



LUND UNIVERSITY

A comparison of different initialization protocols to obtain statistically independent molecular dynamics simulations.

Genheden, Samuel; Ryde, Ulf

Published in:
Journal of Computational Chemistry

DOI:
[10.1002/jcc.21546](https://doi.org/10.1002/jcc.21546)

2011

[Link to publication](#)

Citation for published version (APA):
Genheden, S., & Ryde, U. (2011). A comparison of different initialization protocols to obtain statistically independent molecular dynamics simulations. *Journal of Computational Chemistry*, 32(2), 187-195.
<https://doi.org/10.1002/jcc.21546>

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

A comparison of different initialisation protocols to obtain statistically independent molecular dynamics simulations

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

2012-02-07

Abstract

We study how the results of molecular dynamics (MD) simulations are affected by various choices during the set-up, e.g. the starting velocities, the solvation, the location of protons, the conformation of His, Asn, and Gln residues, the protonation and titration of His residues, and the treatment of alternative conformations. We estimate the binding affinity of ligands to four proteins calculated with the MM/GBSA method (molecular mechanics combined with a generalised Born and surface area solvation energy). For avidin and T4 lysozyme, all variations gave similar results within 2 kJ/mol. For factor Xa, differences in the solvation or in the selection of alternative conformations gave results that are significantly different from those of the other approaches by 4–6 kJ/mol, whereas for galectin-3, changes in the conformations, rotations, and protonation gave results that differed by 10 kJ/mol, but only if residues close to the binding site were modified. This shows that the results of MM/GBSA calculations are reasonably reproducible even if the MD simulations set up with different software. Moreover, we show that the sampling of phase space can be enhanced by solvating the systems with different equilibrated water boxes, in addition to the common use of different starting velocities. If different conformations are available in the crystal structure, they should also be employed to enhance the sampling. Protonation, ionization, and conformations of Asn, Gln, and His may also be used to enhance sampling, but great effort should be spent to obtain as reliable predictions as possible close to the active site.

Keywords: molecular dynamics simulations, reproducibility, conformational sampling, starting structure dependence, protonation, ligand-binding affinities, MM/GBSA.

Introduction

During the latest decades, molecular simulations have emerged as a powerful alternative and complement to experiments for the study of macromolecular dynamics and function. A cornerstone of scientific investigations is reproducibility: An independent scientist should be able to repeat any investigation, getting the same results within statistical uncertainty. Unfortunately, this is a problem for macromolecular simulations. The theory and the algorithms underlying molecular simulations are stable and should give the same results, provided that reasonable settings are used. However, the set-up of macromolecular simulations involve numerous more or less sophisticated guesses, and it is therefore unlikely that two independent scientists using different software would set up the simulations in the same way, which may affect the result. For example, the following choices have to be made:

1. Which starting structure to use (if there are several crystal structures or an ensemble of NMR structures).
2. How to treat residues for which several conformations are visible in the crystal structure.
3. How to add hydrogen atoms to the starting structure. Hydrogen atoms are not visible in crystal structures, except at the highest resolutions. For some residues, e.g. the backbone and the aromatic side-chains, it is quite clear where the protons should be, but for other groups, e.g. the OH groups of Ser and Thr, the proton can be anywhere on a circle, and for crystal-water molecules, the first proton can be added anywhere on a sphere around the oxygen atom. The positions of these protons strongly affect the hydrogen-bond structure. Previous investigations have shown that there is a large variation in how different software protonates crystal structures [1].
4. For neutral His residues, it is not even known to what atom the proton should be added, ND1 or NE2.
5. The net charge of ionisable residues is not known. For most residues, the normal pK_a value in water solution is quite far from 7 (<5 or >9), so that it can be assumed that the majority of the ionisable groups are in their standard protonation states also in proteins (although it is well-known that exceptions exist, especially for groups buried in the protein or close to metal sites [2,3]). However, for His residues, the reference pK_a value is 6.6–7.0 [2] so that the two charge states are almost equally likely.
6. In crystal structures, it is hard to differ between C, N, and O atoms. Therefore, there are two possible conformations of the side-chains of Asn, Gln, and His residues (related by a 180° rotation) that cannot be decided from the electron density. Only one conformation is normally given in the crystal structure, but it is the result of a more or less sophisticated guess that should be checked [4,5].
7. The macromolecule is normally solvated in a cell of water molecule, in which the coordinates of the water molecules are typically taken from an equilibrated water simulation or from water molecules on a grid, using some cut-off distance to avoid overlap with the macromolecule.
8. Starting velocities are typically either calculated from the starting forces or assigned at random according to a Maxwell–Boltzmann distribution.

Most simulation softwares have methods to add protons and water molecules, as well as to assign the starting velocities. On the other hand, it is normally up to the user to decide on issues related to the uncertainty in the crystal structure and the protonation state of His and other ionisable residues. Still, the importance of this problem has been increasingly recognised [6] and software and web-servers start to appear to facilitate these decisions [1,4,5,7,8,9]. However, all such methods are quite crude, typically based on single structures and simplified solvation models, because few investigators are willing to spend weeks of computer time for the set-up of the simulation.

A related problem of simulation methods is sampling. It is well-known that macromolecules present a rugged potential-energy surface with many distinct conformations, separated by sizeable

barriers, which makes an extensive sampling of phase space necessary, but also hard [10,11]. Many methods to enhance the sampling have therefore been proposed [12], e.g., accelerated molecular dynamics [13], self-guided molecular dynamics [14], locally enhanced sampling [15], and replica exchange [16].

However, it has also been recognised that the above listed uncertainties can be employed to enhance the sampling of the phase space by running multiple copies of the same molecular system with slightly altered initial conditions [12]. This is an appealing approach because it also allows the trivial parallelisation of the problem by distributing the computations onto several CPUs. This approach has been used in several areas, e.g. for structural transitions [17], calculations of backbone order parameters [18], and free-energy calculations [10,19,20,21,22]. In particular, it has frequently been shown that it is more favourable to run several short and independent simulations than a single long simulation [10,18,21,23,24].

The simplest and most common method to obtain several independent simulations is to assign different starting velocities to the atoms, by giving different seeds to the random number generator that assigns the velocities [10,18,19,20,21]. A few studies have also tested to increase the sampling by starting from different structures, e.g. from different crystal structures or NMR ensembles [17,25,26,27,28], or even from structures generated by an initial conformational search [25]. Recently, we have also shown that the agreement of calculated and measured order parameters is improved if the MD simulations are started from different structures obtained from conformational disorder available in the crystal structures [18].

