



Effect of solvent model when probing protein dynamics with molecular dynamics



Samuel Genheden

Department of Chemistry and Molecular Biology, University of Gothenburg, Box 462, SE-405 30 Göteborg, Sweden

ARTICLE INFO

Article history:

Received 15 August 2016

Received in revised form 2 November 2016

Accepted 3 November 2016

Available online 9 November 2016

Keywords:

Molecular dynamics

Order parameters

Protein dynamics

Implicit solvent

Coarse-graining

ABSTRACT

We probe the dynamics of the Bpti and Galectin-3 proteins using molecular dynamics simulations employing three water models at different levels of resolution, viz. the atomistic TIP4P-Ewald, the coarse-grained Elba and an implicit generalised Born model. The dynamics are quantified indirectly by model-free order parameters, S^2 of the backbone N–H and selected side-chain bond vectors, which also have been determined experimentally through NMR relaxation measurements. For the backbone, the order parameters produced with the three solvent models agree to a large extent with experiments, giving average unsigned deviations between 0.03 and 0.06. For the side-chains, for which the experimental data is incomplete, the deviations are considerably larger with mean deviations between 0.13 and 0.17. However, for both backbone and side-chains, it is difficult to pick a winner, as all models perform equally well overall. For a more complete set of side-chain vectors, we resort to analysing the variation among the estimates from different solvent models. Unfortunately, the variations are found to be sizeable with mean deviations between 0.11 and 0.15. Implications for computational assessment of protein dynamics are discussed.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Molecular dynamics (MD) simulations are today a common technique to probe the structure and dynamics of biomolecules, their function and activity in order to elucidate biochemical processes and pathways [1–3]. A key to their success is the accurate modelling of interactions in an aqueous environment. Because of the importance of solute–water and water–water interactions, the modelling of solvation effects has been the subject of a large body of theoretical and computational work [4–6]. Currently, the most common approach is to treat solvent effects by modelling each water molecule individually with a molecular mechanics force field. Typically, a water molecule is represented by either three or four particles and therefore, to ensure proper solvation, water molecules constitute somewhere between 80 and 90% of the particles in a protein simulation. Although highly optimized codes are implemented in commonly used software, it is still not surprising that these models are computationally costly. Therefore, a lot of effort have been made in accelerating the MD simulations and thereby reaching longer timescales with less computational effort [7]. Herein, we distinguish between methods that retain the atomistic description

of the solvent and methods employing simplified water models. Within the former group, we find methods that for instance bias the forces [8–10], exchange thermodynamic variables between concurrent simulations [11,12] or combine several shorter simulations [13,14] to mention a few. However, we will in this report focus on the latter group.

A common approach to simplify an atomistic description is coarse-graining (CG) [15,16], where sets of atoms are grouped into pseudo-particles, or beads. This reduces the number of particles that needs to be simulated and as a side effect smoothening the energy landscape and thereby accelerating the sampling further. The CG approach has been applied to all sorts of molecules, e.g. proteins [17,18], DNA [19,20], lipids [21,22] and sugars [23]. In addition, several CG water models have been proposed that group the atoms of between one and five real water molecules into a single bead [24–26]. Some of these models have also been combined with an atomistic representation of a protein [27–29]. In particular the Elba (electrostatics based) model, which maps the atoms of a single water molecule to a CG bead, has proven to be useful in predicting the solvation free energy of small molecules [29,30]. A speed-up about six compared to a fully AA simulation can be expected for a protein solvated in a box of water molecules, and better parallel performance can probably be achieved by improving load-balancing [29,30]. This hybrid approach has been shown to preserve the structure of a small protein, and been used to study

E-mail address: samuel.genheden@gu.se

the folding of a peptide [29]. However, it is unclear how well the dynamics of proteins or protein–ligand complexes are described in longer MD simulations.

A more drastic reduction in the number of simulated particles can be achieved by dropping the particle description of the solvent and use an implicit model [31]. Here, the effect of the solvent molecules is included as an effective force that is a function of only the solute coordinates. The most commonly used implicit models are based on continuum dielectric theory where the system is divided into regions with different dielectric constants [32]: in the case of a hydrated protein, the water molecules are represented by a high dielectric constant and the protein is modelled by a low dielectric constant. The electrostatic contribution to the hydration of a protein can be accurately computed with the Poisson–Boltzmann approach [33], but becomes too computationally costly to use in long MD simulations. Therefore, alternatives have been devised and many are some variant of the generalised Born (GB) formalism [34–38]. The models typically differ in the approach to compute the extent of burial of individual atoms and how to define the boundary between the dielectric regions. Combined with a simple solvent-accessible surface-area (SASA) model for the non-polar solvation part, GB models have been used in a wide range of applications, e.g. solvation free energies [39,40], protein folding [41,42] and binding affinities [43]. In addition, GB models have been used to probe protein dynamics [44]. The computational efficiency of GB simulations is hampered by the necessity of long cut-off and the lack of efficient parallelisation, but the smoothening of the energy landscape typically compensate for this [45].

Because of the attractiveness to use CG and GB models to study proteins it is imperative to compare them to atomistic simulations. This has been done several times in the past [46–48], but the comparisons have mostly been focused on structural aspects. In this paper, we will compare all-atom, CG and GB models on their ability to probe the protein dynamics of two proteins Bpti and Galectin-3. The latter protein will be studied in complex with two different ligands. To quantify the dynamics of the protein, we will compute order parameters of selected bond vectors, which are accessible experimentally through nuclear magnetic resonance (NMR) relaxation measurements [49]. Such order parameters have been used extensively to study protein dynamics [50,51] and conformational entropy [52–54] or in rational drug design [55]. In many instances, MD simulations have been used as complements to the NMR experiments [53], which typically cannot report on the full set of order parameters. Therefore, it is important that computation and experiment agree on the bond vectors available to both. As such, the use of order parameters as a benchmark provides a sensitive test of alternative computational models and in fact, has been used to compare protein force fields [56,57]. Herein, we will instead focus on different solvent models combined with an accurate protein force field. We will attempt to rationalise the deviations observed in the different simulations and to provide recommendations for future studies.

2. Method

2.1. System setup

The prepared protein structures of Bovine Pancreatic trypsin inhibitor (BPTI) and Galectin-3 (Gal3) were taken from a previous study [58]. Gal3 was in complex with either lactose or a synthetic lactose derivative (L02), and these complexes will be referred to as Gal-lac and Gal-l02 to be consistent with previous studies. The protein and ligands are shown in Fig. 1. The proteins, lactose and L02 were described with the Amber99SB [59], Glycam [60] and GAFF force fields [61], respectively. For the all-atom and

all-atom/coarse-grained (AA/CG) hybrid simulations, each of the proteins was solvated in a box of pre-equilibrated TIP4P-Ewald [62] or Elba water, respectively, extending at least 0.9 nm from the solute. The Elba water model is a CG representation of a single water molecule, in which a point dipole is embedded on a single Lennard-Jones site [26].

