conformation by the Metropolis criterion. The second filter is a backbone steric screen that ensures the distance between loop backbone atoms (N, Cα, C, O, and Cβ if not glycine) and all other backbone atoms is greater than the sum of the Lennard-Jones radii of the atoms times an overlap factor (set to 0.7). The accepted solution is returned to the protocol for minimization and scoring. If no solution passes the filters, new values for the non-pivot torsions and N-Cα-C bond angles are drawn and closure is attempted again. Closure calculations execute 2,000 times per second on a 1.8 GHz Opteron processor.

Other kinematics-inspired approaches have been applied to protein modeling[12-20]. Applications have included calculating conformations of cyclic peptides[12], exploring loop motions in one protein test case[15,16], and correlating loop models with spectroscopic observables from nuclear magnetic resonance experiments like order parameters and residual dipolar couplings in two proteins[19]. None of these methods have been rigorously tested on large datasets on the problem of loop reconstruction, and each of these methods has either lacked analytical solution[12,14,17,19], has been applicable only to tripeptides or required consecutive pivot residues[12,13,15], or has not been coupled to a full-atom energy function[12-14,16-18,20].

Polynomial resultants. The details of the geometric steps taken by the algorithm are given in reference[21]. The construction proceeds by identifying 3 atoms before the N-terminus of the missing loop, and 3 atoms after the C-terminus. These two triads are assumed to have known positions in space. Together, they constitute the anchoring hinges for the two ends of the loop. They are denoted h_1 and h_2(Fig. S2a). The loop atoms are augmented by the hinge atoms. Together they form the extended loop, which on the outset is considered to be in an extended conformation with all bond lengths and bond angles set to canonical values and all torsions set to 180.0 degrees. Three nonconsecutive atoms (not on the hinges), indexed p_1, p_2, p_3 with $p_1 + 2 \le p_2 \le p_3 - 2$ are chosen as the pivots for loop closure, and the loop is partitioned into four fragments: (1) $F_{3,b}$ including atoms from $h_{1,1}$ (the first atom of h_1) to p_1; (2) F_1 including atoms from p_1 to p_2; (3) F_2 including atoms from p_2 to p_3; and (4) $F_{3,a}$ including atoms from p_3 to $h_{2,3}$ (the third atom of h_3) (Fig. S2b). Next, the four fragments thus defined are constructed using prescribed values for all their internal degrees of freedom (bond lengths, bond angles, and torsions). Arbitrary values can be chosen. At this stage, the bond angles at the three pivot atoms and the torsions about the bonds adjacent to pivot atoms (i.e., the “pivot bond angles” and the “pivot torsions”) are not defined. Since the two hinges are anchored to the (known) rest of the molecule and thus have known absolute positions in space, the fragments $F_{3,a}, F_{3,b}$ are thus constructed with known positions in space for all their atoms relative to the hinges (and thus to the rest of the molecule). Their end atoms ($p_3, p_3 + 1, p_3 + 2, p_1 - 2, p_1 - 1, p_1$) are now fixed in space (Fig. S2c).

The other two fragments, F_1 and F_2 are completely determined in their own body frames (Fig. S2d), but their placement relative to the molecule is still to be determined. Each fragment is characterized by certain geometrical quantities that will enter as parameters in the loop closure equations. Referring to Figure S2e these are: (1) ξ_i, the angle formed by atoms $(p_i + 1, p_i, p_{i+1})$; (2) η_i, the angle $(p_i, p_{i+1}, p_{i+1} - 1)$; (3) d_i, the virtual bond length (p_i, p_{i+1}); and (4) δ_i, the dihedral angle $(p_i + 1, p_i, p_{i+1}, p_{i+1} - 1)$. The Rosetta implementation uses virtual segments composed from only the first and last triads of atoms in each segment, avoiding unnecessary reconstructions. These virtual segments must be assembled into a closed triangle (Fig. S2f), provided the three lengths d_1, d_2, d_3 satisfy the triangle inequalities. If the triangle can be constructed, the three exterior angles $\alpha_1, \alpha_2, \alpha_3$, are among the parameters defining the loop closure equations below. Note that although in Figure S2f the α_i are shown as interior angles, the formulas below assume they are indeed the exterior angles at the corresponding vertices of the triangle. An additional requirement for the proper assembly of the loop is that the pivot bond angles $\theta_i, i = 1,2,3$ must assume their prescribed values. That may be possible to accomplish by rotating segment F_i about the virtual bond joining pivots (p_i, p_{i+1}) by angle τ_i (Fig. S2g). The additional atoms, $p_i + 2, p_{i+1} - 2$ that are included in each virtual segment allow the calculation of the six pivot torsions, once the virtual segments have been rotated to their correct positions, so that the angle $(p_i - 1, p_i, p_i + 1) = \theta_i$. Note that the loop closure equations are formulated in the body frame of the three pivot atoms. To convert to the space frame of the rest of the molecule, the fragment F_3 is assumed fixed, and the rest of the loop (fragments F_1, F_2) is rotated about (p_3, p_1) by angle $-\tau_3$. Determining the pivot torsions completes the specification of all internal degrees of freedom for the missing loop, which can now be constructed, closing the gap (Fig. S2h).

The bond angle constraints lead to the loop closure equations[15]. These are a system of three polynomials that are quadratic in each of the variables:

$$
\begin{gathered}
L_2(u_3)u_1^2 + L_1(u_3)u_1 + L_0(u_3) = 0, \\
\left(M_{22}u_2^2 + M_{21}u_2 + M_{20}\right)u_1^2 + \left(M_{12}u_2^2 + M_{11}u_2 + M_{10}\right)u_1 + \left(M_{02}u_2 + M_{01}u_2 + M_{00}\right) = 0, \\
N_2(u_3)u_2^2 + N_1(u_3)u_2 + N_0(u_3) = 0.
\end{gathered}
$$

The variables are $u_i = \tan\left(\tau_i/2\right), i = 1,2,3$. The $L_i, N_i, i = 0,1,2$ are quadratic polynomials in u_3, while the $M_{ij}, i,j = 0,1,2$ are constants. Each polynomial depends on only two of the u_i. Throughout, we follow the notation of Coutsias *et al.*[15], and refer the reader to that reference for the values of the polynomial coefficients. The code encodes the atomic coordinates of each virtual segment as a set of triaxial parameters as in Coutsias *et al.*[15] These parameters are used to populate a matrix $R(u_3)$ called the Dixon Resultant (DR) that results from eliminating the variables u_1, u_2 (any two of the variables could have been eliminated in favor of the remaining one). The necessary and sufficient condition that the above system of three polynomials in the three variables u_1, u_2, u_3 has a common solution is expressed by the equation[21]

$$
R(u_3)V(u_1,u_2) := \begin{bmatrix}
0 & A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\
A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 & 0 \\
0 & B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\
B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\
0 & 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \\
0 & 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\
0 & D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\
D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
1 \\ u_1 \\ u_1^2 \\ u_1^3 \\ u_2 \\ u_1u_2 \\ u_1^2u_2 \\ u_1^3u_2
\end{bmatrix} = 0
$$

where

$$\begin{aligned} A_i &:= M_{i1}N_0 - M_{i0}N_1, \\ B_i &:= M_{i2}N_0 - M_{i0}N_2, \\ C_i &:= M_{i2}N_1 - M_{i1}N_2, \\ D_i &:= L_i. \end{aligned}$$

Since its coefficients are quadratic polynomials in u_3 the DR can be written as a matrix polynomial

$$R(u_3) = R_2 u_3^2 + R_1 u_3 + R_0.$$

The above matrix equation can be recast as a generalized eigenvalue problem

$$\left(\begin{bmatrix} I & 0 \\ 0 & R_2 \end{bmatrix} u_3 - \begin{bmatrix} 0 & I \\ -R_0 & -R_1 \end{bmatrix} \right) \begin{bmatrix} V \\ u_3 V \end{bmatrix} = 0.$$

