Optimized potential of mean force calculations of standard binding free energy

Ignasi Buch,[†,‡] S. Kashif Sadiq,[†,‡] and Gianni De Fabritiis[*,†]

Computational Biochemistry and Biophysics Laboratory (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), C/ Doctor Aiguader 88, 08003 Barcelona, Spain

E-mail: gianni.defabritiis@upf.edu

Abstract

The prediction of protein-ligand binding free energies is an important goal of computational biochemistry, yet accuracy, reproducibility and cost remain a problem. Nevertheless, these are essential requirements for computational methods to become standard binding prediction tools in discovery pipelines. Here we present the results of an extensive search for an optimal method based on an ensemble of umbrella sampling all-atom molecular simulations tested on the phosphorylated tetrapeptide, pYEEI, binding to the SH2 domain, resulting in an accurate and converged binding free energy of -9.0 ± 0.5 kcal/mol (compared to experimental value of -8.0 ± 0.1 kcal/mol). We find that a minimum of 300 ns of sampling is required for every prediction, a target easily achievable using new generation accelerated MD codes. Convergence is obtained by using an ensemble of simulations per window each starting from different initial conformations and by optimizing window-width, orthogonal restraints, reaction coordinate harmonic potentials and window-sample time. The use of uncorrelated initial conformations in neighboring windows is important for correctly sampling conformational transitions from the unbound to bound states that affect significantly the precision of the calculations. This methodology thus provides a general recipe for reproducible and practical computations of binding free energies for a class of semi-rigid protein-ligand systems, within the limit of the accuracy of the forcefield used.

*To whom correspondence should be addressed

†Computational Biochemistry and Biophysics Laboratory (GRIB-IMIM), Universitat Pompeu Fabra, Barcelona Biomedical Research Park (PRBB), C/ Doctor Aiguader 88, 08003 Barcelona, Spain

‡These authors contributed equally to this work

1 Introduction

Achieving a standard, reliable, and accurate protocol for the quantitative determination of protein-ligand binding affinities has remained one of the pivotal problems in computational biochemistry; its attainment is set to yield a tremendous gain in the basic understanding of molecular biological processes. Attempts to compute binding affinities have been made since near the inception of computational biomolecular modeling and several notable methods involving molecular dynamics (MD) simulations have arisen.[1,2] The underpinning problem circumvented by all of these methods is that unbiased equilibrium-based free ligand binding using an all-atom model (including solvent) is computationally possible in certain cases although much more expensive than the present calculations. Another route is therefore employed in arriving at a quantitative determination of the binding free energy.

At the high-throughput end, empirically tuned methods such as linear interaction energy (LIE) methods[3–5] are used with the forfeit of compromising some accuracy. One major strategy is to use implicit solvent MD,[6] which drastically re-

duces the computational cost, sometimes at the expense of neglecting crucial structural water mediated interactions.[7] Such continuum solvent methods are often used in conjunction with thermodynamic cycle methods, such as the molecular mechanics Poisson-Boltzmann/Generalized-Born solvent accessible surface area (MMPB/GBSA) methods,[2,8–13] that indirectly compute the binding free energy in solution by separation of the solvation and *in vacuo* interaction components of the free energy. Other more accurate and computationally intensive methods involving "alchemical" mutations, such as free energy perturbation (FEP)[14–16] and thermodynamic integration (TI)[17–19] have been traditionally employed for– but not limited to–calculating relative binding free energies between related protein-ligand combinations, being able to calculate absolute binding free energies.[20] The latter however, having a much larger computational cost.

Methods involving the biased sampling along a set of pre-selected reaction coordinates that follow physically meaningful binding pathways have also found a measure of success. These include metadynamics,[21,22] adaptive force bias,[23] the Jarzynski method[24,25] and umbrella sampling[26,27] amongst others. For example, metadynamics approaches have been used to determine peptidic binding of highly flexible target proteins,[28] whilst biased umbrella sampling methods have shown that accurate binding free energies are indeed possible once conformational (rotational and translational) restraints are properly sampled.[29–33] The overriding problem with such methods is that they require extensive knowledge of the specific system, in order to apply the relevant biases; they are thus costly in human resource, requiring informed and manual selection of appropriate restraints in the configurational space and are thus not scalable in a standard way to the high-throughput domain. Recently an unbiased umbrella sampling method was reported using only a one-dimensional potential of mean force (PMF) calculation[34] and the weighted histogram analysis method (WHAM).[35] Although only modestly precise when applied to the benzamidine-trypsin system, the method does away with conformational biasing and applies only generic restraints, orthogonal to the direction of binding. Furthermore the ease of implementation of this method makes further evaluation of it an attractive prospect for being an optimal method for high-throughput binding free energy determination, provided that the fundamental problem of sufficient sampling can be overcome.

The current age of micro- to millisecond MD brings with it the ability to test the hypothesis that current MD forcefields are accurate enough to reproducibly attain accurate binding free energies, given enough sampling. Aggregate sampling across such timescales has been implemented by several groups,[36–41] primarily with respect to conformational dynamics and protein folding and lends itself naturally to distributed computing initiatives.[42] Furthermore, the recent advances in programmable GPU technology[43–45] have facilitated several initiatives, like ACEMD,[44] a new generation fast MD code exclusively running on GPUs and GPUGRID, a distributed computing project[46] for molecular dynamics simulations. Using this resource, we have previously shown that extensive sampling (over 19 μs of aggregate sampling) using the 1D-PMF method for a larger-ligand system results in accurate binding free energies compared with experiment,[46] while with the optimization reported here only 300 ns are necessary.

