THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Article

Refined Bonded Terms in Coarse-Grained Models for Intrinsically Disordered Proteins Improve Backbone Conformations

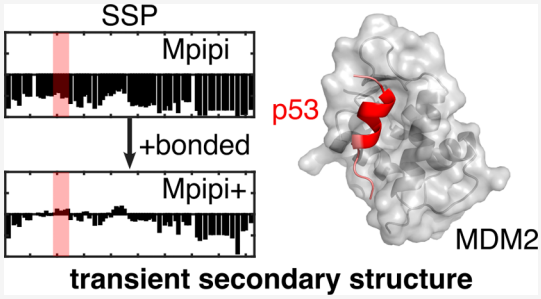
Zixin Hu, Tiedong Sun, Wenwen Chen, Lars Nordenskiöld, and Lanyuan Lu*

Cite This: J. Phys. Chem. B 2024, 128, 6492−6508

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Coarse-grained models designed for intrinsically disordered proteins and regions (IDP/Rs) usually omit some bonded potentials (e.g., angular and dihedral potentials) as a conventional strategy to enhance backbone flexibility. However, a notable drawback of this approach is the generation of inaccurate backbone conformations. Here, we addressed this problem by introducing residue-specific angular, refined dihedral, and correction map (CMAP) potentials, derived based on the statistics from a customized coil database. These bonded potentials were integrated into the existing Mpipi model, resulting in a new model, denoted as the "Mpipi+" model. Results show that the Mpipi+ model can improve backbone conformations. More importantly, it can markedly improve the secondary structure propensity (SSP) based on the experimental chemical shift and, consequently, succeed in capturing transient secondary structures. Moreover, the Mpipi+ model preserves the liquid−liquid phase separation (LLPS) propensities of IDPs.



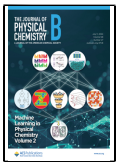
INTRODUCTION

Intrinsically disordered proteins or regions (IDP/Rs) are a class of proteins sampling a broad spectrum of conformations instead of typically folding into a single low-energy state, as folded proteins do. They possess a characteristic amino acid composition, forming stretches of low-complexity (LC) domains. IDP/Rs were found to play pivotal roles in various essential cellular processes, including the cell cycle,[1] cytoskeleton dynamics,[2] translation and transcription,[3] cell communication and signaling,[4] and molecular assembly.[5] Additionally, they are considered to be one of the driving forces of liquid−liquid phase separation (LLPS),[6] a new mechanism governing cell compartmentation, which concentrates specific molecules while excluding others.[7−12] Collectively, IDP/Rs participate in numerous cellular events and are considerably significant to physiological processes within cells.

The functionality of IDP/Rs is largely attributed to short linear motifs (SLiMs), a distinct class of ubiquitous and functionally diverse protein interaction modules.[13] Usually, SLiMs mediate transient protein−protein interactions, existing as hubs for essential cellular functions of IDP/Rs.[14] The backbone conformation is crucial for proper functionality. In some cases, SLiMs remain in disordered states upon binding, taking the dimerization of the ζ chain of T-cell receptor and its interaction with the Nef protein of simian immunodeficiency virus as an example.[15,16] Some other SLiMs adapt transient secondary structures,[17] described as small islands of secondary structures,[18] minimal ordering of SLiMs,[19] etc., as well during functioning,[20] as in the case of MDM2-mediated deactivation of the tumor-suppressor protein, p53, and the rescue by SUMO-specific protease 4, SUSP4.[21] The specific preference toward various conformations, characterized as secondary structure propensity (SSP), is intricately embedded within the sequence of SLiMs.[17−20] Studies on this preference might contribute to the understanding of the roles IDP/Rs play in diverse cellular processes.

The disordered nature presents significant challenges in the study of IDP/Rs. Small-angle X-ray scattering can estimate the radius of gyration (R_g), a common variable, to roughly describe the size of IDP/Rs.[22] Single-molecular Föster resonance energy transfer (smFRET) can examine the dynamics of IDP/Rs.[23] Nuclear magnetic resonance (NMR), a more advanced technique, provides more intricate details regarding the conformations of IDP/Rs.[24] Due to its robustness and informativeness, NMR chemical shifts or chemical-shift-based SSP scores are central judgmental criteria in assessing the presence of transient secondary structures.[20] However, none of these experimental methods directly reveal the detailed conformations of IDP/Rs. This limitation hinders researchers from gaining direct insight into the structure, impeding a more profound understanding of how IDP/Rs precisely modulate numerous physiological cell activities. For acquiring complementary information, investigating IDP/Rs and their asso-

Received: April 30, 2024
Revised: June 13, 2024
Accepted: June 18, 2024
Published: July 1, 2024



https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

ciated functions requires the development of innovative computational tools.

Within the realm of computational biology, molecular dynamics (MD) simulation serves as a popular tool, providing detailed insights into molecular behaviors. The criterion in the selection of a suitable model (i.e., force field) for MD simulation is the balance between precision and computational efficiency. Notable among these modeling paradigms are the atomistic models and coarse-grained models developed for IDP/Rs, both of which have achieved considerable success across a broad spectrum of systems.[25−32] The atomistic model, wherein each atom is represented as one "bead", the interacting unit, is computationally expensive. Consequently, it is more suitable for single-chain systems. Conversely, coarse-grained (CG) models, delineating clusters of atoms as a collective coarse-grained "bead", attains prevalence in larger-scale phenomena, such as LLPS, which involves a large number of proteins.[33] The degree of coarse-graining manifests as a continuum, ranging from multiple beads per amino acid[34] to multiple amino acids per bead.[35] Amidst the protein CG models, the one-bead-per-amino-acid model preserves sequence specificity while minimizing computational cost. There emerged distinctive principles for developing one-bead-per-amino-acid models tailoring specifically to different features of IDP/Rs.

One is the Kim−Hummer (KH) model developed by Kim and Hummer in 2008.[36] It is underpinned by the statistical contact potential originating from the Protein Data Bank (PDB), developed by Miyazawa and Jernigan in 1985[37] and further refined in 1996,[38] and hence subsequently recognized as the Miyazawa−Jernigan (MJ) potential. It is then applied to IDP/Rs engaged in the LLPS system, incorporating adjustments to contact energy parameters.[30]

The second guiding principle pertains to the hydrophobicity scale (HPS), giving rise to the HPS model and subsequently fostering diverse adaptations for IDP/Rs in LLPS systems employing various philosophies. Of particular note are two prominent HPS scales: the HPS-KR scale and the HPS-Urry scale. The former, developed by Kapcha and Rossky in 2014,[39] calculates the weighted sum of binary atomistic hydrophobicity values. The latter, developed by Urry et al. in 1992, relies on inverse temperature transition.[40] HPS-KR was applied on IDP/Rs in LLPS systems incorporating an optimized absolute scale for faithfully reproducing experimental R_g values for IDP/Rs.[30] Simultaneously, the HPS-Urry approach has undergone refinement and practical application in the context of LLPS systems by Regy et al.[41] Beyond mere modifications to the hydrophobicity scale, attempts have been made to introduce previously neglected potentials into the existing HPS model. An illustrative example is the work of Das et al., who introduced cation−π interactions into the HPS-KR model, resulting in the HPS + cation−π model.[42] Furthermore, innovative approaches leveraging machine-learning skills have been employed to derive novel adaptations. Tesei et al., for instance, applied a Bayesian parameter-learning procedure to reparametrize the hydrophobicity scale, resulting in a machine-learning-driven adaptation that exhibits commendable performance in capturing essential features such as R_g and paramagnetic resonance enhancement data.[31]

Another philosophy involves the integration of atomistic simulations and bioinformatics data, as exemplified by the Mpipi model. Conceived by Joseph et al. in 2021, this model distinguishes itself through novel functional forms for non-bonded interactions between amino acid pairs and different assignments of charges to the charged amino acids. The Mpipi model has proved to be effective in accurately predicting both single-chain properties and relative LLPS propensities.[32]

Notably, the investigation of CG models for IDP/Rs, specifically the one-bead-per-amino-acid CG models, has predominantly focused on preserving connectivity while excluding angular, dihedral, and CMAP[43] potentials for bond terms. This selective emphasis, aimed at allowing ample flexibility to accommodate the disordered nature of IDP/Rs, is underpinned by an implicit assumption: either the potential of mean force (PMF) of each can be adequately reproduced by null potentials or their influence on the IDP/Rs behaviors can be deemed negligible. Inaccurate conformations may arise when both statements break down.

To verify the former, we investigated the PMF of each with the existing one-bead Mpipi model, devoid of angular, dihedral, and CMAP potentials. The Mpipi model is a newly developed multiscale CG model applying the parametrization from atomistic simulations and bioinformatics data and has the advantages of applying a novel LJ-like potential avoiding truncation, shifting, and interpolation and recapitulating the LLPS propensity trends among several essential systems.[32] However, the results revealed a significant divergence in PMF profiles when contrasted with the PMF derived from the customized coarse-grained coil database. More intriguingly, while the dihedral and CMAP PMF exhibit a sequence-independent nature, the sequence-dependent characteristics of angular PMF suggest a potential compromise in capturing specific sequence features following the omission of angular potentials. To investigate the latter, we examined the conformations generated by the original Mpipi model and the model implemented with the developed angular, dihedral, and CMAP potentials (denoted as the Mpipi+ model for simplification). Results show that our implementation improved the backbone conformation. This refinement partially improved the Ramachandran plot of each amino acid type and achieved near-quantitative accuracy with experiments on the secondary structure propensity (SSP) based on chemical shifts. In a pivotal case study involving the well-known p53−MDM2 system, our Mpipi+ model precisely predicted the region prone to adopt a transient helical structure upon binding, concordant with experimental evidence derived from the crystallographic structure. Crucially, comprehensive analyses encompassing the R_g, internal scaling profile (ISP), and LLPS behaviors elucidate that the Mpipi+ model, while reinstating specific conformational features and SSP, exerts negligible effects on the overall shape, flexibility, and LLPS propensity compared to the original Mpipi model.

METHODS

Development of a Customized Coil Database. All the atomistic structures sourced from the Coil library[44] underwent a segmentation process at breakpoints corresponding to missing residues. Subsequently, only segments with a length equal to or exceeding 5 residues were retained to form a customized coil database, encompassing a total of 6,523,391 fragments. We kept Cα atoms only to construct a customized coarse-grained coil database. Weights assigned to each structure were derived based on their respective sequences, employing a "crispy", which means "yes" or "no" instead of a value ranging from 0 to 1, sequence similarity strategy. For a given sequence A, if another sequence B contains sequence A

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

or vice versa, sequence A and sequence B are deemed "similar". If there are n sequences (sequence A itself included) similar to sequence A in the customized database, then the weight for sequence A is $1/n$.

Development of Residue-Specific Angular Potentials. Following the work of Ghavami et al,[45] angular values were extracted from the customized coarse-grained coil database with in-house scripts. We then performed amino acid perturbation at the first, second, and third positions of an angle and calculated the corresponding discrete weighted distributions with a bin size of 1°. Due to the substantial influence of amino acid perturbations on these distributions, we applied hierarchical clustering to the three perturbation distributions. The resulting clustering yielded 7 types for amino acids at the first position, 14 types for amino acids at the second position, and 3 types for amino acids at the third position, thus culminating in 294 distinct angle types. We then used a multi-Gaussian distribution with either 3 or 4 terms to fit each type of angular distribution utilizing the particle-swarm optimization (PSO) method integrated into MATLAB.[46] Given the singularity issue associated with dihedral angles at 180°, we added a penalty for angular potentials at 180°, ensuring each above 20 k_BT. We implemented the residue-specific angular potentials in a Gaussian angle style, and the optimization process was grounded in Kullback−Leibler (KL) divergence.[47] Comprehensive details regarding all angle types are provided in the Supporting Information.

The resultant angular distributions for each type were obtained from single-chain simulations, excluding instances related to a specific case study.

Development of Dihedral Potential. Following the work of Ali Ghavami et al,[45] dihedral angle values were extracted from the customized coarse-grained coil database with in-house scripts. We then performed amino acid perturbation at the first, second, third, and fourth positions and calculated the corresponding discrete weighted distributions with a bin size of 1°. Since amino acid perturbation showed minimal impact on distribution (data not shown), we consolidated all dihedral distributions into a single general type with a bin size of 1°. We used iterative Boltzmann inversion (IBI)[48] to get the potential of mean force (PMF) of the dihedral angle. To ensure smoothness and periodicity in the potential, we applied the "sgolay" method in MATLAB with a smoothing factor of 0.1 to the dihedral PMF with three periods. We implemented the general dihedral potential as a tabulated potential with an interval of 1°. The resultant dihedral distribution was obtained from single-chain simulations, excluding instances related to a specific case study.