Thus, it has previously been examined how the first two and the last points in the list above affect the results of molecular simulations and it has been shown that they can be employed to enhance the sampling of conformational space. In this work, we go one step further and investigate how the other five points affect the results, viz. the addition of protons, the solvation of the structure, and the treatment of conformational uncertainty of Asn, Gln, and His residues. As a test case, we use ligand-binding affinities, calculated by the MM/GBSA approach (molecular mechanics combined with generalised Born and surface area solvation) [29,30]) and the calculations are performed for complexes with four different proteins. Thereby, this work is a continuation of our previous study [21] that showed that it was necessary to use several independent trajectories generated by different starting velocities to obtain the same affinity for the four binding sites in the protein avidin. Here, we investigate whether such a rather small perturbation to the system is enough to sample the full space or if larger perturbations are more effective. It will be seen that the effect is somewhat system dependent, but that in general a better sampling can be obtained with larger perturbations.

Methods

Protein and ligand preparations

Four systems were considered in this study: Avidin in complex with biotin, factor Xa (fXa) in complex with a 3-amidinobenzyl-1*H*-indole-2-carboxamide inhibitor (denoted c53 as in ref. no. 31), Galectin-3 (Gal3) with a *N*-acetyllactoseamine derivative (ligand **3** in [32], L3) and the Leu99Ala T4 lysozyme mutant in complex with benzene. The ligands are shown in Figure 1. The avidin–biotin and lysozyme–benzene systems have been the subject of several previous theoretical studies [21,33,34,35,36,37,38,39,40,41,42,43,44,45],

The avidin–biotin simulations were based on the 1avd [46] crystal structure, those of fXa–c53 on 1lpk [31], those of Gal3–L3 on 1kjr [32], and those of lysozyme–benzene on 181l [47]. The preparation of the avidin, fXa, and Gal3 systems has been described previously [22,37,38], and lysozyme was set up in a similar way. In particular, all Asp and Glu residues were assumed to be negatively charged and all Lys and Arg residues were positively charged. Avidin has one histidine

residue in each subunit of the tetrameric protein, and it was normally (i.e. when it was not allowed to titrate) assumed to be protonated on the ND1 atom [39]. FXa has six histidine residues, three of which were normally assumed to be protonated on NE2, two on ND1, and one double protonated [40]. Gal3 has four His residues, one of which was normally assumed to be protonated on ND1 and three on NE2 [22]. T4 lysozyme has one histidine that was normally assumed to be protonated on ND1, according to a consideration of the hydrogen-bonding network and local environment.

All proteins were described by the Amber99SB force field [48], biotin [37] with the Amber99 force field [49] and c53, L3, and benzene with the generalized Amber force field [50]. Charges on the ligands were obtained by the restrained electrostatic potential-procedure [51], using electrostatic potentials calculated at the Hartree–Fock level and the 6-31G* basis set, and sampled according to the Merz–Kollman scheme [52], but using a higher than default density of points (10 concentric layers with 17 points/Å²). The protein–ligand systems were immersed in an octahedral box of TIP4P–Ewald water molecules [53], extending at least 8 Å from the protein.

Generation of independent simulations

Four different approaches to generate independent trajectories were considered

- VIIT (velocity-induced independent trajectories): The simulations were started with different starting velocities.
- SIIT (solvation-induced independent trajectories): The protein–ligand system was solvated in different water boxes.
- CRPIIT (conformation, rotation, and protonation-induced independent trajectories): Variations in the titration, conformations of Asn, Gln, and His, as well as in the rotation of bonds to polar hydrogen atoms were allowed.
- ACIIT (alternative-conformation-induced independent trajectories): The simulations were started from different conformations of disordered residues in the crystal structure.

The different start velocities were obtained by giving different seeds to the random-number generator that assigns velocities according to a Maxwell–Boltzmann distribution. The other three approaches are more involved and are described below.

Generation of different water boxes (SIIT). In Amber, the protein–ligand system is solvated by using a template box of water molecules, equilibrated with the same water model to be used in the simulations. This water box is extended in all dimensions and atoms that overlap with a protein or a ligand atom are removed (we used the default parameters to identify overlap, i.e. if the atoms are closer than the sum of their van der Waals radii). Therefore, we took the pre-equilibrated TIP4P–Ewald water box of 256 water molecules from the Amber 10 library and put it in a periodic system that was equilibrated for 10 ns in the NPT ensemble. Details of the simulations were identical to the protein–ligand simulations, described below. Thereafter, the system was simulated for 500 ns in the same ensemble. Coordinates were sampled every 12.5 ns and these 40 boxes were used to solvate the systems using the Amber 10 tleap module.

Titration, conformational and rotational sampling (CRPIIT). We utilized an in-house Monte Carlo (MC) program (TitProt) to titrate all His and Tyr residues in avidin, but only the His residues in the other proteins. Initial test runs showed that the Asp, Gly, Lys and Arg residues did not titrate. We also sampled two alternative conformations of the side chains of the Asn and Gln residues (180° rotation around the outermost C–C bond) as well as the His residues (180° rotation around the CD–CG bond). We also rotated the polar hydrogen atoms in Ser, Thr, Cys, Lys, and Tyr residues, as well as the protons in crystal-water molecules. Initially, the systems were protonated with the tleap module of Amber 10. The protein was put in a simulation sphere and counter ions were added to neutralize the system and to mimic experimental conditions, see Table 1. The system were equilibrated for 10⁶ MC steps at 300 K and pH 7, and then simulated for another 2·10⁶ (avidin) or

4·10⁶ steps. Snapshots were taken at even intervals from the last simulation and those were immersed in a truncated octahedral box of TIP4P–Ewald waters, using exactly the same water molecules in all systems (i.e. only one snapshots was solvated with tleap and then the same water molecules were used for all snapshots). Note that this approach will change the normal protonation of the His residues, described above.

TitProt uses a fixed protein structure (besides the variations mentioned above), hard spheres of all atoms with an radius of 2.0 Å, no explicit water molecules (besides those available in the crystal structures), atomic charges (the same as in the MD simulations), and a dielectric constant of 80 everywhere. Similar conditions (but with cruder charges) have been used successfully to study the titration of proteins [54], although the approach is questionable for titrable groups inside the protein. However, here we only used it to generate reasonable starting structures for MD simulations. The high dielectric constant ensures that we sample many different conformations, protonation states, and proton locations. The sampling frequencies were 0.77 for ionic moves and 0.08 for titrations, rotations, and conformational changes. Protein charges were taken from the Amber 1994 libraries both for the neutral and ionised residues [55], except for neutral Arg, for which we used charges determined in our group [39] and for negatively charged Tyr, for which the following charges were used (RESP charges, obtained in the same way as for the ligands): N –0.4157, H 0.2719, CA –0.0014, HA 0.0876, CB –0.0152, HB 0.0295, CG –0.0011, CD –0.1906, HD 0.1699, CE –0.2341, HE 0.1656, CZ 0.3226, OH –0.5579, HH 0.3992, C 0.5973, and O –0.5679.

