Supplementary Information: Reweighting Configurations Generated by Transferable, Machine Learned Models for Protein Sidechain Backmapping

Jacob I. Monroe*

Department of Chemical Engineering, University of Arkansas, Fayetteville, Arkansas 72701, USA

E-mail: jacob.monroe@uark.edu

Supporting Figures and Tables

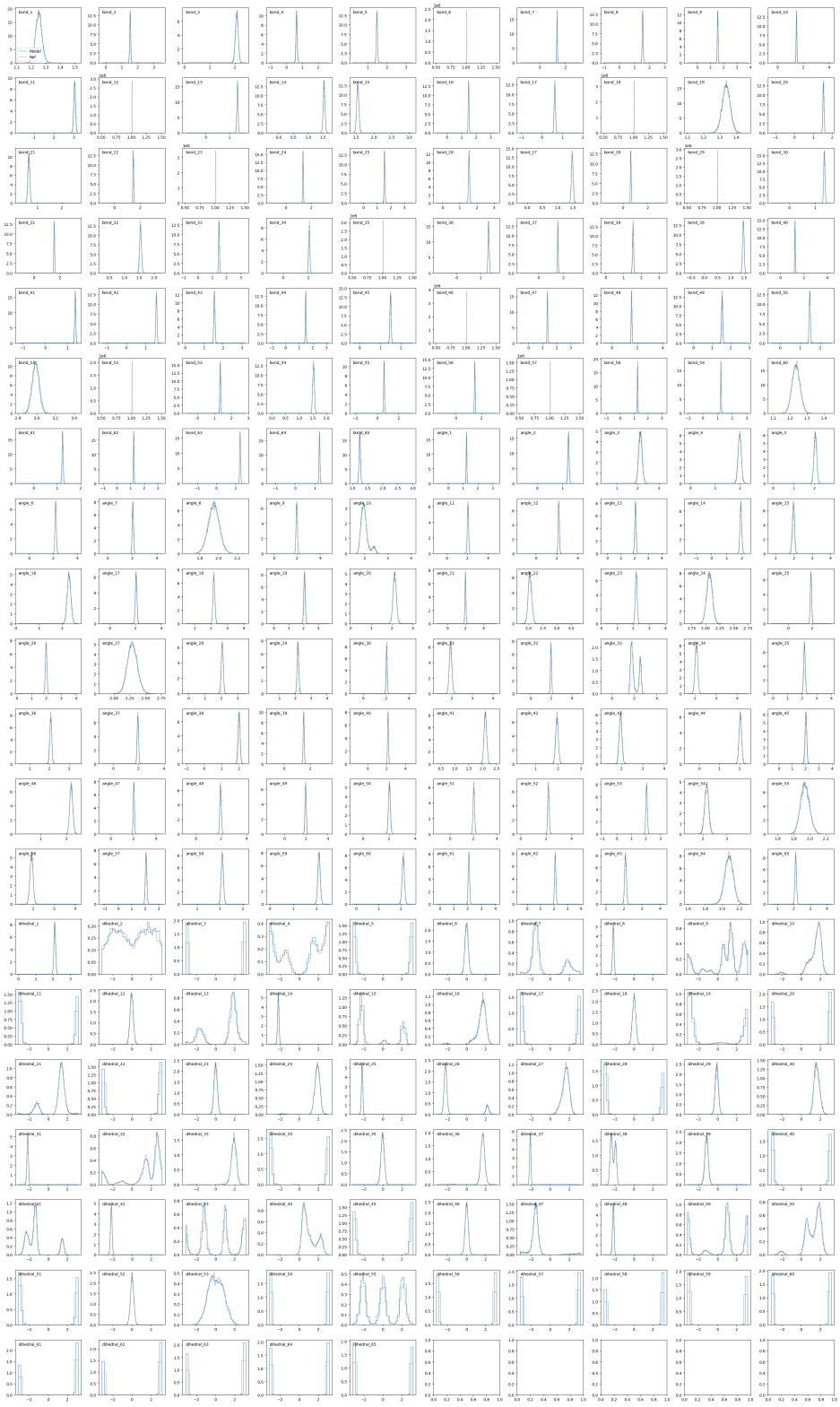

Figure S1: Distributions for all degrees of freedom sampled from the normalizing flow model trained on the CG configurations from an all-atom MD simulations are shown in blue and compared to the training data as dashed black curves.

