Residue-Specific Force Field Based on the Protein Coil Library. RSFF1: Modification of OPLS-AA/L

Fan Jiang,*† Chen-Yang Zhou,‡ and Yun-Dong Wu*†‡

1Laboratory of Computational Chemistry and Drug Design, Laboratory of Chemical Genomics, Peking University Shenzhen Graduate School, Shenzhen 518055, China
2College of Chemistry, Peking University, Beijing 100871, China

ABSTRACT: Traditional protein force fields use one set of parameters for most of the 20 amino acids (AAs), allowing transferability of the parameters. However, a significant shortcoming is the difficulty to fit the Ramachandran plots of all AA residues simultaneously, affecting the accuracy of the force field. In this Feature Article, we report a new strategy for protein force field parametrization. Backbone and side-chain conformational distributions of all 20 AA residues obtained from protein coil library were used as the target data. The dihedral angle (torsion) potentials and some local nonbonded (1-4/1-5/1-6) interactions in OPLS-AA/L force field were modified such that the target data can be excellently reproduced by molecular dynamics simulations of dipeptides (blocked AAs) in explicit water, resulting in a new force field with AA-specific parameters, RSFF1. An efficient free energy decomposition approach was developed to separate the corrections on ϕ and ψ from the two-dimensional Ramachandran plots. RSFF1 is shown to reproduce the experimental NMR J-coupling constants of AA dipeptides better than other force fields. It has a good balance between α-helical and β-sheet secondary structures. It can successfully fold a set of α-helix proteins (Trp-cage and Homeodomain) and β-hairpins (Trpzip-2, GB1 hairpin), which cannot be consistently stabilized by other state-of-the-art force fields. Interestingly, the RSFF1 force field systematically overestimates the melting temperature (and the stability of native state) of these peptides/proteins. It has a potential application in the simulation of protein folding and protein structure refinement.

1. INTRODUCTION

In the last several decades, tremendous efforts have devoted to the bottom-up modeling of complex biomolecular systems, especially the atomistic molecular dynamics (MD) simulations.1,2 Recently, using a powerful special-purpose computer, Shaw’s group demonstrated that ab initio folding of a series of small peptides/proteins can be achieved, providing atomistic-level details of structures and dynamics.3 Besides theoretical understandings of biologically relevant processes, biomolecular simulations have increasing applications such as structural refinement4 and drug discovery.5 However, their reliability and predictive power crucially depend on the accuracy of the force fields used to describe the interactions among atoms.6

Although there are important issues associated with the force field development, such as the solvent effect7–9 and the electronic polarizability,10–11 many recent efforts in improving classical protein force fields (such as AMBER14, CHARMM15 and OPLS-AA16) have been focused on the accurate description of backbone (ϕ, ψ) and side-chain (χ) conformational preferences (Scheme 1), owing to their essential roles in determining peptide and protein conformations. Figure 1 gives a brief summary. Early efforts included the fitting to gas-phase quantum mechanics (QM) ϕ, ψ energy surface of dipeptides (such as Ac-Ala-NHMe) at local-MP2 level (OPLS-AA/Lψ17 CHARMM27,18,19) or the fitting to dipeptide QM energy surface calculated with continuum solvent model (AMBER ff0320). Later, gas-phase QM conformational energies of tetrapeptides (such as Ac-Ala3-NHMe) were used to fit the ϕ, ψ parameters (AMBER ff99SB).21 Recently, side-chain χ potentials of Ile, Leu, Asp, and Asn in ff99SB have been improved by fitting to gas-phase QM (local-MP2 level) energies of dipeptides (ff99SB-ildn).22

Despite these efforts, previous peptide simulations indicated biased secondary structure preferences from various force fields,23–51 which can result in failure of protein folding simulations. For example, the CHARMM27 cannot fold the all-β protein WW domain due to overstabilization of α-helical structures.26 Some more recent efforts (AMBER ff99SB* and ff03* by Best et al.,32 and CHARMM22* from Shaw’s group33) aim to correct this problem by means of a minor adjustment to the backbone potential to reproduce the experimental J-couplings of Alα and the α-helical content of a poly-Ala-based peptide measured from NMR. These corrections can result in more balanced α-helix and β-sheet preferences.34 Very recently, Best et al. empirically optimized the backbone CMAP correction parameters for CHARMM force field on the basis of...
Scheme 1. Definitions of Backbone $\phi$, $\psi$ and Side-Chain $\chi_i$ Dihedral Angles

```
\[ \phi : C-N-Ca-C \]
\[ \phi' : C-N-Ca-C \]
\[ \psi : C-N-Ca-C \]
\[ \psi' : C-N-Ca-C \]
\[ \chi_1 : C-Ca-Ca-C \]
\[ \chi_2 : C-Ca-Ca-C \]
\[ \chi_3 : C-Ca-Ca-C \]
```

“Dipeptide (terminally blocked amino acid) models of alanine (top) and arginine (bottom) are used, with some hydrogen atoms omitted for clarity. The definition of three side-chain rotamers ($g^+/(t/g^-)$) is also given.

Figure 1. Recent developments of all-atom protein force fields. AMBER has more force field variants than CHARMM and OPLS-AA. "ildn" is a modification for side-chain, whereas "-nmr" and "-ff" are modifications for backbone. Therefore, ff99SB-ildn, ff99SB-nmr, and ff99SB-ildn-nmr are also new variants of AMBER force fields.

A second crucial issue is how to incorporate the conformational distributions of 20 AA residues from the coil library into a force field. The coil library statistics are related to free energy surfaces with solvent effect. It should be compared with similar free energies from MD simulation of solvated systems, such as dipeptides in water. However, free energy calculations require equilibrium MD simulations. Too many such calculations are impractical for optimizing parameters if a traditional trial-and-error approach is adapted. Furthermore, if different $\chi_i$ side-chain rotamers are considered separately, totally 55 coil-library Ramachandran plots (one for Gly/Ala, two for Pro, and three for each of the other 17 AAAs) need to be fitted. Because these $\phi$, $\psi$ plots are fairly complicated and quite different, it is very difficult to develop one set of parameters to fit all of them simultaneously. A highly efficient parametrization strategy should be developed to avoid tremendous computational costs and human efforts.

In this paper, we present a new approach for the parametrization of a protein force field. Namely, we develop torsional parameters for each residue independently to fit the statistical conformational distributions derived from the protein coil library. Thus, the torsional parameters for backbone and side chain are residue specific. Although this approach can be applied to improve any available protein force field, we here use OPLS-AA/L as an example. We show that the parametrization is quite simple, and the new force field, named OPLS-AA/L, gives a quite low ($260 \, \text{K}$) $T_m$ of Trp-cage mini-protein. There were also recent attempts to optimize force fields do not reproduce the coil library rotamer distributions well. On the other hand, our QM calculations of model molecules with the solvent effect of water reproduce the side-chain rotamer preferences and rotamer-dependent Ramachandran plots from coil library quite well.
RSFF1, gives much improved simulation results such as the $^3\text{J}$-coupling constants of dipeptides in solution, balance between $\alpha$ and $\beta$ secondary structures, as well as the reliability in folding of peptide and protein structures.

