

# CHARMM36m: an improved force field for folded and intrinsically disordered proteins

Jing Huang<sup>1</sup>, Sarah Rauscher<sup>2</sup>, Grzegorz Nawrocki<sup>3</sup>, Ting Ran<sup>1</sup>, Michael Feig<sup>3</sup>, Bert L de Groot<sup>2</sup>, Helmut Grubmüller<sup>2</sup> & Alexander D MacKerell Jr<sup>1</sup>

The all-atom additive CHARMM36 protein force field is widely used in molecular modeling and simulations. We present its refinement, CHARMM36m ([http://mackerell.umaryland.edu/charmm\\_ff.shtml](http://mackerell.umaryland.edu/charmm_ff.shtml)), with improved accuracy in generating polypeptide backbone conformational ensembles for intrinsically disordered peptides and proteins.

There is increasing interest in intrinsically disordered peptides and proteins (IDPs) due to their abundance and functional importance in eukaryotes, as well as their association with various human disorders ranging from cancer to neurodegenerative diseases. Rather than folding into a single, well-defined three-dimensional structure, an IDP fluctuates between an ensemble of interconverting conformational states, which allows some IDPs to interact with several different binding partners, thereby functioning in protein–protein interaction networks<sup>1</sup>. Experimental characterization of conformational ensembles of IDPs is challenging; assistance from computer simulations is often needed, as the number of degrees of freedom of an IDP far exceed the number of available experimental observables<sup>2</sup>. Recent advances in hardware and software allow molecular simulations to reach relevant timescales for sampling IDP conformations, but a major limiting factor lies in the accuracy of their underlying models, typically empirical force fields (FFs)<sup>3</sup>. Protein FFs were mostly developed to target folded proteins, and their accuracy in modeling IDPs needs to be scrutinized and improved<sup>4–6</sup>.

In a recent benchmark study on the structural ensembles of a disordered arginine–serine (RS) peptide obtained with different force fields<sup>4</sup>, the CHARMM36 (C36) protein FF<sup>7</sup> was found to generate a high population of left-handed  $\alpha$ -helix ( $\alpha_L$ ), inconsistent with NMR spectroscopy and small-angle X-ray scattering (SAXS) experimental measurements. We now present an improved C36 FF based on a refined backbone CMAP potential<sup>8</sup> derived from reweighting calculation (see Online Methods)

and a better description of specific salt bridge interactions (see Online Methods). We validate the modified FF, CHARMM36m (C36m), using a comprehensive set of 15 peptides and 20 proteins with a cumulative simulation time of more than 500  $\mu$ s (Supplementary Table 1).

The sampling of  $\alpha_L$  helical conformations in IDP ensembles generated with the C36m FF is significantly lower than that in ensembles generated with C36, as we demonstrated for four IDPs including the RS peptide, the FG-nucleoporin peptide, a hen egg white lysozyme N-terminal fragment (HEWL19) and the N-terminal domain of HIV-1 integrase (IN) (Table 1). The average  $\alpha_L$  propensity of non-glycine, non-proline residues for these four IDPs changes between the two FFs from 20.0% to 5.7%, much closer to the value of 5.1% obtained from protein-coil libraries<sup>9</sup>. With the bias toward  $\alpha_L$  sampling removed, the C36m FF generates molecular dynamics (MD) ensembles that improve the prediction of experimental observables, for example, the NMR scalar couplings for the central alanine residues in the HEWL19 peptide (Supplementary Table 2). The ensemble obtained with C36m for the RS peptide is in significantly better agreement with NMR data (J couplings, chemical shifts and hydrodynamic radius) than is the RS peptide ensemble obtained with C36 (Supplementary Tables 3 and 4). In addition, the predicted SAXS profile from the C36m ensemble of the RS peptide agrees (within the margin of error) with the experimental SAXS curve (Fig. 1), indicating good agreement between computed and measured chain dimensions.