In this paper, we investigate whether such a method can be made robust, convergent and reproducible, whilst optimizing the protocol to minimize the amount of required computational cost and crucially retaining the accuracy of the result. To allow optimal comparison to other methods[29] and our earlier investigations,[47] the method is applied to the Src homology 2 (SH2) domain binding to the phosphorylated tetrapeptide pYEEI. SH2 domains are non-catalytic domains[48] composed of approximately 100 amino acids,[49] involved in a large variety of tyrosine-kinase signal transduction pathways[50–52] and bind short peptidic sequences containing phosphorylated tyrosine residues.[53,54] Furthermore, many pathological conditions, such as autoimmune diseases, cancer and asthma, can be associated with the incorrect function of SH2-mediated processes, making them an attractive target for structure-based drug design.[55–57] This ubiquitous role in cell function and regulation[48,50] imposes conditions of high affinity and specificity for a range of peptides,[58–63] making them an ex-

cellent template to differentiate various computational methodologies,[29,47,64]

Here, we first re-implement our extensive 1D-PMF sampling protocol used previously,[46] analyzing its convergent properties. Secondly, we adapt the 1D-PMF protocol through a sequence of optimizations. These include window-width and thus corresponding harmonic restraint variation, the use of ensemble trajectory sampling, which has been shown to be advantageous over single trajectory sampling in other methods[13] and the use of multiple initial conditions. At each stage, the computational cost is reduced or the corresponding accuracy and convergence increased. As the sampling required to achieve convergence is related to the conformational freedom and thus the size of the system, the protocol that emerges from this optimization is capable of producing accurate and reproducible binding free energies up to the given size of system implemented here. This result would allow a vast array of ligand-protein binding free energies up to the given molecular weight to be accurately and rapidly determined through high throughput molecular simulation.

2 Materials and Methods

2.1 System preparation

The input model is based on the bound crystallographic structure of the complex of the human $p56^{lck}$ domain and the peptide phosphotyrosine-Glu-Glu-Ile (pYEEI) (PDB:1LKK) using the CHARMM27[65] forcefield. The phosphotyrosine residue was assumed to be in its charged form Y-$PO_3{}^{2-}$ as experimentally determined.[60] Neutral acetylated N-terminus (ACE) and amidated C-terminus (CT2) residues were used to cap the peptide. The complex was solvated in a TIP3P[66] water box with a boundary at least 12 Å from the system in the x and y directions and of 52 Å in the z-direction giving a box-size of $65 \times 62 \times 93$ Å^3, the z axis being larger to allow for the generation of several US initial configurations with the ligand at different distances from the protein (Figure 1a).

The ionic strength was set to 0.15 M of Na^+ and Cl^- and the system charge neutralized. The final system comprised 38,655 atoms. The reaction

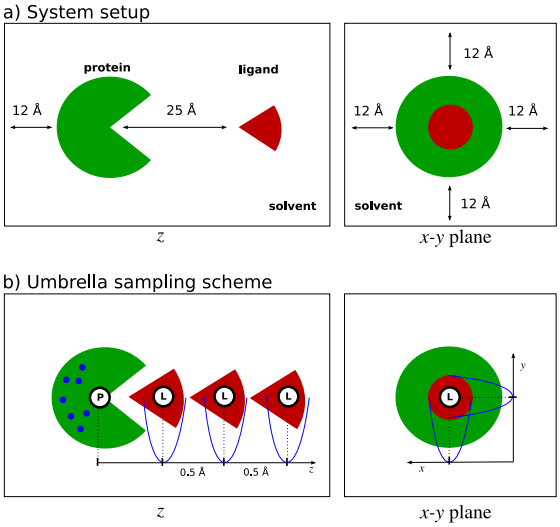

Figure 1: (a) Schematic representation of a system for calculation of free-energy of binding. "P" and "L" are for protein and ligand, respectively. (b) Schematic visualization of the initial configurations for the US of the SH2 domain/pYEEI ligand complex (PDB:1LKK) in the water box.

coordinate z is set to be orthogonal to the plane formed by the binding interface of the complex; the protein was then rotated manually during system preparation with the aim of providing a large water reservoir in the direction of the ligand displacement.

The system was minimized and relaxed under NPT conditions at 1 atm and 298K using a timestep of 2 fs, cutoff of 9 Å, rigid bonds and PME for long range electrostatic with a grid of $64 \times 64 \times 96$. During minimization and equilibration, the heavy protein atoms were restrained by a 10 kcal/mol/Å^2 spring constant. Two rounds of velocity re-initialization for 2 ps were performed under NVT conditions. The magnitude of the restraining spring constant was then reduced to 1 kcal/mol/Å^2 during 10 ps of NVT before the barostat was switched on at 1 atm for a further 10 ps of NPT simulation. A final 40 ps of NPT simulation was conducted with a restraint constant of 0.05 kcal/mol/Å^2. Finally, the volume was allowed to relax for 10 ns under NPT conditions. During this run, only Cα atoms of the complex were restrained with a 1 kcal/mol/Å^2 constant in order to prevent reorientation.[46]

Production simulations were run using ACEMD[44] over GPUGRID.net[46] with the same parameters used for the relaxation but a timestep of 4 fs due to the use of the hydrogen mass repartition scheme[67] implemented in ACEMD. This elegant method[67] uses the mathematical property that individual atom masses do not appear explicitly in the equilibrium distribution, therefore changing them only affects the transport properties of the system marginally but not the equilibrium distribution.[44]

2.2 Initial conformation generation

The umbrella sampling (US) method requires prior generation of initial conformations for each window of the production sampling. Window-centered initial conformations corresponding to the entire range of the reaction coordinate were generated via preliminary MD simulations in which the ligand was displaced by 25 Å along the z-direction towards the bulk from $z = 0$ Å to $z = 25$ Å by applying a linear force $F = -k_d(z - vt)$ to all of its carbon atoms, where $k_d = 10$ kcal/mol/Å^2 and $v = 5$ Å/ns. A second biasing restraint of $k = 0.1$ kcal/mol/Å^2 was applied to the center of mass of the ligand to restrain to the xy plane (with respect to the initial bound position of the ligand). A harmonic restraint of $k = 1$ kcal/mol/Å^2 was applied to every Cα atom residing in an α-helix or β-sheet of the protein further than 9 Å from the ligand. This prevented rotation and translation of the protein during ligand separation while preserving the flexibility of the binding pocket. Snapshots of the system coordinates (Figure 1b) were saved at constant intervals. Two sets of initial conformations were generated using this method. The first set (denoted IC1 hereafter) employed a single preliminary MD run generating a single initial conformation for each window from that run. The second set (denoted IC2 hereafter) employed a total of 10 preliminary MD simulations, thus generating 10 initial conformations for each window. Initial conformations were then chosen by window-sequential selection across the set of 10 preliminary runs thus ensuring that neighboring initial conformations were from different runs. Initial conformations derived from the same preliminary run thus had a 10 window spacing within the set.

