Supporting Information

Chopra *et al.* 10.1073/pnas.0810818105

SI Text

Protein structure refinement is a challenging unsolved problem that affects protein structure determination and homology modeling: it would seem easy to make a near-native structure move closer to the actual native structure but such refinement generally fails. A basic requirement for refinement is that the native structure occurs at a local minimum of the potential energy that is accessible, i.e., it is surrounded by a smooth attractor basin. Such an energy surface will drive near-native decoys toward the native structure and should also result in a high correlation of potential energy with the deviation from the native structure (rmsd). The measure of improvement in refinement relies on two factors: (*i*) accurate predictions of C^{α} atoms that impact the packing of side chains (1) and (*ii*) good structural alignments given by GDT-HA scores for refined models. In this work we aim to test whether solvent inclusion in a MM force field improves protein structure refinement.

SI Results

Decoys Movement for MD with Explicit Solvent. A comparison of directed movement of decoys on the potential energy surface for the good (negative PC value) and bad cases (positive PC value) for MD in explicit solvent using the SPC water model (2) and all-atom OPLS (OPLS-AA) force field (3) is shown in Fig. S2. A 2D representation of a high-dimensional conformational space (see *SI Materials and Methods*) is shown along with the MD path for each decoy at 0 ps (pink points), 100 ps (blue points) and 200 ps (red points). For bad cases (1h99a1 and 1c1ka_1) we see that MD moves the decoys all over the energy surface for both 100 ps and 200 ps. Moreover, the movement away from the native structure (from 100 to 200 ps) is much larger for bad cases compared with the good cases (1lwba and 1whi). For the good cases with MD the near-native decoys move closer to the native structure in the first 100 ps and remain close to the native structure (see red and blue points in Fig. S2).

GBSA Outperforms KB on New Decoys. We wanted to test how GBSA performs on the new decoys set, i.e., the energy minimized near-native decoys (see *SI Materials and Methods*). Because energy minimization with GBSA results in large improvement on the 729 near-native decoys of the 75 native proteins, does this result hold for the new decoys as well? Fig. S3 shows a comparison of PC values for energy minimization of the 729 near-native decoys and the new decoys of the 9 selected proteins with KB and GBSA potentials. For 729 near-native decoys of each of the 9 proteins, the PC values correspond to improvement in wRMS, compared with improvement in cRMS for the PC values of new decoys. In both cases, the definition of PC is the same as described in the main text.

Fig. S3 shows a greater improvement for GBSA compared with KB for the new decoys, in that the ⟨PC⟩ value for GBSA is -32.63% compared to -25.85% with KB. For the 729 near-native decoys the ⟨PC⟩ value for GBSA is -19.55% compared to -22.67% for KB. Thus, the use of energy minimized near-native decoys (new decoys) does not affect the performance of GBSA potential to move near-native decoys closer to the native structure. Clearly, GBSA is a better refinement method than KB.

Fig. S4 shows the flatness of cRMS for the initial unminimized new decoys ($cRMS_{initial}$) and its change upon energy minimization with GBSA for 1nkd. The mean $cRMS_{initial}$ ($\langle cRMS_{initial} \rangle$) for 1nkd is 1.66 ± 0.78 Å and $\langle cRMS_{final} \rangle$ value for energy minimized new decoys of 1nkd is 0.91 Å for the 320 new decoys. The ⟨PC⟩ value for 1nkd is -45.30% for GBSA. Greater improvement is seen for the decoys with large $cRMS_{initial}$, i.e., the initial decoys far from the native structure are improved more compared with decoys close to the native structure upon minimization with GBSA (see Fig. S4). This is similar to the results obtained in Fig. 2*C* for 729 near-native decoys of 1nkd protein where more improvement is seen for initial decoys with smaller value of $GDT\text{-}HA_{initial}$, i.e., decoys far from the native structure.

Additional MD Simulation Destroys Structure. To confirm our conclusion that MD in explicit solvent moves the near-native decoys away from the native structure, we selected a subset of 5 proteins (out of 20 proteins) and performed MD simulations for 10 ns on a randomly selected set of 30 decoys of each of the 5 proteins (see *SI Materials and Methods*). The selected proteins were: 1lwba, 1whi, 1lvfa, 1c1ka_1 and 1h99a1, chosen to get a mix of good cases (1lwba, 1whi), bad cases (1c1ka_1, 1h99a1), and a PC value between the good and bad cases (1lvfa) with 200 ps of MD.

