JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

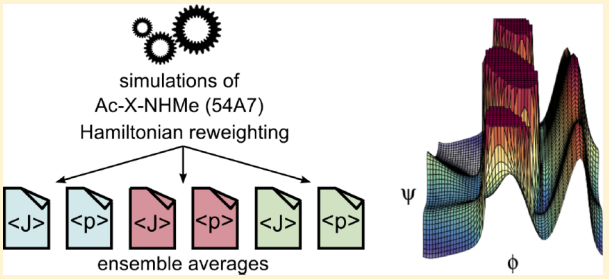
Optimization of Protein Backbone Dihedral Angles by Means of Hamiltonian Reweighting

Christian Margreitter and Chris Oostenbrink*

Institute for Molecular Modeling and Simulation, University of Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria

Supporting Information

ABSTRACT: Molecular dynamics simulations depend critically on the accuracy of the underlying force fields in properly representing biomolecules. Hence, it is crucial to validate the force-field parameter sets in this respect. In the context of the GROMOS force field, this is usually achieved by comparing simulation data to experimental observables for small molecules. In this study, we develop new amino acid backbone dihedral angle potential energy parameters based on the widely used 54A7 parameter set by matching to experimental J values and secondary structure propensity scales. In order to find the most appropriate backbone parameters, close to 100 000 different combinations of parameters have been screened. However, since the sheer number of combinations considered prohibits actual molecular dynamics simulations for each of them, we instead predicted the values for every combination using Hamiltonian reweighting. While the original 54A7 parameter set fails to reproduce the experimental data, we are able to provide parameters that match significantly better. However, to ensure applicability in the context of larger peptides and full proteins, further studies have to be undertaken.



INTRODUCTION

The functional forms that are the basis of virtually all classical force fields are very similar. In general, the interactions between atoms comprising biological macromolecules result from the electrons, which should be described with quantum mechanics. In a force field these interactions are approximated by the assignment to various well-known and simpler equations, for example, harmonic oscillators to account for covalent bonds. The parameters that are used in these simplifications, e.g., force constants and minimum distances, have been subjected to repeated adjustments. Depending on the computational resources that are available, the model (i.e., force field) gets more detailed and accurate but also much more complex. Moreover, these parameters are connected, and changes in one might require adaptation of others as well. Furthermore, as accessible simulation times increase, comparisons to more extensive experimental data become realistic, potentially requiring further updates of the force-field parameters. Another reason for the evolution of force-field parameters over time is the availability of new experimental data, which allows further testing and revalidations. In this context, we attempted to test the protein backbone dihedral angle parameters of the 54A7 parameter set[1] of the GROMOS force field against experimental data and, if necessary, to find better solutions.

In the GROMOS parametrization philosophy, the parametrization relies on the reproduction of experimental data obtained for small molecules such as ethanol, assuming their principle transferability to moieties in bigger systems such as proteins. In contrast, the torsion angles in the protein backbone have been assigned by chemical intuition and subsequent refinement at the peptide and protein level.[1] As experimental reference values, we chose NMR-derived coupling constants[2] ($^3J(\mathrm{HN},\mathrm{H}_\alpha)$) and secondary structure propensities from Raman spectroscopy for all canonical dipeptides.[3] The former have been used in this kind of analysis before,[4,5] but especially the calculation of the J value from the ϕ backbone dihedral angle using the Karplus equation is rather inaccurate. Furthermore, the coupling constants bear only very limited information on the ψ backbone dihedral angle. Thus, the inclusion of the secondary structure information from the latter is a highly relevant addition.

As will become clear, the backbone parameters currently in use are not suited to describe the characteristics of dipeptides, failing in both dimensions of our target values. In order to find proper parameters, we screened close to 100 000 parameter combinations for the 20 amino acids in question, thereby covering a wide range of possible solutions. It is noteworthy that we did not bias our search space except by the selection of the initial parameters. Because of the vast number of combinations considered, we used Hamiltonian reweighting to predict the likely outcome of an actual simulation for a given set of parameters. Provided that the initial simulations using the 54A7 parameters show a sufficient overlap with the (hypothetical) target ones, it is possible to quantitatively predict the

Received: July 10, 2016

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

ensemble averages of structural properties that are to be matched to the experiments.

METHODS

Model Systems. In the current work, we have focused on the backbone dihedral angles of the blocked amino acids, which are formed by blocking the termini of a single amino acid through acetylation of the N-terminus and N-methylation at the C-terminus. Because this compound now has two dipeptide bonds, it is commonly called a "dipeptide". Figure 1 shows the alanine dipeptide. The blocking of the termini is usually applied to exclude interactions between the charged ends and the side chain of the central amino acid, especially in the case of charged side chains. However, it has been reported[6] that blocking with an acetyl group and an *N*-methyl group, respectively, does not significantly change the intrinsic propensities, which in turn means that the influence of charged ends is negligible. However, to stay as close to the experiments as possible, we added an acetyl and *N*-methyl groups to the amino acids as well. The protonation states of the amino acids were set to match these in the respective experiments: histidine, lysine, and arginine were positively charged while aspartate and glutamate were negatively charged.

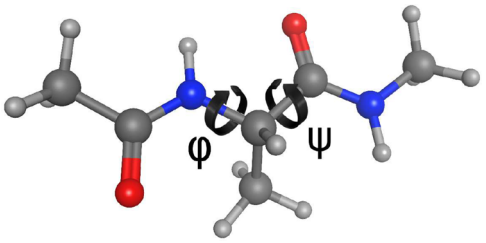


Figure 1. Graphical representation of the two backbone dihedral angles in the model systems. The angles ϕ and ψ are defined by atoms C−N−C_α−C and N−C_α−C−N, respectively. It should be noted that in accordance with the experimental studies, the amino acids have been blocked: an acetyl group at the N-terminus and a methyl moiety at the C-terminus are used to ensure noncharged ends.

Simulation Setup. The simulations were performed using the GROMOS simulation package[7,8] with the 54A7 parameter set[1] of the GROMOS force field or using the backbone parameters reported in Table 1 implemented in the 54A7 set. The force-field parameters for the blocked termini are listed in Table S1. The compounds were placed in a periodic cubic water box in the absence of counterions. The water boxes were initialized with a 1.4 nm minimum distance of the solute to the box walls. Prior to the production simulations, the systems were equilibrated from 60 to 300 K in five discrete steps with a simulation length of 20 ps each. All of the simulations were carried out at 300 K and a constant pressure of 1 atm unless explicitly stated differently. A weak thermostat coupling with two baths for the solute and solvent (τ = 0.1) and a weak barostat coupling (relaxation time of 0.5 ps and an isothermal compressibility of 4.575 × 10^{-4} (kJ·mol^{-1}· nm^{-3})$^{-1}$) were applied.[11] The SHAKE algorithm[12] was used to maintain the bond distances at the energy minimum. An integration time step of 2 fs was used, and the configurations were stored every picosecond. Interactions within 0.8 nm were calculated at every time step from a pair list that was updated every five steps. Intermediate-range interactions up to a distance of 1.4 nm were calculated at pair list updates and kept constant between updates. Long-range interactions were approximated with a reaction field[13] contribution to the energies and forces, accounting for a homogeneous medium with relative dielectric constant[14] of 61 beyond the cutoff of 1.4 nm. The lengths of the trajectories were 100 ns. In all cases, 100% of the trajectories were used for analysis. All of the graphs showing the results were composed using the R packages MDplot (http://cran.r-project.org/package=MDplot) and RPrometheus.

Comparison to Experimental Data. The first step in remodeling the energy potentials of the backbone dihedral angles in amino acids is the sound selection of experimental data to be used for comparison. In the case of this study, we chose the *J*-value measurements of the "dipeptide" (blocked amino acid) series published by Avbelj and co-workers.[2] This observable represents a time average dependent on the ϕ angle in the protein backbone. In conjunction, we also included the propensity scale subsequently published by Grdadolnik and co-workers[3] as our experimental target values (see Table S2). These two studies form a consistent set, since the *J* values from the first study were used to calibrate the Raman spectroscopy measurements in the latter.

In general, there are two ways to define and distinguish secondary structure elements in peptides and proteins. First,

Table 1. Backbone Parameters of the GROMOS Force Field and Those of the Suggested Sets (All Other Combinations Mentioned in the Text Are Provided in Table S5)

	backbone angle potential energy functions						
	ϕ			ψ			
combination	*K* [kJ/mol]	shift [deg]	mult.	*K* [kJ/mol]	shift [deg]	mult.	description
45A3[a]/53A6[a]	1.0	180.0	6	1.0	0.0	6	see refs 9 and 10
54A7/54A8	2.8	0.0	3	3.5	180.0	2	see ref 1
	0.7	180.0	6	0.4	0.0	6	
#81883	1.0	180.0	2	3.0	180.0	2	glycine
	3.0	180.0	1	5.0	180.0	1	
#12572	5.0	0.0	3	1.0	180.0	2	alanine
	5.0	180.0	6	1.0	180.0	3	
#5623	3.0	180.0	2	1.0	0.0	1	common amino acids
	5.0	0.0	3	1.0	180.0	3	
#86516	3.0	0.0	3	5.0	0.0	1	C_β-branched
	5.0	180.0	6	5.0	0.0	2	

[a]For these parameter sets, only one potential energy term is used to describe the ϕ and ψ backbone dihedral angles.

