LUND UNIVERSITY

Will molecular dynamics simulations of proteins ever reach equilibrium?

Genheden, Samuel; Ryde, Ulf

Published in:
Physical chemistry chemical physics : PCCP

DOI:
10.1039/c2cp23961b

2012

Link to publication

Citation for published version (APA):
Genheden, S., & Ryde, U. (2012). Will molecular dynamics simulations of proteins ever reach equilibrium? *Physical chemistry chemical physics : PCCP*, *14*(24), 8662-8677. https://doi.org/10.1039/c2cp23961b

Total number of authors:
2

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Will molecular dynamics simulations of proteins ever reach equilibrium?

Samuel Genheden and Ulf Ryde*

Department of Theoretical Chemistry, Lund University, Chemical Centre,
P. O. Box 124, SE-221 00 Lund, Sweden

*Correspondence to Ulf Ryde, E-mail: Ulf.Ryde@teokem.lu.se,
Tel: +46 – 46 2224502, Fax: +46 – 46 2224543

2013-01-29

We show that conformational entropies calculated for five proteins and protein–ligand complexes with dihedral-distribution histogramming, the von Mises approach, or quasi-harmonic analysis do not converge to any useful precision even if molecular dynamics (MD) simulations of 380–500 ns length are employed (the uncertainty is 12–89 kJ/mol). To explain this, we suggest a simple protein model involving dihedrals with effective barriers forming a uniform distribution and show that for such a model, the entropy increases logarithmically with time until all significantly populated dihedral states have been sampled, in agreement with the simulations (during the simulations, 52–70% of the available dihedral phase space has been visited). This is also confirmed by the analysis of the trajectories of a 1-ms simulation of bovine pancreatic trypsin inhibitor (31 kJ/mol difference in the entropy between the first and second part of the simulation). Strictly speaking, this means that it is practically impossible to equilibrate MD simulations of proteins. We discuss the implications of such a lack of strict equilibration of protein MD simulations and show that ligand-binding free energies estimated with the MM/GBSA method (molecular mechanics with generalised Born and surface-area solvation) vary by 3–15 kJ/mol during a 500 ns simulation (the higher estimate is caused by rare conformational changes), although they involve a questionable but well-converged normal-mode entropy estimate, whereas free energies estimated by free-energy perturbation vary by less than 0.6 kJ/mol for the same simulation.

Keywords: Entropy, molecular dynamics simulations, convergence, equilibration, phase-space sampling, MM/GBSA, thermodynamic integration.

Introduction

Molecular dynamics (MD) simulations are today routinely used to study large biomolecular systems. They can provide an atomic-detail interpretation of biochemical process such as protein folding, enzymatic catalysis, and molecular recognition.[1,2,3] The quantity that drives these reactions is the free energy, ΔG, and computational methods to estimate it for various processes can be derived from the laws of statistical mechanics.[4,5]

However, it is also often of interest to decompose the free energy into enthalpic (ΔH) and entropic (ΔS) parts to obtain additional insight into the process, according to

$$\Delta G = \Delta H - T \Delta S \tag{1}$$

where T is the absolute temperature. Such a decomposition has been useful in drug design, for instance, where the effect of chemical modifications often is discussed in terms of entropy and enthalpy.[6] In particular, isothermal titration calorimetry measurements have popularised such a decomposition.[7] Many approaches have also been developed to estimate these quantities from molecular simulations.[8,9,10,11,12,13,14,15]

Nuclear magnetic resonance (NMR) is another popular experimental technique to study the structure and dynamics of biomolecules.[16,17] The dynamics of protein atoms are typically quantified by order parameters and with some assumptions on the angular fluctuations, the conformational entropy can be derived from these order parameters.[18,19] Unfortunately, such a description of the entropy is incomplete because the NMR experiments only probe a limited subset of the protein atoms, e.g. to the motion of the backbone and some side-chain atoms.[16,17,19,20] Therefore, MD simulations are often used to supplement the NMR measurements to provide the total conformational entropy of the protein.[16,21,22,23,24,25,26,27,28,29] Protein conformational entropies are the main topic of this study, but it should be remembered that there are other important contributions to the total entropy as well, e.g. from the solvent.

It is well-known that it is harder to obtain converged entropies than free energies from molecular simulations.[30,31,32,33] Many different methods to calculate entropies from MD simulations have been proposed[8,9,10,11,12,13,14,15] and they have been evaluated previously several times, but mainly for rather simple and small systems.[12,13,14,34,35] In this article, we perform a detailed analysis of five different protein or protein– ligand systems. We show that even with 380–500 ns simulation length, it is not possible to obtain converged conformational entropies of any usable precision. We present a simple model of a protein that is in agreement with experimental results and explains this poor convergence. An analysis of the recent 1-ms simulation of bovine pancreatic trypsin inhibitor[36] confirms the results. We discuss the implications of this observation for the equilibration of MD simulations in general and study how it affects ligand-binding free energies calculated with the MM/GBSA method and free-energy perturbations.

Methods

System preparation. Three proteins were studied: galectin-3 (Gal3), matrix metalloprotease 12 (MMP12), and bovine pancreatic trypsin inhibitor (BPTI). Two simulations were performed for Gal3, one with a lactose molecule (Lac) bound and the other with a synthetic lactose derivative (L02)[37] bound. Both ligands are shown in Figure 1. Likewise, two simulations were performed for MMP12 with different ligands, viz. two stereoisomers that will be denoted cn1h and cn2h (also shown in Figure 1). The preparation of Gal3 has been described previously.[38] The simulations of MMP12 were based on in-house crystal structures[39] and the simulations of BPTI were based on the 6PTI[40] crystal structure. For all proteins, Asp and Glu residues were assumed to be negatively charged, whereas Lys and Arg residues were assumed to be positively charged. This assignment was checked with the PROPKA tool.[41] The protonation of the His residues was decided from their hydrogen-bond network and solvent accessibility, as has been specified before for Gal3.[38] For

MMP12, His168, 183, 218, 222, and 228 were protonated on the ND1 atom, whereas His112, 172, 196, and 206 were protonated on the NE2 atom. BPTI does not contain any His residues.

The proteins were described by the Amber 99SB force field[42] and the ligands by the general Amber force field[43] with charges calculated by the RESP (restrained electrostatic potential) procedure[44] from electrostatic potentials calculated at the Hartree–Fock/6-31G* level and sampled according to the Merz–Kollman scheme.[45] The proteins were immersed in a box of TIP4P-Ewald[46] water molecules that extended at least 10 Å from the solute. The simulated systems contained 19674**,** ~21800**,** and ~27000 atoms for BPTI, Gal3, and MMP12, respectively.

The MMP12 protein contains three calcium ions and two zinc ions (cn1h and cn2h coordinate directly to one of the zinc ions). The charges of the metal ions and their coordinating residues were computed using RESP calculations at the same level as for the ligands. The zinc ions were modelled by a bonded potential, using parameters taken from quantum mechanical calculations of the optimum structures and frequencies at the B3LYP/6-31G** level, and extracted with the approach of Seminario[47] using the Hess2FF program.[48] For the calcium sites, which contain several water molecules, we instead had to use a restrained non-bonded model.[49] The van der Waals parameters were 1.85 Å and 0.25 kJ/mol for Zn (from OPLS), and 1.60 Å and 0.42 kJ/mol for Ca (from the Amber parm91.dat file). The parameters for the metal sites are included in the supplementary material.

MD simulations. All MD simulations were run using the sander module of Amber 10 or the pmemd module of Amber 11.[50] The temperature was kept constant at 300 K using a Langevin thermostat[51] with a collision frequency of 2.0 ps^{-1} and the pressure was kept constant at 1 atm using a weak-coupling isotropic algorithm with a relaxation time of 1 ps.[52] Particle-mesh Ewald summation with a fourth-order B spline interpolation and a tolerance of 10^{-5} was used to handle long-range electrostatics.[53] The cut-off for non-bonded interactions was set to 8 Å and the non-bonded pair list was updated every 50 fs. The SHAKE algorithm[54] was used to constrains bonds involving hydrogen atoms so that a 2 fs time step could be used.

Two types of simulations were performed. First, ten independent simulations were run by using different random starting velocities. For Gal3 and MMP12 systems, these were taken from previous studies.[38,39] In short, these systems were equilibrated for 200 (Gal3) or 500 ps (MMP12) in the NVT ensemble. BPTI was first minimized using 100 steps of steepest descent with restraints on all atoms except hydrogen atoms and water molecules, followed by 20 ps equilibration in the NPT ensemble and with the same restraints, and 100 ps equilibration in the NVT ensemble. The length of the production runs was 10, 20, or 40 ns for the MMP12, Gal3, and BPTI simulations, respectively, and they were performed in the NVT ensemble. Finally, one (Gal3) or two (BPTI) simulations were extended to 500 ns and one simulation was extended to 380 ns for the larger MMP12. Snapshots were saved every 10 ps if nothing else is stated. Each of the long simulations took ~40 days of CPU time on 64 cores with Intel Xeon 2.26 GHz processors.

Entropy estimates. The conformational entropy of the simulated proteins has been calculated with four different methods. In the first, which we will denote dihedral-distribution histogramming (DDH),[24,55] the protein Cartesian coordinates were converted to internal (bond, angle, and torsion) coordinates. Previous calculations have shown that entropy contributions from the bond and angle fluctuations are negligible.[27] The distribution for each dihedral angle, i, was then approximated by a discrete histogram with 72 bins (other numbers of bins have also been tested, but they gave similar results) and the entropy was calculated from

$$S_i = \frac{R}{2} - R\ln 72 - R\sum_{j=1}^{72} p_i(j)\ln p_i(j) \qquad (2)$$

where R is the gas constant and $p_i(j)$ is the probability that the dihedral angle is found in bin j (i.e. the dihedral angle is between $5(j-1)$ and $5j$ degrees). The first two terms are normalisation factors,

giving the entropy of a free rotor (R/2) for a uniform distribution. They cancel for relative entropies. Thus, each dihedral angle can contribute by an entropy between –3.8 R (–9.4 kJ/mol; a delta function) and 0.5 R (1.2 kJ/mol; a free rotor); a sine function gives 0.2 R (0.5 kJ/mol) independent of the periodicity, whereas a sine function with three minima, of which only one or two are populated gives –0.9 R or –0.2 R (–2.3 or –0.5 kJ/mol), respectively.

Second, we calculated the entropy by the von Mises approach (VMA)[25,56]. Instead of estimating the probabilities in Eqn. 2 from discrete histograms, this approach attempts to estimate the von Mises kernel density (which is the circular analogue of a Gaussian) of the dihedral angle, given by

$$p(\phi)=\frac{1}{N}\sum_{k=1}^{N}\exp(\kappa\cos(\phi-\phi_k))/(2\pi I_0(\kappa)) \tag{3}$$

where N is the total number of snapshots in the MD simulation, κ is the spread of the distribution (we used $\kappa^{-1/2}=1^{\circ}$ in all calculations, which has been shown to give stable results[25]), ϕ_k is the value of the dihedral angle in snapshot k, and I_0 is the modified Bessel function of zeroth order. Then, the entropy can be calculated from the continuous variant of Eqn. 2:

$$S_i=\frac{R}{2}-R\ln 2\pi-R\int_0^{2\pi}p(\phi)\ln p(\phi)\,\mathrm{d}\phi \tag{4}$$

Third, we calculated the entropy using quasi-harmonic analysis (QHA).[57,58] In this approach, one assumes that the total entropy is a multivariate Gaussian distribution. Quasi-harmonic frequencies, ω are calculated from the determinant

$$\det\left(\mathbf{M}^{1/2}\boldsymbol{\sigma}\,\mathbf{M}^{1/2}-\frac{kT}{\omega^2}\right)=0 \tag{5}$$

where $\mathbf{M}$ is the mass matrix (a diagonal matrix with the masses of the atoms on the diagonal) and $\boldsymbol{\sigma}$ is the covariance matrix, which has the elements $\sigma_{ij}=\langle(x_i-\langle x_i\rangle)(x_j-\langle x_j\rangle)\rangle$, where x_i and x_j are coordinates and the brackets indicate an average over the MD simulations. The frequencies are then used in the formula for the harmonic oscillator to estimate the entropy.[59] Thus, QHA does not assume that the various degrees of freedom are uncoupled, but it instead assumes that each normal mode resides in a single harmonic potential, which is a crude approximation for dihedral angles with several minima.[34]

Finally, we have also calculated the entropy using normal-mode analysis (NMA).[10] In this method, the structure is first minimised and then frequencies are calculated. The entropy is calculated from these frequencies, using in the formula for the harmonic oscillator. Because the NMA is computationally intensive, we employed a recent approach,[60] in which the protein is truncated using a 12 Å spherical cut-off from the bound ligand. Residues in the outermost 4 Å and all water molecules were kept fixed during the minimisation and were left out from the frequency calculation. This has been shown to be a stable and accurate approach.[61]

The influence of anharmonic effects and correlation between the various degrees of freedom for entropies calculated with different approaches has been much discussed. Most authors agree that anharmonic effects have a rather small influence on the calculated entropies.[32,35] On the other hand, some authors claim that correlation is insignificant if the entropies are calculated using internal coordinates,[26,62,63] especially for relative entropies (i.e. the correlation is constant e.g. during ligand binding),[64] whereas others report large second-order correlation effects even if the entropies are calculated by QHA.[32,35,65,66] In this investigation, we have ignored correlation effects because we concentrate on the convergence of the entropy – correlation will only further slow down the convergence of the calculated entropy.[35]

The DDH analysis was also performed on trajectories from the recent simulation of BPTI.[36] These trajectories were generously provided by D. E. Shaw research and contained 4 123 000 snapshots sampled every 0.25 ns during a 1.03 ms simulation. The files were converted to Amber format and were analysed the same way as the other simulations.

MM/GBSA calculations. Molecular mechanics with generalised Born and surface-area solvation (MM/GBSA) is an approximate method to estimate the free energy of a binding reaction.[10,67] In this approach, ΔG is estimated as the difference in the free energy between the complex (PL) and the protein (P) and the ligand (L), i.e. $\Delta G = G(\mathrm{PL}) - G(\mathrm{P}) - G(\mathrm{L})$, and each of these free energies are estimated from

$$G = E_{\mathrm{ele}} + E_{\mathrm{vdW}} + G_{\mathrm{pol}} + G_{\mathrm{np}} - TS \tag{6}$$

where the first two terms are the molecular mechanics (MM) electrostatic and van der Waals energy, G_{pol} and G_{np} are the polar and non-polar solvation energy, and the last term is an entropy estimate. Thus, MM/GBSA estimates the free energy by a combination of enthalpies (the two first terms), entropy (the last term), and free energies (the two solvation terms).

The enthalpies in Eqn. 6 were estimated without any cut-off on snapshots in which the water molecules have been stripped off. G_{pol} was estimated using the generalised Born model of Onufriev, Bashford and Case (model I, with $\alpha = 0.8$, $\beta = 0$, and $\gamma = 2.91$).[68] G_{np} was estimated through a relation to the solvent-accessible surface-area (SASA), viz. $\Delta G_{\mathrm{np}} = \gamma\mathrm{SASA} + b$, with $\gamma = 0.0227$ kJ/mol/Å^2 and $b = 3.85$ kJ/mol.[69] The entropy was taken from the NMA analysis, as described above. This is standard in the MM/GBSA approach.[10] The MM/GBSA analysis was based on only every tenth snapshot (sampled every 100 ps) because of the high computational demand of the NMA calculations.

