Supporting Information for:

COFFDROP: A Coarse-Grained Nonbonded Force Field for Proteins Derived from All-Atom Explicit-Solvent Molecular Dynamics Simulations of Amino Acids

Casey T. Andrews, and Adrian H. Elcock*

Department of Biochemistry, University of Iowa, Iowa City, IA 52242

e-mail: adrian-elcock@uiowa.edu
Supporting Methods

Simulations Using The MARTINI CG Force Field

All MD simulations using the MARTINI version 2.2 force field\textsuperscript{1,2} were performed using a previously published protocol\textsuperscript{3} with GROMACS version 4.5.1.\textsuperscript{4,5} For each simulation, solutes were first placed randomly in a $35 \times 35 \times 35$ Å simulation box and then solvated with the MARTINI polarizable water model.\textsuperscript{6} As suggested by the MARTINI developers,\textsuperscript{6} a time step of 25 fs was used, van der Waals interactions were treated with a shift function that acted from 9 to 12 Å, Coulombic interactions were treated with a shift function that acted from 0 to 12 Å, and a constant dielectric of 2.5 was applied. All simulations were performed in the NPT ensemble with temperature and pressure being maintained at 298 K and 1 atm using, respectively, the Berendsen\textsuperscript{7} thermostat and barostat. Each system was first energy minimized for 1000 steps using the steepest descent algorithm, and then subjected to a 300 ps equilibration period during which the temperature was gradually increased from 50 to 298 K. After this, each system was equilibrated for a further period of 1 ns before a production simulation was carried out for 1 μs. The atomic coordinates of the solute molecules were saved every 0.1 ps for analysis. Clustering analysis followed the same protocol as described in the main text, with the exception that solute molecules were considered to be in contact if any pair of pseudoatoms were within 6.0 Å; the cutoff of 6.0 Å corresponds to the first minimum after the first major peak in the solute-solute \textit{g}(r) computed from the 300 mg/ml trp simulation.
Simulations Using The Betancourt-Omovie CG Force Field

All BD simulations using the Betancourt and Omovie nonbonded potential functions were performed using the same protocol described in the main text. The Betancourt-Omovie potential functions, which are binned in intervals of 0.15 Å, are undefined for very short-range interactions so they were first modified to include short-range steric repulsions: this was achieved by increasing the energy of each successively shorter-range bin by an additional 2 kcal/mol, leaving unchanged the energies of bins previously defined by Betancourt & Omovie. Clustering analysis followed the same protocol described in the main text, with the exception that solute molecules were considered to be in contact if any pair of pseudoatoms were within 5.0 Å; the cutoff of 5.0 Å corresponds to the first minimum after the first major peak in the solute-solute g(r) computed from the 300 mg/ml trp simulation.
References

1 Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.* 2008, 4, 819-834.

2 de Jong, D. H.; Singh, G.; Bennett, W. F. D.; Arnarez, C.; Wassenaar, T. A.; Schäfer, L. V.; Periole, X.; Tieleman, D. P.; Marrink, S. J. Improved parameters for the martini coarse-grained protein force field. *J. Chem. Theory Comput.* 2013, 9, 687-697.

3 Stark, A. C.; Andrews, C. T.; Elcock, A. H. Toward optimized potential functions for protein-protein interactions in aqueous solutions: osmotic second virial coefficient calculations using the MARTINI coarse-grained force field. *J. Chem. Theory Comput.* 2013, 9, 4176-4185.

4 van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 2005, 26, 1701-1718.

5 Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* 2008, 4, 435-447.

6 Yesylevskyy, S. O.; Schäfer, L. V.; Sengupta, D.; Marrink, S. J. Polarizable water model for the coarse-grained MARTINI force field. *PLoS Comput. Biol.* 2010, 6, e1000810.

7 Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.* 1984, 81, 3684-3690.

8 Betancourt, M. R.; Omovie, S. J. Pairwise energies for polypeptide coarse-grained models derived from atomic force fields. *J. Chem. Phys.* 2009, 130, 195103.
Supporting Figure Legends

**Figure S1**  **CG mapping scheme used in COFFDROP for amino acids.** Images show the COFFDROP CG mapping scheme in (black) ball and stick representation overlaid onto the all-atom representation of the amino acids shown in licorice representation. In the all-atom representation, hydrogen atoms have been removed for clarity.