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

the hydrogen-bonding pattern of the backbone is often typical and is used for classification, for example, by the widely used Define Secondary Structure of Proteins (DSSP)[15] algorithm. However, this approach requires a minimum fragment length of at least four amino acids, which is needed in order to recognize, e.g., an α-helical conformation. For this reason, this approach cannot be used for the small compounds investigated in this study. The second possibility is to make use of the dihedral angles comprising the Ramachandran plots, where certain areas indicate backbone conformations associated with secondary structure elements.[16] Recently, the DIhedral-based Segment Identification and CLassification (DISICL)[17,18] toolbox has been developed by our group. DISICL is able to process and annotate both proteins and nucleic acids in terms of their secondary structure on the basis of two consecutive pairs of dihedral angles. Because the current systems investigated are very small, we simplified the DISICL secondary structure region definitions (see Table S3). Consequently, in the current work we made no distinction between different types of helices (α, 3_{10}, and π helices), whose basins in the Ramachandran plot are next to each other. This simplification allowed comparisons to the experimental data used as a target. It should be noted that other secondary structure elements were neglected because they were not covered by the experimental study used. In contrast to Hollingsworth's original definitions,[19] we defined the ϕ angle representing the border between the β and P_{II} basins to have the value −100°, which is in closer agreement with the average of the basin centers in the experimental study of Grdadolnik et al.[3] and the observed distributions. Moreover, this value is a compromise of the values used in the literature[20−23] (ranging from −110° to −90°).

Since the Raman experiments used as a target are based on a fitting procedure reporting propensities that sum to 100%, we normalized our results to account for that. This approach avoids locking our systems in the three basins, where other or unclassified regions in the Ramachandran plot might also be visited occasionally, and allows for better comparability. However, in all cases but glycine, which is further discussed below, this renormalization did not affect the best set of parameters, nor did it influence the overall score significantly. Unclassified regions typically show occupancies below about 5%.

One critical experimental observable in the context of investigations on small peptide systems is $^3J(\mathrm{HN},\mathrm{H}_\alpha)$. This coupling constant as determined by NMR spectroscopy represents an ensemble average and is connected to the HN−N−C_α−H_α dihedral angle. It can be calculated according to the Karplus equation (eq 1):

$$^3J(\mathrm{HN},\mathrm{H}_\alpha) = A\cos^2(\phi - 60°) - B\cos(\phi - 60°) + C \quad (1)$$

However, the results depend strongly on the choice of the three parameters A, B, and C in this equation. We chose the Pardi parameters (A = 6.4, B = −1.4, and C = 1.9),[24] which have been used previously for the protein backbone in the context of simulation studies using the GROMOS force field.[25,26] Since H_α is not represented by an actual atom in the GROMOS force field, we approximate the HN−N−C_α−H_α dihedral angle by subtracting 60° from ϕ, making use of the planarity of the peptide bond and the tetrahedral coordination at C_α. It is noteworthy that comparisons of J values from experiments to their equivalents calculated from molecular dynamics simulations are only possible with limited accuracy[27] (typically about 0.5 Hz).

In order to compare the experimental results to our predicted propensities and J values, the deviations of the J values (given in Hz) and the percentages (given in fractional propensities) are weighted equally, meaning that the deviations of the J values and the three secondary structure basins are summed up (see below).

Hamiltonian Reweighting. As will be seen in the Results and Discussion, the 54A7 parameter set did not reproduce the experimentally determined propensities and J values very well. We therefore attempted a systematic reparametrization of the backbone dihedral angle parameters.

In order to determine the effect various torsional angle parameters have on the properties used for calibration, the simplest approach is to perform multiple molecular dynamics simulations in a brute-force fashion. However, the potential number of combinations is vast (roughly 100 000 combinations; see Table 2), and this is beyond reach. Instead, a one-step perturbation protocol was used to predict the observables of interest from a single simulation using the current GROMOS 54A7 parameter set. The overall workflow is described in Figure 2. First, the time series of the experimental observables, i.e., the J value and the propensities as well as the dihedral angles ϕ and ψ (see Figure 1), were calculated from simulations performed for the 20 canonical amino acids using the 54A7 parameter set. Afterward, the time series of the Hamiltonian for a given parameter set as well as the reference (54A7) values were calculated on the basis of the backbone dihedral angles as well. By reweighting the ensemble to the updated Hamiltonian, we were able to project the ensemble average values for the individual quantities, Q, according to eq 2:

$$\langle Q\rangle_\mathrm{A} = \frac{\langle Q\, e^{-(H_\mathrm{A}-H_\mathrm{R})/k_\mathrm{B}T}\rangle_\mathrm{R}}{\langle e^{-(H_\mathrm{A}-H_\mathrm{R})/k_\mathrm{B}T}\rangle_\mathrm{R}} \quad (2)$$

where H_A and H_R represent the Hamiltonians of parameter set A and the reference parameters R, respectively, $k_\mathrm{B}T$ is the Boltzmann constant multiplied by the absolute temperature, and the angular brackets indicate ensemble averages obtained from simulations using the reference Hamiltonian (R) or predicted for the parameter set (A). The obtained $\langle Q\rangle_\mathrm{A}$ values were compared to the target values and ranked by their experimental match. Equation 2 goes back to the umbrella

Table 2. All Combinations Used for the Reweighting Workflow[a]

	force constants	shifts	multiplicities
ϕ	{1, 3, 5}	{0, 180}	{1, 2, 3, 6}
	{0, 1, 3, 5}	{0, 180}	{1, 2, 3, 6}
ψ	{1, 3, 5}	{0, 180}	{1, 2, 3, 6}
	{0, 1, 3, 5}	{0, 180}	{1, 2, 3, 6}

[a]Every angle is described by either one or two potential energy functions. The possible values for the parameters of eq 3 are given in braces. We accounted for the possibility that one potential energy term might suffice by adding a force constant of zero to the second potential energy terms. The total number of unique combinations sums up to 97 344. This number consists of $(3^2\cdot2^2\cdot4^2)/2$ = 288 combinations with two potential energy terms and $(3^1\cdot2^1\cdot4^1)$ = 24 combinations with one potential energy term, which sum to give 312 combinations for each of the angles. Combining these parameter combinations for both the ϕ and ψ angles leads to 312^2 = 97 344 combinations in our set.

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

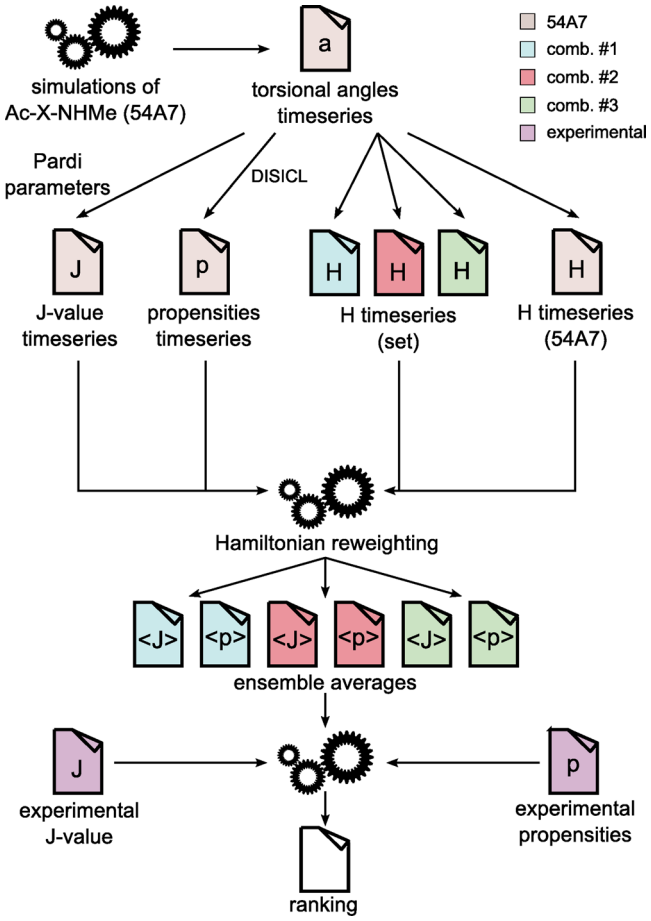


Figure 2. Workflow applied for Hamiltonian reweighting. From the configurations sampled in the simulations, the ensemble averages of dependent observables were predicted. For the calculations of the J values and Hamiltonians, see eqs 1 and 3, respectively. The considerations regarding the ranking of solutions are further described in Ranking.

sampling technique of Torrie and Valleau[28] and is widely used to obtain ensemble averages for Hamiltonians that are slightly different from the simulated one.

The actual calculation of the Hamiltonians from the dihedral angle time series (Figure 2) in our case consists of summing over the four backbone dihedral angle energy potentials (eq 3):

$$H = \sum_{i=0}^{n} k_i[\cos(s_i)\cos(m_i\phi) + 1] + \sum_{j=0}^{r} k_j[\cos(s_j)\cos(m_j\psi) + 1] \quad (3)$$

in which the variables k, s, and m are the force constants (either 1, 3, or 5 kJ/mol), the shifts (either 0° or 180°), and the multiplicities, i.e., the number of minima (either 1, 2, 3, or 6), respectively. Since multiple torsional potential energy terms per angle are possible, the variables n and r represent the numbers of individual ϕ and ψ potential energy terms, respectively. Our set of potential combinations was built by one or two torsional potentials per angle (see Table 2). For every combination, the time series of the Hamiltonian was calculated using the combination's parameters, and the ensemble average of the observables Q was calculated using eq 2.