Free-energy perturbation. The free energy of reducing the magnitude of either the charges or the van der Waals energy parameters of two ligands bound to galectin-3 was estimated using free-energy perturbation (FEP).[70] As will be more discussed below, these are typical steps in a calculation of the binding affinity of the ligands.[71] The system was simulated with full charges and van der Waals parameters on the ligands, and the free energy was then estimated using

$$\Delta G = -RT \langle \exp(-\Delta E / RT) \rangle \tag{7}$$

where ΔE is the difference in total potential energy between a state for which the magnitude of the charges (or van der Waals energy parameters) has been reduced by 10% and a state for which the full charges (or van der Waals parameters) were used, and the brackets indicate an average over the simulation with full charges and van der Waals parameters.

Results and Discussion

Protein conformational entropies. This study was started with the aim of estimating the change in protein conformational entropies upon ligand binding for Gal3 and MMP12 that should complement NMR relaxation data (for simplicity, we will simply say entropies in the following, although we recognize that there are other important contributions to the total entropy of the system, e.g. from the solvent).[39] Such studies have been performed several times before[16,21,22,23,24,25,26,27,28,29] and we followed the protocol developed in those studies: We performed 500 ns simulations of two ligand-bound states of the Gal3 protein and 380 ns simulations of two ligand-bound states of MMP12. The protein conformational entropies were then estimated by the DDH approach. The time evolution of these entropy estimates are shown in Figures 2 and 3, and the results are summarised in Table 1.

Starting with MMP12, it can be seen that the DDH entropy changes by 42 and 89 kJ/mol during the last 100 ns of the simulations with cn1h and cn2h, respectively (throughout the article, we will discuss entropies in energy units, i.e. $T\Delta S$ in kJ/mol at 300 K). For cn1h, it seems that the system has reached a quasi-stable state after 300 ns but it is not clear if the system will stay in this state. The cn2h system reached a quasi-stable state that lasted from 100 to 250 ns, but in the last part of the simulation, the entropy estimate is drifting. Because our initial aim was to study the entropy change upon ligand binding, we present also the entropy difference between cn1h and cn2h simulations in Figure 2 and Table 1. It can be seen that neither the absolute conformational entropies, nor their difference are stable after 380 ns simulations. For example, the entropy difference changes by 47 kJ/mol during the last 100 ns of the simulation and it changes sign after 225 ns simulation.

The results for Gal3 are similar as can be seen in Figure 3. Gal3-Lac attained a quasi-stable state between 150 and 300 ns, but then the entropy started to increase, leading to a change in the entropy of 38 kJ/mol during the last 100 ns of the simulation. On the other hand, Gal3-L02 seems to have reached a quasi-stable state after 300 ns, but the entropy still changes by 22 kJ/mol during the last 100 ns of the simulation. Moreover, in contrast to the other simulations, the entropy decreased during the last 100 ns, so that change in the entropy difference between the two ligands during the last 100 ns of the simulation is actually larger than the changes for the two ligands, 59 kJ/mol. Also for this protein, the difference in entropy between the two ligands changes sign during the simulation.

Previous studies have shown that it is more effective to run several shorter simulation than a single long one and that the uncertainty estimated from a long simulation typically is an underestimate because it stays close to the starting point in the phase space.[38,72,73,74,75,76,77] Therefore, we also calculated the entropy from ten independent simulations of 10 (MMP12) or 20 ns (Gal3) length. The average entropies over these simulations are shown in Table 1 together with the corresponding standard errors (i.e. the standard deviations among the ten simulations divided by $\sqrt{10}$). It can be noted that the precision of the estimates is rather poor, 18–38 kJ/mol. Even worse, there is a large difference in the entropy estimate from the ten short and the single long simulation of 423 to 927 kJ/mol, the short simulations always giving a more negative result. This is much larger than what is expected from the standard error of the short simulations or the convergence of the long simulation. On the other hand, the results in Figures 2 and 3 show that the entropies increase by several hundreds of kJ/mol during the first 100 ns of the simulation. Thus, it is clear that averaging over simulations shorter than 100 ns is not a solution to the sampling problem.

The DDH analysis is based on sampling of a distribution and does not have an inherent notion of time. Therefore, an alternative to averaging over independent simulations is to concatenate the ten short trajectories and estimate the entropy from this concatenated trajectory of 100–200 ns total length. This estimate, which is also shown in Table 1, is closer to that of the long simulation, but it is still 141–482 kJ/mol more negative. This shows that several short simulations is an effective way to sample the phase space, but for entropy estimates, 10–20 ns simulations are much too short to give reliable results.

For Gal3, we also investigated whether the convergence of the entropy could be improved by considering only a subset of the dihedral angles. It is natural to assume that only residues close to the ligand contribute significantly to the entropy change during ligand binding (such an assumption is normally used for the NMA analysis in MM/PBSA[10,67] and it has been shown to be accurate at least for relative entropies[61]). Therefore, we calculated the DDH entropy only for the residues that are within 8 Å of L02 in the crystal structure. The results in Figure S1 in the supplementary material show that such a procedure reduces the uncertainty of the entropy by a factor of ~2, but not at an acceptable level: The entropy changes by 11, 12, and 23 kJ/mol the last 100 ns for Gal3-Lac, Gal3-L02, and the difference, respectively.

We also tried to include in the DDH analysis only dihedral angles that gave converged results, as has been suggested before:[64] For each dihedral angle we calculated the difference in entropy at 400 ns and 500 ns, and included it in the total entropy if the difference was below a certain

threshold. Unfortunately, it turned out that the uncertainty in the entropy is caused by small differences in a great number of dihedrals: To obtain results that are converged to within 4 kJ/mol, 40% of the dihedrals have to be excluded, using a convergence threshold of 0.036 kJ/mol for the individual dihedrals. Excluding only the 25 dihedrals with the poorest convergence (with a threshold of 0.5 kJ/mol) actually deteriorated the convergence for Gal3-L02.

Two other popular methods to estimate conformational entropies are QHA and VMA.[56,58] We have tested also these method to see if they give results that are better converged. However, this turned out not to be the case, as can be seen from the results for Gal3 in Figures S2–S3 in the supplementary material. In fact, the VMA and DDH approaches give very similar results, whereas the QHA results are somewhat more different, but show similar convergence problems, as has also been noted before in a different context.[34,78] Correlation effects were ignored in all these calculations. The significance of such effect have been much discussed with varying conclusions,[26,32,35,62,63,64,65,66] but it is clear that they will slow down, rather than improve the convergence,[35] which is the only thing that matters in this study.

These results clearly show that it is impossible to estimate the protein conformation entropy of MMP12 and Gal3 with a reasonable precision from either DDH, VMA, or QHA analyses based on simulations of 380–500 ns length. This is quite surprisingly, considering that several studies have used this or similar methods to estimate entropies for other proteins.[17,21,22,23,24,25,26,28] Therefore, we run simulations also of one of the previously studies proteins, BPTI.[26] The results in Figure 4 show that this protein gives a somewhat smaller drift in the entropy than the other two proteins, but the DDH entropy still changes by 12 kJ/mol during the last 100 ns of the simulation and there is no indication that the results are stable after 500 ns. In order to test the reproducibility of this calculation, we run a second independent 500 ns MD simulation of the same protein, starting with the exactly the same structure but with different velocities. As can be seen in Figure 4, this simulation gave a similar time course, but the entropies differ by 11–97 kJ/mol (14 kJ/mol at the end of the simulations). This shows that even for this small and extremely stable protein (58 residues and three Cys–Cys linkages), the entropy cannot be determined from a 500 ns simulation with a precision better than 14 kJ/mol (and the entropy is still increasing at the end of the simulations).

A simple protein model explains the convergence problems. To gain understanding of this poor convergence of the conformational entropies from long MD simulations, we suggest in this section a simplified model of a protein. In the MM description, a protein is described by a force field consisting of terms for bonds, angles, dihedrals, as well as Lennard-Jones and electrostatic non-bonded interactions. The energy of each dihedral angle is determined by terms of the type

$$V_i(\phi_i)=\frac{k_i}{2}(1+\sin(n_i\phi_i+\delta_i)) \tag{8}$$

where k_i is a force constant, n_i is the periodicity (typically 1, 2, or 3), and δ_i is a phase shift (0 or 180°) of the dihedral. Thus, the dihedrals are described by a sine function, where n_i determines the number of minima, δ_i determines the location of the minima, and k_i determines the barrier between the minima. However, in the protein, there is a strong coupling (correlation) between the various dihedral angles and other degrees of freedom, e.g. owing to steric effects. In NMA and QHA, this is described by diagonalising the Hessian or covariance matrices, thereby obtaining the normal modes of the protein, which are uncoupled to the first order.

We suggest a model protein consisting of a number, N, of such uncoupled normal modes. For simplicity, we assume that each normal mode has a dihedral-like potential, i.e. that there are two minima (conformations) along the normal mode, separated by a barrier of A_i. To start with, we assume that the two minima have the same energy (we will remove that assumption below). Other potential energy terms are ignored. We assume that the barriers A_i form a uniform (random) distribution with values ranging between 0 and a high number, e.g. A_{max} = 100 kJ/mol. A_{max} is higher

than the highest barriers of dihedrals (21 kJ/mol in Amber force field, disregarding dihedrals in aromatic groups), owing to the coupling of the dihedrals. Thus, after sorting the normal modes, we assume that $A_i = i/N\, A_{max}$. Moreover, we assume that the characteristic time for the barrier passing of each dihedral (τ_i; i.e. the inverse of the rate by which the barrier is passed) can be calculated from the Arrhenius equation, i.e.

$$\tau_i = \tau_0 \exp\left(\frac{A_i}{RT}\right) \tag{9}$$

where τ_0 is a constant of the magnitude of $h/k_B T$ = 0.2 ps (Planck's constant divided by Boltzmann's constant and the absolute temperature), and R is the gas constant. Eqn. 9 indicates that τ_i is exponentially distributed from short to very long times (A_i = 1 kJ/mol corresponds to ~0.2 ps, whereas 100 kJ/mol corresponds to ~12 h). It also means that if we simulate the system for t_{sim}, on average all normal modes with $\tau_i > t_{sim}$ will reside in a single conformation (the starting conformation), whereas all normal modes with $\tau_i < t_{sim}$ will visit both conformations with equal probability. The ratio of the normal modes that visit both conformations increases linearly with the logarithm of t_{sim}, as is shown in Figure 5 (diamonds).

Consequently, we can calculate the entropy of each normal mode from

$$S_i = \begin{cases} 0 & \text{if } \tau_i \le t_{sim} \\ R\ln 2 & \text{if } \tau_i > t_{sim} \end{cases} \tag{10}$$

and the total conformational entropy of the entire protein is simply the sum of these normal-mode entropies. With these assumptions, the entropy of the model protein will increase logarithmically with the simulation time, i.e. grow linearly if t_{sim} is plotted on a logarithmic scale, as is show in Figure 5 (squares). This shows that for such a model protein, the conformational entropy will not converge until $t_{sim} > \tau_i$ for all normal modes, i.e. when t_{sim} > 12 h if A_{max} = 100 kJ/mol. In fact, the entropy will increase by 1.7 kJ/mol every time the barrier of a new normal mode is passed, as can be seen in Figure 5.

We can now go back to our MD simulations and see how realistic such a protein model is. We do this by plotting the ratio of the dihedral angles that visit all their possible minima (i.e. n_i in Eqn. 8) during a certain simulation time. This is similar the ratio that was plotted for the model protein in Figure 5, giving a straight line when plotted against log(t_{sim}). Figure 6 shows that this is approximately also the case for all our protein simulations. At the end of the simulations, 52–69% of the dihedrals in the proteins have visited all their available conformations (excluding dihedrals in aromatic rings from the analysis) and the amount is still increasing with logarithmic time.

Likewise, Figure 7 shows that approximately linear plots are obtained if the entropies for each protein are plotted against logarithmic time, at least for MMP12 and Gal3. This is in accordance with the linear relation predicted for the model protein, as shown in Figure 5. Thus, our simple model of a protein actually seems to cover the relevant physics (the simulation time of 500 ns corresponds to a barrier of 37 kJ/mol). The model indicates that if we increase the simulation time, we will continue to open up new conformations (volumes in the phase space) that have not been visited before and the entropy will continue to increase with simulation time. In fact, according to the model, the entropy will not to converge until all possible conformations have been visited (i.e. when the ratio on the right y axis in Figure 5 reaches 1). In practice, this would mean that we need to simulate until the protein unfolds during the simulation (the folded and unfolded conformations of a protein are normally in equilibrium, so a folded protein will visit also unfolded conformations if the simulations are long enough), because then most dihedrals will become uncoupled and therefore be determined by their dihedral barriers k_i in Eqn. 8, which in the Amber force field are less than 21 kJ/mol (this was confirmed by performing a MD simulation of denatured BPTI, starting from the crystal structure, but sampling at 1000 K; already after 10 ns, it showed 95% sampling of the dihedral angels). Of course, such long simulations can currently not be performed except for

very small, fast-folding polypeptides, indicating that it is currently practically impossible to converge entropy calculations for normal-sized protein. This is in accordance with the results in Figures 2–4, 7, and Table 1.

The reader may object that Figures 2–4 indicate that the entropy actually is converging, so if somewhat longer simulations were used, converged entropies would actually be obtained. However, this is mainly an artefact of plotting a logarithmic function on a linear scale. A logarithmic function will always give the impression of converging if plotted on a linear scale, as is shown in Figure 8, especially if supplemented by noise. This is probably the reason why, scientists always have claimed that their simulations are equilibrated with an equilibration time of 1/10–1/2 of the total simulation time, independently of the actual simulation time. However, plotted on a logarithmic scale, Figures 5–7 show that there hardly is any convergence. Thus, convergence of entropies should always be judged on a logarithmic time scale.

The curves in Figure 6 show some variation between the three proteins. All five curves start from a similar point of 16% at 10 ps. However, between 10 and 100 ps, the two curves of Gal3 show a more rapid increase (to 37%) than the curves of the other two proteins (26%). After that, the five curves run more or less in parallel. This indicates that the rate of the sampling of conformational space is an intrinsic property of each protein, but the rate is quite similar outside the picosecond range. The lower sampling of BPTI is probably related to the presence of three cystine links, which restricts the available conformational space.

Figure 6 can be used to estimate how long simulations are needed to sample all dihedral angles: This is simply the time when the curves in Figure 6 reach 100% sampling. Assuming that all curves are linear for $t_{\mathrm{sim}} > 100$ ps, this will take between 1 ms and 100 hours, corresponding to maximum barriers of 56–105 kJ/mol. On the other hand, as our test simulation of an unfolded protein showed, once the protein unfolds, nearly all available conformations are rapidly visited. The folding of proteins typically occurs on the time scale of milliseconds to hours, so this is probably also the time scale needed to reach full sampling of the conformational space of the dihedrals in proteins.

Some of the curves in Figure 6 indicate that the rate of conformation sampling decreases somewhat at the end of the 380–500 ns simulations, in particular those for BPTI. It is not clear whether this is a random variation or if this will become pronounced if the simulation is extended. However, this only means that it will take a longer time to reach complete sampling of all conformational states.

The convergence of the conformational entropy estimates from MD simulations, obtained with similar methods as in this paper, has been studied before. In most cases, the conformational entropy shows a similar logarithmic time-dependence as in Figures 2–4 and a questionable convergence.[15,25,22,32,35,79,80,81,82,83] For example, a 1.1 µs simulation of a 15-peptide was not enough to reach convergence of the entropy.[35] As longer and longer MD simulations are published, it also becomes more and more apparent that proteins show extensive dynamics at the ns and µs time scales.[61,84] In a few cases, the results indicate that the entropies are converged,[85] although it cannot be excluded that longer simulations would lead to a change in the entropy again, as was observed in our Gal3-Lac simulation. For example, for villin head piece (35 residues), the entropy from several trajectories showed convergence within 4–5 kJ/mol after 100 ns simulation in a generalised Born solvent, although two trajectories gave entropies that increased strongly after 70 ns, which was interpreted as a drift away from the near-native basin.[25] For the bigger ubiquitin protein (76 residues), the entropies after 10 ns simulation showed a significant drift, similar to what we have observed in our simulations and in our protein model, although this was again interpreted as the formation of non-native states owing to a poor force field and the continuum-solvation model.[25] In our simulations, there is no indication that the simulation visits non-native states, as can be seen from an analysis of the root-mean square deviation relative to the starting structure, shown in Figure S4.