Crystal conformations (ACIIT). This approach was only tested for Gal3, because this was the only crystal structure that contained alternative conformations. Five residues had two conformations with an occupancy of 0.5. All 32 permutations of these two conformations were used as initial structures, resulting in 32 simulations of Gal3 with this approach. All permutations were solvated with the same set of water molecules. In the other three approaches, the A conformation was used for all residues.

MD simulations

Molecular dynamics (MD) simulations were run by the Amber 10 sander module [56]. In all simulations, the SHAKE algorithm [57] was used to constrain bonds involving hydrogen atoms, making a time step of 2 fs possible. The temperature was kept constant at 300 K using a Langevin thermostat [58] with a collision frequency of 2.0 ps⁻¹. The pressure was kept constant at 1 atm using a weak-coupling isotropic algorithm [59] with a relaxation time of 1 ps. Long-range electrostatics were handled by particle-mesh Ewald summation [60], with a fourth-order B spline interpolation and a tolerance of 10⁻⁵. The non-bonded cut-off was 8 Å and the non-bonded pair list was updated every 50 fs.

The systems were energy minimized for 500 cycles of steepest descent, with all atoms, except water molecules and hydrogen atoms, restrained to their start position with a force constant of 418 kJ/mol/Å². This was followed by a 20 ps NPT simulation with the same restraints (this was the step in which different starting velocities were used in the VIIT approach) and a 100 ps (avidin and lysozyme) or 1000 ps (fXa and Gal3) unconstrained NPT equilibration. Finally, a 200 ps production run was initiated and coordinates were saved every 5 ps for the MM/GBSA calculations, according to the simulation protocol recently developed to obtain MM/GBSA estimates with an precision close to 1 kJ/mol [21].

MM/GBSA calculations

The binding free energy, ΔG_{bind} , of the ligands to their receptors was estimated with the MM/GBSA method [61,62]. The free energy of the ligand, the protein and the complex, was estimated as a sum of four terms

$$G = \langle E_{MM} \rangle + \langle G_{solv} \rangle + \langle G_{np} \rangle + T \langle S_{MM} \rangle \quad (1)$$

where E_{MM} is the MM energy of the molecule, i.e., a sum of internal energies, van der Waals interactions and electrostatic interactions. G_{solv} is the polar solvation energy, estimated by the generalized Born method of Onufriev et al., model I (GB^{OBCI}) [63], i.e. with $\alpha = 0.8$, $\beta = 0$, and $\gamma = 2.91$. G_{np} is the non-polar solvation energy, estimated from the solvent-accessible surface area (SASA), using the formula $G_{np} = 0.0227 \text{ SASA (in } \text{\AA}^2) + 3.85 \text{ kJ/mol}$ [35,62]. T is the absolute temperature and S_{MM} is an entropy estimate at the MM level. The entropy was estimated on a truncated and buffered system, as described previously, to improve the statistical precision of the estimate [25]. All the terms in Eqn. (1) were averages over 40 MD snapshots. The binding free energy is then calculated as

$$\Delta G_{bind} = G_{complex} - G_{protein} - G_{ligand} \quad (2)$$

The reported ΔG values are the average of ΔG over 20 (avidin) or 40 independent simulations (each with 40 snapshots) and the reported standard error is the standard deviation of this average divided by the square root of the number of independent simulations. To obtain stable energies, the same geometry was used for all three reactants (i.e. only the complex was simulated). The binding free energy was estimated for all four subunits of avidin, treating the other three ligands as a part of the protein (this is the reason why it is enough with 20 independent simulations for this protein – each simulation gives four estimates of the binding affinity). The reported binding free energy is the average over the four subunits. All MM/GBSA calculations were done with the Amber 10 software [56].

Results and Discussion

We calculated MM/GBSA estimates of the binding free energy, ΔG_{bind} , for four test systems, using three or four different methods to obtain independent simulations (depending on whether alternative conformations are present in the crystal structure or not). The results are collected in Table 2. We will discuss the results for the four test systems in turn.

For the biotin–avidin complex, the results in Table 2 show that there is little difference in ΔG_{bind} obtained from simulations started with different velocities (VIIT), with a different solvation (SIIT), or with different protonations and conformations (CRPIIT): All three approaches give a calculated affinity between -114 and -116 kJ/mol, i.e. with differences of only 1–2 kJ/mol. The standard errors of the calculated affinities (estimated from the 20 independent simulations) are 1 kJ/mol (note that we need to use a larger number of energy calculations, $4 \times 20 \times 40 = 3200$, than normally is used with MM/GBSA to obtain such a precision [21]). We performed a two-tailed, unpaired Student's t test (allowing for unequal variances) to test whether there is a significant difference between the simulations based on the VIIT approach and the other two approaches. The resulting p values in Table 3 show that there is no significant difference ($p > 0.42$). All three approaches also give similar estimated affinities for the four subunits of avidin, as can be seen from the results of the individual simulations, plotted in Figure 2.

To check that these small differences also remain with longer simulations, we extended the production simulation time from 0.2 ns to 2 ns. These results are also shown in Table 2 (column avidin/biotin long). It can be seen that the results for the VIIT and CRPIIT approaches hardly changed, whereas those with the SIIT approach changed slightly. However, the results of the three approaches are still the same within the statistical uncertainty, as can be seen from the p values in Table 3.

This is an important and promising result. It indicates that the binding affinity of biotin to avidin estimated with this method is very stable and certain. In particular, it shows that any scientist would get the same results (provided that the precision is good enough), irrespectively what program is used to set up the calculations. In fact, even uncertainties in the crystal structure regarding the location of the C, N, and O atoms in Asn, Gln, and His residues, and the problematic uncertainty in the protonation state of His residues do not have any influence on the calculated

values. This strongly increases the credibility of the results.