It has been shown that these predictions are quite accurate (even quantitatively) provided that a significant part of the sampled configurations are relevant both for the reference and target ensembles.[29] In order to ensure that this approach is accurate enough, we predicted the J values and propensities for three randomly chosen force-field parameter sets using our 54A7 trajectories (see Figure S1). It is noteworthy that we are able to project the values to accuracies within a few percent even when different J values and secondary structure preferences are to be predicted. Moreover, we also report the actual simulated values for the final selected parameter sets for comparison.

Ranking. The vast number of potential candidates for the backbone parameters (see Table 2) and the multidimensional optimization space make the identification of a single best combination highly unlikely. Thus, we instead aimed for a good compromise between contrary optimization goals to ensure applicability in different contexts. Our selection protocol relied on the prediction of the J value and the propensities for every combination based on the 54A7 trajectories. This task was performed by an R library developed for this purpose. Multidimensional optimization invariably requires a weighting of the various properties. Here we defined the overall deviation Δ according to eq 4,

$$\Delta = w_J|\Delta J| + w_\alpha|\Delta P_\alpha| + w_\beta|\Delta P_\beta| + w_{P_{II}}|\Delta P_{P_{II}}| \quad (4)$$

in which ΔJ, ΔP_α, ΔP_β, and $\Delta P_{P_{II}}$ are the deviations of the J value (in Hz) and the α, β, and P_{II} propensities, respectively, and w_J, w_α, w_β, and $w_{P_{II}}$ are the corresponding weights. Unless stated otherwise, we used $w_J = 1\ \mathrm{Hz}^{-1}$ and $w_\alpha = w_\beta = w_{P_{II}} = 1$.

The candidate combinations of dihedral angle parameters were subsequently sorted to identify a promising subset of the best (approximately) 100 results. The mean values of Δ over multiple amino acids were taken when appropriate. These were plotted in a scatter plot (exemplified by Figure S2). Usually, an "arch" is obtained, as no combination perfectly reproduces both experiments simultaneously. In addition, we tried to find solutions that optimized all of the amino acids within a subgroup using bar plots with detailed information. For the remaining candidates, the potential energy functions were plotted to identify and exclude those with very high energy barriers (roughly over 20 kJ/mol). The candidate sets were further analyzed by predicting a low-resolution Ramachandran distribution (with 15 × 15 = 225 bins) using eq 2 in order to visualize any uncommon peaks or very narrow distributions (Figure S3). In our opinion, low and smooth energy surfaces and broader distributions are generally preferable for these backbone parameters because they allow for a certain freedom compared with those "locking" the backbone very tightly into a few narrow minima. Therefore, in cases where our other criteria matched (almost) equally well, we opted for these alternatives.

There are two special amino acids that should be considered more closely: glycine and proline. For glycine, the experimental propensities are not necessarily representative of the real distribution because of its known broad sampling of the Ramachandran plot.[30] Hence, its renormalized propensity values differ tremendously from the calculated ones. Moreover, glycine is known to show a strong bias toward the left-helical region of the Ramachandran plot. In this case, we aimed for a rather homogeneous distribution of the ϕ and ψ angles, in agreement with distributions observed in the Protein Data Bank (PDB) (see ref 30). Since no data were available for proline, we

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

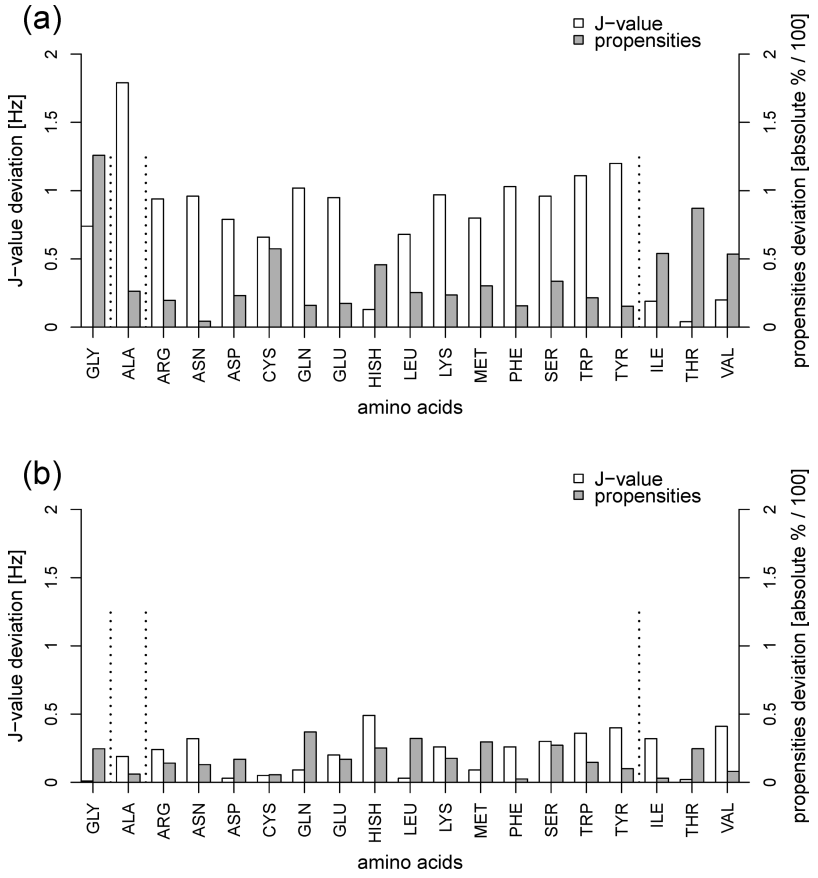


Figure 3. Deviations between simulated and experimental data using (a) the 54A7 parameters and (b) our suggested set. White bars indicate the deviations of the *J* values (left axis), and shaded bars indicate deviations in the secondary structure propensities (right axis). Dotted lines indicate the subgroups used in the optimization.

decided to apply the parameters retrieved for the common amino acids and check the resulting distribution afterward, i.e., proline was not included in the optimization procedure. Furthermore, histidine was excluded from the ranking procedure because the experimental data were reported to be rather uncertain.[2,3] Including histidine in the ranking only slightly shifted the overall results (not shown).

RESULTS AND DISCUSSION

Our analysis shows that the GROMOS 54A7 force-field parameter set fails to reproduce the experimental studies in terms of $^3J(\mathrm{HN},\mathrm{H}_\alpha)$ and secondary structure element propensities in the context of blocked amino acids (see Figure 3A and the lines marked with 54A7 in Tables 3 and 4). The average deviation in *J* values amounts to 0.8 Hz, and the propensities are off by a total of 0.34 (34%). Even when the relatively high uncertainty in comparisons of experimental and Karplus-derived *J* values is considered, the results for 54A7 are quite poor. Half of the amino acids are close to or more than 1 Hz off the respective target values (see Table 3). From a linear regression model, the R^2 for the correlation between the experimental data and the data obtained with 54A7 is 0.404, and removing alanine, the major outlier, improves this only to 0.606 (see Figure S4). The match of secondary structure propensities is better, and the deviations arise mainly from a shifted ratio between the strongly populated β and P_{II} basins and a few cases where artificially high α-helical propensities also play a significant role (e.g., threonine and cysteine). This mismatch against the experimental data is hardly surprising since 54A7 was not parametrized against such data and uses one set of protein backbone potentials for all of the amino acids for the sake of simplicity.

As outlined in Methods, we subsequently embarked on a reparametrization effort. Because of the large number of potential combinations (Table 2), a predictive method (Hamiltonian reweighting; see Methods) was used to estimate the average *J* values and propensities rather than trying to simulate all of the dihedral angle potential energy combinations. The accuracy of this approach was ensured by testing against data obtained from actual simulations for three different, randomly chosen combinations (see Figure S1) for the dialanine peptide and by performing simulations for all of the finally suggested combinations for all of the amino acids. The predicted and simulated values are reported in Table 3 and show a quantitative match in both the *J* values and the propensities. As elaborated in Methods, our selection was biased toward combinations with wide distributions in the respective basins. For asparagine and glutamine, however, the predictions may be slightly biased because in 54A7 both amino acids show no α population at all (experimental estimates amount to 2% and 8%, respectively). Since no configurations are available to be reweighted, our prediction is off in this respect. This shows the general limitation of our approach: if there is no significant overlap between the ensembles for the reference and target states, the prediction is naturally very poor. However, it is noteworthy that the actual simulation of glutamine using the suggested combination #5623 actually leads to better agreement with experiment. The other amino