Fig. S5 compares energy minimization with KB and GBSA with 200 ps of MD in explicit solvent using the SPC water model (2) and OPLS-AA (3) force field (OPLS-AA/SPC) together with 10 ns of MD in explicit solvent using the TIP4P water model (4) and OPLS-AA force field (OPLS-AA/TIP4P). The good cases (1lwba and 1whi) were improved (negative PC) with 200 ps of MD using OPLS-AA/SPC (see Fig. S5(A)) but for 10 ns of MD using OPLS-AA/TIP4P, we see that the decoys move away from the native structure as MD progresses in time (see Table S1). The PC value increases from 29.71% at 1 ns to 63.45% at 10 ns for 1lwba and 42.92% at 1 ns to 79.40% at 10 ns for 1whi. For the bad cases (1c1ka_1 and 1h99a1) with 200 ps MD using OPLS-AA/SPC, the situation gets much worse from 1 ns to 10 ns of MD using OPLS-AA/TIP4P (see Table S1). For 1c1ka_1, the PC value increases from 106.73% at 1 ns to 153.77% at 10 ns compared with 1h99a1 for which MD deteriorates the structure with the PC value of 283% at 1 ns to 452.15% at 10 ns. For 1lva, the PC value at 1 ns is 34.76% and increases to 76.88% at 10 ns.

Thus, in all cases we see that additional MD moves the decoys away from the native structure. This is also clearly shown in Fig. S5(B) where 2 ns curve is always above 1 ns curve and 3 ns curve is always above 2 ns curve and so on. Moreover, 100 ps and 200 ps curves with OPLS-AA/SPC is much lower than the nanosecond curves, which confirms our conclusion that MD in explicit solvent moves the decoys away from the native structure, and additional MD further deteriorates the structure. Nonetheless, there could be an event at a much longer timescale where MD finds the native structure. Because we never know when such an event might occur even for near-native decoys, we conclude MD in explicit solvent does not seem to be a good refinement method to be used in practice.

Future Directions. Better implicit solvent treatments may be important. We are testing the Screened Coulomb Potential Implicit Solvent Model (5, 6) where the reaction field is taken into account as a screened potential with the parameters obtained from experimental data. We are developing a new KB potential that includes a more accurate choice of PMF derivation scheme for a much larger dataset than was used before (7). Before using more computationally expensive simulation techniques like parallel tempering and stochastic optimizations we would like to mimic the environment of protein in a crystal and

test an accurate QMPFF (Quantum Mechanical Polarizable Force Field) potential (8, 9) for refinement.

For this, we are actively working on a "pseudo" implicit model to be used with MM force fields and QMPFF, which might be similar to the solvent environment as observed in protein crystals.

SI Materials and Methods

Here, we provide details of decoy generation followed by methods used for simulation with implicit and explicit solvent and then our analysis procedures.

Generation of Decoy Test Set. The decoy test set used in this work was exactly the same as used by Summa and Levitt (7). Native protein structures solved to high resolution were chosen to be representative so as to not bias the results: specifically, to eliminate redundancy and to have a wide variety of different folds. As explained before (7) a set of 122 proteins was selected from the 27,570 coordinate files in the PDB (release 23 Aug 2004) so as to have good SPACI scores (10, 11) and only contain small heteroatoms such as SO_4, MG, CA, CL, NA, K, NO_3 and NH_4. These 122 proteins were then minimized using the ENCAD potential (12) and a subset of 77 native proteins that moved least were selected (rmsd <2.0 Å). Tirion quasi-elastic modes (13) were calculated for each native structure using single bond torsion angles as degrees of freedom to sample the conformational space near the native structure. A total of 6 quasi-elastic lowest frequency modes were selected to get $3^6 = 729$ near-native decoys for each of the 75 native proteins (out of 77) for which normal modes were generated successfully.

Generation of New Decoys. We selected 9 native proteins (out of 75) and generated near-native decoys using the same methods used by Summa and Levitt (7) but with different mode amplitudes sampled at -1, -0.5, 0, 0.5, 1 compared with -1, 0, 1 in (7). The selected proteins were: 1ail, 1ay7b, 1bkra, 1ctf, 1ln4a, 1mf7a, 1n0r, 1nkd and 1qmr. A total of 6 quasi-elastic lowest frequency modes were selected to get $5^6 = 15{,}625$ near-native decoys for each of the 9 native proteins. We computed cRMS from the native structure for the near-native decoys of each protein and randomly selected a maximum of 10 near-native decoys over cRMS bins of width 0.1 Å. This gives ≈300 near-native decoys per protein with a wide spread in cRMS values (shown for 1nkd in Fig. S4). These selected near-native decoys were energy minimized using ENCAD potential (12) with C^α interatomic distance restraint of 0.1 Å to stay close to the starting structure. These energy minimized decoys are the new decoys; free from bad contacts and have good stereo-chemistry.

Implicit Water Simulations. Energy minimization was done on all of the 729 decoys of each of the 75 proteins and all of the new decoys for the 9 selected proteins (out of 75), using all-atom OPLS (OPLS-AA) parameter set (14). Solvation was modeled by the Generalized Born Surface Area (GBSA) implicit solvent model (15, 16) as implemented in Tinker 3.9 (17). A modified limited memory l-BFGS minimization algorithm (18), also implemented in Tinker 3.9 (19) was used with 0.01 kcal/mole/Å desired RMS per atom gradient for convergence.