This eigenproblem can be solved directly using the QZ factorization algorithm. An attractive feature of this approach is that the remaining variables u_1, u_2 are also found directly from the solution of this generalized eigenproblem, since they appear explicitly as particular components (resp. V_2, V_5) of the corresponding generalized eigenvector while u_3 is the generalized eigenvalue[21]. Sixteen solutions are always found, but some or all may be complex. To have geometrical meaning the solutions must be real, so complex solutions are discarded. The eigenproblem has the advantage of robustness and conceptual simplicity, but it can be computationally expensive, as each step of the iterative QZ algorithm scales with the cube of matrix size. As an alternative, we can get u_3 from the condition that the determinant of the DR must vanish. Having found values for u_3 for which $R(u_3)$ becomes singular, we can determine the desired components of its null-vector V by Cramer's rule. Since the coefficients of R are quadratic polynomials in u_3, its determinant is a polynomial of degree 16 in u_3, and by examining the existence of real solutions only, a substantial speedup can be accomplished. The polynomial conversion has been carried out optimally by careful regrouping of the terms and employing Lagrange expansions in complementary minors. Since this expansion has not been previously reported, we outline it here**.** By a rearrangement of rows, we have the equivalent form

$$\det(R) = \begin{vmatrix} D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 & 0 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 & 0 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ 0 & D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ 0 & A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ 0 & B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{vmatrix}.$$

In this form we get a compact expansion in terms of 4×4 minors

$$\det(R) = -P_{1235}P_{3567} + P_{1256}P_{2367} - P_{1257}P_{2357} - P_{1356}P_{1367} + P_{1357}^2 - P_{1567}P_{1237}$$

where P_{ijkl} is the determinant of the minor formed by rows 1,…,4 and columns *i*, *j*, *k* and *l*.

We have

$$P_{1235}P_{3567} := \begin{vmatrix} D_0 & D_1 & D_2 & 0 \\ A_0 & A_1 & A_2 & B_0 \\ B_0 & B_1 & B_2 & C_0 \\ 0 & 0 & 0 & D_0 \end{vmatrix} \begin{vmatrix} D_2 & 0 & 0 & 0 \\ A_2 & B_0 & B_1 & B_2 \\ B_2 & C_0 & C_1 & C_2 \\ 0 & D_0 & D_1 & D_2 \end{vmatrix} = D_0 D_2 \begin{vmatrix} D_0 & D_1 & D_2 \\ A_0 & A_1 & A_2 \\ B_0 & B_1 & B_2 \end{vmatrix} \begin{vmatrix} B_0 & B_1 & B_2 \\ C_0 & C_1 & C_2 \\ D_0 & D_1 & D_2 \end{vmatrix} =$$

$$D_0 D_2 \left(C_0 \begin{vmatrix} B_1 & B_2 \\ D_1 & D_2 \end{vmatrix} - C_1 \begin{vmatrix} B_0 & B_2 \\ D_0 & D_2 \end{vmatrix} + C_2 \begin{vmatrix} B_0 & B_1 \\ D_0 & D_1 \end{vmatrix} \right) \left(A_0 \begin{vmatrix} D_1 & D_2 \\ B_1 & B_2 \end{vmatrix} - A_1 \begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix} + A_2 \begin{vmatrix} D_0 & D_1 \\ B_0 & B_1 \end{vmatrix} \right)$$

$$P_{1256}P_{2367} := \begin{vmatrix} D_0 & D_1 & 0 & 0 \\ A_0 & A_1 & B_0 & B_1 \\ B_0 & B_1 & C_0 & C_1 \\ 0 & 0 & D_0 & D_1 \end{vmatrix} \begin{vmatrix} D_1 & D_2 & 0 & 0 \\ A_1 & A_2 & B_1 & B_2 \\ B_1 & B_2 & C_1 & C_2 \\ 0 & 0 & D_1 & D_2 \end{vmatrix} =$$

$$\left(\begin{vmatrix} D_0 & D_1 \\ A_0 & A_1 \end{vmatrix} \begin{vmatrix} C_0 & C_1 \\ D_0 & D_1 \end{vmatrix} - \begin{vmatrix} D_0 & D_1 \\ B_0 & B_1 \end{vmatrix} \begin{vmatrix} B_0 & B_1 \\ D_0 & D_1 \end{vmatrix} \right) \left(\begin{vmatrix} D_1 & D_2 \\ A_1 & A_2 \end{vmatrix} \begin{vmatrix} C_1 & C_2 \\ D_1 & D_2 \end{vmatrix} - \begin{vmatrix} D_1 & D_2 \\ B_1 & B_2 \end{vmatrix} \begin{vmatrix} B_1 & B_2 \\ D_1 & D_2 \end{vmatrix} \right)$$

$$P_{1257}P_{2357} := \begin{vmatrix} D_0 & D_1 & 0 & 0 \\ A_0 & A_1 & B_0 & B_2 \\ B_0 & B_1 & C_0 & C_2 \\ 0 & 0 & D_0 & D_2 \end{vmatrix} \begin{vmatrix} D_1 & D_2 & 0 & 0 \\ A_1 & A_2 & B_0 & B_2 \\ B_1 & B_2 & C_0 & C_2 \\ 0 & 0 & D_0 & D_2 \end{vmatrix} =$$

$$\left(\begin{vmatrix} D_0 & D_1 \\ A_0 & A_1 \end{vmatrix} \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix} - \begin{vmatrix} D_0 & D_1 \\ B_0 & B_1 \end{vmatrix} \begin{vmatrix} B_0 & B_2 \\ D_0 & D_2 \end{vmatrix} \right) \left(\begin{vmatrix} D_1 & D_2 \\ A_1 & A_2 \end{vmatrix} \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix} - \begin{vmatrix} D_1 & D_2 \\ B_1 & B_2 \end{vmatrix} \begin{vmatrix} B_0 & B_2 \\ D_0 & D_2 \end{vmatrix} \right)$$

$$P_{1356}P_{1367} := \begin{vmatrix} D_0 & D_2 & 0 & 0 \\ A_0 & A_2 & B_0 & B_1 \\ B_0 & B_2 & C_0 & C_1 \\ 0 & 0 & D_0 & D_1 \end{vmatrix} \begin{vmatrix} D_0 & D_2 & 0 & 0 \\ A_0 & A_2 & B_1 & B_2 \\ B_0 & B_2 & C_1 & C_2 \\ 0 & 0 & D_1 & D_2 \end{vmatrix} =$$

$$\left(\begin{vmatrix} D_0 & D_2 \\ A_0 & A_2 \end{vmatrix} \begin{vmatrix} C_0 & C_1 \\ D_0 & D_1 \end{vmatrix} - \begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix} \begin{vmatrix} B_0 & B_1 \\ D_0 & D_1 \end{vmatrix} \right) \left(\begin{vmatrix} D_0 & D_2 \\ A_0 & A_2 \end{vmatrix} \begin{vmatrix} C_1 & C_2 \\ D_1 & D_2 \end{vmatrix} - \begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix} \begin{vmatrix} B_1 & B_2 \\ D_1 & D_2 \end{vmatrix} \right)$$

$$P_{1357}^2 := \begin{vmatrix} D_0 & D_2 & 0 & 0 \\ A_0 & A_2 & B_0 & B_2 \\ B_0 & B_2 & C_0 & C_2 \\ 0 & 0 & D_0 & D_2 \end{vmatrix}^2 = \left(\begin{vmatrix} D_0 & D_2 \\ A_0 & A_2 \end{vmatrix} \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix} - \begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix} \begin{vmatrix} B_0 & B_2 \\ D_0 & D_2 \end{vmatrix} \right)^2$$

