Supporting Information for:

Higher Accuracy Achieved in the Simulations of Protein Structure Refinement, Protein Folding and Intrinsically Disordered Proteins using Polarizable Force Fields

1. Simulated systems.

For protein structure refinements, 18 targets with different topologies from CASP11 were first selected and then a model structure was constructed using our in-house scripts for each of them based on their sequences. The target IDs, sequence lengths, and model structure's RMSDs compared to experimental structures are listed in Table S3.

For protein folding, four fast-folding proteins including chignolin (PDB 1UAO), Trp-cage (PDB 2JOF), villin headpiece (PDB 1YRF), and WW domain (PDB 2F21) were selected. For the simulation of IDPs, four proteins including pepG (sequence EGAAGAASS), p53 (residues 17-29 of PDB 1YCR), NT9 (residues 1-22 of PDB 2HBB) and RS (sequence GAMGPSYGRSRSRSRSRSRSRS) were selected. The fully extended configurations of fast-folding proteins and IDPs were built in PyMOL[1] software and then used as the starting structures of the simulations.

2. Simulation settings.

OpenMM[2] on GPUs with mixed precision was used in the simulations of AMBER, CHARMM and AMOEBA force fields while NAMD[3] was used in the simulations of Drude force field. For the simulation of AMBER99SB, CHARMM36/CHARMM36m, AMOEBA-2013 and Drude-2013 force fields used in this work, water models matched with these force fields were respectively adopted.

Each protein was solvated into a rectangle box of water molecules with at least 1 nm distance from the box edge, followed by the addition of 0.15 M NaCl to neutralize the system. To reduce the box size in the simulation of fast-folding proteins and IDPs, the fully extended configurations were subjected to 10 ns NVT MD simulation and the slightly collapsed configurations were re-solvated into a smaller water box.

Then the systems were initially equilibrated by 500 ps NVT and 500 ps NPT MD simulations. After that, production runs with NPT ensemble at 300 K were performed. Three independent 50 ns standard MD simulations were performed for each experimental and model structures in protein structure refinements. For the simulations of protein folding and IDPs where more conformational samples are needed, four independent 1 μs aMD enhanced sampling simulations with boost potential added to the dihedral term were performed for each protein.

The Langevin Thermostat (or dual Langevin Thermostat[4] for Drude force field) was applied to couple the temperature to 300 K with a collision frequency of 1.0 ps-1, and Monte Carlo Barostat (or Langevin piston Nose-Hoover for Drude force field) was used to keep the pressure to 1.0 bar with a trial frequency to change the box every 50 MD steps. The non-bonded interaction was cut off at 1.0 nm, whereas the electrostatic and van der Waals interactions beyond that were treated with particle mesh Ewald[5] (PME) and dispersion correction method[6].

For all simulations, the Settle Algorithm[7] was applied to keep the water rigid. For all simulations using Drude force field, the time step was set to 1 fs. For simulations of protein structure refinement using AMBER, CHARMM and AMOEBA, the time step was set to 2 fs.

However, in the simulation of fast-folding proteins and IDPs with AMBER, CHARMM and AMOEBA, all bonds involving hydrogen of protein were fixed using CCMA algorithm, and the hydrogen mass in protein was repartitioned to 4 amu to enable an integration step of 4 fs in these simulation.[8-9]

3. Trajectory analyses.

The Python package MDAnalysis[10] and MDTraj[11] were used to calculate the Cα RMSD, radius of gyration of the backbone atoms, number of hydrogen bonds. The DSSP[12] program was used to assign the secondary structures. Chemical shifts were calculated using SPARTA+[13]. The Karplus equation[14] was used to calculate the $^{3}J_{HNHA}$ couplings, and the widely-used ubiquitin parameters[15-16] was adopted. All of the computed NMR observables were compared to those reported in previous studies.