A prominent difference between the entropy of our simple protein model in Figure 5 and the protein simulations in Figures 2–4 and 7 is that the entropy always increases with time in our

model, whereas the entropy often goes down in the real simulations. The reason for this is the simplified entropy expression in Eqn. 10, which dictates that the entropy increases by $R \ln 2$ = 1.7 kJ/mol every time a barrier is passed. Eqn. 2 provides a more accurate expression for the entropy and it shows that the entropy depends on the probability, $p_i(j)$, of observing a certain minimum, j. In our simple model, we assumed that once the barrier is passed, the two minima have equal probability ($p_i(j)$ = 0.5 for both j = 1 and 2). In a real protein, the various minima have different probabilities, and the estimated probabilities may vary during the simulation until they converge when all minima have been visited many times. For example, if an unusual conformation is found by chance early in the simulation, the entropy will increase, because both conformations will have a high probability (assuming only two possible states). However, if the unusual conformation is not visited again during the rest of the simulation, its probability will successively decrease, and therefore, the entropy will also decrease for that mode, according to Eqn. 2.

We can extend our model by introducing varying probabilities for the conformations of each normal mode. In order to keep the number of parameters to a minimum, we still assume only two states for each mode, but calculate the entropy from

$$S_i = R\left(p_i(1)\ln p_i(1) + p_i(2)\ln p_i(2)\right) \tag{11}$$

where $p_i(1) = 1$ and $p_i(2) = 0$ if $\tau_i < t_{sim}$, but $0 < p_i(1), p_i(2) < 1$ otherwise. Moreover, we assume that the free-energy difference between the two states of each normal mode, ΔG_i, increases linearly with the height of the barrier between them, i.e. $\Delta G_i, = i/N\ \Delta G_{max}$, where ΔG_{max} is the maximum free-energy difference. The probability is then given by

$$p_i(2) = \frac{1}{1 + e^{\frac{\Delta G_i}{RT}}} \tag{12}$$

In Figure 9, the entropy for such a system is shown as a function of t_{sim}, for four different values of ΔG_{max} = 10–40 kJ/mol. It can be seen that for small values of t_{sim}, the entropy still increases linearly with logarithmic time. However, as t_{sim} increases, the entropy curve finally levels of, converging to a certain value. The rate of convergence strongly depends on ΔG_{max}: A convergence to within 4 kJ/mol is obtained at 1 ms for ΔG_{max} = 40 kJ/mol, but not within 10^5 s (28 h) for ΔG_{max} = 10 kJ/mol. Unfortunately, it is hard to settle a proper value of ΔG_{max}. The folded state of a protein is typically 10–60 kJ/mol more stable than the unfolded state.[86] However, it is not clear if this free-energy difference should be related to ΔG_{max} directly or to the sum of ΔG_{max} for all dihedrals (which vary from 7505 to 30020 kJ/mol for the four curves in Figure 9. The most realistic answer is probably in between.

From this extended protein model, we learn several things: The dihedral entropy in protein simulations can converge for two different reasons. One is that we have visited all possible conformations of all dihedrals. This would happen if A_{max} is low, e.g. for a small peptide, for which there are few dihedrals that are not much coupled. Such a convergence can be detected by counting the number of minima visited during the simulations, as in Figure 6. Unfortunately, Eqn. 2 and its simplified variant in Eqn. 11 show that it is not enough to visit all minima once. Instead, they must be visited so many times that accurate estimates of the probabilities of all minima are obtained.

The second possibility is that the remaining (not visited) conformational minima are so unlikely (have so low probabilities, i.e. high ΔG_i) that they no longer contribute to the total entropy of the protein. This is the type of convergence that we might start to see for BPTI in Figures 6 and 7c. However, it is impossible to prove that convergence has really been reached or whether new volumes in phase space may be found if the simulations are extended (so more barriers may be passed) or started with other conditions, without doing longer or additional simulations.[87,88]

Fortunately, we can actually check this for BPTI: Shaw and co-workers have simulated the native state of BPTI for 1 ms, using specialised hardware.[36] We have analysed the trajectories from

this simulation by calculating the DDH entropy using Eqn. 2. Figure 10 shows that the entropy changes by 15 kJ/mol when going from 500 to 750 ns in agreement with our results in Figure 4 and Table 1. Then, the change decreases and between 18 and 34 µs, the entropy varies by only 4 kJ/mol, giving the impression of convergence. However, at 34 µs, the Cys14–Cys38 disulphide bridge changes its conformation, which opens new volumes in the phase space and the entropy increases by 190 kJ/mol during the next 46 µs. After 80 µs Cys14–Cys38 changes back to its old conformation and the entropy decreases again for over 100 µs. The rest of the simulation, the entropy increases again but more slowly, with plateaus mixed with increases and decreases. However, the entropy still changes by 23 kJ/mol between 0.8 and 0.9 ms. The difference in the entropy calculated for the first and second 0.5 ms part of the simulation is 31 kJ/mol. When plotted on a logarithmic time scale (Figure 10b), the entropy increases approximately linearly throughout the entire simulation, with exception for the large jump in the entropy around 34 µs. In particular, there is no indication that the rate decreases at the end of the simulation. Likewise, the sampling of the dihedral phase space increases linearly through the simulation, with 70% of the dihedral sampling all available conformations after 1 ms (Figure 10b). There is no indication of a saturation of the sampling of the dihedral conformations. This analysis strongly supports our conclusions from the 500 ns simulation that dihedral entropies are practically impossible to converge, even for such a small and stable protein as BPTI. In particular, they show that DDH entropies obtained after 500 ns differ by over 250 kJ/mol from those obtained after 1 ms and that possible convergences observed during the simulations are only apparent.

Finally, we note that our simple protein models are quite insensitive to the parameters. In particular, the number of dihedrals, N, will not change the results at all (Figures 5 and 9 were obtained with N = 100 and 1500, respectively). Likewise, the results are completely insensitive to $A_{\max}$ as long as the considered maximum t_{sim} is shorter than the corresponding maximum τ_i. The assumption of a uniform distribution is equivalent to a rectangular random distribution, which is quite reasonable. A normal distribution would probably be more realistic, but then the results would depend on the selected mean and variance of the distribution. The most problematic assumption is ΔG_{i} in the modified model and that it should be proportional to A_i. This leads to a strong dependence on the unknown $\Delta G_{\max}$, as is shown in Figure 9. In fact, the assumption that all $\Delta G_{\mathrm{i}} > 0$ does not agree with the dihedral potential function in Eqn. 8, which shows that in the unfolded state (i.e. when the dihedrals are uncoupled), all ΔG_i should vanish. Still, it allowed us to illustrate the fact that we do not need to sample all points in phase space and that the entropy depends also on the probability for each conformation. It should also be noted that to reach convergence in the entropies, as in Figure 9, increasing free-energy differences (ΔG_i) are needed. A constant difference for all dihedrals (all $\Delta G_i = c$) give the same linear increase in entropy as the original model (i.e. as shown in Figure 5).

Can MD simulations be equilibrated? We have seen that conformational entropies of three typical proteins estimated from MD simulations with the DDH, VMA, and QHA approaches do not converge for simulation times of 380–500 ns and not even for a 1-ms simulation of BPTI. How do such conclusions affect other properties studied by MD simulations?

Strictly speaking, this observation is catastrophic. If we use the normal definition of equilibration of MD simulations, viz. a state when all properties become independent of the simulation time,[89] this means that essentially no protein MD simulation will strictly be equilibrated (entropy is also a property of the system). On the other hand, 35 years experience of protein simulations has shown that quite useful information can be obtained from MD simulations. Therefore, it seems evident that different properties show different sensitivities to the amount of sampling.[87,88] We have shown here that the DDH, VMA, and QHA entropies, both absolute and relative, are extremely sensitive to the equilibration and hard to converge. Below, we will study the convergence of two other types of properties that can be obtained from the same simulations, viz. absolute free energies obtained from the MM/GBSA approach (involving NMA entropies) and free-energy differences obtained from FEP.

MM/GBSA energies. We have calculated the MM/GBSA binding free energies of Lac and L02 to Gal3 using our 500 ns MD simulations. The convergence of those results is shown in Figures 11 and 12, for Lac and L02 respectively. From Figure 11, it can be seen that the net binding free energy of Lac (ΔG trace) converges reasonably, showing a variation of less than 3 kJ/mol throughout the simulation time. On the other hand, 3 kJ/mol is quite much for binding affinities, corresponding to a difference by a factor of 3 in the binding constant. In fact, 300 ns simulation time is needed before the binding free energy is converged to within 1 kJ/mol (a factor of 1.5 in the binding constant) and there is no guarantee that the binding energy will not change if the simulation is extended.

As can be seen from Eqn. 6, the MM/GBSA free energies are obtained from five separate terms, which are also shown in Figure 11. It can be seen that the electrostatic and solvation energy terms show much larger variations than the net free energy, up to 15 kJ/mol. However, these two energies cancel to a large extent, so that their sum show a variation of up to only 5 kJ/mol. The van der Waals energy shows a even smaller variation, up to 2 kJ/mol, but it also partly cancels the variation in the $E_{el} + G_{pol}$ terms. The non-polar solvation energy term shows essentially no variation.

MM/GBSA also includes an entropy term, obtained from an NMA analysis. Interestingly, we see no convergence problem for this term, contrary to the entropies obtained by the DDH, VMA, and QHA methods: The NMA entropy varies by up to 3 kJ/mol during the first 20 ns of the simulation, but after 50 ns, it is converged to within 1 kJ/mol (but 20 and 50 ns is much longer than normal simulation times of MM/GBSA). This shows that the NMA analysis produces much more stable estimates of the entropy than the other three methods. The reason for this is that the NMA analysis ignores the entropy arising from the fact that some groups can attain several different conformations. Instead, it assumes that the protein attains a single conformation (obtained by minimisation) and estimates the entropy from the vibrational spectrum in this conformation, i.e. the width of the potential-energy wells. Apparently, the widths of the wells do not change significantly if different conformations are used to calculate this NMA entropy. Thus, NMA avoids the convergence problem of the conformational entropies by simply ignoring this entropy term, an effective, but not necessarily accurate, approximation. This conformational entropy, is the dominant term for the DDH and WMA approaches, and the only term in our simple protein model (Eqn. 10), whereas the QHA method includes both contributions (conformational entropy and the widths of the wells), although the conformational term comes in as a widening of the harmonic well when a new conformation is sampled, a quite coarse approximation, as has been discussed before.[78] It is likely that replacement of NMA entropies with DDH, VMA, or QHA entropies would improve the accuracy of the MM/GBSA method, but on the other hand it would then also inherit the convergence problems of these methods.

The Gal3-L02 energies in Figure 12 show similar qualitative trends but the variation is much larger, up to 26 kJ/mol in the individual energy terms (the absolute electrostatic energy changes by almost 50 kJ/mol) and 13 kJ/mol in the net binding free energy. The reason for this will be explained below. However, the NMA entropy is still stable with a variation of up to 2 kJ/mol.

We have previously argued that it is more efficient to base MM/GBSA energies on averages from several short simulations than from a single long one.[38] Therefore, we also calculated MM/GBSA estimates on the ten independent Gal3 simulations of 20 ns length. The average binding free energy estimates for Gal3-Lac and L02 are –27±1 and –67±2 kJ/mol, respectively. These estimates agree within statistical uncertainty with the estimates based on the 500 ns simulations, –27±0.4 and –70±0.4 kJ/mol, respectively.

However, it is also interesting to see if the two methods estimate a similar distribution of MM/GBSA binding free energies. In Figures 13 and 14, we have plotted the MM/GBSA estimates for each of the individual snapshots, together with a box plot that summarises the distribution. It can be seen that for both Gal3-Lac and Gal3-L02, the distribution for the long simulation is similar to the collective distributions of the short simulations. However, it is obvious that the long simulations have more outliers than the short simulations, most likely due to rare events.

It can also be seen from the scatter plots in Figure 14 that that the Gal3-L02 simulation shows at least two different quasi-stable conformations, one dominant, giving a binding free energy around

–70 kJ/mol, and a less stable one giving a much lower affinity of about –15 kJ/mol. The latter is observed after 175–200 ns simulation time, but also for a short while after ~460 ns. It is the former period that gives rise to the cusp in the energy curves in Figure 12, which then is slowly averaged away during the rest of the simulation. However, the sampling of this conformation (which is barely seen in the short simulations) is too poor to know whether the sampling of it in the long or the short simulations is most typical. Still, it illustrates that the MM/GBSA method is sensitive to the sampling of rare events that may quite strongly affect the calculated energies if the conformational changes are close enough to the active site. For Gal3-L02, it affects the binding affinity by ~12 kJ/mol or over two orders of magnitude in the binding constant. However, the total binding free energy differs from the final one by 8–10 kJ/mol already before the conformational change is seen (i.e. after 10–175 ns simulation).

FEP energies. Finally, we also studied energies calculated with free-energy perturbation (FEP), which is considered to be one of the most accurate method to estimate free energies. This method can also be used to calculate the binding free energy of a ligand to a protein, but it would typically require simulations of ~10 intermediate states and separate simulations for electrostatic and non-electrostatic interactions. Of course, it would be prohibitively expensive to perform each of these calculations for 500 ns. However, the first step in a FEP study of ligand binding would be to simulate the native complex and calculate the free-energy cost of reducing the ligand charges. Therefore, we have used the available 500 ns simulations and performed this first step in the FEP calculation, using Eqn. 7 and reducing the charges by 10%.

The convergence of this FEP energy result is plotted in Figure 15 for Gal3-L02 and Gal3-Lac. It can be seen that the FEP energy is extremely stable: The free energy varies by less than 0.1 kJ/mol for Lac and 0.17 kJ/mol for L02. The latter curve shows a kink at 180 ns, i.e. at the same time as the cusps are observed in Figure 12 and the differing binding energies are observed in Figure 14b, but the energetic variation in Figure 15 is ~200 times smaller for FEP than for MM/GBSA and there is no indication that the FEP free energies show any long-term trends.

In our experience,[90] the precision of the electrostatic term in FEP is typically of a similar size for the various intermediate perturbations. Assuming that the standard errors are statistical, this means that the net uncertainty in the total electrostatic FEP free energy is approximately $\sqrt{10} = 3.2$ times the uncertainty in Figure 15, i.e. up to 0.5 kJ/mol.

A FEP calculation of the binding affinity would also involve a perturbation of the van der Waals parameters of the ligand. In a practical application, this perturbation is always performed after the charges have been removed, to avoid that charges come too close together when the van der Waals parameters are vanishing. However, this is no problem in the early van der Waals steps and from a thermodynamic point of view, the order of the perturbations should not matter. Therefore, we can get an estimate of the uncertainty in also in the van der Waals step by performing a perturbation decreasing the van der Waals parameters of the ligand by 10%, based on the available 500 ns simulation. The results in Figure 16 show that the variation is even smaller, less than 0.04 kJ/mol for Lac and up to 0.09 kJ/mol for L02.

Using the same assumption of equal contributions for each intermediate step (which is less accurate for the van der Waals step), we get an estimated variation in all van der Waals perturbations of less than 0.3 kJ/mol ($\sqrt{10} * 0.9$), and a total variation in the absolute FEP affinity predicted of ~0.6 kJ/mol ($\sqrt{(10 * 0.17^2 + 10 * 0.09^2)}$), i.e. 20 times smaller than for MM/GBSA. It should be recognized that these estimates are highly approximate, because we only extrapolate the results from the first step in the charge and van der Waals perturbations. On the other hand, it is more common to study only local variations of the ligand (perturbations of one group to another, giving relative binding affinities) with FEP, which would reduce the FEP energies and uncertainties.