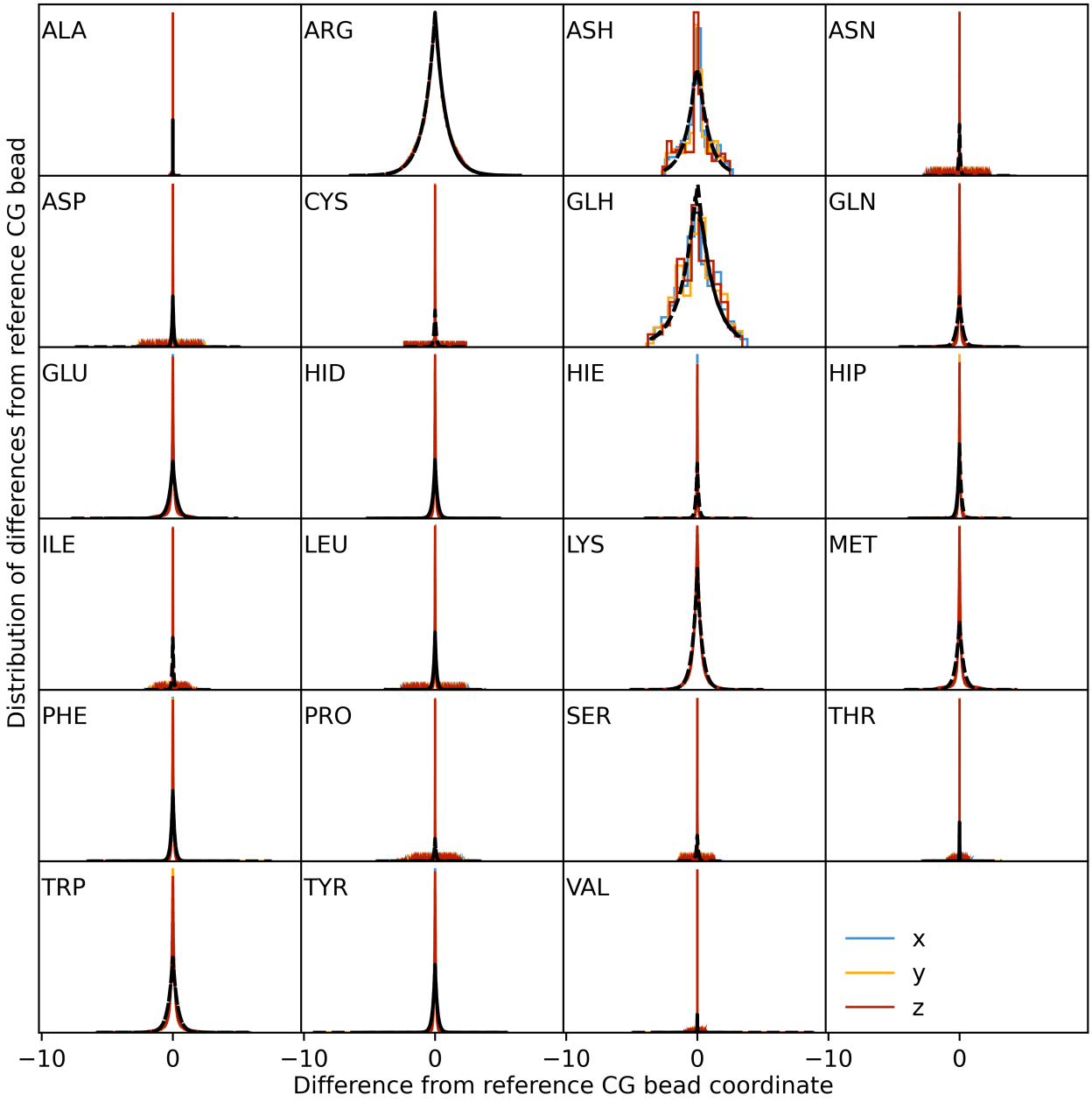

Figure S2: For each residue type, the distributions of Cartesian coordinate deviations from the backmapped CG sidechain site. Bulkier and more flexible sidechains exhibit broader distributions, as well as residues with small amounts of training data. Dashed black curves are fits of Laplace distributions with their centers fixed to zero (i.e., only the scale parameter is fit to the distributions shown). Since distributions in each Cartesian coordinate are nearly identical, they result in nearly identical scale parameters. As such, the average of the x, y, and z scale parameters is used for all Cartesian dimensions and is what is used to plot the dashed black curves and in computing $P_1(\mathbf{R}|\mathbf{r})$.