$$P_{1567}P_{1237} := \begin{vmatrix} D_2 & 0 & 0 & 0 \\ A_2 & B_0 & B_1 & B_2 \\ B_2 & C_0 & C_1 & C_2 \\ 0 & D_0 & D_1 & D_2 \end{vmatrix} \begin{vmatrix} D_0 & D_1 & D_2 & 0 \\ A_0 & A_1 & A_2 & B_0 \\ B_0 & B_1 & B_2 & C_0 \\ 0 & 0 & 0 & D_0 \end{vmatrix} = D_2 D_0 \begin{vmatrix} B_0 & B_1 & B_2 \\ C_0 & C_1 & C_2 \\ D_0 & D_1 & D_2 \end{vmatrix} \begin{vmatrix} D_0 & D_1 & D_2 \\ A_0 & A_1 & A_2 \\ B_0 & B_1 & B_2 \end{vmatrix}$$

$$= P_{3567}P_{1235}.$$

There are only 9 different 2×2 determinants involved in these calculations, each resulting in a quartic polynomial in u_3. The computation of the coefficients of the characteristic polynomial can be carried out in under 1800 flops. The polynomial is solved efficiently by the method of Sturm chains[22] and each solution results in a set of torsions for the pivot residues that close the loop. The determination of the variables u_1, u_2 now requires additional calculation. Briefly, the generalized eigenvectors are null vectors of the DR matrix R. Once u_3 has been found for which $\det(R)$ vanishes, we may determine specific components of the null vectors of $R(u_3)$ by using Cramer's rule. Most of the determinant minors involved in this computation are already known from the calculation of the characteristic polynomial.

$$\begin{bmatrix} D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 & 0 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 & 0 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ 0 & D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ 0 & A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ 0 & B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{bmatrix} \begin{bmatrix} 1 \\ u_1 \\ u_1^2 \\ u_1^3 \\ u_2 \\ u_1 u_2 \\ u_1^2 u_2 \\ u_1^3 u_2 \end{bmatrix} = 0 \Rightarrow$$

$$\begin{bmatrix} D_1 & D_2 & 0 & 0 & 0 & 0 & 0 \\ B_1 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_1^2 \\ u_1^3 \\ u_2 \\ u_1 u_2 \\ u_1^2 u_2 \\ u_1^3 u_2 \end{bmatrix} = \begin{bmatrix} D_0 \\ B_0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

where we have omitted one of the equations since it is dependent on the others and moved the first column to the RHS of the resulting system of 7 equations in the 7 unknowns. We only need to solve for the first and fourth components; in a generic situation one would apply LU factorization. However here Cramer’s rule can be used effectively, since most of the necessary determinantal computations have already been done for finding the characteristic polynomial.

We have

$$d_0 := \begin{vmatrix} D_1 & D_2 & 0 & 0 & 0 & 0 & 0 \\ B_1 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{vmatrix}$$

$$u_1 = \frac{d_1}{d_0} := \frac{1}{d_0}\begin{vmatrix} D_0 & D_2 & 0 & 0 & 0 & 0 & 0 \\ B_0 & B_2 & 0 & C_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & D_0 & D_1 & D_2 & 0 \\ 0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ 0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ 0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{vmatrix}$$

$$u_2 = \frac{d_2}{d_0} := \frac{1}{d_0}\begin{vmatrix} D_1 & D_2 & 0 & D_0 & 0 & 0 & 0 \\ B_1 & B_2 & 0 & B_0 & C_1 & C_2 & 0 \\ 0 & 0 & 0 & 0 & D_1 & D_2 & 0 \\ D_0 & D_1 & D_2 & 0 & 0 & 0 & 0 \\ A_0 & A_1 & A_2 & 0 & B_0 & B_1 & B_2 \\ B_0 & B_1 & B_2 & 0 & C_0 & C_1 & C_2 \\ 0 & 0 & 0 & 0 & D_0 & D_1 & D_2 \end{vmatrix}.$$

These are expanded as follows

$$d_0 = D_0\begin{vmatrix} D_1 & D_2 \\ B_1 & B_2 \end{vmatrix}P_{3567} - D_1\left(\begin{vmatrix} C_0 & C_1 \\ D_0 & D_1 \end{vmatrix}P_{2367} - \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix}P_{2357}\right) + D_2\left(\begin{vmatrix} C_0 & C_1 \\ D_0 & D_1 \end{vmatrix}P_{1367} - \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix}P_{1357}\right)$$

$$d_1 = -D_0\left(\begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix}P_{3567} + \begin{vmatrix} C_0 & C_1 \\ D_0 & D_1 \end{vmatrix}P_{2367} + \begin{vmatrix} C_0 & C_2 \\ D_0 & D_2 \end{vmatrix}P_{2357}\right)$$

$$d_2 = \left(D_1P_{2367} - D_2P_{2357}\right)\begin{vmatrix} D_0 & D_1 \\ B_0 & B_1 \end{vmatrix} - \left(D_1P_{1367} - D_2P_{1357}\right)\begin{vmatrix} D_0 & D_2 \\ B_0 & B_2 \end{vmatrix} - D_0\begin{vmatrix} C_1 & C_2 \\ D_1 & D_2 \end{vmatrix}P_{1237}.$$

As the P_{ijkl} and the various 2×2 determinants in these expressions have already been computed in the calculation of the characteristic polynomial, the computation of the $d_i, i = 1,2,3$ can be accomplished with an additional cost of under 400 flops.

Elimination of native bias. For loop reconstruction to have broad applicability it is important to carry out predictions with minimal knowledge of the native side-chain environment. The Rosetta KC protocol first discards all native side-chain chi angles, bond angles, and bond lengths and repacks the side-chains using conformations from a rotamer library[8]. This initial repacking (without the presence of any native side-chains at any position) is carried out against the native backbone (dataset 1) or on the perturbed backbone in dataset 2 obtained from Sellers *et al*[6]. Subsequently, an initial kinematic closure discards the native loop backbone torsions, bond angles, and bond lengths, and places the loop into a perturbed starting conformation with idealized bond lengths and bond angles (except for N-Cα-C bond angles, which have been sampled without knowledge of the native values). After the protocol completes the centroid stage, all side-chains within 10Å of the predicted loop conformation are discarded and repacked. This step entails that at the beginning of the full-atom stage, side-chains within 10Å of the loop have been optimized against a predicted non-native backbone in dataset 1, and all side-chains have been optimized against an initially perturbed backbone in dataset 2.

II. Supplementary Discussion

Both conformational sampling and accurate scoring are significant challenges in protein loop modeling. In the following sections we discuss examples of successes and failures of loop reconstruction arising in both areas, and evaluate the sensitivity of our KC method to modified sampling parameters and the required computational cost.