2.3 Umbrella sampling optimization

A number of umbrella sampling (US) simulations were performed, each varying a protocol parameter, namely, window width (Δw) and correspondingly the number of windows, sample time per independent simulation per window (t), orthogonal restraints k_{xy}, force constant for the harmonic window potential k_z, the ensemble size or number of independent simulations (N_r) and finally whether the initial starting conformation set was IC1 or IC2 (IC). The reaction coordinate always extended from $z = 0$ Å to $z = 25$ Å with the bound configuration at position $z = 0$ Å used as reference. The parameter sets for the full range of simulations performed here together with the total corresponding sampling time (t_{tot}) are listed in Table 1. All initial US windows were submitted to GPUGRID.net for execution of the US protocol. Each US window simulation was divided into several successive steps, with each step being 4 ns of duration. Each step was run as a separate GPUGRID work unit (WU), where each WU corresponded to about 6 hours of continued computation for a typical GPUGRID volunteer computer, while ACEMD on a top GPU like a GTX480 would perform 50 ns/day for this system. Preliminary runs to generate initial conformations were performed locally. The rationale for the different simulations is explained below.

Set 1 corresponded to the implementation of a previous exhaustive sampling simulation, reported previously,[46] using a small window width 0.1 Å (381 windows) and up to a sampling time of 50 ns per window. Initial conformations were generated from a single preliminary MD run (IC1).

Set 2 corresponded to the optimization procedure for the force constants for the harmonic potentials both for restraining diffusion in the xy plane (k_{xy}) and for the US potential (k_z). It employed a set of 3 US simulations each of up to 80 ns/window for a combination of 10 different permutations of k_{xy} and k_z listed in Table 1 and using a larger window width of 0.5 Å (51 windows). The optimal parameter set (OPS) was determined as k_{xy} = 1 kcal/mol/Å^2 and k_z = 0.5 kcal/mol/Å^2 (Figure 3). The initial conformation set was IC2.

Set 3 corresponded to the ensemble sampling procedure using the OPS. This entailed an ensemble

Table 1: Umbrella sampling simulation parameter variation. N_r, number of complete US replicas; N_w, number of windows per US replica; Δw, US window width; t, simulation time per US window; k_{xy}, force constant for orthogonal restraints; k_z, force constant for US restraints; *IC*, source of initial conformations; t_{tot}, total aggregate simulation time.

ID	N_r	N_w	Δw (Å)	t (ns)	k_{xy} ($\frac{kcal}{mol^2}$)	k_z ($\frac{kcal}{mol^2}$)	*IC*	t_{tot} (μs)
Set 1[46]	1	381	0.1	50	0.1	10	IC1	19
Set 2	10× 3	51	0.5	80	0.1,1	0.5,1,2.5,5,10	IC2	122.4
Set 3	10	51	0.5	20	1	0.5	IC2	10.2
Set 4	10	51	0.5	20	1	0.5	IC1	10.2

of ten identical US simulations for which the PMF and subsequent binding free energy was calculated in order to determine the convergence properties of the method. Initial conformations were generated as for Set 2 (IC2).

Set 4 corresponded to the study of the effect of using a less varied initial conformation set across neighboring windows. A set of ten identical US simulations were performed, similar to Set 3 but using initial conformations generated in a more simple manner, from a single preliminary MD simulation (IC1).

2.4 Free-energy calculation

The PMF over the reaction coordinate for each replica was reconstructed using WHAM[35] with a convergence tolerance of 10^{-4}. From the PMF, the standard free energy of binding was computed using the expression given in:[34]

$$\Delta G^\circ = \Delta W_R - k_B T \ln(\frac{l_b A_{u,R}}{V^\circ}) + \Delta G_R, \quad (1)$$

where ΔW_R is the PMF depth, k_B is the Boltzmann constant, T is the temperature, $l_b = \int_{\text{bound}} \exp(-W_R(z)/k_B T)\, dz$ is the integral of the PMF over the bound length, $A_{u,R} = 2\pi k_B T / k_{xy}$ is the area in the x and y directions of the unbound ligand, $V^\circ = 1{,}661$ Å^3 is the standard volume, and ΔG_R is the free energy to remove the orthogonal restraints (on x and y) when the ligand is bound. ΔG_R is obtained via a free energy perturbation approach from the exponential average.[34]

In order to assess convergence the sampling time per window used to construct the PMF and thus the free energy for each replica was increased up to the maximum sampling time across the 51 windows. The convergence of each replica with increase in time was then charted as well as the corresponding mean and standard deviation.

3 Results and Discussion

We begin by analyzing the exhaustive umbrella sampling (US) method reported previously,[46] called Set 1 here. Optimization of the method requires us to obtain convergent and accurate results with the minimum sampling time and proceeds via alteration of window width and a corresponding systematic parameter search with respect to harmonic restraint values (Set 2). Convergence is investigated by employing an ensemble of simulations and analyzing across an increasing sampling time per window (Set 3) and compared against a similar ensemble with the choice of using less varied initial conformations (Set 4).