Conclusions

This investigation was started as a standard study of conformational entropies in a joint NMR and MD project, similar to several earlier similar studies.[16,21,22,23,24,25,26,27,28] However, we soon realised that it was not possible estimate such entropies for our proteins with any statistical significance (Figures 2–3 and Table 1) for three commonly used methods, DDH, VMA, and QHA. We then went back and studied a protein that had been studied before, BPTI, finding similar severe convergence problems.

To explain these problems, we suggest a simple protein model: A collection of uncorrelated dihedrals with two possible states, separated by an activation barrier with a uniform (random) distribution. According to the Arrhenius equation, such a model gives rise to an entropy that increases linearly with $\log(t_{\mathrm{sim}})$ as is shown in Figure 5. The model is intuitively reasonable and it is in qualitative accordance with the results of the simulations (Figures 6 and 7). Thus, the reason why the entropies do not converge is that as the time of the simulation is extended, new barriers are passed and new conformations are observed for some dihedrals. The entropy of each dihedral is related to the number of conformations sampled (Eqn. 10) and the total entropy is approximately the sum of the individual entropies. Therefore, the entropy will increase (on average) when the simulation time is increased, either until the protein unfolds during the simulation or until all conformations of all dihedrals with a significant probability are sampled, as is illustrated by our extended model. Moreover, it is not enough to sample all conformations once – they have to be sampled so many times that a converged probability of each conformation is obtained. Analysis of a 1-ms simulation of BPTI confirms our results and indicates that even after this time, both the entropy and the sampling of the dihedral angles are still linearly increasing with logarithmic time.

This not only explains the poor convergence of the simulated entropies, but it has far-reaching consequences on all type of MD simulations, because it shows that most MD simulations of proteins will not be equilibrated within 1 ms in the sense that all properties are independent of simulation time.[89] On the other hand, we argue that this does not mean that MD simulations of proteins are meaningless. Instead, different properties show varying sensitivity to this lack of strict equilibration. In this article, we have studied three different properties, viz. conformational entropies, MM/GBSA ligand-binding free energies, and FEP free energies. We show that the conformational entropies (both absolute and relative) are extremely sensitive to the equilibration and they show uncertainties of 12–89 kJ/mol even after 380–500 ns MD simulations and the entropy changes by over 250 kJ/mol when the BPTI simulation is extended to 1 ms. The MM/GBSA free energies are intermediate in sensitivity, showing a variation of 3 and 12 kJ/mol during a 500 ns simulation for the two test systems studied here. The larger variation is caused by an important conformation that is sampled only occasionally during the long simulation. Such energies may or may not be useful, depending on the application. However, for the FEP free energies, the variation is less than 0.2 kJ/mol for a 10% charge perturbation and less than 0.1 kJ/mol for a 10% van der Waals perturbation over the 500 ns simulation, indicating that the total FEP free energy would show a variation of ~0.6 kJ/mol. Thus, FEP free energies are extremely stable and not at all affected by the lack of equilibration. Strictly, such a convergence investigation is needed for each property of interest to ensure that it is independent of the length of the simulation. Unfortunately, the convergence may depend on the simulated protein and it may change if the simulation is extended.

Our message is not completely new. A few groups have already pointed out that it is hard to converge MD simulations of protein and that in most cases, equilibrium sampling is not obtained.[77,88,91] It has also been discussed how the completeness of sampling can be measured and several measures of convergence have been suggested.[77,87,88,92,93,94] There are methods to show that the sampling is incomplete, but there does not seem to be any way to detect whether a simulation has failed to visit some important parts of the phase space, unless these parts are known beforehand. We have here used an objective measure of how large part of the (dihedral) phase space that has been visited, viz. a plot of the portion of the dihedral minima that has been visited during the simulation (Figures 5, 6, and 10b). However, as noted above, complete sampling of the dihedral

space is not always needed, nor is it enough to ensure convergence of the simulations. Still, even if different proteins may show different slopes and total ratios in such plots, it is clear that as long as they show a steady increase in the number of visited minima, the simulations cannot be considered converged or fully equilibrated (i.e. at convergence, the ratio of fully visited minima should either reached 100% or a fixed stable value).

This study has important implications on how MD simulations should be analysed.[77,88,89] Much effort has been spent on developing methods to determine the required equilibration time.[95,96,97] If we agree that simulations never will be strictly equilibrated, the procedure can be simplified by discarding only parts of the simulation that give erroneous results, i.e. the first short part, during which unphysical interactions caused by the set-up of the simulation are removed. This part is characterised by monotonic trends during the start of the simulation and typically takes a few 100 ps.[38,98] All simulation time after this initial equilibration is equally important and should be used for analysis, even if some properties show significant variations. The sampling can typically be improved by running several short independent simulations, exploiting the uncertainty inherent in the set-up of the simulations, such as the starting velocities, the placement of solvent molecules, alternative configurations in the crystal structure, rotation of groups, or the protonation states of the protein residues.[76] In fact, the main challenge is to decide the statistical precision of the results.[77]Moreover, there is no way to detect whether we have missed important parts of the phase space. Instead, the interpretation of MD simulations has to rely on the assumption that we sample enough parts of the phase space around the native structure to obtain reliable estimates of our property of interest.

Acknowledgements
This investigation has been supported by grants from the Swedish research council (project 2010-5025) and from the Research school in pharmaceutical science. It has also been supported by computer resources of Lunarc at Lund University, C3SE at Chalmers University of Technology, NSC at Linköping University, and HPC2N at Umeå University. We are grateful to D. E. Shaw research for providing the 1-ms BPTI trajectories.

References

1 Adcock, S. A.; McCammon, J. A. *Chem. Rev.* **2006**, 106, 1589-1615
2 van Gunsteren, W. F.; Dolenc, J.; Mark, A. E. *Curr. Opin. Struct. Biol.* **2008**, 18, 149-153
3 van der Kamp, M. W.; Shaw, K. E.; Woods, C. J.; Mulholland, A. J. *J. R. Soc. Interface* **2008**, 5, S173-S190
4 Simonson, T. *Curr. Opin. Struct. Biol.* **2001**, *11* 243-252.
5 Michel, J, Essex, J W *J. Comput. Aided Mol. Des.*, **2010**, 24, 639-658
6 Ferenczy, G. G.; Keser, G. M. *Drug Discov. Today* **2010**, *15*, 919-932
7 Falconer, R. J.; Penkova, A.; Jelesarov, I.; Collins, B. M. *J. Mol. Recog.* **2010,** *23*, 395-413
8 Brady, G. P.; Sharp, K. A. *Curr. Opin. Struct. Biol.* **1997**, 7, 215-221.
9 Meirovitch, H. *Curr. Opin. Struct. Biol.* **2007**, 17, 181-186.
10 Kollman, P., A., Massova, I., Reyes, C., Kuhn, B., Huo, S., Chong, L., Lee, M., Lee, T., Duan, Y., Wang, W., Donini, O, Cieplak, P., Srinivasan, J., Case, D. A., Cheatham III, T., E., *Acc. Chem. Res.* **2000**, 33, 889-897
11 Schäfer, H.; Mark, A. E.; van Gunsteren, W. F. *J. Chem. Phys.* **2000**, 113, 7809-7817.
12 Peter, C.; Oostenbrink, C.; van Dorp, A.; van Gunsteren, W. F. *J Chem. Phys.* **2004,** 120, 2652-2661.
13 Carlsson, J., Åqvist, J. *J. Phys. Chem. B* **2005**, *109*, 6448-6456
14 Carlsson, J.; Åqvist, J. *Phys. Chem. Chem. Phys*. **2006**, *8*, 5385-5395
15 Minh, D. D.; Hamelberg, D.; McCammon, J. A. *J. Chem. Phys.* **2007**, 127, 154105.
16 Jarymowycz, V. A.; Stone, M. J. *Chem. Rev.* **2006**, 106, 1624-1671.

17 Igumenova, T. I.; Fredrick, K. K.; Wand, A. J. *Chem. Rev.* **2006**, 106, 1672-1699.
18 Akke, M., Brüschweiler, R.; Palmer, A. G. *J. Am. Chem. Soc.* **1993**, *115*, 9832-9833.
19 Yang, D.; Kay, L. E. *J Mol. Biol.* **1996**, *263*, 369-382
20 Best, B., R.; Clarke, J.; Karplus, M. *J. Mol. Biol.* **2005**, *349*, 185-203
21 Wrabl, J. O.; Shortle, D.; Woolf, T. B. *Proteins: Struct., Funct., Gen.* **2000**, *38*, 123-133
22 Hsu, S.-T. D.; Peter, C.; van Gunsteren, W. F.; Bonvin, A. M. J. J. *Biophys. J.* **2005**, *88*, 15-24.
23 Deng, N. J.; Yan, L.; Singh, D.; Cieplak, P. *Biophys. J* **2006**, *90*, 3865-3879
24 Trbovic, N.; Cho, J.-H.; Abel, R.; Friesner, R. A., Rance, M.; Palmer III, A. G. *J. Am. Chem. Soc.* **2009**, *131*, 615-622.
25 Li, D.-A.; Brüschweiler, R. *Phys. Rev. Lett.* **2009**, *102*, 118108
26 Li, D.-A.; Brüschweiler, R. *J. Am. Chem. Soc.* **2009**, 131, 7226–7227.
27 Diehl, C.; Genheden, S.; Modig, K.; Ryde, U.; Akke, M. *J. Biomol. NMR*, **2009**, 45, 157-169
28 Evans, D. A., Bronowska, A. K. *Theor. Chem. Acc.* **2010**, *125*, 407-418
29 Syme, N. R.; Dennis, C.; Bronowska, A.; Paesen, G. C.; Homans, S. W. *J. Am. Chem. Soc.* **2010**, *132*, 8682-8689
30 Fleischman, S. H.; Brooks, C. L. **1987**, *67*, 3029-3037.
31 Wan, S.; Stote, R. H.; Karplus, M. *J. Chem. Phys.* **2004**, *121*, 9539-9548.
32 Baron, R.; van Gunsteren, W. F.;Hünenberger, P. H. *Trends Phys. Chem.* **2006**, *11*, 87-122.
33 van Gunsteren, W. F.; Bakowies, D.; Baron, R.; Chandrasekhar, I.; Christen, M.; Daura, X.; Gee, P.; Geerke, D. P.; Glättli, A.; Hünenberge, P. H.; Kastenholz, M. A.; Oostenbrink, C.; Schenk, M.; Trzesniak, D.; van der Vegt, N. F. A.; Yu, H. B. *Angew. Chemie. Int. Ed.* **2006**, *45*, 4064-4092.
34 Chan, C.-E.; Chen, W.; Gilson, M. K. *J. Chem. Theory Comput.* **2005**, *1*, 1017-1028
35 Baron, R.; Hünenberg, P. H.; McCammon, J. A. *J. Chem. Theory Comput*. **2009**, *5*, 3150-3160
36 Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. *Science* **2010**, *330*, 341-346.
37 Sörme, P.; Arnoux, P.; Kahl-Knutsson, B.; Leffler, H.; Rini, J. M.; Nilsson, U. J. *J. Am. Chem. Soc.* **2005**, 127, 1737-1743.
38 Genheden, S.; Ryde, U. *J. Comput Chem.* **2010**, 31, 837–846
39 Diehl, C. *Conformational Entropy and Protein Flexibility in Drug Design Studies by NMR Spectroscopy*, Ph.D. thesis, Lund University 2010
40 Wlodawer, A., Nachman, J., Gilliland, G.L., Gallagher, W., Woodward, C. *J. Mol. Biol.* **1987**, *198*, 469-480
41 Li, H.; Robertson, A. D.; Jensen, J. H. *Proteins*, **2005**, *61*, 704-721.
42 Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct., Funct. Bioinform.* **2006**, *65*, 712-725.
43 Wang, J. M., Wolf, R. M., Caldwell, K. W., Kollman, P. A., Case, D. A. *J. Comput. Chem.*, **2004**, 25, 1157-1174
44 Bayly, C. I. , Cieplak, P. , Cornell, W. D., Kollman, P. A. *J. Phys. Chem.* **1993**, *97*, 10269-10280.
45 Besler, B. H. , Merz, K. M., Kollman, P. A. *J. Comput. Chem.* **1990**, 11, 431-439.
46 Horn, H. W.; Swope, W. C.; Pitera, J. W.; Madura, J. D.; Dick, T. J.; Hura, G.; Head-Gordon, T. *J. Chem. Phys.* **2004**, *120*, 9665-9678.
47 Seminario, J. M. *Int. J. Quantum Chem.* **1996**, 60, 1271.
48 Nilsson, K.; Lecerof, D.; Sigfridsson; E.; Ryde, U. *Acta Crystallogr. D*, **2003**, *59*, 274-289.
49 L. Hu, U. Ryde *J. Chem. Theory Comput.*, **2011**, *7*, 2452-2463.
50 Case, D.A.; Cheatham, III, T.E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, Jr., K.M.; Onufriev, A.; Simmerling, C.; Wang B.; Woods, R.. *J. Comput. Chem.* **2005**, 26, 1668-1688
51 Wu, X., Brooks, B.R. *Chem. Phys. Lett.* **2003**, 381, 512-518
52 Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A., Haak, J.R. *J Chem Phys*, 1984, 81, 3684–3690
53 Darden, T., York, D., Pedersen, L. *J. Chem. Phys.* **1993**, 98, 10089-10092
54 Ryckaert, J. P, Ciccotti, G., Berendsen, H. J. C. *J. Comput. Phys.*, **1977**, 23, 327-341
55 Edholm, O.; Berendsen, H. J. C. *Mol. Phys.* **1984**, *51*, 1011-1028.