Conformational sampling. Accurate loop reconstruction requires substantial conformational sampling, owing to the extensive conformational space accessible to protein loops. If conformations near the crystallographic loop are not sampled, reconstruction accuracy will be poor. Even if near-native conformations are sampled, the scoring function must discriminate them from the ensemble of conformations sampled in the course of the simulation (insofar as the crystallographic structure represents the lowest free energy conformation of the protein). We sought to determine which failure cases (reconstruction accuracy ≥1.0Å) were attributable to insufficient sampling, and which suffered from incorrect scoring for both Rosetta methods on datasets 1[3] and 2[6] (accuracy is measured as global loop rmsd to the native backbone N, Cα, C, O atoms throughout the manuscript). To do so, we compared the scores of the lowest-scoring reconstructions to the scores of the crystallographic loops. If the crystallographic loop scored lower (better) than the lowest-scoring model, the failure resulted from insufficient conformational sampling, because the scoring function would have discriminated very-near crystallographic conformations had they been sampled. Conversely, if the lowest-scoring reconstruction was lower in score than the crystallographic loop, the failure was attributable to the scoring function, since near-crystallographic conformations scored worse then conformations ≥1.0Å away. The scores of the crystallographic loops were obtained by relaxing the repacked, minimized input structures through 100 independent trajectories of the full-atom stage of the KC protocol fixed at a temperature of 0.5 *k*T and recording the lowest scoring conformations within 0.5Å of each crystallographic loop. On dataset 1, we found that 16 out of 18 failure cases were due to poor conformational sampling using the standard protocol, compared to 5 out of 10 such cases using KC (Table S4). On dataset 2, all 15 failures were attributable to insufficient sampling using the standard protocol, compared to 6 out of 10 using the KC protocol (Table S5). Contributions from scoring and sampling cannot be completely decoupled since Metropolis Monte Carlo simulations accept or reject conformations with a probability dependent on the score. Since both protocols use the same number of steps over identical simulated annealing schedules with the same scoring function, however, these results suggest that the KC protocol, while imperfect, substantially improves conformational sampling compared to the standard protocol. Additionally, the results show that the enhanced torsion sampling enabled by KC can reveal scoring errors by finding low scoring structures distant from the crystallographic loop. Cases 4i1b and 1tgh in Table S4 and 1my7, 2pia, 1m3s and 1oyc in Table S5 are examples where scoring errors become apparent when sampling is enhanced with the KC protocol.

Dataset 3 provides an additional perspective on conformational sampling with KC because all the loops are crystallized in multiple conformations bound to different protein partners. Since protein modeling and design methods frequently transplant existing structures into new contexts as templates, it is useful to know how often the predicted loop more closely resembles the crystallographic loop than the same loop crystallized with different partner proteins. We pairwise-superimposed the cores of all loop proteins in dataset 3 and computed the global backbone N, Cα, C, O rmsds between the conformations of the loops bound to different partners. These rmsds, which show that even shorter 7-residue loops are capable of assuming significantly different conformations across complexes, are reported in Table S3. In 57 of 68 cases, the predicted loop was closer to the crystallographic loop than the same loop crystallized with another partner (shown in bold), suggesting the KC method provides a significant advantage over using another crystal structure as a template for modeling interface loops in new complexes.

Factors not modeled. Errors in loop reconstruction can result from structural features that are not explicitly considered by the modeling method. For the Rosetta KC protocol, such factors include crystallization conditions at pH values outside the neutral range, amino acid residues with shifted ionization constants, and residues with *cis* peptide bonds (since the protocol currently does not sample *cis* peptide bonds). Other errors could result from interactions between loop residues and neighboring protein copies in the crystal lattice, since the simulations are not performed within the crystallographic unit cell. To check for possible crystal packing effects, we reconstructed the crystal lattice using Pymol[23] and computed the changes in solvent accessible surface area (SASA) with and without the crystal context using Surface Racer[24] (1.2Å probe radius, using Richards 1977[25] van der Waals radii) for all loop residues in datasets 1 and 2. Cases where the delta SASA with and without the crystal context was >200Å^2 were considered to have significant crystal packing. Tables S6 and S7 show which failure cases had significant crystal packing by this measure, *cis* peptide bonds, or pH values well outside the neutral range.

Energy function simplifications and errors. The most significant scoring function failure in Table S5 involves a protein loop with specific interactions with a buried water molecule (Old Yellow Enzyme, pdb code 1oyc, 2.0Å resolution). The crystal structure suggests that this water molecule (H_2O 609), which has a B-factor (~24) that is lower than the average B-factor for waters in this structure (~29), forms a hydrogen bonding network with the backbone carbonyl of loop residue Ser 206, the side-chain hydroxyl group of Ser 136, and the backbone amide of Ser 138 (Fig S3). The loop reconstruction deviates substantially from the crystallographic loop in the region where the buried water molecule interacts with the loop backbone. Interactions with water molecules are a common source of error associated with the use of an implicit solvent model such as the one implemented in Rosetta (see next paragraph) that ignores effects resultant from the discrete size and asymmetry of a water molecule and the geometric constraints of water-mediated hydrogen bonding interactions.

Even when all atoms are explicitly represented, evaluating the energetic contribution of charged and polar interactions is a significant challenge for any scoring function. The Rosetta all-atom scoring function uses a combination of an orientation-dependent hydrogen bond term[26] with an implicit solvation model[27] to assess hydrogen bonding in protein structures. Due to the delicate energetic balance between forming inter-residue hydrogen bonds and losing hydrogen bonds to solvent (in addition to the absence of polarization effects that are ignored by most methods), it can be difficult to reconstruct the complex hydrogen bonding networks observed in some protein structures. Figure S4a shows an example of two loop residues that participate in a hydrogen bonding network with two other residues in human 5'-deoxy-5'-methylthioadenosine phosphorylase (pdb code 1cb0). A loop side-chain (Asp 43) accepts hydrogen bonds from the backbone and side-chain of Arg 63, which in turn donates hydrogen bonds to Glu 31 and another loop side-chain, Tyr 33. In the KC reconstruction of this loop and the surrounding side-chain environment (Fig S4b, 0.6Å accuracy), the hydrogen bonds between loop residue Asp 43, and neighbors Arg 63 and Glu 31 are recovered, suggesting that Rosetta sufficiently samples the side-chain conformations and that the hydrogen bonding terms can successfully evaluate the hydrogen bonding interactions in the presence of a perturbed backbone. Nevertheless, the reconstruction orients the side-chain of Tyr 33 out into bulk solvent, demonstrating that some electrostatic effects are too subtle for the Rosetta hydrogen bonding and solvation terms to model accurately.

Sensitivity to simulation parameters. As described in the Supplementary Methods, the Rosetta KC protocol samples N-Cα-C bond angles. To assess the importance of bond angle sampling to reconstruction accuracy, we re-ran the simulations on the combined 45 loops from datasets 1 and 2 using the same KC protocol except we fixed the loop N-Cα-C bond angles at their canonical values (110.86°). The fixed bond angle protocol achieved similar performance (1.0Å median rmsd) as the protocol that sampled bond angles (0.9Å rmsd), suggesting that the contribution of bond angle sampling is small (Table S8). This result of the relatively small effect of bond angle sampling is consistent with the overwhelmingly greater variability of backbone torsions compared to bond angles, and Laskowski *et al.*[28] have shown that the observed variability of bond angles decreases at very high crystallographic resolution. Additionally, Coutsias *et al.*[15] showed in an earlier analysis of their KC method that while N-Cα-C bond angle sampling increases the number of closable loops, it does not produce more native-like conformations. These results also suggest that bond angle sampling is not a significant bottleneck to the performance of the standard Rosetta method. In general, high-resolution structure prediction may not require sampling far from canonical bond angles, although it may be important in cases of experimentally observed bond angle strain, as in small cyclic peptides.

We also considered the effect of the simulated annealing schedule, which was performed originally on a fairly narrow range, on reconstruction performance. Rather than varying the temperature as described in the Supplementary Methods, we fixed the temperature for both centroid and full-atom stages at 1.0 *k*T. Again, the modified protocol performed nearly as well as the original protocol with simulated annealing (Table S8), achieving a median accuracy of 1.0Å with fixed temperature compared to 0.9Å with annealed temperature. Taken together, these results show that the protocol is quite robust to changes in some simulation parameters, and suggest that the most important feature is the enhanced torsion sampling provided by KC.

Computational cost. As noted in the main text, the KC protocol requires ~320 CPU-hours on a single 2.2 GHz Opteron processor to generate 1,000 models, while the standard protocol requires ~280 CPU-hours to generate the same number of models on the same processor. To assess the performance of both Rosetta methods as a function of CPU time, we performed shorter constant-time simulations on datasets 1 and 2. Each protocol was run for 120 CPU-hours on each protein using the same parameters as in the longer simulations. The rmsd of the best-scoring reconstruction to the crystallographic loop was computed in the same manner as the longer simulations. We found that using equal computational time, KC improved the median reconstruction accuracy to 0.9Å from 1.9Å using the standard protocol on dataset 1 and improved median accuracy to 1.2Å from 1.9Å using the standard protocol on dataset 2. When both protocols were started from the perturbed loops from ref[6] on dataset 2, KC improved median accuracy to 1.2Å from the standard protocol value of 2.2Å.