**Figure S2**  **Bond length probability distribution using different harmonic spring constant strengths.** Plot shows the bond length probability distribution for the Cβ–Cy pseudo-bond during the simulation of leu in MD (black circles), and BD using spring constants of 10 (cyan squares), 50 (blue upward triangles), 100 (green downward triangles), 150 (yellow diamonds), and 200 (red hexagons) kcal/mol/Å².

**Figure S3**  **Average and standard deviation of angle and dihedral probability distributions calculated from MD for trp.** A. Angle probability distributions; each color and symbol represents a different angle; standard deviations were calculated by splitting the trajectory into three 333 ns blocks. B. Same as A but showing dihedral probability distributions.

**Figure S4**  **Comparison of angle and dihedral probabilities obtained from MD and BD for single amino acids.** A. Angle bin probabilities obtained from MD (x-axis) versus those obtained from BD (y-axis); symbols are shown for all angle bins for all amino acids. B. Same as A but showing dihedral bin probabilities.
Figure S5  Improper dihedral probability distributions obtained from MD and BD for ala. A. Graph showing improper dihedral probability distributions obtained from MD (lines) and IBI BD (circles) when improper dihedral potential functions are not included in BD simulations; red and blue indicate different improper dihedrals. B. Same as A but showing results obtained when including and optimizing improper dihedral potential functions.

Figure S6  Reoptimization of angle and dihedral potential functions for leu. Row A: Plots shows angle potential functions obtained from the original COFFDROP derivation (blue line), the ‘wrong’ initial potential function used for the reoptimization of the potential function (green line), and the reoptimized angle potential function (red triangles). Each panel represents a different angle. Row B: Plots show the angle probability distribution obtained from MD (blue line), obtained from the ‘wrong’ initial potential function (green line), and obtained from the reoptimized angle potential function (red triangles). Row C: Same as row A, but showing results for dihedral potential functions. Row D: Same as row B, but showing results for dihedral probability distributions.

Figure S7  Reoptimization of angle and dihedral potential functions for lys. Row A: Same as row A of Figure S6, but showing results for lys. Row B: Same as row B of Figure S6, but showing results for lys. Row C: Same as row C of Figure S6, but showing results for lys. Row D: Same as row D of Figure S6, but showing results for lys.
**Figure S8** MD simulation snapshots, converted to their CG representations, showing intramolecular nonbonded interactions in trp. **A.** Snapshot showing that a short distance for the Ace-C\(\delta\) interaction corresponds to a long distance for the Nme-C\(\delta\) interaction. **B.** Snapshot showing that a short distance for the Nme-C\(\delta\) interaction corresponds to a long distance for the Ace-C\(\delta\) interaction.

**Figure S9** Intramolecular nonbonded probability distribution for trp. **A.** Plot showing the nonbonded probability distribution obtained from MD (line) and BD (circles) for the interaction of the Ace and C\(\delta\) pseudoatoms in trp. **B.** Same as A but showing results for the interaction of Nme and C\(\delta\) pseudoatoms.

**Figure S10** Comparisons of g(r)s obtained from independent replicates of all-atom MD simulations. **Row A.** g(r) for the trp-trp system; each plot shows the g(r) computed from an independent 1 \(\mu\)s simulation using only the closest pair of heavy atoms. **Row B.** Same as for row A but showing results for the asp-glu system.

**Figure S11** Clustering of asp, cys, gly, lys, tyr, and val solutions in MD and BD. The plots show the fraction of solute molecules that are members of clusters of various sizes. Blue circles represent results from MD, green upward triangles represent results from BD using COFFDROP’s derived potential functions and red downward triangles represent results from BD using steric nonbonded potential functions.
Figure S12  Clustering of ala, leu, asn, and trp 50 and 100 mg/ml concentrations solutions in MD and BD. Same as Figure 7 of the main text but showing results for lower solute concentrations.