56 Demchuk, E.; Singh, H. *Mol. Phys.* **2001**, *99*, 627.636
57 Karplus,M.; Kushick, J. N. *Macromol.* **1981,** *14*), 325-332.
58 Andricioaei, I.; Karplus, M. *J. Chem Phys.* **2001**, *115*, 6289
59 Jensen, F. *Introduction to Computational Chemistry* J. Wiley & Sons Ltd. Chichester **1999**, pp. 298-307.
60 Kongsted, J., Ryde, U. *J. Comput-Aided Mol. Design.* **2009**, 23, 63-71
61 Genheden, S.; Kuhn, O.; Mikulskis, P.; Hoffmann, D.; U. Ryde "The normal-mode entropy in the MM/GBSA method: Effect of system truncation, buffer region, and dielectric constant" J. Comput.-Aided Mol. Design, submitted.
62 Prompers, J.J.; Brüschweiler, R. *J. Phys. Chem. B* **2000**, 104, 11416-11424.
63 Li, D.-W.; Khanlarzadeh, M.; Wang, J.; Huo, S., Brüschweiler, R. *J. Phys. Chem. B* **2007**, 111, 13807-13813.
64 Li, D.-W.; Showalter, S. A.; Brüschweiler, R. *J Phys. Chem. B.* **2010**, *114*, 16036–16044
65 Wang, J.; Brüschweiler R *J. Chem. Theory Comput.* **2006**, 2, 18-24.
66 Somani, S.; Killian, B. J.; Gilson, M. K. *J. Chem. Phys.* **2009**, 130, 134102.
67 Srinivasan, J., Cheatham III, T. E., Cieplak, P., Kollman, P. A., Case, D. A. *J. Am. Chem. Soc.* **1998**, 37, 9401-9809
68 Onufriev, A., Bashford, D., Case, D. A. *Proteins* **2004**, 55, 383-394
69 Pearlman, D. A., Charifson, P. S. *J. Med. Chem.* **2001**, 44, 3417
70 Zwanzig, R. W. *J. Chem. Phys.* **1954**, *22*, 1420-1426
71 Deng, Y.; Roux, B. *J. Chem. Theory Comput.* **2006**, *2*, 1255-1273.
72 Elofsson, A.; Nilsson, L. *J Mol Biol* **1993**, 233, 766-780
73 Caves, L. S. D.; Evanseck, J. D.; Karplus, M. *Prot Sci*, **1998**, 7, 649-666
74 Zagrovic, B.; van Gunsteren, W. F. *J Chem Theory Comput* **2007**, 3, 301-311
75 S. Genheden, C. Diehl, M. Akke, U. Ryde *J. Chem. Theory Comput.*, **2010**, 6, 2176-2190
76 S. Genheden, U. Ryde *J. Comput. Chem.* **2011**, 32, 187-195
77 Grossfield, A.; Zuckerman, D. M. *Annu. Rep. Comput. Chem.* **2009**, *5*, 23-48.
78 Gohlke, H.; Case, D. A. *J. Comput. Chem.* **2003**, *25*, 238-250.
79 Schäfer, H.; Mark, A. E.; van Gunsteren, W. F. *J. Chem. Phys.* **2000**, *113*, 7809-7817.
80 Schäfer, H.; Daura, X.; Mark, A. E.; van Gunsteren, W. F. *Proteins* **2002**, *46*, 215-224.
81 Schäfer, H.; Smith, L. J..; Mark, A. E.; van Gunsteren, W. F. *Proteins* **2001**, *43*, 45-56.
82 Harpole, K. W.; Sharp, K. A. *J. Phys. Chenm. B* **2011**, *115*, 9461-6472.
83 Suárez, E.; Dia, N.; Suárez, D. *J. Chem. Theory Comput.* **2011** *7*, 2638-2653.
84 Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E. *Curr. Opin. Struct. Biol.* **2009**, *19* 120-127.
85 Killian, B. J.; Kravitz, J. Y.; Somani, S.; Dasgupta, P.; Pang, Y.-P.; Gilson, M. K. *J. Mol. Biol.* **2009**, *389*, 315-335.
86 Telium, K.; Olsen, J. G.; Kragelund, B. B. *Biochem. Biophys. Acta* **2011**, *1814*, 969-976.
87 Lyman, E.; Zuckerman, D. M. *J. Phys. Chem. B* **2007**, *111*, 12876-12882.
88 Zuckerman, D. M. *Annu. Rev. Biophys.* **2011**, *40*, 41-62.
89 Schiferi, S. K.; Wallace, D. C. *J. Chem. Phys.* **1985**, *83*, 5203-5209.
90 Genheden, S.; Ryde, U., J. Chem. Theory Comput., in press; DOI 10.1021/ct200853g.
91 Grossfield, A.; Feller, S. E.; Pitman, M. C. *Proteins*, **2009**, *67*, 31-40.
92 Mountain, R. D.; Thirumalai, D. *J. Phys. Chem.* **1989**, *93*, 6975-6979.
93 Hess, B. *Phys. Rev. E* **2002**, *65*, 031910.
94 Zhang, X.; Bhatt, D.; Zuckerman, D. M. *J. Chem. Theory Comput.* **2010**, *6*, 3048-3057.
95 Flyvberg, Hl; Petersen, H. G. *J. Chem. Phys.* **1989**, *91*, 461.
96 Allen, M. P.; Tildesley, D. J. *Computer simulation of liquids.* Clarendon Press, Oxford, **1991**.
97 Yang, W.; Nymeyer, H.; Zhou, H.X.; Berg, B.; Brüschweiler, R. *J. Compu. Chem.* **2008**, *29*, 668-672.
98 Genheden, S., Ryde, U. *J. Chem. Theory Comput.*, **2011**, 7, 3768-3778.

Table 1. DDH estimates of the protein conformational entropy ($T\Delta S$ in kJ/mol at 300 K). Δ is the difference in the entropy of the two ligands (cn2h–cn1h or L02–Gal3).

Protein	MMP12			Gal3			BPTI	
Ligand	cn1h	cn2h	Δ	Lac	L02	Δ	Run1	Run2
10 short [1]	-8499±18	-8572±20	-74±26	-7113±15	-7085±35	28±38	-2888±12	
Concatenated [2]	-8054	-8168	-114	-6878	-6792	86	-2824	
1 long [3]	-7572	-7965	-393	-6690	-6651	40	-2845	-2830
Convergence [4]	42	89	47	38	-22	-59	12	6

[1] The results of ten independent simulations of 10 ns (MMP12), 20 ns (Gal3), or 40 ns (BPTI) length. The reported uncertainty is the standard error over the ten simulations, i.e. the standard deviation of the results of the ten simulations divided by $\sqrt{10}$.
[2] The ten short simulations were concatenated and the entropy was calculated for the concatenated trajectory.
[3] The results of a single long simulation (380 ns for MMP12, 500 ns for the other two proteins).
[4] The difference between the estimates obtained with the full trajectory and when excluding the last 100 ns of the trajectory. A positive value indicates that the entropy increases the last 100 ns.

Figure 1. Ligands of Gal3 (Lac and L02) and MMP12 (cn1h and cn2h) studied in this paper.

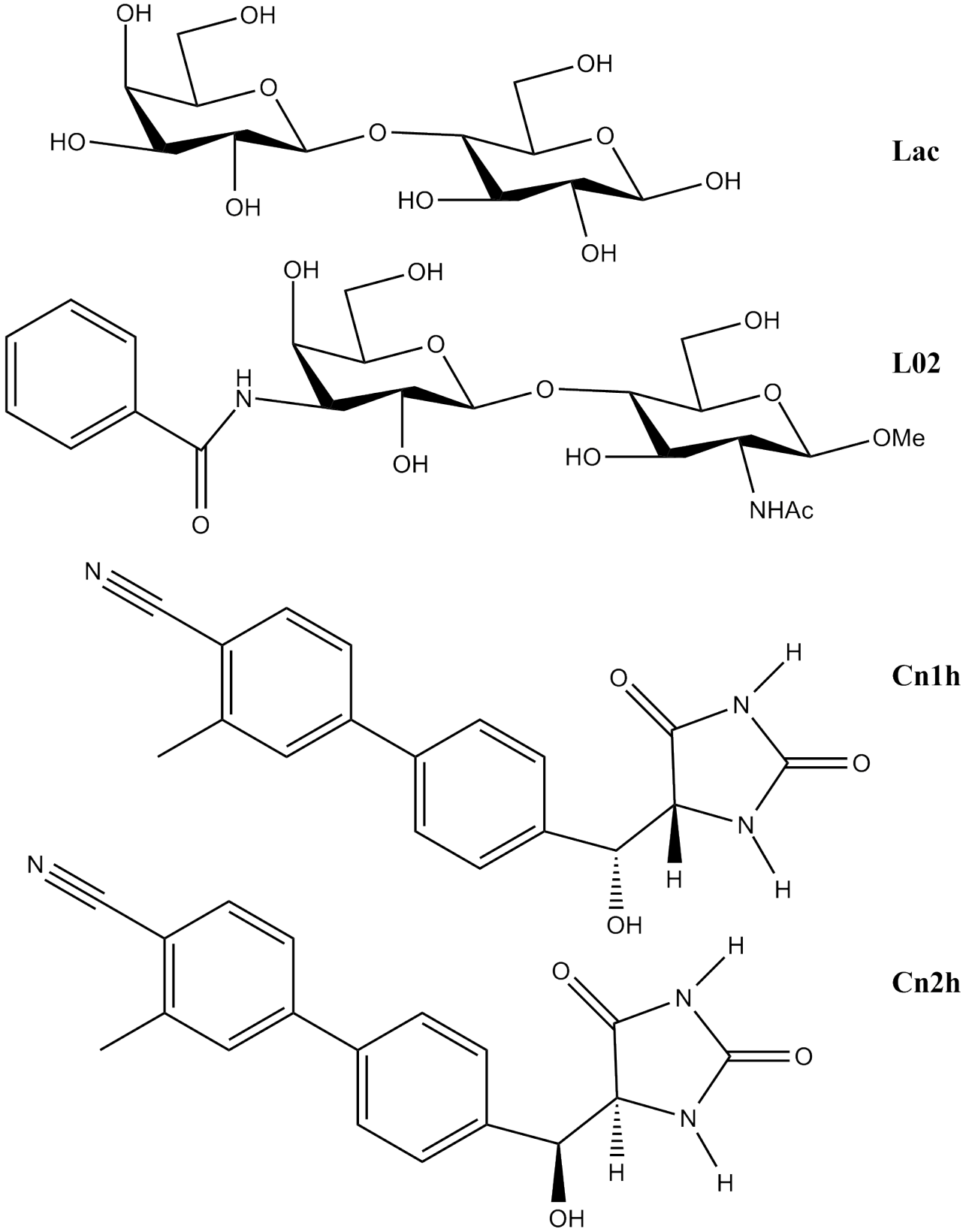


Figure 2. Protein conformational entropy from DDH analysis for the MMP12 simulations with the cn1h and cn2h ligands, as well as the difference in this entropy of the two states, as a function of the length of the simulation. The entropy ($T\Delta S$ in kJ/mol at 300 K) is relative to the estimate at 380 ns. At each time, the entropy was calculated by Eqn. 2, using all snapshots up to that time.

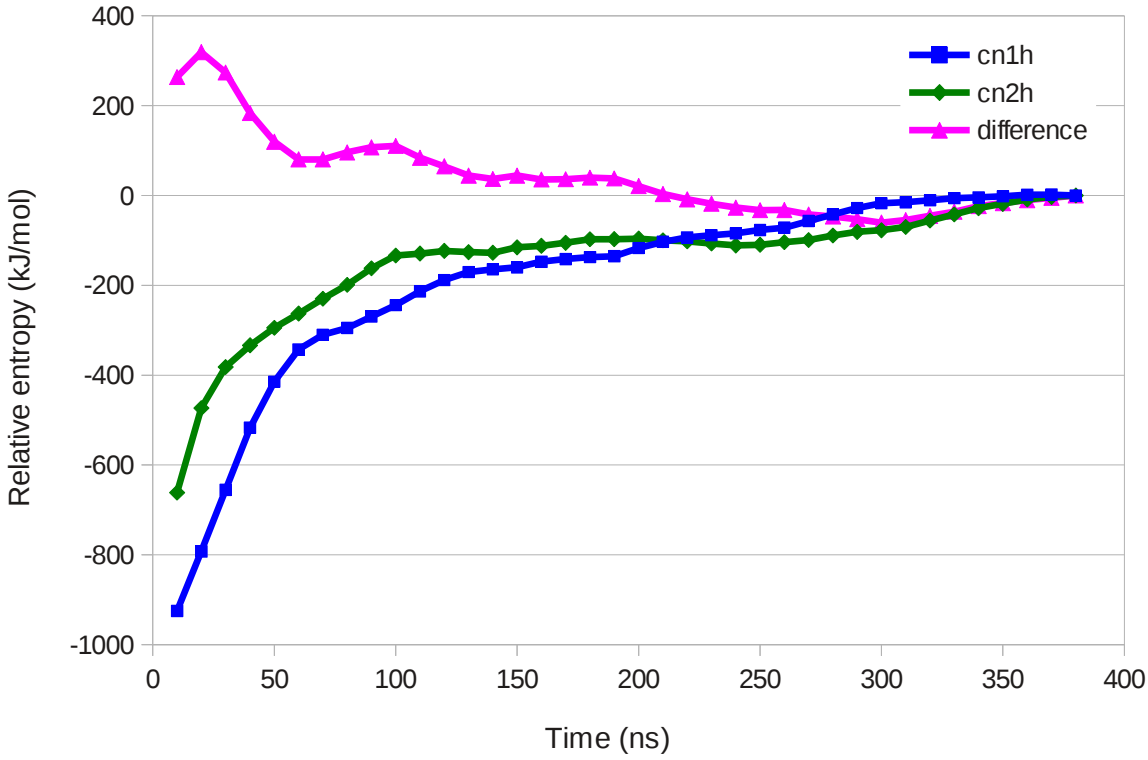


Figure 3. Protein conformational entropy from DDH analysis for the Gal3 simulations with the Lac and L02 ligands, as well as the difference in entropy of the two states, as a function of the length of the simulation. The entropy ($T\Delta S$ in kJ/mol at 300 K) is relative to the estimate at 500 ns.

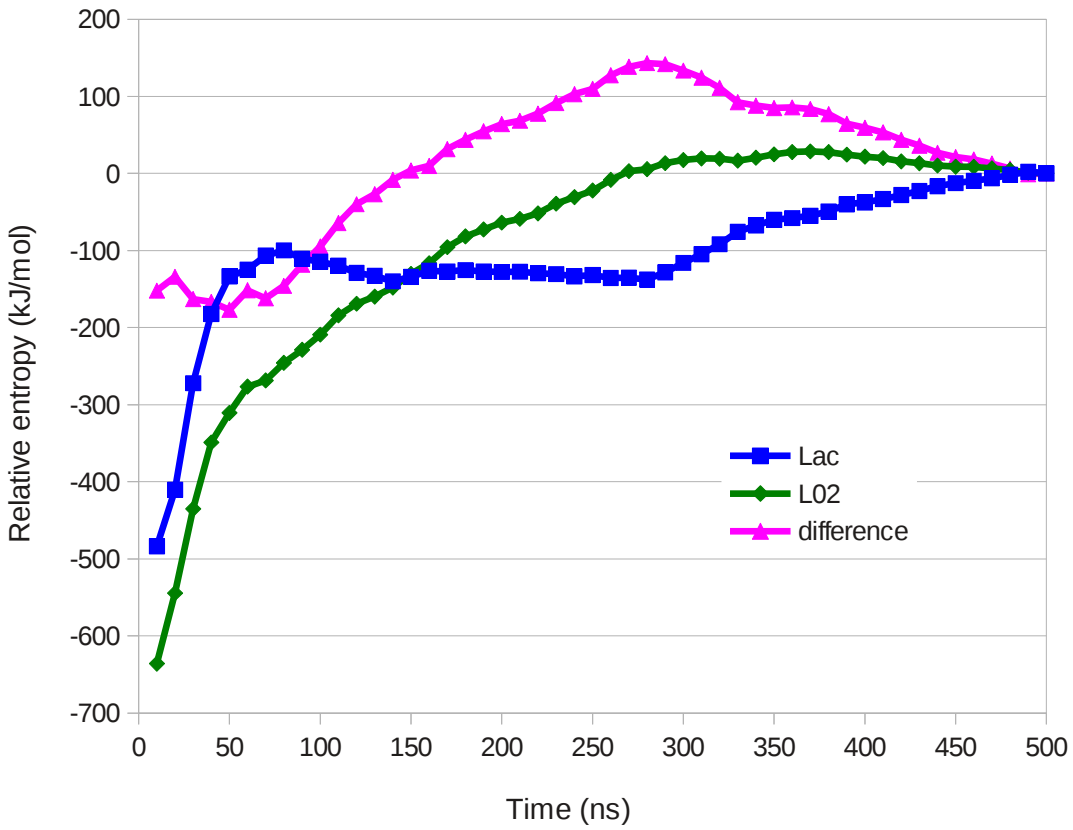


Figure 4. Protein conformational entropy from DDH analysis of two 500 ns MD simulations of BPTI as a function of the length of the simulation. The entropy ($T\Delta S$ in kJ/mol at 300 K) is relative to the estimate at 500 ns of the first simulation.