The molecular mechanics method required ~260 CPU-hours for each 12-residue loop simulation (B. Sellers, personal communication). As noted by Sellers *et al.*[6], the reported results employ side-chain optimization in a 7.5Å shell around the reconstructed loops. The Rosetta KC and standard protocols optimize side-chains within 10.0Å of the loops. As additionally noted in Figure 3 in reference[6], the molecular mechanics method requires roughly twice the computational time to optimize side-chains within 10.0Å of the loop compared to 7.5Å on 8-residue loops. We can thus expect that the molecular mechanics method will require at least as much computational time as the KC protocol when optimizing side-chains within 10.0Å of 12-residue loops.

Figure S1 – The Rosetta KC loop reconstruction protocol.

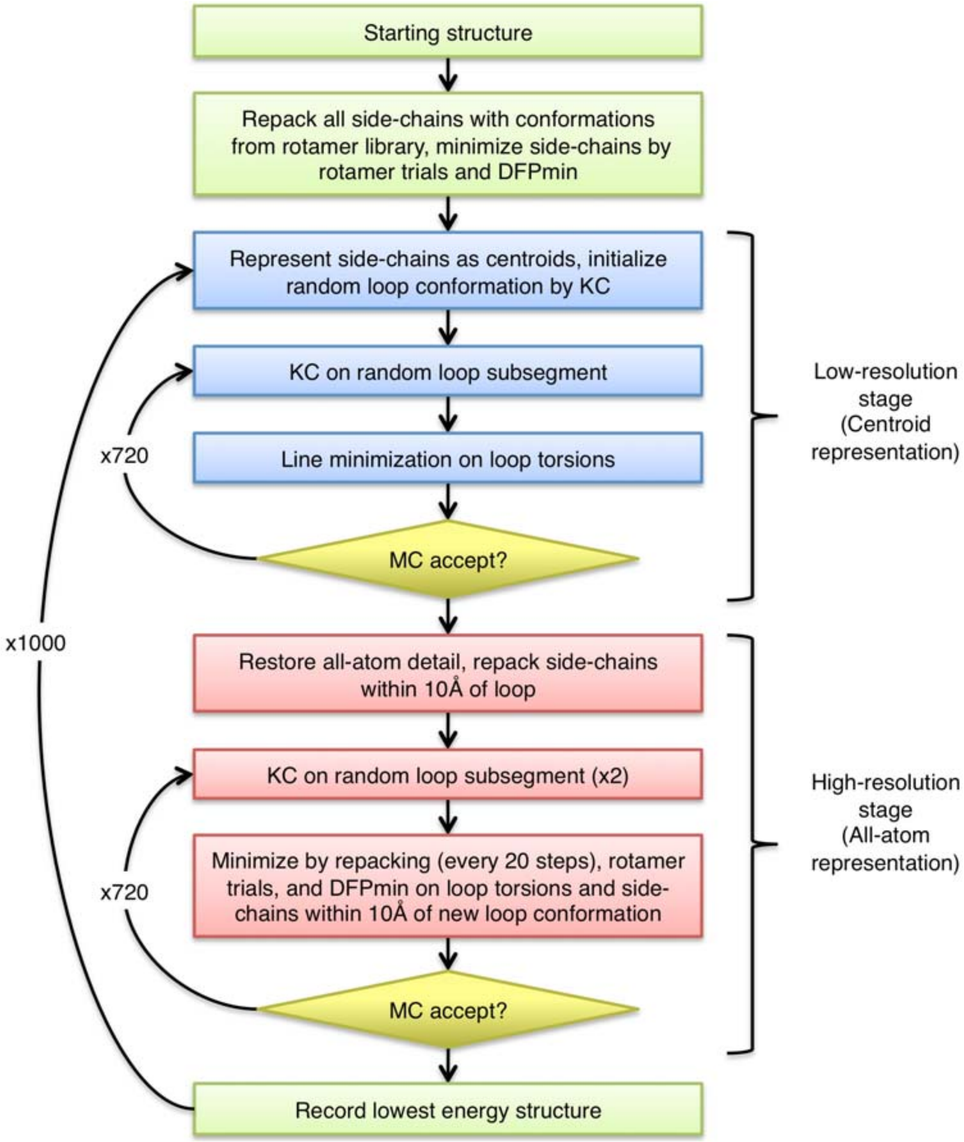
Starting structure
Repack all side-chains with conformations from rotamer library, minimize side-chains by rotamer trials and DFPmin
Represent side-chains as centroids, initialize random loop conformation by KC
KC on random loop subsegment
x720
Line minimization on loop torsions
MC accept?
Low-resolution stage (Centroid representation)
x1000
Restore all-atom detail, repack side-chains within 10Å of loop
KC on random loop subsegment (x2)
x720
Minimize by repacking (every 20 steps), rotamer trials, and DFPmin on loop torsions and side-chains within 10Å of new loop conformation
MC accept?
High-resolution stage (All-atom representation)
Record lowest energy structure

Figure S2 – Geometric steps taken by the kinematic closure solver. (**a**) Hinge N-Cα-C triads h_1, h_2 are defined flanking an arbitrary peptide chain. (**b**) The chain is partitioned into four fragments F_1, F_2, $F_{3,a}$, and $F_{3,b}$, defined by the three pivot Cα atoms p_1, p_2 and p_3. (**c**) The hinges are fixed in space, and the fragments $F_{3,a}$ and $F_{3,b}$ are constructed from the hinges using prescribed geometry. (**d**) The other two fragments, F_1 and F_2, are determined in their body frame with prescribed internal bond lengths, bond angles, and torsions, but are yet to be positioned with respect to $F_{3,a}$ and $F_{3,b}$. (**e**) Geometrical parameters for the kinematic closure equations are defined for the 4 fragments. (**f**) The fragments are assembled into a triangle such that three lengths d_1, d_2, and d_3 satisfy the triangle inequality. The resulting exterior angles corresponding to interior angles α_1, α_2, and α_3 form additional parameters for the loop closure equations. (**g**) The atoms of the 3 segments connecting two adjacent pivot atoms are rotated about the axis between the two pivots by an angle τ_i so that the prescribed pivot bond angles θ_i are satisfied. (**h**) The chain is converted from the body frame of the pivots to the space frame of the hinges by assuming the fragment F_3 is fixed and rotating the remaining fragments by angle $-\tau_3$.