3.1 Exhaustive umbrella sampling

An exhaustive umbrella sampling method utilizing 19 μs aggregate sampling was used to determine the free energy of binding of SH2 to pYEEI (Set 1[46] in Table 1, 1 run). The PMF depth (see Figure 2), is $\Delta W_R = -10.8$ kcal/mol, with a small variation for different times in the US runs, the bound distance is $l_b = 0.93$Å and the area explored by the ligand in the xy plane $A_{u,R} = 37.07$Å^2. The free energy to remove the constraints have a negligible contribution $\Delta G_R = -0.0124$ kcal/mol due to the low restrain applied. The standard free energy of binding for the pYEEI ligand is computed from Eq. (1) as $\Delta G^\circ = -8.5$ kcal/mol which compares with a reported experimental value of -8.0 kcal/mol.[61]

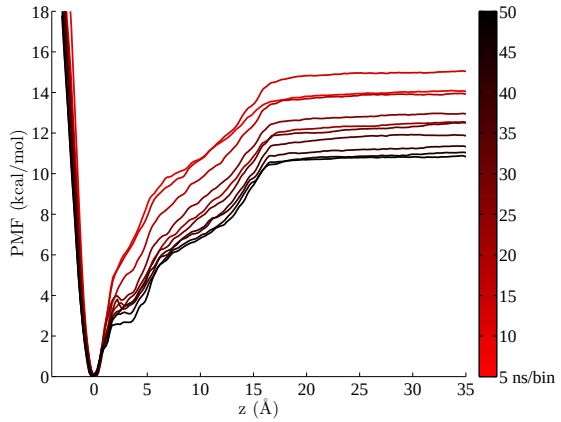

Figure 2: Reconstructed potential of mean force of the SH2 domain/pYEEI ligand complex along the reaction coordinate, calculated from 381 completed US configurations of 50 ns each. The PMF is reconstructed over increasing sample time windows along the US trajectories showing the long relaxation time of the US simulations. The reference ΔW_R value (PMF depth) computed from the last PMF is 10.8 kcal/mol producing a standard free energy of binding of -8.5 ± 0.5 kcal/mol, accounting for the standard volume and biasing factors. The experimental value for this system is -8.0 kcal/mol. Simulation is termed Set 1.

Construction of the PMF over the entire data set thus results in a single value for free energy without specification of the error. In order to compute the error it is first instructive to determine the amount of sampling time per window required to stabilize the free energy. Computing the PMF for increasing sample time within each window (Figure 2) we see that convergence is achieved after approximately 50 ns with a value of -8.5 kcal/mol. This firstly indicates that the sampling can be reduced to 12 μs by considering a shorter (25 Å) reaction coordinate. However, it also indicates that a long equilibration time is necessary using an approach with a single simulation per window. An associated error is then determined by discretizing the post-equilibration region into 5 ns blocks and computing the block average (as done in previous studies[34]). This results in a binding free energy of $\Delta G^\circ = -8.5 \pm 0.5$ kcal/mol and compares well with a reported experimental value of -8.0 kcal/mol.[61] However, the accuracy of the above result comes at a substantial sampling cost (19 μs); it is thus desirable to lower these costs by optimizing the method.

3.2 Determining the optimal parameter set (OPS)

The first optimization strategy is to reduce the number of windows by increasing the window width to 0.5 Å. However, alteration of window width requires further optimization of harmonic restraints, in particular that of the umbrella sampling potential (k_z). To determine the optimal choice of k_{xy} and k_z, an ensemble of three umbrella sampling simulations for every window is performed for each of 10 permutations of k_{xy} and k_z (Set 2 in Table 1, 30 runs).

It is clear from Figure 3(a) and (b) that a stabilized PMF is exhibited for various selections of k_{xy} and k_z. For example, for $k_{xy} = 0.1$ kcal/mol/Å^2 and $k_z = 1$ kcal/mol/Å^2 (blue lines in Figure 3(a)), each of the three members of the ensemble show unchanging PMF values after 60 ns but vary amongst themselves over a range of 2.5 kcal/mol.

We have shown thus far that binding free energies attained using single runs exhibit stable PMFs with respect to themselves at 50 ns. However, convergence requires that multiple replicas of the same run converge to the same value. Here, even for 80 ns of sampling per window, no harmonic constraint permutation yields convergent results between the three members of its corresponding ensemble, except that of $k_{xy} = 1$ kcal/mol/Å^2 and $k_z = 0.5$ kcal/mol/Å^2 (green lines in Figure 3(a)). This set converges to within 0.5 kcal/mol within 50 ns of sampling per window at the given window width of 0.5 Å, establishing it as the optimal parameter set (OPS). After 50 ns of sampling the OPS thus attains an accurate binding free energy of -9.0 ± 0.5 kcal/mol, within 1 kcal/mol of experiment.

The chosen OPS exhibits the best convergence but only for an ensemble of three. Whilst sufficient to discriminate it from the other parameter sets, the absolute binding free energy convergence properties can be investigated better using a larger ensemble.

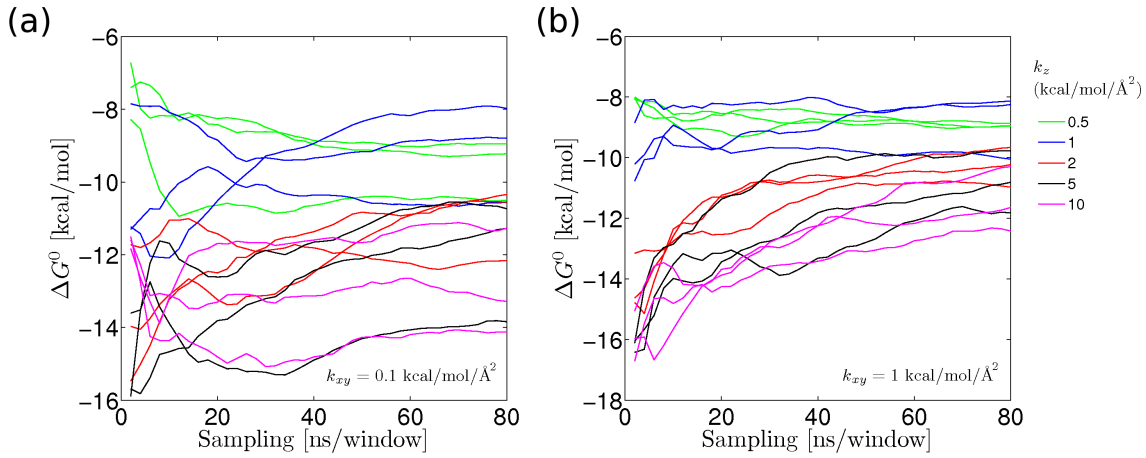

Figure 3: Harmonic constraint optimization over a range of k_z = 0.5,1,2.5,5,10 kcal/mol/Å^2 for k_{xy} = 0.1 (left) and k_{xy} = 1 kcal/mol/Å^2(right). Simulation is termed Set 2.