2. METHODS

2.1. General Strategies. The general energy expression (eq 1) of the RSFF1 force field is within the framework of a classical force field:

$$V_{\text{total}} = V_{\text{bond}} + V_{\text{angle}} + V_{\text{torsion}} + V_{\text{local-lJ}} + V_{\text{lj}} + V_{\text{Coulomb}}$$

(1)

All bond stretching ($V_{\text{bond}}$) and angle bending ($V_{\text{angle}}$) potentials were adopted from the OPLS-AA/L without modification. The atomic $\sigma$ and $\varepsilon$ parameters of Lennard-Jones potential ($V_{\text{lJ}}$) and atomic charges of Coulomb potential ($V_{\text{Coulomb}}$) were also fully adopted from OPLS-AA/L. On the other hand, torsional potentials ($V_{\text{torsion}}$) for each AA residue were developed independently according to coil library data. Besides, local Lennard-Jones potentials ($V_{\text{local-lJ}}$) between atoms separated by three covalent bonds (1-4 interactions) were treated differently from ordinary $V_{\text{lJ}}$, and some 1-5 and 1-6 interactions were also treated specially (included in the $V_{\text{local-lJ}}$). These are described in more detail below:

**Torsional Parameters.** Similar to potentials of most physical-based force fields, the torsion potentials in RSFF1 are Fourier expansions:

$$V(\theta) = \sum k_n \cos(n\theta) = \sum c_n \cos^n(\theta)$$

(2)

where $\theta$ is a given dihedral angle and the coefficients $k_n$ or $c_n$ are parameters. There are $3 \times 3 = 9$ coupled torsion terms for each $\text{sp}^3-\text{sp}^3$ bond rotation, and $3 \times 3 = 6$ terms for $\text{sp}^3-\text{sp}^2$ rotation. To reduce the number of adjustable parameters, the torsion terms involving hydrogen atoms were kept the same as in OPLS-AA/L, such as $\text{H-N-C}=C$ for $\phi$ and $\text{N-C}=C=\text{H}_2$ for $\psi$. Because there are enough data from the coil library, AA-dependent torsion parameters were used for RSFF1. The torsion potential of peptide bond $V(\omega)$ and improper torsions were kept the same as OPLS-AA/L, because their purpose is to maintain planar geometry of conjugated systems.

**Local Lennard-Jones Parameters.** Besides the torsion terms, those 1-4 interactions shown in Scheme 2 also affect the bond rotation. Some of them were modified in RSFF1, because we found that the original values might give too strong effects. Furthermore, to optimize coupling between neighboring torsions, instead of using dihedral angle cross-terms or 2D grid-based CMAP-like corrections, we choose to directly modify the related 1-5 or 1-6 $V_{\text{local-lJ}}$ parameters ($\varepsilon$, $\sigma$) whenever necessary. In most cases, we set $\varepsilon = 0.1 \text{ kJ/mol}$ and only manually adjusted $\sigma$ for different strength of repulsion. Same $\varepsilon$ and $\sigma$ parameters are used for different AAs whenever possible, to reduce the efforts in manually adjusting them.

2.2. Parameterization Flow. All parameterizations were based on replica exchange molecular dynamics (REMD) simulations of various AA dipeptides (Ac-X-NHMe) in water.

Generally, initial parameters were assigned for a given AA residue and were then updated by repeating the procedure shown in Figure 2, until the simulated results are difficult to be further improved. One cycle of parameter optimization can be regarded as two successive transformations.

**From Force Field Parameters to Probability Distributions.** After REMD simulation using current parameters, the obtained trajectory is analyzed to obtain various statistical distributions (free energy surfaces), including the Ramachandran plot $p(\phi,\psi)$, the $\chi_i$-rotamer-dependent backbone conformational preference, percentages of three $\chi_i$-rotamers, and potentials of mean force (PMFs) for $\chi$ torsions. They are then compared with corresponding data from coil library statistics. The details are described in sections 2.3 and 2.4.

**From Probability Distributions to Updated Parameters.** As shown in Figure 2, various parameters were updated in parallel. In essence, the torsion potential $V(\theta)$ is updated according to the difference between the coil library PMF and the simulated PMF:

$$\Delta G(\theta) = G_{\text{coil}}(\theta) - G_{\text{MD}}(\theta)$$

$$\rightarrow \Delta V(\theta)$$

$$V_{\text{new}}(\theta) = V_{\text{old}}(\theta)$$

(3)

This strategy is similar to the iterative Boltzmann inversion (IBI) method. The required changes of the related Fourier coefficients can be obtained by fitting $\Delta V(\theta)$ to discrete $\Delta G(\theta)$ values. Before applying eq 3, a decomposition method is applied to derive corrections on $\phi$ and $\psi$ potentials from the 2D $\phi, \psi$ distributions. The details are described in section 2.5 for backbone torsions and section 2.6 for side-chain torsions. Besides, local L-J parameters ($\varepsilon$, $\sigma$) were adjusted manually only when necessary.

2.3. Molecular Dynamics Simulations. When OPLS-AA/L or our new force field was used, each dipeptide molecule was solvated with 319–330 TIP4P/Ew water molecules. For the simulations using AMBER and CHARMM force fields, similar numbers of TIP3P water molecules were used. The ionic Arg, Lys, Asp, and Glu side chains were neutralized with counterion (Cl$^-$ or Na$^+$). REMD simulations were performed using Gromacs version 4.5.4, with 12 replicas ranging from 298 to 451 K. The temperatures of intermediate replicas were calculated according to a recent study to give uniform exchange rates of ~16%. Swaps between neighboring replicas were attempted every 0.45 ps. The velocity rescaling thermostat was used with $\tau_T = 0.2$ ps was used to maintain the NVT ensemble. The periodic box size was obtained from averaging last 1 ns of a 3 ns NPT preproduction run at 300 K and 1 atm. Electrostatics were treated using the particle-mesh Ewald (PME) method with a real-space cutoff of 0.9 nm. van der Waals interactions were cut off at 0.9 nm with the long-range dispersion correction for energy and pressure.

In all simulations, the mass of water oxygen atom was reduced from 16 to 2 amu to increase the sampling efficiency without altering the thermodynamics equilibrium properties. All bond lengths involving hydrogen atoms were constrained by
the LINCS, allowing a time step of 3 fs. The REMD simulation of each dipeptide was carried out for \( \sim 90 \) ns per replica, and the structures were recorded every 0.6 ps. Trajectories from 298 K replica were used for statistical analysis, with the first 20 ns discarded. One such REMD simulation required \( \sim 24 \) h in real time on a 12-core 2.4 GHz Intel Xeon node.

2.4. Protein Coil Library and Statistical Analysis. Briefly, 6178 protein crystal structures with resolution <2.0 Å and R factor <0.2 were retrieved from the Protein Data Bank (PDB) database with 50% sequence identity cutoff. The popular DSSP program was used to assign secondary structures. Residues within any secondary structures—including the DSSP codes G(3\(_{10}\)-helix), H(α-helix), I(π-helix), B(β-bridge), E(β-sheets), and T(turn)—were all excluded from the coil library. Residues preceding proline or containing backbone atoms with B factor >35 were also excluded. Following our previous work, residues with short polar side-chains (Asp, Asn, Ser, Thr) preceding any residue with \(-60^\circ < \psi < +60^\circ\) were excluded, to avoid inter-residue H-bonding between their side-chain O atom and the successive backbone amide H atom.