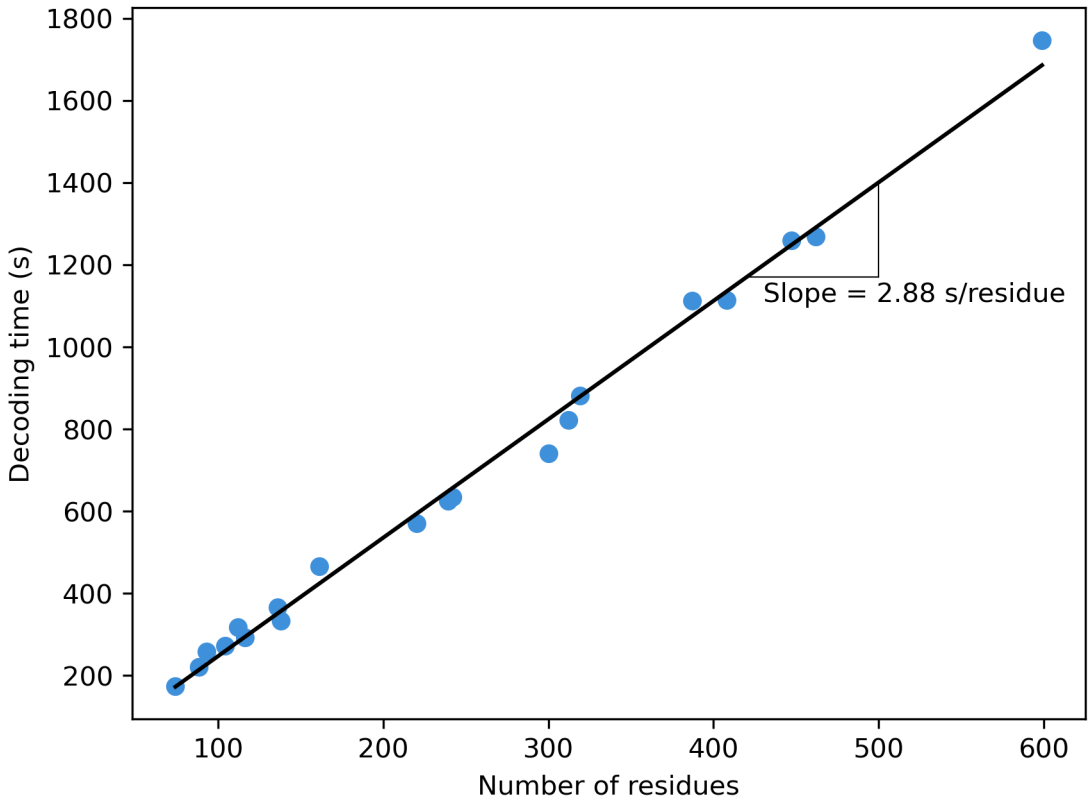

Figure S3: Computational time (in seconds) is shown for backmapping all of the proteins in the test set described in Section 3.2 of the main text. As expected, timings are linear in the length of the protein sequence at approximately 2.88 seconds per residue. For each protein backmapped, these timings apply to parallel production of 100 independent configurations.