Reference:

1. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.
2. Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput Biol* **2017,** *13* (7), e1005659.
3. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J Comput Chem* **2005,** *26* (16), 1781-802.
4. Jiang, W.; Hardy, D. J.; Phillips, J. C.; Mackerell, A. D., Jr.; Schulten, K.; Roux, B. High-performance scalable molecular dynamics simulations of a polarizable force field based on classical Drude oscillators in NAMD. *J Phys Chem Lett* **2011,** *2* (2), 87-92.
5. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J Chem Phys* **1995,** *103* (19), 8577-8593.
6. Shirts, M. R.; Mobley, D. L.; Chodera, J. D.; Pande, V. S. Accurate and efficient corrections for missing dispersion interactions in molecular Simulations. *J Phys Chem B* **2007,** *111* (45), 13052-13063.
7. Miyamoto, S.; Kollman, P. A. Settle - an Analytical Version of the Shake and Rattle Algorithm for Rigid Water Models. *J Comput Chem* **1992,** *13* (8), 952-962.
8. Hopkins, C. W.; Le Grand, S.; Walker, R. C.; Roitberg, A. E. Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J Chem Theory Comput* **2015,** *11* (4), 1864-74.
9. Eastman, P.; Pander, V. S. Constant Constraint Matrix Approximation: A Robust, Parallelizable Constraint Method for Molecular Simulations. *J Chem Theory Comput* **2010,** *6* (2), 434-437.
10. Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **2011,** *32* (10), 2319-27.
11. McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernandez, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys J* **2015,** *109* (8), 1528-32.
12. Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983,** *22* (12), 2577-2637.
13. Shen, Y.; Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR* **2010,** *48* (1), 13-22.
14. Karplus, M. Vicinal Proton Coupling in Nuclear Magnetic Resonance. *J Am Chem Soc* **1963,** *85* (18), 2870-2871.
15. Wang, A. C.; Bax, A. Determination of the backbone dihedral angles phi in human ubiquitin from reparametrized empirical Karplus equations. *J Am Chem Soc* **1996,** *118* (10), 2483-2494.
16. Hu, J. S.; Bax, A. Determination of phi and chi(1) angles in proteins from C-13-C-13 three-bond J couplings measured by three-dimensional heteronuclear NMR. How planar is the peptide bond? *J Am Chem Soc* **1997,** *119* (27), 6360-6368.

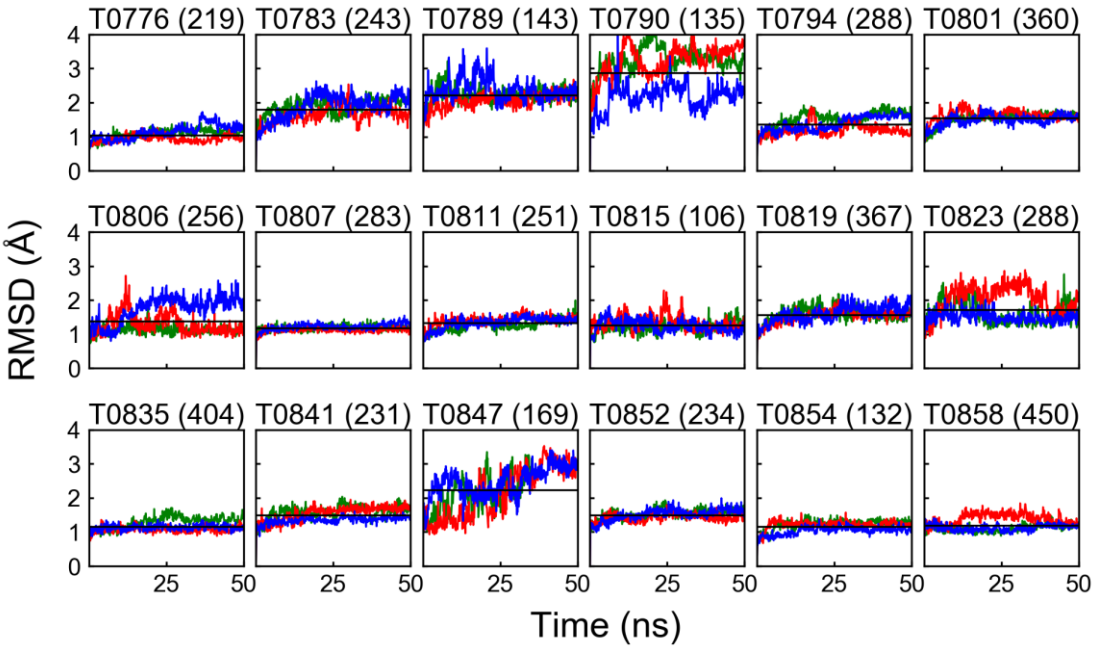

Figure S1. Stability of experimental structures in MD simulations with AMBER force field. For each protein simulated, Cα RMSDs of three independent trajectories as a function of time are colored in red, blue and green. Trajectories stopped due to machine failure were not continued for few systems.