Development of Correction MAP Potentials. Certain atomistic force fields have incorporated the correction map (CMAP) to rectify the (φ, ψ) distribution within individual amino acids, which was proven to improve the backbone conformations, thereby leading to improvements in secondary structures.[43,49,50] There are also atomistic and multiple-bead-per-amino-acid models incorporating a customized CMAP, especially for IDP/Rs.[25,26,51−53] In the one-bead-per-amino-acid CG model, individual amino acids are represented as single beads and an explicit (φ, ψ) space ceases to exist as internal coordinates. According to the works of Tozzini, Rocchia, and McCammon,[54] information stored in the (φ, ψ) space can be analytically mapped to internal variables describing backbone conformation. In a comprehensive book providing a review celebrating 50 years of the Ramachandran map, a succinct description of backbone conformation through internal variables, including a pseudoangle and a pseudodihedral angle, denoted as (α, θ) space, or two continuous pseudodihedral angles, namely, (θ_1, θ_2) space can be found.[55] For ease of implementation, we selected the continuous pseudodihedral angle (θ_1, θ_2) space as our internal variables. The (α, θ) space and (θ_1, θ_2) space were used for the secondary structure analysis from the free-energy landscape view for the p53 protein in the later part of this article.

Dihedral angle values were extracted from the customized coarse-grained coil database with in-house scripts. We then performed amino acid perturbation at the first, second, third, fourth, and fifth positions and calculated the corresponding weighted distributions. Since amino acid perturbation showed minimal impact on distribution (data not shown), we consolidated all the (θ_1, θ_2) distributions into a single, general type with a mesh size of 15° × 15°. We implemented the general CMAP potential as a 24 × 24 matrix with a mesh size of 15° × 15°.

Iterative Boltzmann Inversion for CMAP. We employed iterative Boltzmann inversion (IBI) to obtain the (θ_1, θ_2) CMAP potential. To alleviate the biasing toward amino acid types or lengths, we selected 8 IDPs with the highest diversity of amino acid types and varying lengths from DisProt[56] (Table 1). To assess the convergence and isolation of different potentials, we additionally generated test CMAP potentials exclusively for A5HC98 (designated as 5 in Table 1) and P0AG63 (designated as 2 in Table 1).

Table 1. IDPs for CMAP Iteration

#	DisProt ID	UniProtKB	PDB ID	length	# amino acid types
1	P02754	DP00193	1BSQ	162	20
2	P0AG63	DP00242	2YKR	80	20
3	P62925	DP00340	1CBI	136	20
4	Q928V6	DP02495	4KIS	214	20
5	A5HC98	DP02521	6Z8K	315	20
6	P07221	DP00132	1A8Y	324	19
7	P01555	DP00250	1S5F	154	19
8	P47047	DP01867	5OOQ	104	19

We initialized the CMAP potential with a zero potential as the starting point. Assuming the target distribution is $g_{ref}(\theta_1, \theta_2)$, the potential implemented at iteration i is $V_i(\theta_1, \theta_2)$, and the resulting distribution is $g_i(\theta_1, \theta_2)$, the implemented potential at iteration $i + 1$ is calculated as follows

$$V_{i+1}(\theta_1, \theta_2) = V_i(\theta_1, \theta_2) + \alpha \cdot \frac{g_i(\theta_1, \theta_2)}{g_{ref}(\theta_1, \theta_2)} \quad (1)$$

where α represents the step length. Following the work of Jiang et al., we used the similarity coefficient (S) between the (θ_1, θ_2) distribution from the customized coarse-grained coil database as $n_{ref}(\theta_1, \theta_2)$ and that from simulations $n_{MD}(\theta_1, \theta_2)$ to measure the similarity between them, assuming that two identical distributions give $S = 1$.[49,50] The similarity coefficient is calculated as

$$S = \frac{\sum n_{ref}(\theta_1, \theta_2) \cdot n_{MD}(\theta_1, \theta_2)}{\sqrt{\sum n_{ref}(\theta_1, \theta_2)^2} \cdot \sqrt{\sum n_{MD}(\theta_1, \theta_2)^2}} \quad (2)$$

Following 7 rounds of iterations, the similarity coefficient reached 99.47% (Table 2). The CMAP potential derived from

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

Table 2. Convergence of the CMAP Potential Iteration[a]

	distance M	distance E	convergence extent ave	convergence extent	similarity coefficient (%)
Mpipi	0.618	0.034	0.984	0.6333	77.39
Iter 0	0.305	0.020	0.565	0.3637	93.22
Iter 1	0.191	0.012	0.350	0.2256	97.43
Iter 2	0.127	0.008	0.234	0.1508	98.86
Iter 3	0.099	0.0067	0.191	0.1231	99.24
Iter 4	0.088	0.0060	0.170	0.1096	99.40
Iter 5	0.083	0.0057	0.162	0.1041	99.46
Iter 6	0.082	0.00564	0.161	0.1036	99.46
Iter 7	0.080	0.00561	0.160	0.1031	99.47

[a]Distance M stands for Manhattan distance. Distance E stands for Euclidean distance. Convergence extent ave stands for the ratio of the Euclidean distance between current distribution and reference distribution with that between even distribution and reference distribution. This ratio provides an intuitive understanding of the degree of convergence. The smaller it is, the closer the optimization it is to the end.

the seventh iteration, utilizing data from 8 IDPs, was designated as the final CMAP potential.

Coarse-Grained Model. We used the Mpipi model for all the Mpipi simulations in this paper. Each amino acid was depicted by a single bead positioned at its Cα atom. We adopted the methodology of Joseph et al.[32] to calculate the potential energy of a given protein as

$$E_{\text{Mpipi}} = E_{\text{bond}} + E_{\text{elec}} + E_{\text{pair}} \tag{3}$$

where the bond energy is computed by harmonic bond potentials when the bond length is r_i

$$E_{\text{bond}} = \sum_{\text{bond}} k(r_i - r_{\text{ref}})^2 \tag{4}$$

with spring constant k = 9.6 kcal/(mol·Å^2) and reference bond length r_{ref} = 3.81 Å. The electrostatic potential applied a Coulombic term with Debye−Hückel electrostatic screening

$$E_{\text{elec}} = \sum_{i,j} \frac{q_1 q_2}{4\pi\epsilon_r\epsilon_0 r_{ij}} e^{-\kappa r_{ij}} \tag{5}$$

where ϵ_r = 80 is the relative dielectric constant of water, ϵ_0 is the electric constant, and κ = 0.126 is the recipient of the Debye screening length, corresponding to a monovalent salt concentration of 150 mM. The Coulombic cutoff is 35 Å.

The nonbonded interactions between proteins were modeled via the Wang−Frenkel (WF) potential,[57] which between two beads of type i and j at distance r writes

$$\phi_{ij}(r) = \epsilon_{ij}\alpha_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{2\mu_{ij}} - 1\right]\left[\left(\frac{r_{ij}}{r}\right)^{2\mu_{ij}} - 1\right]^{2\nu_{ij}} \tag{6}$$

where

$$\alpha_{ij} = 2\nu_{ij}\left(\frac{R_{ij}}{\sigma_{ij}}\right)^{2\mu_{ij}}\left\{\frac{2\nu_{ij}+1}{2\nu_{ij}\left[\left(\frac{R_{ij}}{\sigma_{ij}}\right)^{2\mu_{ij}} - 1\right]}\right\}^{2\nu_{ij}+1} \tag{7}$$

Here, σ_{ij}, ϵ_{ij}, and μ_{ij} are parameters specified for each pair of interacting beads. The exponent μ_{ij} is a positive integer. We used ν_{ij} = 1 and $R_{ij} = 3\sigma_{ij}$. The total pairwise potential E_{pair} was taken as the sum over all pairs of beads within the respective interaction ranges.

For the Mpipi+ model, there are extra terms, including developed angular, dihedral, and CMAP potentials, i.e.

$$E_{\text{Mpipi+}} = E_{\text{bond}} + E_{\text{angular}} + E_{\text{dihedral}} + E_{\text{CMAP}} + E_{\text{elec}} + E_{\text{pair}} \tag{8}$$

Ramachandran Plots for Different Amino Acids from the Customized Coil Database and Simulation Trajectories. We employed the in-built commands within the GROMACS 5.1.2 package[58] to convert the format of simulation trajectories. Then, we utilized PULCHRA[59] to reconstruct coarse-grained simulation trajectories into atomistic structures. The calculation of Ramachandran plots for each amino acid, based on both the customized coil database and simulation data, was executed by using in-house scripts.

Single-Chain Simulations. For structures with experimental validation, we utilized the entries from the PDB (https://www.rcsb.org/) as the initial structures. In instances where experimental structures were unavailable, model structures were generated from the amino acid sequences using Swiss-PdbViewer[60] (http://www.expasy.org/spdbv/). The process of coarse-graining was executed through SMOG2.[61−63] The box size was 1000 Å × 1000 Å × 1000 Å. The single protein chain was centered within the simulation box.

Each simulation was conducted using LAMMPS,[64] maintaining a temperature of 300 K, unless otherwise specified, through a Langevin thermostat[65,66] with a damping coefficient of 100 ps and an integration time step of 10 fs under the periodic boundary condition. Frames were saved at intervals of 1000 steps. For CMAP iteration and Ramachandran plot analyses, an equilibrium run of 1 ns was followed by a production run of 300 ns. CMAP iteration analysis took 200 to 300 ns of each trajectory. Ramachandran plot analysis utilized the final 300 ns of each trajectory. Chemical shift analysis, secondary structure propensity (SSP) analysis, radius of gyration (R_g) calculation, and ISP were performed with a total run of 1101 ns, which consists of an equilibrium run of 1 ns, a test run of 100 ns, and a production run of 1 μs. SSP analysis spanned 500 to 1000 ns for analysis, while R_g and ISP analyses utilized the final 500 ns of each trajectory, saving data at intervals of 500 frames. In direct-coexistence simulations, the initial conformation was derived from a preparation run of 101 ns, preserving the last frame as the starting configuration. Additionally, we conducted the simulation on the testing system, p53 residues 1−93, to investigate the potential manifestation of secondary structure propensity within the region known to adopt a transient secondary structure upon binding to MDM2[67] (PDB ID: 1YCR).

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

Direct-Coexistence Simulations. Direct-coexistence simulations were employed to construct the phase diagrams for different protein systems. The simulations for both the dense phase and the dilute phase were conducted within an elongated simulation box. For each protein, a single-chain simulation was executed to acquire the initial conformation. Subsequently, a specific number of copies of the initial conformation were randomly inserted into the elongated box using in-built commands in the GROMACS 5.1.2 package[58] to generate a multichain dispersed conformation. More detailed settings can be found in Table 3.

Table 3. Direct-Coexistence Simulation Settings for LLPS Systems

protein	length (# amino acid)	# chains	box size (Å^3)
LAF-1 RGG	168	100	150 × 150 × 2800
DDX4	236	100	170 × 170 × 3000
DDX4 cs	236	100	170 × 170 × 3000
DDX4 9FtoA	236	100	170 × 170 × 3000
DDX4 14FtoA	236	100	170 × 170 × 3000
DDX4 24RtoK	236	100	170 × 170 × 3000
FUS	526	100	200 × 200 × 5000
FUS 27R	526	100	200 × 200 × 5000
FUS PLD	163	100	200 × 200 × 5000
FUS PLD 6D	526	100	200 × 200 × 5000
FUS PLD YtoF	526	100	200 × 200 × 5000
FUS RBD RtoG	526	100	200 × 200 × 5000

The multichain dispersed conformation underwent a simulation to obtain the multichain initial slab conformation. This equilibration simulation extended over 30 ns in the NpT ensemble at 200 K, maintained by a Langevin thermostat with a damping coefficient of 100 ps and an integration time step of 10 fs under the periodic boundary condition. Anisotropic pressure coupling was applied along the z-axis during equilibration. The conformation following equilibration was positioned in an elongated box, serving as the multichain initial conformation for production simulations conducted at various temperatures. All production simulations were maintained by a Langevin thermostat with a damping coefficient of 100 ps and an integration time step of 10 fs under periodic boundary condition. Trajectory frames were saved at intervals of 1000 steps. Each simulation commenced at 50 K for the initial 1 ns, followed by a gradual temperature increase of 50 K per nanosecond, with a subsequent stabilization period of 1 ns at each temperature increment until reaching the target temperature. An equilibration run of 101 ns was succeeded by a production run lasting at least 2 μs. Trajectories after 1 μs were utilized for phase diagram analysis.