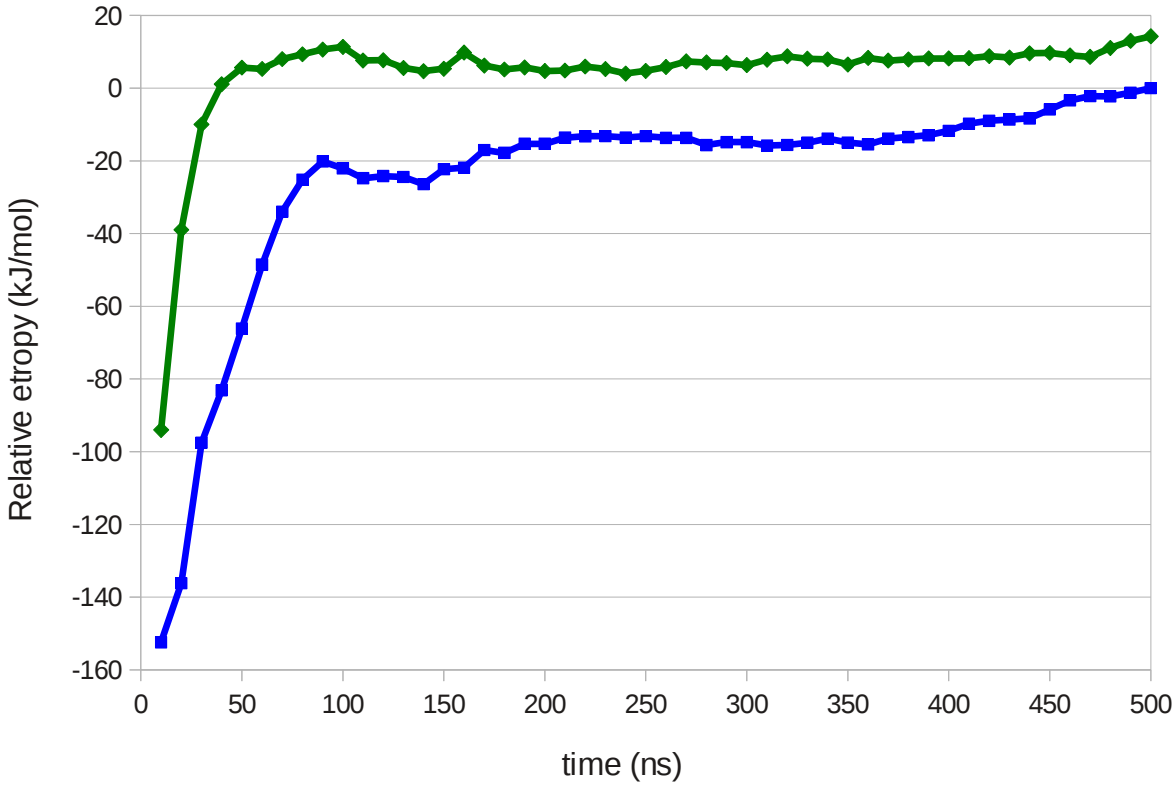


Figure 5. Conformational entropy (left axis and blue squares) and ratio of normal modes with two conformations (right axis and green diamonds) of our simple model protein (Eqns. 9 and 10) with N = 100, A_{max} = 100 kJ/mol, and $\tau_0 = h/k_BT$ as a function of t_{sim} on a logarithmic scale.

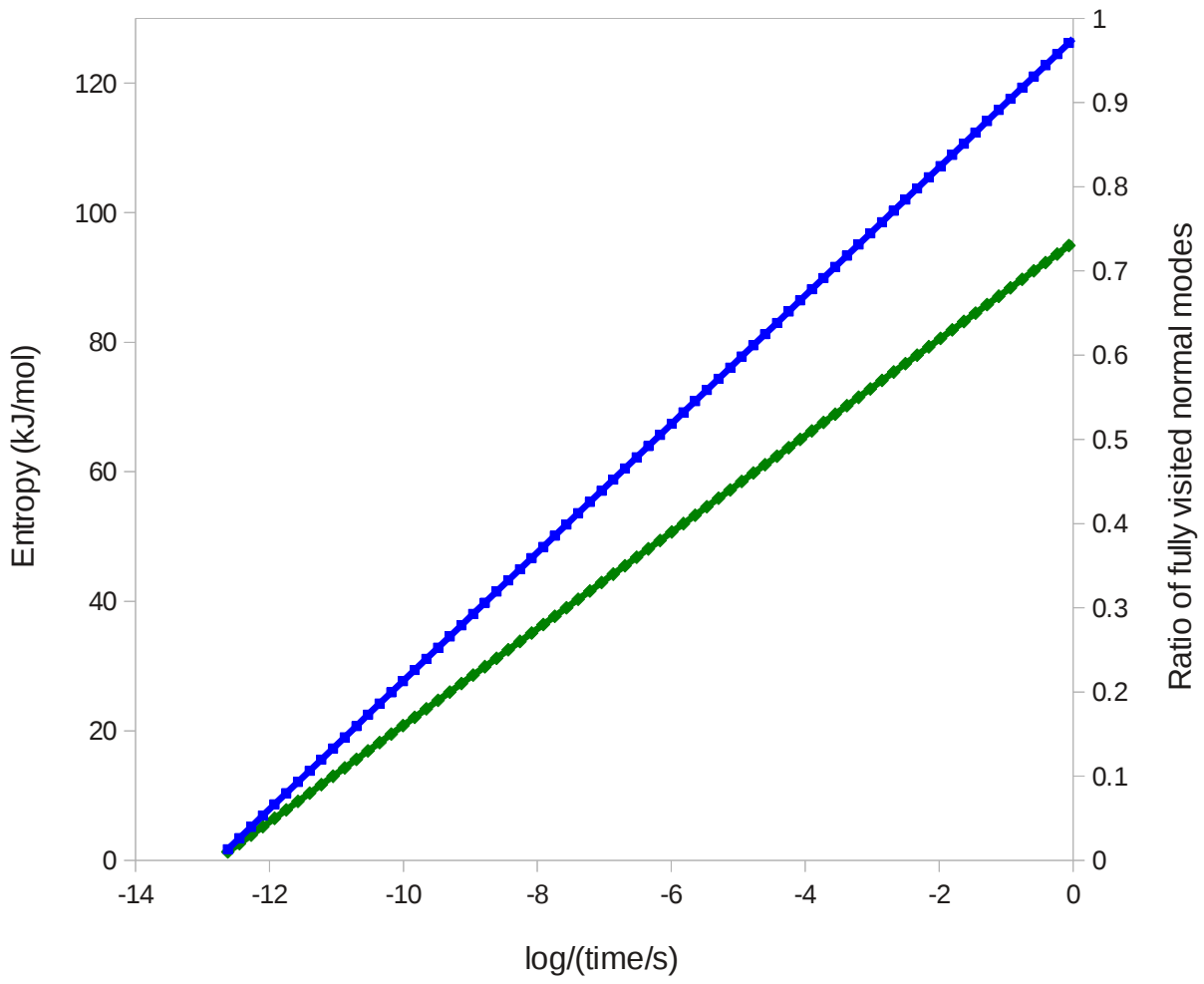


Figure 6. Dihedral phase-space sampling in the various simulations, i.e. the ratio of the dihedrals for which all available minima are visited during a certain simulation time. Each dihedral is described by a potential of the form in Eqn. 8. The domain of each minimum is defined as a $360/n_i$ interval around each minimum and it is checked whether all these domains are visited at least once during the simulation for each dihedral. Dihedrals of aromatic groups were excluded from the analysis. Note the logarithmic time scale on the *x*-axis.

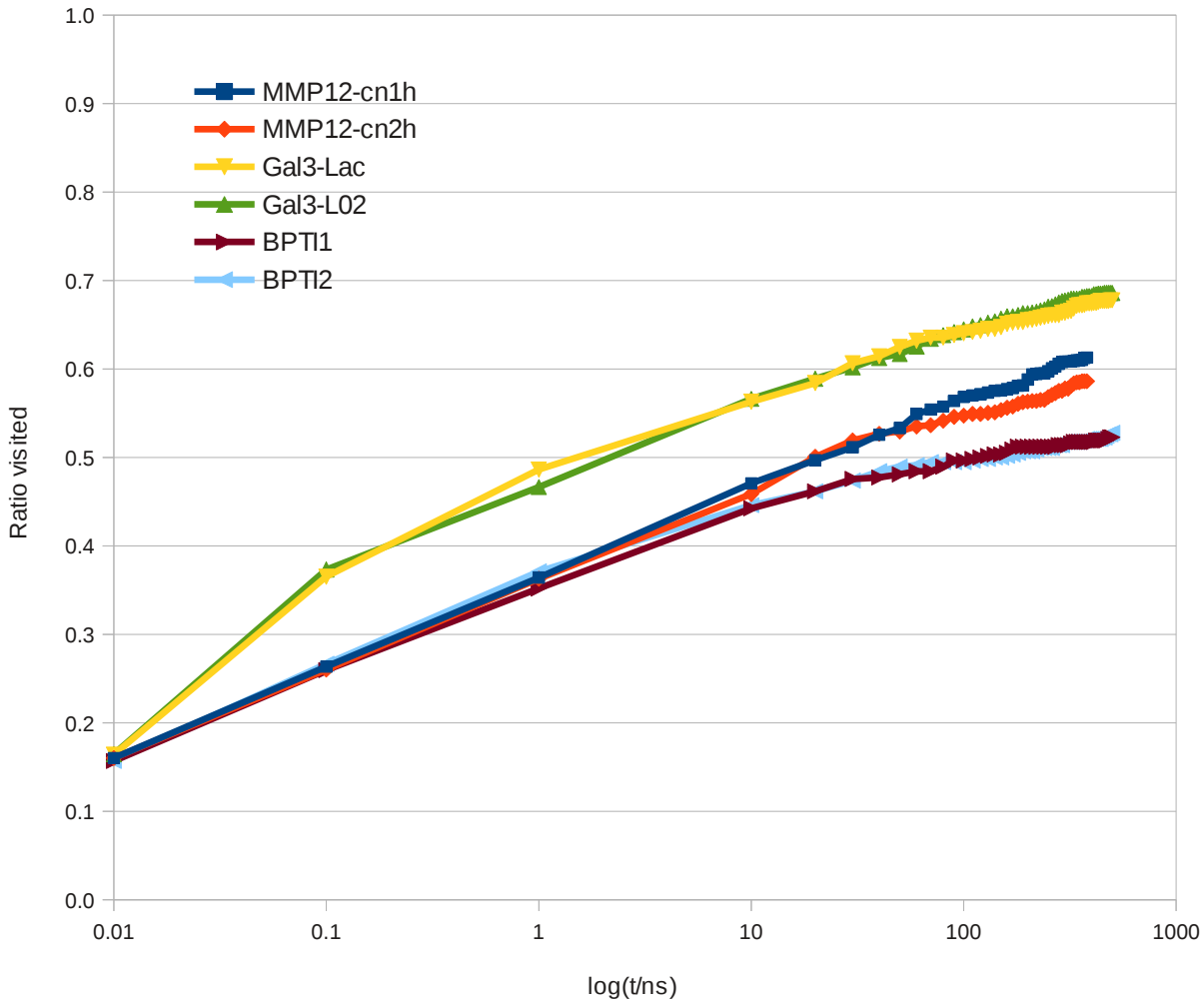


Figure 7. Protein conformational entropies from the DDH analysis for the six protein simulations as a function of the length of the simulation, plotted on a logarithmic scale. The entropy ($T\Delta S$ in kJ/mol at 300 K) is relative to the estimate at the end of the simulation. a) MMP12 simulations with the cn1h and cn2h ligands, b) Gal3 simulations with the Lac and L02 ligands, and c) the two BPTI simulations.

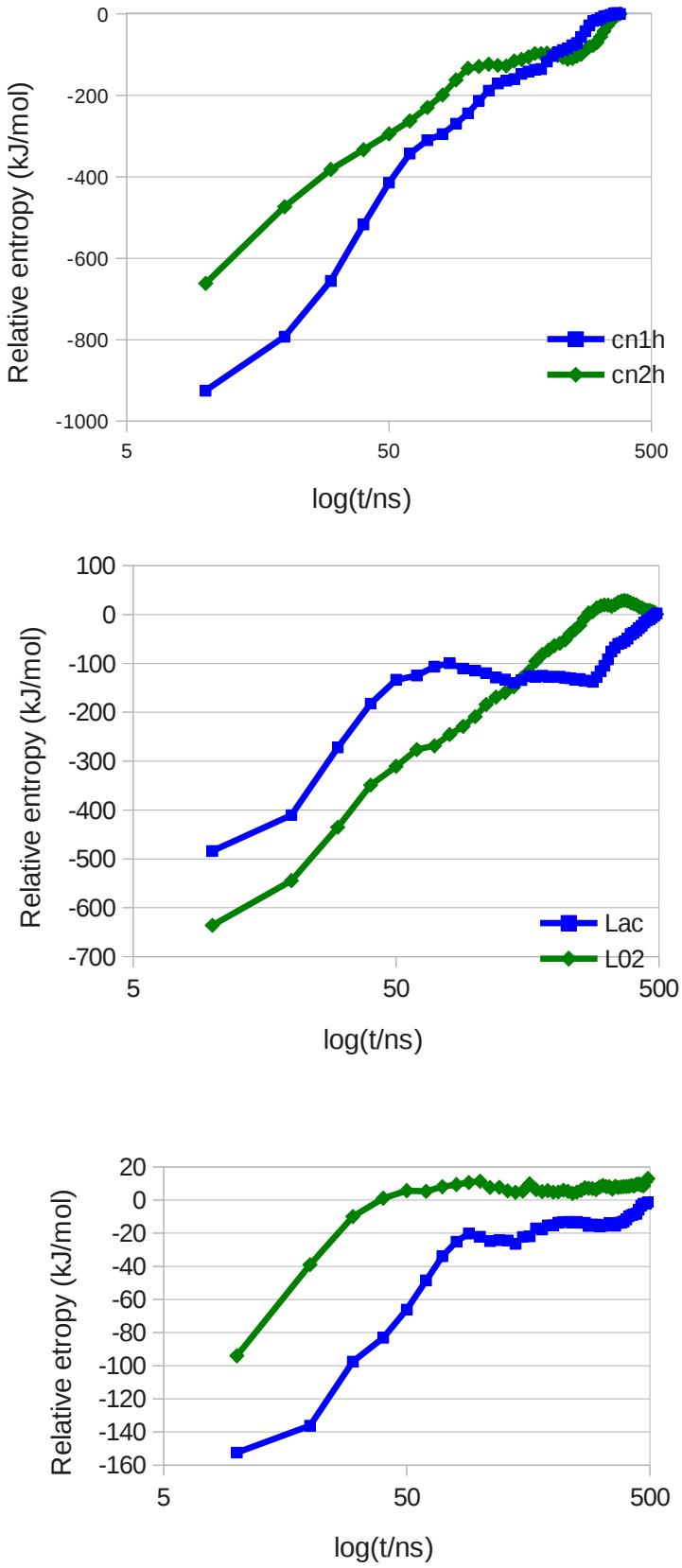
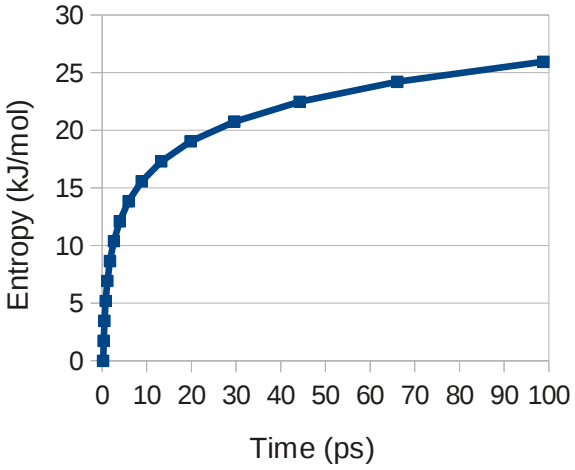


Figure 8. Plot of the conformational entropy of our model protein on two different linear time scales.