2.2. Simulations

The simulations with CG solvent were performed with the LAMMPS software [63]. The masses of the hydrogen atoms were increased three-fold and the masses of the neighbouring atoms were decreased through mass repartitioning [64]. This allows the all-atom particles to be propagated with a 4 fs timestep. A multiple timestep integrator was used and the CG water beads were propagated with a 8 fs timestep [65]. The SHAKE algorithm [66] was used to constrain bonds with hydrogen atoms. The non-bonded cut-off was 1.2 nm and long-range electrostatics between all-atom particles were calculated with the particle–particle particle–mesh algorithm [67]. The interactions between CG beads and between CG beads and atoms were calculated with a shifted-force potential and no long-range correction was applied. The temperature was kept constant at 300 K using a Langevin thermostat [68] and the pressure was controlled with an isotropic weak-coupling algorithm [69].

All generalised Born (GB) simulations were performed with the Amber suite of programs [70]. The polar part was the GB model of Onufriev, Bash and Case [36] (model 2) and the non-polar part was computed by γ SASA, where SASA is the solvent-accessible surface area and $\gamma = 2.1$ kJ/mol/nm². The modified Bondi radii [36] were used and SASA was computed with the LCPO method [71]. The atoms were propagated with a 2 fs timestep and SHAKE [66] was used to constrain bonds with hydrogen atoms. Both the non-bonded cut-off and the Born cut-off was 1.5 nm. The temperature was kept constant at 300 K using a Langevin thermostat [68].

In both the CG and GB setup, the systems were subjected to 100 steps of steepest descent followed by 20 ps equilibration in the NPT ensemble where the heavy atoms of the protein were restrained to the starting position with a 4184 kJ/mol/nm² harmonic force. This was followed by a 200 ps equilibration in the NVT ensemble without any restraints and a 500 ns production run in which snapshots were saved every 10th ps. Two independent repeats were initiated by assigning different initial velocities to the atoms.

The AA simulations were taken from a previous study [58]: Bpti was equilibrated for 20 ps in the NPT ensemble, following a 100 ps equilibration in the NVT ensemble and a 500 ns production run, also in the NVT ensemble. Two independent repeats were initiated by assigning different initial velocities. The Gal-lac and Gal-l02 complexes were equilibrated for 200 ps in the NVT ensemble, followed by a 500 ns production run, also in the NVT ensemble. Only a single simulation for each complex was performed. The sampling frequency for all three systems was 10 ps.

2.3. Analysis

Order parameters for different sets of bond vectors were calculated with the isotropic reorientational eigenmode dynamics (iRED) approach [72]. First, the following covariance matrix was computed

$$M_{ij} = \frac{1}{2} \langle 3(\bar{\mu}_i \bar{\mu}_j)^2 - 1 \rangle \quad (1)$$

where $\bar{\mu}$ is a bond vector and the brackets indicate an ensemble average. Second, the eigenvalues, λ_m and eigenvectors \bar{m} were cal-

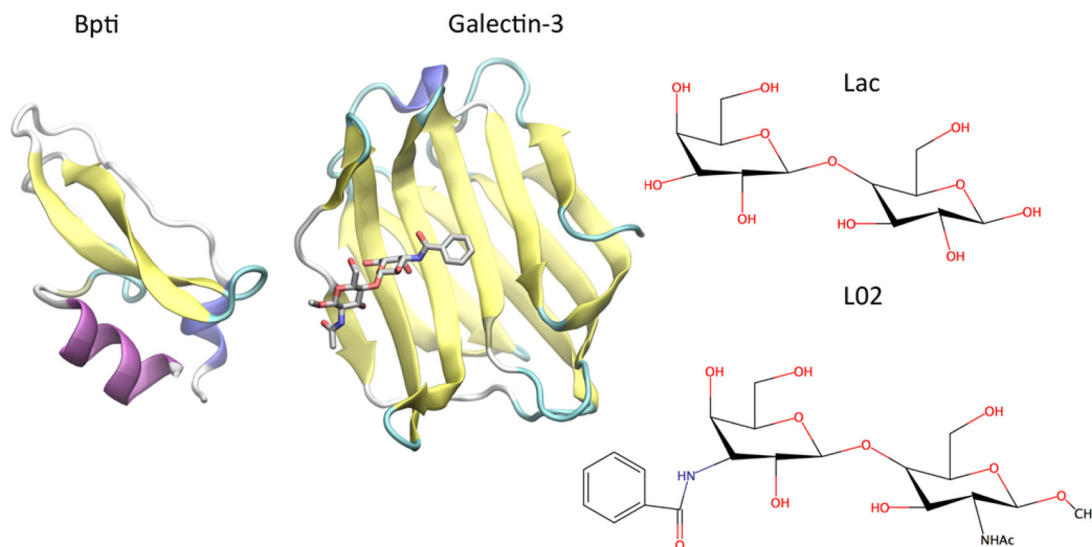


Fig. 1. The protein and ligands used in the study. Galectin-3 is bound with L02.

culated and the order parameter, S_i^2 , for a residue i was computed from

$$S_i^2 = 1 - \sum_{m=6}^n \lambda_m |m_i|^2 \quad (2)$$

where m_i is the i th element of m . The order parameters were computed for sub-segments of the trajectories and averaged [73]. The length of these sub-segments or windows are discussed in the text. Three different sets of order parameters were calculated: 1) backbone N–H bond vectors, 2) side-chain C–C or S–C bond vectors in Ala, Ile, Met, Leu, Thr and Val residues, and 3) side-chain vectors defined in the dictionary proposed by Li and Brüschweiler [44]. These sets will be referred to as NH, Met and Dict, respectively.

The root mean square fluctuation (RMSF) for selected atoms was computed from

$$\text{RMSF}_i = \frac{8}{3} \pi^2 \text{Var}(\vec{r}_i) \quad (3)$$

where \vec{r} is the position vector of the atom after removal of overall motion by fitting the backbone C, C_α and N atoms onto the initial structure. The RMSF was estimated for backbone C_α atoms or the initial atom of the bond vectors in the Dict set [44].

Systematic errors stemming from particular amino acids were analysed using the approach suggested by Mobley et al. for the study of solvation free energies [74]. Therefore, the BEDROC (Boltzmann-enhanced discrimination of receiver-operating characteristic) metric [75] was computed for different amino acid. A labelled list of individual residues was made using the unsigned error compared to either experiment or AA as scores for the BEDROC analysis, which was computed with the CROC python package (version 1.1) [76]. The uncertainty of the BEDROC metric was estimated by 500 bootstrap iterations. The estimates for the different solvent models as well as for the different proteins were pooled together.