Same as the previous work of Amir et al. and ours, a 2D Gaussian kernel estimator was used to extract \( \phi, \psi \) distributions, considering the periodicity of the dihedral angles:

\[
n(\phi, \psi) = \sum_i w_i \exp\left[\min(\phi_i - \phi, 360^\circ - \phi_i - \phi)^2 + \min(\psi_i - \psi, 360^\circ - \psi_i - \psi)^2 / 2\sigma^2\right]
\]

Here \( i \) counts for all residues of given type in the coil library, and \( w_i \) is the \( 1/m \) weighting for \( m \) identical chains in one PDB structure. \( \phi_i \) and \( \psi_i \) are the observed backbone dihedral angles for residue \( i \). \( 10^\circ \times 10^\circ \) grids of \( (\phi, \psi) \) and \( \sigma = 10^\circ \) were used. For AAs other than Ala and Gly, the statistics were carried out for three side-chain \( \chi_1 \) rotamers separately. The obtained distributions were shown in the Supporting Information Figure.
S1. The same approach was used to analyze the trajectories from REMD simulations with \( w_i = 1 \) for each structural frame. For side-chain \( \chi \) distributions, a 1D Gaussian estimator was used with grid space of \( 6^\circ \) and \( \sigma = 7^\circ \).

The similarity coefficient \( S \) between \((\phi, \psi)\) distribution from coil library \( n_{\text{coil}}(\phi, \psi) \) and that from simulation \( n_{\text{MD}}(\phi, \psi) \) can be calculated without normalization:

\[
S = \frac{\sum n_{\text{coil}}(\phi, \psi) n_{\text{MD}}(\phi, \psi)}{\sqrt{\sum n_{\text{coil}}(\phi, \psi)^2 \cdot \sum n_{\text{MD}}(\phi, \psi)^2}}
\]

(5)

Only two identical distributions will give \( S = 1 \).

The probability distributions were obtained by normalizing the statistical counts \( n(\phi, \psi) \):

\[
p(\phi, \psi) = \frac{n(\phi, \psi) + \epsilon}{\sum n(\phi, \psi)}
\]

(6)

To avoid infinity free energy when \( n = 0 \), pseudocount \( \epsilon = 0.02 \) is used with negligible changes in the allowed Ramachandran regions.

2.5. Optimization of Backbone Dihedral Potentials.

The difference between 2D \( \phi, \psi \) free energy surfaces from the coil library and REMD simulations is separated into the corrections for 1D \( \phi \) component and \( \psi \) component:

\[
\Delta G^*_{\phi}(\phi) + \Delta G^*_{\psi}(\psi) = G^*_{\text{coll}}(\phi, \psi) - G^*_{\text{MD}}(\phi, \psi)
\]

(7)

To solve this equation, we can rewrite eq 7 into eq 8 by using hypothetical probability distributions for both \( \Delta G^*_{\phi} \) and \( \Delta G^*_{\psi} \):

\[
\delta p^*_{\phi}(\phi) \delta p^*_{\psi}(\psi) = p^*_{\text{coll}}(\phi, \psi)/p^*_{\text{MD}}(\phi, \psi)
\]

(8)

Then \( \delta p^*_{\phi} \) and \( \delta p^*_{\psi} \) can be solved from the known \( p^*_{\text{coll}}(\phi, \psi) \) and \( p^*_{\text{MD}}(\phi, \psi) \) by applying following two equations iteratively:

\[
\delta p^*_{\phi}(\phi) = \frac{\sum_{\psi} p^*_{\text{coll}}(\phi, \psi) \delta p^*_{\psi}(\psi)}{\sum_{\psi} p^*_{\text{MD}}(\phi, \psi) \delta p^*_{\psi}(\psi)}
\]

(9a)

\[
\delta p^*_{\psi}(\psi) = \frac{\sum_{\phi} p^*_{\text{coll}}(\phi, \psi) \delta p^*_{\phi}(\phi)}{\sum_{\phi} p^*_{\text{MD}}(\phi, \psi) \delta p^*_{\phi}(\phi)}
\]

(9b)

Uniform distribution of \( \delta p^*_{\phi} \equiv 1 \) was used as the initial guess, and the convergence can always be achieved within 10 iterations. Then the \( \delta p^*_{\phi} \) and \( \delta p^*_{\psi} \) are converted to free energy scale separately:

\[
\Delta G^*_{\phi}(\phi) = -RT \ln \delta p^*_{\phi}(\phi)
\]

(10a)

\[
\Delta G^*_{\psi}(\psi) = -RT \ln \delta p^*_{\psi}(\psi)
\]

(10b)

The obtained \( \Delta G^*_{\phi} \) and \( \Delta G^*_{\psi} \) are discrete functions with 10° interval. They are fitted to analytical dihedral potentials in the force field:

\[
\Delta V_{\phi}(\phi) + \Delta V^*_{\phi}(\phi) = \sum_{n=0}^{5} \Delta \epsilon_{\phi, n} \cos^n(\phi) + \sum_{n=0}^{5} \Delta \epsilon_{\phi, n} \cos^n(\phi')
\]

(11a)

\[
\Delta V_{\psi}(\psi) + \Delta V^*_{\psi}(\psi) = \sum_{n=0}^{5} \Delta \epsilon_{\psi, n} \cos^n(\psi) + \sum_{n=0}^{5} \Delta \epsilon_{\psi, n} \cos^n(\psi')
\]

(11b)

where \( \Delta \epsilon_{\phi, n} \), \( \Delta \epsilon_{\phi, n} \), \( \Delta \epsilon_{\psi, n} \), and \( \Delta \epsilon_{\psi, n} \) are the changes of Fourier coefficients related to \( V_{\phi}, V_{\phi'}, V_{\psi}, \) and \( V_{\psi'} \) terms, respectively. Of course, there are no \( V_{\phi} \) and \( V_{\psi} \) terms for Gly. The zeroth-order \( (n = 0) \) terms are constant and do not affect the force field, but they are necessary as an offset to minimize the difference with target \( \Delta G_{\phi} \) or \( \Delta G_{\psi} \). Assuming the relationships \( \phi = \phi = 120^\circ \) and \( \psi = \psi + 120^\circ \), the parameters were fitted by minimizing the following penalty functions:

\[
s_{\phi} = \sum w(\phi) \left[ \Delta V_{\phi}(\phi) + \Delta V^*_{\phi}(\phi - 120^\circ) - \Delta G^*_{\phi}(\phi) \right]^2
\]

(12a)

\[
s_{\psi} = \sum w(\psi) \left[ \Delta V_{\psi}(\psi) + \Delta V^*_{\psi}(\psi + 120^\circ) - \Delta G^*_{\psi}(\psi) \right]^2
\]

(12b)

The \( \phi \) and \( \psi \) values with higher occurrences have higher weight \( w \) in the fitting:

\[
w(\phi) = \left( \sum p_{\text{coll}}(\phi, \psi) \right)^{-1/2}
\]

(13a)

\[
w(\psi) = \left( \sum p_{\text{coll}}(\phi, \psi) \right)^{-1/2}
\]

(13b)

The square root of the probability corresponds to a Boltzmann weight at 600 K, similar to the 500 K previously used by Lindorff-Larson et al.\textsuperscript{22} It is a compromise between equal weight (infinite \( T \)) that cannot ensure a high accuracy at the most probable conformations and the 300 K Boltzmann weight that will lead to large errors in the barrier regions. Besides, \( w(\psi) \) is doubled at \( \psi = -40^\circ \) to achieve better fitting at the \( \alpha \)-helix conformation, which is highly populated in protein structures but much less favored in a coil library. A very simple version of self-adaptive evolution strategy was used in the parameter fitting. In each iteration, one parameter is randomly chosen for mutation, by adding a random value of normal distribution with standard deviation \( \sigma \). Only a mutation that improves the fitness is accepted. Following the one-fifth success rule, if the acceptance rate of mutating a certain parameter is >5%, the corresponding \( \sigma \) is increased to 1.5 \( \sigma \); otherwise, the \( \sigma \) is reduced to 0.6 \( \sigma \). Excellent convergence can be achieved within 10\textsuperscript{5} iterations. Although Fourier expansion up to fifth-order was used for backbone dihedral angles, the fitting is well overdetermined due to 10\textsuperscript{5} interval for \( \Delta G_{\phi} \) and \( \Delta G_{\psi} \). An actual example of the fitting was given in Figure S2 (Supporting Information). It is an important feature that our new force field places higher precision on describing the conformations with low free energies.