We tested secondary structure sampling for a number of model peptides. The fraction of right-handed  $\alpha$ -helices in the Ac-(AAQAA)<sub>3</sub>-NH<sub>2</sub> peptide simulated with the C36m FF is 17%, which is larger than the C36 result of 13% and closer to the NMR estimates of ~19% and ~21% at 300 K (Supplementary Fig. 1 and Supplementary Table 5). We carried out folding simulations of four  $\beta$ -hairpins (GB1, chignolin, CLN025 and Nrf2); starting from unfolded, fully extended conformations, native-like  $\beta$ -hairpin structures were sampled for all four peptides (Supplementary Figs. 2–5). For the GB1  $\beta$ -hairpin, the MD ensembles generated by the C36m and C36 FFs were found to be very similar (Supplementary Figs. 2 and 3) and consistent with NMR estimate of the folded state population<sup>10</sup>. However, the folded state populations of chignolin and CLN025 are substantially lower than the NMR estimates (Supplementary Table 6), indicating that the C36m FF may underestimate the stability of some  $\beta$ -hairpins.

To directly compare to previous C36 results<sup>11</sup>, we tested C36m by simulating polyglutamine (polyQ) peptide, an IDP with relatively compact collapsed states due to the hydrogen bonding interactions between polar side chains. Similar to the C36 results, the 30-residue polyQ peptide (Q<sub>30</sub>) was found to be disordered

<sup>1</sup>Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, Maryland, USA. <sup>2</sup>Department of Theoretical and Computational Biophysics, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany. <sup>3</sup>Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, Michigan, USA. Correspondence should be addressed to A.D.M. ([amackere@rx.umaryland.edu](mailto:amackere@rx.umaryland.edu)).

RECEIVED 3 MAY; ACCEPTED 5 OCTOBER; PUBLISHED ONLINE 7 NOVEMBER 2016; DOI:10.1038/NMETH.4067

**Table 1** |  $\alpha_L$  conformational sampling in four IDP systems in MD simulations with the C36 and the C36m FFs.

System	Simulation	$\alpha_L$ probability (%)	$\alpha_L$ propensity (%)	Max. $\alpha_L$ length
FG-nucleoporin peptide	C36	32 ± 6	22 ± 2	14 aa
	C36m	1.1 ± 0.3	6.2 ± 0.2	5 aa
RS peptide	C36	80 ± 2	41 ± 1	17 aa
	C36m	1.8 ± 0.5	5.5 ± 0.2	5 aa
IN	C36	64 ± 18	14 ± 2	7 aa
	C36m	3 ± 2	5.6 ± 0.5	4 aa
HEWL19 peptide	C36	11 ± 7	12 ± 2	8 aa
	C36m	0.5 ± 0.4	6.1 ± 0.7	3 aa

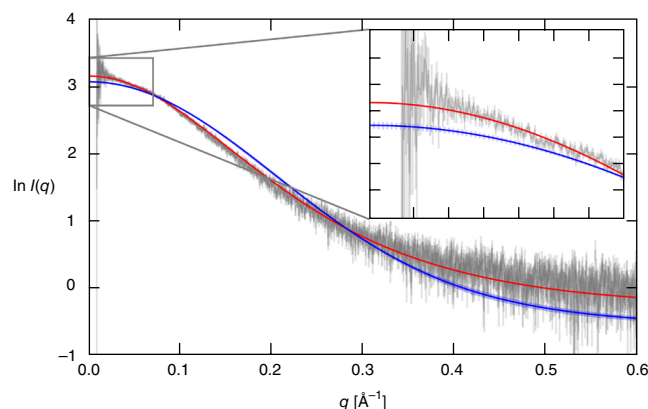
The  $\alpha_L$  probability is computed as the fraction of the ensemble containing left-handed  $\alpha$ -helices, and the  $\alpha_L$  propensity is computed as the probability for non-glycine, non-proline residues to sample the  $\alpha_L$  region. The maximum length of the  $\alpha_L$  helices observed in the simulations is also listed.

during 98% of the simulation time. On average, 7% of  $\phi$  and  $\psi$  torsion angles were in the  $\alpha_L$  region with C36m. A broad conformational ensemble was sampled, with the most favorable states being relatively compact (first major minimum at the end-to-end distance of 25 Å; **Supplementary Fig. 6**). Extrapolation of fluorescence resonance energy transfer (FRET) measurements of shorter polyglutamine peptides<sup>12</sup> to a length of 30 residues leads to an estimate of end-to-end distance of 25 Å, which compares favorably with the C36m simulation results.