However, from Table 2, it can also be seen that the MM/GBSA estimate is quite far from the experimental ΔG , -85 kJ/mol [64]. This has been observed and discussed in several studies [22,39]. In particular, absolute binding affinities strongly depend on the method to used to calculate the polar solvation energies and other solvation methods than the generalised Born approach (or another variant of this approach) may shift the results by up to ~ 200 kJ/mol [65], but the relative solvation energies between similar ligands are typically stable within $3\text{--}5$ kJ/mol [66]. However, this is not of any concern in this study, because we are only interested in differences in the results between various approaches to set up the simulations. It is also notable that all approaches give a similar standard error, ~ 1 kJ/mol, indicating that all the results are converged and that none of the three approaches give a larger spread of the results.

However, for the other proteins, the results are somewhat different. For fXa, the VIIT and CRPIIT approaches give the same results (-61 to -63 kJ/mol; Table 2) within the statistical uncertainties, with a p value from the t test of 0.2 (Table 3). However, the SIIT approach gives a slightly less negative affinity, -57 kJ/mol. According to the t test, this difference is statistically significant with a confidence of 98% . A scatter plot of ΔG_{bind} values from all simulations (Figure 3) shows that there is a general weak trend of the SIIT simulations to give slightly less negative values. Solvation is probably more important for fXa than for avidin because the fXa active site is more solvent exposed than the buried binding site in avidin. The standard error of ΔG_{bind} is slightly lower with the VIIT approach than with the other two approaches, indicating that the VIIT approach does not sample the conformational space as well as the other two approaches. The calculated binding affinities are always ~ 40 kJ/mol more favourable than the experimental estimate [31].

The Gal3 system is even more challenging than the fXa system. For this system, the VIIT and SIIT approaches give similar results, -62 kJ/mol. This is interesting because the Gal3 binding site is fully exposed to the bulk water and it could be expected that the different water structures might result in significantly different solvation of the exposed ligand.

The 32 simulations started with different conformations observed in the crystal structure give somewhat less negative result (-58 kJ/mol), but the confidence of this difference is only 92% (Table 3). This is also interesting, because the selection of what alternative conformation to use in the MD simulations is arbitrary (because both conformations have the same occupancy). The present results show that this has a small (4 kJ/mol) influence on the final results. Previously, we have shown that MD-derived order parameters for Gal3 are somewhat improved if several independent simulations are used, starting from different conformations, compared to simulations started from a single conformation [18]. The standard error for ΔG_{bind} is also slightly larger with the ACIIT approach than with the VIIT and SIIT approaches, indicating that the conformational space is better sampled if the alternative conformations are employed.

The challenge with the Gal3 system becomes clear when we consider the CRPIIT approach, for which the estimated binding affinity is -54 kJ/mol, i.e. 8 kJ/mol more positive than for the VIIT approach. This difference is significant with practically 100% confidence. This extensive difference is somewhat surprising, considering the results for avidin and fXa. To understand the cause of this difference, we tested to exclude the titration of the His-158 residue, which makes a hydrogen bond with the ligand and we also excluded the conformational sampling of two Asn residues and the rotational sampling of one Ser residue, which all are within 5 Å of the ligand. This resulted in a ΔG_{bind} of -63 kJ/mol, 9 kJ/mol more negative than if these four residues were included and with no significant difference from the results obtained with the VIIT and SIIT approaches. This shows that one has to be careful when preparing the protein close to the binding site. The results of the three approaches are plotted in Figure 4.

Finally, for T4 lysozyme, all three approaches gave similar binding energies, -1 to -3 kJ/mol. However, considering the low standard error of these estimates (~ 0.5 kJ/mol), the differences are actually significant with $96\text{--}99\%$ confidence. Thus, our results again indicate that it

is advantageous to increase the conformational sampling by using the SIIT or CRPIIT approaches, which also give slightly larger standard errors. The results of the three approaches are shown in Figure 5. All three approaches underestimate the experimental binding free energy [67] by almost 20 kJ/mol. It is interesting that MM/GBSA gives a too positive estimate for this system, whereas it is too negative for the other three proteins. This may be related to the fact that the binding site in lysozyme is entirely hydrophobic, whereas the binding of the other three ligands involves several hydrogen bonds.

Conclusions

In this paper, we have studied two related questions, viz. how reproducible ligand-binding affinities predicted by the MM/GBSA method are, if set up by independent scientists using different software, and how conformational sampling can be enhanced by employing the rather arbitrary choices made during the set-up of MD simulations of proteins. As test cases, we have used the binding of ligands to four different proteins, avidin, factor Xa, galectin 3, and T4 lysozyme.

For the first question, we obtain rather encouraging results: All calculated binding affinities agree within 10 kJ/mol, and all except three estimates (CRPIIT and ACIIT for Gal3 and SIIT for fXa) actually agree within 2 kJ/mol. This must be considered quite impressive, considering the quite large changes made to the systems, including the total charge of the protein, nearly free rotation of crystal waters and polar groups, flips of Asn, Gln, and His residues, and variation in the protonation of His residues. These variations should include, and probably surpass, possible differences encountered if the systems were set up by different scientists using different software. The 2-kJ/mol difference, is close to the statistical precision of the approach (0.4–1.5 kJ/mol). Moreover, the 10 kJ/mol difference is well within the accuracy of the MM/GBSA approach, even if only the same protein and similar ligands are considered; for example, we obtained a mean absolute deviation of 15 kJ/mol for the binding of seven biotin analogues to avidin, using the same simulation protocol [21]. Thus, we can conclude that even if the set-up of MD simulations of proteins involve a large number of hard choices, the results are reasonable reproducible, considering the current precision and accuracy, at least for binding affinities predicted by MM/GBSA. However, great care should be taken in the set-up of the calculations close to the active site, as will be more discussed below.

Regarding the second question, our results show that in two of the four test cases, significantly different results are obtained if the protein is solvated in different water boxes than if exactly the same structure is started with different velocities. In three of the four cases, the standard errors of the estimated ligand affinities are also larger with the SIIT approach. Therefore, it seems that the VIIT approach somewhat undersamples the phase space and therefore also overestimates the precision of the calculated binding affinities. Considering that the SIIT approach is almost as easy to implement as the VIIT approach, and neither of them induce any increase in the computational load, we see no reason not to use it as a means to increase the conformational sampling with the independent-trajectories approach. Of course, the two approaches should be combined by also using different random number seeds for the generation of the starting velocities.

Likewise, we see no reason not to use conformational disorder in the crystal structure, if available, to enhance the conformational sampling. This is most conveniently obtained by clustering the residues with alternative conformations according to distance (so that conformations that depend on each other go into the same group) and then simply select conformations by random (possibly weighted by the occupancies, if different). In this investigation, the results obtained with the ACIIT approach gave only barely significantly different results from the VIIT approach (but a 97% significance compared to the SIIT approach). However, in a previous study, we obtained significantly improved order parameters with this approach [18].