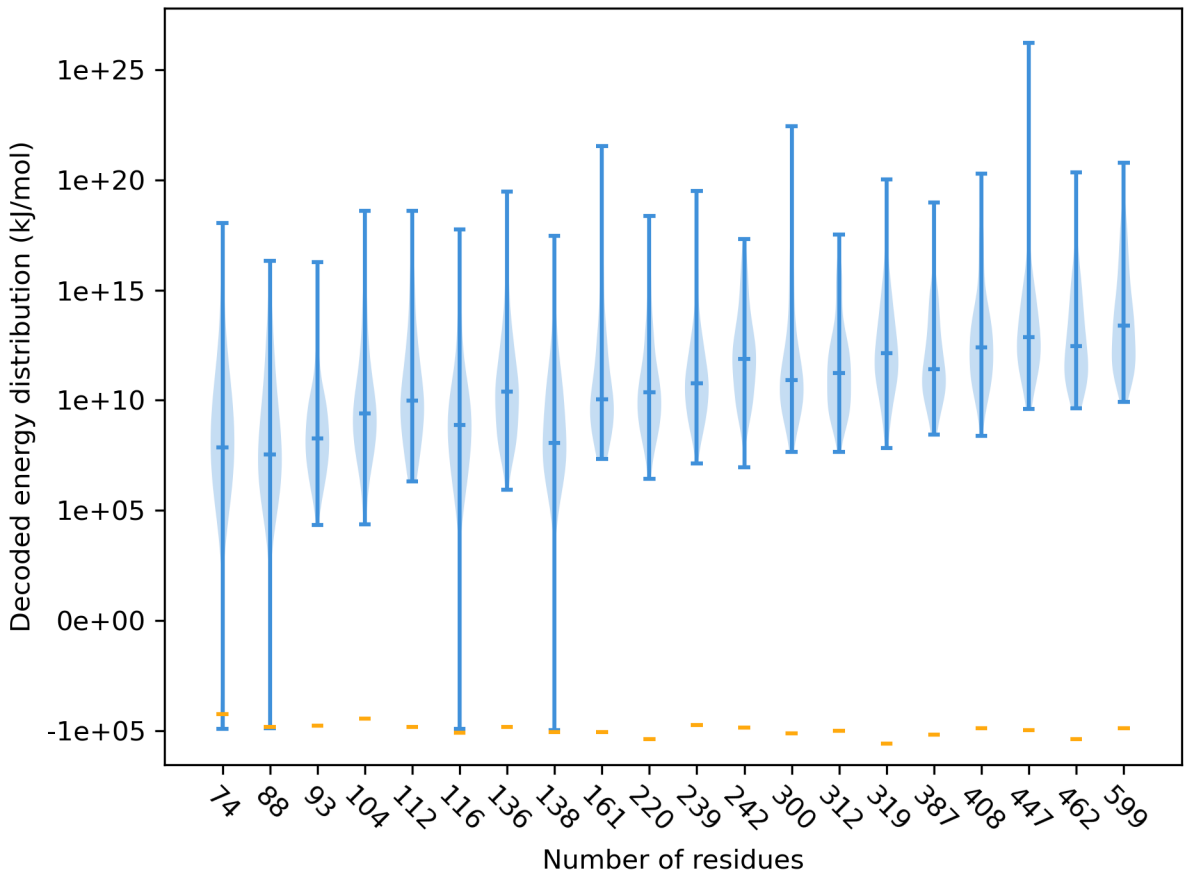

Figure S4: Distributions of potential energies evaluated using the AMBER14SB forcefield with implicit solvent for 100 backmapped configurations of each of 20 test set proteins. Labels indicate the number of residues in the protein, which demonstrates that large proteins are backmapped with higher energy.

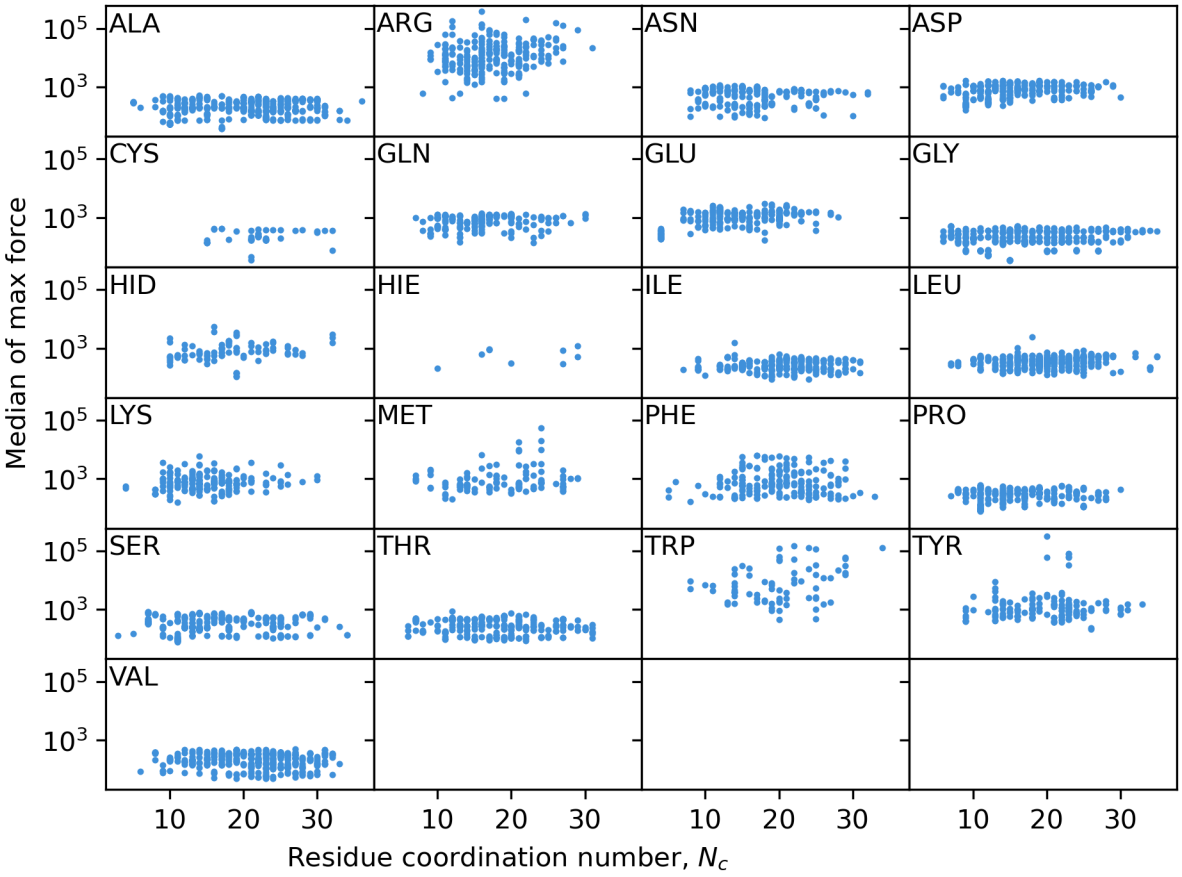

Figure S5: For each residue type, the median over 100 backmapped samples of the maximum mean force experienced by any atom in a residue is plotted against that residue's coordination in the protein test set. Note that the coordination is based on the residue's C-alpha position and does not change with the decoding.