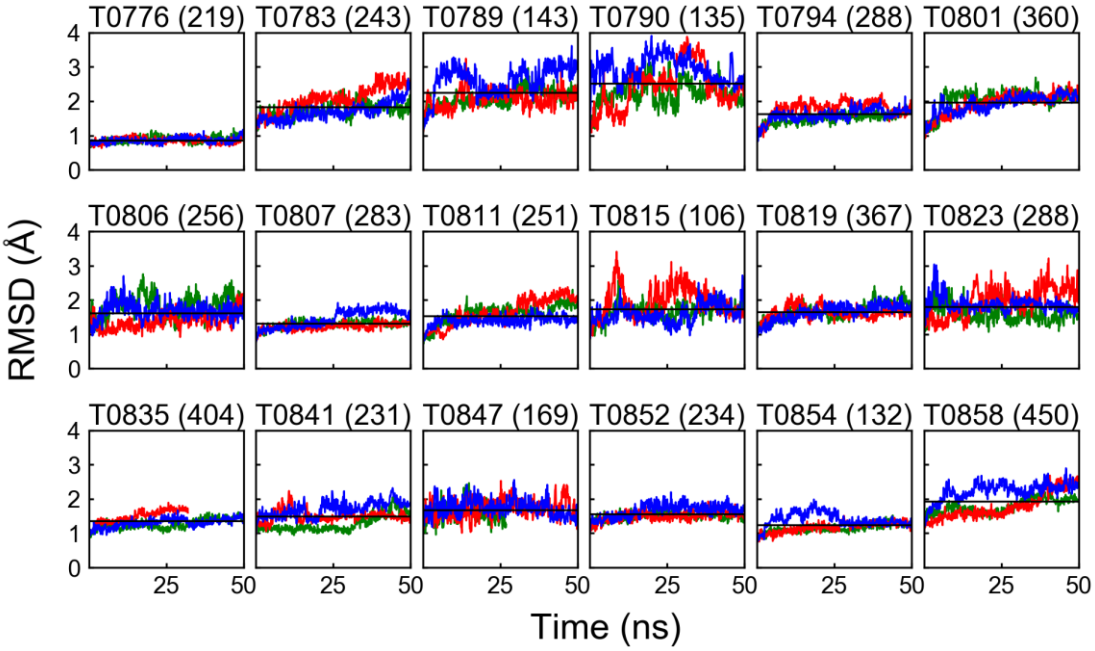

Figure S2. Stability of experimental structures in MD simulations with CHARMM force field. For each protein simulated, Cα RMSDs of three independent trajectories as a function of time are colored in red, blue and green. Trajectories stopped due to machine failure were not continued for few systems.