The total energy E_{TOT} of the system is given by

$$E_{TOT} = E_{OPLS\text{-}AA} + E_{pol} + (E_{cav} + E_{VdW})$$

where $E_{\mathrm{OPLS\text{-}AA}}$ is the energy of all-atom OPLS force field, E_{pol} is the solute-solvent electrostatic polarization term, E_{cav} is the sum of solvent-solvent cavity term and E_{VdW} is the solute-solvent van der Waals term. Together, $E_{\mathrm{cav}} + E_{\mathrm{VdW}}$ constitute the total solvent accessible surface area of all atoms

$$(E_{cav} + E_{VdW}) = \sum_i \sigma_i SA_i$$

where σ_i is an empirically determined atomic solvation parameter, and SA_i is the total surface area of the solute calculated with a spherical water molecule of radius 1.4 Å. The electrostatic energy of the system is the columbic part of OPLS-AA potential with the polarization energy term E_{pol} defined as the energetic contribution of transfer of the protein from vacuum to water, a continuum medium with dielectric constant ε_{w}. Specifically in kcal/mol we have

$$E_{pol} = -166\left(1 - \frac{1}{\varepsilon_w}\right)\sum_i \sum_j \frac{q_i q_j}{\sqrt{r_{ij}^2 + \alpha_i \alpha_j \exp(-r_{ij}^2/4\alpha_i\alpha_j)}}$$

where α_i is the Born radius of atom i, q_i is the charge of the atom i, and r_{ij} is the distance between atoms i and j. The Born radius is calculated as shown by Still *et al.* (15, 16).

Explicit Water Simulations. Energy minimization and MD simulations were performed on all 729 decoys of a subset of 20 proteins selected from all of the 75 proteins. This subset was chosen to minimize bias with a mix of PC values for KB potential (4). We used Gromacs 3.3 (20–22) with a triclinic periodic box of explicit SPC water molecules (2) and the OPLS-AA (3) force field (OPLS-AA/SPC). For all 729 decoys of each of the 20 proteins, the length of each side of the water box is 80 Å with a 90° angle between any 2 sides. The SPC water molecules were then added and the system was neutralized by adding NA^+ or CL^- counterions to balance the net charge of the protein. To prepare the decoys for MD, we ran 400 steps of Steepest Descent energy minimization with initial step size of 0.1 Å and a maximum force tolerance of 1 KJ/mole/Å. A grid search algorithm for the neighbor searching was used, constructing a new neighbor list every 10 steps with a cut-off distance of 10 Å for the short-range neighbor list. The Particle Mesh Ewald scheme (PME) was used for long-range electrostatics (23) using the Coulomb cut-off distance of 10 Å. A van der Waals (VdW) cut-off type was specified with the Lennard-Jones cut-off distance of 10 Å. All other parameters in energy minimization were set as default values in Gromacs 3.3.

After energy minimization, MD was run at 300K for 200 ps with a time step of 2 fs using the Berendsen thermostat (24) at a reference temperature of 300K and a time constant of 1 ps for the entire system and the Berendsen barostat (24) at a reference pressure of 1 bar, a time constant of 20 ps and a compressibility value of 4.5×10^{-5} $\mathrm{bar}^{(-1)}$ to get the NPT canonical ensemble (constant number of molecules, pressure and temperature). The neighbor searching, PME and VdW parameters were the same as used in energy minimization. MD was run with the initial velocities generated by Maxwell distribution at 300K and a random seed of 11111. Moreover, all of the bonds in the MD runs were constrained using the LINCS algorithm (25) in Gromacs 3.3. All other parameters in the MD runs were set to the default values. In addition we performed energy minimization on the 729 decoys of all 75 proteins using the same protocol that was used to prepare the decoys for MD. This corrects stereochemical deformations like atomic clashes, strained bonds and angles that are formed in the near-native decoy generation procedure.

We also performed 10 ns long MD simulations on a selected subset of 5 proteins (out of 20) each with 30 randomly selected near-native decoys (out of 729 decoys). This subset was chosen to minimize bias with a mix of PC values for previous MD runs of 200 ps in explicit solvent using the SPC water model and OPLS-AA force field (OPLS-AA/SPC). Here, we used Gromacs 4.0 (26) with a triclinic periodic box of explicit TIP4P water

molecules (4) and the OPLS-AA (3) force field (OPLS-AA/TIP4P). For all 30 decoys of each of the 5 proteins, the length of each side of the water box is 80 Å with a 90° angle between any 2 sides. The TIP4P water molecules were then added and the system was neutralized by adding NA^+ or CL^- counterions to balance the net charge of the protein. To prepare the decoys for MD, we ran 400 steps of Steepest Descent energy minimization and 50 ps of position restrained MD where the protein is restrained and water structure relaxes by moving freely in the box. The energy minimization protocol was exactly the same as used by energy minimization with OPLS-AA/SPC. All other parameters in energy minimization were set as default values in Gromacs 4.0.