Hydrogen bonds were analysed by monitoring distances less than 0.3 nm and angles larger than 120° between a pre-set list of standard hydrogen bond acceptor and donors in protein residues. The analysis was performed on atoms in 1) the backbone, 2) the side-chains or 3) all atoms.

Analyses of the trajectories were performed with in-house tools based on the MDAnalysis framework [77].

3. Results and discussion

We performed long (500 ns) molecular dynamics simulations of the Bpti and Galectin-3 (Gal3) proteins and analysed the dynamics using the iRED approach [72]. The all-atom proteins were simulated in all-atom (AA) TIP4P-Ewald water, in coarse-grained (CG) Elba water, or with an implicit generalised Born (GB) model.

3.1. iRED analysis of backbone motion

The iRED approach is a robust method to compute order parameters for selected bond vectors from a molecular dynamics trajectory [72]. Here we use it to indirectly quantify the dynamics of the proteins because we can make direct comparisons to order parameters derived from NMR relaxation experiments.

The order parameters for the backbone N–H vectors (NH set) were compared to parameters from ^{15}N relaxation experiments, [55,78]. To make a fair comparison, the computed order parameters must correspond to roughly the same timescales as in the experiments. The experimental timescales are limited by the overall rotational timescales, which in the case of Bpti and Gal3 is between 3 and 8 ns [55,78]. It has been argued previously that the computed order parameters should be on the order of five times this length [79], i.e. 15–40 ns. However, these recommendations were made for AA simulations, and it is unclear to what the extent the dynamics in the CG and GB simulations are accelerated. Therefore, we compared the experimental order parameters to iRED estimates with windows of 2.5, 5, 10, 25, 50, 100 and 250 ns. The 2.5 ns window was considered to be the minimum appropriate window size as only 250 snapshots are used to estimate the order parameters in each window. The order parameters are compared with the average unsigned deviation (AUD) and mean signed deviation (MSD) in Table S1. To make the analysis easier, we pooled the data for the two Gal3 systems, as they were very similar. For Bpti, we find a minimum AUD with a window size of 2.5 ns with all solvent models. However, t -tests show that the AUDs for window sizes of 5 to 50 ns are not significantly different at the 95% confidence level. If we instead look at the MSD, we find that the smallest deviations are found with window sizes of 250 ns, 10 ns and 2.5 ns for AA, CG and GB, respectively. For AA the MSD is increasing with increased window size, whereas the opposite is true for GB. For AA the MSDs for window sizes of 5–100 ns are in fact not significantly different from the minimum MSD. For CG, this is true for window sizes

of 2.5–25 ns, and for GB for window sizes of 2.5–10 ns. Looking at Gal3, we see similar trends: the minimum AUD is found for window sizes of 5, 5 and 2.5 for AA, CG and GB, respectively. For AA and CG, the AUDs for other window sizes are not significantly different from the minimum AUD, whereas for GB this is only true for window sizes of 5 and 10 ns. The minimum MSD is found for 100, 50 and 2.5 ns for AA, CG and GB, respectively. For AA, all window sizes between 25 and 250 gives a MSD not significantly different from the minimum MSD. For CG, this range is also 25–250 ns and for GB it is 2.5–10 ns. Thus we can conclude that the optimal window size is highly dependent on both the protein and solvent model, which is expected considering the different smoothness of the energy landscape. For AA it seems to be appropriate with a long window of 50–100 ns, for CG it seems appropriate with a shorter window of 10–50 ns, and for GB it should probably be less than 10 ns. Therefore, to be consistent, we will use a window size of 50, 25 and 5 ns for AA, CG and GB in the rest of this paper.

In Fig. 2a, we plot the average unsigned deviation (AUD) for the three different protein systems simulated with three different solvent models. The uncertainty of the AUD was estimated with bootstrapping and is less than 0.01 for all comparisons. This implies that the analysis is insensitive to which part of the trajectory is being analysed. From Fig. 2a, it is clear that the all-atom (AA) TIP4P-Ewald water model best reproduces the experimental order parameters with AUDs between 0.03 and 0.06. The Elba coarse-grained (CG) model gives AUDs between 0.05 and 0.06, whereas the generalised Born (GB) model gives an AUD of 0.06 for both Bpti and Gal-I02. Unfortunately, the Gal-lac complex dissociated when simulating with the GB model and therefore we do not have any estimates for this trajectory. This is further analysed and discussed in the Supporting information. Still, it seems that for the NH set, no model outperforms the others with a considerable margin.

Naturally, the AUDs could be skewed by a few bad estimates, and therefore we list in Table S2, the median unsigned deviation (MUD). It is clear that this metric tend to be smaller than the AUD with 0.01–0.02 units. For both Bpti and Gal-lac, the MUD is identical for each of the solvent models, whereas there is a very small difference between them for Gal-I02.

It is also of interest to investigate if the deviations are systematic or random in nature and to this end, we list in Table S2, the mean signed deviation (MSD). The AA and CG estimates show small MSD for all three proteins, indicating that these estimates do not contain any systematic deviation. The GB simulations of Bpti and Gal-I02 give estimates that seem to be systematically and slightly more negative than the experimental data, with MSDs of 0.02–0.04.

We also compared the computed order parameters to the experimental data but only for those vectors that are on residues in helices or sheets (i.e. excluding residues in loops and bridges). This is a simple approach to see if the discrepancies between NMR and MD are located to regions of the proteins that are inherently more flexible. The metrics for this analysis are shown in Table S2. It is clear that the AUDs are reduced significantly with up to 0.03 units, whereas the MUDs and MSDs are less affected. It is interesting to note that the MSD is increased for AA when looking at Bpti and Gal-lac.

Finally, we performed a BEDROC analysis [74] to investigate if there are some amino acids that are estimated significantly worse than expected. The BEDROC analysis will indicate if there are some amino acids that tend to have a larger error compared to the other amino acids. An analytical expression can be computed for the BEDROC value assuming the predictive power is equal for all amino acids. Therefore, if the observed BEDROC metric for an amino acid is significantly larger than the analytical value, it is an indication that the MD simulation is particularly bad at predicting the dynamics of this amino acid. The analytical and observed BEDROC values for the amino acids are plotted in Fig. 3. For only two amino acids we observe a BEDROC value that is significantly larger than the analytical value, viz. Cys and Trp. The observed value for Cys, 0.54 is in fact rather small and the large observed value for Trp, 0.74 could be attributed to the few sampling points (only 5 Trp residues in total). Thus, it seems that there is no strong evidence that some amino acids are harder to model.