For the CRPIIT approach, the interpretation is more troublesome: Whereas, the starting velocities and the solvation is in principle completely arbitrary, this is not the case for ionisations,

protonations, and conformations. On the contrary, in most cases there are probably correct and erroneous choices, i.e. most residues reside in one titration state or one conformation the great majority of the time, or at least the various choices do not have equal probability. The problem is only that we do not know which choice is correct, and we do not want to spend the time necessary to needed to obtain a more certain answer, owing to the large number of possibilities and residues with uncertain choices. The CRPIIT approach, as implemented by the TitProt program may give us a first indication of how important these choices are for the final results. It gives a simple and rather fast estimate of the involved energies (the calculations in this paper were run over a night and it is possible that even shorter simulations could be used), and this approach has been shown to give good results for titration curves of solvent-exposed residues [54]. In fact, a MC approach seems to well suited for the problem, because it gives completely random sampling if no information is available about the relative energies of the various states, whereas it will correctly bias the sampling to the more probable states if the energies differ.

However, we want to emphasize that the results of the current implementation of TitProt are highly approximate and that we have selected this simple approach mainly to obtain an extensive sampling, rather than on the basis on accuracy. In fact, we have deliberately used a high dielectric constant also for buried interactions to enhance the sampling of different conformations and rotations. The approach could easily be adapted for more accurate energy calculations, although at a higher expense. In particular, our results for Gal3 show that for residues close to the active site, the choice of ionization, protonation, conformation, and rotation is especially important. The ideal approach is probably to use more accurate and detailed methods to decide the protonation and conformations close to the site of interest, whereas the CRPIIT approach can be used more distant residues. The same probably applies to the solvation of water-exposed active sites, like that in fXa. Then, enhanced sampling of different solvation states should be tested close to the active site.

The results in Figures 2–5 clearly illustrate the importance of sampling for MM/GBSA: Even if the results of each simulation is based on an average over 40 snapshots, they may differ by up to 60 kJ/mol. Therefore, stable and reliable results can only be obtained by averaging over a significant number of independent simulations [21].

Finally, we want to point out that the present results were obtained for the calculation of ligand-binding affinities with the MM/GBSA approach. However, the results are most likely directly applicable also to studies of other properties, although the differences and the accuracy may be different (for example, the use of the continuum solvation model in the MM/GBSA energies probably makes this method less sensitive to the solvent structure than free-energy perturbations with explicit water). Therefore, we strongly encourage the use of the SIIT and ACIIT approaches to enhance the sampling in the simulations and to increase the reproducibility of the results. To simplify such calculations, we provide on our home page the 40 equilibrated TIP4P-Ewald water boxes (<http://www.teokem.lu.se/~ulf/waterboxes.html>). The TitProt program is also available for free upon request. For residues not inside the site of interest, we also recommend the use of different protonation states, conformations Asn, Gln, and His residues, and location of the polar hydrogen atoms, although based on as sophisticated guesses as can be afforded.

Acknowledgements

This investigation has been supported by grants from the Swedish research council and from the Research school in pharmaceutical science. It has also been supported by computer resources of Lunarc at Lund University and HPC2N at Umeå University.

References

- 1 Forrest, L. R.; Honig, B. *Proteins: Struct, Funct, Bioinf* 2005, 61, 296-309
- 2 Ullman, GM.; Knapp, E.-W. *Eur Biophys J* 1999, 28, 533-551
- 3 Schutz, C. N.; Warshel, A. *Prot Struct Funct Gen* 2201, 44, 400-417