After energy minimization, we performed positioned restrained MD simulation at 300K for 50 ps with a time step of 2 fs. The simulation parameters used by position restrained MD was exactly the same as used by previous MD runs with OPLS-AA/SPC. The position restrained MD was run in parallel on 20 nodes. For all MD runs with OPLS-AA/TIP4P, we used $3 \times 2 \times 2$ domain decomposition grid and 8 separate PME nodes. The dynamical load balancing feature in Gromacs 4.0 (26) was set to auto. Because the number of PME nodes should be divisible by the FFT grid size, we set this equal to $64 \times 64 \times 64$ in x, y and z directions. During position restrained MD, the protein was fixed with all bonds constrained using the P-LINCS algorithm (27) in Gromacs 4.0. All other parameters in the MD runs were set to the default values in Gromacs 4.0.

After 50 ps of position restrained MD, we performed 10 ns of MD simulation on all 30 selected decoys of each of the 5 proteins using exactly the same simulation parameters that was used to for position restrained MD without restraining the protein.

MultiDimensional Scaling and Energy Contours. Multidimensional scaling (28) is a technique used to visualize high-dimensional data. Starting with the N by N matrix of distances between N points, it provides 2-dimensional coordinates for each point such that the distance between points in 2-dimensions is a good approximation to the actual distances given in the N by N matrix. In general the approximation will be distorted but we have found the implementation by Kamada and Kawai (29) in the open source program GraphViz (30) to work well. The 2-dimensional coordinates allow the points to be plotted on the printed page and visualize a space that would require N-1 dimensions to be free of any distortion (any line, any triangle and any tetrahedron can be drawn without distortion in 1, 2, or 3 dimensions). To simplify the pictures, we select a random subset of 30 decoys (out of 729) and show the initial (in green) and final (red) positions of each point together with the native structure (pink) in Fig. 4. This gives a total of 61 points. Here, each point has an associated potential energy; in Fig. 4 we also use MATLAB's 4-point smoothing to generate energy contours that show the energy surface in the vicinity of the native structure. This representation of the hills and valleys of the PES gives insight into how the decoys move around the native structure. In Fig. 5 we omit energy contours and use 4 points for each MD of the 30 runs (initial, at 0 ps, at 100 ps and at 200 ps) as well the native structure (cyan) to get a total of 121 structures for each plot and in Fig. S2 we use same 4 points for each MD of the 20 runs with the native structure (cyan) to get 81 structures for each plot.

1. Keskin O, Bahar I (1998) Packing of sidechains in low-resolution models for proteins. *Fold Des* 3:469–479.
2. Berendsen HJC, Postma JPM, van Gunsteren WF, Hermans J (1981) *Intermolecular Forces*, ed Pullman B (D. Reidel, Dordrecht, The Netherlands), pp 331–342.
3. Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (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.
4. JorgensenWL, Chandrasekhar J, Madura JD, Impey RW, Klein ML (1983) Comparison of simple potential functions for simulating liquid water. *J Chem Phys* 79:926–935.
5. Hassan SA, Guarnieri F, Mehler EL (2000) A general treatment of solvent effects based on screened coulomb potentials. *J Phys Chem B* 104:6478–6489.
6. Hassan SA, Mehler EL (2002) A critical analysis of continuum electrostatics: The screened Coulomb potential-implicit solvent model and the study of the alanine dipeptide and discrimination of misfolded structures of proteins. *Proteins* 47:45–61.
7. Summa C, Levitt M (2007) Near-native structure refinement using in vacuo energy minimization. *Proc Natl Acad Sci USA*, 104:3177–3182.
8. Donchev AG, Ozrin VD, Subbotin MV, Tarasov OV, Tarasov VI (2005) A quantum mechanical polarizable force field for biomolecular interactions. *Proc Natl Acad Sci USA* 102:7829–7834.
9. Donchev AG, Galkin SG, Pereyaslavets LB, Tarasov VI (2006) Quantum mechanical polarizable force field (QMPFF3): Refinement and validation of the dispersion interaction for aromatic carbon. *J Chem Phys* 125:244107.
10. Brenner SE, Koehl P, Levitt M (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res* 28:254–256.
11. Chandonia JM, *et al.* (2004) The ASTRAL Compendium in 2004. *Nucleic Acids Res* 32:D189–192.
12. Levitt M, Hirshberg M, Sharon R, Daggett V (1995) Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution. *Comput Phys Commun* 91:215–231.
13. Tirion MM (1996) Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett* 77:1905–1908.
14. Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J Am Chem Soc* 118:11225–11236.
15. Still WC, Tempczyk A, Hawley RC, Hendrickson T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* 112:6127–6129.
16. Qiu Q, Shenkin PS, Hollinger FP, Still WC (1997) The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii. *J Phys Chem A* 101:3005–3014.
17. Ponder JW, *et al.* (2001) TINKER: *Software Tools for Molecular Design* (Washington Univ, St Louis), Version 3.9.
18. Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. *Math Program B* 45:503–528.
19. Ponder JW, Richards FM (1987) An efficient newton-like method for molecular mechanics energy minimization of large molecules. *J Comput Chem* 8:1016–1026.
20. Berendsen HJC, van der Spoel D, van Drunen R (1995) GROMACS: A message-passing parallel molecular dynamics implementation. *Comput Phys Commun* 91:43–56.
21. Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J Mol Model* 7:306–317.
22. van der Spoel D, *et al.* (2005) GROMACS: Fast, flexible, and free. *J Comput Chem* 26:1701–1718.
23. Essmann U, *et al.* (1995) A smooth particle mesh Ewald potential. *J Chem Phys* 103:8577–8592.
24. Berendsen HJC, Postma JPM, van Gunsteren WF, DiNola A, Haak JR (1984) Molecular dynamics with coupling to an external bath. *J Chem Phys* 81:3684.
25. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM (1997) LINCS: A linear constraint solver for molecular simulations. *J Comput Chem* 18:1463–1472.
26. Hess B, Kutzner C, van der Spoel D, Lindahl E (2008) GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* 4:435–447.
27. Hess B (2008) P-LINCS: A Parallel Linear Constraint Solver for Mol Simul. *J Chem Theory Comput* 4:116–122.
28. Kruskal JB, Wish M (1978) *Multidimensional Scaling*, ed Uslaner EM (Sage Publications, Inc), pp 07–011.
29. Kamada T, Kawai S (1989) An algorithm for drawing general undirected graphs. *Informat Proc Lett* 31:7–15.
30. Ellson J, Gansner ER, Koutsofios E, North SC, Woodhull G (2003) *Graph Drawing Software*, eds Jünger M, Petra M (Springer, New York), pp 127–148.