We also tested C36m in modeling the kinetics of protein dynamics with the disordered Ac-C(AGQ)<sub>n</sub>-W-NH<sub>2</sub> peptides. The C-terminal tryptophan of C(AGQ)<sub>n</sub>W peptides can be optically excited into the triplet state, with the rate of quenching of that state due to contact formation with the N-terminal cysteine corresponding to the rate of loop closure<sup>13,14</sup>. The decay of triplet survival probabilities calculated from MD simulations compares favorably with experiments for four C(AGQ)<sub>n</sub>W peptides with  $n = 1-4$  at 293 K (**Supplementary Fig. 7**). The computed loop closure rates, as well as both their diffusion-limited parts and their reaction-limited parts, agree with results for shorter peptides ( $n = 1$  or 2) (**Supplementary Table 7**)<sup>14</sup>. The calculated diffusion-limited rates for longer peptides ( $n = 3$  or 4) are higher than experimental estimations, indicating that the simulation ensembles are too collapsed.

We also validated C36m using 15 different folded proteins. All were stable during the 1- $\mu$ s simulation time (**Supplementary Fig. 8**), and the distribution of backbone  $\phi$  and  $\psi$  dihedral angles from these simulations closely resembles the Ramachandran plot of the 'top 500' protein structures<sup>15</sup> (**Supplementary Fig. 9**), indicative of the high quality of the C36m FF in treating the backbone conformational properties of folded proteins. NMR observables of ubiquitin computed from microsecond MD simulations with the C36m FF correlate well with both the C36 results and experimental data (**Supplementary Tables 8-12**).

We additionally performed a folding free energy calculation of villin headpiece (HP) 36 (**Supplementary Table 13** and **Supplementary Fig. 10**); conformational sampling of an N-terminal fragment of HP36, HP21 (**Supplementary Figs. 11-14**); and MD simulations of designed proteins GA95 and GB95 with 95% sequence identity but different folds (**Supplementary Fig. 15**). The HP21 peptide folds to the correct folded state (**Supplementary Fig. 11**), consistent with the fact that the peptide is partially folded and has a preference for its native structure<sup>16</sup>. The most predominant secondary structure is  $\alpha$ -helix (**Supplementary Fig. 12**),



**Figure 1** | SAXS profiles of the RS peptide. Ensemble-averaged scattering curves from the C36 simulation (blue) and the C36m simulation (red) are plotted, with the experimental curve<sup>4</sup> shown with error in gray. The nonweighted error function  $\chi^2$ , as defined in ref. 20, was 0.63 using C36 and 0.12 using C36m. The error bars represent the s.d., computed by dividing the conformational ensembles in two and computing the average SAXS profile for each half separately. For the C36m ensemble, the error bars are smaller than the line width.

as expected based on NMR studies<sup>16</sup> and a recent simulation study of this peptide<sup>17</sup>. The helical state is slightly destabilized (by  $\sim 1kT$ ) as compared to the population, as inferred from the chemical shifts and secondary structure propensity (**Supplementary Fig. 13** and **Supplementary Tables 14** and **15**). For the HP36 protein, we find that C36m gives improved agreement in the folding free energy compared to C36 (**Supplementary Table 13**).

Finally, we comment on the general problem of overly compact IDP ensembles, a problem that is encountered with most physics-based atomistic models<sup>5,6,18</sup>. While the C36m FF, based on the CHARMM-modified TIP3P water model, leads to good agreement between computed and experimental chain dimensions for the RS peptide, the ensemble averaged radii of gyration ( $R_g$ ) of IN and the cold-shock protein from *Thermotoga maritima* (CspTm) with the C36m FF are  $13.8 \pm 0.2$  Å and  $12.8 \pm 0.2$  Å (data presented as mean  $\pm$  s.e.m.). These dimensions are much smaller than the experimental estimates of 24 Å and 15 Å (**Supplementary Table 16**)<sup>19</sup>, respectively, which are inferred from FRET measurements assuming a Gaussian chain model. An approach to correct for this bias is to increase the dispersion interactions between the protein and water. One study applied a general scaling factor to scale up the total protein–water van der Waals interactions<sup>5</sup>, while another proposed a reparametrized water model with the oxygen Lennard–Jones (LJ) well depth  $\epsilon_O$  increased by 50% (ref. 6). Motivated by the difference between the CHARMM-modified water model and the original TIP3P water model, we propose here an alternative water model in which the LJ well depth parameter  $\epsilon_H$  of the water hydrogen atoms is increased (from  $-0.046$  kcal/mol in CHARMM TIP3P) while the oxygen LJ parameters and the water–water interactions are maintained. This approach specifically makes the dispersion part of protein–water interactions more favorable, with minimal perturbation on the repulsive part versus the larger impact of altering the oxygen LJ parameters (see Online Methods).