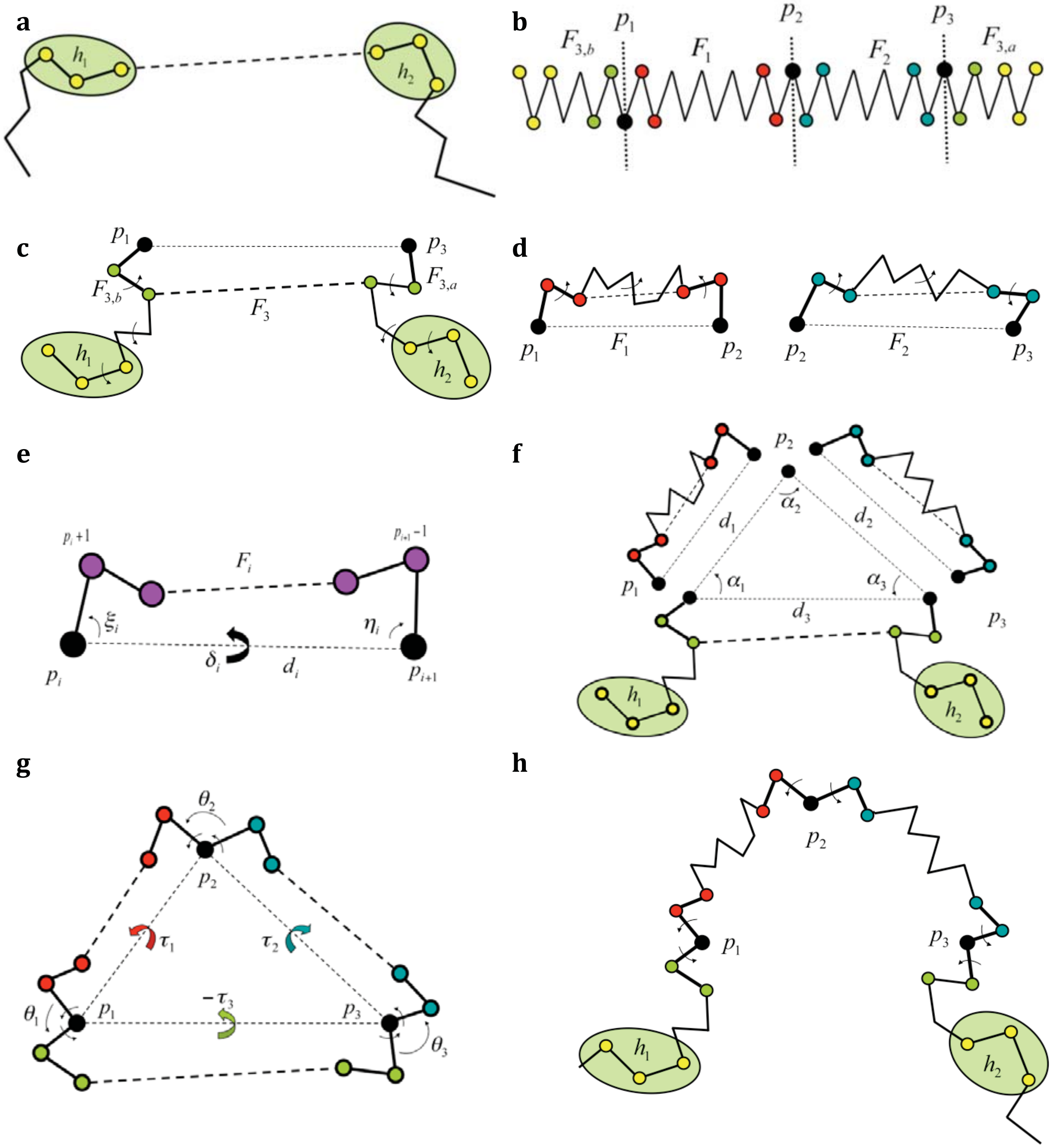

Figure S3 – Specific interactions with a buried water molecule in Old Yellow Enzyme (pdb code 1oyc). The loop residues subject to reconstruction are colored cyan in the crystal structure conformation, and the reconstruction is shown in blue. The backbone carbonyl of loop residue Ser 206 is shown in sticks, along with the side-chains of Ser 206, Ser 136, and the backbone amide of Ser 138 in the crystal structure. The backbone carbonyl and side-chain of Ser 206 on the reconstructed loop are also shown in sticks. Hydrogen atoms are included in the crystal structure. Explicit water molecules are not included in the loop reconstruction simulations, and this protein produces the most significant scoring error with the KC protocol on dataset 2[6].

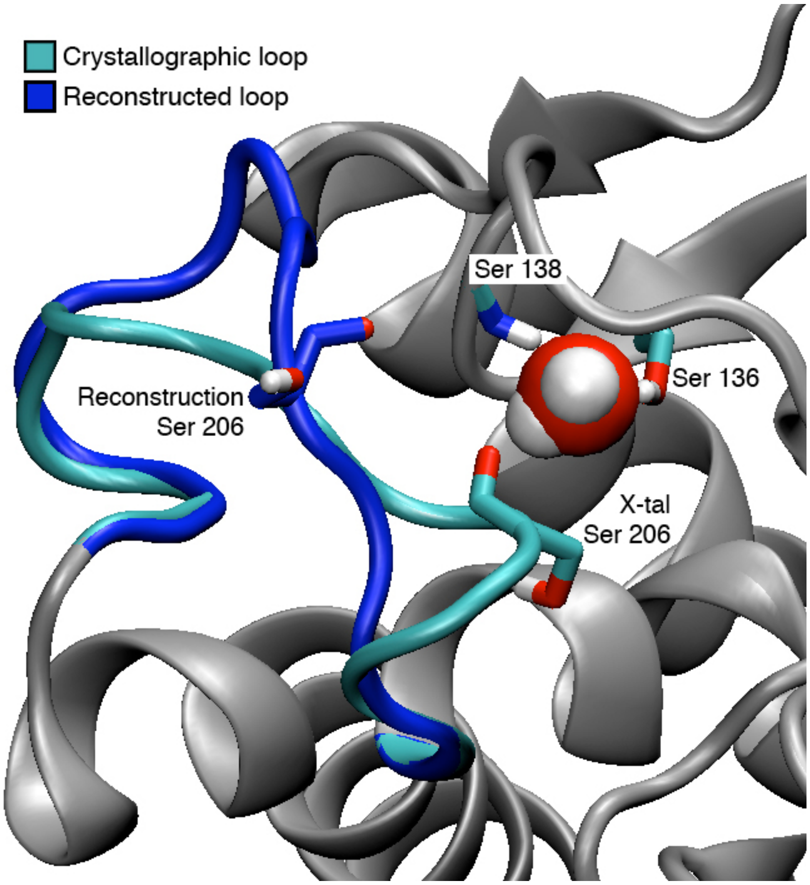

Figure S4 – Loop reconstruction with a complex hydrogen bonding network. (**a**) Loop residues Asp 43 and Tyr 33 form a hydrogen bond network with the backbone amide of Arg 63 and the side-chains of Arg 63 and Glu 31 in the crystal structure of human 5'-deoxy-5'-methylthioadenosine phosphorylase (pdb code 1cb0). (**b**) The loop reconstruction of 1cb0 recovers the hydrogen bonds between loop residue Asp 43 with the amide backbone of Arg 63 and the side-chains of Arg 63 and Glu 31, but orients the side-chain of Tyr 33 towards bulk solvent. The loop was reconstructed to 0.6Å accuracy.

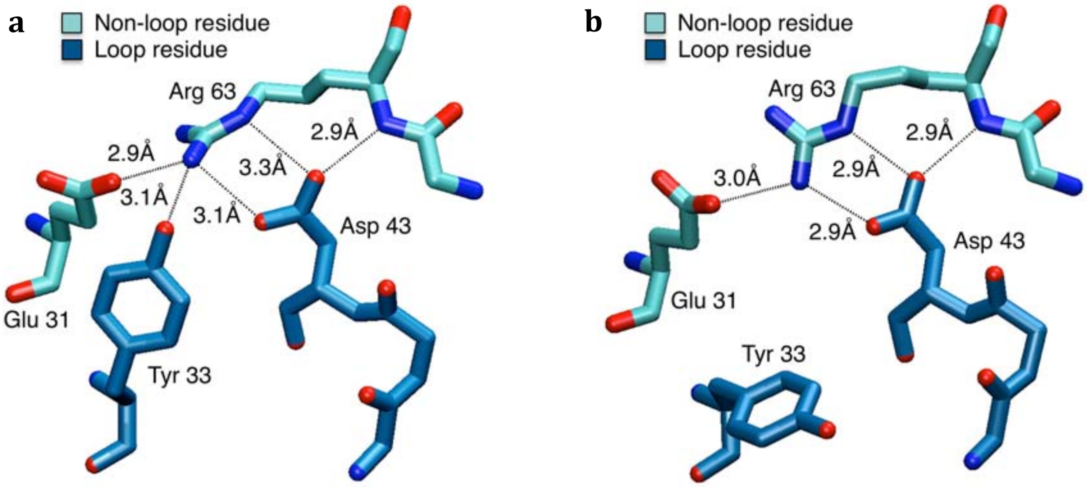

Table S1 – KC and standard protocol loop reconstruction accuracy on dataset 1[3]. Ligand/ion filter is passed if all loop heavy atoms are ≥4.0Å from neutral ligand heavy atoms and ≥6.5Å from charged ions.