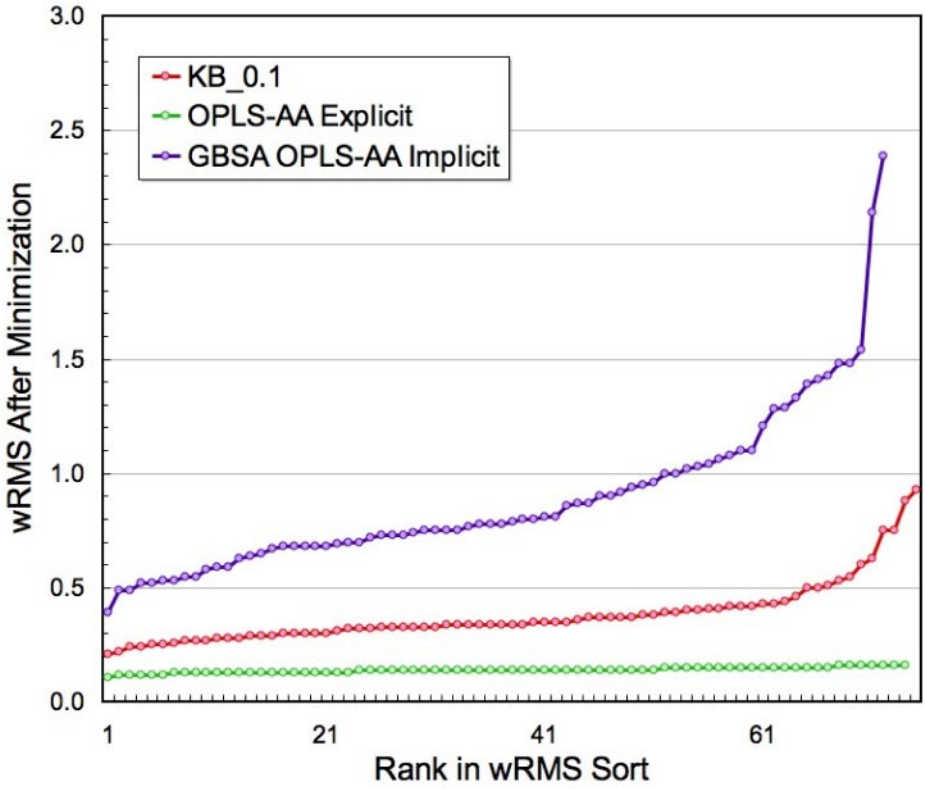

Fig. S1. Potential energy minimization of native structures in implicit and explicit solvent compared with the KB potential. The wRMS value for the 75 native structures is shown for energy minimization runs of KB with Encad (12), GBSA implicit solvent with Tinker (17) and MD in explicit solvent using the SPC water model and the OPLS-AA force field with Gromacs 3.3 (20–22). The wRMS values (*y* axis) are sorted in ascending order and plotted against their rank in the sort. Note, that the native structure at rank 1 does not necessarily reference the same native structure for every condition. The KB potential causes least perturbation to the native structure and the OPLS bulk explicit solvent indicates failure in refinement approach as movement in bulk solvent is greatly restricted; energy minimization does not move the native structure for OPLS bulk explicit solvent.