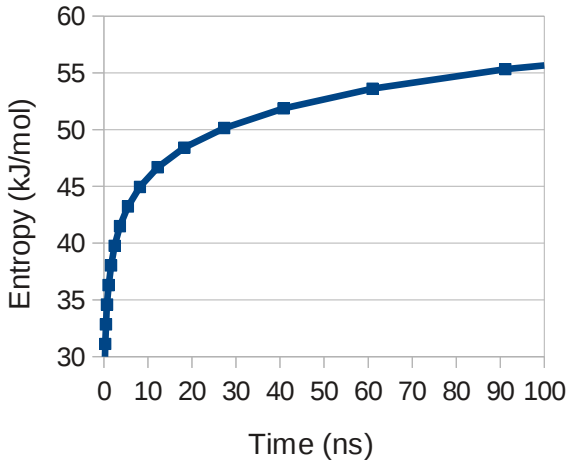


Figure 9. The conformational entropy of our extended model protein (Eqns. 9, 11, and 12) with N = 1500 (similar to what is found for Gal3 and MMP12) and A_{max} = 100 kJ/mol as a function of t_{sim} on a logarithmic scale. Four different values of ΔG_{max} were tested, 10, 20, 30, and 40 kJ/mol.

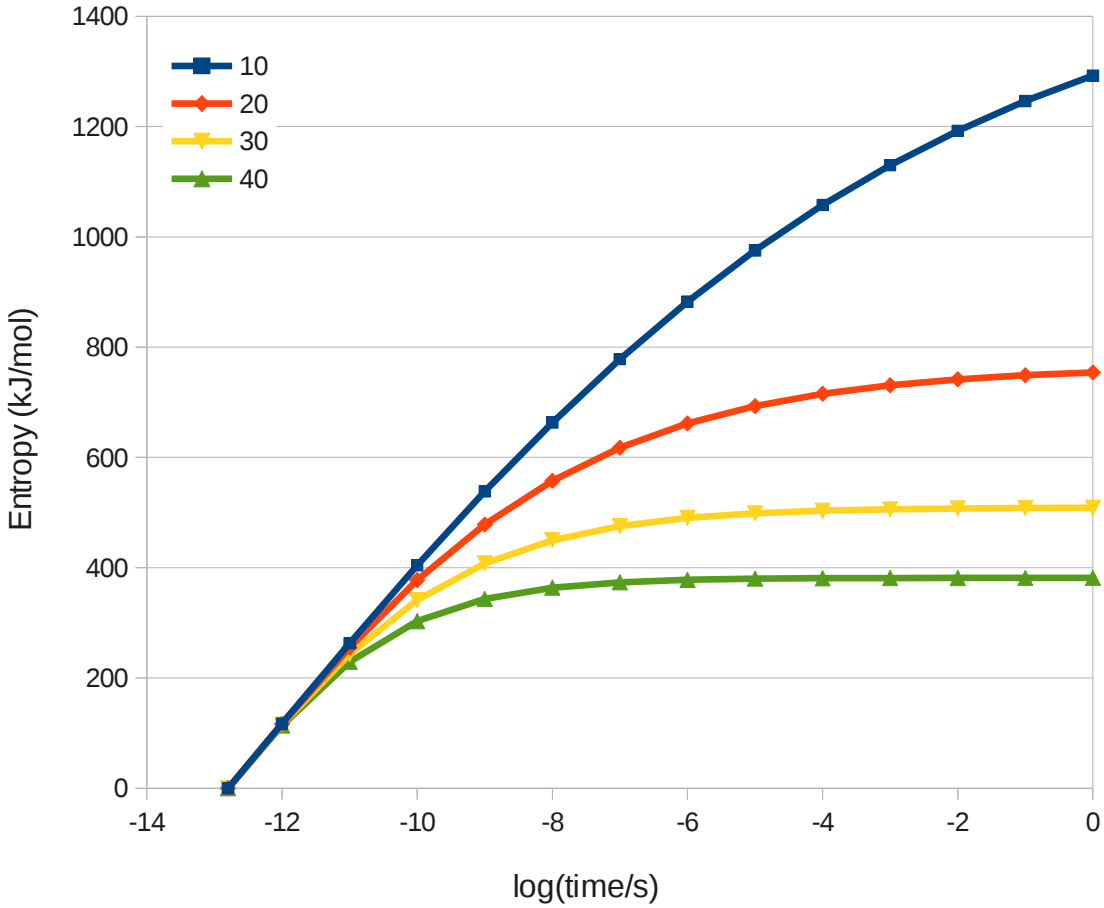


Figure 10. Protein conformational entropy from DDH analysis of the BPTI simulation of Shaw et al.[36] as a function of the length of the simulation on a linear (a) and a logarithmic (b; blue squares and left axis) time scale. The entropy ($T\Delta S$ in kJ/mol at 300 K) is relative to the estimate at 1 ms. In (b), the dihedral phase-space sampling (calculated the same way as in Figure 6) is also shown (green diamonds and right axis).

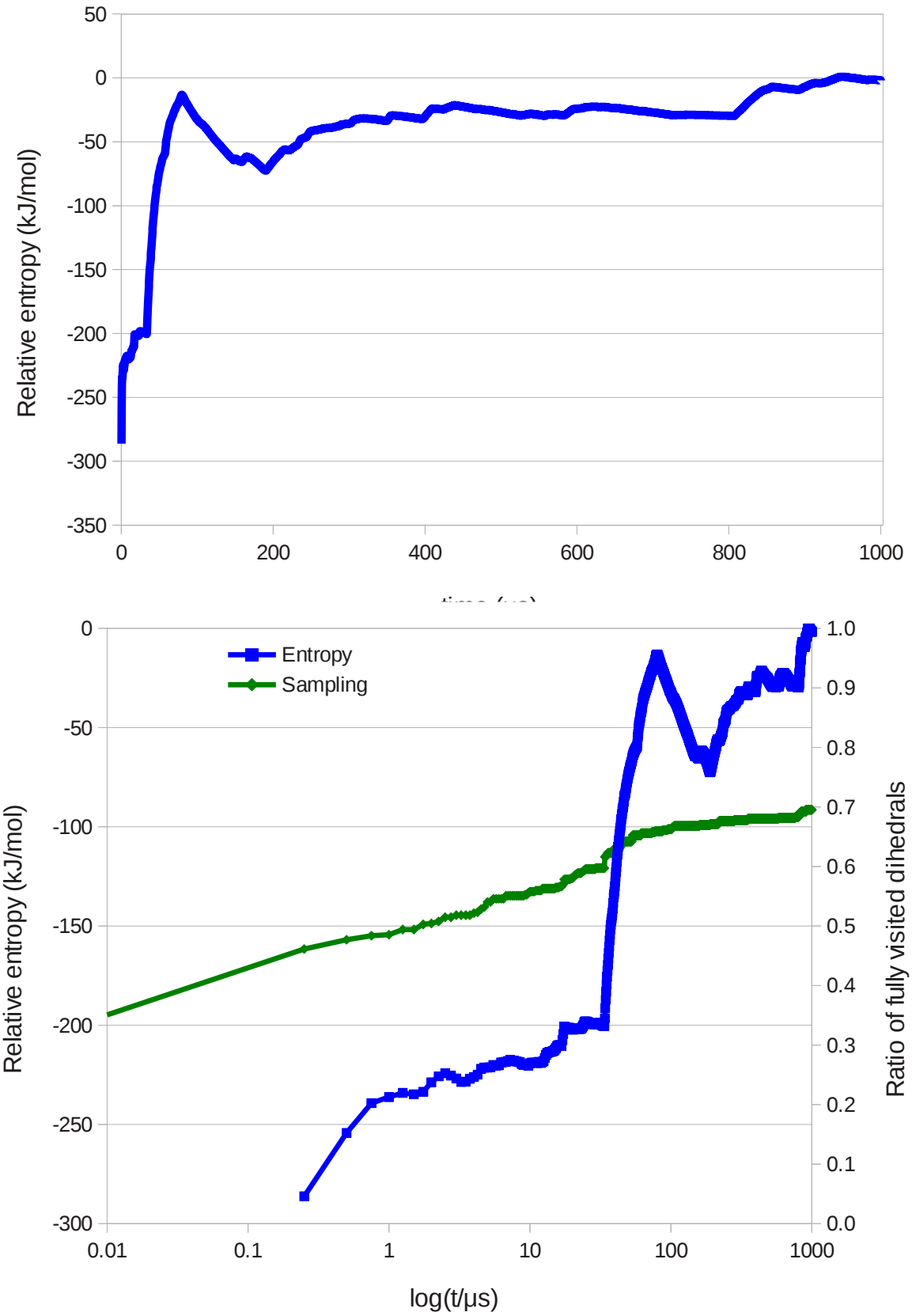


Figure 11. MM/GBSA estimates of the binding free energy of Lac to Gal3, together with the components (referring to Eqn. 6, the terms are from top to bottom in the legend, E_{vdW}, E_{ele}+ G_{pol}, G_{np}, $-TS$, E_{ele}, G_{pol}, and ΔG). For all terms, the cumulative average in kJ/mol is plotted relative to the average at 500 ns.

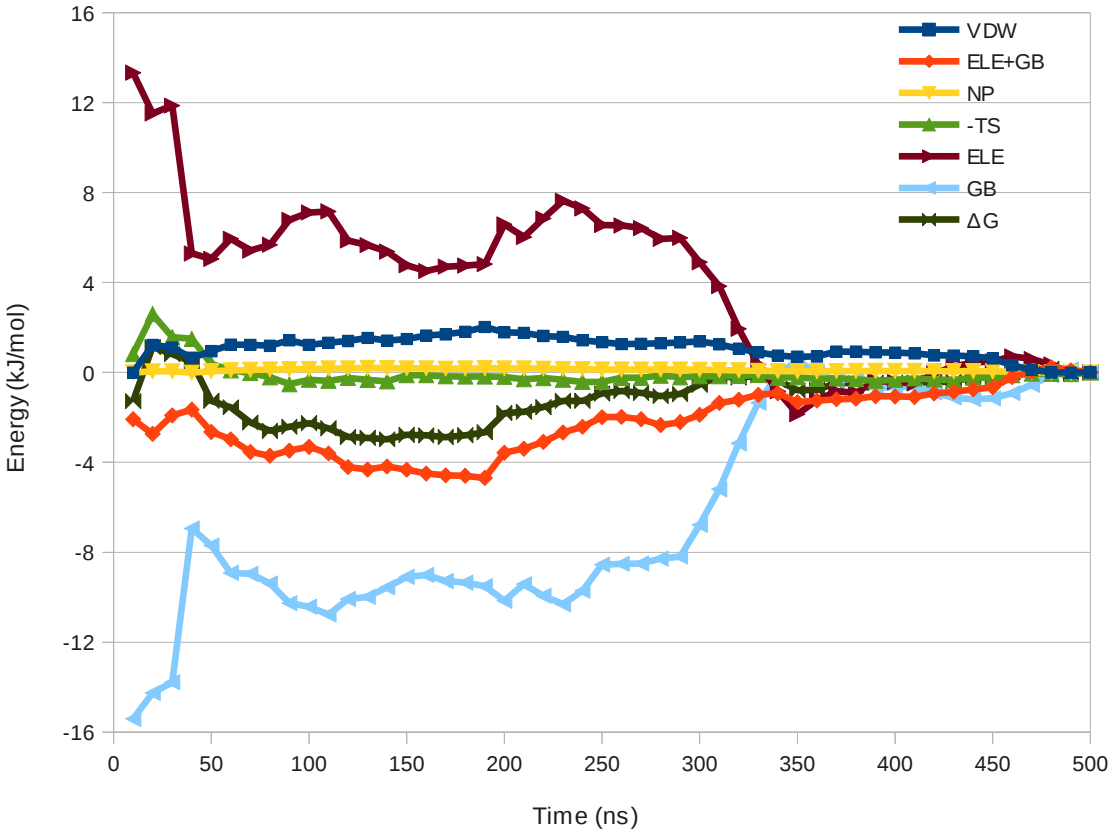


Figure 12. MM/GBSA estimates for binding free energy of L02 to Gal3 together with the components. The terms are the same as in Figure 11.

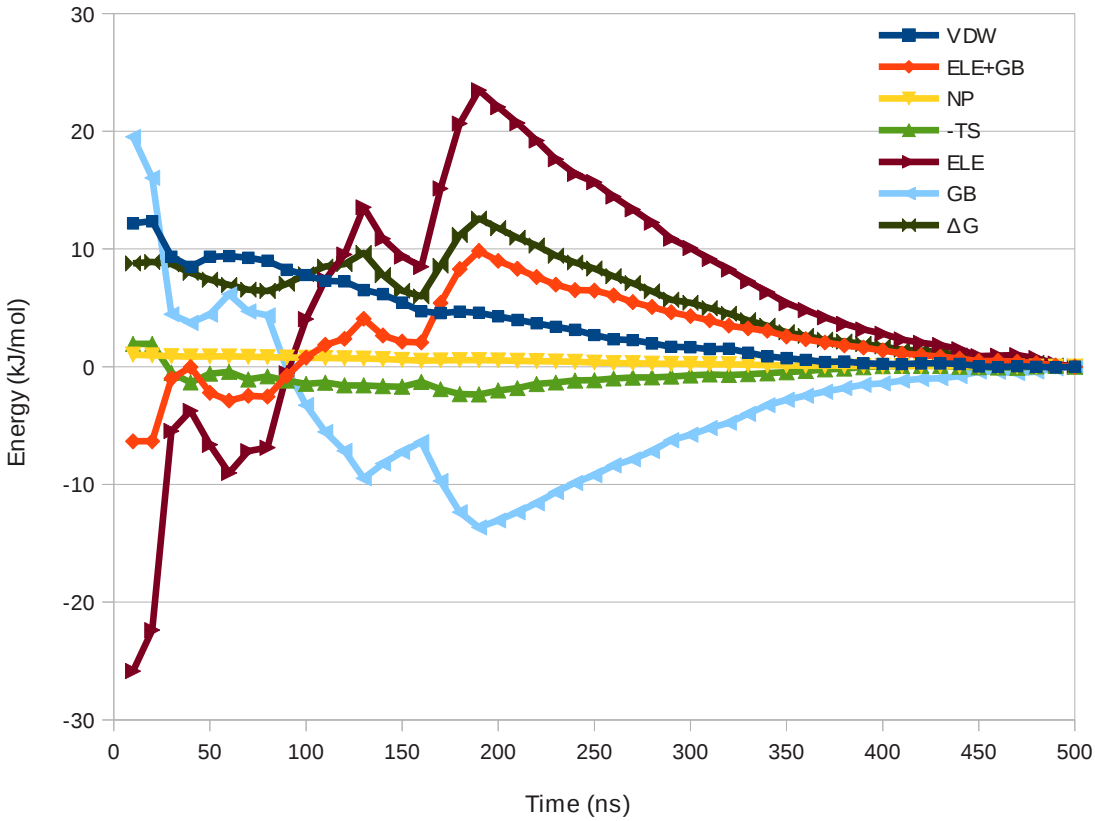


Figure 13. MM/GBSA energy for each snapshot in the ten independent simulations of Gal3-Lac (top) and the single long simulation (middle), as well as a box plot of values (bottom), with independent simulations to the left and long simulation to the right. In the box plot, the median is marked by a vertical line at the centre of the box, and the edges of the box are determined by the 25th and 75th percentiles. Finally, the whiskers extend such that 99% of the data is covered, and the plus signs outside are deemed outliers.