For AAs other than Ala/Gly/Pro, there are three side-chain rotamers, which can give quite different \( \phi, \psi \) plots. Under the standard forms of current force fields, we cannot use different \( \phi, \psi \) potentials for different rotamers. However, if we directly use eq 3 to obtain \( \phi, \psi \) distribution regardless of the rotameric state, the most abundant rotamer will weight more for the optimized parameters and the Ramachandran plot of the least favored rotamer may not be well reproduced. To reduce this bias, we use the following to combine the three rotamer-dependent \( \phi, \psi \) distributions:

\[
n(\phi, \psi) = n_{g+}(\phi, \psi)/\sqrt{N_{g+}} + n_{g}(\phi, \psi)/\sqrt{N_n} + n_{g-}(\phi, \psi)/\sqrt{N_{g-}}
\]

(14)
Contours are drawn every \( \chi \) for manually adjusted to reproduce the rotational barriers. Thus, the two C\(_\alpha\) alanine (upper) and glycine (lower), force field optimization and the results from the final optimized force field (c, g) are given. The similarity coefficient (S) with respect to each simulated plot. Contours are drawn every \( k_B T \) free energy difference. The same scale is used throughout the paper.

where \( N_{g_+}, N_0 \) and \( N_{g-} \) are the total numbers of the \( g_+, t \), and \( g- \) rotamers, respectively. The square root of total number ensures that the more abundant rotamer still weights more in the fitting. The obtained \( n_{\text{coll}}(\phi,\psi) \) and \( n_{\text{MD}}(\phi,\psi) \) are then normalized to \( p_{\text{coll}}(\phi,\psi) \) and \( p_{\text{MD}}(\phi,\psi) \) using eq 5. They are directly related to \( \phi,\psi \) free energy surfaces on the basis of the Boltzmann distribution law.

2.6. Optimization of Side-Chain Torsion Potentials. Less Fourier terms (up to third order) were used for side-chain \( \chi \) torsions. For the side-chain \( \chi \) and \( \chi' \) potentials, the new force field use fewer terms than OPLS-AA/L to reduce the number of parameters:

\[
V(\chi) + V(\chi') = k_1 \cos(\chi) + k_2 \cos(3\chi) + k_3 \cos(3\chi')
\]

The update of parameters \( k_1 \) and \( k_1' \) is based on comparing the simulated and target populations of \( g_+/t/g- \) rotamers:

\[
k_{1,\text{new}} - k_{1,\text{old}} = \Delta k_1 = aRT \ln \left( \frac{P_{g_+} / P_{g_0}}{P_{\text{coll}} / P_{\text{MD}}} \right)
\]

\[
k_{1',\text{new}} - k_{1',\text{old}} = \Delta k_1' = aRT \ln \left( \frac{P_{g_+} / P_{g_0}}{P_{\text{coll}} / P_{\text{MD}}} \right)
\]

In practice, we found \( \alpha = 0.6 \) is a good choice. For \( \beta \)-branched Val and Ile, there are two N--C\(_\alpha\)--C\(_\beta\)--C\(_\gamma\) dihedral angles and two C--C\(_\alpha\)--C\(_\beta\)--C\(_\gamma\) dihedral angles on one C\(_\alpha\)--C\(_\beta\) rotation. Thus, the \( p_{g_+}, p_{g^0} \) in eq 16a and \( p_{g_+}, p_{g^0} \) in eq 16b were replaced by \( p_{g_+}, p_{g^0} \) and \( p_{g_+}, p_{g^0} \), respectively. For Thr, The N--C\(_\alpha\)--C\(_\beta\)--C\(_\gamma\) dihedral potentials were set to zero, so eq 16a and eq 16b can be applied. Unlike \( k_1 \) and \( k_1' \), \( k_3 \) is manually adjusted to reproduce the rotational barriers.

We use the functional form similar to OPLS-AA/L force field for \( \chi_{\alpha-1} \) torsion potentials:

\[
V(\chi) = k_1 \cos(\chi) + k_2 \cos(2\chi) + k_3 \cos(3\chi)
\]

In most cases, \( k_1 \) controls the trans/gauche preference, and \( k_3 \) controls the rotational barrier. Except for the \( \chi_2 \) of Asx and \( \chi_3 \) of Glx, we found \( k_3 = 0 \) can be used for all side-chain torsions. The rotation of terminal \(-CO-NH\) group (\( \chi_2 \) of Asn and \( \chi_3 \) of Gln) involves two coupled dihedral angles: C--C--C\(_\alpha\)--C--C--C--N. The potential for the latter is set to zero to simplify the parametrization. The updates of Fourier coefficients \( \Delta k_3 \) were obtained from minimizing:

\[
s = \sum \left[ \Delta V(\chi) - RT \ln \frac{P_{\text{coll}}(\chi)}{P_{\text{MD}}(\chi)} \right]^2
\]

The \( \chi \) values with relative free energy >20 kJ/mol from coil library are not included in the fitting. The fitting was also carried out using self-adaptive evolution strategy.

3. RESULTS AND DISCUSSION

3.1. Alanine and Glycine. We began our studies with Ala because most AAs are its derivatives. The reparameterization began with setting all four \( \phi, \phi', \psi, \psi' \) potentials to zero, with \( \sigma = 0.270 \) nm and \( \varepsilon = 0.1 \) kJ/mol for all the six 1-4 L-J interactions in Scheme 2. As shown in Figure 3a, the obtained \( \phi, \psi \) distribution (a) was very different from the target coil library data (S = 0.49). However, after only one cycle of optimization using our \( \phi, \psi \) decomposition approach, the simulated \( \phi, \psi \) plot (b) was significantly improved to \( S > 0.97 \). This agreement is already close to the final RSFF1 force field (S = 0.985). This highly efficient approach makes the optimization of torsion potentials no longer the bottleneck in our force field development.