3.3 Convergence and sampling properties of the optimal parameter set

Here, we perform a larger ensemble of simulations using the OPS (Set 3 in Table 1, 10 runs) and analyze the corresponding convergence properties with an increase in sample time per window. We also investigate whether using less varied initial conformations across the US profile confers any difference to the accuracy or convergence properties the binding free energy. This entails a second ensemble of similar size and sampling time (Set 4 in Table 1, 10 runs).

The mean binding free energy across the ensemble set as a function of sample time is analyzed (Figure 4) for each. The free energy for Set 3 exhibits convergence at 6 ns with a free energy of -9.0 ± 0.9 kcal/mol and convergence to within 0.4 kcal/mol at 20 ns. By contrast, Set 4 does not exhibit true convergence even up to 20 ns, even though it yields a flattened mean binding free energy of -8.7 ± 1 kcal/mol. This is because unlike for Set 3, the error does not diminish significantly with increased sampling. Examination of the convergence of each single US run with increasing sample time (see Supporting Information) shows that whilst all single runs converge to the same value for Set 3, they do not for Set 4. Instead single runs stabilize on a particular binding free energy and this results in the error not diminishing for the latter whilst it does for the former.

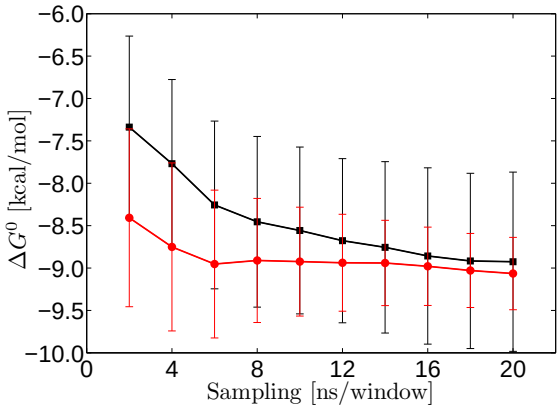

Figure 4: Comparison of two ensembles of 10 US simulations, one where each simulation was started from a different initial conformation per window (red), termed Set 3, and the other where each simulation was started from a single initial conformation per window (black), termed Set 4.

The above analysis draws us to conclude that, provided convergence is demonstrated through an ensemble of runs, a single run using the OPS and an aggregate sampling time of 300 ns is sufficient to provide a result to within 1 kcal/mol accuracy and precision to within 1 kcal/mol and 1 μs to within the same accuracy but a tighter convergence of within 0.5 kcal/mol. However, it is important to note that such an aggregate timescale also requires a sufficient relaxation time to be met within each window; in this case relaxation leads to con-

vergence of the binding free energy within 6 ns of sample time per window.

Furthermore, the analysis also demonstrates that the choice of initial conformations play a significant role in the attainment of convergence. More specifically, it is the sensitivity to the correlation between the initial conformations along a single profile of US windows that affects the convergence of the binding free energy. Furthermore, deriving initial conformations from a single preliminary MD run would require marginally less computation; however, the loss of convergence due to the correlated nature of the initial conformations prevents such a choice being optimal.

3.4 Structural correlates of differential sampling

Convergent sampling depends on the flexibility of the ligand and the protein across the reaction coordinate. Very flexible ligands/proteins or those capable of accessing multiple distinct conformations increase the convergence time because it requires to sample across all the relevant conformational degrees of freedom. The flexibility of both the protein and the ligand is thus assessed in terms of root mean squared fluctuations (RMSF) relative to the average structure in each window of the reaction coordinate (Figure 5).

In Figure 5, it is shown that the ligand is more rigid closer to the surface of the protein (RMSF $\sim$ 0.6 Å), and more flexible in the unbound state (RMSF $\sim$ 1.2 Å). The protein shows similar flexibility upon binding (RMSF $\sim$ 1.3 Å) with the ligand as when unbound. There is a sharp transition in flexibility in the ligand RMSF between 4.5 and 5.5 Å along the reaction coordinate, while the ligand transit from being bound at 4.5 Å to more flexible and unbound over a short distance of 1 Å.

The sensitivity of convergence to the correlation between initial conformations is investigated in more detail and the structural correlates pertaining to the variation of binding free energies for correlated PMF profiles determined. Firstly, the PMF profiles (Figure 6(a)) of all individual US runs belonging to Set 4 (black lines) show greater variability than the converged PMF profile range of Set 3 (red band). It is this variation that causes the 3 kcal/mol deviation between the highest and the lowest value for the free energy. Crucially, significant variation with respect to the convergence band commences across windows centered at 4.5 Å, 5 Å and 5.5 Å, corresponding to the sharp transition region exhibited in ligand flexibility, suggesting that it is in this region where orientational and conformational degrees of freedom play a more important role.

Three profiles from Set 4, corresponding to the binding free energies of -6.7 kcal/mol, -9.2 kcal/mol and -10.4 kcal/mol from individual simulations denoted $r1$, $r2$ and $r3$ respectively, are investigated more closely on the basis that the first underestimates the free energy, the second lies within the convergence band and the third overestimates the free energy. An examination of the integrated normalized probability distribution of the ligand center of mass across the windows centered at 4.5Å, 5 Å and 5.5 Å (Figure 6(b)) shows substantially different sampling compared to Set 3 (red). Whilst the converged ensemble samples a trimodal distribution consisting of a sharp peak at 0 Å, and two shallower peaks at 2.5 and 3.5 Å respectively, the three individual simulations $r1$, $r2$ and $r3$ each predominantly simulate a different mode from each other. Each one, however, corresponds to a mode within the trimodal distribution of Set 3. This confirms that the region with window centers between $z = 4.5$ Å and $z = 5.5$ Å thus corresponds to a sensitive transition region between bound and unbound states of pY for the two protocols.