- 4 Hooft, R. W.; Vriend, G.; Sander, C. Abola, E. E. *Nature* 1996, 381, 272.
- 5 Word J.M.; Lovell S. C.; Richardson J. S.; Richardson D. C. *J Mol Biol* 1999, 285, 1735–1747
- 6 Senn, H. M.; Thiel, W. *Top Curr Chem* 2007, 268, 173-290.
- 7 Vriend, G. *J Mol Graph* 1990, 8, 52–56
- 8 Li H.; Robertson A.D.; Jensen, J.H. *Proteins Struct Funct Bioinf* 2005, 61, 704–721
- 9 Gordon, J. C.; Myers, J. B.; Folta, T.; Shoja V., Heath, L. S.; Onufriev, A. *Nucleic Acids Res*, 2005, 33, W368-371
- 10 Zagrovic, B.; van Gunsteren, W. F. *J Chem Theory Comput* 2007, 3, 301-311
- 11 Straub, J. E.; Rashkin, A. B.; Thirumalai, D. *J Am Chem Soc* 1994, 116, 2049-2063
- 12 Adcock, S. A.; McCammon, J. A. *Chem Rev* 2006, 106, 1589-1615
- 13 Hamelberg, D.; Mongan, J.; McCammon, J. A. *J Chem Phys*, 2004, 120, 11919-11929
- 14 Wu, X.; Wang, S. M. *J Phys Chem B*, 1992, 102, 7238-7250
- 15 Simmering, C.; Fox, T.; Kollman, P. A. *J Am Chem Soc* 1998, 120, 5771–5782
- 16 Woods, C. J.; Essex, J. W.; King, M. A. *J Phys Chem B*, 2003, 107, 13703-13710
- 17 Loccisano, A. E.; Acevedo, O.; DeChancie, J.; Schulze, B. G.; Evanseck, J. D. *J Mol Graph Model* 2004, 22, 369-376
- 18 Genheden, S.; Diehl, C.; Akke, M.; Ryde, U. 2010, *J Chem Theory Comput*, submitted.
- 19 Lawrenz, M.; Baron, R.; McCammon, J. A. *J Chem Theory Comput* 2009, 5, 1106–1116
- 20 Fujitani, H.; Tanida, Y.; Ito, M.; Jayachandran, G.; Snow, C. D.; Shirts, M. R.; Sorin, E. J.; Pande, V. S. *J Chem Phys* 2005, 123, 0804108-5
- 21 Genheden, S.; Ryde, U. *J Comput Chem* 2010, 31, 837-846
- 22 Olsson, M. H. M.; Hong, G.; Warshel, A. *J Am Chem Soc* 2003, 125, 5025-5039
- 23 Elofsson, A.; Nilsson, L. *J Mol Biol* 1993, 233, 766-780
- 24 Caves, L. S. D.; Evanseck, J. D.; Karplus, M. *Prot Sci*, 1998, 7, 649-666
- 25 Worth, G. A.; Nardi, F.; Wade, R. C. *J Phys Chem B* 1998, 102, 6260-6272
- 26 Gorfe, A. A.; Ferrara, P.; Cafilisch, A.; Marti, D. N.; Bosshard, H. R.; Jelesarov, I. *Proteins: Struct, Funct, Genet* 2002, 46, 41-60
- 27 Sherer, E. C.; Cramer, C. J. *J Phys Chem B* 2002, 106, 5075-5085
- 28 Koller, A. N.; Schwalbe, H.; Gohlke, H. *Biophys J* 2008, 95, L04-L06
- 29 Srinivasan, J.; Cheatham III, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J Am Chem Soc* 1998, 37, 9401-9809
- 30 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
- 31 Matter, H.; Defossa, E.; Heinelt, U.; Blohm, P.-M.; Schneider, D.; Müller, A.; Herok, S.; Schreuder, H.; Liesum, A.; Brachvogel, V.; Lönze, P.; Walser, A.; Al-Obeidi, F.; Wildgoose, P. *J Med Chem* 2002, 45, 2749-2769
- 32 Sörme, P.; Arnoux, P.; Kahl-Knutsson, B.; Leffler, H.; Rini, J. M.; Nilsson, U. J. *J Am Chem Soc* 2005, 127, 1737-1743
- 33 Miyamoto, S.; Kollman, P. A. , *Proteins, Struct, Funct, Genet* 1993, 16, 226-245
- 34 Wang, J.; Dixon, R.; Kollman, P. A., *Proteins, Struct, Funct, Genet*, 1999, 34, 69-81
- 35 Kuhn, B.; Kollman, P. A. *J Med Chem* 2000, 43, 3786-3791
- 36 Brown, S. P.; Muchmore, S. W., *J Chem Inf Model* 2006, 46, 999-1005
- 37 Weis, A.; Katebzadeh, K.; Söderhjelm, O.; Nilsson, I.J Ryde, U. *J Med Chem* 2006, 49, 6595-6606
- 38 Kongsted, J.; Ryde, U., *J Comp-Aided Mol Des*, 2009, 23, 63-71
- 39 DeChancie, J.; Houk, K. N. *J Am Chem Soc* 2007, 129, 5419-5429
- 40 Söderhjelm, P.; Kongsted, J.; Ryde, U. *J Chem Theory Comput*, submitted
- 41 Hermans, J.; Wang, L. *J Am Chem Soc* 1997, 119, 2707–2714
- 42 Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *J Phys Chem, B*, 2003, 107, 9535–9551
- 43 Mann, G.; Hermans, J. *Mol Biol* 2000, 302, 979–989
- 44 Deng, U.; Roux, B. *J Chem Theory Comput* 2006, 2, 1255-1273
- 45 Mobley, D. L.; Graves, A. P. Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A., *J Mol Biol* 2007, 371, 1118-1134
- 46 Pugliese, L.; Coda, A.; Malcovati, M.; Bolognesi, M. *J Mol Biol* 1993, 231, 698-710
- 47 Morton, A.; Matthews, B. W. *Biochem* 1995, 34, 8576-8588
- 48 Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. *Proteins: Struct, Funct, Bioinform* 2006, 65, 712-725
- 49 Wang, J.; Cieplak, P.; Kollman, P. A. *J Comput Chem* 2000, 21, 1074-1074
- 50 Wang, J. M.; Wolf, R. M.; Caldwell, K. W.; Kollman, P. A.; Case, D. A. *J Comput Chem* 2004, 25, 1157-1174
- 51 Bayly, C. I.; Cieplak, P.; Cornell, W. D.; Kollman, P. A. *J Phys Chem* 1993, 97, 10269-10280
- 52 Besler, B. H.; Merz, K. M.; Kollman, P. A. *J Comput Chem* 1990, 11, 431-439
- 53 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
- 54 Lund, M.; Jönsson, B. *Biochem* 2005, 44, 5722-5727

-
- 55 Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. *J Am Chem Soc* 1995, 117, 5179-5197
- 56 Case, D. A.; Darden, T. A.; Cheatham III, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Crowley, M.; Walker, R.; C.; Zhang, W.; Merz, K. M.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossvary, I.; Wong, K.; F.; Paesani, F.; Vanicek, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Yang, L.; Tan, C.; Mongan, J.; Hornak, V.; Cui, G.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Kollman, P. A. *Amber 10*, University of California, San Francisco, 2008
- 57 Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. *J Comput Phys*, 1977, 23, 327-341
- 58 Wu, X.; Brooks, B.R. *Chem Phys Lett* 2003, 381, 512-518
- 59 Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J Chem Phys*, 1984, 81, 3684-3690
- 60 Darden, T.; York, D.; Pedersen, L. *J Chem Phys* 1993, 98, 10089-10092
- 61 Srinivasan, J.; Cheatham III, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. *J Am Chem Soc* 1998, 120, 9401-9809
- 62 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
- 63 Onufriev, A.; Bashford, D.; Case, D. A. *Proteins* 2004, 55, 383-394
- 64 Green, N. M. *Adv Protein Chem* 1975, 29, 85-133
- 65 Genheden, S.; Lychko, T.; Gusarov, S.; Kovalenko, S.; Ryde, U. 2010, *J. Phys. Chem. B.*, submitted
- 66 Kongsted, J.; Söderhjelm, P.; Ryde U. *J Comp-Aided Mol Design*, 2009, 23, 395-409
- 67 Morton, A.; Baase, W. A.; Matthews, B. W. *Biochem*, 1995, 34, 8564-8575

Table 1. Details of the Monte Carlo simulations with TitProt.

System	Sphere size (Å)	# negative ions	# positive ions	[protein] (mM)	[ions] (mM)
avidin	203	269	251	4	100
fXa	200	65	60	4	25
Gal3	200	30	26	10	400
lysozyme	1500	2600	2597	0.001	25

Table 2. MM/GBSA estimates of ΔG_{bind} for various approaches to generate independent simulations in kJ/mol. Standard errors are given in parenthesis.

Approach	avidin/biotin	avidin/biotin (long ^b)	fXa/c53	Gal3/L3	lysozyme/benzene
VIIT	-114.2 (1.1)	-114.8 (1.0)	-61.2 (1.1)	-61.5 (0.9)	-3.1 (0.4)
SIIT	-113.7 (0.8)	-115.6 (0.8)	-56.9 (1.5)	-62.4 (1.0)	-1.8 (0.5)
CRPIIT	-115.5 (1.1)	-115.4 (0.9)	-63.4 (1.4)	-53.6 (1.1)	-1.1 (0.6)
CRPIIT ^a				-63.3 (1.2)	
ACIIT				-58.4 (1.5)	
Experimental	-85.4 [64]	-85.4 [64]	-19.3 [31]	-34.8 [32]	-21.7 [67]

^a No titration and no sampling of residues within 5 Å of L3.