In simulations using a water model in which the  $\epsilon_H$  value was set to  $-0.10$  kcal/mol, we obtained good agreement with the experimentally estimated  $R_g$  for CspTm (**Supplementary Tables 16**

and 17). In contrast, computed ensemble averaged  $\langle R_g \rangle$  values were found to be larger than the experimental value for the RS peptide and smaller than the experimental value for the IN protein (Supplementary Fig. 16 and Supplementary Table 16). These results suggest that no universal  $\epsilon_H$  can be found to be applicable to all IDP systems and that IDP specific water models may be of utility; further studies are required to address this issue.

**Requests for materials.** Requests should be directed to [amackere@rx.umaryland.edu](mailto:amackere@rx.umaryland.edu).

## METHODS

Methods and any associated references are available in the [online version of the paper](#).

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper.*

## ACKNOWLEDGMENTS

Financial support from the NIH (GM072558 to A.D.M.) and (GM084953 to M.F.) and computational support from the University of Maryland Computer-Aided Drug Design Center, XSEDE (TG-MCA98N017 to A.D.M.) and (TG-MCB090003 to M.F.) and the SuperMUC supercomputer at the Leibniz Rechenzentrum in Garching provided through an allocation by the Gauss Supercomputing Center to S.R. and H.G. are acknowledged. We thank V. Gapsys for helpful discussions. S.R. is supported by a postdoctoral fellowship from the Alexander von Humboldt Foundation.

## AUTHOR CONTRIBUTIONS

J.H. performed the force field optimization. J.H., S.R., G.N. and M.F. ran simulations. J.H., S.R., G.N., T.R. and M.F. analyzed data. J.H., S.R., M.F., B.L.d.G., H.G. and A.D.M. wrote the manuscript. A.D.M. conceived and initiated the research.

## COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the [online version of the paper](#).

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>.

1. Wright, P.E. & Dyson, H.J. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
2. Brucale, M., Schuler, B. & Samori, B. *Chem. Rev.* **114**, 3281–3317 (2014).
3. Mackerell, A.D. Jr. *J. Comput. Chem.* **25**, 1584–1604 (2004).
4. Rauscher, S. *et al.* *J. Chem. Theory Comput.* **11**, 5513–5524 (2015).
5. Best, R.B., Zheng, W. & Mittal, J. *J. Chem. Theory Comput.* **10**, 5113–5124 (2014).
6. Piana, S., Donchev, A.G., Robustelli, P. & Shaw, D.E. *J. Phys. Chem. B* **119**, 5113–5123 (2015).
7. Best, R.B. *et al.* *J. Chem. Theory Comput.* **8**, 3257–3273 (2012).
8. Mackerell, A.D. Jr., Feig, M. & Brooks, C.L. III. *J. Am. Chem. Soc.* **126**, 698–699 (2004).
9. Fitzkee, N.C., Fleming, P.J. & Rose, G.D. *Proteins* **58**, 852–854 (2005).
10. Fesinmeyer, R.M., Hudson, F.M. & Andersen, N.H. *J. Am. Chem. Soc.* **126**, 7238–7243 (2004).
11. Fluit, A.M. & de Pablo, J.J. *Biophys. J.* **109**, 1009–1018 (2015).
12. Walters, R.H. & Murphy, R.M. *J. Mol. Biol.* **393**, 978–992 (2009).
13. Lapidus, L.J., Eaton, W.A. & Hofrichter, J. *Proc. Natl. Acad. Sci. USA* **97**, 7220–7225 (2000).
14. Buscaglia, M., Lapidus, L.J., Eaton, W.A. & Hofrichter, J. *Biophys. J.* **91**, 276–288 (2006).
15. Lovell, S.C. *et al.* *Proteins* **50**, 437–450 (2003).
16. Meng, W., Shan, B., Tang, Y. & Raleigh, D.P. *Protein Sci.* **18**, 1692–1701 (2009).
17. Baltzis, A.S. & Glykos, N.M. *Protein Sci.* **25**, 587–596 (2016).
18. Jensen, M.R. & Blackledge, M. *Proc. Natl. Acad. Sci. USA* **111**, E1557–E1558 (2014).
19. Müller-Späth, S. *et al.* *Proc. Natl. Acad. Sci. USA* **107**, 14609–14614 (2010).
20. Chen, P.C. & Hub, J.S. *Biophys. J.* **107**, 435–447 (2014).