Pdb	Loop residues	Standard protocol[b] (Å rmsd of lowest scoring model)	KC protocol (Å rmsd of lowest scoring model)	Pass ligand/ion filter?
1541	153-164	1.6	3.3	No
1arp	201-212	2.3	0.5	No
1ctm	9-20	5.4	2.9	No
1cyo	12-23	0.8	5.2	Yes
1dts	41-52	5.8	6.4	Yes
1eco	35-46	0.6	0.4	Yes
1ede	150-161	1.2	0.7	Yes
1ezm	122-133	2.4	2.7	Yes
1hfc	165-176	8.5	8.2	No
1ivd	365-376	7.4	2.1	No
1msc	9-20	3.7	3.2	Yes
1onc	23-34	3.8	0.5	Yes
1pbe	129-140	2.0	0.6	Yes
1pmy	77-88	2.6	2.6	No
1prn	15-26	7.0	6.6	No
1rcf	88-99	5.0	0.6	No
1rro	17-28	2.2	0.4	Yes
1scs	199-210	2.3	2.9	No
1srp	311-322	2.6	0.6	Yes
1tca	305-316	2.6	0.6	Yes
1thg	127-138	1.6	1.1	Yes
1thw	178-189	2.4	2.7	Yes
1tib	99-110	0.7	1.2	Yes
1tml	243-254	0.7	0.4	Yes
1xif	203-214	1.8	0.7	Yes
2cpl	145-156	0.4	0.2	Yes
2cyp	191-202	0.8	0.5	No
2ebn	136-147	3.9	2.1	Yes
2exo	293-304	1.2	0.8	Yes
2pgd	361-372	3.3	5.1	No
2rn2	90-101	1.1	0.8	Yes
2sil	255-266	2.0	1.0	Yes
2sns	111-122	3.3	3.6	No
2tgi	48-59	4.6	3.1	Yes
3cla	176-187	0.7	1.0	Yes
3cox	478-489	1.0	1.1	No
3hsc	72-93	0.5	0.5	Yes
451c	16-27	4.7	5.8	No
4enl	372-383	2.7	3.6	No
4ilb	46-57	5.1	3.8	Yes
	mean	2.8	2.3	
	median	2.4	1.2	
	filtered mean[a]	2.2	1.6	
	filtered median[a]	2.0	0.8	

[a] Mean/median of cases that pass the ligand/ion filter. [b] Values regenerated from simulations rather than copied from ref[3].

Table S2 – Performance of KC and standard Rosetta protocols, and published results on dataset 2[6]. "Perturbed" columns means simulations begin with the starting structures used in ref[6].

Pdb	Loop residues	Standard protocol de novo rmsd (Å)	KC protocol de novo rmsd(Å)	Standard protocol perturbed rmsd(Å)	KC protocol perturbed rmsd (Å)	Molecular mechanics perturbed rmsd (Å)[a]
1a8d	155-166	5.4	6.9	5.3	0.6	2.8
1arb	182-193	1.6	1.0	5.1	1.4	2.6
1bhe	121-132	7.1	0.8	4.9	0.7	0.7
1bn8	298-309	2.5	0.7	1.7	0.6	2.6
1c5e	82-93	0.8	0.5	5.1	0.4	1.7
1cb0	33-44	1.0	0.6	1.1	0.7	0.3
1cnv	188-199	2.3	1.4	2.8	2.1	3.3
1cs6	145-156	2.5	3.0	4.0	3.0	3.5
1dqz	209-220	1.9	0.7	1.8	2.6	0.6
1exm	291-302	0.6	0.9	2.8	0.9	0.5
1f46	64-75	2.1	2.5	0.7	2.3	1.1
1i7p	63-74	0.7	2.7	0.8	0.4	0.3
1m3s	68-79	3.6	6.3	2.2	5.6	5.6
1ms9	529-540	2.5	0.4	2.8	1.0	2.5
1my7	254-265	2.0	2.3	0.6	2.3	0.9
1oth	69-80	0.6	0.6	1.9	0.6	0.7
1oyc	203-214	3.2	4.0	1.7	3.9	1.2
1qlw	31-42	3.3	1.0	5.0	0.9	1.4
1t1d	127-138	0.5	0.8	0.6	0.8	1.0
2pia	30-41	1.1	1.0	1.0	0.9	0.5
	mean	2.3	1.9	2.6	1.6	1.7
	median	2.1	1.0	2.0	0.9	1.2

[a] Values taken directly from Table S4 in ref[6]

Table S3 – Performance of the KC protocol on the dataset 3.

Pdb	Loop protein	Partner	Chains	rmsd (Å)	Loop RMSDs to the loop conformation in the other complex structures (in order listed)[a]	Pdb loop	Length	Ligand	Cofactor
1doa	Cdc42	Rho GDI	A,B	2.2	**3.7**, **2.9**, 1.7, **4.1**	30-40	11	GDP	Mg2+
1grn	Cdc42	CDC42 GAP	A,B	0.9	**3.7, 5.0, 3.6, 1.1**	30-40	11	GDP/AF3	Mg2+
1gzs	Cdc42	SOP-E (Toxin)	A,B	2.1	**2.9, 5.0, 2.3, 6.2**	30-40	11	none	none
1ki1	Cdc42	Intersectin	A,B	4.3	1.7, 3.6, 2.3, **5.4**	30-40	11	none	none
1nf3	Cdc42	Par	A,C	1.5	**4.1**, 1.1, **6.2**, **5.4**	30-40	11	GNP/MG	Mg2+
1g4u	Rac	GAP SPTP	R,S	0.7	**0.8, 4.6**	30-39	10	GDP/AF3	Mg2+
1he1	Rac	Toxin	C,A	0.4	**0.8, 4.6**	30-39	10	GDP/AF3	Mg2+
1hh4	Rac	Rho GDI	A,D	0.8	**4.6, 4.6**	30-39	10	GDP	Mg2+
1bkd	Ras	Son of Sevenless-I	R,S	6.4	**9.9, 9.8, 9.6**	28-37	10	none	none
1he8	Ras	PI-3 Kinase	B,A	1.7	**9.9**, 0.5, 1.0	28-37	10	GNP	Mg2+
1k8r	Ras	BRY-2RBD	A,B	1.5	**9.8**, 0.5, 0.9	28-37	10	GNP	Mg2+
1wq1	Ras	Ras-GAP	R,G	0.6	**9.6, 1.0, 0.9**	28-37	10	GDP/AF3	Mg2+
1cmx	Ubiquitin	Modified Ubiquitin	B,A	0.3	**3.3, 2.3, 1.3, 1.0, 1.1**	306-312	7	none	none
1fxt	Ubiquitin	Conjugating Enzyme	B,A	0.7	**3.3, 2.3, 2.5, 3.1, 1.5**	6-12	7	none	none
1nbf	Ubiquitin	Deubiquitinating Enzyme	D,A	1.0	**2.3, 2.3, 2.4, 3.0, 2.7**	306-312	7	none	none
1wr6	Ubiquitin	GGA3-GAT	E,A	0.5	**1.3, 2.5, 2.4, 4.3, 0.9**	6-12	7	none	none
1wrd	Ubiquitin	TOM-GAT	B,A	0.6	**1.0, 3.1, 3.0, 4.3**, 0.6	6-12	7	none	none
2d3g	Ubiquitin	HRS-UIM	A,B+P	0.6	**1.1, 1.5, 2.7, 0.9**, 0.6	6-12	7	none	none
			mean	1.5					
			median	0.8					

[a] The core of the loop protein was pairwise-superimposed onto the structures of the loop protein bound to other partners. Global loop rmsds to the loop protein in the other structures are shown in the order listed in the table (descending from top). Cases where the predicted loop rmsd is less than the rmsd to the loop bound to another partner are shown in bold (57 / 68 cases).