Phase Diagrams. The simulation box was divided into 200 trunks along the z-axis to determine the average density within each trunk. Employing the built-in k-means clustering function in MATLAB, we subjected all the densities in each simulation to clustering, effectively differentiating between dense and dilute phases and obtaining their respective densities. Critical temperatures were subsequently estimated using the law of coexistence densities

$$(\rho_{\text{dense}}(T) - \rho_{\text{dilute}}(T))^{3.06} = d\left(1 - \frac{T}{T_c}\right) \tag{9}$$

and critical densities were computed by assuming that the law of rectilinear diameters holds, namely

$$\rho_{\text{dense}}(T) + \rho_{\text{dilute}}(T) = \rho_c + 2A(T - T_c) \tag{10}$$

where $\rho_{\text{dense}}(T)$, $\rho_{\text{dilute}}(T)$, and ρ_c are the densities of the dense and dilute phases and the critical density, respectively; T_c is the critical temperature; and d and A are free fitting parameters. We employed the in-built PSO method in MATLAB.[46,68,69] During the optimization process, constraints were imposed to ensure that the fitted coexistence curves exceeded 0, the coexistence curve for the dilute phase exhibited a monotonic increase, and the critical temperature surpassed the specified temperature threshold. The optimization process was grounded in minimizing the summation of Euclidean distances between dense and dilute densities and their corresponding predicted values. The standard error of mean (SEM) for the estimated critical temperatures (Table S8) and critical concentrations (Table S9) was calculated over 3 blocks using the block average method by dividing each production run for analysis into 3 blocks.

Radius of Gyration (R_g) Analysis. We conducted 41 single-chain simulations of chosen IDPs (details in Table 4) sourced from the Small Angle Scattering Biological Data Bank (SASBDB, https://www.sasbdb.org/) under the IDP tag. The selected IDPs were specifically chosen to be free of ligands, rare amino acids, or modifications. Calculations of the R_g were performed using in-house scripts.

Internal Scaling Profile. For the ISP analysis, we utilized the trajectories obtained from the single-chain simulations conducted for R_g analysis. In-house scripts were used to calculate the distance $R^2_{|i-j|} = \langle\langle r^2_{ij}\rangle\rangle_{\text{ens}}$, averaging over all pairs of amino acids separated by $|i-j|$ amino acid(s) from each other and subsequently across all conformations in the ensemble. The apparent scaling exponent ν was estimated by fitting it to the ISP

$$\ln(R_{|i-j|}) = \nu \ln(|i - j|) + A_0 \tag{11}$$

Only residues located at least 15 amino acids away from the N terminal and 5 amino acids away from the C terminal were considered in the analysis.

Chemical Shift Prediction and Secondary Structure Propensity Analysis. We conducted single-chain simulations, chemical shift prediction, and secondary structure propensity (SSP) analysis on 13 IDPs (details in Table 5) from the Biological Magnetic Resonance Data Bank[70,71] (https://bmrb.io/). The selected IDPs were specifically chosen to be free from unusual ligands, rare amino acids, modifications, and conditions that are far from neutrality. Trajectory formats were converted using MDTraj,[72] and chemical shifts of Cα and Cβ were predicted using LARMORCA.[73] Subsequently, SSP analysis utilized SSP[74] to predict the secondary structure propensity scores based on the chemical shifts with RefDB[75] serving as the reference. Mean-squared error (MSE) was calculated to measure the deviation from the experimental value.

Secondary Structure Analysis for p53 from the Free-Energy Landscape View. Selecting the previous successful case of p53, we used the (α, θ) space and the (θ_1, θ_2) space mentioned in the previous section on CMAP development to profile how the refined angular, dihedral, and CMAP potentials modulate the free-energy landscapes of IDPs.

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

Table 4. IDPs for Radius of Gyration (R_g) Analysis and Internal Scaling Profile Analysis

#	SASBDB ID	name	T (K)	R_g (nm)			ν	
				exp	Mpipi	Mpipi+	Mpipi	Mpipi+
1	SASDDF6	myelin basic protein	283	3.3	2.99	2.75	0.406	0.435
2	SASDET6	PQBP-1 XLID m1	283	3.5	2.62	2.39	0.174	0.185
3	SASDED2	hPQBP1	293	3.7	3.07	2.90	0.262	0.301
4	SASDEE2	ER-NTD	283	3.0	2.89	2.70	0.355	0.382
5	SASDEX4	UL11	298	2.4	2.33	2.13	0.399	0.421
6	SASDEF6	EBNA2 T1 381−455	298	2.7	2.21	2.04	0.475	0.501
7	SASDEG6	EBNA2 T2 348−422	298	2.8	2.29	2.07	0.466	0.476
8	SASDEU6	PQBP-1 XLID m2	283	3.6	2.76	2.53	0.262	0.273
9	SASDF34	N-CoRNID	293	4.7	4.43	4.00	0.464	0.475
10	SASDFK8	GON7	288	3.1	2.73	2.43	0.418	0.416
11	SASDHH8	histatin5	298	1.5	1.25	1.06	0.427	0.463
12	SASDJB5	protein P	293	6.2	5.41	4.86	0.440	0.448
13	SASDJ86	Fep1	293	3.5	3.74	3.35	0.462	0.462
14	SASDJJ6	N-FATZ-1	293	3.5	3.55	3.12	0.476	0.468
15	SASDJK6	Δ91-FATZ-1	293	3.9	3.46	3.19	0.431	0.456
16	SASDKD6	NURS red1	293	2.5	2.05	1.82	0.446	0.418
17	SASDKH8	C-Tir	298	3.8	3.24	3.05	0.428	0.453
18	SASDK68	TIF2	283	3.7	3.55	3.20	0.490	0.504
19	SASDKC9	SMAD4	283	4.4	4.26	4.08	0.313	0.344
20	SASDKG9	SMAD4MH2	283	2.1	3.36	3.17	0.350	0.388
21	SASDKF9	SMAD4 SADMH2	283	2.5	3.27	3.11	0.302	0.336
22	SASDKE9	SMAD4 linkerMH2	283	3.3	4.02	3.70	0.362	0.368
23	SASDK29	SMAD2	283	3.8	3.62	3.57	0.264	0.303
24	SASDKT8	VWF 1596−1668	293	3.1	2.48	2.24	0.415	0.429
25	SASDLS4	Tau35	293	4.6	4.76	4.40	0.494	0.517
26	SASDLT4	Tau2N3R	293	6.3	6.25	5.62	0.497	0.498
27	SASDLU4	Tau2N4R	293	6.7	6.58	5.90	0.509	0.498
28	SASDLL5	frataxin	273	2.2	2.84	2.58	0.450	0.473
29	SASDLM5	frataxin	293	2.1	2.93	2.64	0.451	0.478
30	SASDLN5	frataxin	323	2.5	3.06	2.76	0.484	0.498
31	SASDL79	syndecan-2 Ecto	293	4.3	3.46	3.13	0.471	0.488
32	SASDL89	syndecan-3 Ecto	293	6.5	5.84	5.28	0.505	0.501
33	SASDL99	syndecan-4 iso2	293	4.2	3.69	3.33	0.486	0.503
34	SASDL69	syndecan-1 Ecto	293	5.3	4.64	4.20	0.486	0.496
35	SASDLF9	protein W(PNT3)	293	3.4	2.75	2.57	0.371	0.419
36	SASDBY5	FI1050p(Met)	288	5.1	4.20	3.73	0.482	0.492
37	SASDBZ6	draxin	293	4.2	3.90	3.61	0.326	0.335
38	SASDC62	TRF2	277	1.7	1.71	1.53	0.456	0.470
39	SASDC53	ColN-T	277	2.8	2.12	1.95	0.392	0.400
40	SASDCY4	RNase E 603−850	288	5.3	3.44	3.13	0.343	0.351
41	SASDDC2	MAP2c	293	6.7	6.49	5.80	0.485	0.486

Table 5. IDPs for Chemical Shift Prediction and Secondary Structure Propensity Analysis

#	BMRB ID	name	temperature (K)
1	26672	FUS LC WT	298
2	27125	FUS 12E LC	298
3	26719	Ash	278
4	51353	hsTRPV1-IDR	298
5	51354	hsTRPV2-IDR	298
6	51355	hsTRPV3-IDR	298
7	7244	γ-synuclein	278
8	15397	ACTR	304.15
9	15409	T-cell receptor ζ-chain	288
10	19191	eIF4E	278
11	7279	HMGA1	298
12	15131	MBP	277
13	25185	FG-N-6His	298

Angles and dihedral angle values were extracted from the last 1.1 μs of each simulation trajectory and the crystal structure with in-house scripts. The mesh size was 5° × 5° for both the (α, θ) space and the (θ_1, θ_2) space.

Plotting Method. The protein structure figures were rendered using PyMOL,[76] while all other plots were created using MATLAB. In the CMAP figures, we used the colormaps "batlowW" and "vik" from Crameri (https://www.fabiocrameri.ch/colourmaps/) to avoid distortion of data and accommodate readers with color blindness or color vision deficiency.[77,78] For the Ramachandran plot, the built-in colormap "jet" was applied to enhance visual comparison. In the secondary structure analysis from the free-energy landscape view for p53, we used the colormap "davos" from Crameri. In the phase diagram, we applied the colormap "virdis" for DDX4 and its variants and the colormap "romaO" from Crameri for FUS and its variants to avoid distortion of data and

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

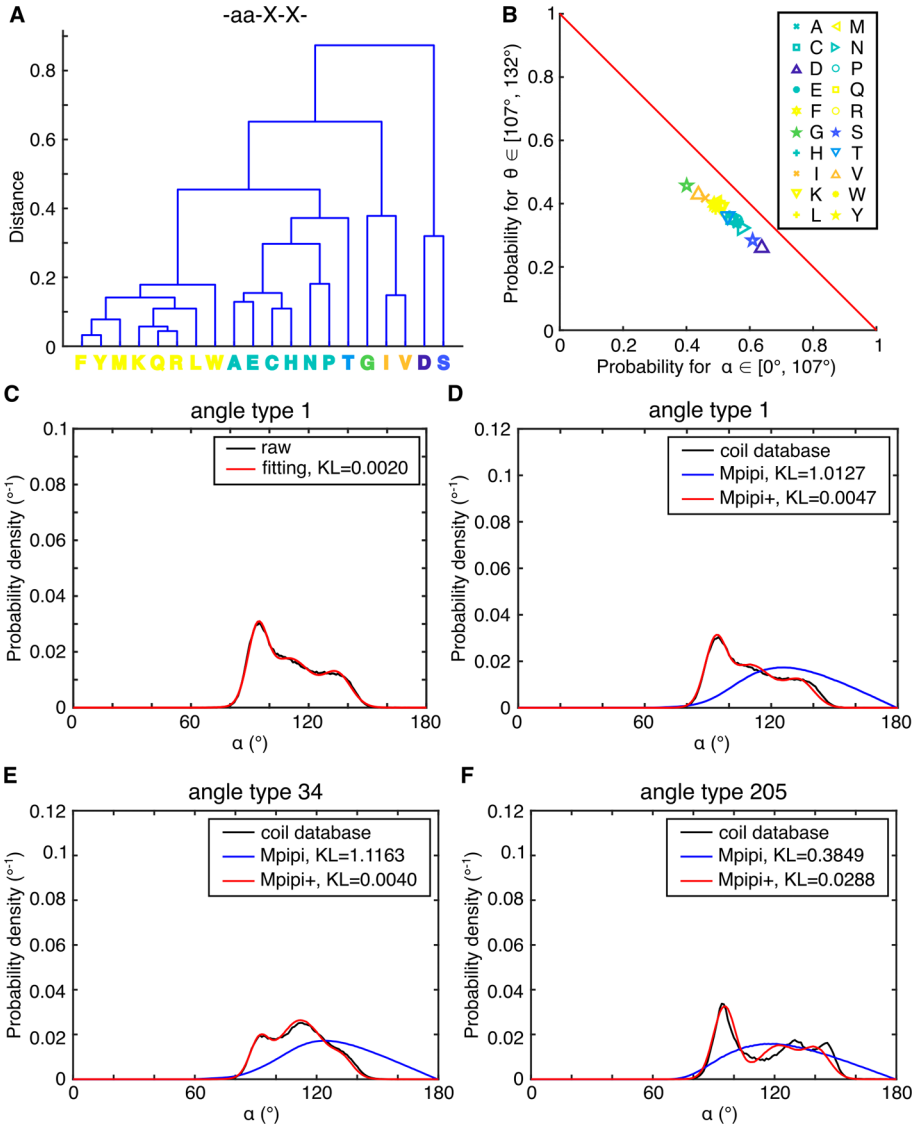


Figure 1. Development of statistical potentials for residue-specific angular terms. (A) Hierarchical clustering of probability distributions for varied amino acids at the first position of an angle. (B) Mapping angles with varied amino acids at the first position onto a 2-dimensional panel, where x represents the probability of an angle in the range of 0−107° and y represents the probability of an angle in the range of 132−180°. (C) Demonstration of the fitting for angle type 1. (D−F) Distributions for angle types 1, 34, and 205 from simulations with the Mpipi model (blue curve) and the Mpipi+ model (red curve) compared with statistics over the customized coarse-grained coil database (black curve).