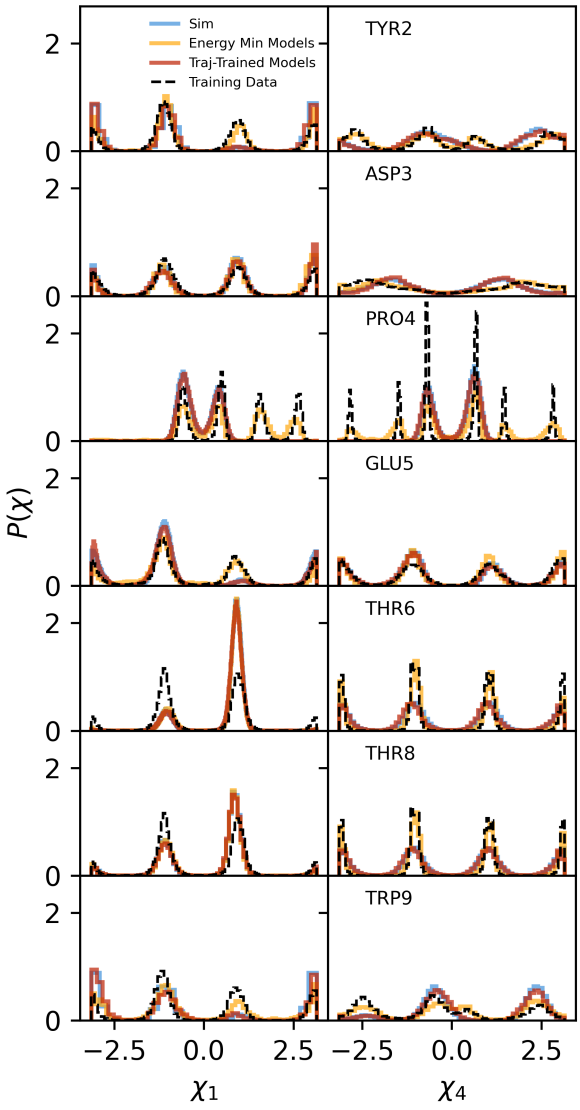

Figure S6: For each residue (excluding glycines) in chignolin, distributions of the first and fourth dihedrals are compared between the MD simulation (blue), generated configurations from models trained on energy-minimized PDB structures (orange) or trajectory-trained models (red), and the training data set (black-dashed). See Fig. 4 of the main text for a description of sidechain dihedrals 1 and 4.

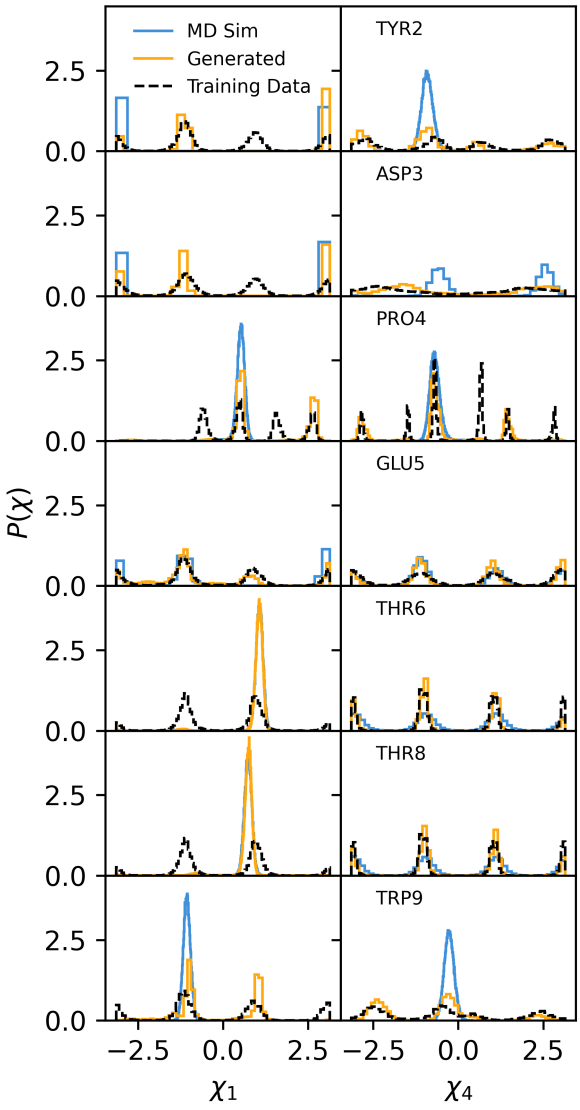

Figure S7: Dihedral distributions from restrained simulation and decoding model.