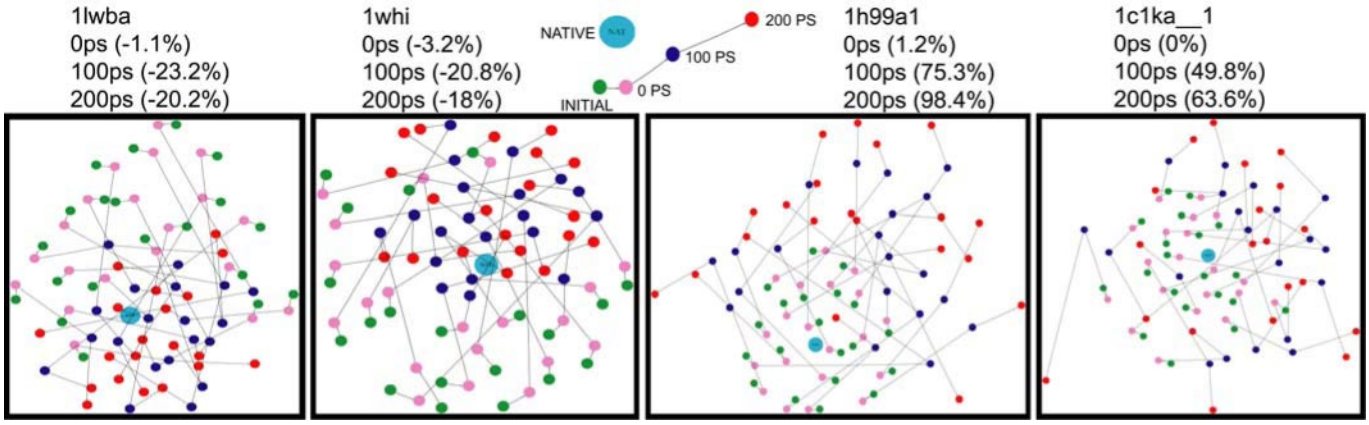

Fig. S2. Showing directionality of movement on the potential energy surface for MD in explicit solvent using the SPC water model and OPLS-AA force field. The green points are the starting initial near-native decoy structures, the pink points are the energy minimized preparatory structure for OPLS explicit solvent MD at 0 ps, the blue points are the decoys at 100 ps and the red points are the decoys at 200 ps of MD. These projections of multidimensional energy surface are made by first selecting a random subset of 20 decoys from the 729 near-native decoys for each protein. The MD path for each of the 20 decoy set selected is shown. An 81 $\times$ 81 matrix is constructed containing the pairwise wRMS values of the 81 structures (the 20 initial starting decoys, the 20 corresponding decoys at 0 ps, 20 decoys at 100 ps, 20 decoys at 200 ps and the native structure, shown as a big cyan disk). Multidimensional scaling is done using Graphviz (30) to get the 2D representation of the 81-dimensional space defined by wRMS matrix. In this plot, points close on the paper are generally also similar with a low wRMS, whereas points far apart on the paper are different with a large wRMS. The points that are smaller in size (1h99a1 and 1c1ka_1) are further apart from the corresponding larger size point pair (1whi) for same length of edge between them. Good (1lwba and 1whi) and bad (1h99a1 and 1c1ka_1) cases are selected using the PC values for each of the protein as shown in parenthesis. For the good cases the protein moves closer to the native structure but the average movement from 100 to 200 ps moves the decoys away from the native structure. For the bad cases, MD moves the structure all over the energy surface with the 200 ps points in red, father apart from the native structure than that of 100 ps points in blue and initial decoys points in green.