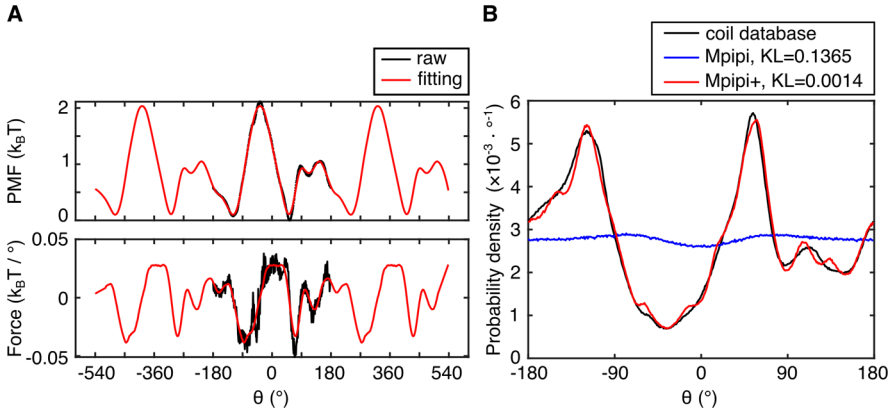


Figure 2. Development of statistical potential for the general dihedral term. (A) Fitting of the potential of mean force (PMF) for three periodic dihedral angles. (B) Distribution for the general dihedral angle from simulations with the Mpipi model (blue curve) and the Mpipi+ model (red curve) compared with statistics over the customized coarse-grained coil database (black curve).

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

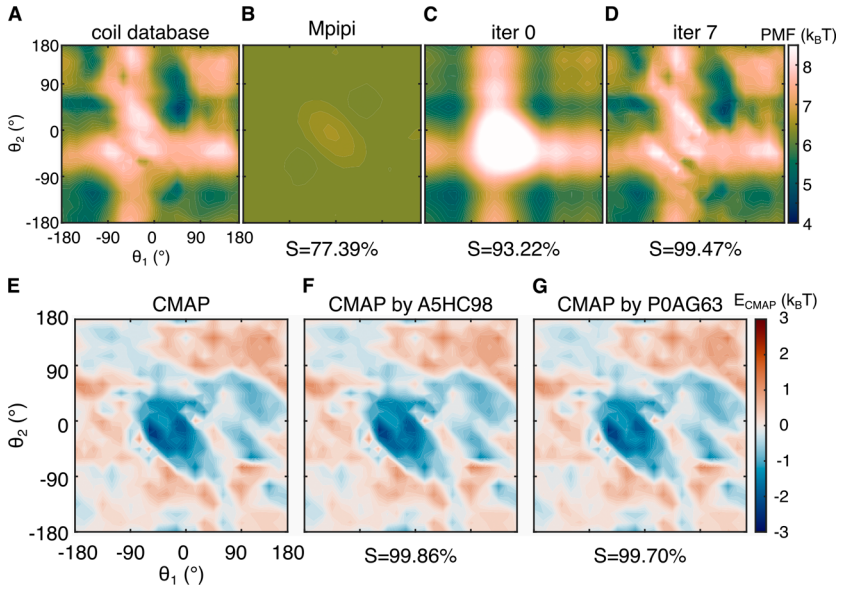


Figure 3. Development of statistical potentials for the general CMAP term using IBI. (A) Potential of mean force (PMF) of (θ_1, θ_2) space extracted from the customized coarse-grained coil database, i.e., the reference. (B) PMF of (θ_1, θ_2) space extracted from the simulations with the Mpipi model. (C) PMF of (θ_1, θ_2) space extracted from the simulations with zero CMAP potential (iteration 0, denoted as "iter 0"). The value S represents the similarity coefficient compared with the reference (θ_1, θ_2) PMF. (D) PMF of (θ_1, θ_2) space extracted from the simulations with the final CMAP potential (iteration 7, denoted as "iter 7"). The value S represents the similarity coefficient compared with the reference (θ_1, θ_2) PMF in (A). (E) Final CMAP potential used in this paper. (F) CMAP potential generated by iterations using the protein A5HC98 only. The value S represents the similarity coefficient compared with the final CMAP potential in (E). (G) CMAP potential generated by iterations using the protein P0AG63 only. The value S represents the similarity coefficient compared with the final CMAP potential in (E).

accommodate readers with color blindness or color-vision deficiency.[77,78] Additionally, the table for angle clusters (Figure S4) used the colormaps "oleron", "bukavu", and "bam" from Crameri, to avoid distortion of data and accommodate readers with color blindness or color-vision deficiency.[77,78] The bud green and pink colors were from https://www.color-hex.com/color-palette/1011103. The color code for all of the special single colors can be found in Table S10.

RESULTS AND DISCUSSION

Angular PMF for IDPs Shows Sequence Dependence. During the development of angular potentials, the amino acid perturbations at the first, second, and third positions of the angles revealed that all three of these positions are of consequence (Figures S2 and S4). The distributions typically exhibited three discernible peaks: one spanning from 75 to 107°, another ranging from 107 to 132°, and the remaining cluster extending from 132 to 160°. We observed that the second position displayed the highest sensitivity to amino acid types, while the third position exhibited exceptional sensitivity to proline. To streamline angle types without losing specificity, we applied hierarchical clustering to amino acid types at each position (Figures 3A and S4A,C). The results of clustering led to the categorization of amino acids at the first position into 7 distinct clusters, while the second and third positions yielded 14 and 3 clusters (Table S6). As anticipated, proline forms a unique cluster at the third position.

To simplify our analysis, we partitioned each distribution into distinct regions corresponding to the three aforementioned peaks (Figures 3B and S4B,D−G). A two-dimensional panel was then utilized to plot each amino acid, with the x-axis representing the probability of an angle ranging from 0 to 107°, the y-axis representing the probability of an angle ranging from 107 to 132°, and the distance from a point to the red line (counter-diagonal) indicating the probability of an angle ranging from 132 to 180°. The results illustrated that each cluster falls within a reasonable range with distinct clusters well separated on each panel.

Therefore, we derived a total of 294 angle types based on the amino acid types in three positions. Employing a mixed Gaussian distribution, we fitted each angular potential, showcasing the fitting result for angle type 1 (Figure 3C). Further details for all angle types are available in the Supporting Information.

An examination of the angular distribution outcomes for different angle types revealed that, unfortunately, 12 angle types were absent in our IDP data set. For the remaining 282 angle types, all exhibited impressive agreement with the reference, showcasing minimal KL divergence. The excellent agreement achieved using a simple mixture Gaussian potential led us to speculate that angular potentials are almost related only to the residue types in the three positions. Consequently, the mixture Gaussian potentials will likely work on the 12 missing angle types, as well. Notably, all available angle types displayed significant improvement in their distributions with our implementation (Table S7). As illustrative examples, angle types 1, 34, and 205 exhibited substantial enhancement with our implementation (Figure 1D−F). Angle type 34 (Figure 1E) emerged as the most improved, attaining a minimal KL divergence of 0.0040 among all available angle types. Angle type 205 (Figure 1F) demonstrated the least improvement, with the largest KL divergence of 0.0288, which still falls within an acceptable range among all available angle types.

Dihedral Potential for IDPs Adopts a General Form. During the development of the dihedral potential, the amino acid perturbations at the first, second, third, and fourth positions of dihedrals exhibited minimal differences (data not shown). This outcome implies that the amino acid type does

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

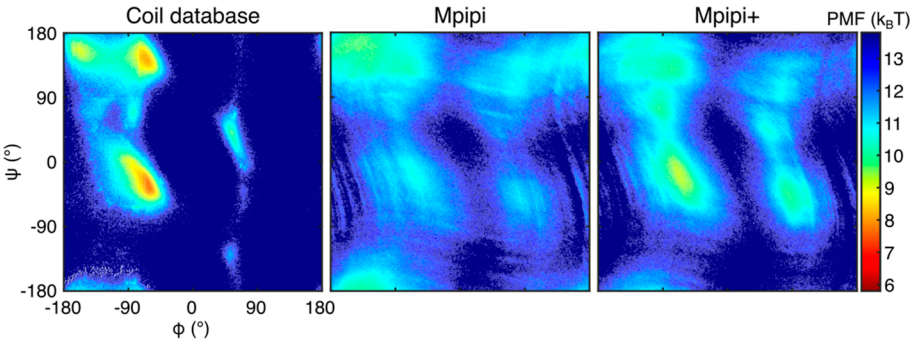


Figure 4. Comparison for Ramachandran plots of alanine. The Ramachandran plot from the customized coil database (left panel) was directly calculated from atomistic structures in the database. Ramachandran plots for the Mpipi model (middle panel) and the Mpipi+ model (right panel) were calculated from atomistic structures reconstructed from coarse-grained structures obtained from simulations.

not significantly influence the dihedral angle distribution at any position of the four. Consequently, we formulated a general dihedral potential in a tabulated form. To mitigate the risk of abnormally large forces arising from an unsmooth or a discontinuous PMF surface, we linked three copies of PMF to account for periodicity and smoothed the PMF (Figure 2A, top panel) while monitoring the force change (Figure 2A, bottom panel). The dihedral distribution extracted from simulations utilizing the Mpipi+ model closely resembled the reference, yielding a minimal KL divergence of 0.0014. In contrast, without our implementation, the KL divergence value surged to 0.1365, and the curve exhibited a nearly flat, evenly distributed pattern (Figure 2B).

Development of the CMAP Potential for IDPs. During the development of the CMAP potential, we initially analyzed the (θ_1, θ_2) free-energy landscape without incorporating developed bonded potentials, specifically using the Mpipi model (Figure 3B). Comparatively, by referencing the PMF of (θ_1, θ_2) extracted from the customized coil coarse-grained database (Figure 3A), we observed a notably flat (θ_1, θ_2) PMF pattern for the Mpipi model with a similarity coefficient of 77.39%. Furthermore, even with the implementation of solely developed angular and dihedral potentials (Table 2, Figure 3C), the (θ_1, θ_2) PMF showed a partial improvement, reflecting a similarity coefficient of 93.22%.

Amino acid perturbation at the first, second, third, fourth, and fifth positions of pentads exhibited minimal differences (data not shown). This outcome suggested that the amino acid type exerts no discriminative influence on the (θ_1, θ_2) distribution. Given that CMAP involves a higher order than dihedral angles, this result aligns with our expectations. Consequently, we formulated a general CMAP potential to capture the (θ_1, θ_2) distribution feature.

After the first iteration, the (θ_1, θ_2) PMF has been largely improved (Table 2, Figure S5, iter 1), achieving an impressive similarity coefficient of 97.43%. Remarkably, excellent convergence was attained within 10 iterations (Table 1, Figures 3D, and S5). The CMAP potential derived from iteration 7 was adopted as the final CMAP potential (Figure 3E). Furthermore, to assess whether the CMAP potentials would differ when employing different sets of IDP data sets, we constructed two extreme situations where only a single protein with 20 amino acid types exists. One is the long IDP, A5HC98, with 315 residues. The other is P0AG63 with a medium length of 80 amino acids. The results demonstrated that the CMAP potentials generated using A5HC98 (Figure 3F) and P0AG63 (Figure 3G) were generally indistinguishable when compared with our final CMAP potential, with similarity coefficients of 99.86 and 99.70%, respectively. This compelling evidence underscores the isolation from other terms and superb convergence across diverse IDP databases of the CMAP potential.