As shown in Figure 3b, the densities for \( \phi < -160^\circ \) with \( \psi \) in the range +40° to −80° are still higher than the coil library distributions whereas the C\(_5\) basin is not deep enough. These can be improved by adding a weak repulsion between H\(_\alpha\)--N\(_\alpha\) to destabilize the \( \alpha' \) conformation and a weak attraction between H\(_\beta\)--O\(_i\) to stabilize the C\(_5\) conformation. We also reduce the repulsion between O\(_\alpha\)--C\(_\beta\) to reduce the barrier at \( \phi = 0^\circ \). When proper modifications shown in Table 1 are introduced, with several additional cycles of optimization, the \( \phi, \psi \) distribution from the Ala dipeptide simulation agrees excellently with the coil library distributions.
Interestingly, with optimized backbone (Figure 3g). Similar to the case of Ala, the charged terminal group has limited electronic transferability of the OPLS-AA/L parameters optimized on Ala. Indeed, in the gas phase the most stable conformation of Glu dipeptide forms a H-bond between side chain and backbone,17 which is not favored in protein structures. From the coil library, Glu and Gln have very similar rotamer preferences,41 indicating a large deviation from the original ones in OPLS-AA/L (Table 3).

Among 20 AAs, Gly has a special conformational flexibility due to its lack of a side chain. We then used Gly to examine the transferability of the $V_{\text{local-LJ}}$ parameters optimized on Ala. Interestingly, with optimized $\phi$, $\psi$ parameters, RSFF1 gives the $\phi$, $\psi$ distribution very similar to the coil library distribution (Figure 3g). Similar to the case of Ala, the $S$ value increases from 0.43 to 0.932 after just one cycle of optimization (Figure 3e,f). Because of the success on Gly residue, the same backbone $V_{\text{local-LJ}}$ parameters were used for all other AAs.

### 3.2. Side-Chain $\chi$ Torsions.

As shown in Table 2, the first-order Fourier coefficients $k_1$ and $k_1'$ for the side-chain $\chi_1$ and $\chi_1'$ torsions (Scheme 1) in OPLS-AA/L have large deviations, with the ranges $-9.4$ to $+11.4$ and $-11.8$ to $+2.2$ kJ/mol, respectively. The large deviations might be partly resulted from applying eq 16, these parameters were optimized to achieve excellent agreement with the coil library rotamer distributions (Figure S3, Supporting Information). The ranges of $k_1$ and $k_1'$ are reduced to $-3.7$ to $+2.3$ and $-0.9$ to $+5.5$ kJ/mol, respectively.

As shown in Figure 4, RSFF1 reproduces the whole free energy profiles (or PMFs) from the coil library quite well. A few rotational barriers from RSFF1 simulations are slightly lower than those from the coil library. The OPLS-AA/L only reproduces PMFs well for some hydrophobic residues (such as Leu, Phe, Tyr, Trp, Val, Ile), but it does not describe the rotamer distributions (relative free energies of the three minima) well in many cases. The problem is most serious for Glu, His, Cys, Ser, Thr, and Asp. In RSFF1, besides the $k_1$ parameter in eq 15, the $1-4$ L-J parameters between non-hydrogen atoms are also adjusted to give a good description of the rotational barriers. These parameters are shared between different $\chi$ torsions and different AA types. These modified $1-4$ L-J parameters resulted in much weaker interactions compared with the original ones in OPLS-AA (Table 3).

Besides $\chi_1$ rotation, all side-chain $\chi_{1-3}$ rotational free energy profiles were optimized to match coil library PMFs (Figure S4, AA/L force field tend to compensate the unbalanced 1-4/1-5 electrostatic interactions. As shown in Scheme 3, the $g^+$ rotamer of Asp has attractive 1-4 interaction and repulsive 1-5 interaction, whereas the $t$ rotamer of Asp has 1-4 repulsion and 1-5 attraction. In the original OPLS-AA/L force field, the 1-4 electrostatic interactions are scaled down by a factor of 0.5, which significantly favors the $t$ rotamer and disfavors the $g^+$ rotamer. In OPLS-AA/L, negative $k_1$ ($-9.4$ kJ/mol) for Asp still cannot fully compensate the strong $t$ preference from the unbalanced electrostatic 1-4/1-5 interactions. A similar situation also occurred for Ser and Thr, in an opposite way to Asp (also shown in Scheme 3).

It might be more appropriate not to scale down the 1-4 electrostatic interactions. Indeed, Smith and Karplus found that reducing the 1-4 electrostatic interactions by 50% led to qualitatively incorrect trans-gauche energy for n-butane.68 In developing ECEPP-05 force field, Scheraga et al. found that no scaling of 1-4 electrostatic interactions provided the best results for conformational energies of 1,3-propanediol.69 In RSFF1 force field, we do not scale down the 1-4 electrostatics, resulting in more consistent torsion parameters. Within a few cycle of applying eq 16, these parameters were optimized to achieve excellent agreement with the coil library rotamer distributions (Figure S3, Supporting Information). The ranges of $k_1$ and $k_1'$ are reduced to $-3.7$ to $+2.3$ and $-0.9$ to $+5.5$ kJ/mol, respectively.

As shown in Figure 4, RSFF1 reproduces the whole $\chi_1$ free energy profiles (or PMFs) from the coil library quite well. A few rotational barriers from RSFF1 simulations are slightly lower than those from the coil library. The OPLS-AA/L only reproduces PMFs well for some hydrophobic residues (such as Leu, Phe, Tyr, Trp, Val, Ile), but it does not describe the rotamer distributions (relative free energies of the three minima) well in many cases. The problem is most serious for Glu, His, Cys, Ser, Thr, and Asp. In RSFF1, besides the $k_1$ parameter in eq 15, the 1-4 L-J parameters between non-hydrogen atoms are also adjusted to give a good description of the rotational barriers. These parameters are shared between different $\chi$ torsions and different AA types. These modified 1-4 L-J parameters resulted in much weaker interactions compared with the original ones in OPLS-AA (Table 3).

Besides $\chi_1$ rotation, all side-chain $\chi_{1-3}$ rotational free energy profiles were optimized to match coil library PMFs (Figure S4, Table 1. Parameters for All Modified 1-5 and 1-6 L-J Interactions

<table>
<thead>
<tr>
<th>pair</th>
<th>type</th>
<th>$\sigma$ (nm)</th>
<th>$\epsilon$ (kJ/mol)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H--O</td>
<td>1-5</td>
<td>0.180</td>
<td>5.0</td>
</tr>
<tr>
<td>H--N$_{\text{N}}$</td>
<td>1-5</td>
<td>0.290</td>
<td>0.1</td>
</tr>
<tr>
<td>O$_1$--C</td>
<td>1-5</td>
<td>0.270</td>
<td>0.1</td>
</tr>
<tr>
<td>C$_1$--O</td>
<td>1-5</td>
<td>0.230</td>
<td>5.0</td>
</tr>
<tr>
<td>C$_1$--O</td>
<td>1-5</td>
<td>0.230</td>
<td>3.0</td>
</tr>
<tr>
<td>O$_2$--C$_1$</td>
<td>1-6</td>
<td>0.250</td>
<td>1.0</td>
</tr>
<tr>
<td>O$_2$--C</td>
<td>1-5</td>
<td>0.270</td>
<td>1.5</td>
</tr>
<tr>
<td>C$_2$--H</td>
<td>1-5</td>
<td>0.290</td>
<td>0.1</td>
</tr>
<tr>
<td>C$_2$--H</td>
<td>1-5</td>
<td>0.260</td>
<td>0.1</td>
</tr>
<tr>
<td>C$_2$--C$_1$</td>
<td>1-5</td>
<td>0.350</td>
<td>0.1</td>
</tr>
<tr>
<td>O$_2$--N</td>
<td>1-5</td>
<td>0.310</td>
<td>0.1</td>
</tr>
<tr>
<td>C$<em>3$--N$</em>{\text{N}}$</td>
<td>1-5</td>
<td>0.350</td>
<td>0.1</td>
</tr>
<tr>
<td>N$_{\text{N}}$--O</td>
<td>1-6</td>
<td>0.375</td>
<td>0.1</td>
</tr>
<tr>
<td>N$_{\text{N}}$--N</td>
<td>1-5</td>
<td>0.320</td>
<td>0.1</td>
</tr>
<tr>
<td>O$<em>{\text{N}}$--C$</em>{\text{N}}$</td>
<td>1-5</td>
<td>0.320</td>
<td>0.1</td>
</tr>
<tr>
<td>O$_{\text{N}}$--O</td>
<td>1-5</td>
<td>0.330</td>
<td>0.1</td>
</tr>
<tr>
<td>O$<em>{\text{N}}$--N$</em>{\text{N}}$</td>
<td>1-5</td>
<td>0.345</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 2. Side-Chain $\chi$ Fourier Coefficients (kJ/mol) of Some AA Residue from the OPLS-AA/L and the RSFF1 Force Fields