The three sampling peaks exhibited along the reaction coordinate (Figure 6(b)) correspond to three distinct structural conformations (Figure 6(c) and (d)), which are all sampled correctly in the converged simulations but incorrectly in the individual runs. The most bound conformation (I), at $z = 0$ Å, consists of an extremely tight hydrogen bond network (6 hydrogen bonds) between the phosphotyrosine (pY) of the ligand and the R154, S156, E157 and S158 residues in SH2. This is due to the favorable conformation of flexible loop between residues 156 to 162 of SH2 (cyan). The second conformation (II), at $z = 2.5$ Å, corresponds to a slight retraction of the loop coupled with the increased separation of the ligand and results in the loss of 2 hydrogen bonds with S158. The third conformation (III) at $z = 3.5$ Å corre-

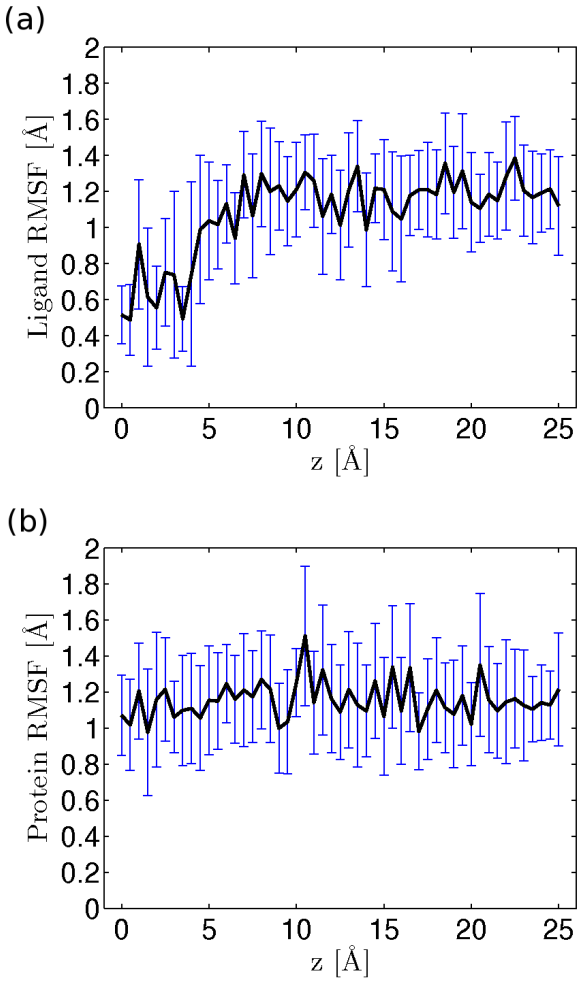

Figure 5: Backbone flexibility in terms of root mean squared fluctuations (RMSF) relative to the average structure in each window of the reaction coordinate. (a) The ligand is more rigid closer to the surface of the protein (RMSF $\sim$ 0.6 Å), and more flexible in the unbound state (RMSF $\sim$ 1.2 Å). A sharp transition in flexibility is seen between 4.5 and 5.5 Å along the reaction coordinate. Within 1 Å distance, the ligand transits from being rigidly bound to being flexible and unbound. (b) The protein instead, shows similar flexibility (RMSF $\sim$ 1.3 Å) between its bound and unbound conformations.

sponds to a more significant retraction of the flexible loop region losing all of its hydrogen bonds with pY; only a single hydrogen bond is maintained with R154. The individual run $r1$ thus oversamples the most bound conformation, increasing the PMF at that point and resulting in an excessive binding free energy. Conversely, $r3$ oversamples the more unbound conformation, flattening the PMF at that point and eventually resulting in a smaller free energy. Finally, $r2$ oversamples conformation II which compensates the loss in bound-conformation sampling with a loss in more unbound sampling and results in a PMF change within the convergence band leading to an accurate binding free energy.

The above analysis shows that the accuracy of the binding free energy calculation can be significantly affected by differently sampled structural events that occur in each window, especially in windows that correspond to sharp transitions in ligand flexibility and/or binding that have their root in discrete structural events such as hydrogen bonding. Whilst both stably bound and unbound states are easier to sample correctly by several approaches, convergence is more profoundly tested in the transition region between the two. How-

ever, it is not only the occurrence of such events in each window that matters but the overall integration across a number of relevant neighboring windows. Use of correlated initial conformations increases the chance of maintaining insufficient sampling across a set of neighboring windows resulting in an incorrect shift in PMF depth ($r1$ and $r3$) which is then propagated along the reaction coordinate. Even though this can occur for uncorrelated neighboring windows too, the latter exhibit far more sampling across windows resulting in a tighter convergence of the PMF.

4 Conclusions

In this work, we showed that it is possible to determine accurate, reproducible and scalable absolute protein-ligand binding free energies using molecular dynamics simulations, at least for the specific case used here. Our optimized protocol employs a simply-biased 1D-PMF umbrella sampling method applied using an ensemble of simulations, initiated from uncorrelated initial conformations across neighboring windows and an optimal parameter set (OPS) describing orthogonal restraints, force constant for the sampling potential, window width and sampling time per window.

Applied to the SH2 domain binding to the pY-EEI ligand, we obtain an absolute binding free energy of -9.0 $\pm$ 0.5 kcal/mol, in good agreement with experiment (1 kcal/mol deviation), demonstrating the accuracy of the method. The minimum aggregate sampling time to compute an accurate result is 300 ns with the OPS, a significant improvement over the 25 μs aggregate sampling of a previous method.