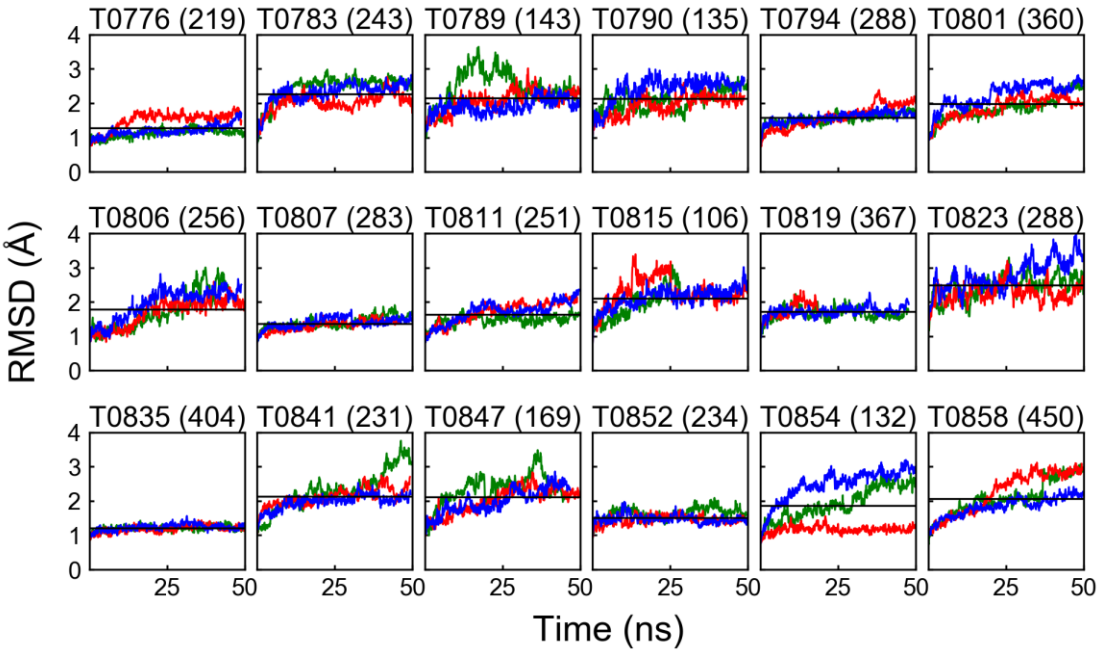

Figure S3. Stability of experimental structures in MD simulations with AMOEBA force field. For each protein simulated, Cα RMSDs of three independent trajectories as a function of time are colored in red, blue and green. Trajectories stopped due to machine failure were not continued for few systems.

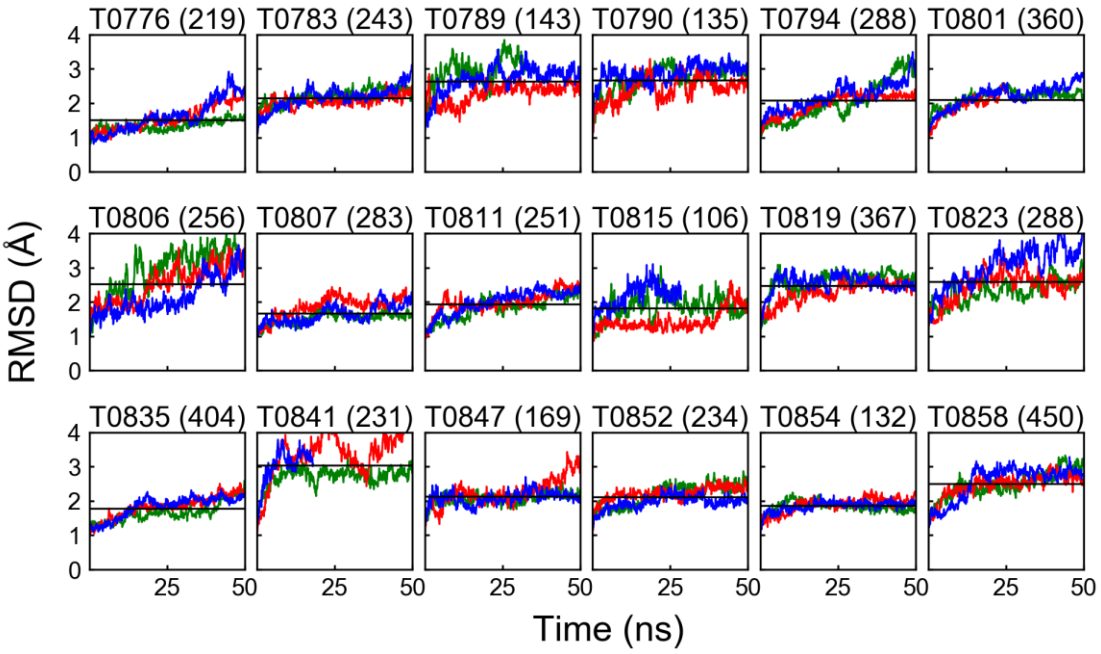

Figure S4. Stability of experimental structures in MD simulations with Drude force field. For each protein simulated, Cα RMSDs of three independent trajectories as a function of time are colored in red, blue and green. Trajectories stopped due to machine failure were not continued for few systems.