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. Detailed Results of the Predicted and Simulated Parameter Combinations: Deviations with Respect to the Experimental Target Values Are Reported in Parentheses, with Negative (Positive) Values Indicating That the Computed Values Are Too High (Too Low); The Combinations Given in Italics Indicate the Best Hits When the Amino Acids Were Optimized Individually and Thus Give the Best Hits Possible with Our Screening Set of Parameters

			propensities [%/100]			Δ^a	
		J value [Hz]	P_α	P_β	$P_{P_{II}}$	abs.	ren.
GLY	54A7[b]	5.1 (+0.7)	0.014 (+0.646)	0.241 (−0.121)	0.076 (+0.144)	1.651	1.976
	#81883	5.8 (+0.0)	0.210 (+0.450)	0.062 (+0.058)	0.032 (+0.188)	0.736	0.270
	#81883[b]	5.9 (+0.0)	0.221 (+0.439)	0.058 (+0.062)	0.030 (+0.190)	0.701	0.256
	#5623	4.4 (+1.4)	0.002 (+0.658)	0.202 (−0.082)	0.008 (+0.212)	2.362	3.076
ALA	54A7[b]	7.9 (−1.8)	0.111 (−0.001)	0.388 (−0.098)	0.444 (+0.156)	2.045	2.048
	#12572	6.3 (−0.2)	0.114 (−0.004)	0.259 (+0.031)	0.592 (+0.008)	0.283	0.283
	#12572[b]	6.2 (−0.1)	0.127 (−0.017)	0.239 (+0.051)	0.597 (+0.003)	0.211	0.224
	#5623	7.1 (−1.1)	0.039 (+0.071)	0.530 (−0.240)	0.375 (+0.225)	1.616	1.623
ARG	54A7[b]	7.8 (−0.9)	0.156 (−0.086)	0.291 (+0.099)	0.497 (+0.043)	1.168	1.131
	#5623	7.1 (−0.2)	0.067 (+0.003)	0.306 (+0.084)	0.583 (−0.043)	0.370	0.380
	#5623[b]	7.1 (−0.2)	0.080 (−0.010)	0.303 (+0.087)	0.565 (−0.025)	0.362	0.380
	#67656	6.9 (+0.0)	0.116 (−0.046)	0.259 (+0.131)	0.392 (+0.148)	0.355	0.192
ASN	54A7[b]	8.4 (−1.0)	0.000 (+0.020)	0.590 (−0.010)	0.387 (+0.013)	1.003	1.008
	#5623	7.8 (−0.3)	0.000 (+0.020)	0.613 (−0.033)	0.346 (+0.054)	0.437	0.448
	#5623[b]	7.8 (−0.3)	0.000 (+0.020)	0.619 (−0.039)	0.341 (+0.059)	0.438	0.450
	#96795	7.5 (+0.0)	0.000 (+0.020)	0.600 (−0.020)	0.389 (+0.011)	0.061	0.064
ASP[c]	54A7[b]	7.7 (−0.8)	0.093 (−0.043)	0.324 (+0.136)	0.517 (−0.027)	0.996	1.017
	#5623	7.0 (−0.1)	0.045 (+0.005)	0.372 (+0.088)	0.533 (−0.043)	0.216	0.222
	#5623[b]	7.0 (+0.0)	0.056 (−0.006)	0.354 (+0.106)	0.533 (−0.043)	0.185	0.199
	#31541	7.0 (−0.1)	0.045 (+0.005)	0.383 (+0.077)	0.541 (−0.051)	0.213	0.217
CYS	54A7[b]	8.0 (−0.7)	0.179 (−0.149)	0.381 (+0.159)	0.393 (+0.037)	1.005	0.976
	#5623	7.3 (+0.0)	0.055 (−0.025)	0.488 (+0.052)	0.408 (+0.022)	0.139	0.096
	#5623[b]	7.3 (+0.1)	0.038 (−0.008)	0.488 (+0.052)	0.427 (+0.003)	0.113	0.106
	#50989	7.3 (+0.0)	0.029 (+0.001)	0.531 (+0.009)	0.424 (+0.006)	0.026	0.012
GLU[c]	54A7[b]	7.6 (−1.0)	0.126 (−0.076)	0.264 (+0.096)	0.559 (+0.031)	1.153	1.116
	#5623	6.9 (−0.2)	0.057 (−0.007)	0.263 (+0.097)	0.644 (−0.054)	0.378	0.394
	#5623[b]	6.8 (−0.2)	0.046 (+0.004)	0.269 (+0.091)	0.652 (−0.062)	0.357	0.368
	#87099	6.6 (+0.1)	0.091 (−0.041)	0.224 (+0.136)	0.435 (+0.155)	0.382	0.192
GLN	54A7[b]	8.2 (−1.0)	0.000 (+0.080)	0.489 (−0.009)	0.491 (−0.051)	1.160	1.180
	#5623	7.5 (−0.4)	0.000 (+0.080)	0.499 (−0.019)	0.468 (−0.028)	0.507	0.540
	#5623[b]	7.1 (+0.1)	0.064 (+0.016)	0.295 (+0.185)	0.597 (−0.157)	0.448	0.458
	#83856	7.1 (+0.0)	0.000 (+0.080)	0.462 (+0.018)	0.535 (−0.095)	0.193	0.194
HIS[c]	54A7[b]	8.0 (−0.1)	0.211 (−0.171)	0.331 (+0.249)	0.397 (−0.017)	0.567	0.585
	#5623	7.4 (+0.5)	0.078 (−0.038)	0.383 (+0.197)	0.470 (−0.090)	0.835	0.848
	#5623[b]	7.4 (+0.5)	0.060 (−0.020)	0.421 (+0.159)	0.446 (−0.066)	0.735	0.742
	#89861	7.9 (+0.0)	0.040 (+0.000)	0.250 (+0.330)	0.154 (+0.226)	0.586	0.130
ILE	54A7[b]	7.5 (−0.2)	0.118 (−0.098)	0.237 (+0.283)	0.603 (−0.143)	0.714	0.735
	#86516	7.6 (−0.3)	0.005 (+0.015)	0.502 (+0.018)	0.452 (+0.008)	0.311	0.299
	#86516[b]	7.7 (−0.3)	0.015 (+0.005)	0.509 (+0.011)	0.427 (+0.033)	0.369	0.350
	#5623	6.7 (+0.6)	0.042 (−0.022)	0.145 (+0.375)	0.797 (−0.337)	1.374	1.386
	#28191	7.3 (+0.0)	0.016 (+0.004)	0.455 (+0.065)	0.504 (−0.044)	0.123	0.124
LEU	54A7[b]	7.6 (−0.7)	0.174 (−0.074)	0.214 (+0.136)	0.562 (−0.012)	0.902	0.93
	#5623	6.9 (+0.0)	0.077 (+0.023)	0.213 (+0.137)	0.667 (−0.117)	0.277	0.294
	#5623[b]	6.9 (+0.0)	0.047 (+0.053)	0.231 (+0.119)	0.683 (−0.133)	0.335	0.352
	#9315	6.9 (+0.0)	0.105 (−0.005)	0.321 (+0.029)	0.493 (+0.057)	0.101	0.039
LYS[c]	54A7[b]	7.8 (−1.0)	0.153 (−0.113)	0.293 (+0.117)	0.496 (+0.054)	1.254	1.215
	#5623	7.1 (−0.3)	0.066 (−0.026)	0.307 (+0.103)	0.579 (−0.029)	0.428	0.445
	#5623[b]	7.1 (−0.3)	0.071 (−0.031)	0.297 (+0.113)	0.554 (−0.004)	0.408	0.436
	#80612	6.8 (+0.1)	0.028 (+0.012)	0.328 (+0.082)	0.583 (−0.033)	0.207	0.222
MET	54A7[b]	7.8 (−0.8)	0.152 (−0.122)	0.288 (+0.182)	0.471 (+0.029)	1.133	1.108
	#5623	7.1 (−0.1)	0.059 (−0.029)	0.261 (+0.209)	0.528 (−0.028)	0.346	0.405
	#5623[b]	7.1 (−0.1)	0.044 (−0.014)	0.307 (+0.163)	0.603 (−0.103)	0.370	0.386
	#28751	7.1 (−0.1)	0.031 (−0.001)	0.399 (+0.071)	0.433 (+0.067)	0.249	0.126
PHE	54A7[b]	8.2 (−1.0)	0.133 (−0.073)	0.43 (+0.060)	0.381 (+0.069)	1.232	1.191
	#5623	7.5 (−0.3)	0.048 (+0.012)	0.473 (+0.017)	0.433 (+0.017)	0.316	0.290
	#5623[b]	7.4 (−0.3)	0.046 (+0.014)	0.467 (+0.023)	0.441 (+0.009)	0.306	0.284