Figure S13  Comparison of snapshots of the 300 mg/ml trp solution obtained from MD and BD. A. All-atom snapshot from MD; hydrogen atoms removed for clarity B. Coarse-grained representation of the same snapshot shown in A. C. Coarse-grained snapshot from BD. The blue box in each image represents one simulation cell.

Figure S14  Spatial disposition of the Cδ pseudoatom of tryptophan in MD and BD. First row: Red contours show preferred locations of the Cδ pseudoatom of a trp molecule interacting with a second trp molecule (shown in black) sampled from MD; each of the panels A-C shows the same image viewed from a different orientation. Second row: same as first row but with the two trp molecules interchanged. Third row: same as first row but showing results from BD. Fourth row: same as third row but with the two trp molecules interchanged.

Figure S15  Clustering of ala, leu, asn, and trp at 50, 100, 200, and 300 mg/ml concentrations using different coarse grain force fields. The plots show the log of the fraction of solute molecules that are members of clusters of various sizes. Blue circles represent results from using COFFDROP, green upward triangles represent results from using Betancourt and Omovie’s derived potential functions and red downward triangles represent results from using the MARTINI version 2.2 force field.
Table S1  Mapping scheme listing which atoms of the all-atom model are used to determine the positions of pseudoatoms of the CG model. Heavy atom names are taken from the Amber force field. The placement of each pseudoatom was computed from the center of geometry of the listed heavy atoms. The residue Hip refers to protonated histidine.
Figure S1
Figure S2
Figure S3
Figure S4
Figure S5
Figure S6
Figure S7
Figure S8
Figure S9
Figure S10
Figure S11
Figure S12
Figure S13
Figure S14
Figure S15
| Res | Pseudoatom | Heavy atom(s) name | Res | Pseudoatom | Heavy atom(s) name |
|-----|------------|-------------------|-----|------------|-------------------|
| Ace | X1         | CH3               | Ile | Cβ         | Cβ Cγ1 Cγ2       |
| Nme | Y1         | CH3               | Ile | Cγ         | Cδ/Cδ1           |
| Ala | Cα         |                   | Leu | Cα         |                   |
| Ala | Cβ         |                   | Leu | Cβ         | Cβ               |
| Arg | Cα         |                   | Leu | Cγ         | Cγ Cδ1 Cδ2       |
| Arg | Cβ         | Cβ Cγ Cδ         | Lys | Cα         | Cα               |
| Arg | Cγ         | Ne Cζ NH1 NH2     | Lys | Cβ         | Cβ Cγ Cδ         |
| Asn | Cα         |                   | Lys | Cγ         | Cε Nζ            |
| Asn | Cβ         |                   | Met | Cα         | Cα               |
| Asp | Oγ         | Cγ Oδ1 Oδ2       | Phe | Cβ         | Cβ               |
| Cys | Cα         |                   | Phe | Cγ         | Cγ Cδ1 Cδ2 Cε1 Cε2 Cζ |
| Cys | Cβ         | Cβ Sγ             | Pro | Cα         | Cα               |
| Gln | Cα         |                   | Pro | Cβ         | Cβ Cγ Cδ         |
| Gln | Cβ         | Cβ Cγ             | Ser | Cα         | Cα               |
| Gln | Cγ         | Cδ Oε1 Nε2       | Ser | Cβ         | Cβ Oγ Cγ2        |
| Glu | Cα         |                   | Thr | Cα         | Cα               |
| Glu | Cβ         | Cβ Cγ             | Trp | Cα         | Cα               |
| Gly | Cα         |                   | Trp | Cβ         | Cβ               |
| His | Cα         |                   | Trp | Cγ         | Cγ Cδ1 Cδ2 Nε1 Cε2 |
| His | Cβ         |                   | Trp | Cδ         | Cδ2 Cε2 Cε3 Cζ2 Cζ3 CH2 |
| His | Cγ         | Cγ Cε1 Cδ2 Nδ1 Nε2 | Tyr | Cα         | Cα               |
| Hip | Cα         |                   | Tyr | Cβ         | Cβ               |
| Hip | Cβ         |                   | Tyr | Cγ         | Cγ Cδ1 Cδ2 Cε1 Cε2 Cζ OH |
| Ile | Cα         |                   | Val | Cα         | Cα               |

**Table S1**