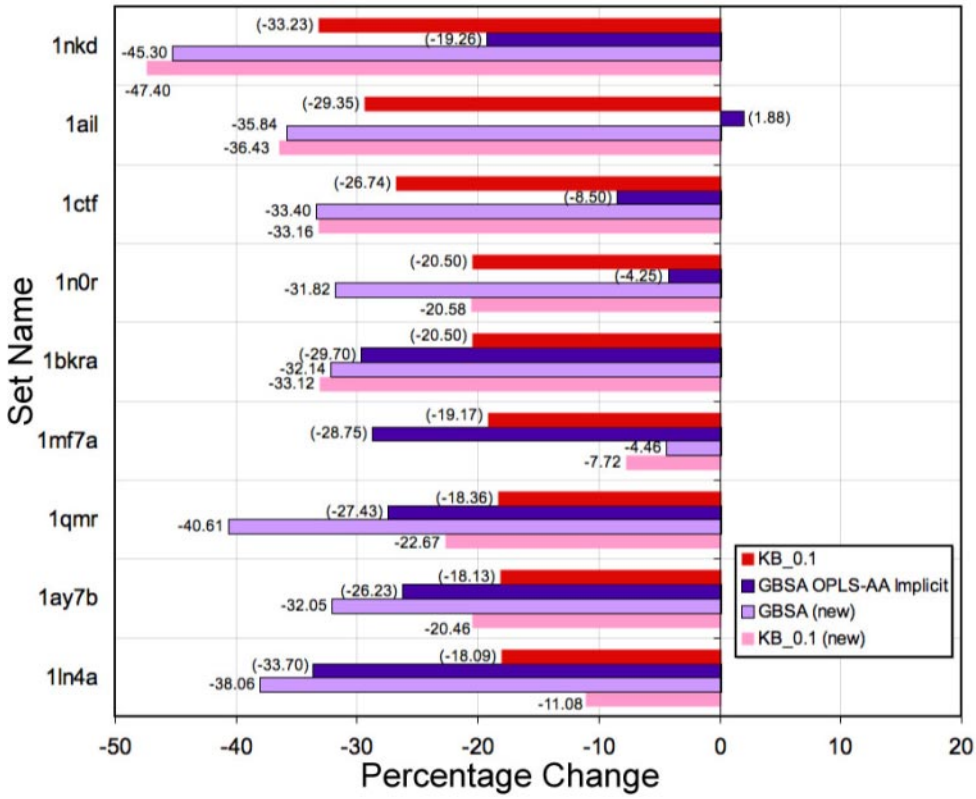

Fig. S3. Potential energy minimization of the 729 near-native decoys and the new decoys for 9 selected native structures with GBSA implicit solvent and KB potential. The 75 proteins are arranged in decreasing order of improvement for the KB potential on 729 near-native decoys for each of the 9 native structures. A negative value of PC corresponds to improvement in the energy minimized structure compared with the starting structure. The PC values (shown in parenthesis) are calculated using wRMS and averaged over 729 near-native decoys for each of the 9 native structures, compared with PC values (shown without parenthesis) calculated using cRMS and averaged over all new decoys for each of the 9 native structure. It is clearly seen that GBSA outperforms KB for 5 of the 9 native structures for both 729 near-native decoys and the new decoys.

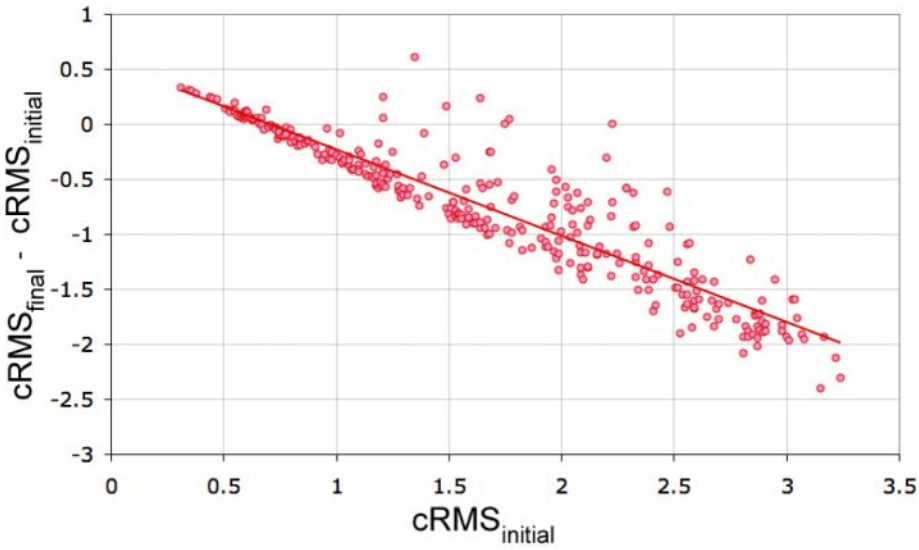

Fig. S4. Showing the improvement in cRMS for each near-native decoy for the new decoys set of 1nkd protein, a good case with GBSA. It is seen that the decoys with large $cRMS_{initial}$ value move toward the native structure ($cRMS_{final}$ value less than $cRMS_{initial}$ value), compared with the initial near-native decoys that are close to the native structure, i.e., decoys with small $cRMS_{initial}$ from native. This makes sense as it is harder to improve a good structure, i.e., near-native decoy close to the native structure compared with a near-native decoy far away from the native structure.