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. continued

			propensities [%/100]			Δ[a]	
		J value [Hz]	P_α	P_β	$P_{P_{II}}$	abs.	ren.
	#25538	7.2 (+0.0)	0.086 (−0.026)	0.390 (+0.100)	0.444 (+0.006)	0.142	0.142
PRO	54A7[b]	6.6	0.256	0.002	0.729	–	–
	#5623[b]	5.0	0.112	0.000	0.857	–	–
SER	54A7[b]	8.0 (−1.0)	0.195 (−0.155)	0.389 (+0.081)	0.370 (+0.120)	1.316	1.288
	#5623	7.3 (−0.3)	0.050 (−0.010)	0.545 (−0.075)	0.347 (+0.143)	0.518	0.534
	#5623[b]	7.3 (−0.3)	0.043 (−0.003)	0.564 (−0.094)	0.333 (+0.157)	0.554	0.572
	#67652	7.0 (+0.0)	0.036 (+0.004)	0.382 (+0.088)	0.479 (+0.011)	0.103	0.088
THR	54A7[b]	7.4 (−0.1)	0.407 (−0.377)	0.137 (+0.443)	0.421 (−0.031)	0.891	0.916
	#86516	7.3 (+0.0)	0.037 (−0.007)	0.410 (+0.170)	0.464 (−0.074)	0.281	0.290
	#86516[b]	7.3 (+0.0)	0.052 (−0.022)	0.409 (+0.171)	0.434 (−0.044)	0.257	0.266
	#5623	6.6 (+0.8)	0.192 (−0.162)	0.110 (+0.470)	0.655 (−0.265)	1.647	1.680
	#92991	7.4 (−0.1)	0.021 (+0.009)	0.462 (+0.118)	0.475 (−0.085)	0.262	0.262
TRP	54A7[b]	8.0 (−1.1)	0.121 (−0.101)	0.401 (+0.039)	0.424 (+0.116)	1.366	1.326
	#5623	7.3 (−0.4)	0.051 (−0.031)	0.461 (−0.021)	0.451 (+0.089)	0.521	0.524
	#5623[b]	7.3 (−0.4)	0.041 (−0.021)	0.472 (−0.032)	0.449 (+0.091)	0.504	0.507
	#93571	6.9 (+0.0)	0.008 (+0.012)	0.391 (+0.049)	0.572 (−0.032)	0.103	0.108
TYR	54A7[b]	8.3 (−1.2)	0.114 (−0.044)	0.473 (−0.003)	0.356 (+0.104)	1.351	1.365
	#5623	7.6 (−0.5)	0.041 (+0.029)	0.505 (−0.035)	0.411 (+0.049)	0.533	0.536
	#5623[b]	7.5 (−0.4)	0.031 (+0.039)	0.498 (−0.028)	0.429 (+0.031)	0.498	0.500
	#6094	7.1 (+0.0)	0.088 (−0.018)	0.399 (+0.071)	0.470 (−0.010)	0.099	0.106
VAL	54A7[b]	7.5 (−0.2)	0.124 (−0.104)	0.233 (+0.277)	0.601 (−0.131)	0.712	0.733
	#86516	7.6 (−0.3)	0.010 (+0.010)	0.499 (+0.011)	0.445 (+0.025)	0.356	0.337
	#86516[b]	7.7 (−0.4)	0.003 (+0.017)	0.527 (−0.017)	0.428 (+0.042)	0.486	0.490
	#5623	6.7 (+0.6)	0.052 (−0.032)	0.141 (+0.369)	0.787 (−0.317)	1.328	1.342
	#24957	7.3 (+0.0)	0.015 (+0.005)	0.457 (+0.053)	0.509 (−0.039)	0.107	0.108

[a]The overall Δ is the sum of the absolute values of the deviation of the J value (in Hz) and the discrepancies in the propensities (in %/100) (see eq 4). The "abs." column refers to the deviation when the absolute occurrences of the three secondary structure classes were used, while the "ren." column refers to the average deviations when the propensities were first renormalized to 100%. [b]These values were computed from real simulations and were not projected. [c]In terms of protonation states, we used GROMOS parameters for HISH (+1 charge, doubly protonated), LYSH (+1 charge, protonated), ARG (+1 charge), and the dissociated versions of glutamic acid (GLU) and aspartic acid (ASP).

Table 4. Summary of the Averaged Performance of the Different Sets in Terms of Agreement with Experimental Data: The Absolute Values of the Individual Deviations Were Averaged and Are Reported Together with Their Standard Deviations; For the Propensities, the Renormalized Data Are Reported

				propensities [%/100]			Δ[a]	
subset	combination	description	⟨|ΔJ|⟩ [Hz]	⟨|ΔP_α|⟩	⟨|ΔP_β|⟩	⟨|Δ$P_{P_{II}}$|⟩	⟨abs.⟩	⟨ren.⟩
glycine	54A7[b]	see ref 1	0.74 ± 0.00	0.646 ± 0.000	0.121 ± 0.000	0.144 ± 0.000	1.651 ± 0.000	1.976 ± 0.000
	#81883	suggested	0.04 ± 0.00	0.450 ± 0.000	0.058 ± 0.000	0.188 ± 0.000	0.736 ± 0.000	0.270 ± 0.000
	#81883[b]	simulated	0.01 ± 0.00	0.439 ± 0.000	0.062 ± 0.000	0.190 ± 0.000	0.701 ± 0.000	0.256 ± 0.000
	#5623	common[c]	1.41 ± 0.00	0.658 ± 0.000	0.082 ± 0.000	0.212 ± 0.000	2.362 ± 0.000	3.076 ± 0.000
alanine	54A7[b]	see ref 1	1.79 ± 0.00	0.001 ± 0.000	0.098 ± 0.000	0.156 ± 0.000	2.045 ± 0.000	2.048 ± 0.000
	#12572	suggested	0.24 ± 0.00	0.004 ± 0.000	0.031 ± 0.000	0.008 ± 0.000	0.283 ± 0.000	0.283 ± 0.000
	#12572[b]	simulated	0.14 ± 0.00	0.017 ± 0.000	0.051 ± 0.000	0.003 ± 0.000	0.211 ± 0.000	0.224 ± 0.000
	#5623	common[c]	1.08 ± 0.00	0.071 ± 0.000	0.240 ± 0.000	0.225 ± 0.000	1.616 ± 0.000	1.623 ± 0.000
common	54A7[b]	see ref 1	0.87 ± 0.26	0.093 ± 0.045	0.098 ± 0.072	0.052 ± 0.037	1.115 ± 0.212	1.103 ± 0.198
	#5623	suggested	0.25 ± 0.15	0.024 ± 0.019	0.083 ± 0.062	0.058 ± 0.038	0.416 ± 0.170	0.425 ± 0.179
	#5623[b]	simulated	0.22 ± 0.15	0.018 ± 0.014	0.092 ± 0.053	0.067 ± 0.054	0.401 ± 0.153	0.410 ± 0.156
	Table S5	indiviual	0.03 ± 0.04	0.019 ± 0.023	0.087 ± 0.081	0.065 ± 0.068	0.201 ± 0.152	0.131 ± 0.066
C_β-branched	54A7[b]	see ref 1	0.14 ± 0.09	0.193 ± 0.159	0.334 ± 0.094	0.102 ± 0.061	0.772 ± 0.103	0.795 ± 0.105
	#86516	suggested	0.20 ± 0.15	0.011 ± 0.004	0.066 ± 0.090	0.036 ± 0.034	0.316 ± 0.038	0.309 ± 0.025
	#86516[b]	simulated	0.25 ± 0.20	0.015 ± 0.009	0.066 ± 0.091	0.040 ± 0.006	0.371 ± 0.115	0.369 ± 0.113
	#5623	common[c]	0.67 ± 0.07	0.072 ± 0.078	0.405 ± 0.057	0.306 ± 0.037	1.450 ± 0.172	1.469 ± 0.184
	Table S5	individual	0.02 ± 0.02	0.006 ± 0.003	0.079 ± 0.035	0.056 ± 0.025	0.164 ± 0.085	0.165 ± 0.085

[a]The overall Δ is the sum of the absolute values of the deviation of the J value (in Hz) and the discrepancies in the propensities (in %/100) (see eq 4). The ⟨abs.⟩ column refers to the deviation when the absolute occurrences of the three secondary structure classes were used, while the ⟨ren.⟩ column refers to the average deviations when the propensities were first renormalized to 100%. [b]These values were computed from the respective actual simulations and were not predicted by reweighting. [c]For comparison, the values projected for the common set are reported for the other subgroups as well. It is clear that #5623 does not perform well for the other groups.