<table>
<thead>
<tr>
<th></th>
<th>OPLS-AA/L</th>
<th>RSFF1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k_1$</td>
<td>$k_2$</td>
</tr>
<tr>
<td>Glu</td>
<td>10.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Gln</td>
<td>2.4</td>
<td>1.1</td>
</tr>
<tr>
<td>Lys</td>
<td>1.3</td>
<td>0.4</td>
</tr>
<tr>
<td>Ser</td>
<td>11.4</td>
<td>1.8</td>
</tr>
<tr>
<td>Thr</td>
<td>11.4</td>
<td>1.8</td>
</tr>
<tr>
<td>Asp</td>
<td>-9.4</td>
<td>-2.0</td>
</tr>
<tr>
<td>Asn</td>
<td>-7.3</td>
<td>1.4</td>
</tr>
</tbody>
</table>
Supporting Information). Using the fitting scheme described by eq 18 in section 2.6, the target PMF can be achieved by only one cycle of optimization (Figure S5, Supporting Information). As a result, fewer χ parameters are used in RSFF1 compared with OPLS-AA/L (Table S1, Supporting Information). The PMF of χ2 rotation in Asp is shown in Figure 5. From coil library, χ2 around 0° and 180° are mostly favored, which may be stabilized by n → π* interaction between Asp side-chain carboxylic O atom and backbone C atom, as indicated from QM calculations.70 In the contrary, OPLS-AA/L and AMBER ff99SB gave maximum free energies at χ2 near 165°. The χ2 PMF from ff03 force field (not shown) is very similar to that from ff99SB. Simulation using improved version ff99SB-ILDN still gives χ2 PMF different from the coil library PMF, although it is better than ff99SB. Indeed, the parametrization for Asp is most difficult because the side-chain has strong interactions with the backbone.

3.3. Rotamer-Dependent Ramachandran Plots. Because the high efficiency of our parametrization methods and sufficient data from coil library, it is very convenient to use different ϕ, ψ parameters for different AAs. Still, same set of ϕ, ψ parameters are used for AAs with very similar local conformational features: (1) Glu/Gln/Lys/Arg/Met/Leu with single sp3 Cγ atom, (2) Phe/Tyr/Trp with single nonpolar sp2 Cγ atom, and (3) Val/Ile with nonpolar β-branched side chains. Other AAs with more polar γ atoms use their special ϕ, ψ parameters, including Ser, Thr, Cys, His, Asp, and Asn. The final ϕ, ψ torsion parameters are given in the Supporting Information (Table S2).

To account for the coupling between side-chain and backbone conformations, an additional 1-5 L-J interaction between the Cγ atom and the backbone amide H atom is added. This may be related to the fact that β-branched AAs (Val, Ile) with two nonpolar Cγ atoms have intrinsically highest β-sheet propensities and low α-helix propensities (Scheme 4), as discussed earlier by Han et al.71 In addition, Ala without the Cγ atom has the highest propensity for α-helix formation. The steric repulsions between the polar H atom and nonpolar atoms can be missing if the van der Waals radius of amide H atom was ignored. Like most Vlocal-LJ in RSFF1, we set ε = 0.1 kJ/mol and adjusted the σ value. Finally, σ = 0.320 nm was used for the interactions with the sp3 Cγ atom (most AAs) or Sγ atom (Cys), and slightly weaker repulsion of σ = 0.310 nm was used for the interactions with sp2 Cγ atom in aromatic side chains. With this additional H···Cγ/Sγ interaction and optimized backbone ϕ, ψ potentials, a significant increase of similarities (S) with coil library data can be achieved. As shown in Table 4, OPLS-AA/L simulations give S < 0.8 for g− rotamers of most AAs, and S < 0.9 for all t rotamers. On the other hand, RSFF1
Scheme 4. Newman Projections along the N–Ca Bond (ϕ Torsion)\textsuperscript{a}

```
\begin{align*}
g^+ & \quad \text{C}_\gamma \quad \text{g}^- \\
g^+ & \quad \text{C}_\gamma \quad \text{g}^-
\end{align*}
```

\*The amide H atom is close to Cβ and Cγ atoms when in \textit{α}-helical conformation. This H···Cγ repulsion also depends on the side-chain rotamer.

Table 4. Similarity Coefficients (S) between Simulated Rotamer-Dependent ϕ, ψ Distributions and Coil Library Statistics

<table>
<thead>
<tr>
<th></th>
<th>OPLS-AA/L</th>
<th>RSFF1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>g+ t g− all</td>
<td>g+ t g− all</td>
</tr>
<tr>
<td>A</td>
<td>0.83</td>
<td>0.985</td>
</tr>
<tr>
<td>G</td>
<td>0.35</td>
<td>0.939</td>
</tr>
<tr>
<td>P</td>
<td>0.88</td>
<td>0.94</td>
</tr>
<tr>
<td>E</td>
<td>0.90</td>
<td>0.85</td>
</tr>
<tr>
<td>Q</td>
<td>0.93</td>
<td>0.91</td>
</tr>
<tr>
<td>K</td>
<td>0.92</td>
<td>0.85</td>
</tr>
<tr>
<td>R</td>
<td>0.92</td>
<td>0.88</td>
</tr>
<tr>
<td>M</td>
<td>0.95</td>
<td>0.85</td>
</tr>
<tr>
<td>L</td>
<td>0.95</td>
<td>0.86</td>
</tr>
<tr>
<td>F</td>
<td>0.90</td>
<td>0.81</td>
</tr>
<tr>
<td>Y</td>
<td>0.90</td>
<td>0.82</td>
</tr>
<tr>
<td>W</td>
<td>0.90</td>
<td>0.85</td>
</tr>
<tr>
<td>C</td>
<td>0.90</td>
<td>0.82</td>
</tr>
<tr>
<td>V</td>
<td>0.90</td>
<td>0.80</td>
</tr>
<tr>
<td>I</td>
<td>0.90</td>
<td>0.89</td>
</tr>
<tr>
<td>S</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>T</td>
<td>0.86</td>
<td>0.74</td>
</tr>
<tr>
<td>D</td>
<td>0.81</td>
<td>0.47</td>
</tr>
<tr>
<td>N</td>
<td>0.87</td>
<td>0.51</td>
</tr>
</tbody>
</table>

simulations give S > 0.92 for all except a few cases. As shown in Figure 6, there is an additional α’ basin (ϕ, ψ \approx \{−140°, 30°\}) in the ϕ, ψ plots of g− and t rotamers of Lys from OPLS-AA/L simulations, which is absent in coil library results. This additional α’ basin agrees with the OPLS-AA/L simulation of Ala dipptide, which was suppressed in RSFF1 by modified local interactions. All the χ1-dependent ϕ, ψ plots from RSFF1 simulations are given in the Supporting Information (Figure S6).