^b These simulations had a production simulation time of 2 ns, whereas it was 0.2 ns for the other simulations.

Table 3. Calculated p values from a two-tailed, unpaired Student's t test, assuming unequal variances, for ΔG_{bind} , obtained with the various approaches compared to the VIIT approach.

Approach	avidin/biotin	avidin/biotin (long ^b)	fXa/c53	Gal3/L3	lysozyme/benzene
SIIT	0.69	0.58	0.02	0.49	0.04
CRPIIT	0.42	0.69	0.20	0.00	0.01
CRPIIT ^a				0.24	
ACIIT				0.08	

^a No titration and no sampling of residues within 5 Å of L3.

^b These simulations had a production simulation time of 2 ns, whereas it was 0.2 ns for the other simulations.

Figure 1. Ligands considered in this study. A) biotin, B) c53, C) L3 and D) benzene.

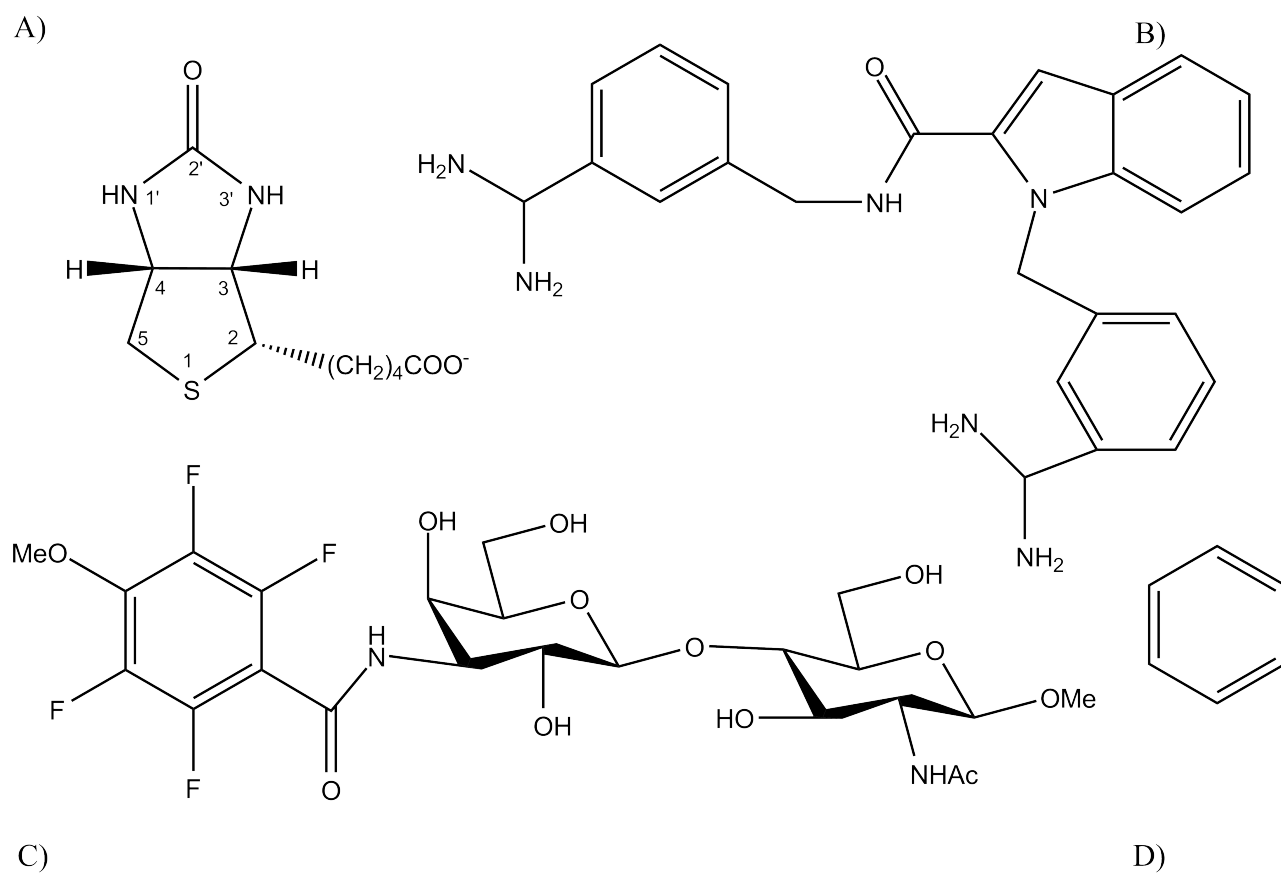


Figure 2. Individual binding affinities in the three sets of simulations for avidin/biotin. The first 20 simulations are for subunit A, the next 20 for subunit B, and so on.

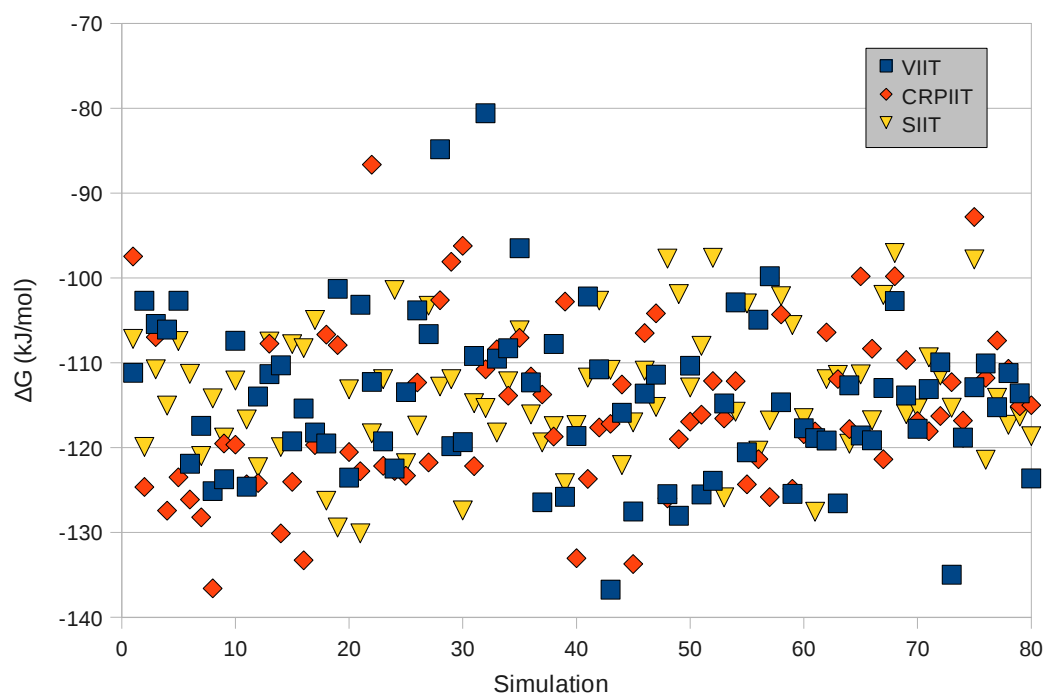


Figure 3. Individual binding affinities in the three sets of simulations for fXa/c53.