In summary, the experimental backbone order parameters are reproduced well by all three solvent models with deviations comparable to other studies [53,57]. Furthermore, it is encouraging that the deviations do not stem from any particular amino acid or from

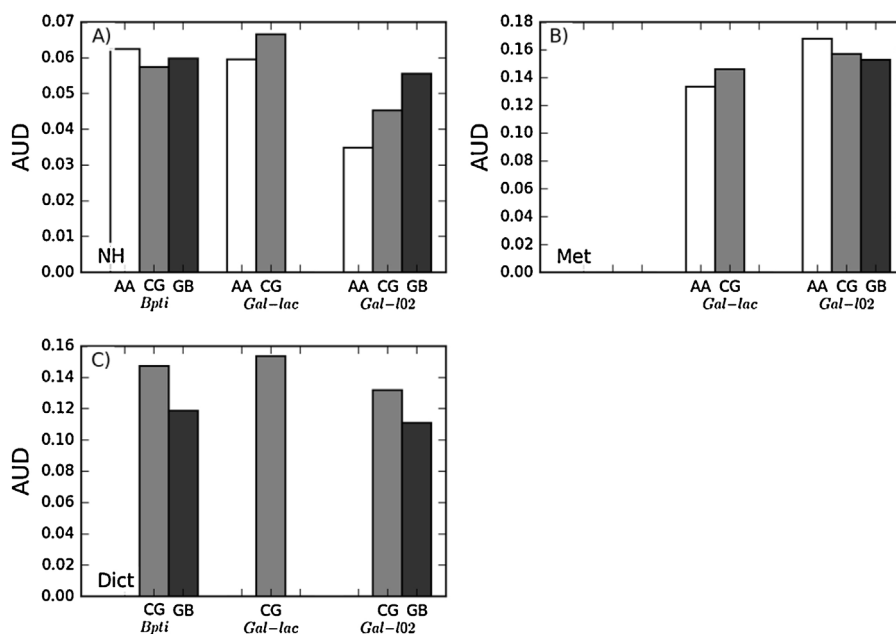


Fig. 2. Average unsigned deviations (AUD) for S^2 estimates in the sets A) NH (N–H bond vectors), B) Met (C–C bond vectors) and C) Dict (side-chain dictionary vectors). The solvent model is indicated below the bars, and the proteins are highlighted in italics under each set of bars. In A) and B), the reference is experimental data, and in C), the reference is AA. In B) there is no data for Bpti due to a lack of experimental reference.

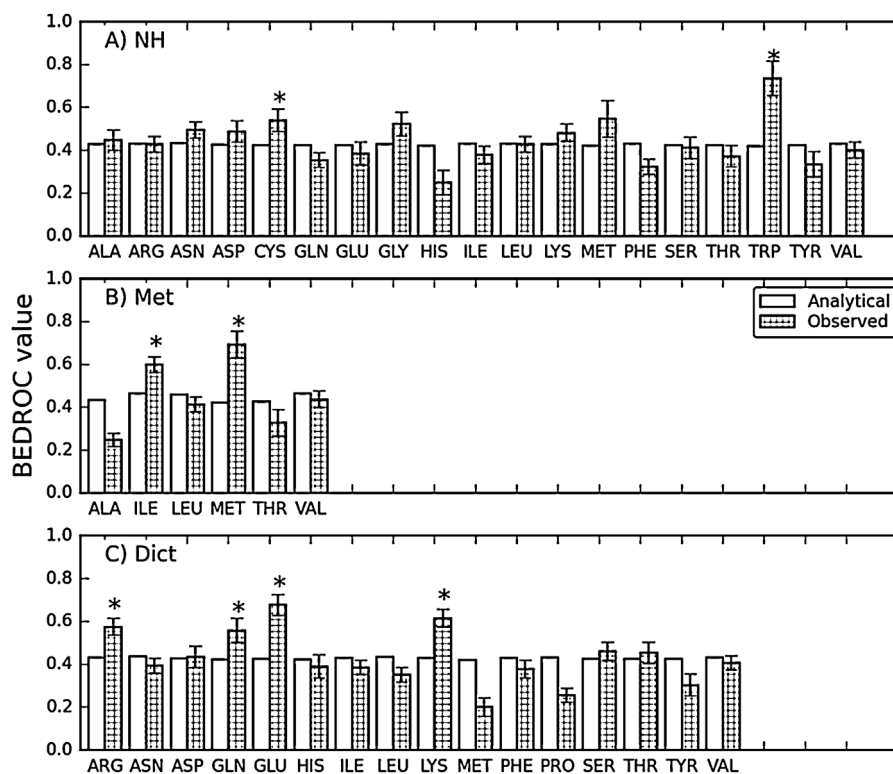


Fig. 3. BEDROC values for S^2 estimates of individual amino acids in a) NH set, b) Met set and c) Dict set. One standard deviation of the observed BEDROC is indicated with errors bars, and the observed values that are significantly larger (at 95% confidence) than the analytical value are marked with a star.

highly mobile regions, e.g. loops. This shows that MD simulations can reliably probe the backbone dynamics of proteins.

3.2. iRED analysis of side chain motion

For the Galectin-3 systems, we can compare the side-chain dynamics for a limited set of vectors that probe methyl groups with ^2H relaxation experiments [55]. Unfortunately, no such data is available for Bpti, making this comparison impossible. The AUDs of the computational estimates compared to experiments are plotted in Fig. 2b. The observed deviations are much larger than for the backbone parameters: using AA, the AUD is 0.13 and 0.17 for Gal-lac and Gal-I02, respectively. In light of this, the simpler water models give encouraging results; CG gives an AUD of 0.15 and 0.16, for Gal-lac and Gal-I02, respectively and the GB estimates for Gal-I02 give an AUD of 0.15. A t -test shows that the differences between the solvent models are not statistically significant at the 95% confidence, and hence the simpler models are of similar accuracy as the AA model. In Table S3, we list the MUDs and although they are smaller, indicating that there are a few outliers, the magnitude of the MUDs is still worrisome. The MUDs ranges between 0.9 and 0.13, and at least for Gal-I02, no water model performs better than the other. Looking at the MSD in Table S3, it is clear that neither the AA nor the CG model produces estimates with a strong systematic component. However, the GB estimates for Gal-I02 seem to be slightly more negative than the experimental data, indicating more flexible side-chains.

We also performed the same analysis when excluding residues in loops and bridges. The metrics for this analysis are listed in Table S3 and the results are mixed. For AA and GB estimates there are no significant differences compared to the analysis of all residues, whereas the CG estimates become significantly better. The MSD does not change significantly, when excluding residues, indicating

that for the methyl motion, the deviations between simulations and experiment are not localised to helices or sheets. We also performed a BEDROC analysis and the results for the different amino acids are plotted in Fig. 3b. Only Ile and Met display a significantly larger observed BEDROC value of 0.60 and 0.70, respectively; it is hard to observe any significant trend from this.