## ONLINE METHODS

**The origin of left-handed  $\alpha$ -helices in the C36 protein FF.** Left-handed helices are very rare in peptides and proteins due to the steric clash between amino acid side-chains and the CO groups, which are bulkier than the NH group, in the case of common right-handed helices. Though present in both models (**Supplementary Fig. 17**), such steric effect is weaker in the C36 FF compared to the C22/CMAP FF, because of the inclusion of refined Lennard–Jones (LJ) parameters for aliphatic carbon atoms in the C36 FF that give improved condensed-phase properties of alkanes<sup>21</sup>. Specifically, the van der Waals (vdW) radius of alanine C $\beta$  atoms changed from 2.06 Å in C22/CMAP to 2.04 Å in C36; for the C $\beta$  atoms in Ile, Thr and Val from 2.275 Å to 2.0 Å; and for the C $\beta$  atoms in the remaining non-Gly, non-Pro amino acids from 2.175 Å to 2.01 Å. Notably, these changes also represent improved treatment in the FF of intramolecular interactions, as the CMAP corrections to the original C22 FF contain large negative values in the  $\alpha_L$  region to account for the steric clash disfavoring  $\alpha_L$  being overestimated due to the larger vdW radii. As improved LJ parameters were adopted in the C36 FF, a decrease in the contribution of the CMAP correction that favors  $\alpha_L$  was needed. However, as the original C36 CMAP (**Supplementary Fig. 17**) has the same CMAP potential in the  $\alpha_L$  region as C22/CMAP, oversampling of the  $\alpha_L$  region occurs, requiring the present additional refinement and subsequent validation of the model.

**Optimization of the CMAP potential.** Central to any optimization problem is its target function. While left-handed  $\alpha$ -helices are very rare, there is little qualitative experimental information on the probability or the average length of left-handed helices or on how often an amino acid populates the  $\alpha_L$  conformation. Protein-coil libraries<sup>9,15</sup> estimate an average  $\alpha_L$  propensity of 6.4% for all amino acids and 5.1% for non-Gly, non-Pro residues. It is anticipated that  $\alpha_L$  propensity in different IDPs will depend on their primary sequence<sup>22,23</sup>, and arguments on the amount of  $\alpha_L$  sampled in specific peptides date back to earlier days of molecular mechanics force fields<sup>24–26</sup>. As experimental target data on the correct amount of  $\alpha_L$  population is lacking, we instead attempted to answer a related and more general question: what is the minimal perturbation of the current CMAP potential energy,  $E$ , that reduces  $\alpha_L$  sampling to an approximate target value? This corresponds to minimizing the following target function:

$$E = kT \ln P + w \text{RMS}_{\text{CMAP}} \quad (1)$$

where  $P$  represents the probability of conformations in a structural ensemble containing a left-handed  $\alpha$ -helix ( $\alpha_L$  probability),  $w$  is an adjustable weighting factor and  $\text{RMS}_{\text{CMAP}}$  is the root mean square difference between the two CMAPs:

$$\text{RMS}_{\text{CMAP}} = \sqrt{\frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n (\text{CMAP}_{i,j}^{\text{new}} - \text{CMAP}_{i,j}^{\text{orig}})^2} \quad (2)$$

where  $m = n = 24$  are the two dimensions of the tabulated CMAP potentials.