Our methodology is also demonstrated to be reproducible; that is, ensemble based repetition of the calculation shows convergence to within 1 kcal/mol amongst independent simulations for the above-mentioned aggregate sampling time. Furthermore, we show that it is correct sampling of sensitive bound-unbound transition regions, corresponding to various phosphotyrosine interactions in the binding groove, that determine the convergence of the result. Structural correlates of differential sampling account for the discrepancies between different methodologies, and the optimal methodology presented here overcomes such sensitivities.

The protocol reported concerns the calculation of the PMF for systems where there is a direct path from the bulk to the binding site, thus making it amenable to the 1D-PMF method. Calculations for more complex binding processes that involve multiple reaction coordinates and/or significant protein-ligand conformational changes upon binding are beyond the remit of the method. Within the remit, however, as the protocol reported here does away with system specific conformational restraints and the corresponding human choices of system construct, it is readily scalable to a large number of protein-ligand systems. There may be limits of transferability for the parameters optimized on this system when applied to other systems. A priori it is difficult to determine whether certain classes of system will exhibit transferable parameters, but it is likely that flexibility, ligand size and binding pathway will play an important role. Cases where the protein is very flexible,[68] much more so than the SH2 domain, may cause a problem because the umbrella sampling would need to sample correctly all the conformations. This was possible here where the conformational fluctuation was limited to a loop of the SH2 domain by properly sampling the initial conformations of the umbrella sampling. The same problem of conformational sampling applies for very flexible ligands. Also, if the exit pathway of the ligand is very narrow, care has to be used in the selection of the exit direction. In summary, we would expect this methodology and parameter set to work for semi-rigid proteins (small loop movements) and semi-rigid ligands with an easy access pathway to the binding site. In the case where these parameters may not be directly transferable (free energy of binding very different from the experimental value), we believe that a good approach is to enhance the creation of initial configurations for the umbrella sampling before undergoing any deep optimization study. Once the limit of this protocol is reached, additional optimization methods like Hamiltonian replica exchange[69] would need to be considered. Finally, as the accuracy and precision obtained is, in this case, very high, it supports the accuracy of the forcefield for the given ligand. However, ligand forcefield accuracy is not

the general case, which means that the extensive and convergent sampling provided by this methodology may allow the validation and improvement of forcefield accuracy for different ligands.

Acknowledgement The authors are grateful to the volunteers of GPUGRID who donate GPU computing time to the project. IB acknowledges support from the Obra Social Fundació "La Caixa". SKS acknowledges support from a European Commission FP7 Marie Curie IEF. GDF acknowledges support from the Ramón y Cajal scheme and from the Spanish Ministry of Science and Innovation (Ref. FIS2008-01040).

References

(1) Kollman, P. *Chemical Reviews* **1993**, *93*, 2395–2417.

(2) 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, T. E. r. *Acc. Chem. Res.* **2000**, *33*, 889–897.

(3) Wang, W.; Wang, J.; Kollman, P. A. *Proteins: Structure, Function and Genetics* **1999**, *34*, 395–402.

(4) Aqvist, J.; Medina, C.; Samuelsson, J.-E. *Protein Eng.* **1994**, *7*, 385–391.

(5) Carlson, H. A.; Jorgensen, W. L. *Journal of Physical Chemistry* **1995**, *99*, 10667–10673.

(6) Kuhn, B.; Kollman, P. A. *Journal of Medicinal Chemistry* **2000**, *43*, 786–3791.

(7) Wittayanarakul, K.; Hannongbua, S.; Feig, M. *J. Comput. Chem.* **2008**, *29*, 673–685.

(8) Schwarzl, S. M.; Tschopp, T. B.; Smith, J. C.; Fischer, S. *Journal of Computational Chemistry* **2002**, *23*, 1143–1149.

(9) Rizzo, R. C.; Toba, S.; Kuntz, I. D. *Journal of Medicinal Chemistry* **2004**, *47*, 3065–3074.

(10) Wang, W.; Kollman, P. A. *Journal of Molecular Biology* **2000**, *303*, 567–582.

(11) Zoete, V.; Michielin, O.; Karplus, M. *Journal of Computer-Aided Molecular Design* **2003**, *17*, 861–880.

(12) Stoica, I.; Sadiq, S. K.; Coveney, P. V. *J. Am. Chem. Soc.* **2008**, *130*, 2639–2648.

(13) Sadiq, S. K.; Wright, D. W.; Kenway, O. A.; Coveney, P. V. *Journal of Chemical Information and Modeling* **2010**, *50*, 890–905.

(14) Lu, N.; Singh, J. K.; Kofke, D. A. *Journal of Chemical Physics* **2003**, *118*, 2977–2984.

(15) Price, D.; Jorgensen, W. *Journal of Computer-Aided Molecular Design* **2001**, *15*, 681–695.

(16) Reddy, M.; Erion, M. *J. Am. Chem. Soc* **2001**, *123*, 6246–6252.

(17) Shirts, M. R.; Pande, V. S. *Journal of Chemical Physics* **2005**, *122*, 144107:1–16.

(18) Fowler, P. W.; Jha, S.; Coveney, P. V. *Phil. Trans. R. Soc. A* **2005**, *363*, 1999–2015.

(19) Wan, S.; Coveney, P. V.; Flower, D. R. *Phil. Trans. R. Soc. A* **2005**, *363*, 2037–2053.

(20) Mobley, D. L.; Chodera, J. D.; Dill, K. A. *The Journal of Chemical Physics* **2006**, *125*, 084902.

(21) Gervasio, F.; Laio, A.; Parrinello, M. *J. Am. Chem. Soc* **2005**, *127*, 2600–2607.

(22) Fidelak, J.; Juraszek, J.; Branduardi, D.; Bianciotto, M.; Gervasio, F. *The Journal of Physical Chemistry B* **2010**, *114*, 9516–9524.