To investigate the side-chain motion further, we used a more complete set of side chain vectors as defined in a proposed dictionary of Li and Brüschweiler [44]. Unfortunately, we are now unable to compare to experimental data and instead we choose to look at the deviations compared to the AA estimates. The AUDs are plotted in Fig. 2c and the MUDs and MSDs are listed in Table S4. The observed differences compared to the AA model are on the same order as with the Met set. For CG, the AUDs are between 0.13 and 0.15, whereas the AUDs for GB are between 0.11 and 0.12. Thus for both Bpti and Gal-I02, GB is closer to AA. This trend is removed if we instead look at the MUD. If we look at the MSD, we see that the systematic deviation between the AA and GB estimates are smaller in magnitude than the systematic deviations between the AA and CG estimates. Interestingly, the GB estimates indicate a more flexible protein than the protein in AA water, whereas the opposite is true for the CG estimates.

The analysis on residues in helices and sheets follows the trend observed for the Met set (see Table S4). The AUDs and MUDs are up to 0.04 units significantly smaller when excluding some residues. However, the observed systematic shifts quantified in the MSD remains. Finally, we also performed a BEDROC analysis and the values are plotted in Fig. 3c. The observed BEDROC values for four of the residues with the largest side chains, Arg (0.57), Gln (0.56), Glu (0.68) and Lys (0.62) are significantly larger than the analytic value. Thus, it seems that the length of the side chain might influence the predictive power of the iRED approach although this is far from established.

Table 1

Average RMSD in nm compared to the initial structure for the first and second half of the trajectories.

	<i>Bpti</i>			<i>Gal-lac</i>		<i>Gal-I02</i>		
	AA	CG	GB	AA	CG	AA	CG	GB
First half	0.088	0.201	0.170	0.096	0.176	0.103	0.175	0.168
Second half	0.089	0.231	0.175	0.099	0.202	0.104	0.214	0.179
Difference	0.001	0.029	0.005	0.003	0.026	0.001	0.039	0.011

3.3. Protein stability and atomic fluctuations

In order to corroborate and complement the results from the iRED approach, we performed some simpler analyses. First, we computed the root mean square deviation (RMSD) of the backbone C, C α and N atoms compared to the initial structure throughout the trajectories, and then compared the average RMSD of the first and second half of the trajectory in Table 1 (the traces are shown in Fig. S1). The simulations with AA are the most stable with small differences between the first and second half of the simulations, between 1.0 and 3.5%. The simulations with the CG model are significantly more dynamic with differences between 0.03 and 0.04 nm, or 14 to 20%. Finally, the simulations with the GB model are intermediate with differences of about 0.01 nm or between 3 and 6%. Therefore, it is clear that the simpler CG and GB models lead to more fluctuating trajectories. From Fig. S1, it is seen that this is partly due to a larger variance from the mean RMSD, and partly from a longer burn-in period (especially for the CG Bpti simulation). There is no apparent drift in the RMSD, which would indicate a gradual degradation of the protein structure. It could be argued that the difference between the first and second half for the CG trajectories implies that an initial portion of the trajectories should be excluded from the simulations. However, RMSD is generally not a fair indication of thermodynamic equilibration, something that anyway is very hard to achieve.⁵⁸ Furthermore, the standard deviation from the block averaging of the order parameters, indicates that the iRED analysis is not sensitive to the observed differences in RMSD (see e.g. Table S1).

Next, we computed the root mean square fluctuation (RMSF) of selected atoms. This analysis was performed after the overall motion had been removed through fitting. We calculated the RMSF for the C α atoms and compared the fluctuations in the AA simulations with the fluctuations in the CG and GB simulations. The AUD is plotted in Fig. 4a, whereas MUD and MSD are listed in Table S5. The AUDs when comparing the CG and AA simulations are between 0.08 nm² for Gal-I02 and 0.21 nm² for Bpti. The sizable deviation for the Bpti protein is due to a systematic shift as indicated by the MSD of -0.18 nm², and it also seems to be dominated by a few residues as indicated by the significantly smaller MUD of 0.11 nm². Although the CG simulations of Bpti show more flexible C α atoms than the AA simulations, this is not reflected in a large difference in the predictive power of S² parameters (see Fig. 2a and Table S2). The large deviation between AA and CG for Bpti seems furthermore to be located to loop regions as indicated by the AUD of 0.10 nm² (see Table S5). Interestingly, the situation is reversed when looking at the analysis of the GB trajectories; the AUD for Bpti is smaller (0.10 nm²) than AUD for Gal-I02 (0.17 nm²). The discrepancies between AA and GB seem to be systematic, as indicated by the sizable MSD (-0.1 to -0.16 nm²). Furthermore, the deviations seem to stem from a few outliers as indicated by the MUD (0.05–0.08 nm²) and located in loop regions as indicated by the smaller AUD (0.04–0.10 nm²) if such residues are excluded.

We also computed the RMSF of side-chain atoms included in the Dict set. The initial atoms of each bond vector were analysed. As is clear from Fig. 4b the deviations compared to AA are larger for side-chain atoms than for backbone C α atoms. It is also clear that the differences between CG and GB are smaller. The AUDs for CG range

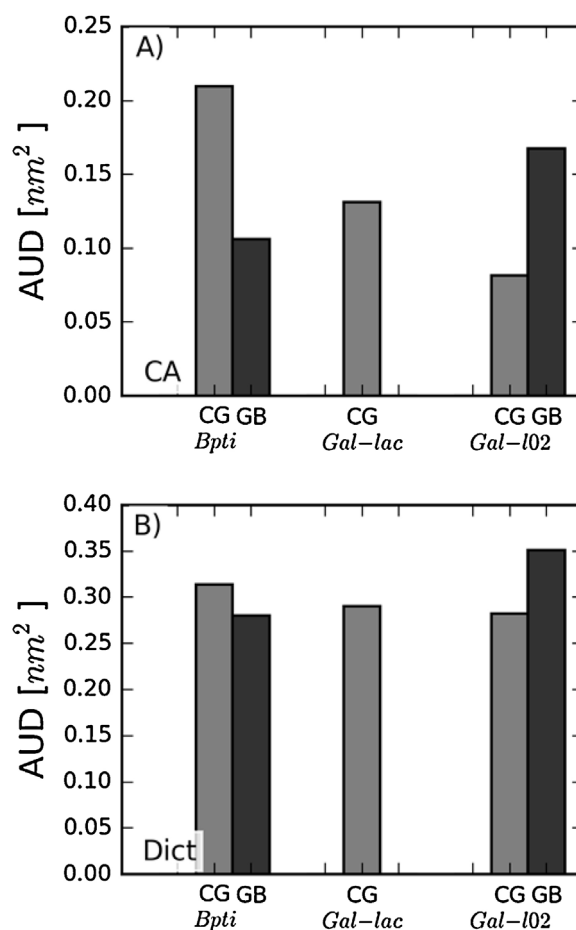


Fig. 4. Average unsigned deviations (AUD) for RMSF compared to AA simulations for a) C α atoms, b) atoms in Dict set. The solvent model is indicated below the bars, and the proteins are highlighted in italics under each set of bars.