Reweighting has emerged as a powerful tool in force field parametrization<sup>27–29</sup>. Given a well-converged conformational ensemble generated by a force field parameter set  $\lambda$ , the ensemble

average of a certain property  $A$  under a new force field parameter set  $\lambda + \Delta\lambda$  is given by

$$\langle A_{\lambda + \Delta\lambda} \rangle = \left\langle A_{\lambda} e^{-\beta(E_{\lambda + \Delta\lambda} - E_{\lambda})} \right\rangle / \left\langle e^{-\beta(E_{\lambda + \Delta\lambda} - E_{\lambda})} \right\rangle \quad (3)$$

where  $E_{\lambda}$  and  $E_{\lambda + \Delta\lambda}$  are the potential energies with FF parameters  $\lambda$  and  $\lambda + \Delta\lambda$  for each sampled conformation, and  $\beta$  is the reciprocal of the thermodynamic temperature. In the context of left-handed helices,  $\alpha_L$  probability  $P$  associated with a certain CMAP modification can be computed as

$$P = \frac{\sum_{i=1}^n h_i e^{-\beta \Delta E_i^{\text{CMAP}}}}{\sum_{i=1}^n e^{-\beta \Delta E_i^{\text{CMAP}}}} \quad (4)$$

where  $n$  is the number of frames,  $h_i$  is a binary that equals 1 if the  $i$ th conformation contains a left helix and 0 if not, and  $\Delta E_i^{\text{CMAP}}$  is the potential energy change associated with the CMAP modification at the  $i$ th conformation. The reweighted  $\alpha_L$  probability was determined based on  $\Delta E_i^{\text{CMAP}}$ , as this was the only energy term adjusted in the FF that directly impacts backbone conformational sampling, which allows efficient evaluation of the target function.

A Monte Carlo simulated annealing (MCSA) simulation is combined with the reweighting equation (equation (4)) to derive the optimized CMAP potential for the target function (**Supplementary Fig. 18**). We carried out  $10^5$  MCSA steps with a starting MCSA temperature of 10 K, and random CMAP revisions between  $-0.01$  kcal/mol and  $+0.01$  kcal/mol were added to individual grid points in a broadly defined  $\alpha_L$  region (**Supplementary Fig. 18**) in each MC step. The full set of  $\phi$  and  $\psi$  values of the FG-nucleoporin (FG) peptide from the MD ensemble generated with the C36 FF at 298 K (ref. 4) was used as the input data, and MCSA optimizations were run with different weighting factor  $w$  values (**Supplementary Table 18**). Smaller  $w$  values lead to more pronounced reduction of  $\alpha_L$  probability, indicating that  $w$  balances between the amount of  $\alpha_L$  and the magnitude of the CMAP modification. The predicted  $\alpha_L$  probability drops to 1.1% with  $w = 2kT$ , and further decreases of  $w$  bring little improvement in  $\alpha_L$  reduction (**Supplementary Table 18**). The CMAP resulting from a  $10^5$  step MCSA run with  $w = 2kT$  is determined to be used as the CMAP for non-Gly and non-Pro residues in the C36m FF. The revision to the original C36 CMAP is localized to the  $\alpha_L$  region around  $\phi = 60^\circ$  and  $\psi = 45^\circ$ , and is much smaller than with the full region allowed to optimize, as indicated by the black lines in **Supplementary Figure 18**. The final number of parameters (i.e., CMAP grid points) being modified is much smaller than the number of parameters allowed to freely change during fitting, which is a good indication that overfitting has been avoided. The penalty term ( $\text{RMS}_{\text{CMAP}}$ ) in the optimization target function helps maintain minimal revision to the CMAP potential while maximally reducing the  $\alpha_L$  probability.

**Improved modeling of the guanidinium and carboxylate salt bridge.** Another refinement in the C36m FF concerns improved description of salt bridge interactions involving guanidinium and carboxylate functional groups with a pair-specific



nonbonded LJ parameter (NBFI term in CHARMM) between the guanidinium nitrogen in arginine and the carboxylate oxygen in glutamate, aspartate and the C terminus. This salt bridge interaction was found to be too favorable in the CHARMM protein force fields, as indicated by the overestimation of the equilibrium association constant of a guanidinium acetate solution<sup>30,31</sup> and by the underestimation of its osmotic pressure (B. Roux, personal communication). The added NBFI term increases the  $R_{\min}$  from 3.55 Å (based on the Lorentz-Berthelot rule) to a larger value of 3.637 Å (R. Shen and B. Roux, personal communication), which we subsequently showed improves the agreement with the experimental osmotic pressure of guanidinium acetate solutions (**Supplementary Fig. 19**). We note that the NBFI approach employed here differs from Piana *et al.*'s work<sup>25</sup>, where the CHARMM22 charges of the Arg, Asp and Glu side chains were reduced in magnitude, with both approaches leading to weaker and more realistic salt-bridge interactions. The NBFI term ensures that only the specific interaction between Arg and Asp/Glu is modified, while the interactions of these residues with other amino acids, water or ions are maintained as in the C36 FF. Again, our aim is to improve the C36 FF with minimal changes in the model.