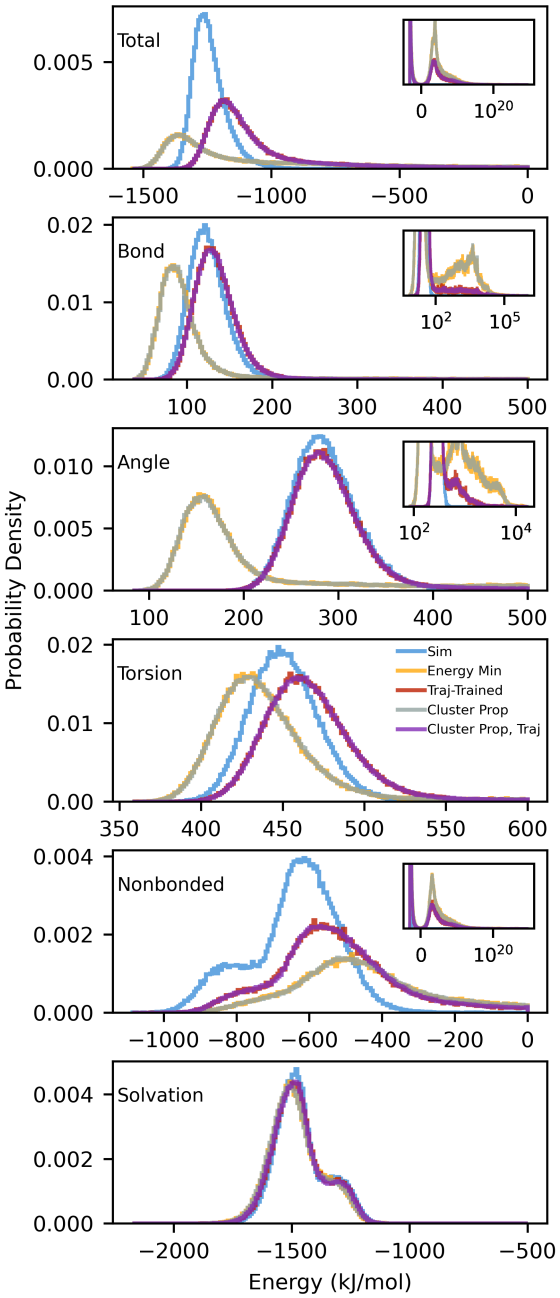

Figure S8: Probability densities of potential energies of chignolin, broken down into contributions from bonds, angles, torsions, nonbonded interactions (LJ and electrostatic), and implicit solvation for MD simulations (blue). Configurations generated from a CG representation of the all-atom MD trajectory are shown for models trained on energy minimized PDB structures (orange) or chignolin trajectories (red). Distributions for configurations generated by each model type from our proportional-clustering CG model overlap closely with configurations generated by backmapping CG representations of the all-atom MD trajectories. Primary axes are truncated to more easily compare to simulation distributions while insets show densities utilizing base-ten logarithms of the potential energy to include the full generated distributions.