between 0.28 and 0.31 nm² and for GB between 0.28 and 0.35 nm². The analysis of the GB trajectories and the CG trajectory of Bpti indicate systematically more flexible side-chains compared to AA as quantified by the MSD in Table S6. The MSDs of the CG simulations of Gal3 are considerably smaller. As with the RMSF of the C α atoms, the AUDs are skewed by a few outliers, which is indicated by the significantly smaller MUD (between 0.08 and 0.18 nm² smaller). It is also clear that the AUD tends to be smaller if we exclude residues in loops (between 0.10 and 0.15 nm² smaller).

3.4. Hydrogen bond analysis

A difference in hydrogen bond pattern is a potential rationale for the observed difference in dynamics between simulations with different solvent models. For instance, it has been previously observed that the Protein G displays significantly more hydrogen bonds between side-chain atoms in the CG solvent compared to all-atom solvent [29]. This could be explained by a lack of hydrogen bonds between protein and solvent, a behaviour that should also be observable in the GB model. Therefore, we counted the average number of intra-protein hydrogen bonds during the simulation and divided the hydrogen bonds into backbone–backbone (BB–BB), sidechain–sidechain (SC–SC) and backbone–sidechain (BB–SC). The averages are plotted in Fig. 5. Looking at backbone atoms, we observe that both CG and GB give rise to fewer bonds than AA, and in most cases the difference is statistically significant. CG gives about 3 fewer hydrogen bonds for all proteins, whereas GB gives between 1 and 3 fewer bonds. If we then look at bonds between side-chain

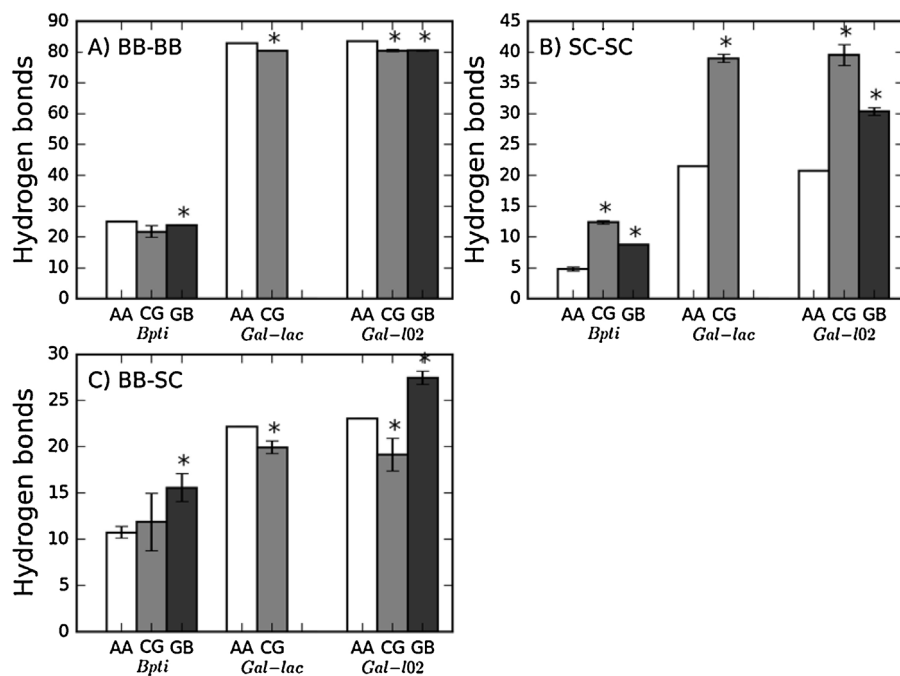


Fig. 5. Observed number of hydrogen between A) backbone atoms, B) side-chain atoms, C) backbone and side chain atoms. Errors bars indicate one standard deviation, and statistically significant deviations compared to AA (at 95% confidence) are marked with a star. The solvent model is indicated below the bars, and the proteins are highlighted in italics under each set of bars.

atoms, the differences between AA and the simpler models are even larger, and all of the differences are statistically significant. The CG model gives averages that are between 1.8 and 2.6 times larger than AA (or between 8 and 19 bonds), whereas the corresponding increase for GB is 1.5–1.8 (or between 4 and 10 bonds). Finally, if we look at hydrogen bonds between backbone and side-chain atoms, we again observe significant differences. However, CG seems to produce simulations with fewer hydrogen bonds than AA (between 1 and 4 bonds), whereas the opposite is true for GB (between 4 and 5 more bonds). Taking all together, the CG simulations gives between 5 and 13 more hydrogen bonds in total than AA, whereas the GB simulations give 8 more hydrogen bonds in total for both Bpti and Gal-102. Interestingly, an increased number of intra-protein hydrogen bonds should arguably lead to a more stable trajectory. However, as we have seen in the analysis of S^2 and RMSF, the proteins in both CG and GB tends to be more flexible than in AA. Thus, it seems that although the solvent model influences the hydrogen bond pattern, this change cannot be the only explanation for the observed difference in dynamics.

4. Conclusion

We analysed the protein dynamics in MD simulations of Bpti and Galectin-3 by computing order parameters that are accessible experimentally. The MD simulations were performed in water at three different resolutions, viz. all-atom (AA), coarse-grained (CG) and an implicit generalised Born (GB). The aim was to investigate how the choice of solvent model affects the protein dynamics, and in particular how well the simpler CG and GB models can reproduce NMR data. We found that for backbone dynamics, the agreement between NMR and simulations is good, with reasonable absolute deviations. Overall, all models are equally good in reproducing the NMR order parameters. It is interesting to note that different models sample dynamics on different time scale, as evidenced by the different optimal averaging window. The disagreement between simulations and NMR is worse if we look at side-chain dynamics, but encouragingly, all models give similar deviations. Furthermore,