Also from Table 4, OPLS-AA/L gives especially low S values (0.5 or less) for t and g– rotamers of Asx. From the coil library results (Figure 6), the t Asp favors the Cβ-like conformation around ϕ, ψ \approx −80°, +80° and ϑt conformation with ϕ > 0. For g– Asp, the bottom of the ϑt basin is shifted to ϕ, ψ \approx −110°, +10° with a higher population than extended conformations. These conformations are not well stabilized in OPLS-AA/L force field. Asn also has similar special features.

Modifications of some 1-5/1-6 L-J interactions for Asx (Table 1) were introduced to achieve significantly better agreement with coil library observations. All these modified interactions involve pairs between polar atoms and may function to compensate possible small inaccuracies in the water-mediated electrostatic interactions. Compared with the case for other AAs, full optimization of these parameters is rather difficult, which required most of the efforts in our RSFF1 parametrization. Interestingly, different from the majority situations of ε = 0.1 kJ/mol, the modified L-J interactions between polar C and O atoms in Asx are attractive. The L-J potential with σ = 0.230 nm and ε = 5.0 kJ/mol gives an energy of −2.8 kJ/mol at the distance of 0.31 nm. There it can be stabilization from the n \rightarrow π* interaction between the oxygen long pair and antibond π orbital of the C=O group.\textsuperscript{72} Still, RSFF1 gives the ϕ, ψ plot of t Asp not in exact agreement with the coil library (S = 0.81) (Figure 6). Simple L-J potential may not fully account for the directionality of the n \rightarrow π* interaction.

For Ser and Thr, our QM calculations indicated that their Oγ and Cγ-1 atoms can have short distance of <3.0 Å. This can lead to >4 kJ/mol repulsion when default OPLS-AA L-J parameters are used. We thus set ε = 0.1 kJ/mol and σ = 0.32 nm to give reduced repulsion at short C=O distance. This allows strong electrostatic attraction between the two oppositely charged atoms, which is sufficient to give satisfactory results.

3.4. NMR J Couplings of Dipeptides. To compare the performance of RSFF1 with that of other force fields, we calculate the NMR \( J_{\text{HdHl}} \) couplings of all 19 dipeptides (except Pro) from their ϕ torsions sampled in the simulations and compared them with experimental data reported by Avbelj et al.\textsuperscript{50} The \( J_{\text{HdHl}} \) scalar coupling has been widely used in experimental characterization of conformations of short peptides in solution. It is sensitive to the distribution of backbone ϕ angle through the Karplus relationship:

\[
J_{\text{HdHl}} = A \cos^2(\phi - 60°) + B \cos(\phi - 60°) + C
\]

Several different sets of empirical Karplus parameters (A, B, C in eq 19) were reported, from different fittings of experimental \( J_{\text{HdHl}} \) values of different proteins to their X-ray or NMR structures.

From Table 5, for all Karplus parameter sets used, RSFF1 force field gives the lowest RMSD values, indicating better agreement with experimental \( J_{\text{HdHl}} \). Especially, when the 2007 parameter set is used, only RSFF1 gives the RMSD value (0.19 Hz) smaller than the estimated uncertainty (σ = 0.36 Hz) in deriving the Karplus parameters. Unlike the RMSD values, different Karplus parameter sets give nearly the same correlation coefficients (R) between calculated and experimental \( J_{\text{HdHl}} \). The RSFF1 gives significantly higher r values (>0.9) than other force fields.

As shown in Figure 7, experimental \( J_{\text{HdHl}} \) coupling of the Ala dipptide is significantly smaller than its derivatives (non-Gly/Ala AAs) by 0.6–1.8 Hz. Different from experiments, A99sb-ildn force field gives similar \( J_{\text{HdHl}} \) coupling (within 0.7 Hz) for Ala and its derivatives (except Val). The \( J_{\text{HdHl}} \) for Ala and some AAs such as Glu and Val are significantly overestimated. The overestimation of Ala \( J_{\text{HdHl}} \) is also observed in OPLS-AA/L simulation. Compared with A99sb-ildn-NMR force field consistently reduced the \( J_{\text{HdHl}} \) of all AAs, resulting in good results for some AAs, but \( J_{\text{HdHl}} \) values for Gly, Cys, Asn, and His were considerably underestimated. Unlike other force fields, RSFF1 can reproduce the gap between \( J_{\text{HdHl}} \) of Ala and its derivatives. The excellent performance of RSFF1 agrees with the previous finding that
the $\gamma_{\text{H}_2\text{H}_2}$ directly calculated from coil library $\phi$ distributions agree very well with those in dipeptides.$^{50}$

To better understand these results, the $\phi$ distributions of two representative cases are shown in Figure 8. From the coil library, the $\phi$ distribution of Ala has a highest peak around $-67^\circ$ and a lower shoulder around $-153^\circ$, which agree with previous work of Avbelj et al.$^{47}$ On the other hand, the coil library $\phi$ distribution of $g^+\text{Gln}$ has a much higher population around $-100^\circ$ and only one peak. (Scheme 4). This agrees with much higher $\gamma_{\text{H}_2\text{H}_2}$ of Gln dipeptide. This large difference cannot be fully reproduced by current force fields. At $\phi = -120^\circ$, both OPLS-AA/L and CHARMM22* well reproduce the coil library value for Ala but cannot fully describe the increased population for $g^+\text{Gln}$. At $\phi = -65^\circ$, ff99SB-NMR agrees with coil library results for Ala but cannot fully follow the decrease in the population for $g^+\text{Gln}$. Indeed, the coil library $\phi, \psi$ distribution of Gln (ordinary AA) differ from that of Ala with $S = 0.86$. In comparison, ff99SB variants and OPLS-AA/L give higher $S = 0.94–0.97$. Our results imply that current force fields may underestimate the backbone conformational differences between Ala and its derivatives, suggesting that same parameters on all AAs may not be enough.

### 3.5. Folding of Both $\alpha$-Helix Proteins and $\beta$-Hairpin Peptides.

The ability to fold peptides and small proteins is a stringent test of a force field, because even minor inaccuracies at single-residue level can lead to a significant perturbation of delicate balance among different structures. We carried out the folding simulations of Trp-cage80 (Tc5b, a designed 20-residue $\alpha$-helix mini-protein), Trpzip-281 (a designed tryptophan zipper $\beta$-hairpin), and GB1 hairpin82 (residues 41–56 of protein G B1 domain), using the RSFF1 and OPLS-AA/L, and the two state-of-the-art force fields83 CHARMM22* and AMBER ff99SB*. We also carried out folding simulation of a much larger three-helix bundle protein Engrailed Homeodomain (1ENH)84 using RSFF1. All folding simulations were carried out using REMD, initiated from unfolded structures.