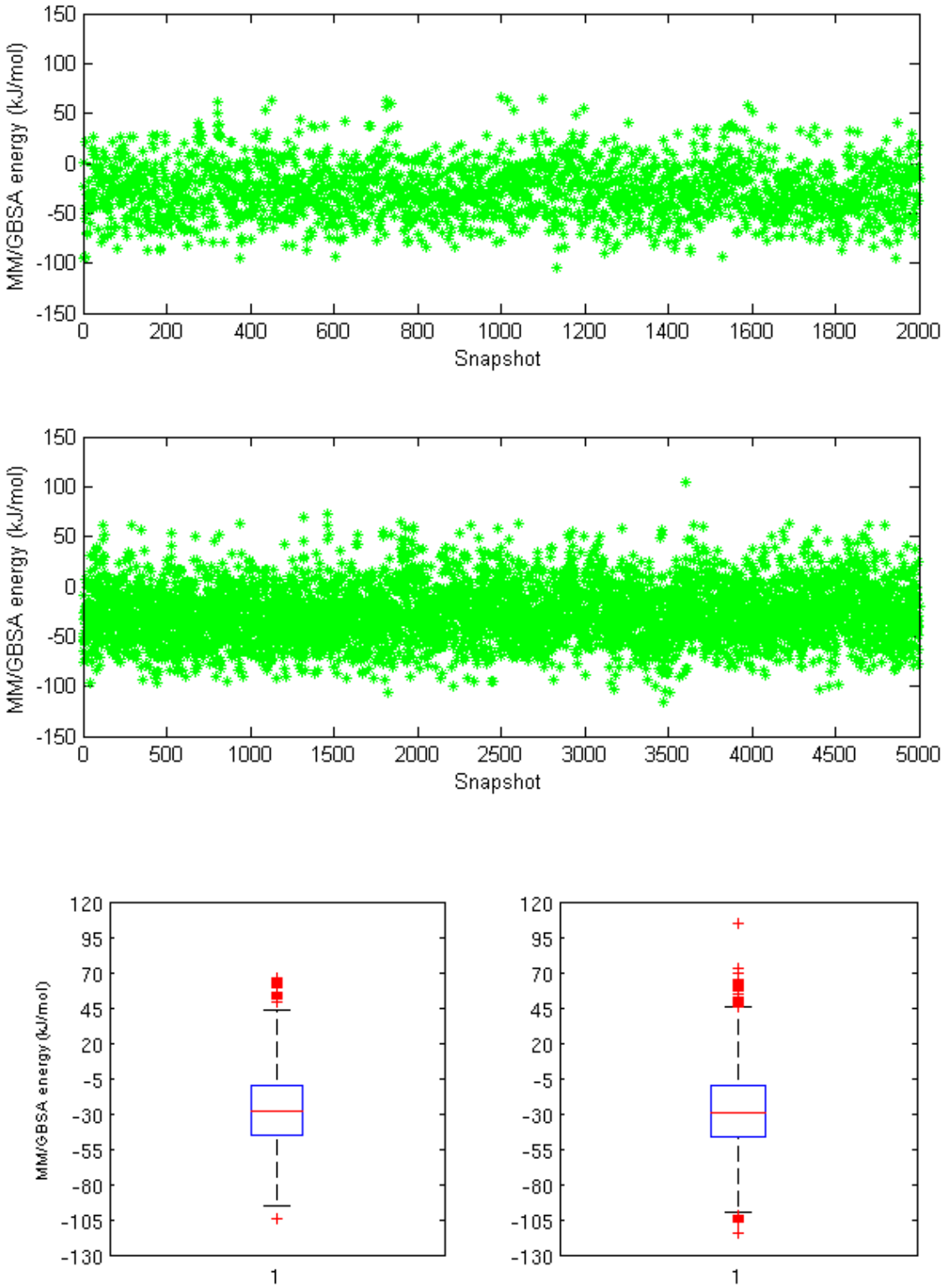


Figure 14. MM/GBSA energy for each snapshot in the ten independent simulations of Gal3-L02 (top) and the single long simulation (middle), as well as a box plot of values (bottom) with independent simulations to the left and long simulation to the right (constructed in the same way as in Figure 13).

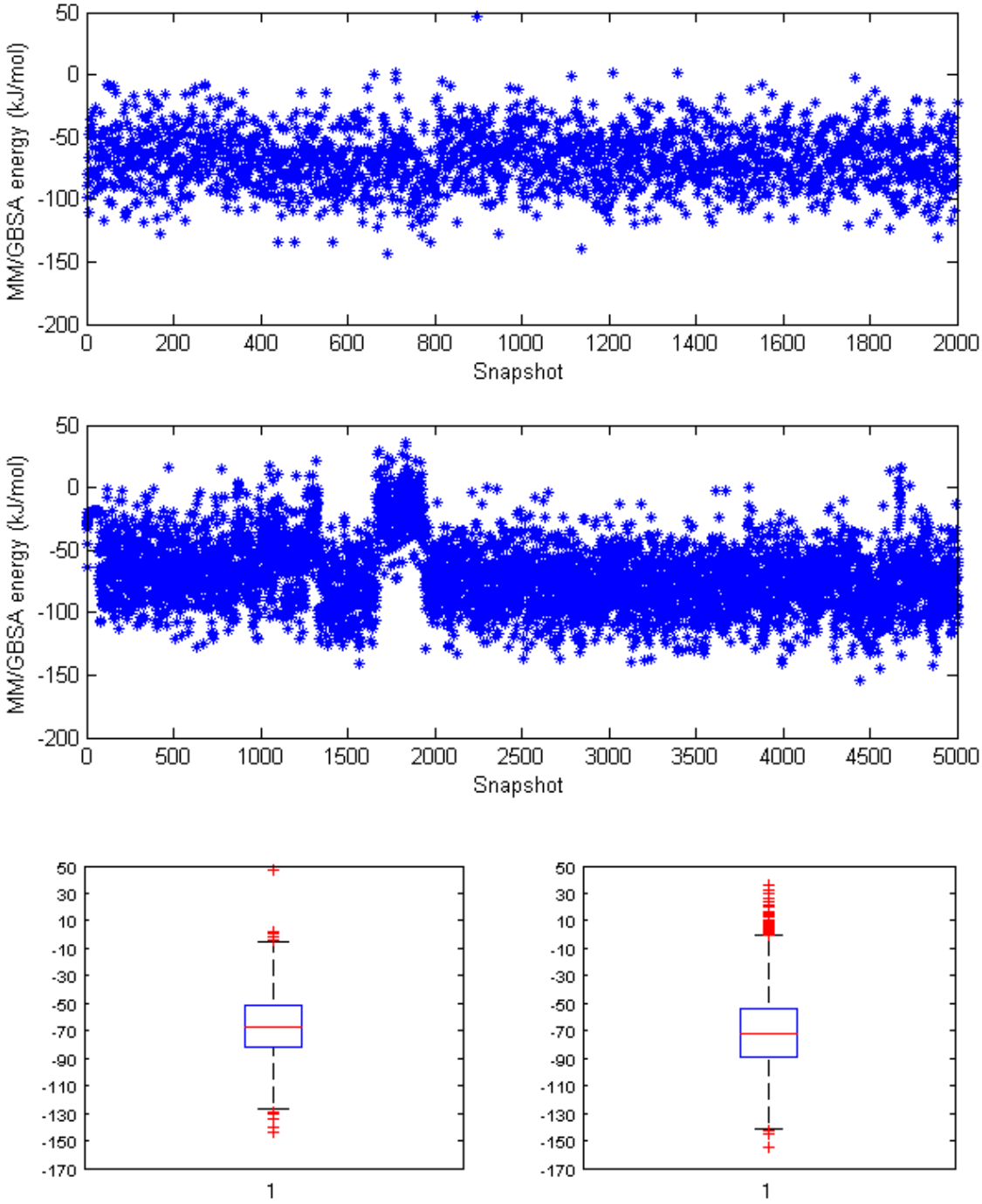


Figure 15. Cumulative estimates of the free energy of reducing the ligand charges with 10% in kJ/mol in the simulation of Gal3 with the Lac and L02 ligands.

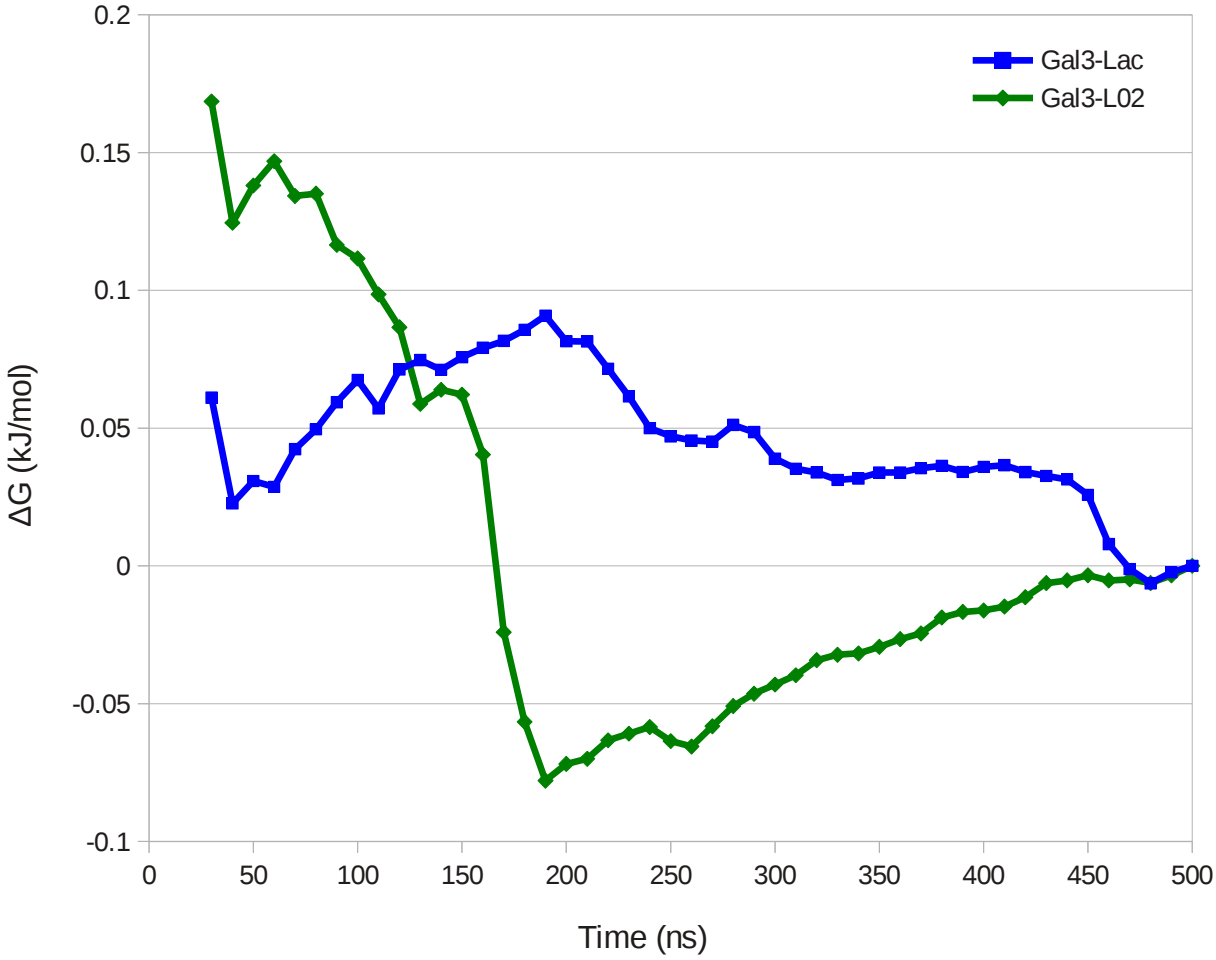
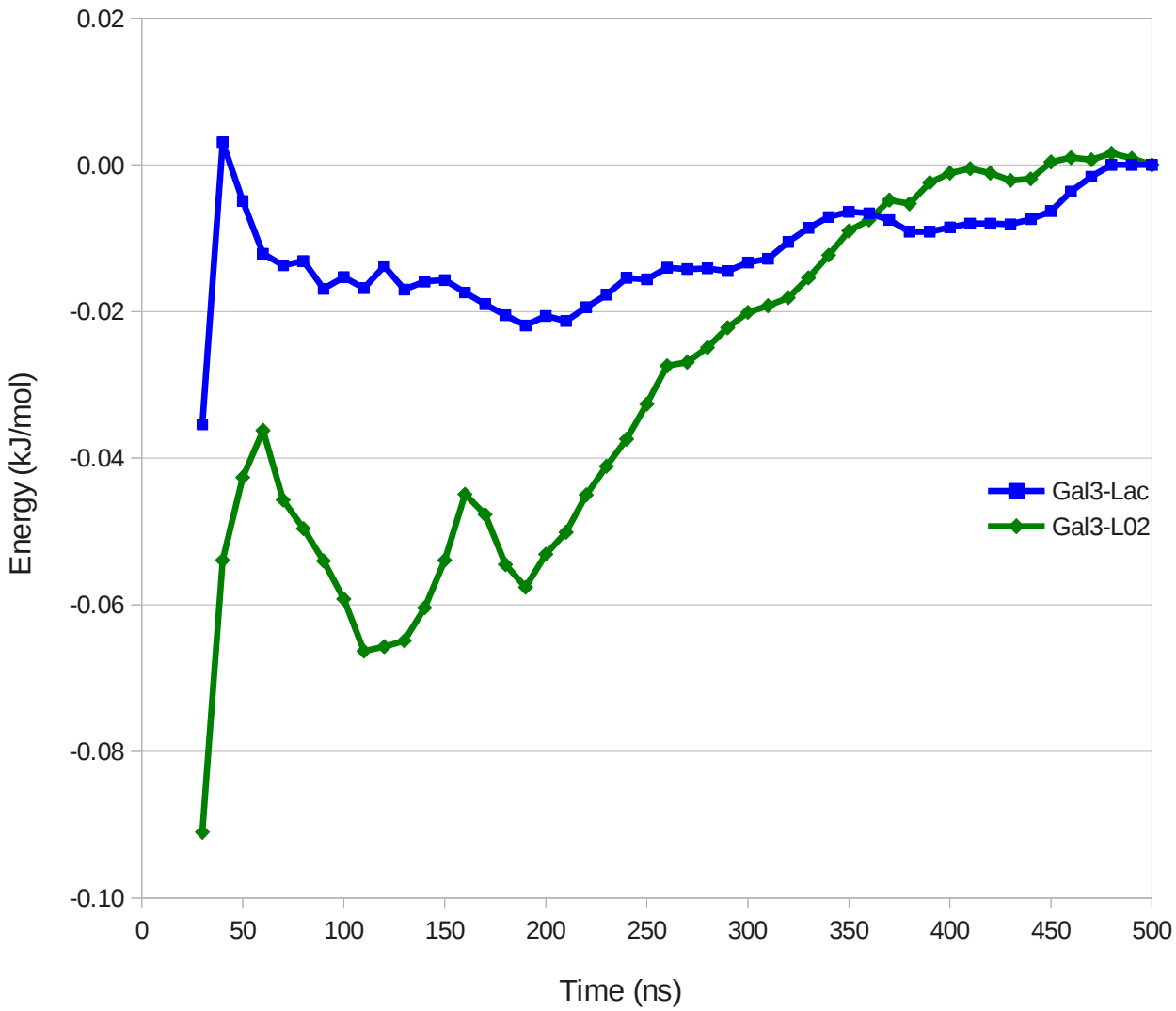
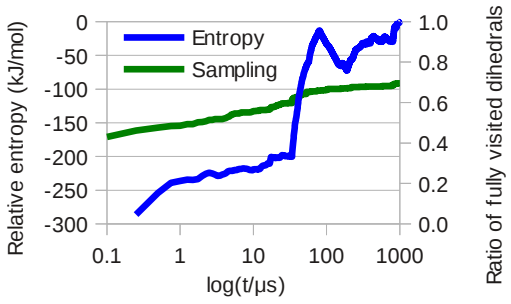


Figure 16. Cumulative estimates of the free energy of reducing the ligand van der Waals energy parameters with 10% in kJ/mol in the simulation of Gal3 with the Lac and L02 ligands.



TOC Graphics

Conformational entropies estimated from molecular dynamics do not converge to any usable precision even after 1 ms simulations.