**Molecular dynamics simulations.** The C36m FF was validated using a variety of systems including peptides, IDPs, unfolded states of proteins and globular proteins. The CHARMM-modified TIP3P model<sup>32</sup> was used in all simulations, unless noted. All the systems studied here are in high dilution, such that the systems did not test the force fields with respect to aggregation. A summary of the validation simulations is given in **Supplementary Table 1**, and detailed information of setup and analysis for each simulation system is given in the **Supplementary Note**. Briefly, temperature replica exchange (T-REX) simulations were carried out with GROMACS<sup>33</sup> for the RS peptide (0.63  $\mu$ s  $\times$  34 replicas), the GB1 hairpin (0.8  $\mu$ s  $\times$  32 replicas), the Nrf2 hairpin (1  $\mu$ s  $\times$  28 replicas), chignolin (6  $\mu$ s  $\times$  29 replicas) and CLN025 (6  $\mu$ s  $\times$  29 replicas). Hamiltonian replica exchange (H-REX) simulation was carried out with CHARMM<sup>34</sup> for polyQ using the end-to-end distance as the biasing reaction coordinate. Harmonic umbrella potentials with a force constant of 0.2 kcal/mol/Å<sup>2</sup> were applied to target end-to-end distances ranging from 5 to 75 Å, spaced at 5-Å intervals. A similar H-REX protocol, using distance as the biasing reaction coordinate, was applied to study the folding free energy of HP21. Conformations were also sampled using single, long-MD trajectories with OpenMM<sup>35</sup>, including 5- $\mu$ s simulations for the HEWL19 peptide, IN and CspTm, 10- $\mu$ s simulations for the (AGQ)<sub>n</sub> peptides and 16- $\mu$ s simulations for (AAQAA)<sub>3</sub>. A 1.2- $\mu$ s simulation of ubiquitin was carried out with NAMD to compare with previous results using the C36 FF<sup>36</sup>. Alternative water models were tested with the RS peptide using T-REX simulations (0.63  $\mu$ s  $\times$  34 replicas) and with IN and CspTm using 5- $\mu$ s MD simulations.

Analysis of MD trajectories was carried out using GROMACS<sup>33</sup> or CHARMM<sup>34</sup>. A left-handed  $\alpha$ -helix is defined as having at least three consecutive residues with  $\phi$  and  $\psi$  falling in the  $\alpha_L$  region ( $30^\circ < \phi < 100^\circ$  and  $7^\circ < \psi < 67^\circ$ ; **Supplementary Fig. 20**). The  $\alpha_L$  probability is computed as the fraction of the ensemble containing left-handed  $\alpha$ -helices, as in ref. 4. We also compute the  $\alpha_L$  fraction as the probability for residues to be in a left-handed  $\alpha$ -helix and calculate the  $\alpha_L$  propensity as the probability for residues'  $\phi$  and  $\psi$  to be in the  $\alpha_L$  region as an additional measurement of left-handed  $\alpha$ -helix sampling.

**Statistics.** Observables were computed as the ensemble average of  $\sim 10^4$  to  $10^6$  frames in MD trajectories. Unless otherwise noted, uncertainties were estimated with block analysis by partitioning MD trajectories into five blocks ( $n = 5$ ).

#### Sampling extended states of IDPs with alternative water model.

A promising way to obtain larger  $\langle R_g \rangle$  is to introduce stronger dispersion interactions between the protein and water. We first test the approach suggested by Best *et al.*<sup>5</sup>, which employs a general scaling factor for the vdW interaction between protein and water. A scaling factor of 1.05 is tested with the C36m FF and confirms that scaling up protein–water vdW interaction leads to more extended conformational states for the RS peptide, the IN proteins and the CspTm proteins (**Supplementary Fig. 16** and **Supplementary Table 16**).