Table S4 – KC and standard protocol sampling and scoring errors on dataset 1[3]. Cases where reconstruction accuracy is ≥1.0Å are shown. Gray boxes are primarily scoring errors, white boxes are primarily due to insufficient sampling.

KC Protocol		Standard Protocol	
Pdb	best scoring model score – crystallographic loop score	Pdb	best scoring model score – crystallographic loop score
1ezm	9.47	1tca	21.74
2ebn	4.94	1ezm	15.75
2tgi	4.41	1srp	15.32
1thw	2.03	2exo	12.77
1tib	1.63	1pbe	10.73
4i1b	-0.76	2ebn	8.28
1thg	-1.17	2tgi	7.19
1cyo	-1.58	2rn2	7.18
1dts	-8.64	1thw	6.87
1msc	-39.53	1thg	6.64
		1ede	6.38
		1rro	3.09
		1xif	3.05
		2sil	2.55
		4i1b	2.52
		1onc	1.46
		1dts	-3.50
		1msc	-36.09

Table S5 – KC and standard protocol sampling and scoring errors on dataset 2[6]. Cases where reconstruction accuracy is ≥1.0Å are shown. Gray boxes are primarily scoring errors, white boxes are primarily due to insufficient sampling.

KC Protocol		Standard Protocol	
Pdb	best scoring model score – crystallographic loop score	Pdb	best scoring model score – crystallographic loop score
1a8d	9.36	1a8d	21.98
1f46	7.91	1cnv	18.88
1cnv	5.25	1qlw	16.86
1i7p	4.69	1dqz	16.66
1qlw	4.18	1bhe	14.89
1cs6	2.41	1f46	9.63
1my7	-0.56	1arb	6.32
2pia	-1.63	1bn8	6.12
1m3s	-3.89	1cs6	5.36
1oyc	-5.58	1cb0	5.33
		1m3s	2.82
		1ms9	1.39
		1oyc	1.13
		1my7	0.61
		2pia	0.61

Table S6 – Potential error sources from benchmark dataset 1[3]. Cases where reconstruction accuracy is ≥1.0Å using the KC protocol are shown. Scoring errors are shaded gray as defined in Table S4.

Pdb	Non-modeled factor(s)	Reconstruction rmsd (Å)
1dts	Crystal packing	6.4
1cyo	Crystal packing	5.2
4i1b		3.8
1msc	Crystal packing	3.2
2tgi	Crystal packing, low pH (4.2)	3.1
1ezm		2.7
1thw		2.7
2ebn	*Cis* proline	2.1
1tib	low pH (4.0)	1.2
1thg		1.1

Table S7 – Potential error sources from benchmark dataset 2[6]. Cases where reconstruction accuracy is ≥1.0Å using the KC protocol are shown. Scoring errors are shaded gray as defined in Table S5.

Pdb	Non-modeled factor(s)	Reconstruction rmsd (Å)
1a8d		6.9
1m3s	Crystal packing	6.3
1oyc		4.0
1cs6	*Cis* proline	3.0
1i7p		2.7
1f46	Crystal packing, *Cis* proline	2.5
1my7		2.3
1cnv	low pH (3.0-5.0)	1.4
1qlw		1.0
2pia	Crystal packing	1.0

Table S8 – Sensitivity of reconstruction accuracy to simulation parameters. Mean and median rmsds are shown for the 3 protocols on all 45 loops from dataset 1 (filtered) and dataset 2. The input structures for dataset 2 are the perturbed starting structures used in ref[6].

	KC Protocol	KC Protocol with fixed N-Cα-C bond angles	KC Protocol with temperature fixed at 1.0 *k*T
Mean (Å)	1.6	1.7	1.6
Median (Å)	0.9	1.0	1.0

Supplemental References

1. Fiser, A., Do, R.K., and Sali, A. *Protein Sci* **9**, 1753--1773 (2000).
2. Rohl, C.A., Strauss, C.E.M., Chivian, D., and Baker, D. *Proteins* **55**, 656--677 (2004).
3. Wang, C., Bradley, P., and Baker, D. *J Mol Biol* **373**, 503--519 (2007).
4. Zhu, K., Pincus, D.L., Zhao, S., and Friesner, R.A. *Proteins* **65**, 438--452 (2006).
5. Jacobson, M.P. et al. *Proteins* **55**, 351--367 (2004).
6. Sellers, B.D., Zhu, K., Zhao, S., Friesner, R.A., and Jacobson, M.P. *Proteins* **72**, 959--971 (2008).
7. Jacobson, M.P., Loop decoy sets, Available at http://jacobsonlab.org/decoy.htm, (2008).
8. Dunbrack, R.L.J. and Cohen, F.E. *Protein Sci* **6**, 1661--1681 (1997).
9. Kuhlman, B. et al. *Science* **302**, 1364-1368 (2003).
10. Press, W., Teukolsky, S., and Vetterling, W., *Numerical Recipes: The Art of Scientific Computing, Third Edition*. (Cambridge University Press, Cambridge, 2007).
11. Rohl, C.A., Strauss, C.E.M., Misura, K.M.S., and Baker, D. *Methods Enzymol* **383**, 66--93 (2004).
12. Go, N. and Scheraga, H.A. *Macromolecules* **3**, 178-187 (1970).
13. Wedemeyer, W.J. and Scheraga, H.A. *Journal of Computational Chemistry* **20**, 819-844 (1999).
14. Canutescu, A.A. and Dunbrack, R.L.J. *Protein Sci* **12**, 963--972 (2003).
15. Coutsias, E.A., Seok, C., Jacobson, M.P., and Dill, K.A. *J Comput Chem* **25**, 510-528 (2004).
16. Cortes, J., Simeon, T., Remaud-Simeon, M., and Tran, V. *J Comput Chem* **25**, 956-967 (2004).
17. Lee, A., Streinu, I., and Brock, O. *Phys Biol* **2**, S108-115 (2005).
18. Noonan, K., O'Brien, D., and Snoeyink, J. *The International Journal of Robotics Research* **24**, 971-982 (2005).
19. Shehu, A., Clementi, C., and Kavraki, L.E. *Proteins* **65**, 164--179 (2006).
20. Milgram, R.J., Liu, G., and Latombe, J.C. *J Comput Chem* **29**, 50-68 (2008).
21. Coutsias, E.A., Seok, C., Wester, M.J., and Dill, K.A. *International Journal of Quantum Chemistry* **106**, 176-189 (2005).
22. Hook, D.G. and McAree, P.R., in *Graphics gems* (Academic Press, New York, 1990).
23. Delano, W.L., The PyMOL Molecular Graphics System, Available at http://www.pymol.org.
24. Tsodikov, O.V., Record, M.T., and Sergeev, Y.V. *Journal of Computational Chemistry* **23**, 600-609 (2002).
25. Richards, F.M. *Annu Rev Biophys Bioeng* **6**, 151-176 (1977).
26. Kortemme, T., Morozov, A.V., and Baker, D. *J Mol Biol* **326**, 1239-1259 (2003).
27. Lazaridis, T. and Karplus, M. *Proteins* **35**, 133-152 (1999).
28. Laskowski, R.A., Moss, D.S., and Thornton, J.M. *J Mol Biol* **231**, 1049-1067 (1993).