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

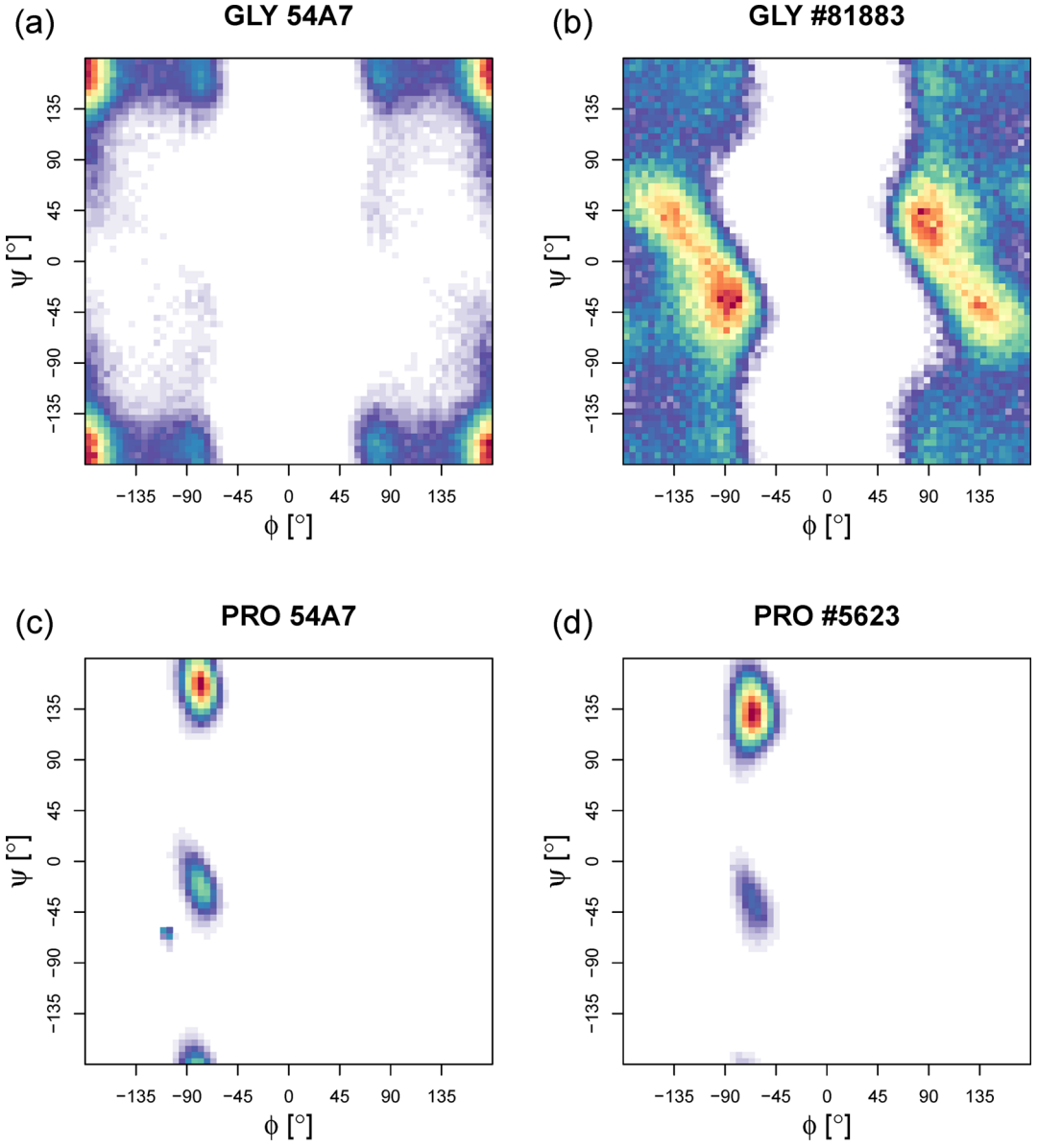


Figure 4. Ramachandran plots for glycine and proline with different parameters. As shown in (a), 54A7 fails to reproduce a distribution that is in agreement with the literature (see ref 30), while the distribution for #81883 in (b) seems to be in better agreement. Moreover, proline in 54A7 shows an unexpected peak at ϕ and ψ of approximately −107° and −65°, respectively (c). For our combination #5623, which is suggested for the common amino acids, we observe (d) a significant shift toward the P_{II} conformation, leaving only a minor fraction in the helical basin.

acids in this subgroup are sufficient to drive the conformational ensemble of glutamine toward the α population. In Figure S3 the prediction of serine using a coarse 15 × 15 bin resolution for several combinations is shown as an example. As long as the number of configurations is large enough, it is clearly possible to predict the sampling of the backbone angles with a higher resolution than just three secondary structure basins.

It appeared difficult to identify a single combination of dihedral angle parameters that could uniformly satisfy the experimental data for all amino acids. This suggests that different potential energy terms may be required for different amino acids. In our opinion this is hardly surprising, as the size, shape, and polarity of the side chain will most likely have an effect on the torsional potential of the backbone and thus need to be taken into account. On the other hand, while it is theoretically possible to optimize the amino acids individually, that approach leads to a complex and potentially overfitted result. Instead, we chose a somewhat intermediate solution: we identified subgroups based on the differences in the experimental data and the possibilities of reproducing these subgroups with common potential energy functions. The subgroups nicely appeared to correspond to the substitution pattern at the C_β side-chain atom. Concretely, the identified commonalities between outliers led to the separate optimization of glycine, alanine, the "common" amino acids, and a subgroup with a CH group at C_β (the C_β-branched amino acids valine, isoleucine, and threonine). Proline, for which the backbone degrees of freedom are much restricted, was excluded from the optimizations.

The rationale for this subdivision of the amino acids is likely to be found in their respective side-chain characteristics. Glycine does not have a real side chain, and this lack of certain steric hindrances allows a broad sampling of angles (see Figure 4 and ref 30). Alanine is the only amino acid that has no

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

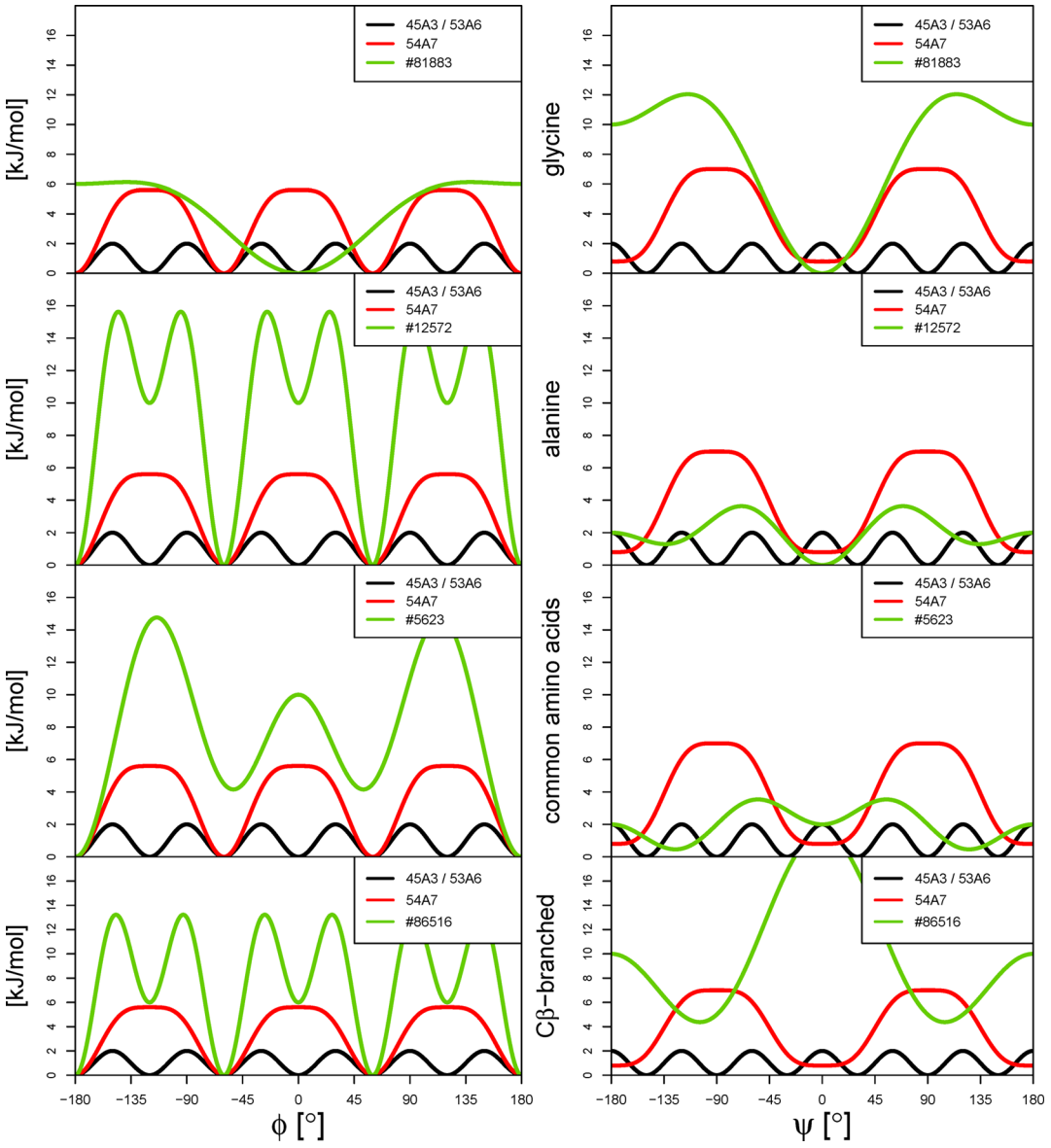


Figure 5. Potential energy terms of Table 1. The top row represents the suggested parameters for glycine, the second row those for alanine, the third row those for the common amino acids subset, and the last row those for the C_β-branched amino acids, respectively. The left column shows the ϕ angle and the right column the ψ angle.

extension on C_β. The common amino acids have one non-hydrogen extension at C_β, and the β-branching in valine, isoleucine, and threonine leads to special rotational profiles. In order to illustrate that, we report the values predicted using combination #5623 (the suggested solution for the common amino acids) in Table 3 for alanine, glycine, and the C_β-branched amino acids. Given the experimental data at which we aimed and the vast set of possibilities that we screened, it might be expected that a solution adequate for all amino acids simultaneously would have been found if it were possible. Instead, we were not able to optimize all of the canonical amino acids together. Glycine is described best by low potentials, granting it its natural extraordinary freedom in sampling the Ramachandran space. The score obtained for alanine using the "common" parameters (#5623) shows a deviation of 1 Hz in the J value and a strong bias toward the α-helical basin. The three β-branched amino acids also do not perform well with this set, and in all three cases the summed deviation is much worse than even 54A7, arising from both the J value and the propensities. For threonine, valine, and isoleucine alike, the common parameters lead to an extreme overemphasis of the P_{II} basin at the cost of β conformations. Accordingly, the optimized torsional energy profiles are quite different for both ϕ and ψ.