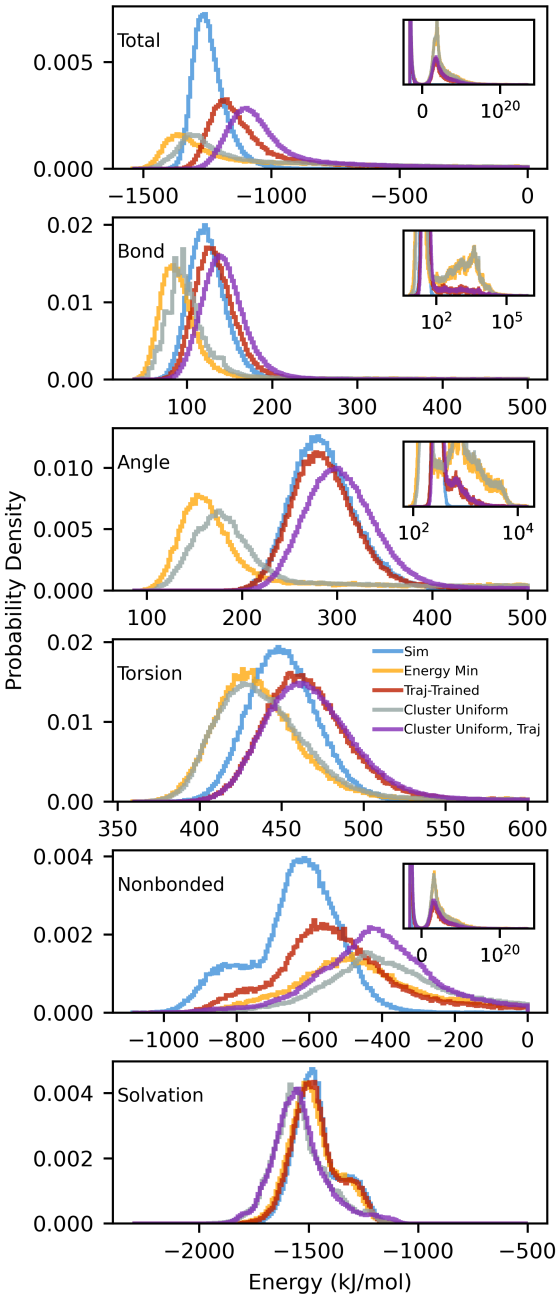

Figure S9: Probability densities of potential energies of chignolin, broken down into contributions from bonds, angles, torsions, nonbonded interactions (LJ and electrostatic), and implicit solvation for MD simulations (blue). Configurations generated from a CG representation of the all-atom MD trajectory are shown for models trained on energy minimized PDB structures (orange) or chignolin trajectories (red). Distributions for configurations generated by each model type from our uniform-clustering CG model are shifted to slightly higher values compared with configurations generated by backmapping CG representations of the all-atom MD trajectories. Primary axes are truncated to more easily compare to simulation distributions while insets show densities utilizing base-ten logarithms of the potential energy to include the full generated distributions.

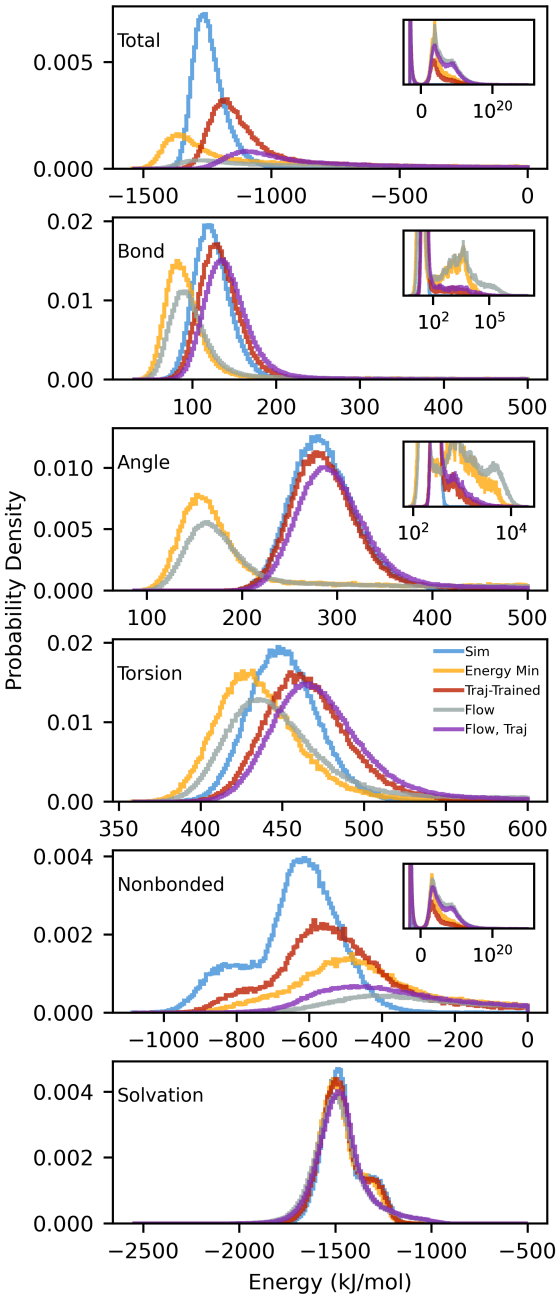

Figure S10: Probability densities of potential energies of chignolin, broken down into contributions from bonds, angles, torsions, nonbonded interactions (LJ and electrostatic), and implicit solvation for MD simulations (blue). Configurations generated from a CG representation of the all-atom MD trajectory are shown for models trained on energy minimized PDB structures (orange) or chignolin trajectories (red). Distributions for configurations generated by each model type from our normalizing flow CG model are shifted to higher values, especially for non-bonded energies. Shapes also differ compared to those generated by backmapping CG representations of the all-atom MD trajectories. Primary axes are truncated to more easily compare to simulation distributions while insets show densities utilizing base-ten logarithms of the potential energy to include the full generated distributions.

