Force Field Effects in Simulations of Flexible Peptides with Varying Polyproline II Propensity

Jephthah, Stéphanie; Pesce, Francesco; Lindorff-Larsen, Kresten; Skepö, Marie

Published in:
Journal of Chemical Theory and Computation

DOI:
10.1021/acs.jctc.1c00408

Publication date:
2021

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Force Field Effects in Simulations of Flexible Peptides with Varying Polyproline II Propensity

Stéphanie Jephthah, Francesco Pesce, Kresten Lindorff-Larsen,* and Marie Skepö*

ABSTRACT: Five peptides previously suggested to possess polyproline II (PPII) structure have here been investigated by using atomistic molecular dynamics simulations to compare how well four different force fields sample PPII structures and to determine how well four different force fields sample PPII structures. The simulations were compared to circular dichroism and nuclear magnetic resonance spectroscopy experiments and to the predictions made by several software packages available for analyzing CD data and secondary structure predictions. The results revealed that all force fields sample PPII structures but to different extents and with different propensities toward other secondary structure elements, in particular, the β-sheet and “random coils”. A cluster analysis of the simulations of histatin 5 also revealed that the conformational ensembles of the force fields are quite different. We compared the simulations to circular dichroism and nuclear magnetic resonance spectroscopy experiments and conclude that further experiments and methods for interpreting them are needed to assess the accuracy of force fields in determining PPII structure.

1. INTRODUCTION

Intrinsically disordered proteins and regions (IDPs and IDRs)—also recognized as natively unfolded proteins and peptides—are characterized by the lack of a well-defined tertiary structure in aqueous solution. Their structural properties are known to vary significantly, which makes it difficult to study them by standard methods. For example, because of their flexible behavior, IDPs and IDRs cannot be crystallized, and until relatively recently, many molecular simulations of flexible peptides and proteins had a strong bias for sampling α-helical and β-sheet structures as well as too compact and stable structures.

Despite IDPs often being thought of as “unordered”, studies originating from the 1970s discovered that a few natively unfolded peptides possessed some degree of local order in their backbones, identified as the left-handed polyproline II (PPII) helix. Although the PPII helix was discovered a long time ago, it is still substantially less known than, for example, the α-helix and the β-sheet. One of the more recognized occurrences of the PPII helix might be as being part of the triple-helix structure of collagen, where it helps stabilize the collagen structure, and indeed specific efforts have been devoted to optimizing computational models for collagen.

Another well-known and important property of the PPII helix has been observed in the binding to SH3 domains, where it facilitates and mediates protein–protein interactions. As a secondary structure element, the PPII helix is decidedly different from the α-helix and the β-sheet and perhaps less well-known, although frequently occurring in many proteins. The PPII helix has backbone dihedral angles of approximately (φ, ψ) = (−75°, +145°) with a helical pitch of 9.3 Å/turn and 3.0 residues/turn, which causes it to become quite extended. The peptidal bond in a PPII helix is always found as the trans isomer (ω = +180°) since the cis isomer (ω = 0°) yields a different secondary structure known as the polypropylene 1 (PPII) helix ((φ, ψ) = (−75°, +160°)) with a helical pitch of 5.6 Å/turn and 3.3 residues/turn.

One popular experimental technique for determining the secondary structure of proteins is circular dichroism (CD) spectroscopy. In the CD spectrum, the PPII helix is often associated with a strong band with negative ellipticity around 198 nm and a weak positive band around 218 nm. There are several software packages available for analyzing CD data and providing estimates of the relative secondary structure content, at least in terms of α-helices and β-sheets. Unfortunately, these algorithms may fall short when it comes to analyzing spectra of more disordered proteins that contain several less common secondary structure elements, including the PPII helix, and where the structural elements are not fixed in time. In such cases, the remaining secondary structure elements are lumped together and categorized as “others” or “random coils”. A similar problem is also encountered when the secondary structure...
content of protein structures is determined from simulations. Although the widely used DSSP (Dictionary of Secondary Structure of Proteins) program is able to identify and quantify a wider collection of secondary structure elements, it does not include the PPII helix. Fortunately, there is other software available that utilizes modified DSSP assignment to also include the PPII structure.22,23

The development of force fields for simulating IDPs is constantly evolving to help alleviate the problem of overly collapsed structures in simulations and to make the simulations as computationally efficient and accurate as possible.24 So far in force field development, the focus has mainly been on optimizing two aspects of IDP simulations.25 The first aspect is the secondary structure propensities, which are often modified by adjusting the protein backbone dihedral parameters.26−28 The second aspect concerns the balance of the protein−solvent interactions,29,30 which is crucial to not sampling too compact IDP conformations and to accurately capturing the more extended conformations. This is generally controlled by increasing and fine-tuning the interaction between the protein and the water in the simulations.