Implementing the Developed Angular, Dihedral, and CMAP Potentials Improves the Ramachandran Plot of Amino Acids for IDPs. We systematically examined the Ramachandran plot for each amino acid within the customized coil database (Figure S6), employing alanine as an illustrative example (Figure 4, left panel). Subsequently, we reconstructed atomistic structures from simulations and profiled the Ramachandran plots for each amino acid (Figures S7 and S8), again using alanine as a representative case (Figure 4, middle and right panels). Notably, the Ramachandran plot constructed from simulations employing the Mpipi model mostly clusters in the β-sheet region, whereas the Mpipi+ model exhibits a more balanced distribution between the α-helix and β-sheet conformations. Our newly developed angular, dihedral, and CMAP potentials exhibit the capacity to enhance the Ramachandran plots for individual amino acids. However, the Ramachandran plots of atomistic structures reconstructed from both the Mpipi model and the Mpipi+ model lack specificity for amino acid types (Figures S7 and S8). The implementation also failed to capture the (φ, ψ) distributions for proline and glycine, which demonstrate distinct landscapes compared to alanine. Furthermore, an unexpected region emerges around (90, −90°) for both models.

The loss of residue specificity in the Ramachandran plot is an anticipated outcome. The φ and ψ distributions characterize the conformation within a residue, and the coarse-graining process inherently discards information about side-chain conformations. When attempting to map a rough model back to a finer model, the challenge lies in reconstructing information stored in dimensions absent from the coarse-grained model. Intriguingly, our implementation within a coarse-grained model exhibits a partial restoration of (φ, ψ) distributions. A plausible explanation is that the implementation procedure facilitates the more accurate placement of atoms in both the side chains and backbones. The observation hints at this idea that the overall conformation of IDPs is a collective outcome influenced by conformations at various levels. It further suggests that the correlation between these levels might be more robust than initially anticipated.

Implementing the Developed Angular, Dihedral, and CMAP Potentials Improves the Secondary Structure Propensity of IDPs. The difference between the deviation of

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

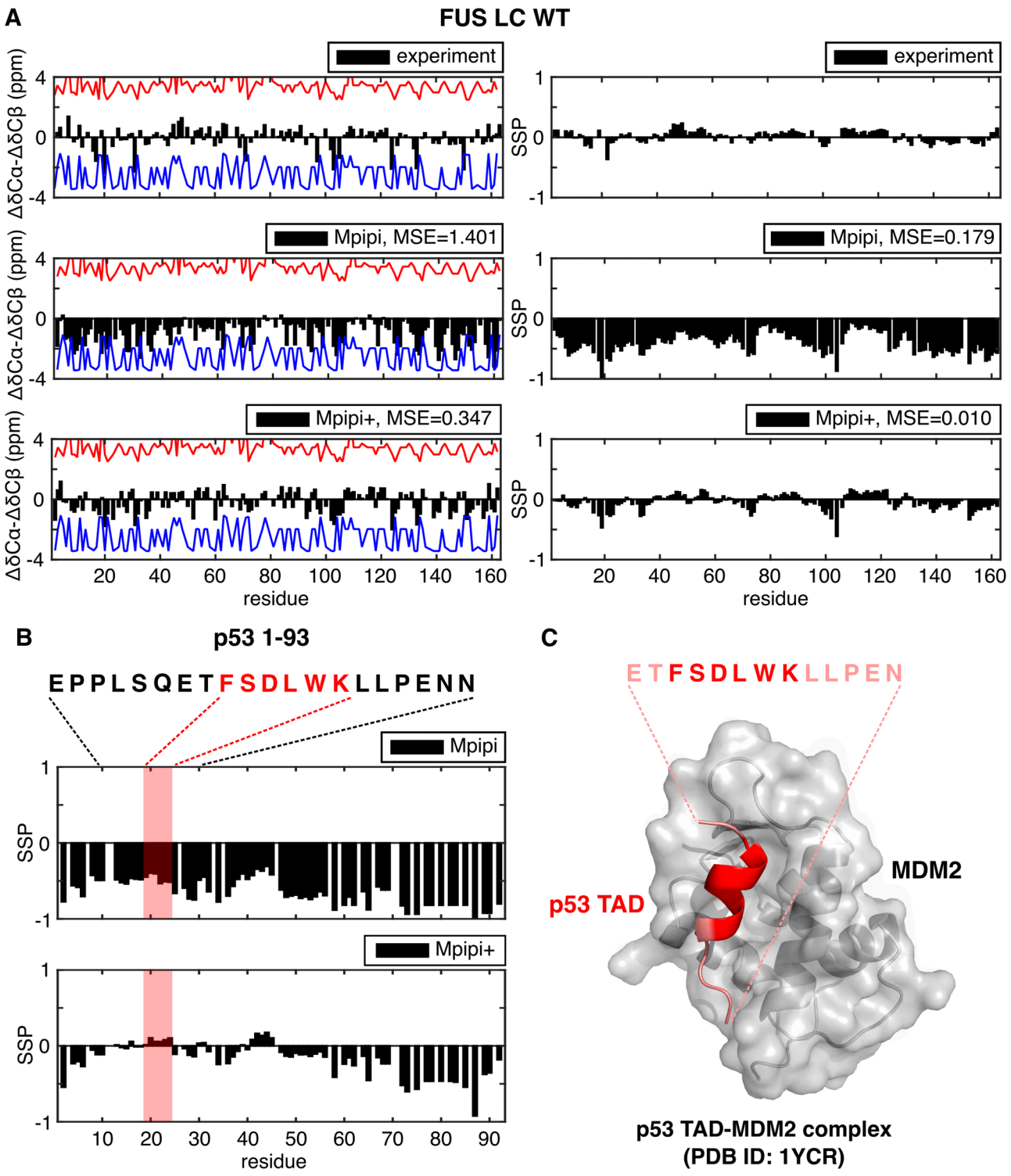


Figure 5. Implementation of developed bonded potentials improves secondary structure propensity (SSP) and captures the transient secondary structure. (A) Differences in Cα and Cβ ^{13}C chemical shift deviations from a random coil reference ($\Delta\delta C\alpha - \Delta\delta C\beta$, left panel) and SSP scores (right panel) of FUS LC WT from the experiment (top panel), simulations with the Mpipi model (middle panel), and simulations with the Mpipi+ model (bottom panel). The red curves in the left panel represent the reference value for α-helix, while the blue curves in the left panel represent the reference value for β-sheet. (B) SSP scores of p53 residues 1−93 truncation from simulations with the Mpipi model (top panel) and the Mpipi+ model (bottom panel). The red shades imply one region (residues 19−24) with α-helical propensity. The sequence for residues 11−30 was displayed, and the corresponding region with α-helical propensity was marked as red. (C) Crystal structure (PDB ID: 1YCR) of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain (TAD). MDM2 is displayed in gray, while the p53 TAD is displayed in salmon red. In the structure, part of the p53 TAD peptide was observed to adopt a transient helical structure. The region predicted to have α-helical propensity in simulation with the Mpipi+ model is displayed in red. The sequence for the p53 TAD peptide is displayed, except for two missing residues (SQ) in the N-terminus. The region with α-helical propensity is highlighted in red.

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

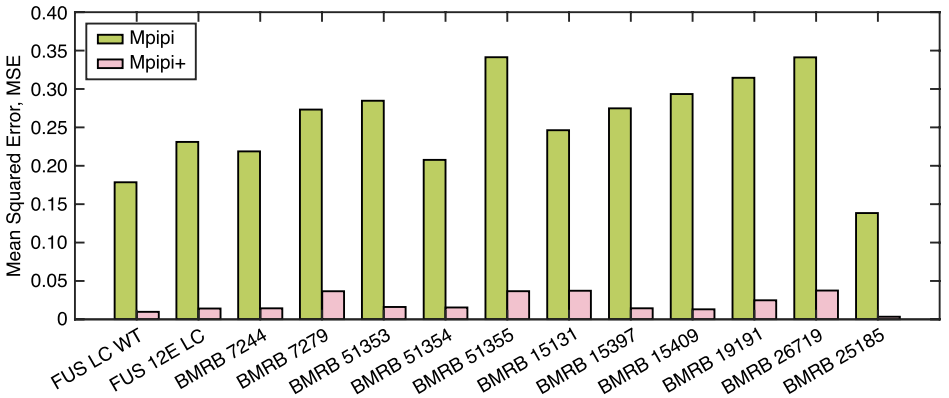


Figure 6. Mean-squared error (MSE) of secondary structure propensity (SSP) values derived from simulations with the Mpipi model (bud green) and the Mpipi+ model (pink) compared with experiments for 13 IDPs.

the chemical shift of Cα and Cβ from RefDB, a reference chemical shift database for IDPs, denoted as $\Delta\delta C\alpha - \Delta\delta C\beta$, along with the secondary structure propensity (SSP) based on chemical shifts and RefDB, serves as an effective means to scrutinize the conformational preferences of the protein backbone. Illustrated in Figure 5A, taking the protein FUS LC WT as an example, both $\Delta\delta C\alpha - \Delta\delta C\beta$ and SSP scores closely approach zero, indicating disordered structures, for experiments (top panel) and the Mpipi+ model (bottom panel). In contrast, for the Mpipi model (mid panel), these two metrics are mostly negative, implying a preference for β structures. This pattern holds for the remaining 12 IDPs examined in our study, and the Mpipi+ model can largely reduce the deviation from experiments (Figures 6 and S9).

The disparities in conformational preferences between simulations conducted with the Mpipi model and the Mpipi + model align with the insights from the Ramachandran plots. The Mpipi+ model demonstrated superb predictive accuracy regarding the conformational preferences. These nuanced "preferences" might be the cuddle of transient structures, intricately contributing to the orchestration of ubiquitous vital cellular processes.[13] These "locally transiently ordered" fragments nourish SLiMs, elucidating the transient interactions and multifaceted functionalities inherent in IDP/Rs. Consequently, the inherent preferences encoded in the sequence are indispensable for unraveling the mechanisms of the IDP/R functions.

To investigate how well our SSP analysis aligns with the experimental reality, we focused on the well-established p53−MDM2 system. In this system, MDM2, a ubiquitin ligase, orchestrates ubiquitin-mediated protein degradation, while p53, renowned as a tumor-suppressor protein, acts as a defender combatting the overgrowth of the cell, with its N-terminal existing as an IDR.[67] As depicted in Figure 5C, the binding motif within TAD in the flexible N-terminal of p53 adopts a helical structure when forming a complex with MDM2,[67] indicating a transient helical structure. We truncated the 1−93 region of p53 and examined its SSP as a single chain. Interestingly, the SSP analysis revealed continuous positive values in the region 19−24 for simulation with the Mpipi+ model (Figure 5B), suggesting a preference for the α-helical structure. Remarkably, this aligns with the α-helical structure observed in the p53 TAD−MDM2 complex. This special case study underscores the meaningfulness of the SSP values introduced by our implementation in capturing biologically relevant structural preferences.

To obtain comprehensive understanding of the effects of developed angular, dihedral, and CMAP potentials on the secondary structure, we analyzed the free-energy landscapes of (α, θ) space for the FSDL motif (Figure 7A,B) and (θ_1, θ_2)

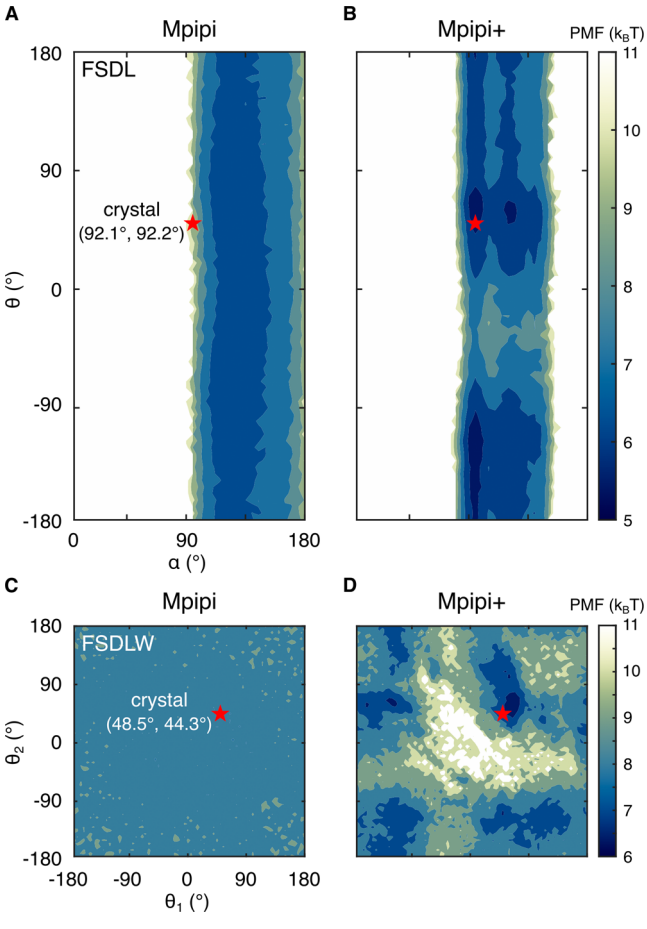


Figure 7. (A,B) Free-energy landscapes in (α, θ) space for the FSDL motif from simulation trajectories with the Mpipi model (A) and Mpipi+ model (B). The solid red star represents the FSDL motif in the crystal structure. PMF means potential of mean forces. Free-energy values higher than 11 k_BT are in white color. (C,D) Free-energy landscapes in (θ_1, θ_2) space for the FSDLW motif from simulation trajectories with the Mpipi model (C) and Mpipi+ model (D). The solid red star represents the FSDLW motif in the crystal structure.