we observed sizable differences between different solvent models when comparing estimates for a more extensive set of vectors than experimentally available. Still, the sample size is rather small and care should be taken to conclude too much from that analysis. Structural analyses show a strong indication that the side chains are more flexible in the CG and GB simulations than in the AA simulations, despite that there are an increased number of intra-protein hydrogen bonds with CG and GB compared to AA. Therefore, we can conclude that the solvent model has a considerable effect on the protein dynamics. Unfortunately, it also has to be concluded that the effect is very much system dependent, which rules out a simple fix to the observed discrepancies. Furthermore, it is for these test cases, not fruitful to make a detailed analysis of conformation subspaces, as the proteins are globular and fairly rigid. Although this study is limited in the number of proteins and solvent models tested, it is nevertheless clear that the solvent model has a crucial influence on protein dynamics and care has to be taken when choosing the model. It has to be mentioned though that the GB model used herein is only one of many, and it cannot be ruled out that other models are more appropriate. It has for instance been shown that different GB models produce significantly different backbone fluctuations [48]. There are also other CG models of water, but they are even more coarse-grained and problems with protein–solvent contacts have been described [27]. Further progress could potentially be made by comparing to other NMR observables, e.g. residual dipole couplings or scalar couplings, which has been done previously when benchmarking protein force fields [80]. However, it is worth to realize that AA is far from perfect in probing dynamical properties as shown recently by comparing different AA force fields [81]. Lastly, it is also worth emphasizing the protein force field was developed to be used predominantly with AA water and has been extensively validated in this context. Nothing guarantees that it is compatible with other solvent models and mixing different models should always be carried out with care. Therefore it might be necessary to adjust the protein force field to be useful with simpler water models. However, from this study the recommendation has to be, although the structural analyses show an increased flexibility

in the simpler models, that if the objective is to only compute NMR order parameters, any model will do.

Acknowledgments

SG acknowledges the Wenner-Gren Foundations for funding. C3SE at Chalmers University of Technology and iSolutions at the University of Southampton are acknowledged for computational resources. Professor David Goldberg is acknowledged for providing the raw N–H order parameters for Bpti. Ulf Ryde and Anna Reymer are acknowledged for fruitful discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.jmglm.2016.11.002>.