To evaluate new force fields for simulations of flexible peptides and proteins, different properties can be considered. Nuclear magnetic resonance (NMR) observables, such as scalar couplings and chemical shifts, are used for assessing force field accuracy by comparing simulated and experimental values and are in particular sensitive to local structural properties. Comparisons of scattering curves and the radius of gyration are used to evaluate the global compactness of the simulated proteins. In addition to such direct comparison to experimental observables, secondary structure propensities are often also assessed, and although more comprehensive analyses sometimes are used, they are most often restricted to the α-helix, the β-sheet, and the “random coil”. Conformational clustering is also sometimes used as a tool in force field analyses.

Here we present a study where five different short peptides (7−24 residues long) with varying PPII propensities, as well as five variants of one of the peptides, have been simulated with four different force fields that are known to work relatively well for simulating IDPs. Our analyses were mainly focused on differences regarding the secondary structure content across peptides and force fields. Our findings revealed that, although all the chosen force fields give rise to conformational ensembles with some level of PPII structure, they do so to different extents and with different propensities toward other secondary structure elements. Additionally, all force fields capture a trend showing that the PPII content increases with the number of Pro residues in peptide chains consisting of only Ala and Pro residues.

2. METHODS

2.1. Molecular Dynamics Simulations. Five different peptides known to possess PPII structure, as well as five different variants of one of the peptides, were simulated using atomistic molecular dynamics (MD) simulations. Names and amino acid sequences of the selected peptides are shown in Table 1.

The simulations were performed with the GROMACS package (ver. 4.6.7),31−33 with four different force fields that have also previously been used for simulating IDPs: (A) the AMBER 899SB-disp force field with its own TIP4P-D-type water model,34 (B) the AMBER 899SB-ILDN force field35 with the TIP4P-D water model,36 (C) the CHARMM36IDPSFF force field with the CHARMM-modified TIP3P water model,37 and (D) the CHARMM36m force field,38 also with the

| Table 1. Names, Number of Amino Acid Residues (#aa), and Amino Acid Sequences of the Peptides Used in This Study* |
|----------------|----------------|----------------|
| Name           | #aa            | Sequence       |
| A7             | 7              | AAAAA          |
| P-113          | 12             | AKRHHGKYRRK   |
| P13            | 13             | PPPPPPAPP PPP  |
| V1: P6AP6      | 13             | PPPPPPAPP PPP  |
| V2: (PA)6P5    | 13             | PAPAPAPAPAPAP  |
| V2: 6               | 13             | AAAAAPP PPPPPP |
| V3: (PA2)2P3    | 13             | PPPAPPAP APP   |
| V4: (A3P2)2A3   | 13             | AAAPPAAPPAAA   |
| Pep3           | 13             | DRAEGIPELNQG   |
| Hst5           | 24             | DSAAKIHGYKRKFKHHSHHKGY |

*Positively charged amino acids are shown in blue, and negatively charged amino acids are shown in red.

<table>
<thead>
<tr>
<th>Table 2. Force Field Notations</th>
</tr>
</thead>
<tbody>
<tr>
<td>force field</td>
</tr>
<tr>
<td>AMBER 899SB-disp</td>
</tr>
<tr>
<td>AMBER 899SB-ILDN</td>
</tr>
<tr>
<td>CHARMM36IDPSFF</td>
</tr>
<tr>
<td>CHARMM36m</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3. Force Fields (FF), Total Production Run Times (t), and Temperatures (T) of All the Simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>peptide(s)</td>
</tr>
<tr>
<td>A7, P-113, P13, V1, V2, V3, V4, Pep3</td>
</tr>
<tr>
<td>A7, P13, V1, V2, V2.2, V3, V4, Pep3</td>
</tr>
<tr>
<td>A7, P-113, P13, V1, V2, V3, V4, Pep3</td>
</tr>
<tr>
<td>A7, P-113, P13, V1, V2, V3, V4, Pep3</td>
</tr>
<tr>
<td>P-113</td>
</tr>
<tr>
<td>Hst5</td>
</tr>
<tr>
<td>Hst5</td>
</tr>
<tr>
<td>Hst5</td>
</tr>
<tr>
<td>Hst5</td>
</tr>
</tbody>
</table>

Figure 1. Example of a Ramachandran map illustrating the four different secondary structure regions analyzed in this study. The β-sheet region is defined by −180° ≤ ϕ < −104° and 180° ≤ ψ ≤ 104°, the PPII helix region is defined by −104° ≤ ϕ ≤ −46° and 116° ≤ ψ ≤ 174°, the αS-helix region is defined by −90° ≤ ϕ ≤ −30° and −90° ≤ ψ ≤ 0°, and the αL-helix region is defined by 30° ≤ ϕ ≤ 90° and 0° ≤ ψ ≤ 90°. Anything outside of these regions is classified as “random coil”. The plot is normalized for a maximum intensity of 1. The displayed angles of the example were obtained from simulations of Hst5 with A99SB-ILDN.
CHARMM-modified TIP3P water model. When discussed in the text, the force fields are referred to by their abbreviations, and when described in the figures and tables, they are referred to by their one-letter code (see Table 2). A rhombic dodecahedron was used as a simulation box, with periodic boundary conditions in all directions. A minimum distance of 1 nm was set between the solute and the box edges. The initial, linear protein structures were built with the use of PyMOL.

The Verlet leapfrog algorithm, with a time step of 2 fs, was used to integrate the equations of motion. Nonbonded interactions were computed with a Verlet list cutoff scheme, short-ranged interactions were calculated by using a pair list with a cutoff of 1 nm, and long-range electrostatics were evaluated by using particle mesh Ewald summation with cubic interpolation and a grid spacing of 0.16 nm. Long-range dispersion interactions were applied to the energies and pressures of the simulated systems. All bond lengths were constrained by using the LINCS algorithm. A velocity-rescaling thermostat with a relaxation time of 0.1 ps was used to keep a temperature of 293 or 300 K (see Table 3 for details), and a Parrinello–Rahman barostat kept the pressure at 1 bar throughout the simulations. A relaxation time of 2 ps was used, and the isothermal compressibility was set to that of water, i.e., \( 4.5 \times 10^{-5} \) bar\(^{-1} \).

Energy minimization was done by using the steepest descent algorithm. Equilibration of the temperature and pressure was done in two steps and with position restraints on the proteins: (1) 500 ps in the NVT ensemble and (2) 1000 ps in the NPT ensemble.

### Table 4. Radius of Gyration, \( R_g \) (nm), for the Five Peptides with the Four Different Force Fields

<table>
<thead>
<tr>
<th>FF</th>
<th>A7</th>
<th>P-113</th>
<th>P-13</th>
<th>Pep3</th>
<th>Hst5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.62</td>
<td>0.97 ± 0.01</td>
<td>1.14</td>
<td>0.91 ± 0.03</td>
<td>1.29 ± 0.08</td>
</tr>
<tr>
<td>B</td>
<td>0.63</td>
<td>0.91 ± 0.03</td>
<td>1.14</td>
<td>0.94 ± 0.01</td>
<td>1.29 ± 0.05</td>
</tr>
<tr>
<td>C</td>
<td>0.60</td>
<td>0.86 ± 0.01</td>
<td>1.13</td>
<td>0.87 ± 0.01</td>
<td>1.18 ± 0.02</td>
</tr>
<tr>
<td>D</td>
<td>0.61</td>
<td>0.92 ± 0.02</td>
<td>1.11</td>
<td>0.97 ± 0.01</td>
<td>1.35 ± 0.03</td>
</tr>
</tbody>
</table>

The values of A7 and P-13 are reported without error margins because their errors are smaller than 0.005 nm.

**Figure 2.** Free energy surfaces as a function of the first (\( x \)-axis) and second (\( y \)-axis) principal components. As a result of performing the PCA on the aggregated trajectories, the PC coordinates for all the force fields on a row are the same.
ensemble. Five replicates with different starting seeds were used for each simulation. The final production runs were performed in the NPT ensemble for a total of 5 μs (5 × 1 μs) for the majority of the peptides. Hst5 with A99SB-ILDN was run for a total of 7 μs (1 × 3 μs + 2 × 2 μs), and P-113 with A99SB-ILDN was run for a total of 12 μs (1 × 8 μs + 2 × 2 μs). The differences among all the simulations are summarized in Table 3. Simulation data of Hst5 with A99SB-ILDN and Hst5 with C36m have previously been published in the paper by Jephthah et al.48

2.2. Simulation Analyses. The GROMACS tool g_analyse was used to obtain averages, autocorrelation functions, and error estimates (block averaging) of the radius of gyration and the end-to-end distance of the simulated peptides. Results from these analyses were used to assess convergence and are presented in the Supporting Information (Figures S1–S15). The GROMACS tool g_cluster was used with the GROMOS algorithm50 to obtain conformational clusters, and to obtain frames for representative structures. All protein structures were visualized and rendered with PyMOL.37

2.2.1. Principal Component Analysis. Principal component analysis (PCA) is a dimensionality reduction method that makes it possible to represent a fraction of the information contained in a large set of variables (or features) in a smaller set. This is achieved by calculating the eigenvectors (or principal components) of the variables’ covariance matrix. A PCA was performed, for each peptide, on an aggregated trajectory made by concatenating the trajectories resulting from the four different force fields. This, as suggested in ref 51, ensures a robust comparison of the force fields by projecting the resulting trajectories onto common principal components. PCA calculations were performed with pyEMMA32 using as features the cosine and sine of each backbone dihedral. The analyses were based on the first two principal components.

2.2.2. Secondary Structure Analysis. The secondary structure was analyzed by using three methods: (i) the DSSP algorithm,53 (ii) the DSSP-PPII algorithm,20 and (iii) estimations from the Ramachandran plot. The last method is described in the following paragraph. The GROMACS tool g_chi was used to check the ω angles in the simulations (see Figure S16).

The local structural preferences were also estimated from the dihedral angles of the peptide backbones, which were obtained by using the GROMACS tool g_rama. Only the α-helical (both the right-handed and the left-handed), the β-sheet, and the PPII helical regions of the Ramachandran map were considered for this analysis. Similar to what has previously been done in many other studies,9,20,26,27,30,31,54 a residue was considered to be in the right-handed α-helix (αR) region of the Ramachandran plot when −90° ≤ ϕ ≤ −30° and −90° ≤ ψ ≤ 0°, as illustrated in the Ramachandran map in Figure 1. Correspondingly, 30° ≤ ϕ ≤ 90° and 0° ≤ ψ ≤ 90° were used for the left-handed α-helix (αL) region, −180° ≤ ϕ < −104° and 180° ≤ ψ ≤ 104° were used for the β-sheet region, and −104° ≤ ϕ ≤ −46° and 116° ≤ ψ ≤ 174° were used for the PPII helix region. Residues not belonging to any of the aforementioned regions were unclassified but categorized as “bend/coil/turn” for simplification of the plots. We note that this classification is not based on secondary structure elements, but simply examines which regions of the Ramachandran map the different residues populate.

2.2.3. CD Prediction. To predict CD spectra from structural ensembles, we employed SESCA.59 The SESCA algorithm has two steps:

1. The first step is per residue secondary structure assignment. We use DISICL60 as the secondary structure prediction algorithm as it explicitly takes into account PPII conformations.
2. Spectral contributions from each secondary structure element in a conformation are combined to produce the CD spectra. In SESCA, the set of spectral contributions assigned to subsets of secondary structures are stored in the “basis sets”. Different basis sets for a given secondary structure assignment are available, which differ in the resolution of the spectral

![Figure 3. Structural analysis of free energy minima in A99SB-disp (blue) and C36m (red). (a) Side view and (b) C-terminal view of two representative P13 structures from A99SB-disp and C36m. (c) Average per residues ϕ and ψ dihedrals, compared to the PPII helix range defined in DSSP-PPII. (d) Per residue probability of PPII conformations as assigned by DSSP-PPII.](https://doi.org/10.1021/acs.jctc.1c00408)
contributions definition. Optionally, the spectral contributions of the side chains may be added. We tested several of the available basis sets, but mainly used DS6-1SC1 (DS6-1 with side chain contribution), as this gave rise to the predicted spectra that resembled the most in shape the experimental spectra.

Finally, the CD spectra from each conformation of the ensemble are linearly averaged.

2.3. Calculation of $J$-Couplings. We used Karplus-like equations\textsuperscript{61} to calculate the backbone NMR scalar ($J$) couplings. This equation has the functional form

$$J(\theta) = A \cos^2(\theta + \delta) + B \cos(\theta + \delta) + C$$  \hspace{1cm} (1)

In eq 1, $\theta$ is the torsional angle that determines the $J$-coupling constant, while $A$, $B$, and $C$ are fitting parameters and $\delta$ is a shift used in some calculations. For $3J_{HNH}^\alpha$, $2J_{NC}^\alpha$, $1J_{NC}^\alpha$, and $3J_{HNC}^\beta$, we used the parametrization from Lindorff-Larsen et al.\textsuperscript{62} The same parametrizations employed in the work of Graf et al.\textsuperscript{63} were used for $3J_{NC}^\alpha$, $2J_{NC}^\alpha$, and $3J_{HNC}^\beta$.

To compare the calculated coupling constants with the experimental values we used

$$\chi^2 = \sum_{i=1}^{N} \left( \frac{J_{i,\text{calc}} - J_{i,exp}}{\sigma_i^2} \right)^2$$  \hspace{1cm} (2)
where \( \langle J \rangle_{\text{calc}} \) is the time average of the \( i \)th \( J \)-coupling constant from the frames of a simulation, \( J_{\text{exp}} \) is the respective experimentally observed \( J \)-coupling constant, and \( \sigma_i \) is the error associated with the parametrization of the Karplus relationship.\(^{55-67}\) Experimental errors and the errors on simulated averages are smaller than those from the Karplus parametrizations and were thus not included.

3. RESULTS AND DISCUSSION

3.1. Effects of Force Field in Simulations of Five Peptides. An initial comparison of the effects of the force fields was done by analyzing the resulting average radius of gyration of the five peptides (Table 4). All force fields resulted in similar average values for the radius of gyration for each of the individual peptides, although C36IDPSFF on average resulted in slightly more compact conformations compared to the other force fields. This was, however, not the case for P13, for which C36m sampled more compact conformations compared to the other force fields. Interestingly, the PPII region seems to be somewhat shifted in Ramachandran plots depicting all dihedral angles of each simulated peptide are presented in Figure 4. A few differences were observed when the force fields were compared. First, the Amber force fields (especially A99SB-ILDN) clearly show a more distinct \( \beta \)-sheet region and the populations in both the \( \alpha \)-helix region and the \( \beta \)-sheet region seem to be confined to smaller and more concentrated regions in the Amber simulations compared to the CHARMM simulations, where they seem to be spread out over larger areas. It is also worth noting that all Amber simulations have similar appearances/distributions over the Ramachandran space. The same is observed for the CHARMM simulations, and their appearance/distribution is slightly different from that of the Amber simulations. This strongly suggests that the different force field families (as might be expected) sample different structures.

Ramachandran plots depicting all dihedral angles of each simulated peptide were presented in Figure 4. A few differences were observed when the force fields were compared. First, the Amber force fields (especially A99SB-ILDN) clearly show a more distinct \( \beta \)-sheet region and the populations in both the \( \alpha \)-helix region and the \( \beta \)-sheet region seem to be confined to smaller and more concentrated regions in the Amber simulations compared to the CHARMM simulations, where they seem to be spread out over larger areas. It is also worth noting that all Amber simulations have similar appearances/distributions over the Ramachandran space. The same is observed for the CHARMM simulations, and their appearance/distribution is slightly different from that of the Amber simulations. This strongly suggests that the different force field families (as might be expected) sample different structures. Interestingly, the PPII region seems to be somewhat shifted in the case of P13, with C36m, which is not seen for the other peptide simulations with the same force field. From studying these aggregated distributions across the Ramachandran plots alone, it is nearly impossible to obtain any detailed information on secondary structure propensities. Thus, the region populations have been quantified and are presented in Figure 5, where they are also compared to secondary structure estimates from the DSSP and the DSSP-PPII methods.

Comparisons of the average secondary structure content of each simulated peptide using three different methods are depicted in Figure 5. The DSSP analysis suggested that all peptides were fully disordered and dominated by bends, coils, and turns. Further investigation using DSSP-PPII did, however,
reveal that approximately 10−35% of the secondary structure content was, in fact, PPII structure for most of the peptides. The P13 peptide was found to possess even more PPII structure (∼50−80%), which is reasonable since it is expected to mainly possess PPII structure in aqueous solution.68,69

By comparing the secondary structure contents obtained from the different force fields using the DSSP-PPII method, A99SB-disp was found to sample more PPII structure than the other force fields for all peptides (except P13) and A99SB-disp and C36IDPSFF sampled slightly more α/310/β-sheet/bridge content. A7 and P13 possessed no other secondary structure elements according to this analysis. The amount of unstructured content (coil/bend/turn) was highest for C36m in the case of P13. No other obvious secondary structure propensities and trends were discerned.

Average populations of different regions of the Ramachandran map (corresponding to typical dihedral angles in different secondary structure elements) were also estimated from all of the simulations. Although this method was able to identify angles corresponding to PPII, it also showed large populations in the α-helical and β-sheet regions, which were not observed in the DSSP and the DSSP-PPII analyses. This is not surprising since the Ramachandran map includes all angles regardless of position, whereas secondary structures need several consecutive amino acids with the same classification for them to register as a secondary structure. The Ramachandran analysis indicated that C36IDPSFF provided less sampling in the structured regions for
all peptides except P13, whereas A99SB-ILDN sampled larger populations in the $\beta$-sheet region than any other force field. Similar to what was observed in the DSSP-PPII analysis, A99SB-disp and C36IDPSFF were found to have larger sampling in the $\alpha$-helical regions than the other two force fields.

Each simulated peptide was investigated with DSSP-PPII to identify the average secondary structure content per amino acid residue (Figure 6). All force fields gave similar secondary structure profiles for A7 and P-113, although a small difference was observed between the Amber and the CHARMM force fields. In P-113 the largest PPII content was centered around Arg-3 and Arg-9, and the lowest PPII content was found around Gly-6. The secondary structure profiles of P13 were similar for all force fields except for C36IDPSFF, where the PPII content was lower. For the Pep3 simulations, the largest PPII content was centered around Ala-3 and Pro-7 for all force fields. A small increase in PPII content was also observed around Asn-11 in Pep3 for all force fields except C36IDPSFF. The PPII content followed a sharper decrease toward the C-terminus in the simulations with A99SB-ILDN and C36IDPSFF, and the lowest PPII content was centered around Gly-5. The $\alpha/310/\pi$-helical content in Pep3 was found mainly around Val-10 except for in the C36IDPSFF simulation, where a larger portion was centered around Lys-4. For Hst5, the largest PPII content was found around Lys-5 and Lys-13 for all force fields, in addition to a smaller peak around Lys-17. A99SB-ILDN and C36m had their $\alpha/310/\pi$-helical content around Tyr-10, whereas it was located closer to the termini in the C36IDPSFF simulations and around His-19 in the A99SB-disp simulations. The simulation of Hst5 with A99SB-disp also gave rise to a low amount of $\alpha/310/\pi$-helical content throughout most of the peptide. The $\beta$-sheet/bridge content in Hst5 was found around Lys-5 and Lys-13 for all force fields, in addition to a smaller peak around Lys-17. A99SB-ILDN and C36m had their $\alpha/310/\pi$-helical content around Tyr-10, whereas it was located closer to the termini in the C36IDPSFF simulations and around His-19 in the A99SB-disp simulations. The simulation of Hst5 with A99SB-disp also gave rise to a low amount of $\alpha/310/\pi$-helical content throughout most of the peptide. The $\beta$-sheet/bridge content in Hst5 was found around Lys-5 and Arg-22 in the Amber simulations but was more randomly distributed in the CHARMM simulations.

Differences between the force fields were further analyzed for Hst5. Cluster analysis was performed for each individual force field, as well as for a concatenated trajectory in which all four force fields were included. The analysis was done using a root-mean-square-deviation (RMSD) cutoff of 0.5 nm. This value...
was chosen by examining the total population of the eight first clusters and varying the cluster radius (in steps of 0.05 nm) until their total population was closest to 50%. Additionally, using the same cutoff for all force fields made it easier to compare them. We do, however, note that because clustering methods compare “central structures”, they are not optimal to use for analyzing flexible peptides. Therefore, the number of clusters and their sizes that are presented here are not representations of the “truth”—they are simply used to compare the conformational sampling of the force fields.

The representative structures of the top eight conformational clusters of Hst5 with the four force fields, as well as the force field mix, are presented in Figure 7. Visual inspection of the representative structures immediately reveals that the first cluster conformers are different for the different force fields. Comparing the combined percentage sizes of the top eight conformation clusters gives some indication of the relative variability of the conformations sampled by the four different force fields. A higher value means that there are fewer conformations sampled in the remaining clusters, which suggests a lower degree of conformational variability. By this reasoning, of the four selected force fields, A99SB-disp provides the smallest amount of conformational variability with the remaining three force fields being comparable.

To get a more quantitative comparison among the force fields, their trajectories were concatenated, followed by a new cluster analysis where each structure could be traced back to its individual force field. The relative cluster populations of the individual force fields in the top eight clusters are illustrated in Figure 8. Although all force fields are represented in each cluster, they are not evenly distributed. For example, the first cluster is dominated by C36IDPSFF, whereas the second cluster mainly contains conformations from A99SB-disp and C36m. The fifth cluster is the most evenly distributed cluster across the force fields, and the sixth cluster is heavily dominated by A99SB-ILDN. From this analysis it is safe to say that, although the average properties of different force fields may be similar, the force fields’ individual conformational ensembles are rather different, which naturally leads to different secondary structure contents.

### 3.2. CD Prediction Using SESCA

Since the four force fields give rise to different conformational ensembles, one may reasonably ask which of the force fields is more representative of the real conformational ensemble in solution. To answer this question, we may attempt to compare the simulations to experimental data. This ideally requires a forward model to predict the experimental observables from an ensemble of structures. We here used data from CD spectroscopy, as CD is highly sensitive of secondary structure composition, and we used SESCA as a forward model. Experimental data for A7, P-113, and Hst5 were obtained from Graf et al., Han et al., and Jephthah et al. respectively. Unfortunately, and as also noted for other IDPs in the papers by Fagerberg et al. and Gopal et al., it was not possible to obtain a meaningful agreement between the experimental CD spectra and those predicted by SESCA. This can be due to the fact that the main negative signature peak of a PPII helix may appear in experiments between 190 and 210 nm because of non-secondary-structure contributions, while the spectral contribution associated with a PPII conformation in SESCA has a fixed position. Also, given the relative scarcity of PPII structure in folded proteins, it may be difficult to deconvolute its

![Figure 9. Comparison of experimental CD spectra with the ones predicted with SESCA from the conformational ensembles produced employing different force fields. The y-axes show the ellipticity, θ (deg cm²/dmol).](image)

<table>
<thead>
<tr>
<th>FF</th>
<th>χ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>19.8</td>
</tr>
<tr>
<td>B</td>
<td>98.1</td>
</tr>
<tr>
<td>C</td>
<td>39.6</td>
</tr>
<tr>
<td>D</td>
<td>53.1</td>
</tr>
</tbody>
</table>

Table 5. Result of Comparison between the Experimental J-Coupling Constants of A7 and Those Calculated from the Conformational Ensembles Resulting from the Different Force Fields
contribution when developing prediction methods for CD spectra. Also, a qualitative analysis based on the intensity of the main negative peak does not provide a decisive suggestion of what force field may be the most reliable. This is also complicated by some intensity scaling that may be needed to take into account uncertainty in the estimate of the concentration of the sample used for the experimental CD data. At this stage it is not clear if the source of the problem may be the force fields’ inaccuracy, finite sampling, or the inaccuracy of the CD calculation for these kinds of systems.

3.3. Comparing A7 Simulations to Experimental NMR Scalar Couplings. A perhaps more accurate way of comparing simulations and experiments of flexible proteins and peptides is by investigating NMR scalar couplings. Since experimental J-coupling constants are available in the literature for A763 and J-couplings are more commonly used in comparison with computational models, we decided to calculate J-couplings from our A7 simulations (see Methods). Simulations with A99SB-disp show the best agreement with the experimental data, followed by C36IDPSFF, while C36m and A99SB-ILDN give rise to a less good fit with the experimental data (Table 5).

This result may in part reflect that the target data for optimizing A99SB-disp included the experimental J-couplings for Ala5. With 39 scalar couplings used to calculate the \( \chi^2 \) values, it appears that the deviations observed in C36m and A99SB-ILDN are greater than what would be expected by chance.

We examined in more detail the \( ^3J_{HNH} \) and \( ^3J_{NC} \) couplings, as \( ^3J_{HNH} \) can help discriminate the \( \psi \) dihedral regions of \( \beta \) and \( \alpha_i \)/PPII elements, and \( ^3J_{NC} \) can help discriminate the \( \psi \) dihedral regions.
regions of $\beta$/PPII and $\alpha$. We observe that the $\psi$ angle distributions are relatively similar, with a strong peak corresponding to the $\beta$/PPII region, and the agreement with $\gamma_{J_{\text{NC}}}$ is very good (Figure 10c,d). More differences can be observed for $\gamma_{J_{\text{HNH}}}$, where especially A99SB-ILDN and C36m show a lower population in the $\alpha$/$\beta$ (4−7 Hz) and a higher population in an unclassified region, resulting in a lower agreement with the experimental $\gamma_{J_{\text{HNH}}}$ (Figure 10a,b).

3.4. Effect of Proline Residue Content. A few variants of $P_{13}$ V1−V4 (see Table 1), were investigated to see how the Pro content affected the PPII propensities, as estimated by how much the residues sampled the PPII regions of the Ramachandran maps. Figure 11 shows the PPII content as a function of the number of Pro residues in $P_{13}$, and the peptide variants. All force fields yielded significant correlation ($p < 0.05$, see Table 6) between the PPII content and the number of Pro residues for $P_{13}$ and the chosen variants, where an increased number of Pro residues provided a larger PPII content. The slopes of these trends in the linear regression, however, differ depending on what force field was used, with C36m having the smallest increase and A99SB-ILDN having the largest increase. Furthermore, A99SB-ILDN was the force field that provided the strongest correlation (see Table 6).

$V_{2.2}$ (see Table 1) was simulated using the A99SB-ILDN force field to investigate if the PPII content is affected by the relative position of the Pro residues in the amino acid sequence. The average PPII contents of $V_{2}$ and $V_{2.2}$ from the DSSP-PPII analysis were found to be essentially the same: 56 and 55%, respectively. Although the PPII content at first did not seem to be affected by the patterning of the Pro residues, histograms of the secondary structure per amino acid residue of $V_{2}$ and $V_{2.2}$ (Figure 12) revealed that the PPII content is significantly more localized to the Pro residues in $V_{2.2}$, whereas it was more evenly distributed in $V_{2}$. However, further investigation is needed to fully characterize this trend. It is, for example, necessary to study the effect of patterning of Pro in the other $P_{13}$ variants as well. It would also be of interest to see how the trend is affected by the peptide length.

4. CONCLUSIONS

In this study we have evaluated the differences among four different force fields in simulations of five short peptides with varying PPII propensities. All force fields gave similar ensemble averages of the radius of gyration, although the averages by the C36IDPSFF force field were generally slightly smaller compared to the other force fields. All force fields appeared to sample comparable regions of conformational spaces (as probed by the first two principal components) for each individual peptide, although with slightly different probabilities. Similarly, all force fields sampled the PPII structure, but to different extents. Additionally, some force fields were more prone to sampling other secondary structure elements. For example, A99SB-disp and C36IDPSFF sampled more $\alpha$/$\beta$-helical and $\beta$-sheet/bridge content, C36m sampled less structured content than the other force fields, and A99SB-disp often had the highest PPII content. Direct comparison by conformational clustering revealed that the force fields have a bias toward different conformational clusters. CD prediction using SECA was performed to examine which force field provided a more accurate conformational ensemble. Unfortunately, the method was not able to match the predicted and experimental spectra. We also calculated scalar couplings and compared them to experimental results for A7 and found two force fields (A99SB-disp and C36IDPSFF) that gave agreement roughly within expected error. We note that the calculations of both CD and scalar couplings contain contributions from all types of local structures and do not solely report on the accuracy of the PPII content. Finally, we investigated the effect of Pro residue content on the PPII content of short peptides containing only Ala and Pro, and we observed a correlation between the number of Pro residues in the amino acid sequence and the PPII content.

We conclude by highlighting that we need better methods to calculate experimental observables that are sensitive to secondary structure preferences for flexible peptides. Such methods are often parametrized using folded protein structures and, thus, may be difficult to apply or inaccurate for disordered peptides and proteins. In particular, we stress the need for better methods to link populations of PPII-like structures in simulations to a broader range of experimental observables and note that NMR chemical shifts can also be used for this purpose. This is especially needed for simulations of proteins in which PPII might have a significant role, such as for example Hst5, SH3-binding peptides, and collagen.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.1c00408.

Probability distributions, autocorrelation functions, block average error estimates, and $\omega$ angle as a function of time (PDF)

AUTHOR INFORMATION

Corresponding Authors

Kresten Lindorff-Larsen – Structural Biology and NMR Laboratory & the Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark; orcid.org/0000-0002-4750-6039; Phone: +45 35 32 20 27; Email: lindorff@bio.ku.dk

Marie Skepö – Division of Theoretical Chemistry, Lund University, SE-221 00 Lund, Sweden; orcid.org/0000-0002-8639-9993; Phone: +46 46 222 33 66; Email: marie.skepo@teokem.lu.se

Authors

Stéphanie Jephthah – Division of Theoretical Chemistry, Lund University, SE-221 00 Lund, Sweden; orcid.org/0000-0003-4287-506X

Francesco Pesce – Structural Biology and NMR Laboratory & the Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark

Complete contact information is available at: https://pubs.acs.org/doi/10.1021/acs.jctc.1c00408

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was funded by the NordForsk’s Nordic Neutron Science Programme (Grant No. 81912) and the Lundbeck Foundation BRAINSTRUC initiative in structural biology (R155-2015-2666,lundbeckfonden.com). Computer resources
for the simulations were provided by the Swedish National Infrastructure for Computing (SNIC) at the Center for Scientific and Technical Computing at Lund University (LUNARC). We also acknowledge access to computational resources from the ROBUST Resource for Biomolecular Simulations (supported by the Novo Nordisk Foundation Grant No. NF18OC0032608) and from the Biocomputing Core Facility at the Department of Biology, University of Copenhagen.

REFERENCES


(42) The PyMOL Molecular Graphics System, ver. 1.2r1; Schrödinger LLC. 2009.