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

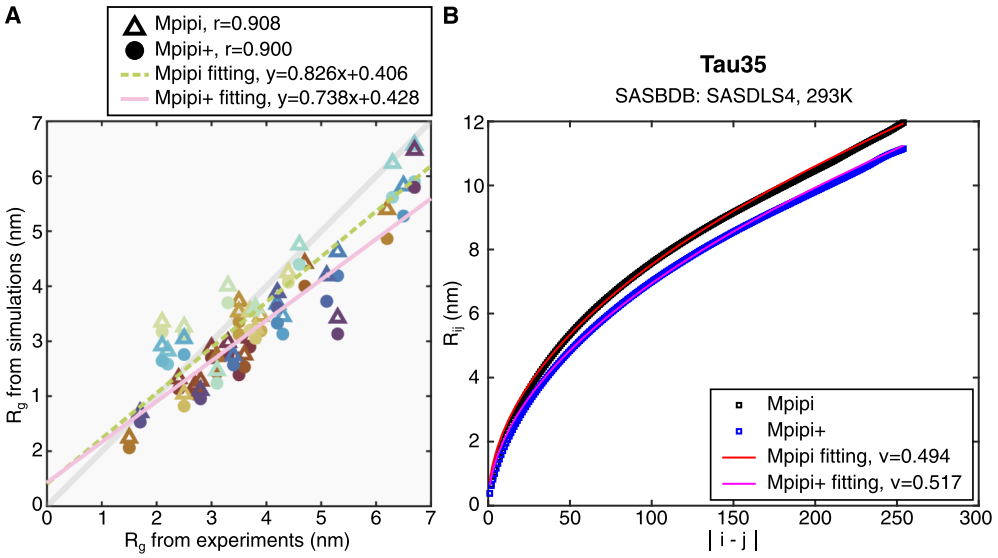


Figure 8. Impact of the implementation of developed bonded potentials on the radius of gyration (R_g) and ISP. (A) Correlation of R_g for 41 simulations of IDPs in SASBDB using the Mpipi model (hollow triangles) and the Mpipi+ model (solid circles). A pair of a hollow triangle and a solid circle in the same color represent simulations for the same IDP. Pearson's coefficient (r) between simulation and experimental values is indicated. The dashed bud green line represents the linear fitting curve for simulations with the Mpipi model, and the solid pink line represents the linear fitting curve for simulations with the Mpipi+ model. (B) ISP for Tau35 (SASBDB: SASDL4) at 293 K. Hollow black squares represent values from simulations with the Mpipi model, and hollow blue squares represent values from simulations with the Mpipi+ model. The red curve is the fitting curve for values from simulations with the Mpipi model, while the magenta curve is the fitting curve for values from simulations with the Mpipi+ model.

space for the FSDLW motif (Figure 7C,D). Here, one additional residue W is included for (θ_1, θ_2) because this CG coordinate pair involves five residues [four residues for (α, θ)]. As shown in Figure 7A, for the (α, θ) space, the free-energy landscape of the Mpipi+ model contains an evident basin around (90°, 50°), a characteristic hallmark of α-helix in (α, θ) space.[55] The crystal structure also falls in the basin corresponding to the α-helix. On the contrary, the crystal structure is located at a high free-energy area in the free-energy landscape of the Mpipi model (Figure 7B), suggesting little chance of being sampled in simulation. The analysis of the (θ_1, θ_2) space reached similar conclusions. The (θ_1, θ_2) free-energy landscape of the Mpipi model is relatively flat (Figure 7C), while, as depicted in Figure 7D, that of the Mpipi+ model forms a deep basin around (50°, 50°), representing α-helix in (θ_1, θ_2) space.[55] Similarly, the point representing the crystal structure of the FSDLW motif lies in the basin.

All these results suggest that the refined angular, dihedral, and CMAP potentials shape basin biasing certain secondary structures in the free-energy landscape, which consequently improves the secondary structure propensities of IDPs.

Implementing the Developed Angular, Dihedral, and CMAP Potentials Has Minor Effects on the Overall Shape and Flexibility of IDPs. R_g serves as a widely used metric to characterize the overall shape of proteins. As depicted in Figure 8A, both the Mpipi model and the Mpipi + model exhibit a correlation coefficient exceeding 0.9 when compared to the experimental data in terms of R_g. The linear fitting slopes for the Mpipi model and the Mpipi+ model are 0.826 and 0.738, respectively, suggesting a shared tendency of underestimating the R_g for both models.

ISP provides insights into the polymer behavior of protein chains. Specifically, with regard to ISP, the incorporation of the developed angular, dihedral, and CMAP potentials resulted in a subtle reduction in the distance between two residues equidistant from each other (Figure S10), taking Tau35 as an example (Figure 8B).

These findings indicate that the conformations generated by the Mpipi model tend to be more expanded on average. This can be explained by the differences in the angular distributions of the two models. For the Mpipi model, which is free of any angular potential, the distribution is broader and centers around 120°. In contrast, in the Mpipi+ model, the implemented angles concentrate within the range of 80 to 120°. This resulted in a smaller average value over angles in the Mpipi+ model and, consequently, less-expanded conformations.

Implementing the Developed Angular, Dihedral, and CMAP Potentials Has a Subtle Effect on the Relative LLPS Propensity of IDPs. We conducted direct-coexistence simulations to characterize the phase diagrams for three protein families: LAF-1 RGG (Figure 9, top panel), DDX4 and its variants (Figure 9, middle panel), and FUS and its variants (Figure 9, bottom panel). Critical temperatures obtained from simulations with the Mpipi model and the Mpipi+ model exhibited negligible differences. However, it is noteworthy that the implementation induced an increase in density for dense phases across all three families, resulting in a corresponding elevation in the estimated critical density. The variation in density is explicable by distinct degrees of expansion of the conformations of IDPs generated by the two models. The Mpipi model tends to generate expanded structures. Instead, the Mpipi+ model yields relatively compact configurations, contributing to higher density in dense phases. Despite the subtle impact on critical density, this implementation did not perturb the ranking of LLPS propensity across three protein families.

The Mpipi model demonstrates exceptional qualitative alignment with the LLPS propensities of the examined systems.[32] Notably, the implementation of developed angular,

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

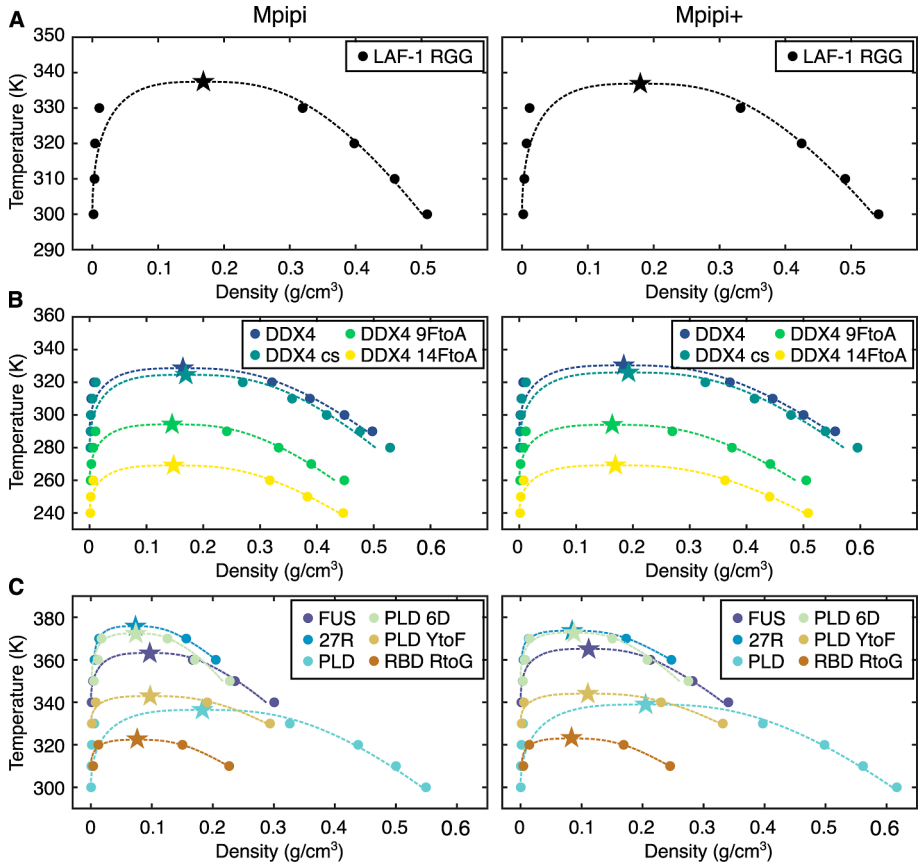


Figure 9. Phase diagram for three LLPS systems. (A) LAF-1 RGG domain; (B) DDX4 (Chinese blue) and three variants: charge-scrambled mutant (DDX4 cs, celadon green), 9 phenylalanine mutated to the alanine mutant (9FtoA, emerald), and 14 phenylalanine mutated to the alanine mutant (14FtoA, yellow sun); (C) FUS (UCLA blue) and 5 variants: FUS mutant with insertions and residues mutated to arginine (27R, silver lake blue), FUS PLD domain (PLD, middle blue green), FUS mutant with 6 residues mutated to asparagine in PLD domain (PLD 6D, tea green), FUS with insertions and tyrosine mutated to phenylalanine in the PLD domain (PLD YtoF, dark khaki), and FUS with insertions and arginine mutated to glycine in the RBD domain (RBD RtoG, dark gold). Solid circles mark the concentrations for dilute and dense phases estimated from the direct-coexistence simulations at the corresponding temperature. Dash lines represent the estimated coexistence curves. Stars indicate the estimated critical points. The left panels are from simulations with the Mpipi model, and the right panels are from simulations with the Mpipi+ model. For all the simulations, the SEM for estimated critical temperatures and critical concentrations are smaller than 1.3 K and 0.004 g/cm^3, respectively (error bar not shown).

dihedral, and CMAP potentials did not alter the ordering of different IDPs, thereby preserving the perfect ranking of the LLPS propensity.

Nevertheless, some discrepancies with the experimental values were observed. The predicted protein concentrations in the dense phase by both models deviated significantly from the experimental values. In the case of the LAF-1 RGG domain, experimental findings suggested an astonishingly low protein concentration in the dense phase of approximately 7.32 mg/mL in 125 mM NaCl.[79] However, the estimated value in this study was on the order of 100 mg/mL, almost 2 orders of magnitude higher than those observed in experiments.

Temperature is another issue. Taking DDX4 and its variant system as an example, the estimated critical temperatures deviate significantly from the experimental values. In experiments, DDX4, at a concentration of around 300 to 400 mg/mL in 100 mM NaCl, underwent LLPS at 343 K (70 °C).[80] The estimated critical temperatures for 100 and 200 mM NaCl are around 393 K (120 °C) and 368 K (95 °C), respectively. However, the estimated critical temperature is 330 K (57 °C), at least 40 K lower than the real value. Similar disparities were observed in the case of the DDX4 charge-scrambled (DDX4 cs) system. The experimental critical temperature was around 313 K (40 °C),[80] while the value in our simulations approached 325 K (52 °C), more than 10 K higher. Also, experimentally, the critical temperatures for DDX4 and DDX4 cs differed around 45 K. In simulations, however, the difference is only 5 K. This suggests that our model is insensitive to the presence of scrambled charges.