The AMBER ff99SB*-ildn, CHARMM22* and OPLS-AA/L cannot consistently stabilize the native structures of the four systems (Figure 9) as the dominant cluster. AMBER ff99SB*-ildn simulations of the two $\beta$-hairpins gave many quite different structures without a dominant cluster, although a small fraction...
of near-native structures was found for each. Indeed, even ff99SB with β-sheet propensity higher than ff99SB* still significantly underestimate the stability of Trpzip-2. In a previous study, ff99SB simulations could not stabilize the folded structure of another β-hairpin peptide (Mbh12). However, we noticed that previous folding simulations of GB1 hairpin using ff99SB* gave the β-hairpin structure as the most populated cluster (21%). CHARMM22* can sample the native structure of the Trp-cage but did not give it as the dominant cluster in our simulation. Also, the native state of Engrailed Homeodomain is unstable in a previous MD simulation using CHARMM22*. The OPLS-AA/L force field, which share same nonbonded parameters with RSFF1, cannot fold the Trp-cage. The simulation gave different structures without a dominant cluster, and structures from most populated clusters lack regular secondary structures.

The backbone torsion parameters for non-Gly/Pro AAs in current force fields were usually parametrized on Ala residue. Especially, the AMBER ff99SB*-ildn and CHARMM22* force fields were optimized to reproduce experimental data on Ala-based peptides, intended to achieve more balanced conformational preferences. For Figure 10a, they indeed give similarly α-helicities of Ac-Ala4-NHMe, reasonably agree with experiments for $T > 300$ K. However, they give very different melting curves for Trp-cage and Trpzip-2 without any Ala residue. It seems that, at least for these systems, ff99SB*-ildn prefers α-helix structure whereas CHARMM22* prefers β-hairpin structure. Therefore, current strategy of deriving backbone correction based on Ala-based peptides cannot fully solve the secondary structure biases existed in current protein force fields.

On the other hand, RSFF1 can successfully fold the two α-helical proteins and two β-hairpins, each with the dominant cluster very similar to the experimental structure. The reliability of RSFF1 in stabilizing the native structures of various sequences may come from its ability to accurately describe different intrinsic conformational preferences of different AA residues by using the residue-specific parameters.

As shown in Figure 10, RSFF1 consistently overestabilizes the α-helical structure of Ac-Ala4-NHMe and the folded states of Trp-cage and the two β-hairpins. RSFF1 also overestabilizes the three-helix bundle Homeodomain, because a high population (80%) of folded structures is observed at temperature (330 K, the lowest replica) higher than its experimental $T_m$ (325 K). Interestingly, the $\phi, \psi$ distributions from coil library were originally used to model the denatured and intrinsically disordered peptides and proteins. However, because the underlying local interactions determining these intrinsic conformational features also exist in the folded states, the RSFF1 does not bias toward the unfolded state. On the contrary, it actually somehow overestimates the stability of the native state, which is better than uncertain secondary structure biases for some applications such as the structure prediction and refinement. The underline reason is still unknown, but it is possible that RSFF1 can be fine-tuned to achieve better agreement with experiments.

4. CONCLUSION AND OUTLOOK

In this work, we present our efforts in developing a new protein force field RSFF1, based on the $\phi, \psi$, and $\chi$ free energy surfaces (PMFs) of all 20 amino acids (AAs) from statistical analysis of protein coil library. A set of new methods has been established, by which excellent agreement can be achieved between PMFs from dipeptide simulations and the target PMFs. Especially, backbone torsion parameters, which are AA-dependent in RSFF1, can be easily optimized using our new $\phi, \psi$ decomposition approach. This work demonstrates that it is
feasible to parametrize all rotatable torsions in an all-atom force field based on free energy surfaces instead of potential energy surfaces.

During the parametrization, we found that not scaling 1-4 electrostatic interactions while significantly reducing 1-4 van der Waals (L-J) interaction is a good choice. Also, adding only three 1-5 L-J interactions (Hᵢ···Oᵢ, Hᵢ···Nᵢ₊₁, Oᵢ₋₁···Cᵢ), which are the same for all AAs, is enough for the coupling between backbone ϕ and ψ torsions. For Asp, Asn, Ser, and Thr, modifications of local polar interactions such as additional O···C attraction may also be necessary.
We show that RSFF1 gives significantly improved simulation results for a variety of peptides and proteins. It well reproduces the NMR $J$ coupling constants of AA dipeptides, better than its parent OPLS-AA/L and some recent force fields (AMBER ff99SB-ildn, ff99SB-ildn-NMR, and CHARMM22*). The different intrinsic conformational preferences of various AA residues cannot be fully captured using a single set of backbone parameters. RSFF1 can also consistently fold a set of peptides and proteins including both $\alpha$-helix (Ac-Ala14-NHMe, Trp-cage, Homeodomain) and $\beta$-sheet (Trpzip-2, GB1 hairpin) ones, with similar overstabilization. In comparison, other force fields cannot correctly fold all of them simultaneously. This indicates that RSFF1 not only achieves a good balance between $\alpha$-helical and $\beta$-sheet structures but also is transferable among different sequences. Indeed, RSFF1 can successfully fold all of the 12 small fast-folding proteins recently studied by Lindorff-Larson et al., $^3$ which will be reported elsewhere.

**ASSOCIATED CONTENT**

1. **Supporting Information**
   The detailed methods for the folding simulations, various free energy surfaces from the coil library, and RSFF1 simulations (Figure S1–S6), and all new torsion parameters in the RSFF1 force field (Tables S1, S2). This material is available free of charge via the Internet at http://pubs.acs.org. The implementation of RSFF1 in Gromacs will be provided upon request.

**AUTHOR INFORMATION**

**Corresponding Authors**
*F. Jiang: e-mail, jiangfan@pku.edu.cn.
*Y.-D. Wu: e-mail, wuyd@pkusz.edu.cn.

**Notes**
The authors declare no competing financial interest.

**Biographies**

Fan Jiang received his B.Sc. and Ph.D. from Peking University (PKU) in 2003 and 2008, respectively. He worked in Yun-Dong Wu’s lab both as a student at PKU and as a research associate (2008–2010) at HKUST. He is currently a research staff in PKU Shenzhen Graduate School. His current research focuses on developing peptide/protein simulation and structure prediction methods, especially on the efforts to merge physical-based and knowledge-based approaches.

Chen-Yang Zhou is a Ph.D. candidate in the College of Chemistry at Peking University working with Prof. Yun-Dong Wu. He received his B.Sc. in Chemistry from Peking University in 2010. His current research focuses on parametrization and validation of protein force fields.

Yun-Dong Wu is currently Chair Professor of Chemistry at Peking University. He is a member of the Chinese Academy of Science. He received his B.Sc. in chemistry from Lanzhou University in 1981, and his Ph.D. from University of Pittsburgh in 1986. After postdoctoral research with Prof. K. N. Houk at UCLA, he joined the faculty at the HongKong University of Science & Technology, becoming Chair Professor in 2007. His research focuses on understanding the mechanisms of organic reactions, molecular design with peptides, modeling of protein folding, and protein–protein interactions.

**ACKNOWLEDGMENTS**

We are grateful for the financial supports from the National Natural Science Foundation of China (Grant No. 21133002 for Y.-D.W. and 21203004 for F.J.), the Shenzhen Peacock Program (KQTD201103), and Peking University Shenzhen Graduate School.

**REFERENCES**


Conformational Preferences of Dipeptides Using Vibrational Spectroscopy.


2011-06-23 10:59:00

Amino Acid Dipeptides.