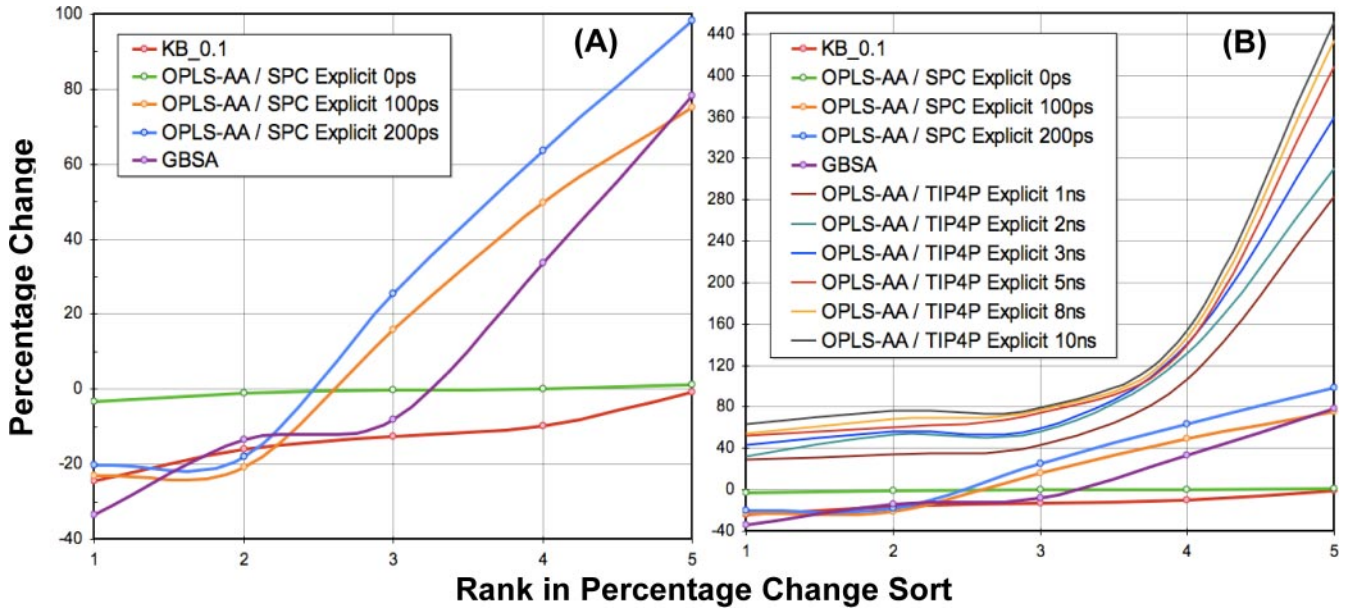

Fig. S5. (A) MD of decoys in explicit solvent using the SPC water model and the OPLS-AA force field (OPLS-AA/SPC) compared with energy minimization with the GBSA and KB potentials. The PC values (*y* axis) are shown for the sampled subset of 5 proteins, sorted in ascending order and plotted against their rank in the sort. Negative values of PC correspond to improvement of the structure. For the 5 proteins, KB gives the largest number of improved proteins, but GBSA does better than KB with a more negative PC value. There is almost no change for MD at 0 ps and additional MD moves the structure away from the native, in that the 200 ps curve is always above those of 100 ps and 0 ps. (B) MD of decoys in explicit solvent using the TIP4P water model and the OPLS-AA force field (OPLS-AA/TIP4P) compared with different scenarios in (A). Energy minimization with KB and GBSA and 200 ps of MD using OPLS-AA/SPC gives better PC values than nanosecond long MD simulations using OPLS-AA/TIP4P as all of the nanosecond simulation curves are above all of the curves shown in (A). Moreover, additional MD using OPLS-AA/TIP4P also moves the structure away from the native, in that 2 ns curve is always above 1 ns curve and so on.

Table S1. Showing ⟨wRMS⟩ values at every nanosecond for a 10-ns trajectory of MD run using OPLS-AA force field and TIP4P explicit water model for a subset of 5 proteins

Set name	$\langle wRMS_{initial} \rangle$ of 30 near-native decoys, Å	⟨wRMS⟩ of 30 Near-Native Decoys for a 10-ns MD trajectory using OPLS-AA/TIP4P, Å									
		1 ns	2 ns	3 ns	4 ns	5 ns	6 ns	7 ns	8 ns	9 ns	10 ns
1lwba	0.8032	1.042	1.064	1.152	1.194	1.225	1.269	1.269	1.243	1.247	1.313
1whi	0.7363	1.052	1.149	1.174	1.212	1.281	1.306	1.287	1.304	1.281	1.321
1lvfa	1.1731	1.581	1.794	1.835	1.867	1.886	1.945	2.011	1.972	2.027	2.075
1c1ka_1	0.9778	2.021	2.266	2.356	2.406	2.342	2.346	2.397	2.409	2.451	2.481
1 h99a1	0.6955	2.664	2.855	3.197	3.430	3.540	3.627	3.704	3.711	3.792	3.840
MEAN	0.877	1.672	1.826	1.943	2.022	2.055	2.098	2.134	2.128	2.160	2.206
STDDEV	0.197	0.688	0.757	0.862	0.935	0.949	0.967	1.002	1.009	1.046	1.042
MIN	0.696	1.042	1.064	1.152	1.194	1.225	1.269	1.269	1.243	1.247	1.313
MAX	1.173	2.664	2.855	3.197	3.430	3.540	3.627	3.704	3.711	3.792	3.840

The ⟨wRMS⟩ values are an average of 30 selected decoys of each of the 5 proteins. The mean and standard deviation of ⟨wRMS⟩ is also shown.