We propose an alternative water model with specific modification of water hydrogen LJ parameters. This is inspired by the difference in conformational sampling with the CHARMM modified TIP3P model and with the original TIP3P model. The CHARMM modified TIP3P contains additional LJ parameters on the hydrogen atoms ( $\epsilon_H = -0.046$  kcal/mol and  $R_{\min}/2 = 0.2245$  Å), so it has more favorable dispersion interactions, which stabilize extended conformations, leading to less-structured conformational ensembles as compared to the original TIP3P water model<sup>4,37</sup>. By further increasing the  $\epsilon_H$  value while maintaining the LJ parameters of the water oxygen atom and resetting the water–water interaction to be same as the CHARMM-modified TIP3P water with NBFI terms, one can specifically make the dispersion part of the vdW interactions between protein and water more favorable while not perturbing the water properties. The advantage of altering the  $\epsilon_H$  value is due to the LJ potential containing both repulsion ( $r^{-12}$ ) and dispersion ( $r^{-6}$ ) terms. Therefore, altering the water oxygen atom LJ parameters will affect its effective size based on the repulsive term such that, for example, the change would alter the balance between the attractive and repulsive interactions of water–protein interactions. In contrast, the water hydrogen atom has a very small LJ radius, so its repulsive wall remains inside the repulsive wall of the oxygen atom in such a way that its LJ term only contributes favorable dispersion interactions. Thus, by only modifying the hydrogen LJ  $\epsilon_H$  parameter, we ensure minimal perturbation of the Hamiltonian, i.e., only the dispersion interaction of the protein with water in the simulation systems is changed. As our goal in the present study was to verify that such an approach would lead to improved sampling of IDPs, we approximately doubled the  $\epsilon_H$  value from  $-0.046$  to  $-0.1$  kcal/mol.

**Code availability.** The computer code used to perform optimization of the CMAP potentials via reweighting is deposited at <https://github.com/jing-huang/CMAPoptimizer>.

**Data availability.** The C36m FF is available along with the remainder of the CHARMM force fields at [http://mackerell.umaryland.edu/charmm\\_ff.shtml](http://mackerell.umaryland.edu/charmm_ff.shtml). More specifically, the parameter file (par\_all36m\_prot.prm) is provided in the toppar\_c36\_jul16.tgz file. The C36m FF is also included in the CHARMM program (version c41 and onward). In addition, the FF may be used in a number of open source molecular simulation programs including NAMD, GROMACS and OpenMM.

21. Vorobyov, I.V., Anisimov, V.M. & MacKerell, A.D. Jr. *J. Phys. Chem. B* **109**, 18988–18999 (2005).
22. Mukrasch, M.D. *et al. J. Am. Chem. Soc.* **129**, 5235–5243 (2007).
23. Mantsyzov, A.B. *et al. Protein Sci.* **23**, 1275–1290 (2014).
24. Gibson, K.D. & Scheraga, H.A. *Proc. Natl. Acad. Sci. USA* **83**, 5649–5653 (1986).
25. Brooks, B.R., Pastor, R.W. & Carson, F.W. *Proc. Natl. Acad. Sci. USA* **84**, 4470–4474 (1987).
26. Roterman, I.K., Gibson, K.D. & Scheraga, H.A. *J. Biomol. Struct. Dyn.* **7**, 391–419 (1989).
27. Li, D.-W. & Brüschweiler, R. *Angew. Chem. Int. Ed. Engl.* **49**, 6778–6780 (2010).
28. Di Pierro, M. & Elber, R. *J. Chem. Theory Comput.* **9**, 3311–3320 (2013).
29. Wang, L.-P. *et al. J. Phys. Chem. B* **117**, 9956–9972 (2013).
30. Piana, S., Lindorff-Larsen, K. & Shaw, D.E. *Biophys. J.* **100**, L47–L49 (2011).
31. Debiec, K.T., Gronenborn, A.M. & Chong, L.T. *J. Phys. Chem. B* **118**, 6561–6569 (2014).
32. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W. & Klein, M.L. *J. Chem. Phys.* **79**, 926–926 (1983).
33. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. *J. Chem. Theory Comput.* **4**, 435–447 (2008).
34. Brooks, B.R. *et al. J. Comput. Chem.* **30**, 1545–1614 (2009).
35. Eastman, P. *et al. J. Chem. Theory Comput.* **9**, 461–469 (2013).
36. Huang, J. & MacKerell, A.D. Jr. *J. Comput. Chem.* **34**, 2135–2145 (2013).
37. Boonstra, S., Onck, P.R. & van der Giessen, E. *J. Phys. Chem. B* **120**, 3692–3698 (2016).