The optimal set of parameters for the four subgroups are given in Table 1, and a graphical comparison of the potential energy contributions is given in Figure 5. It should be noted that these figures are somewhat misleading because the main contribution to the potential energy landscape is determined by the intramolecular nonbonded interactions. This is exemplified by Figure 6, which shows the nonbonded energy of the alanine dipeptide resulting from a systematic scan of ϕ and ψ starting from a minimized conformation in vacuum. The torsional dihedral angle profiles in Figure 5 merely modulate the intrinsic potential energy surface, leading to the appropriate shifts in J value and secondary structure propensities. It is not surprising

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

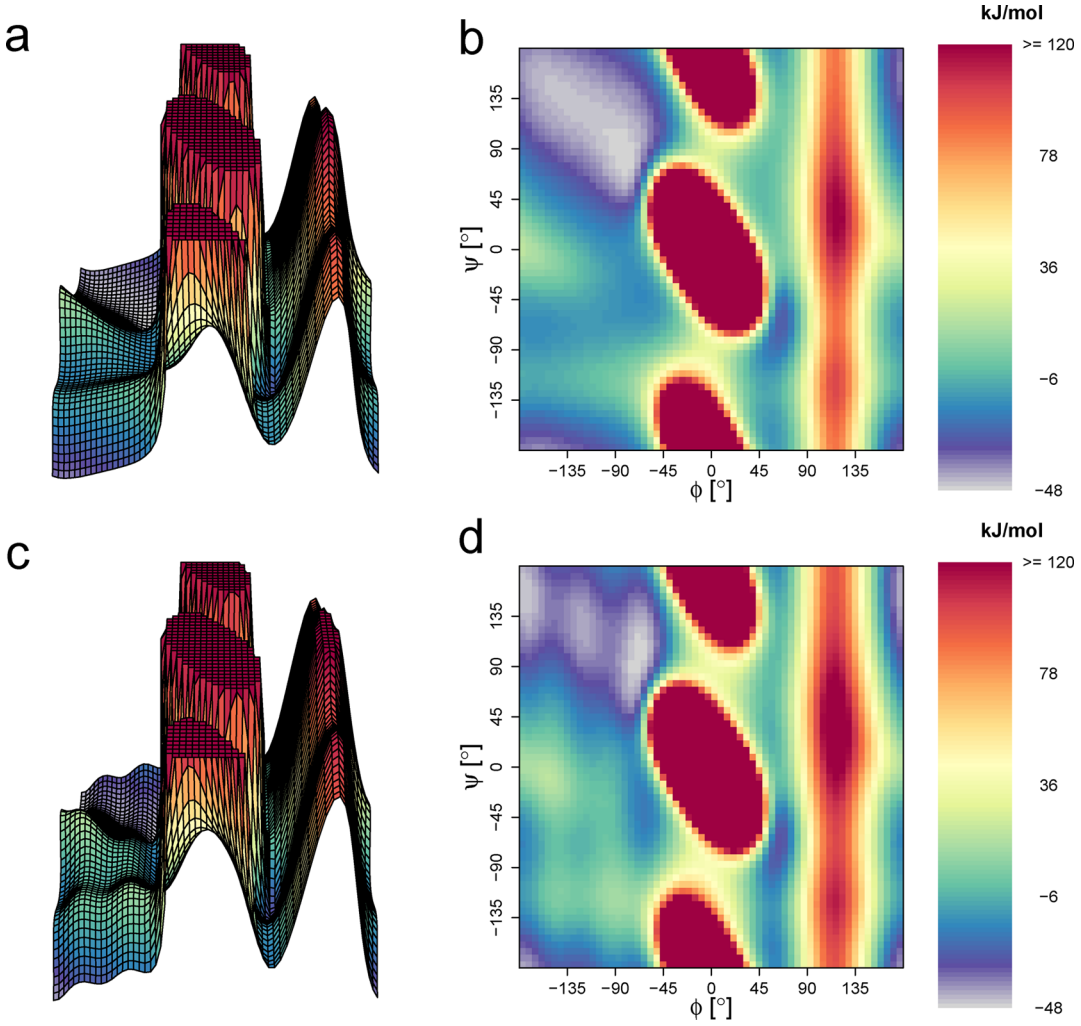


Figure 6. Backbone energy surface for Ac-A-NHMe (using parameter combination #12572) obtained by varying the ϕ and ψ angles. In (a) and (b), the energy surface arising from the nonbonded interactions only is shown in 3D and 2D representations, respectively. It is clear that the addition of the dihedral angle potential energy terms contributes only a limited but still crucial amount, as shown in (c) and (d). This modulation of the energy landscape particularly leads to a P_{II}/β separation and forms a small α-helical basin, which is 12.7% populated (estimated by simulation).

that there are many alternative combinations that could lead to similar shifts in the conformational preference. To ensure that the seemingly high barriers in Figure 5 do not lead to artificial locking of molecules in local minima, we computed the mean residence times in particular basins for one representative amino acid in each subgroup (Table S4). Compared with those for 54A7, the residence times in the helical conformation seem to be reduced while those in β and P_{II} are slightly increased, with the most extreme increment observed for the P_{II} conformation of alanine from 2.1 to 13.7 ps. With currently available simulation times, we do not consider this a significant reduction of the dynamic behavior of the molecule.

In general, it appears that matching of the J value alone is not sufficient to ensure a representative distribution. For example, the J value for threonine in 54A7 matches perfectly, but in terms of propensities, the agreement with the experimental values is rather poor; reweighting combinations #33268 and #75480 for glycine share the J value but differ 11% in propensities (data not shown). The concomitant optimization of both the J value and the secondary structure propensities reduces the number of combinations to be considered: although the J values were used for calibration of the Raman spectroscopy experiments, the latter hold additional information.

Performing the stepwise selection of potential candidates (see Methods), we were able to provide a set of suggested parameters (Table 4) that performs significantly better than the GROMOS 54A7 parameters in terms of agreement with experimentally determined J values and propensities (Figures 3 and S4). In addition to these values, we also sought out the best parameter set for every amino acid individually. Table S5 shows the individually optimized combinations. In general, these perform significantly better than the ones optimized for the subgroups (see Tables 3 and 4), suggesting that our set of parameters contains enough variety to account for the special features of all 20 canonical amino acids.

Of the four subgroups, the β-branched one is the most likely to be further improved. For a group of only three members, the deviation is still quite high for our suggested set of parameters. It appears from the results shown in Figure 3 that all three amino acids contribute to the same extent to this deviation. Their best individual hits share the ϕ potential, but it is threonine that is significantly different in ψ (Table S5) and prohibits a further subgroup-wise optimization. Indeed, for

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

many of the best parameter sets of this subgroup, we identified threonine to contribute most to the deviations.

Among the 20 canonical amino acids in this study, there are two that require special consideration: glycine and proline. Since the propensity values for glycine are most likely only comparable when the crucial effect of the normalization in this case is taken into account, we propose a parameter set for this special amino acid that samples wide areas of the Ramachandran plot, as also observed in the PDB structures.[30] Figure 4 shows the obtained distribution for 54A7 and the selected parameter set (#81883). For proline, no experimental values were available, so we adopted the "common" parameters (set #5623) for this amino acid, leading to a more pronounced sampling of the P_{II} basin and the removal of a surprising (artificial) peak at $\phi \approx -107°$ and $\psi \approx -76°$ (see Figure 4).

Further tests could include other small compound series (e.g., it has been reported in an earlier study[31] that $^3J(HN,H_\alpha)$ increases for alanine inserted in the XAO peptide; see studies mentioned before as well). As a preliminary test of the parameters in a protein environment, we performed four independent 50 ns simulations of hen egg-white lysozyme (HEWL) (initial structure taken from PDB entry 4B0D[32]) using both the GROMOS 54A8 parameter set and a parameter set based on 54A8 with the suggested backbone dihedral parameters. The overall structure seems to be equally well maintained in terms of both root-mean-square deviation (Figure 7) and the persistence of secondary structure elements.

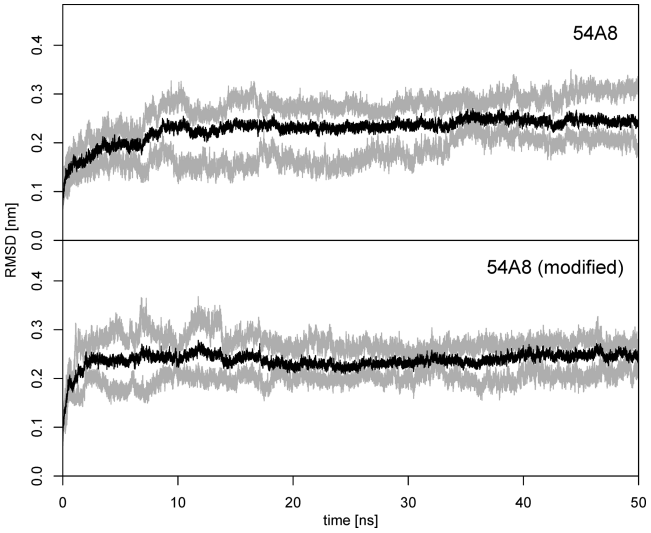


Figure 7. Average root-mean-square deviation (RMSD) with respect to the original structure for four independent simulations of HEWL using (top) parameter set 54A8 and (bottom) 54A8 with our suggested dihedral parameters. The RMSD was calculated on the basis of the backbone atoms C, N, and O exclusively. Both parameter sets show a stable structure over 50 ns. In each panel, the black curve denotes the mean values and the gray ones the respective minimum and maximum values at every time point. The plots were generated using the R package MDplot.

While the simulations with 54A8 show helical conformations of 36.6 ± 1.1% (averaged over sequence, time, and individual simulations), this amounts to 37.6 ± 2.8% for the simulations with the modified backbone dihedral angle parameters. It should be noted that this is merely a preliminary confirmation that the updated set of parameters does not disrupt the protein structure. Future analyses will involve detailed comparisons to NMR observables (nuclear Overhauser effect distance restraints, J values) and a more extensive set of proteins. It is not unlikely that further refinements will be necessary at the protein level. It can be expected that this kind of analysis will have to be performed for every major iteration of the force field's (nonbonded) parameter set because of the interdependence of parameters mentioned before.