References

- [1] D.A. Case, *Acc. Chem. Res.* 35 (2002) 325–331.
- [2] T. Schlick, R. Collepardo-Guevara, L.A. Halvorsen, S. Jung, X. Xiao, *Q. Rev. Biophys.* 44 (2011) 191–228.
- [3] R.O. Dror, R.M. Dirks, J.P. Grossman, H. Xu, D.E. Shaw, *Annu. Rev. Biophys.* 41 (2012) 429–452.
- [4] R.G. Bryant, *Annu. Rev. Biophys. Biomol. Struct.* 25 (1996) 29–53.
- [5] N. Prabhu, K. Sharp, *Chem. Rev.* 106 (2006) 1616–1623.
- [6] M. Tarek, D.J. Tobias, *J. Biophys.* 79 (2000) 3244–3257.
- [7] D. Perez, B.P. Uberuaga, Y. Shim, J.G. Amar, A.F. Voter, *Annu. Rep. Comput. Chem.* 5 (2009) 79–98.
- [8] T. Huber, A.E. Torda, W.F. van Gunsteren, *J. Comput. Aided. Mol. Des.* 8 (1994) 695–708.
- [9] D. Hamelberg, J. Mongan, J.A. McCammon, *J. Chem. Phys.* 120 (2004) 11919–11929.
- [10] A. Barducci, G. Bussi, M. Parrinello, *Phys. Rev. Lett.* 100 (2008) 020603.
- [11] R. Swendsen, J. Wang, *Phys. Rev. Lett.* 57 (1986) 2607–2609.
- [12] C.J. Woods, J.W. Essex, M.A. King, *J. Phys. Chem. B* 107 (2003) 13703–13710.
- [13] A.M.A. West, R. Elber, D. Shalloway, *J. Chem. Phys.* 126 (2007) 145104.
- [14] J.D. Chodera, F. Noé, *Curr. Opin. Struct. Biol.* 25 (2014) 135–144.
- [15] P.J. Bond, J. Holyoake, A. Ivetac, S. Khalid, M.S.P. Sansom, *J. Struct. Biol.* 157 (2007) 593–605.
- [16] M.G. Saunders, G.A. Voth, *Annu. Rev. Biophys.* 42 (2013) 73–93.
- [17] D.H. de Jong, G. Singh, W.F.D. Bennett, C. Arnarez, T.A. Wassenaar, L.V. Schäfer, X. Periole, D.P. Tieleman, S.J. Marrink, *J. Chem. Theory Comput.* 9 (2013) 687–697.
- [18] P. Kar, S.M. Gopal, Y.-M. Cheng, A. Predeus, M. Feig, *J. Chem. Theory Comput.* 9 (2013) 3769–3788.
- [19] T.A. Knotts, N. Rathore, D.C. Schwartz, J.J. de Pablo, *J. Chem. Phys.* 126 (2007) 084901.
- [20] M.C. Linak, K.D. Dorfman, *J. Chem. Phys.* 133 (2010) 125101.
- [21] S.J. Marrink, A.H. de Vries, A.E. Mark, *J. Phys. Chem. B* 108 (2004) 750–760.
- [22] M. Orsi, D.Y. Haubertin, W.E. Sanderson, J.W. Essex, *J. Phys. Chem. B* 112 (2008) 802–815.
- [23] V.H. Rusu, R. Baron, R.D. Lins, *J. Chem. Theory Comput.* 10 (2014) 5068–5080.
- [24] S. Riniker, W.F. van Gunsteren, *J. Chem. Phys.* 134 (2011) 084110.
- [25] S.J. Marrink, H.J. Risselada, S. Yefimov, D.P. Tieleman, A.H. de Vries, *J. Phys. Chem. B* 111 (2007) 7812–7824.
- [26] M. Orsi, J.W. Essex, *PLoS One* 6 (2011) e28637.
- [27] S. Riniker, A.P. Eichenberger, W.F. van Gunsteren, *J. Eur. Biophys.* 41 (2012) 647–661.
- [28] T.A. Wassenaar, H.I. Ingólfsson, M. Priess, S.J. Marrink, L.V. Schäfer, *J. Phys. Chem. B* 117 (2013) 3516–3530.
- [29] M. Orsi, W. Ding, M. Palaiokostas, *J. Chem. Theory Comput.* 10 (2014) 4684–4693.
- [30] S. Genheden, *J. Chem. Theory Comput.* 12 (2016) 297–304.
- [31] J. Chen, C.L. Brooks, J. Khandogin, *Curr. Opin. Struct. Biol.* 18 (2008) 140–148.
- [32] N.A. Baker, *Curr. Opin. Struct. Biol.* 15 (2005) 137–143.
- [33] N.A. Baker, *Methods Enzymol.* 383 (2004) 94–118.
- [34] M.S. Lee, F.R. Salsbury, C.L. Brooks, *J. Chem. Phys.* 116 (2002) 10606.
- [35] W. Im, M.S. Lee, C.L. Brooks, *J. Comput. Chem.* 24 (2003) 1691–1702.
- [36] A. Onufriev, D. Bashford, D.A. Case, *Proteins* 55 (2004) 383–394.
- [37] U. Haberthür, A. Cafilisch, *J. Comput. Chem.* 29 (2008) 701–715.
- [38] E. Gallicchio, K. Paris, R.M. Levy, *J. Chem. Theory Comput.* 5 (2009) 2544–2564.
- [39] D.L. Mobley, K.A. Dill, J.D. Chodera, *J. Phys. Chem. B* 112 (2008) 938–946.
- [40] J.L. Knight, C.L. Brooks, *J. Comput. Chem.* 32 (2011) 2909–2923.
- [41] R. Zhou, *Proteins* 53 (2003) 148–161.
- [42] H. Nymeyer, A.E. García, *Proc. Natl. Acad. Sci. U. S. A.* 100 (2003) 13934–13939.
- [43] J. Srinivasan, T.E. Cheatham, P. Cieplak, P.A. Kollman, D.A. Case, *J. Am. Chem. Soc.* 120 (1998) 9401–9409.
- [44] D.-W. Li, R. Brüschweiler, *J. Am. Chem. Soc.* 131 (2009) 7226–7227.
- [45] R. Anandakrishnan, A. Drozdetski, R.C. Walker, A.V. Onufriev, *J. Biophys.* 108 (2015) 1153–1164.
- [46] X. Li, S.A. Hassan, E.L. Mehler, *Proteins* 60 (2005) 464–484.
- [47] C. Tan, L. Yang, R. Luo, *J. Phys. Chem. B* 110 (2006) 18680–18687.
- [48] D.P. Hua, H. Huang, A. Roy, C.B. Post, *Protein Sci.* 26 (2016) 204–218.
- [49] M. Akke, R. Brueschweiler, A.G. Palmer, *J. Am. Chem. Soc.* 115 (1993) 9832–9833.
- [50] T.I. Igumenova, K.K. Frederick, A.J. Wand, *Chem. Rev.* 106 (2006) 1672–1699.
- [51] Y. Gu, D.-W. Li, R. Brüschweiler, *J. Chem. Theory Comput.* 10 (2014) 2599–2607.
- [52] N. Trbovic, J.-H. Cho, R. Abel, R.A. Friesner, M. Rance, A.G. Palmer, *J. Am. Chem. Soc.* 131 (2009) 615–622.
- [53] C. Diehl, S. Genheden, K. Modig, U. Ryde, M. Akke, *J. Biomol. NMR* 45 (2009) 157–169.
- [54] O. Allnér, N. Foloppe, L. Nilsson, *J. Phys. Chem. B* 119 (2015) 1114–1128.
- [55] C. Diehl, O. Engström, T. Delaine, M. Håkansson, S. Genheden, K. Modig, H. Leffler, U. Ryde, U.J. Nilsson, M. Akke, *J. Am. Chem. Soc.* 132 (2010) 14577–14589.
- [56] T.A. Soares, X. Daura, C. Oostenbrink, L.J. Smith, W.F. van Gunsteren, *J. Biomol. NMR* 30 (2004) 407–422.
- [57] S.A. Showalter, R. Brüschweiler, *J. Chem. Theory Comput.* 3 (2007) 961–975.
- [58] S. Genheden, U. Ryde, *Phys. Chem. Chem. Phys.* 14 (2012) 8662–8677.
- [59] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, C. Simmerling, *Proteins* 65 (2006) 712–725.
- [60] K.N. Kirschner, A.B. Yongye, S.M. Tschampel, J. González-Outeiriño, C.R. Daniels, B.L. Foley, R.J. Woods, *J. Comput. Chem.* 29 (2008) 622–655.
- [61] J. Wang, R.M. Wolf, J.W. Caldwell, P.A. Kollman, D.A. Case, *J. Comput. Chem.* 25 (2004) 1157–1174.
- [62] H.W. Horn, W.C. Swope, J.W. Pitera, J.D. Madura, T.J. Dick, G.L. Hura, T. Head-Gordon, *J. Chem. Phys.* 120 (2004) 9665–9678.
- [63] S. Plimpton, *J. Comput. Phys.* 117 (1995) 1–19.
- [64] K.A. Feenstra, B. Hess, H.J.C. Berendsen, *J. Comput. Chem.* 20 (1999) 786–798.
- [65] S. Genheden, J.W. Essex, *J. Chem. Theory Comput.* 11 (2015) 4749–4759.
- [66] J.-P. Ryckaert, G. Ciccotti, H.J. Berendsen, *J. Comput. Phys.* 23 (1977) 327–341.
- [67] R.W. Hockney, J.W. Eastwood, *Computer Simulation Using Particles*, CRC Press, Boca Raton, FL, 1989, pp. 26–3047.
- [68] P.H. Hünenberger, *Adv. Polym. Sci.* 173 (2005) 105–147.
- [69] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, J.R. Haak, *J. Chem. Phys.* 81 (1984) 3684.
- [70] R. Salomon-Ferrer, D.A. Case, R.C. Walker, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* 3 (2013) 198–210.
- [71] J. Weiser, P.S. Shenkin, W.C. Still, *J. Comput. Chem.* 20 (1999) 217–230.
- [72] J.J. Prompers, R. Brüschweiler, *J. Am. Chem. Soc.* 124 (2002) 4522–4534.
- [73] S. Genheden, M. Akke, U. Ryde, *J. Chem. Theory Comput.* 10 (2014) 432–438.
- [74] D.L. Mobley, C.I. Bayly, M.D. Cooper, M.R. Shirts, K.A. Dill, *J. Chem. Theory Comput.* 5 (2009) 350–358.
- [75] J.-F. Truchon, C.I. Bayly, *J. Chem. Inf. Model.* 47 (2009) 488–508.
- [76] S.J. Swamidass, C.-A. Azencott, K. Daily, P. Baldi, *Bioinformatics* 26 (2010) 1346–1348.
- [77] N. Michaud-Agrawal, E.J. Denning, T.B. Woolf, O. Beckstein, *J. Comput. Chem.* 32 (2011) 2319–2327.
- [78] W.M. Hanson, S.A. Beeser, T.G. Oas, D.P. Goldenberg, *J. Mol. Biol.* 333 (2003) 425–441.
- [79] Y. Gu, D.-W. Li, R. Brüschweiler, *J. Chem. Theory Comput.* 10 (2014) 2599–2607.
- [80] K. Lindorff-Larsen, P. Maragakis, S. Piana, M.P. Eastwood, R.O. Dror, D.E. Shaw, *PLoS One* 7 (2012) e32131.
- [81] F. Vitalini, A.S.J.S. Mey, F. Noé, B.G. Keller, *J. Chem. Phys.* 142 (2015) 084101.