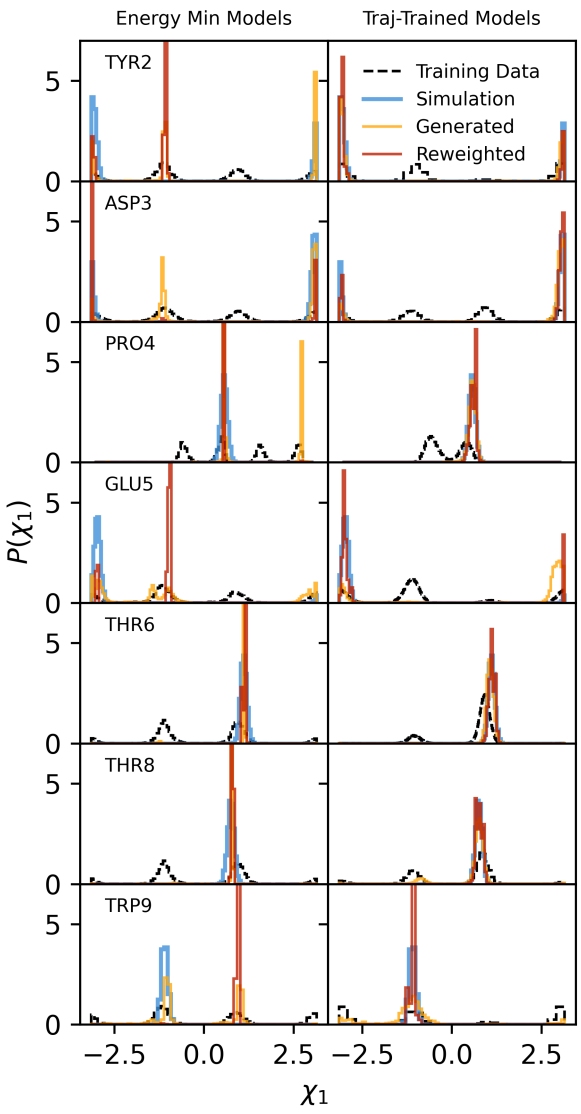

Figure S11: For all residues in chignolin, distributions of the first dihedral from the backbone (ends in the first atom bonded to the beta carbon) are shown from trajectories (blue) with tight restraints on all atoms outside the indicated residue sidechain, as well as a restraint on the sidechain atoms to their CG bead location in the energy-minimized PDB structure. Training data for each model type (based on energy-minimized PDB structures on the left or chignolin-trajectory-trained on the right) are in dashed black. Distributions from generated configurations are shown in orange, while reweightings of those distributions are shown in red.

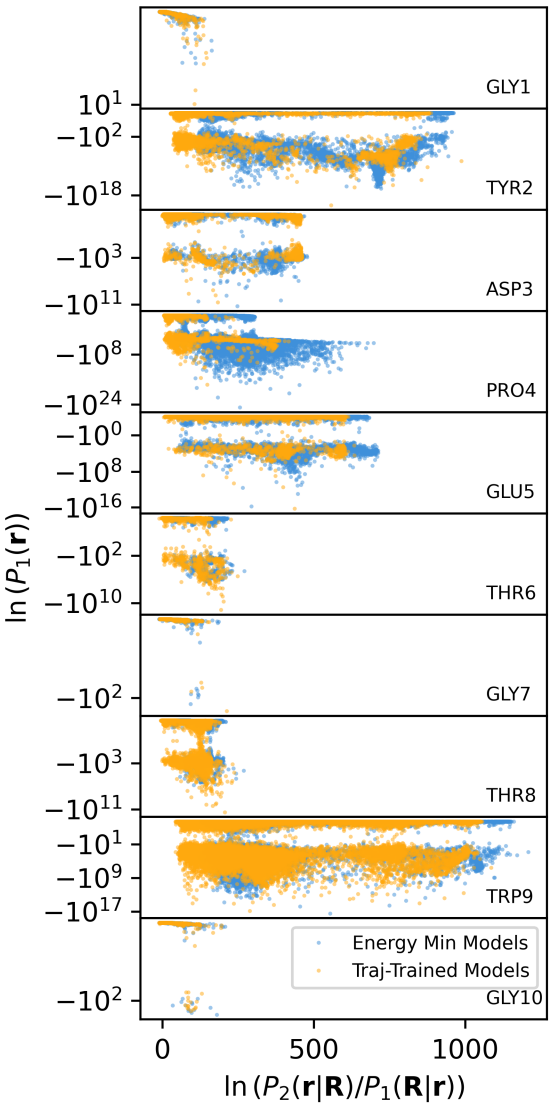

Figure S12: The unweighted log-probability in the desired ensemble (e.g., negative potential energy) is compared to the probability under the generative model for configurations of individual residue sidechains produced from a single CG configuration. Note that this means that $P_2(\mathbf{R}) = 1$.