Furthermore, for the DDX4 14FtoA variant, experimental evidence indicated demixing at 278 K (5 °C) while still undergoing LLPS on ice, namely, 273 K (0 °C).[80] However, our simulations estimated the critical temperature to be around 269 K (−4 °C), slightly lower than 273 K. Limited experimental data for DDX4 9FtoA and DDX4 24RtoK only indicated that they could not undergo LLPS at temperatures above 303 K (30 °C) and 273 K (0 °C), respectively.[80] Our simulations predicted a critical temperature of around 294 K (21 °C) for DDX4 9FtoA and a demixed phase above 260 K (−13 °C) for DDX4 24RtoK. These results align with the observed phenomena for these two mutants. Inadequate experimental data for DDX4 9FtoA and DDX4 24RtoK prevent a quantitative comparison with our simulations.

This discrepancy of temperature was reported in previous studies and may arise from the absence of anisotropic characteristics.[32] The much smaller difference, about 5 K, in

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

estimated critical temperature between DDX4 and DDX4 cs also suggested that our model is insensitive to the scrambled charges.

The discrepancies in density and temperature underscore a critical limitation in our current coarse-grained modeling approach. Another issue is related to the system size. While adept at simulating hundreds of intermediate-length IDPs through a one-bead-per-residue paradigm, it struggles with larger systems as the real condensates. A condensate with a protein concentration of 2 mM (taking 2017 μM for FUS-EGFP[81]) and a size of 1 μm^3 (radius around 0.62 μm) contains millions of chains. Even if we take the maximal size of our model as 1000 chains, the real number is 3 orders of magnitude higher. To overcome this challenge, a refined coarse-grained model tailored explicitly for droplet simulations is essential. It should strike a delicate balance between resolution and accuracy, demanding collaborative efforts between experimentalists and computational scientists. Given the sparse data available on IDP/Rs, this endeavor calls for an ambitious pursuit, requiring innovative methodologies and an interdisciplinary approach to revolutionizing our understanding of these complex systems.

CONCLUSIONS

Though they appear to be "random", IDP/Rs are far from simple polymers connected by bonds. The exclusion of three commonly overlooked bonded potentials—angular, dihedral, and CMAP—reveals intricate and unique profiles. We suggest that researchers take great caution when considering the omission of angular, dihedral, and CMAP potentials. Our study represents a significant leap forward by successfully developing residue-specific angular potentials, general dihedral potentials, and general CMAP potentials specifically for IDP/Rs. These bonded terms, despite their complexity, show isolation from each other and excellent convergence.

Remarkably, our Mpipi+ model exhibits an intriguing capability to partially restore the atomistic conformation within a single residue, even in the absence of such detailed dimensions in a coarse-grained model. The most pivotal finding lies in the discovery of the SSP encoded in the sequence. Our Mpipi+ model successfully reconstructed the SSP comparable to experiments for IDPs. In our case study of p53 residues 1−93, the region recognized to adopt a transient helical structure in experiments was found to align with the SSP in simulations with the Mpipi+ model. This implies that the preference for a secondary structure hidden in disordered regions can materialize under specific conditions.

Additionally, it appears that our implementation exerts a minimal influence on other critical properties of IDPs, including the overall size, flexibility, and LLPS propensity. Collectively, the Mpipi+ model emerges as a straightforward and effective solution to rectify the conformations of IDPs, presenting a promising avenue without any evident side effects.

ASSOCIATED CONTENT

Data Availability Statement

The data underlying this study are available in the published article and its Supporting Information. A demo can be found at https://github.com/ZixinHu-Apple-Manzana/Mpipi_plus_model_2024. Any further data are available from the corresponding author upon reasonable request.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.4c02823.

Amino acid perturbation at the first, second, and third positions in an angle; clustering information for amino acid perturbation at three positions in an angle; (θ_1, θ_2) PMF during the CMAP iterations; Ramachandran plots for different amino acids from the customized coil database, the Mpipi model, and the Mpipi+ model; SSP for the remaining 12 IDPs; ISP for the remaining 40 IDPs; sequence information; angle-type information; KL divergence values for distribution of 294 angle types from simulations (with the Mpipi model and Mpipi+ model) compared with the corresponding statistical distribution extracted from the customized coarse-grained coil database; SEM for estimated critical temperatures and critical concentrations from the phase diagrams for IDPs; and the color palette for special colors (PDF)

AUTHOR INFORMATION

Corresponding Author

Lanyuan Lu − *School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore;* orcid.org/0000-0003-4808-2431; Phone: +65 6316 2866; Email: lylu@ntu.edu.sg

Authors

Zixin Hu − *School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore*

Tiedong Sun − *School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore*

Wenwen Chen − *UHL no. 05-01, Tan Chin Tuan Wing, Office of the President, University Hall, National University of Singapore, Singapore 119077, Singapore*

Lars Nordenskiöld − *School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore;* orcid.org/0000-0002-3681-209X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.4c02823

Author Contributions

Zixin Hu performed all the simulations and analyses in this work. Tiedong Sun contributed to the troubleshooting and provided valuable suggestions on the analysis. Wenwen Chen's early exploration of bonded potentials in folded proteins served as inspiration for this research. The manuscript was drafted by Zixin Hu, Lars Nordenskiöld, and Lanyuan Lu, with all authors approving the final version. Supervision for this study was provided by Lars Nordenskiöld and Lanyuan Lu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore (https://www.nscc.sg). This study was supported by the Singapore Ministry of Education (MOE) Tier 3 (MOE-2019-T3-1-012). The writing used ChatGPT and Grammarly to improve the wording. The content of the final version was reviewed and edited by the authors.

■ ABBREVIATIONS

IDP/Rs, intrinsically disordered proteins/regions; CG, coarse-grained; MD simulation, molecular dynamics simulation; SLiM, short linear motif; SSP, secondary structure propensity; LLPS, liquid−liquid phase separation; CMAP, correction MAP; LC domain, low-complexity domain; TCR, T-cell receptor; SIV, simian immunodeficiency virus; SUSP4, SUMO-specific protease; SAXS, small-angle X-ray scattering; R_g, radius of gyration; smFRET, single-molecular Föster resonance energy transfer; NMR, nuclear magnetic resonance; KH model, Kim−Hummer model; PDB, protein data bank; MJ potential, Miyazawa−Jernigan potential; HPS, hydrophobicity scale; HPS-KR, HPS-Kapcha−Rossky; PRE, paramagnetic resonance enhancements; BI, Boltzmann inversion; PMF, potential of mean force; MDM2, mouse double minute 2 homologue; ISP, internal scaling profile; IBI, iterative Boltzmann inversion; WF potential, Wang−Frenkel potential; CTD, C-terminal domain; LAF-1, Clr6-associated factor 1; DDX4, DEAD-box helicase 4; FUS, fused in sarcoma; PLD, prior-like domain; RBD, RNA-binding domain; PSO, particle-swarm optimization; KL divergence, Kullback−Leibler divergence; SASBDB, Small-Angle Scattering Biological Data Bank; BMRB, Biological Magnetic Resonance Data Bank; MSE, mean-squared error; SEM, standard error of mean; WT, wild type

■ REFERENCES

(1) Das, R. K.; Huang, Y.; Phillips, A. H.; Kriwacki, R. W.; Pappu, R. V. Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 5616−5621.

(2) Takeda, S.; Koike, R.; Nitanai, Y.; Minakata, S.; Maéda, Y.; Ota, M. Actin capping protein and its inhibitor CARMIL: how intrinsically disordered regions function. *Phys. Biol.* **2011**, *8*, 035005.

(3) Das, R. K.; Crick, S. L.; Pappu, R. V. N-Terminal Segments Modulate the α-Helical Propensities of the Intrinsically Disordered Basic Regions of bZIP Proteins. *J. Mol. Biol.* **2012**, *416*, 287−299.

(4) Bhattacharyya, R. P.; Reményi, A.; Good, M. C.; Bashor, C. J.; Falick, A. M.; Lim, W. A. The Ste5 Scaffold Allosterically Modulates Signaling Output of the Yeast Mating Pathway. *Science* **2006**, *311*, 822−826.

(5) Alberti, S.; Gladfelter, A.; Mittag, T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell* **2019**, *176*, 419−434.

(6) Shin, Y.; Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **2017**, *357*, No. eaaf4382.

(7) Yamazaki, T.; Yamamoto, T.; Yoshino, H.; Souquere, S.; Nakagawa, S.; Pierron, G.; Hirose, T. Paraspeckles are constructed as block copolymer micelles. *EMBO J.* **2021**, *40*, No. e107270.

(8) Yamazaki, T.; Souquere, S.; Chujo, T.; Kobelke, S.; Chong, Y. S.; Fox, A. H.; Bond, C. S.; Nakagawa, S.; Pierron, G.; Hirose, T. Functional Domains of NEAT1 Architectural lncRNA Induce Paraspeckle Assembly through Phase Separation. *Mol. Cell* **2018**, *70*, 1038−1053.e7.

(9) Yamazaki, T.; Hirose, T. Control of condensates dictates nucleolar architecture. *Science* **2021**, *373*, 486−487.

(10) Wu, M.; Xu, G.; Han, C.; Luan, P.-F.; Xing, Y.-H.; Nan, F.; Yang, L.-Z.; Huang, Y.; Yang, Z.-H.; Shan, L.; et al. lncRNA SLERT controls phase separation of FC/DFCs to facilitate Pol I transcription. *Science* **2021**, *373*, 547−555.

(11) Nosella, M. L.; Tereshchenko, M.; Pritišanac, I.; Chong, P. A.; Toretsky, J. A.; Lee, H. O.; Forman-Kay, J. D. O-Linked-N-Acetylglucosaminylation of the RNA-Binding Protein EWS N-Terminal Low Complexity Region Reduces Phase Separation and Enhances Condensate Dynamics. *J. Am. Chem. Soc.* **2021**, *143*, 11520−11534.

(12) Brangwynne, C. P.; Eckmann, C. R.; Courson, D. S.; Rybarska, A.; Hoege, C.; Gharakhani, J.; Julicher, F.; Hyman, A. A. Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. *Science* **2009**, *324*, 1729−1732.

(13) Van Roey, K.; Uyar, B.; Weatheritt, R. J.; Dinkel, H.; Seiler, M.; Budd, A.; Gibson, T. J.; Davey, N. E. Short Linear Motifs: Ubiquitous and Functionally Diverse Protein Interaction Modules Directing Cell Regulation. *Chem. Rev.* **2014**, *114*, 6733−6778.

(14) Kumar, M.; Michael, S.; Alvarado-Valverde, J.; Mészáros, B.; Sámano-Sánchez, H.; Zeke, A.; Dobson, L.; Lazar, T.; Örd, M.; Nagpal, A.; et al. The Eukaryotic Linear Motif resource: 2022 release. *Nucleic Acids Res.* **2022**, *50*, D497−D508.

(15) Schaefer, T. M.; Bell, I.; Fallert, B. A.; Reinhart, T. A. The T-Cell Receptor ζ Chain Contains Two Homologous Domains with Which Simian Immunodeficiency Virus Nef Interacts and Mediates Down-Modulation. *J. Virol.* **2000**, *74*, 3273−3283.

(16) Sigalov, A. B.; Zhuravleva, A. V.; Orekhov, V. Y. Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form. *Biochimie* **2007**, *89*, 419−421.

(17) Nováček, J.; Janda, L.; Dopitová, R.; Žídek, L.; Sklenář, V. Efficient protocol for backbone and side-chain assignments of large, intrinsically disordered proteins: transient secondary structure analysis of 49.2 kDa microtubule associated protein 2c. *J. Biomol. NMR* **2013**, *56*, 291−301.

(18) Laptenko, O.; Prives, C. Transcriptional regulation by p53: one protein, many possibilities. *Cell Death Differ.* **2006**, *13*, 951−961.

(19) Mittag, T.; Orlicky, S.; Choy, W.-Y.; Tang, X.; Lin, H.; Sicheri, F.; Kay, L. E.; Tyers, M.; Forman-Kay, J. D. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 17772−17777.

(20) Kim, D.-H.; Han, K.-H. Transient Secondary Structures as General Target-Binding Motifs in Intrinsically Disordered Proteins. *Int. J. Mol. Sci.* **2018**, *19*, 3614.

(21) Kim, D.-H.; Lee, C.; Lee, S.-H.; Kim, K.-T.; Han, J. J.; Cha, E.-J.; Lim, J.-E.; Cho, Y.-J.; Hong, S.-H.; Han, K.-H. The Mechanism of p53 Rescue by SUSP4. *Angew. Chem., Int. Ed.* **2017**, *56*, 1278−1282.

(22) Zheng, W. W.; Best, R. B. An Extended Guinier Analysis for Intrinsically Disordered Proteins. *J. Mol. Biol.* **2018**, *430*, 2540−2553.