CONCLUSION

In this work, we focused on finding optimized parameters for the backbone dihedral angles of amino acids. Screening of a vast library of potential combinations using Hamiltonian reweighting provided a set of parameters that optimizes the experimental target data. We have proven that Hamiltonian reweighting is a useful tool in the parametrization process for molecular dynamics force fields and demonstrated its general accuracy for small to medium changes in the Hamiltonian. Our optimization procedure might serve as a useful basis for further optimizations of other force fields, especially in the range from very small to medium-sized peptides and intrinsically unstructured regions in proteins. Sets of parameters have been suggested for four subsets of amino acids that differ in the substitution at C_β, but various other potential parameter sets are available for further evaluation. Future studies will concentrate on the reproduction of experimental data for other small systems, the performance of the selected parameter sets in proteins, and the agreement against, e.g., NMR data for such simulations.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.6b00399.

Nonbonded parameters used for blocking the amino acids, list of the experimental target values, secondary structure basin definitions, dynamical features of 54A7 and the suggested parameters, parameter combinations for the individually optimized amino acids, validation of the Hamiltonian reweighting approach, an example plot showing the J value and secondary structure deviations for combination #5623, an example of the performed Ramachandran plot predictions, and a correlation plot of the experimental and calculated J values (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: chris.oostenbrink@boku.ac.at.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Financial support by the Vienna Science and Technology Fund (WWTF) (LS08-QM03) and the Austrian Science Fund (FWF) (P25056) is gratefully acknowledged.

REFERENCES

(1) Schmid, N.; Eichenberger, A. P.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A. E.; van Gunsteren, W. F. Definition and Testing of the GROMOS Force-Field Versions 54A7 and 54B7. *Eur. Biophys. J.* **2011**, *40*, 843−856.

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

(2) Avbelj, F.; Grdadolnik, S. G.; Grdadolnik, J.; Baldwin, R. L. Intrinsic Backbone Preferences Are fully Present in Blocked Amino Acids. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 1272−1277.
(3) Grdadolnik, J.; Mohacek-Grosev, V.; Baldwin, R. L.; Avbelj, F. Populations of the three major Backbone Conformations in 19 Amino Acid Dipeptides. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 1794−1798.
(4) Best, R. B.; Zhu, X.; Shim, J.; Lopes, P. E. M.; Mittal, J.; Feig, M.; MacKerell, A. D. Optimization of the additive CHARMM All-Atom Protein Force Field Targeting improved Sampling of the Backbone ϕ, ψ and Side-Chain χ1 and χ2 Dihedral Angles. *J. Chem. Theory Comput.* **2012**, *8*, 3257−3273.
(5) 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.
(6) Toal, S.; Meral, D.; Verbaro, D.; Urbanc, B.; Schweitzer-Stenner, R. pH-Independence of Trialanine and the Effects of Termini Blocking in Short Peptides: A Combined Vibrational, NMR, UVCD, and Molecular Dynamics Study. *J. Phys. Chem. B* **2013**, *117*, 3689−3706.
(7) Schmid, N.; Christ, C. D.; Christen, M.; Eichenberger, A. P.; van Gunsteren, W. F. Architecture, Implementation and Parallelisation of the GROMOS Software for Biomolecular Simulation. *Comput. Phys. Commun.* **2012**, *183*, 890−903.
(8) Christen, M.; Hunenberger, P. H.; Bakowies, D.; Baron, R.; Burgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Krautler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. The GROMOS Software for Biomolecular Simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719−1751.
(9) Schuler, L. D.; Daura, X.; van Gunsteren, W. F. An improved GROMOS96 Force Field for aliphatic Hydrocarbons in the condensed Phase. *J. Comput. Chem.* **2001**, *22*, 1205−1218.
(10) Oostenbrink, C.; Villa, A.; Mark, A. E.; van Gunsteren, W. F. A biomolecular Force Field Based on the Free Enthalpy of Hydration and Solvation: the GROMOS Force-Field Parameter Sets 53A5 and 53A6. *J. Comput. Chem.* **2004**, *25*, 1656−1676.
(11) 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.
(12) Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327−341.
(13) Tironi, I. G.; Sperb, R.; Smith, P. E.; van Gunsteren, W. F. A Generalized Reaction Field Method for Molecular Dynamics Simulations. *J. Chem. Phys.* **1995**, *102*, 5451−5459.
(14) Heinz, T. N.; van Gunsteren, W. F.; Hunenberger, P. H. Comparison of four Methods to Compute the Dielectric Permittivity of Liquids from Molecular Dynamics Simulations. *J. Chem. Phys.* **2001**, *115*, 1125−1136.
(15) Kabsch, W.; Sander, C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* **1983**, *22*, 2577−2637.
(16) Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of Polypeptide Chain Configurations. *J. Mol. Biol.* **1963**, *7*, 95−99.
(17) Nagy, G.; Oostenbrink, C. Dihedral-Based Segment Identification and Classification of Biopolymers I: Proteins. *J. Chem. Inf. Model.* **2014**, *54*, 266−277.
(18) Nagy, G.; Oostenbrink, C. Dihedral-Based Segment Identification and Classification of Biopolymers II: Polynucleotides. *J. Chem. Inf. Model.* **2014**, *54*, 278−288.
(19) Hollingsworth, S. A.; Lewis, M. C.; Berkholz, D. S.; Wong, W.-K.; Karplus, P. A. (ϕ,ψ)-Motifs: a purely Conformation-Based, Fine-Grained Enumeration of Protein Parts at the Two-Residue Level. *J. Mol. Biol.* **2012**, *416*, 78−93.
(20) Hagarman, A.; Measey, T. J.; Mathieu, D.; Schwalbe, H.; Schweitzer-Stenner, R. Intrinsic Propensities of Amino Acid Residues in GxG Peptides Inferred from Amide I Band Profiles and NMR Scalar Coupling Constants. *J. Am. Chem. Soc.* **2010**, *132*, 540−551.
(21) Tran, H. T.; Wang, X.; Pappu, R. V. Reconciling Observations of Sequence-Specific Conformational Propensities with the Generic Polymeric Behavior of Denatured Proteins. *Biochemistry* **2005**, *44*, 11369−11380.
(22) Pizzanelli, S.; Forte, C.; Monti, S.; Zandomeneghi, G.; Hagarman, A.; Measey, T. J.; Schweitzer-Stenner, R. Conformations of Phenylalanine in the Tripeptides AFA and GFG Probed by Combining MD Simulations with NMR, FTIR, Polarized Raman, and VCD Spectroscopy. *J. Phys. Chem. B* **2010**, *114*, 3965−3978.
(23) Beck, D. A. C.; Alonso, D. O. V.; Inoyama, D.; Daggett, V. The Intrinsic Conformational Propensities of the 20 naturally Occurring Amino Acids and Reflection of these Propensities in Proteins. *Proc. Natl. Acad. Sci. U. S. A.* **2008**, *105*, 12259−12264.
(24) Pardi, A.; Billeter, M.; Wüthrich, K. Calibration of the Angular Dependence of the Amide Proton-C alpha Proton Coupling Constants, 3JHN alpha, in a globular Protein. Use of 3JHN alpha for Identification of helical Secondary Structure. *J. Mol. Biol.* **1984**, *180*, 741−751.
(25) Stocker, U.; van Gunsteren, W. F. Molecular Dynamics Simulation of Hen Egg White Lysozyme: A Test of the GROMOS96 Force Field against Nuclear Magnetic Resonance Data. *Proteins: Struct., Funct., Genet.* **2000**, *40*, 145−153.
(26) Soares, T. A.; Daura, X.; Oostenbrink, C.; Smith, L. J.; van Gunsteren, W. F. Validation of the GROMOS Force-Field Parameter Set 45A3 against Nuclear Magnetic Resonance Data of Hen Egg Lysozyme. *J. Biomol. NMR* **2004**, *30*, 407−422.
(27) Best, R. B.; Buchete, N.-V.; Hummer, G. Are Current Molecular Dynamics Force Fields too helical? *Biophys. J.* **2008**, *95*, L07−L09.
(28) Torrie, G. M.; Valleau, J. P. Nonphysical Sampling Distributions in Monte Carlo Free-Energy Estimation: Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187−199.
(29) Lin, Z.; Oostenbrink, C.; van Gunsteren, W. F. On the Use of One-Step Perturbation to Investigate the Dependence of NOE-Derived Atom-Atom Distance Bound Violations of Peptides upon a Variation of Force-Field Parameters. *Eur. Biophys. J.* **2014**, *43*, 113−119.
(30) Lovell, S. C.; Davis, I. W.; Arendall, W. B.; de Bakker, P. I. W.; Word, J. M.; Prisant, M. G.; Richardson, J. S.; Richardson, D. C. Structure Validation by Calpha Geometry: ϕ,ψ and Cβ Deviation. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 437−450.
(31) Shi, Z.; Olson, C. A.; Rose, G. D.; Baldwin, R. L.; Kallenbach, N. R. Polyproline II Structure in a Sequence of seven Alanine Residues. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 9190−9195.
(32) Cipriani, F.; Röwer, M.; Landret, C.; Zander, U.; Felisaz, F.; Marquez, J. A. CrystalDirect: a new Method for automated Crystal Harvesting Based on laser-induced Photoablation of thin Films. *Acta Crystallogr., Sect. D: Biol. Crystallogr.* **2012**, *68*, 1393−1399.

DOI: 10.1021/acs.jcim.6b00399
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX