JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

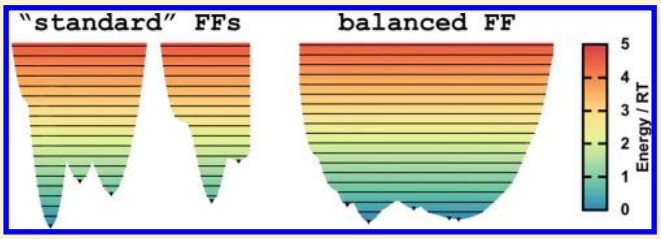
Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment

João Henriques,* Carolina Cragnell, and Marie Skepö

Division of Theoretical Chemistry, Lund University, Post Office Box 124, S-221 00 Lund, Sweden

Supporting Information

ABSTRACT: An increasing number of studies using molecular dynamics (MD) simulations of unfolded and intrinsically disordered proteins (IDPs) suggest that current force fields sample conformations that are overly collapsed. Here, we study the applicability of several state-of-the-art MD force fields, of the AMBER and GROMOS variety, for the simulation of Histatin 5, a short (24 residues) cationic salivary IDP with antimicrobial and antifungal properties. The quality of the simulations is assessed in three complementary analyses: (i) protein shape and size comparison with recent experimental small-angle X-ray scattering data; (ii) secondary structure prediction; (iii) energy landscape exploration and conformational class analysis. Our results show that, indeed, standard force fields sample conformations that are too compact, being systematically unable to reproduce experimental evidence such as the scattering function, the shape of the protein as compared with the Kratky plot, and intrapeptide distances obtained through the pair distance distribution function, $p(r)$. The consistency of this deviation suggests that the problem is not mainly due to protein−protein or water−water interactions, whose parametrization varies the most between force fields and water models. In fact, as originally proposed in [Best et al. *J. Chem. Theory Comput.* **2014**, *10*, 5113−5124.], balanced protein−water interactions may be the key to solving this problem. Our simulations using this approach produce results in very good agreement with experiment.



1. INTRODUCTION

Recent years have seen a sharp increase in the amount of computational studies of unstructured proteins, such as intrinsically disordered proteins (IDPs) and unfolded proteins.[1] Many of those are atomistic molecular dynamics (MD) simulations using standard combinations of force fields and water models that have been primarily developed for classical proteins, i.e. stable, structured proteins with a well-defined native state. *A priori* it is not obvious why such force fields may not be suited for the simulation of unstructured, flexible proteins, as they are constituted by the exact same building blocks, i.e. the 23 proteinogenic amino acid residues. Yet, more and more studies of IDPs and unfolded proteins, using varied simulation settings and parameters, draw the same conclusion: the resulting structures in explicit solvent are too collapsed relative to experiment.[2−6]

A previous study by our group on the IDP Histatin 5, using Metropolis-Hastings Monte Carlo (MC) and constant-pH molecular dynamics (CpHMD) simulations,[2] showed that while both methods produce similar results in respect to electrostatics (net charge and charge capacitance as a function of pH and salt), they differ slightly regarding structure prediction, despite yielding similar relative changes in response to variations in pH and salt. In more detail, the atomistic CpHMD simulations produce conformations with lower average radii of gyration, $\langle R_g \rangle \approx 0.9-1.0$ nm, than the coarse-grained (CG) MC counterpart. At that time, it was concluded that this was probably due to extra detail present in the atomistic model, i.e. smaller excluded volumes combined with the possibility of hydrogen bonding. The high entropic cost of compact conformations combined with lack of hydrogen bonding favors random coil conformations for the CG model and thus larger R_g. The true R_g was hypothesized to be somewhere in between both models. Older studies using nuclear magnetic resonance (NMR)[7,8] and circular dichroism (CD)[8−10] indicate that Histatin 5 behaves like an extended, flexible protein in aqueous solution. In this study, we investigate the applicability of standard, popular combinations of force fields and water models for the simulation of nonclassical, unstructured proteins using MD simulations.

AMBER and GROMOS are two popular force field classes which differ in the parametrization process and level of detail. AMBER force fields are derived from quantum mechanical calculations and are all-atom, in the sense that they comprise full atomistic detail. GROMOS force fields are, on the other hand, usually free-form in terms of parametrization process, with the main effort being driven toward the reproduction of the thermodynamic properties of pure liquids and the solvation free enthalpies of amino acid analogs in different solvents. They are united-atom, i.e. most nonpolar hydrogen atoms are not explicitly represented (collapsed into their respective heavy atom). The force fields of interest are AMBER ff99SB-ILDN,[11] AMBER ff99SBNMR1-ILDN,[12] GROMOS 53A6,[13] and GROMOS 54A7.[14]

Received: December 23, 2014
Published: June 16, 2015

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431

The AMBER ff99[15] (A99) is a very popular force field for the simulation of biomolecules, such as proteins, which has been subject to several improvements over the years. The addition of improved protein backbone parameters lead to the creation of AMBER ff99SB (A99SB),[16] and the subsequential improvement of side-chain torsion potentials resulted in AMBER ff99SB-ILDN (A99SB-ILDN).[11] The latter is one of the current state-of-the-art force fields for the simulation of proteins. Even though validation of force fields by comparison with experimental NMR data is quite common nowadays, it is uncommon for NMR parameters of proteins to be used to actively guide the improvement of protein potentials. One exception is the AMBER ff99SBNMR1 (A99SBNMR1),[12] initially derived from A99SB by modifying the potential of the backbone dihedral angles using existing NMR parameters for a small set of proteins. The AMBER ff99SBNMR1-ILDN (A99SBNMR1-ILDN) is obtained by combining the previously mentioned ILDN side-chain parameters with A99SBNMR1. Other examples of force fields optimized against NMR data are AMBER ff99SB* and ff03*, CHARMM22*[17], and CHARMM36.[18]

In the same fashion, successive parametrizations of the GROMOS force field have been performed over the years in order to achieve better agreement with experimental data. The two latest versions, dedicated to the simulation of biomolecules in water, are the GROMOS 53A6[13] and 54A7[14] force fields (G53A6 and G54A7 in short). Briefly, G53A6 introduces a new set of partial charges to better reproduce hydration free enthalpies in water. This modification has one unfortunate side effect, which is the systematic underestimation of the helical propensities of proteins.[14,19,20] G54A7, among other things, adds adjusted torsional angle terms in order to correct the aforementioned issue with G53A6.

The SPC[21] and TIP3P[22] water models are some of the simplest and most popular models one can use in a MD simulation of biomolecules. Both are similar in nature, i.e. 3-site models with minor differences in their Lennard-Jones parameters and H−O−H angle. Yet, AMBER force fields were intended for use with TIP3P and the GROMOS variety with SPC. More complex multisite, flexible water models do exist, but those are generally more suitable for pure water simulations, where the accurate determination of thermodynamic properties is key. For most general purpose MD *protein in water* simulations it is common practice to use simpler water models in order to drastically decrease the number of degrees of freedom of the system. The computation of nonbonded interactions between water molecules constitutes the bulk of all operations in such a system, and the ratio between accuracy and computational cost does not always justify the use of such advanced models.

Apart from the force field and water model, the specific long-range electrostatics treatment used in a simulation can also have a large influence on its outcome. Particle mesh Ewald (PME)[23] has become the *de facto* methodology for treating the aforementioned type of interactions, yet we are also interested in testing the generalized reaction field (GRF) method[24] to account for the typical setup used in our CpHMD simulations,[2] using the stochastic titration method of Baptista et al.[25]

As mentioned earlier, when used to study IDPs, most state-of-the-art force fields produce structures in explicit solvent that are too collapsed relative to experiment.[2−6] This might, for example, be due to unbalanced interaction(s) which distort the topography of the energy landscape, making more extended conformations energetically unfavorable. In fact, Best et al. recently published an article where it is shown that by improving the balance of protein−water interactions, it is possible to improve the properties of disordered proteins and nonspecific protein association, while not affecting the stability of the native states of structured proteins.[3] Their approach is based on the assumption that proteins are insufficiently well solvated in simulations using current state-of-the-art force fields. Poor solvation arises either indirectly from an inadequate representation of the solvent or directly from protein−solvent interactions. Since current water models are known to accurately reproduce the structure of liquid water,[26−28] the simplest way to address this issue is by modifying the short-range protein−water pair interactions, leaving all other parameters unchanged. On the other hand, Piana et al. hypothesize that water models typically used in MD simulations significantly underestimate London dispersion interactions, and a new water model that corrects for these deficiencies was developed (TIP4P-D).[29] This approach is reported to result in disordered states that are substantially more expanded and in better agreement with experiment.[29]

There are clear similarities between both approaches, as increasing the water dispersion coefficient in TIP4P-D also results in increased protein−water (dispersion) interactions. Furthermore, the Lennard-Jones σ_{OW} and ε_{OW} parameters of both methods are virtually identical (99.80 and 99.95%, respectively). However, it remains unclear whether these approaches are describing different underlying physics or not.

In this work, several popular force field and water model combinations for the simulation of Histatin 5, the model IDP, are studied. The quality of the results is assessed by comparison with small-angle X-ray scattering (SAXS). These *standard* simulations are shown to perform poorly, and the consistency of the resulting deviations across simulation sets is noteworthy. The latter strengthens the notion that the problem is not specific to a given force field, water model, or even long-range electrostatics method *per se*. In fact, additional simulations show that the critical parameter may be the strength of protein−water interactions, as proposed by Best et al.,[3] and implied in a recent study by Piana et al.[29]

2. METHODS

2.1. Model. Histatin 5 (Asp_1-Ser-His-Ala_4-Lys-Arg-His-His_8-Gly-Tyr-Lys-Arg_{12}-Lys-Phe-His-Glu_{16}-Lys-His-His-Ser_{20}-His-Arg-Gly-Tyr_{24}) initial structure was built with PyMOL[30] and assumed linear to avoid biasing subsequent conformational sampling.

2.2. Standard Molecular Dynamics (MD) Simulations. Standard MD simulations were performed with the GROMACS package,[31−34] version 4.6.2, using the GROMOS 53A6 (G53A6)[13] and 54A7 (G54A7)[14], the AMBER ff99SB-ILDN (A99SB-ILDN),[11] and ff99SBNMR1-ILDN (A99SBNMR1-ILDN)[12] force fields. Histatin 5 was solvated with 26244 TIP3P[22] and 26366 SPC[21] water molecules - AMBER and GROMOS force fields, respectively - in a rhombic dodecahedral box, with periodic boundary conditions and a minimum distance between the peptide and the box of 1 nm. The system was neutralized with 5 Cl^- ions. The equations of motion were numerically integrated using the Verlet leapfrog algorithm with a time step of 2 fs. The nonbonded interactions were treated with a Verlet list cutoff scheme in order to make use of the program's native GPU acceleration. The short-range interactions were calculated using a nonbonded pairlist with a single cutoff of 1 nm, updated every 100 fs. Long-range dispersion corrections

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431

were applied to the system's energy and pressure. PME[23] was used to handle the long-range electrostatic interactions with cubic interpolation and a grid spacing of 0.16 nm. Solvent and solute were separately coupled to temperature baths at 300 K, with the velocity rescaling thermostat[35] and a relaxation time of 0.1 ps. A Parrinello–Rahman pressure coupling[36] was used at 1 bar, with a relaxation time of 2 ps and isothermal compressibility of 4.5×10^{-5} bar^{-1}. All bond lengths were constrained using the LINCS algorithm.[37]

The minimization procedure used a combination of steepest descents[38] and l-BFGS methods.[39] Initiations were performed in a three steps scheme of 100, 200, and 300 ps. The first step is performed under NVT conditions to stabilize the temperature, while the second and third steps are performed under the NPT ensemble, with increasingly stronger coupling parameters.

Each set was simulated for a total of 1.5 μs, in triplicates of 500 ns. Snapshots were saved every 10 ps. Residue charges were set to be representative of those at pH 7 (pH value for which Histatin 5 charge capacitance is at its minimum), in accordance to a previous study using CpHMD simulations.[2] Hence, all histidines were modeled in the deprotonated form, yielding a protein net charge of +5e.

The MD trajectories were used in their entirety for analysis, meaning that the initial equilibration time was not clipped out due to being statistically negligible.

2.3. Modified MD Simulations. Modified MD simulations - with settings and parameters analogous to the CpHMD simulations in ref 2 - were performed with a nonstandard GROMACS package, version 4.0.7, modified in order to add the ionic strength as a parameter in the mdp file,[40] using the GROMOS 54A7 force field. Nonbonded interactions were treated using a twin-range cutoff of 8 and 14 Å, updating the neighbor lists every five steps or 10 fs. Long-range electrostatic interactions were treated using the generalized reaction field (GRF) method[24] with a relative dielectric constant of 54 and an ionic strength of zero. Pressure was kept constant using Berendsen's barostat[41] at 1 bar, with a relaxation time of 0.5 ps and isothermal compressibility of 4.5×10^{-5} bar^{-1}. Unless specifically stated, all other settings and parameters are identical to those used for the standard MD simulations (Section 2.2).

2.4. MD Simulations Using Balanced Protein–Water Interactions. Additional simulations - mimicking Best et al. approach - using the AMBER ff03WS (A03WS) force field[3] were performed with GROMACS version 4.6.7 and 26090 TIP4P/2005 water molecules.[27] All other settings and parameters as described in Section 2.2.

2.5. Analyses. All analyses were done using the GROMACS software package, the DSSP program (version 2.2.1),[42] and in-house tools. The calculations of correlation-corrected errors for averages over a single simulation replica have been determined using standard methods.[43] The final errors of the averages over replicates have been estimated using the law of total variance.

Theoretical SAXS intensities, $I(q)$, were computed with the program CRYSOL (version 2.8.2),[44] and the experimental pair distance distribution function, $p(r)$, was constructed from the form factor using GNOM.[45]

Principal component analysis (PCA) and, subsequentially, central structure determination, probability density estimation, and energy landscape calculation were performed using Campos and Baptista approach.[46] The main difference with respect to their approach was the use of g_rms binary dump of the comparison matrix for the determination of the central structure using a NumPy script. In the original method, g_rms was modified in order to produce the comparsion matrix in the xvg file format, which was then used for the central structure determination.

The nomenclature of each set of simulations is as follows: the shortened force field name is used in conjuction with the suffixes _PME or _GRF, e.g. G54A7_GRF. The suffixes are allusions to the long-range electrostatics method used therein. _PME for particle mesh Ewald electrostatics and _GRF for the generalized reaction field method.

2.6. Figures. All figures were generated using GNUPLOT,[47] except for the protein snapshots, which were rendered using PyMOL[30] with ray tracing.

3. RESULTS AND DISCUSSION

3.1. Standard Simulations. *3.1.1. Protein Shape and Size.* In order to assess the validity of the MD simulations using popular, state-of-the-art force field and water model combinations, hereupon referred to as *standard simulations* (i.e., no additional correction to protein–water interactions), comparison has been made with experimental data obtained by small-angle X-ray scattering, SAXS (BM29, ESRF Grenoble). This method provides important dimensional parameters and structural insights of proteins and is often employed when studying unstructured proteins, such as IDPs.[48−51]

Figures 1a-b.1 show the experimental form factor and respective Kratky representation, obtained with SAXS on a solution of Histatin 5 (1 mg/mL) in a buffer (10 mM Tris-HCl, pH 7) and monovalent salt (140 mM NaCl). Figure 1c.1 presents the pair distance distribution function, $p(r)$, obtained as described in Section 2.5. For a more detailed experimental description see the work of Cragnell et al.[52]

The Krakty representation allows easy discernment between different proteins shapes, such as compact, globular structures (Gaussian, q^{-4}), random coil (sigmoid, q^{-2}), and stiff rods (linear, q^{-1}).[53] The experimental curve exhibits asymptotic behavior with $\sim q^{-1.7}$, reaching a plateau at high q values. This is indicative of a flexible, random coil-like structure typical of unstructured proteins such as IDPs. This finding is supported by the work of Raj et al. and Brewer et al., using (NMR)[7,8] and CD.[8,9] Histatin 5 radius of gyration, R_g, is estimated to range between ∼1.34–1.38 nm[52] (obtained from the Guinier analysis and $p(r)$, respectively).

The second row of Figure 1 shows the same analyses as before but now including simulation results. Comparison between the experimental and theoretical form factors (Figure 1a.2), shows that the theoretical curves deviate appreciably, which implies that the protein structures determined by SAXS and simulations differ. More specifically, in the low q region the experimental curve decays faster, indicating a larger R_g. At higher q the opposite is observed, and the theoretical curves acquire a steeper slope, indicating a more compact average structure. The deviation between experiment and simulation is specially evident in Figure 1b.2, where the theoretical curves show a markedly different profile, incompatible with evidence showing that Histatin 5 exists as a flexible, random-coil-like structure in aqueous solution.[7−9] Instead, these curves are characteristic of compact, globular structures with limited flexibility. They exhibit a maximum around $q \cdot R_g \approx \sqrt{3}$, and beyond $q \cdot R_g \geq 3$ their profiles increase or level out, depending on the simulation. A99SB-ILDN_PME seems to give rise to the most extended conformations and G54A7_GRF the least (in agreement with Figure 2 and Table 1). Their profiles flatten out and decay with $\sim q^{-2.5}$ and $\sim q^{-3.5}$, respectively, in the intermediate $q \cdot R_g$ region.

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431

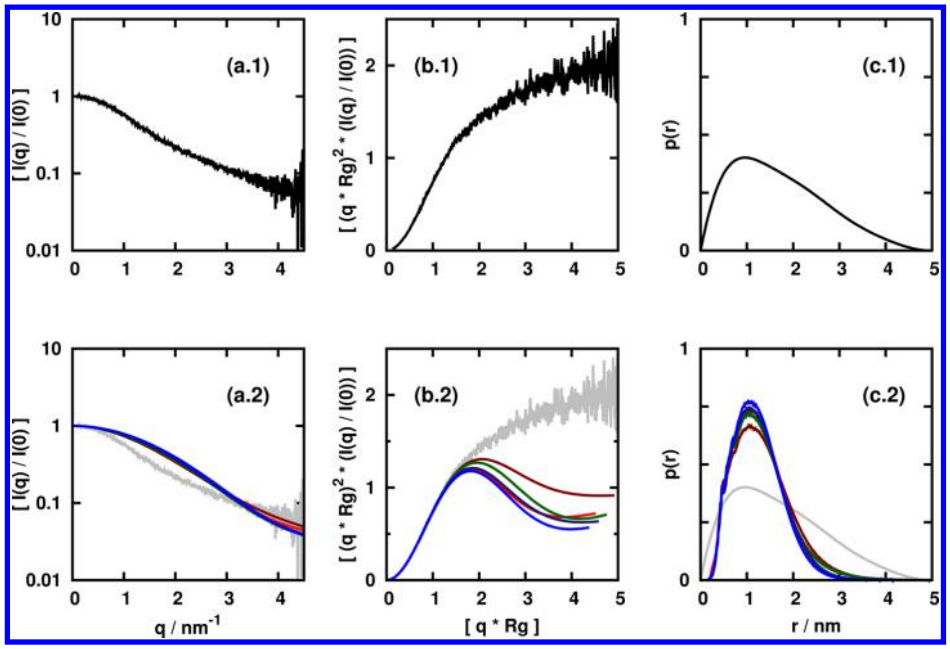


Figure 1. (a) Form factor determined by SAXS and simulations. (b) Kratky plots using the experimental and computed $I(q)$, q, and R_g. R_g(exp) ≈ 1.38 nm and the respective computed radii are presented in Table 1, under $\langle R_g \rangle_{CRYSOL}$. (c) Nonweighted radial distribution functions, for each set of simulations. These are directly comparable to the indirect Fourier transform of the experimental SAXS intensity (**black**/**gray** curve), which represents the density of intraprotein pair distances, $p(r)$. Color code: A99SB−ILDN_PME - **maroon**; A99SBNMR1−ILDN_PME - **red**; G53A6_PME - **green**; G54A7_PME - **navy blue**; G54A7_GRF - **blue**; experimental data - **black**/**gray**.

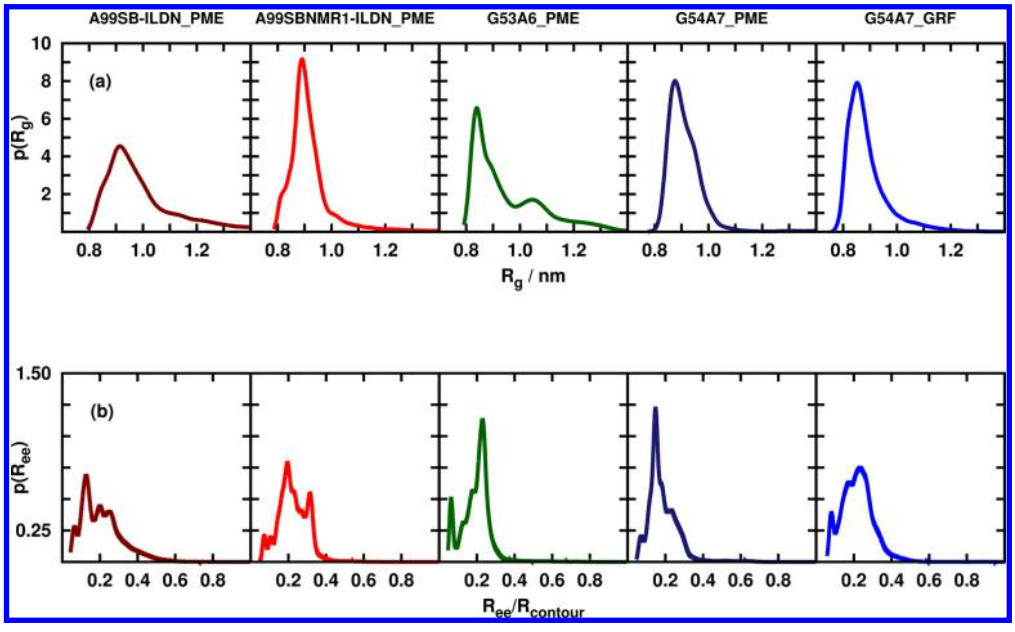


Figure 2. (a) Density estimates of the radius of gyration, $p(R_g)$, and (b) end-to-end distance, $p(R_{ee})$, for each set of simulations, using a Gaussian kernel. The latter is normalized with the contour length of the protein, $R_{contour}$, and thus a ratio of 1 indicates that the protein is fully extended.

In sum, the structures obtained from the investigated standard simulations are more compact than the experimentally determined average conformation, although some flexibility is observed. This notion is reinforced by the results shown in Figure 1c.2, where experiment and simulations differ greatly on the range of pair distances sampled, with all simulations showing similar profiles among themselves. This is a clear indication that (i) the simulations oversample shorter intraprotein atomic distances, and thus more compact structures for Histatin 5; (ii) and this is invariant to the force field, water model, and long-range electrostatics method used therein.

Figure 2 presents the density estimates of the radius of gyration, $p(R_g)$, and end-to-end distance, $p(R_{ee})$, for each set of simulations. A99SBNMR1−ILDN_PME, G54A7_PME, and G54A7_GRF show relatively similar $p(R_g)$ profiles (see Figure 2a), with a sharp single peak of maximum value at ∼0.85−0.90 nm. The computed $\langle R_g \rangle$ values presented in Table 1 are marginally larger than the values shown here. This is a direct consequence of the asymmetry of these curves. A99SB−ILDN_PME differs from the previous examples in the width of its peak, suggesting that it samples a more varied array of R_g values. Notice the limited x-axis range, 0.8−1.4 nm, with very

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431

Table 1. Average Radius of Gyration, $\langle R_g \rangle$, and Standard Deviation, σ, for Each Set of Simulations, Using Different Methods[a]

force field	$\langle R_g \rangle_{MD} \pm \sigma$ (nm)	$\langle R_g \rangle_{CRYSOL} \pm \sigma$ (nm)
A99SB-ILDN_PME	1.00 ± 0.15	0.96 ± 0.18
A99SBNMR1-ILDN_PME	0.92 ± 0.09	0.88 ± 0.10
G53A6_PME	0.96 ± 0.16	0.93 ± 0.17
G54A7_PME	0.93 ± 0.12	0.90 ± 0.13
G54A7_GRF	0.89 ± 0.09	0.85 ± 0.10

[a] $\langle R_g \rangle_{MD}$ was computed with GROMACS g_gyrate using 150 000 conformations. $\langle R_g \rangle_{CRYSOL}$ was computed by averaging the R_g values from the slope of the net intensities obtained with CRYSOL. For the latter, 10 000 evenly spaced frames were picked from the original MD trajectory.

low $p(R_g)$ values for $R_g \geq 1.1$ nm. This is confirmation that these simulations are indeed oversampling rather compact structures of similar radii, a finding which has been documented in several other atomistic MD studies of IDPs and unfolded proteins.[2−6] G53A6_PME is the only set to present two distinct maxima, ∼0.85 and 1.05 nm. While the first is a better representation of its structural ensemble, the latter has repercussion on the high standard deviation of its respective entry in Table 1.

$\langle R_g \rangle$ estimates presented in Table 1 are statistically similar within the associated error, both between simulation sets and within the same set, depending on whether the estimation was performed directly from the MD trajectory or indirectly through CRYSOL. Numerical differences between columns are consistent along each row, pointing out toward a systematic difference between methods due to the value used for the contrast of hydration shell in CRYSOL, i.e. zero $e/Å^3$.

The $p(R_{ee})$ plots shown in Figure 2b are normalized by the contour length in order to give a relative indication of the expansion of the protein. As can be seen, most of these curves show rugged profiles, specially when compared with the $p(R_g)$ curves (top row). This may be an indication of limited sampling. In fact, depending on the property being analyzed, the simulations may or may not be fully converged. The order of convergence for our system - with respect to simulation time - is as follows: (1) energy; (2) radius of gyration; (3) end-to-end distance (R_{ee}); and (4) secondary structure (see Figures S1−S5 in the Supporting Information). The radii of gyration of all simulation sets tested here are converged, whereas the end-to-end distances show larger variance with simulation time and in between replicates.

Nevertheless, most simulation sets rarely sample more expanded conformations, spanning more than ∼40% of the protein contour length, which is in agreement with all previously discussed results, confirming the compact nature of the sampled conformations.

3.1.2. Secondary Structure. The secondary structure can be described as the general three-dimensional form of local segments of biopolymers such as proteins. It is often defined by the patterns of hydrogen bonds between backbone amide groups but can also be established by regular patterns of backbone dihedral angles in a particular region of the Ramachandran plot.[54,55] In the previous Section (3.1.1) it was mentioned that several independent experimental studies using SAXS,[52] CD,[8−10] and NMR[7,8] indicate that Histatin 5 behaves as a random coil in aqueous solution. Hence, it is interesting to test whether the standard simulation sets studied here are able to produce results compatible with experimental evidence. More so, when considering that some force fields have been previously described as having an intrinsic bias toward certain secondary structure classes.[19,20,56]

Figures 3a-e provide the secondary structure histograms for each set of standard simulations. As clearly shown, most simulations yield structured conformational ensembles, with little to no agreement between them, both in terms of type and percentage. These results are in contradiction with the available experimental evidence for Histatin 5[7−10,52] and IDPs in general.

Rule-based secondary structure assignment programs, such as DSSP, are known for their limited accuracy and systematic errors (see ref 57). However, the latter is not the sole cause for the predictions shown in Figure 3b, for example, where unreasonably high helical content is present, with some residues being involved in such structural motif for almost 50% of the total simulation time. Additionally, it cannot account for the high degree of structural variance between sets, of which G53A6_PME and G54A7_PMEa are two good examples. The former indicates that Histatin 5 has a tendency to form β-sheets, and the latter leans in favor of helical content. This contradiction may be explained by the fact that, as addressed earlier in the Introduction (Section 1), G53A6 is known to underestimate helical content.[14,19,20] This is the most likely reason for the lack of helical content, as shown in Figures 3c and 3f, but it does not explain the high content in β-sheet. As such, the results for G53A6_PME could be explained by lack of convergence. This seems plausible taking into consideration the asymmetric form of the $p(R_g)$ (Figure 2a), and the fact that in order to quantitatively estimate structural properties such as the secondary structure, one may need substantially longer simulation time scales, as suggested by Figures S3−S5 in the Supporting Information. In fact, Piana et al. show that secondary structure rearrangements can take up to several tens of microseconds of simulation time.[29]

There is also a contrast between G54A7_PME and G54A7_GRF, Figures 3d-f, respectively. These simulations differ mainly regarding how the long-range electrostatics are handled (particle mesh Ewald vs generalized reaction field), and this was not apparent in the results shown in Section 3.1.1.

Finally, there is a significant deviance between AMBER and GROMOS force fields. The AMBER varieties consistently overestimate the structural character of Histatin 5, with large emphasis on helical content and turns, a trend which is not seen in CD studies.[9] DSSP assigns, erroneously, 3_{10}-helix-like turns to the turn class.[57] Because it is possible that some percentage of the turn class could be a part of the helical content, the former is explicitly shown in Figure 3. Figures 3a and 3b cannot be described as similar within the associated uncertainty (see Figure 3f). Despite their intrinsic preferences, GROMOS force fields generally yield less structured conformations for Histatin 5 and in this sense display better agreement with experimental evidence.

It should be mentioned that AMBER ff99SB*-ILDN[11,16,17] may provide a better description of the α-helical propensities when compared to AMBER ff99SB-ILDN (or A99SB-ILDN), due to the inclusion of additional backbone energy corrections.

Apart from the force field and long-range electrostatics method, the water models used herein may also play a role in the divergence of results for the secondary structure prediction between AMBER and GROMOS force field varieties. This requires further investigation and will not be addressed in this work.

3.1.3. Principal Component Analysis (PCA). Principal component analysis, PCA, is a mathematical method whose main purpose is to reduce the dimensionality of a data set while retaining most of its variation.[58,59] For most systems it is

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431

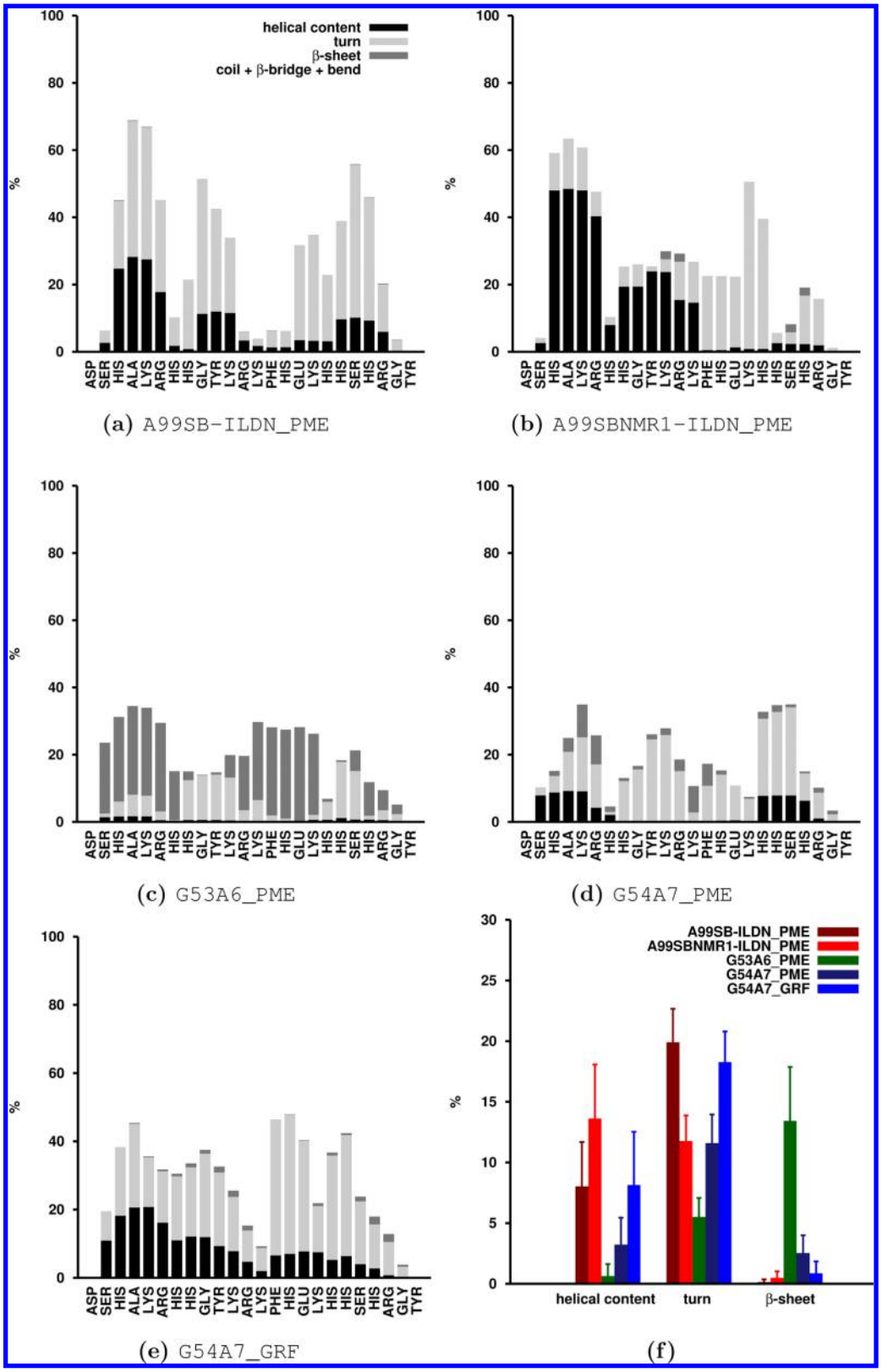


Figure 3. (a-e) Stacked secondary structure histograms per residue for each set of simulations. All histograms sum up to 100%, but coil, β-bridge, and bend percentages are not shown for visual simplicity. (f) Average percentage (and associated error) of each secondary structure motif per simulation set.

unfeasible to map their complete $(3N - 6)$-dimensional energy landscape and thus the use of PCA in order to map the conformational space onto a low representation space that retains the most important features of the distribution of conformations. It is common practice to analyze the backbone atoms only, thus removing several dimensions of the complete landscape. Protein translation and rotation can also be partially removed by fixing its center of mass and performing a least-squares fitting of the

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431

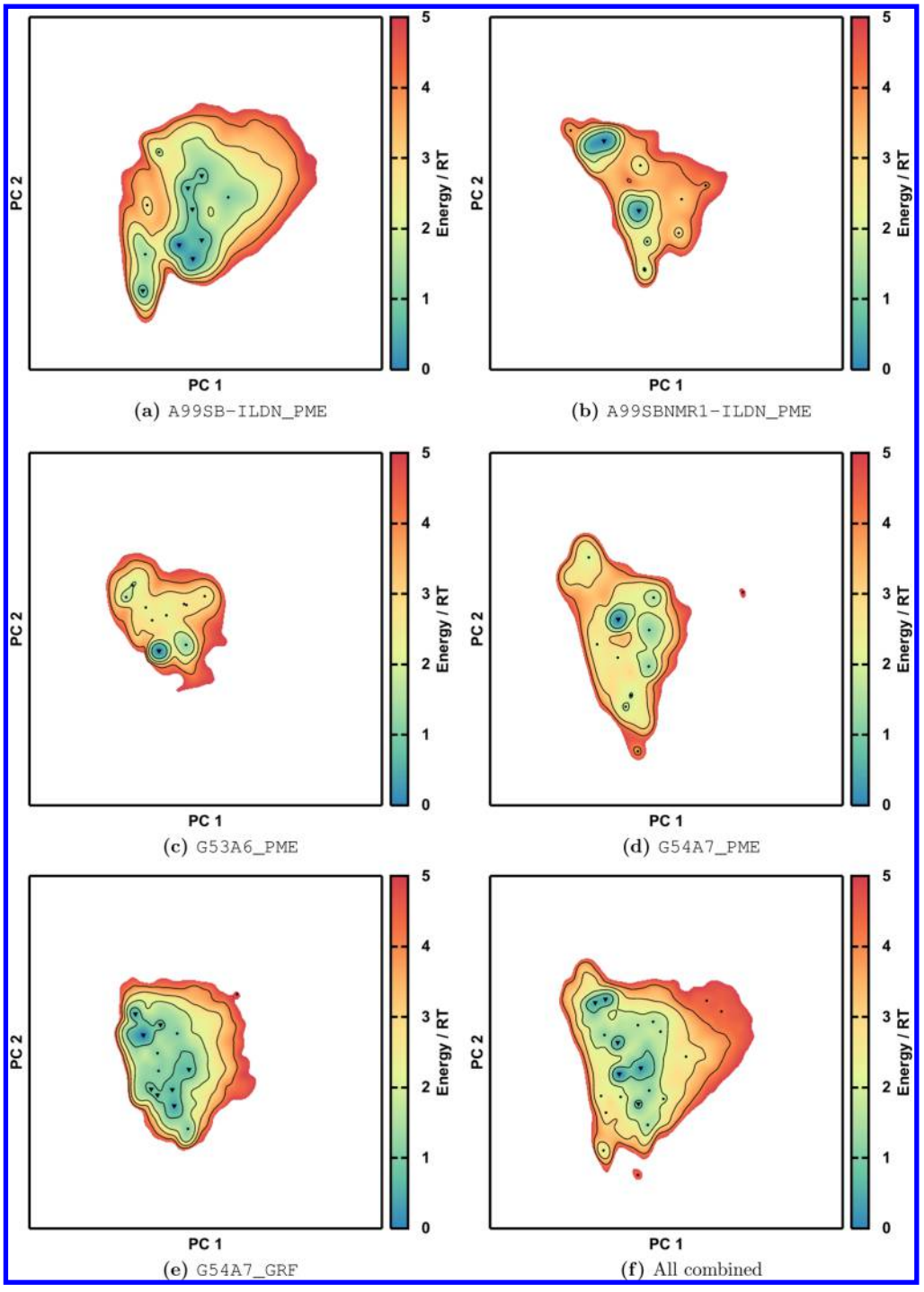


Figure 4. Energy landscapes for each simulation using the first two PCs. Contour lines are drawn for energies between $0 \leq RT \leq 5$, with 1 RT increment. Triangles and circles mark the center of each basin with $RT \leq 1$ and $1 < RT \leq 5$, respectively.

trajectory to a reference structure. Here, this structure is the central structure of each simulation, as defined by Campos and Baptista.[46]

The number of principal components (PCs) needed to capture a substantial percentage of the total variance depends on the system. Here, to attain between 70− 90% of the total variance, at least ∼5 PCs per simulation need to be considered, which means that the energy landscapes should contain five dimensions or more. Nevertheless, our aim for this section is to simply understand whether the standard simulations under consideration are able to produce energy landscapes typical of flexible proteins, such as IDPs, or not. The two first PCs, i.e. two dimensions, are sufficient for such endeavor.

Figure 4 shows the energy landscapes computed from the probability density projected along the first two PCs for each set of simulations, using a common basis set. Flexible proteins are expected to have energy landscapes with multiple energy minima, and this is captured by all standard simulations tested here. Yet, only A99SB-ILDN_PME, G54A7_GRF, and the combination of all simulations display small energy barriers between the lowest energy minima. This is specially visible for G54A7_GRF in Figure 4, where all minima are contained within a central

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431

plateau with energy $\leq 2RT$. G53A6_PME and G54A7_PME (Figures 4c-d) comprise several energy minima but one is much deeper than the others, and there are considerably high energy barriers between regions/minima. This is in line with the results presented before, suggesting that the system may be energetically trapped in some cases (see Figure S6 in the Supporting Information), resulting in a limited sampling of the conformational landscape. The same applies to A99SBNMR1-ILDN_PME, where two regions/minima are separated by a $\sim 4RT$ energy barrier. These types of energy landscapes are more characteristic of classical proteins with metastable intermediate states than those of flexible proteins.

Table 2 comprises all lowest energy basins with $E_{min} \leq 1RT$ (triangular marks in Figure 4), the respective percentage of structures therein, their average radius of gyration, $\langle R_g \rangle$, and associated standard deviation, σ. Figure 5 depicts the representative structure of each basin shown in Table 2. As shown, most basins comprise structures with similar $\langle R_g \rangle$ (within statistical uncertainty), with some representative structures from different simulation sets exhibiting strikingly similar features (A #7, B #2 and E #6; or D #1 and E #1, for example). This implies that either two PCs are not enough to discern between conformational classes or that each simulation is (over)sampling very closely related structures. The latter is supported by the results shown in Figures 1c and 2, discussed earlier in Section 3.1.1. Some of the central structures obtained by combining all simulations are easily identified from the individual constituents (F #3 → B #1; F #4 → C #1; and F #6 → E #3).

Table 2. Energy Landscape Basins with $E_{min} \leq 1RT$, Sorted by the Number of Structures They Comprise, in Percentage[a]

name	basin no.	$E_{min}(RT)$	no. structures (%)	$\langle R_g \rangle \pm \sigma$ (nm)
A	1	0.78	18.00	1.03 ± 0.12
	2	0.54	13.45	0.98 ± 0.07
	3	0.16	11.27	0.91 ± 0.06
	4	0.00	10.32	0.89 ± 0.06
	5	0.65	8.00	0.92 ± 0.06
	6	0.75	6.12	0.94 ± 0.06
	7	0.81	4.42	0.95 ± 0.06
B	1	0.00	44.53	0.90 ± 0.05
	2	0.20	33.06	0.90 ± 0.05
C	1	0.00	25.77	0.85 ± 0.04
D	1	0.00	26.15	0.88 ± 0.04
E	1	0.63	23.79	0.95 ± 0.10
	2	0.00	16.38	0.84 ± 0.04
	3	0.35	12.72	0.89 ± 0.05
	4	0.83	8.48	0.89 ± 0.07
	5	0.83	7.10	0.84 ± 0.03
	6	0.61	6.87	0.87 ± 0.05
	7	0.42	5.42	0.87 ± 0.03
	8	0.86	1.91	0.84 ± 0.04
F	1	0.24	22.68	0.84* ± 0.09*
	2	0.37	9.53	0.78* ± 0.07*
	3	0.48	8.42	0.80* ± 0.06*
	4	0.00	8.41	0.76* ± 0.05*
	5	0.38	7.71	0.80* ± 0.06*
	6	0.81	6.83	0.77* ± 0.06*

[a]These represent the lowest energy basins in Figure 4 (filled triangles), enclosed by the lowest energy contour line. Keys: A - A99SB-ILDN_PME; B - A99SBNMR1-ILDN_PME; C - G53A6_PME; D - G54A7_PME; E - G54A7_GRF; and F - All combined. Values with an asterisk refer to statistical quantities computed using the backbone only. All others use all protein atoms comprised in the MD topology (all atoms for AMBER, and all but nonpolar hydrogens for GROMOS variants).

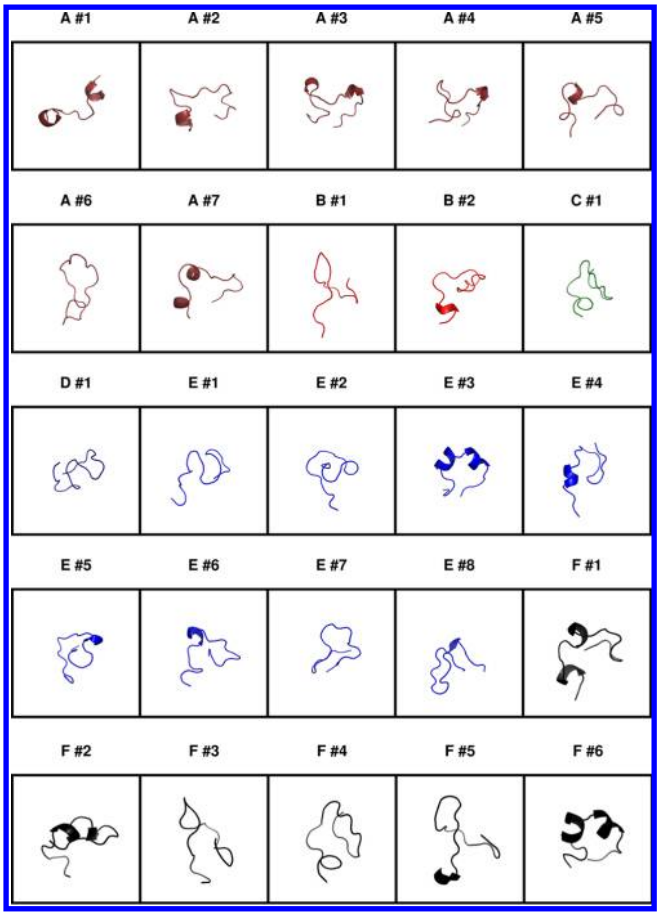


Figure 5. Central structure of each basin in Table 2. Keys: A - A99SB-ILDN_PME; B - A99SBNMR1-ILDN_PME; C - G53A6_PME; D - G54A7_PME; E - G54A7_GRF; and F - All combined.

The representative structures of each low energy basin show a tendency to exhibit helical secondary structure motifs. The two exceptions are G53A6_PME and G54A7_PME (C #1 and D #1, respectively), a finding which is also in close agreement with the results presented in Figure 3 (Section 3.1.2).

3.2. Simulations with Balanced Protein–Water Interactions. Recently Best et al. published an article[3] regarding the importance of having a proper balance in protein–water interactions. By modifying the protein–water Lennard-Jones parameters in AMBER ff03W,[60] the authors obtained AMBER ff03WS, which, when used in combination with the TIP4P/2005 water model, is reported to recover the correct dimensions of IDPs and unfolded proteins, while not affecting the stability of the native states of structured proteins.

Figure 6 comprises the same type of analysis as in Figure 1 but now focusing on the aforementioned approach (A03WS_PME). A99SB-ILDN_PME simulations are included as reference to the standard simulations analyzed earlier in this work (Section 3.1). The experimental results are now reproduced with a higher level of accuracy. Both A03WS_PME and experimental SAXS form factor curves match one another with precision, and this is well reflected in the respective Kratky representation, where the curve exhibits asymptotic behavior with q^{-2}, a perfect example of random coil behavior.[48] Consequently, A03WS_PME simulations also sample a more

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420–3431

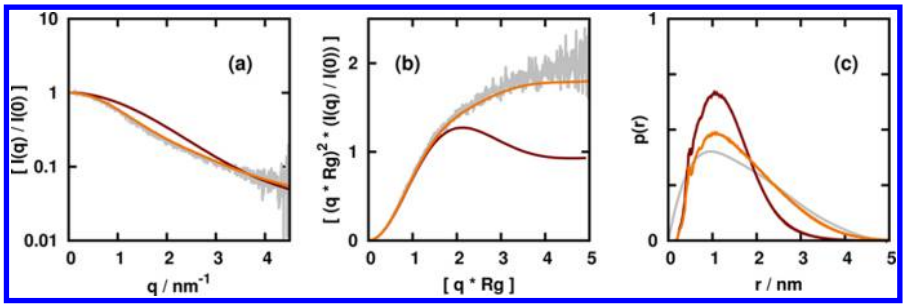


Figure 6. (a) Form factor determined by SAXS and simulations. (b) Kratky plots using the experimental and computed $I(q)$, q, and R_g. R_g(exp) ≈ 1.38 nm, and the respective computed radius is presented in Tables 1 and 3, under $\langle R_g \rangle_{CRYSOL}$. (c) Computed nonweighted radial distribution functions vs the indirect Fourier transform of the experimental SAXS intensity, $p(r)$. These are directly comparable and represent the density of intraprotein pair distances. Color code: A03WS_PME - **orange**; A99SB-ILDN_PME - **maroon**; experimental data - **gray**.

realistic distribution of intraprotein pair distances (Figure 6c), resulting in a considerable improvement upon previous results shown in Figure 1c.

In Figure 2, we saw that the density estimates of the R_g were too narrow, with a range between ∼0.8−1.0 nm. A modest increase of the protein−water interactions results in the sampling of a broader array of radii (∼0.9−1.9 nm), in agreement with what would be expected for a random coil (see Figure 7a). Notice the sharp difference between A03WS_PME and A99SB-ILDN_PME, despite the fact that the latter has the broadest profile of all standard simulations presented earlier.

As a consequence the $\langle R_g \rangle$ values presented in Table 3 are now statistically indifferentiable from the proposed experimental values (∼1.34−1.38 nm[52]).

The density estimate of the end-to-end distance, $p(R_{ee})$ - normalized with the contour length (Figure 7b) - now displays a symmetric form, with more expanded conformations being sampled. This raises interesting questions regarding force field dynamics, as it seems like standard force fields possess slower dynamics when compared with the approach discussed herein (see Figure S2 in the Supporting Information, for example).

As discussed earlier (Section 3.1.2), different force fields exhibit marked preferences (biases) toward certain secondary structure motifs. The same may apply to AMBER ff03WS, and from Figures 7b-d it is noticeable that some transient structure exists in the form of helical content. Interestingly, all force fields tested in this work show a predisposition toward helicity for this system, with the exception being GROMOS 53A6, which is known to underestimate this structural class.[14,19,20] On the other hand, A03WS_PME produces less structured ensembles when compared to its next of kin, i.e. A99SB-ILDN_PME and A99SBNMR1-ILDN_PME. Experimental evidence[7] does not seem to accommodate the notion that Histatin 5 may exhibit transient structure in aqueous solution, although it is known that Histatin 5 exhibits helical structure in organic solutions (2,2,2-trifluoroethanol, for example) and in lipid bilayers.[7,8,61] Hence, it may be possible that Histatin 5 presents transient helical content in aqueous solution.

The energy landscape shown in Figure 7c attests, once more, to the appropriateness of A03WS_PME for the simulation of IDPs. It contains one single large and flat central plateau ($E \leq 1RT$) where all energy minima are located. These are separated by shallow energy barriers of no more than $0.5RT$. The values presented in Table 4 show that each basin comprises structures of statistically distinct $\langle R_g \rangle$ and thus distinct conformational classes. This suggests that high dimensionality is not always necessary in order to properly characterize the density states in the principal component space.

The representative structures of each basin are depicted in Figure 8. These structures exhibit quite extended conformations, compatible with the notion of a flexible unstructured protein. Direct visual comparison with Figure 5 reinforces this conclusion. The similarities between the representative structures in basins #1 and #2 are due to their close proximity within the energy landscape (upper two minima in Figure 7e). Both basins belong to a single larger basin that comprises a considerable area of the landscape. The energy barrier between both conformational classes is lower than one hundredth RT, a value which is statistically insignificant. Even with the very fine mesh used for the density estimation, it is difficult to make a distinction between the values of these energy minima.

As mentioned earlier, an alternative approach by Piana et al.[4] has recently been proposed. Judging from the published results,[3,4] there is good reason to expect both methods to produce similar results for the system at hand. Furthermore, the fact that both approaches share common ground could be an indication that regardless of the path chosen, be it the development of a brand new water model or just some minor modifications to existing force field(s), stronger protein−water interactions are a requirement for the appropriate MD simulation of IDPs, such as Histatin 5.

4. CONCLUSION

All *standard* simulations studied in this work are shown to be equally inept for the simulation of Histatin 5, the model IDP. Moreover, the results prove that the force fields used in these simulations exhibit considerable bias toward overly compact conformational ensembles (force field independent) and certain secondary structure motifs (force field dependent). This could, in part, explain why they work well for structured proteins, which are mostly globular and compact, and exhibit a preference for helical structure. While these features may (over)stabilize structured proteins, they also considerably hinder their transferability for the study of unstructured, flexible proteins, ending up working in the opposite direction, by restricting their natural highly flexible character.

The water model plays an equally important role in the quality of the simulations. In fact, water molecules constitute the bulk of any explicitly solvated simulation of proteins, and the common bias toward overly collapsed structures may be a sign of poor solvation instead. Indeed, the specific modification of protein−water Lennard-Jones parameters in AMBER ff03W, as proposed by Best et al.,[3] leads to marked improvements. The use of this approach results in more extended and flexible conformational ensembles, and, more importantly, better

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431

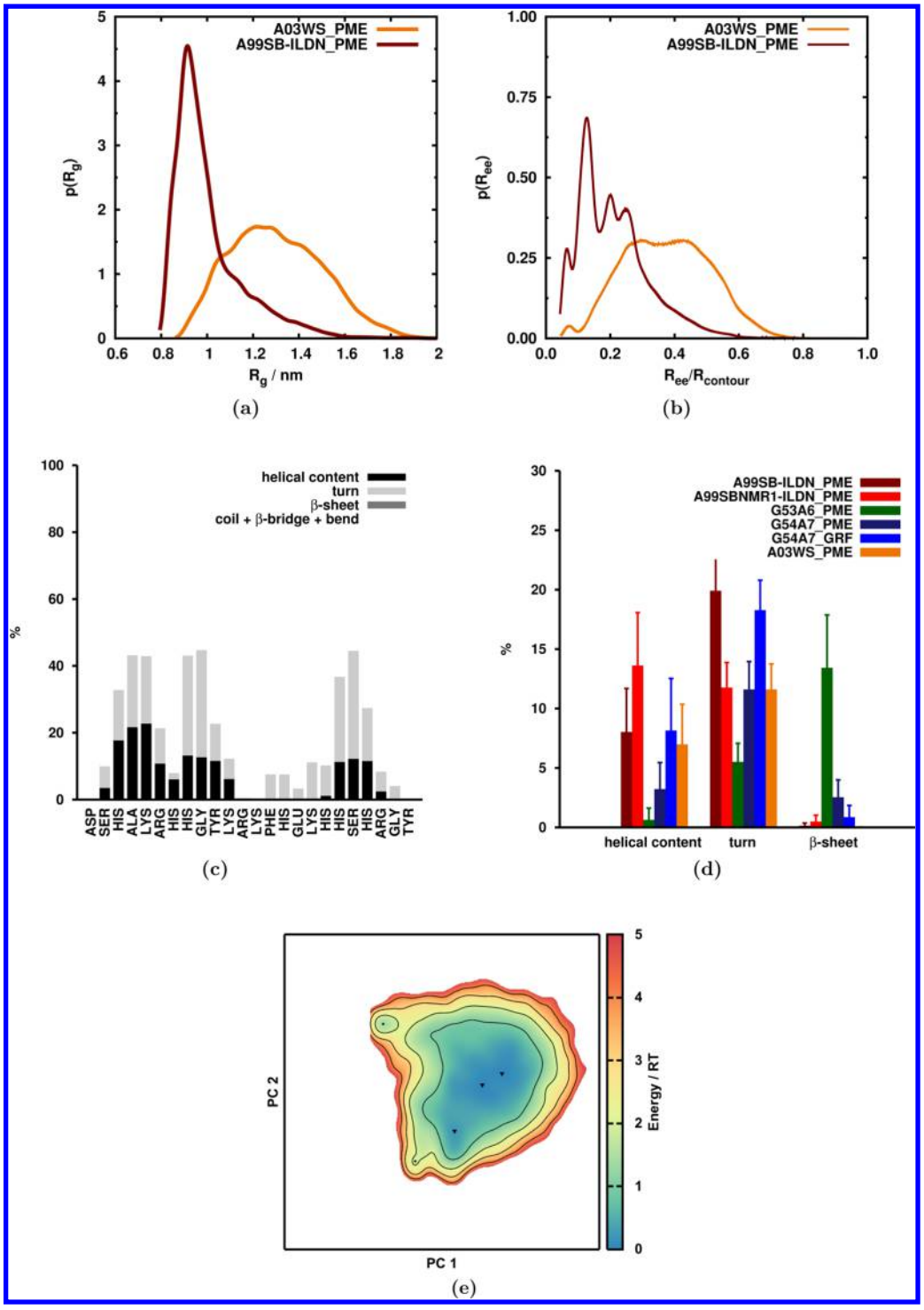


Figure 7. (a) Density estimate of the radius of gyration, $p(R_g)$, and (b) end-to-end distance, $p(R_{ee})$, using a Gaussian kernel for A03WS_PME and A99SB-ILDN_PME. (c) Stacked secondary structure histograms per residue. All histograms sum up to 100%, but coil, β-bridge, and bend percentages are not shown for visual simplicity. (d) Average percentage (and associated error) of each secondary structure motif per simulation set. (e) Energy landscape using the first two PCs. Contour lines are drawn for energies between $0 \leq RT \leq 5$, with 1 RT increment. Triangles and circles mark the center of each basin with $RT \leq 1$ and $1 < RT \leq 5$, respectively. Subfigures (c) and (e) are relative to A03WS_PME only.

reproduces the experimental SAXS data. The theoretical $\langle R_g \rangle$ obtained from the simulations, 1.30 ± 0.20 nm (see Table 3), is statistically indifferentiable from the experimental estimate, ~1.34−1.38 nm. Its respective energy landscape presents a large central plateau with several energy minima of similar depth, separated by very small energy barriers, attesting to the highly flexible character of Histatin 5. The conformational classes comprised in each minima's basin are statistically distinguishable in terms of their radii, which is not the case for the standard simulations tested here. Their energy minima are separated by higher energy barriers (over $4RT$ in some cases), often with one minimum deeper than the others, containing a considerable share of the total conformational ensemble. The conformational classes are also less well-defined in terms of their size, presenting

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431

Table 3. Average Radius of Gyration, $\langle R_g \rangle$, and Standard Deviation, σ, for A03WS_PME, Using Different Methods[a]

force field	$\langle R_g \rangle_{MD} \pm \sigma$ (nm)	$\langle R_g \rangle_{CRYSOL} \pm \sigma$ (nm)
A03WS_PME	1.30 ± 0.20	1.23 ± 0.23

[a] $\langle R_g \rangle_{MD}$ was computed with GROMACS g_gyrate using 150 000 conformations. $\langle R_g \rangle_{CRYSOL}$ was computed by averaging the R_g values from the slope of the net intensities obtained with CRYSOL. For the latter, 10 000 evenly spaced frames were picked from the original MD trajectory.

Table 4. A03WS_PME Energy Landscape Basins with $E_{min} \leq 0.5\ RT$, Sorted by the Number of Structures They Comprise, in Percentage[a]

basin no.	$E_{min}(RT)$	no. structures (%)	$\langle R_g \rangle \pm \sigma$ (nm)
1	0.00	43.57	1.23 ± 0.12
2	0.00	37.14	1.50 ± 0.13
3	0.21	17.46	1.09 ± 0.09

[a] These represent the lowest energy basins in Figure 7e (filled triangles), enclosed by the lowest energy contour line.

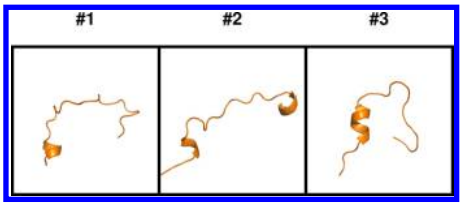


Figure 8. Central structure of each A03WS_PME energy landscape basin presented in Table 4.

similar $\langle R_g \rangle$ within statistical uncertainty. This is a direct consequence of the rather limited range of radii being sampled or, in other words, the oversampling of excessively collapsed structures. This seems to be insensitive to the force field and water model used. Simulations using the generalized reaction field, GRF, method for treating long-range electrostatics produce more flexible ensembles than their respective counterpart using PME (in agreement with ref 40).

While the discrepancies in the results obtained with standard simulations and those using balanced protein−water interactions are mostly a direct consequence of the more favorable protein−water interactions of the latter (lessening the overly hydrophobic character of water in typical MD simulations), it is not possible to infer the same in regard to the ubiquitous presence of transient secondary structure in all simulations analyzed here. This particular matter seems to be subject to a more prominent influence of the force field rather than the water model or even the protein−water interaction strength, which is not entirely unexpected.[19,20,56]

■ ASSOCIATED CONTENT

Supporting Information

This section contains a comment on the estimation of the contrast of hydration shell and additional analyses of the individual replicates of each simulation set. Figures S1−S2 and S4−S5: Time evolution of the radius of gyration, end-to-end distance, and secondary structure (α-helical and β-sheet content), respectively. Figure S3: Time average and error of the radius of gyration and end-to-end distance. Figure S6: Energy landscapes. Figure S7: Contrast of hydration shell parameter scan. The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/ct501178z.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 46 462 220 530. E-mail: joao.henriques@teokem.lu.se.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We acknowledge financial support from Organizing Molecular Matter (OMM), Vinnova, the Vinnmer program, The Royal Physiographic Society in Lund, Per-Eric and Ulla Schybergs Foundation, and the Crafoord Foundation. The authors would like to thank António Baptista and Sara Campos for providing their topography analysis tool, and Robert Best, Wenwei Zheng, and Jeetain Mittal for making AMBER ff03WS available before its public release. Finally, the authors would also like to thank António Baptista, Miguel Machuqueiro, Diogo Vila Viçosa, Luís Filipe, Paulo Costa and João Damas for fruitful discussions, and the reviewers for their insightful comments and suggestions during the review process.

■ REFERENCES

(1) Uversky, V. N.; Habchi, J.; Tompa, P.; Longhi, S.; et al. *Intrinsically Disordered Proteins (IDPs)*; ACS Publications: 2014; Vol. *114*, pp 6557−6948.
(2) Kurut, A.; Henriques, J.; Forsman, J.; Skepö, M.; Lund, M. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 657−667.
(3) Best, R. B.; Zheng, W.; Mittal, J. *J. Chem. Theory Comput.* **2014**, *10*, 5113−5124.
(4) Piana, S.; Klepeis, J. L.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2014**, *24*, 98−105.
(5) Nettels, D.; Müller-Späth, S.; Küster, F.; Hofmann, H.; Haenni, D.; Rüegger, S.; Reymond, L.; Hoffmann, A.; Kubelka, J.; Heinz, B.; et al. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 20740−20745.
(6) Palazzesi, F.; Prakash, M. K.; Bonomi, M.; Barducci, A. *J. Chem. Theory Comput.* **2015**, *11*, 2−7.
(7) Raj, P. A.; Marcus, E.; Sukumaran, D. K. *Biopolymers* **1998**, *45*, 51−67.
(8) Brewer, D.; Hunter, H.; Lajoie, G. *Biochem. Cell Biol.* **1998**, *76*, 247−256.
(9) Raj, P. A.; Edgerton, M.; Levine, M. *J. Biol. Chem.* **1990**, *265*, 3898−3905.
(10) Helmerhorst, E. J.; Van't Hof, W.; Breeuwer, P.; Veerman, E. C.; Abee, T.; Troxler, R. F.; Amerongen, A. V. N.; Oppenheim, F. G. *J. Biol. Chem.* **2001**, *276*, 5643−5649.
(11) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950−1958.
(12) Li, D.-W.; Brüschweiler, R. *Angew. Chem.* **2010**, *122*, 6930−6932.
(13) Oostenbrink, C.; Villa, A.; Mark, A. E.; Van Gunsteren, W. F. *J. Comput. Chem.* **2004**, *25*, 1656−1676.
(14) Schmid, N.; Eichenberger, A.; Choutko, A.; Riniker, S.; Winger, M.; Mark, A.; Van Gunsteren, W. *Eur. Biophys. J.* **2011**, 843−856.
(15) Wang, J.; Cieplak, P.; Kollman, P. A. *J. Comput. Chem.* **2000**, *21*, 1049−1074.
(16) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct., Bioinf.* **2006**, *65*, 712−725.
(17) Best, R. B.; Hummer, G. *J. Phys. Chem. B* **2009**, *113*, 9004−9015.
(18) Huang, J.; MacKerell, A. D. *J. Comput. Chem.* **2013**, *34*, 2135−2145.
(19) Matthes, D.; de Groot, B. L. *Biophys. J.* **2009**, *97*, 599−608.
(20) Villa, A.; Fan, H.; Wassenaar, T.; Mark, A. E. *J. Phys. Chem. B* **2007**, *111*, 6015−6025.
(21) Berendsen, H.; Postma, J.; van Gunsteren, W.; Hermans, J. *Intermol. Forces* **1981**, *11*, 331−342.
(22) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431

(23) Darden, T.; York, D.; Pedersen, L. *J. Chem. Phys.* **1993**, *98*, 10089.
(24) Tironi, I.; Sperb, R.; Smith, P.; van Gunsteren, W. *J. Chem. Phys.* **1995**, *102*, 5451−5459.
(25) Baptista, A.; Teixeira, V.; Soares, C. *J. Chem. Phys.* **2002**, *117*, 4184−4200.
(26) Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G. L.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665−9678.
(27) Abascal, J. L.; Vega, C. *J. Chem. Phys.* **2005**, *123*, 234505.
(28) Wang, L.-P.; Martínez, T. J.; Pande, V. S. *J. Phys. Chem. Lett.* **2014**, *123*, 234505.
(29) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. *J. Phys. Chem. B* **2015**, *119*, 5113−5123.
(30) *The PyMOL Molecular Graphics System*, Version 1.3; Schrödinger LLC.
(31) Berendsen, H.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43−56.
(32) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306−317.
(33) van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A.; Berendsen, H. *J. Comput. Chem.* **2005**, *26*, 1701−1718.
(34) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435−447.
(35) Bussi, G.; Donadio, D.; Parrinello, M. *J. Chem. Phys.* **2007**, *126*, 014101.
(36) Parrinello, M.; Rahman, A. *J. Appl. Phys. (Melville, NY, U. S.)* **1981**, *52*, 7182.
(37) Hess, B.; Bekker, H.; Berendsen, H.; Fraaije, J. *J. Comput. Chem.* **1997**, *18*, 1463−1472.
(38) Fletcher, R.; Powell, M. *Comp. J.* **1963**, *6*, 163−168.
(39) Liu, D.; Nocedal, J. *Math. Prog.* **1989**, *45*, 503−528.
(40) Machuqueiro, M.; Baptista, A. *J. Phys. Chem. B* **2006**, *110*, 2927−2933.
(41) Berendsen, H.; Postma, J.; van Gunsteren, W.; DiNola, A.; Haak, J. *J. Chem. Phys.* **1984**, *81*, 3684−3690.
(42) Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577−2637.
(43) Allen, M.; Tildesley, D. *Computer simulation of liquids*; Oxford University Press: 1989; Vol. *18*.
(44) Svergun, D.; Barberato, C.; Koch, M. *J. Appl. Crystallogr.* **1995**, *28*, 768−773.
(45) Svergun, D. *J. Appl. Crystallogr.* **1992**, *25*, 495−503.
(46) Campos, S. R.; Baptista, A. M. *J. Phys. Chem. B* **2009**, *113*, 15989−16001.
(47) Williams, T.; Kelley, C.; et al. *GNUPLOT: An interactive plotting program*, Version 4.6.
(48) Mittag, T.; Forman-Kay, J. D. *Curr. Opin. Struct. Biol.* **2007**, *17*, 3−14.
(49) Eliezer, D. *Curr. Opin. Struct. Biol.* **2009**, *19*, 23−30.
(50) Uversky, V. N. *J. Biomol. Struct. Dyn.* **2003**, *21*, 211−234.
(51) Boze, H.; Marlin, T.; Durand, D.; Pérez, J.; Vernhet, A.; Canon, F.; Sarni-Manchado, P.; Cheynier, V.; Cabane, B. *Biophys. J.* **2010**, *99*, 656−665.
(52) Cragnell, C.; Durand, D.; Cabane, B.; Skepö, M. Submitted for publication.
(53) Glatter, O.; Kratky, O. *Small angle X-ray scattering*; Academic Press: 1982.
(54) Chou, P. Y.; Fasman, G. D. *Biochemistry* **1974**, *13*, 211−222.
(55) Munoz, V.; Serrano, L. *Proteins: Struct., Funct., Bioinf.* **1994**, *20*, 301−311.
(56) Yoda, T.; Sugita, Y.; Okamoto, Y. *Chem. Phys.* **2004**, *307*, 269−283.
(57) Zacharias, J.; Knapp, E. W. *J. Chem. Inf. Model.* **2014**, *54*, 2166−2179.
(58) Jolliffe, I. *Principal component analysis*, 2nd ed.; Springer-Verlag: New York, 2002.
(59) Rencher, A. C.; Christensen, W. F. *Methods of multivariate analysis*, 3rd ed.; John Wiley & Sons: 2012.
(60) Best, R. B.; Mittal, J. *J. Phys. Chem. B* **2010**, *114*, 14916−14923.
(61) Situ, H.; Balasubramanian, S. V.; Bobek, L. A. *Biochim. Biophys. Acta, Gen. Subj.* **2000**, *1475*, 377−382.

DOI: 10.1021/ct501178z
J. Chem. Theory Comput. 2015, 11, 3420−3431