(23) Gomes, G.-N. W.; Krzeminski, M.; Namini, A.; Martin, E. W.; Mittag, T.; Head-Gordon, T.; Forman-Kay, J. D.; Gradinaru, C. C. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J. Am. Chem. Soc.* **2020**, *142*, 15697−15710.

(24) Monahan, Z.; Ryan, V. H.; Janke, A. M.; Burke, K. A.; Rhoads, S. N.; Zerze, G. H.; O'meally, R.; Dignon, G. L.; Conicella, A. E.; Zheng, W. W.; et al. Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *EMBO J.* **2017**, *36*, 2951−2967.

(25) Liu, H.; Song, D.; Lu, H.; Luo, R.; Chen, H. F. Intrinsically disordered protein-specific force field CHARMM36IDPSFF. *Chem. Biol. Drug Des.* **2018**, *92*, 1722−1735.

(26) Song, D.; Wang, W.; Ye, W.; Ji, D. J.; Luo, R.; Chen, H. F. ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins. *Chem. Biol. Drug Des.* **2017**, *89*, 5−15.

(27) Yang, S.; Liu, H.; Zhang, Y. P.; Lu, H.; Chen, H. F. Residue-Specific Force Field Improving the Sample of Intrinsically Disordered Proteins and Folded Proteins. *J. Chem. Inf. Model.* **2019**, *59*, 4793−4805.

(28) Vitalis, A.; Pappu, R. V. ABSINTH: A New Continuum Solvation Model for Simulations of Polypeptides in Aqueous Solutions. *J. Comput. Chem.* **2009**, *30*, 673−699.

(29) Choi, J. M.; Pappu, R. V. Improvements to the ABSINTH Force Field for Proteins Based on Experimentally Derived Amino Acid Specific Backbone Conformational Statistics. *J. Chem. Theory Comput.* **2019**, *15*, 1367−1382.

(30) Dignon, G. L.; Zheng, W. W.; Kim, Y. C.; Best, R. B.; Mittal, J. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Comput. Biol.* **2018**, *14*, No. e1005941.

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508

(31) Tesei, G.; Schulze, T. K.; Crehuet, R.; Lindorff-Larsen, K. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2111696118.
(32) Joseph, J. A.; Reinhardt, A.; Aguirre, A.; Chew, P. Y.; Russell, K. O.; Espinosa, J. R.; Garaizar, A.; Collepardo-Guevara, R. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nat. Comput. Sci.* **2021**, *1*, 732−743.
(33) Dignon, G. L.; Zheng, W. W.; Mittal, J. Simulation methods for liquid-liquid phase separation of disordered proteins. *Curr. Opin. Chem. Eng.* **2019**, *23*, 92−98.
(34) Benayad, Z.; Von Bulow, S.; Stelzl, L. S.; Hummer, G. Simulation of FUS Protein Condensates with an Adapted Coarse-Grained Model. *J. Chem. Theory Comput.* **2021**, *17*, 525−537.
(35) Garaizar, A., Espinosa, J. R., Joseph, J. A., Krainer, G., Shen, Y., Knowles, T. P. J., Collepardo-Guevara, R. Intermolecular reorganisation of single-component condensates during ageing promotes multiphase architectures. **2021**, bioRxiv:10.09.463670. bioRxiv.
(36) Kim, Y. C.; Hummer, G. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *J. Mol. Biol.* **2008**, *375*, 1416−1433.
(37) Miyazawa, S.; Jernigan, R. L. Estimation of Effective Interresidue Contact Energies From Protein Crystal Structures: Quasi-chemical Approximation. *Macromolecules* **1985**, *18*, 534−552.
(38) Miyazawa, S.; Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **1996**, *256*, 623−644.
(39) Kapcha, L. H.; Rossky, P. J. A Simple Atomic-Level Hydrophobicity Scale Reveals Protein Interfacial Structure. *J. Mol. Biol.* **2014**, *426*, 484−498.
(40) Urry, D. W.; Gowda, D. C.; Parker, T. M.; Luan, C. H.; Reid, M. C.; Harris, C. M.; Pattanaik, A.; Harris, R. D. Hydrophobicity scale for proteins based on inverse temperature transitions. *Biopolymers* **1992**, *32*, 1243−1250.
(41) Regy, R. M.; Thompson, J.; Kim, Y. C.; Mittal, J. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Sci.* **2021**, *30*, 1371−1379.
(42) Das, S.; Lin, Y. H.; Vernon, R. M.; Forman-Kay, J. D.; Chan, H. S. Comparative roles of charge, pi, and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc. Natl. Acad. Sci. U.S.A.* **2020**, *117*, 28795−28805.
(43) Mackerell, A. D.; Feig, M.; Brooks, C. L. Improved treatment of the protein backbone in empirical force fields. *J. Am. Chem. Soc.* **2004**, *126*, 698−699.
(44) Fitzkee, N. C.; Fleming, P. J.; Rose, G. D. The protein coil library: A structural database of nonhelix, nonstrand fragments derived from the PDB. *Proteins: Struct., Funct., Genet.* **2005**, *58*, 852−854.
(45) Ghavami, A.; Van Der Giessen, E.; Onck, P. R. Coarse-Grained Potentials for Local Interactions in Unfolded Proteins. *J. Chem. Theory Comput.* **2013**, *9*, 432−440.
(46) *MATLAB*, version 2022a; The Math Works, Inc.: Natick, MA, 2022. https://www.mathworks.com/.
(47) Kullback, S.; Leibler, R. A. On Information and Sufficiency. *Ann. Math. Statist.* **1951**, *22*, 79−86.
(48) Reith, D.; Pütz, M.; Müller-Plathe, F. Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.* **2003**, *24*, 1624−1636.
(49) Jiang, F.; Zhou, C. Y.; Wu, Y. D. Residue-Specific Force Field Based on the Protein Coil Library. RSFF1: Modification of OPLS-AA/L. *J. Phys. Chem. B* **2014**, *118*, 6983−6998.
(50) Zhou, C. Y.; Jiang, F.; Wu, Y. D. Residue-Specific Force Field Based on Protein Coil Library. RSFF2: Modification of AMBER ff99SB. *J. Phys. Chem. B* **2015**, *119*, 1035−1047.
(51) Ye, W.; Ji, D. J.; Wang, W.; Luo, R.; Chen, H. F. Test and Evaluation of ff99IDPs Force Field for Intrinsically Disordered Proteins. *J. Chem. Inf. Model.* **2015**, *55*, 1021−1029.
(52) Zhang, Y. P.; Liu, H.; Yang, S.; Luo, R.; Chen, H. F. Well-Balanced Force Field ff03CMAP for Folded and Disordered Proteins. *J. Chem. Theory Comput.* **2019**, *15*, 6769−6780.
(53) Choi, J. M.; Pappu, R. V. Experimentally Derived and Computationally Optimized Backbone Conformational Statistics for Blocked Amino Acids. *J. Chem. Theory Comput.* **2019**, *15*, 1355−1366.
(54) Tozzini, V.; Rocchia, W.; Mccammon, J. A. Mapping all-atom models onto one-bead Coarse Grained Models: general properties and applications to a minimal polypeptide model. *J. Chem. Theory Comput.* **2006**, *2*, 667−673.
(55) Bansal, M.; Srinivasan, N. *Biomolecular Forms and Functions*; World Scientific/Indian Institute of Science: India, 2013.
(56) Quaglia, F.; Mészáros, B.; Salladini, E.; Hatos, A.; Pancsa, R.; Chemes, L. B.; Pajkos, M.; Lazar, T.; Peña-Díaz, S.; Santos, J.; et al. DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Res.* **2022**, *50*, D480−D487.
(57) Wang, X.; Ramírez-Hinestrosa, S.; Dobnikar, J.; Frenkel, D. The Lennard-Jones potential: when (not) to use it. *Phys. Chem. Chem. Phys.* **2020**, *22*, 10624−10633.
(58) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **2015**, *1−2*, 19−25.
(59) Rotkiewicz, P.; Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **2008**, *29*, 1460−1465.
(60) Guex, N.; Peitsch, M. C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **1997**, *18*, 2714−2723.
(61) Noel, J. K.; Levi, M.; Raghunathan, M.; Lammert, H.; Hayes, R. L.; Onuchic, J. N.; Whitford, P. C. SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput. Biol.* **2016**, *12*, No. e1004794.
(62) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **2000**, *298*, 937−953.
(63) Whitford, P. C.; Noel, J. K.; Gosavi, S.; Schug, A.; Sanbonmatsu, K. Y.; Onuchic, J. N. An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields. *Proteins: Struct., Funct., Genet.* **2009**, *75*, 430−441.
(64) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; In 'T Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **2022**, *271*, 108171.
(65) Schneider, T.; Stoll, E. Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions. *Phys. Rev. B* **1978**, *17*, 1302−1322.
(66) Dünweg, B.; Paul, W. Brownian Dynamics Simulations Without Gaussian Random Numbers. *Int. J. Mod. Phys. C* **1991**, *02*, 817−827.
(67) Kussie, P. H.; Gorina, S.; Marechal, V.; Elenbaas, B.; Moreau, J.; Levine, A. J.; Pavletich, N. P. Structure of the MDM2 Oncoprotein Bound to the p53 Tumor Suppressor Transactivation Domain. *Science* **1996**, *274*, 948−953.
(68) *Global Optimization Toolbox*, version 2022a; The MathWorks Inc.: Natick, MA, 2022. https://www.mathworks.com/help/stats/index.html.
(69) *Parallel Computing Toolbox*, version 2022a; The MathWorks Inc.: Natick, MA, 2022. https://www.mathworks.com/help/stats/index.html.
(70) Hoch, J. C.; Baskaran, K.; Burr, H.; Chin, J.; Eghbalnia, H. R.; Fujiwara, T.; Gryk, M. R.; Iwata, T.; Kojima, C.; Kurisu, G.; et al. Biological Magnetic Resonance Data Bank. *Nucleic Acids Res.* **2023**, *51*, D368−D376.
(71) Ulrich, E. L.; Baskaran, K.; Dashti, H.; Ioannidis, Y. E.; Livny, M.; Romero, P. R.; Maziuk, D.; Wedell, J. R.; Yao, H.; Eghbalnia, H.

R.; et al. NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. *J. Biomol. NMR* **2019**, *73*, 5−9.
(72) Mcgibbon, R. t.; Beauchamp, K. a.; Harrigan, M. p.; Klein, C.; Swails, J. m.; Hernández, C.; 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*, 1528−1532.
(73) Frank, A. T.; Law, S. M.; Ahlstrom, L. S.; Brooks, C. L. Predicting Protein Backbone Chemical Shifts From Cα Coordinates: Extracting High Resolution Experimental Observables from Low Resolution Models. *J. Chem. Theory Comput.* **2015**, *11*, 325−331.
(74) Marsh, J. A.; Singh, V. K.; Jia, Z.; Forman-Kay, J. D. Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: Implications for fibrillation. *Protein Sci.* **2006**, *15*, 2795−2804.
(75) Zhang, H.; Neal, S.; Wishart, D. S. RefDB: a database of uniformly referenced protein chemical shifts. *J. Biomol. NMR* **2003**, *25*, 173−195.
(76) *PyMOL Molecular Graphics System*, version 1.8; Schrodinger, LLC.: New York, NY, 2015. https://www.pymol.org/.
(77) Crameri, F.; Shephard, G. E.; Heron, P. J. The misuse of colour in science communication. *Nat. Commun.* **2020**, *11*, 5444.
(78) Crameri, F. *Scientific Colour Maps*, version 7.0.1; Zenodo, 2021., (accessed 2021−02−04).
(79) Wei, M. T.; Elbaum-Garfinkle, S.; Holehouse, A. S.; Chen, C. C. H.; Feric, M.; Arnold, C. B.; Priestley, R. D.; Pappu, R. V.; Brangwynne, C. P. Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nat. Chem.* **2017**, *9*, 1118−1125.
(80) Brady, J. P.; Farber, P. J.; Sekhar, A.; Lin, Y.-H.; Huang, R.; Bah, A.; Nott, T. J.; Chan, H. S.; Baldwin, A. J.; Forman-Kay, J. D.; et al. Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, No. E8194.
(81) Ahlers, J.; Adams, E. M.; Bader, V.; Pezzotti, S.; Winklhofer, K. F.; Tatzelt, J.; Havenith, M. The key role of solvent in condensation: Mapping water in liquid-liquid phase-separated FUS. *Biophys. J.* **2021**, *120*, 1266−1275.

https://doi.org/10.1021/acs.jpcb.4c02823
J. Phys. Chem. B 2024, 128, 6492−6508