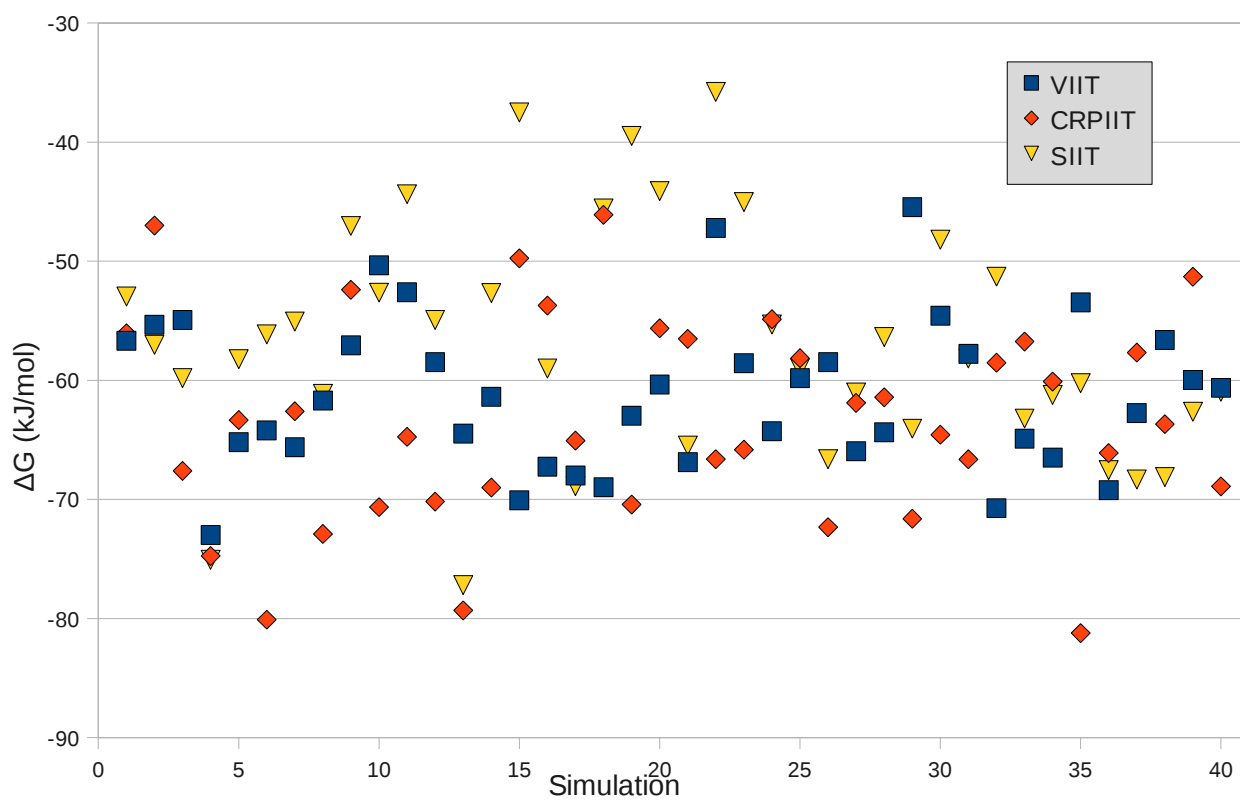


Figure 4. Individual binding affinities in the five sets of simulations for gal3/L3.

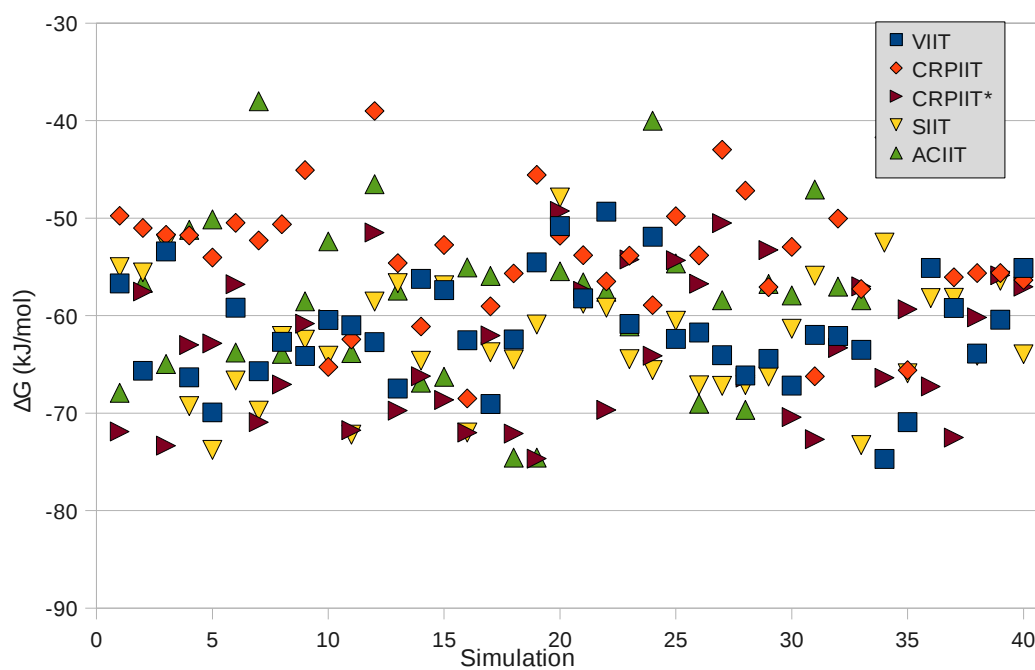


Figure 5. Individual binding affinities in the three sets of simulations for lysozyme/benzene.