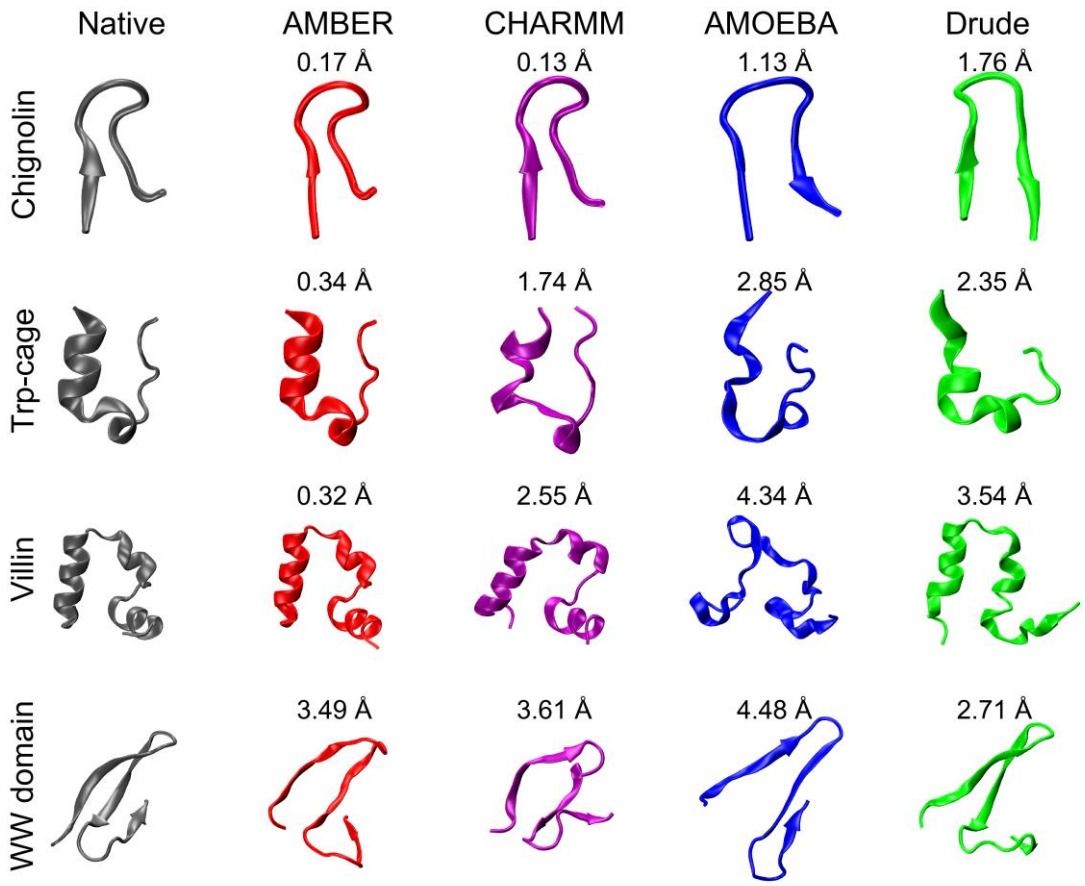

Figure S5. Native structures and conformations with minimal Cα RMSD sampled during MD simulations of each force fields. The native structures are shown in the first column. MD conformations generated from AMBER, CHARMM, AMOEBA, Drude are shown in the second to the fifth column with the corresponding RMSD value labelled above the conformation.

Table S1. The average RMSDs[a] and corresponding fluctuations with all FFs during the MD simulation of native structures.

	AMBER		CHARMM		AMOEBA		Drude	
Target	avg.	std.	avg.	std.	avg.	std.	avg.	std.
T0776	1.05	0.19	0.87	0.09	1.27	0.25	1.51	0.38
T0783	1.80	0.32	1.83	0.33	2.26	0.36	2.15	0.26
T0789	2.22	0.31	2.24	0.41	2.14	0.44	2.60	0.45
T0790	2.88	0.65	2.53	0.55	2.14	0.37	2.67	0.38
T0794	1.37	0.25	1.63	0.23	1.58	0.23	2.08	0.45
T0801	1.55	0.19	1.97	0.30	1.99	0.39	2.13	0.32
T0806	1.38	0.39	1.63	0.33	1.79	0.49	2.53	0.66
T0807	1.18	0.11	1.32	0.19	1.38	0.17	1.68	0.28
T0811	1.34	0.18	1.54	0.30	1.63	0.29	1.94	0.34
T0815	1.27	0.21	1.74	0.37	2.14	0.41	1.77	0.43
T0819	1.57	0.23	1.65	0.22	1.71	0.19	2.47	0.33
T0823	1.71	0.40	1.81	0.35	2.49	0.43	2.60	0.55
T0835	1.15	0.16	1.33	0.18	1.20	0.10	1.77	0.31
T0841	1.49	0.22	1.49	0.28	2.13	0.40	3.04	0.59
T0847	2.24	0.62	1.69	0.27	2.11	0.42	2.13	0.28
T0852	1.48	0.17	1.55	0.17	1.51	0.17	2.12	0.27
T0854	1.15	0.17	1.23	0.20	1.85	0.66	1.85	0.18
T0858	1.19	0.19	1.93	0.36	2.06	0.51	2.49	0.41
avg.	1.56	0.28	1.67	0.29	1.85	0.35	2.20	0.38

[a]All RMSDs are in the unit of angstrom.

Table S2. Overall refinement results for 18 CASP targets with three independent 50 ns MD simulations.

Target	N_{res} [a]	Model RMSD[b]	AMBER ΔRMSD[c]	CHARMM ΔRMSD[c]	AMOEBA ΔRMSD[c]	Drude ΔRMSD[c]
T0776	219	5.73	-1.78	-0.74	-0.76	-0.84
T0783	243	7.61	-0.19	-1.21	-1.27	-0.76
T0789	143	3.93	0.52	0.31	0.12	-0.31
T0790	135	3.63	0.60	0.82	-0.53	-0.26
T0794	288	7.45	-0.09	-0.33	-0.22	-0.77
T0801	360	5.96	0.12	-0.78	-0.48	-0.84
T0806	256	3.91	0.26	0.24	0.34	0.42
T0807	283	4.72	-1.00	-0.38	-0.76	-0.88
T0811	251	4.31	-0.55	-0.30	-0.62	-0.83
T0815	106	2.72	-0.56	0.74	0.30	-0.21
T0819	367	5.2	-0.41	-0.19	-1.34	-1.13
T0823	288	5.3	-0.65	-1.07	-0.79	-1.17
T0835	404	8.25	-1.45	-0.73	-0.79	-0.58
T0841	231	3.63	0.91	-0.11	-0.10	-0.08
T0847	169	4.51	-0.68	0.63	-0.46	-0.39
T0852	234	4.02	-0.22	0.32	0.09	-0.13
T0854	132	3.1	-0.24	-0.09	-0.45	0.00
T0858	450	6.36	0.11	0.62	-0.46	-0.23
avg.	254	5.02	-0.29	-0.13	-0.45	-0.50
No. success[d]			12/18	11/18	14/18	16/18

[a]Number of residues in each target protein. [b]Cα RMSD of the model structure relative to its experimental structure. All RMSDs are in the unit of angstrom. [c]ΔRMSD denotes the difference between representative structure's RMSD and model structure's RMSD for each FFs. [d]Number of successful refinements (ΔRMSD < 0).

Table S3. Overall refinement results for 18 CASP targets with three independent 5 ns MD simulations.

Target	N_{res} [a]	Model RMSD[b]	AMBER ΔRMSD[c]	CHARMM ΔRMSD[c]	AMOEBA ΔRMSD[c]	Drude ΔRMSD[c]
T0776	219	5.73	-1.13	-0.95	-0.85	-0.92
T0783	243	7.61	-0.88	-1.19	-1.19	-0.93
T0789	143	3.93	0.05	0.16	-0.24	-0.45
T0790	135	3.63	0.06	-0.65	-0.62	-0.43
T0794	288	7.45	-0.68	-0.65	-0.36	-0.73
T0801	360	5.96	-0.69	-0.54	-0.75	-0.67
T0806	256	3.91	-0.22	0.34	-0.32	-0.22
T0807	283	4.72	-0.74	-0.35	-0.60	-0.96
T0811	251	4.31	-0.68	-0.26	-0.69	-0.69
T0815	106	2.72	-0.05	0.17	-0.05	-0.38
T0819	367	5.2	-1.12	-1.54	-1.11	-1.06
T0823	288	5.3	-0.51	-0.91	-0.44	-1.10
T0835	404	8.25	-0.91	-0.69	-1.19	-0.70
T0841	231	3.63	-0.03	-0.19	-0.43	-0.27
T0847	169	4.51	-0.61	0.18	-0.52	-0.52
T0852	234	4.02	-0.34	-0.45	-0.32	-0.13
T0854	132	3.1	-0.08	-0.08	-0.51	-0.16
T0858	450	6.36	-0.30	-0.23	-0.40	-0.36
avg.	254	5.02	-0.49	-0.44	-0.59	-0.59
No. success[d]			16/18	14/18	18/18	18/18

[a]Number of residues in each target protein. [b]Cα RMSD of the model structure relative to its experimental structure. All RMSDs are in the unit of angstrom. [c]ΔRMSD denotes the difference between representative structure's RMSD and model structure's RMSD for each FFs. [d]Number of successful refinements (ΔRMSD < 0).