(23) Babin, V.; Roland, C.; Sagui, C. *The Journal of chemical physics* **2008**, *128*, 134101.

(24) Jarzynski, C. *Physical Review E* **2002**, *65*, 046122.

(25) Jarzynski, C. *Physical Review Letters* **1997**, *78*, 2690–2693.

(26) Roux, B. *Computer Physics Communications* **1995**, *91*, 275–282.

(27) Virnau, P.; M"uller, M. *The Journal of chemical physics* **2004**, *120*, 10925.

(28) Pietrucci, F.; Marinelli, F.; Carloni, P.; Laio, A. *J. Am. Chem. Soc.* **2009**, *131*, 11811–11818.

(29) Woo, H.-J.; Roux, B. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 6825–6830.

(30) Wang, J.; Deng, Y.; Roux, B. *Biophysical journal* **2006**, *91*, 2798–2814.

(31) Shivakumar, D.; Deng, Y.; Roux, B. *Journal of Chemical Theory and Computation* **2009**, *5*, 919–930.

(32) Gan, W.; Roux, B. *Proteins: Structure, Function, and Bioinformatics* **2009**, *74*, 996–1007.

(33) Deng, Y.; Roux, B. *The Journal of Physical Chemistry B* **2009**, *113*, 2234–2246.

(34) Doudou, S.; Burton, N.; Henchman, R. *Journal of Chemical Theory and Computation* **2009**, *5*, 909–918.

(35) Kumar, S.; Rosenberg, J.; Bouzida, D.; Swendsen, R.; Kollman, P. *Journal of Computational Chemistry* **1992**, *13*, 1011–1021.

(36) Snow, C. D.; Nguyen, H.; Pande, V. S.; Gruebele, M. *Nature* **2002**, *420*, 102–106.

(37) 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*, 084108.

(38) Jayachandran, G.; Vishal, V.; Pande, V. S. *J. Chem. Phys.* **2006**, *124*, 164902.

(39) Voelz, V.; Bowman, G.; Beauchamp, K.; Pande, V. *Journal of the American Chemical Society* **2010**, *132*, 1526–1528.

(40) 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.

(41) Klepeis, J.; Lindorff-Larsen, K.; Dror, R.; Shaw, D. *Current opinion in structural biology* **2009**, *19*, 120–127.

(42) Shirts, M.; Pande, V. S. *Science* **2000**, *290*, 1903–1904.

(43) Giupponi, G.; Harvey, M.; De Fabritiis, G. *Drug discovery today* **2008**, *13*, 1052–1058.

(44) Harvey, M.; Giupponi, G.; Fabritiis, G. *Journal of Chemical Theory and Computation* **2009**, *5*, 1632–1639.

(45) Harvey, M.; De Fabritiis, G. *Journal of Chemical Theory and Computation* **2009**, *5*, 2371–2377.

(46) Buch, I.; Harvey, M.; Giorgino, T.; Anderson, D.; De Fabritiis, G. *Journal of chemical information and modeling* **2010**, *50*, 397–403.

(47) Fabritiis, G.; Geroult, S.; Coveney, P.; Waksman, G. *Proteins: Structure, Function, and Bioinformatics* **2008**, *72*, 1290–1297.

(48) Sadowski, I.; Stone, J.; Pawson, T. *Molecular and Cellular Biology* **1986**, *6*, 4396.

(49) Waksman, G.; Shoelson, S.; Pant, N.; Cowburn, D.; Kuriyan, J. *Cell* **1993**, *72*, 779–790.

(50) Botfield, M.; Green, J. *Annual Reports in Medicinal Chemistry* **1995**, *30*, 227–237.

(51) Bradshaw, J.; Mitaxov, V.; Waksman, G. *Journal of molecular biology* **1999**, *293*, 971–985.

(52) Tong, L.; Warren, T.; King, J.; Betageri, R.; Rose, J.; Jakes, S. *Journal of molecular biology* **1996**, *256*, 601–610.

(53) Zhou, S. et al. *Cell* **1993**, *72*, 767–778.

(54) Sheinerman, F.; Al-Lazikani, B.; Honig, B. *Journal of molecular biology* **2003**, *334*, 823–841.

(55) Brugge, J. *Science* **1993**, *260*, 918–919.

(56) Gibbs, J.; Oliff, A. *Cell* **1994**, *79*, 193–198.

(57) Sawyer, T. *Peptide Science* **1998**, *47*, 243–261.

(58) Morelock, M.; Ingraham, R.; Betageri, R.; Jakes, S. *Journal of medicinal chemistry* **1995**, *38*, 1309–1318.

(59) Cousins-Wasti, R.; Ingraham, R.; Morelock, M.; Grygon, C. *Biochemistry* **1996**, *35*, 16746–16752.

(60) Bradshaw, J.; Waksman, G. *Biochemistry* **1998**, *37*, 15400–15407.

(61) Lee, T.; Lawrence, D. *J. Med. Chem* **2000**, *43*, 1173–1179.

(62) Lubman, O.; Waksman, G. *Journal of molecular biology* **2003**, *328*, 655–668.

(63) Nam, N.; Ye, G.; Sun, G.; Parang, K. *J. Med. Chem* **2004**, *47*, 3131–3141.

(64) Fowler, P.; Geroult, S.; Jha, S.; Waksman, G.; Coveney, P. *J. Chem. Theory Comput* **2007**, *3*, 1193–1202.

(65) MacKerell Jr, A.; Banavali, N.; Foloppe, N. *Biopolymers* **2000**, *56*, 257–265.

(66) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926–935.

(67) Hess, K.; Berendsen, H. *Journal of Computational Chemistry* **1999**, *20*, 786–798.

(68) Sadiq, S. K.; De Fabritiis, G. *Proteins* **2010**, *78*, 2873–2885.

(69) Sindhikara, D.; Meng, Y.; Roitberg, A. E. *The Journal of Chemical Physics* **2008**, *128*, 024103.

Supporting Information Available: This material is available free of charge via the Internet at `http://pubs.acs.org/`.

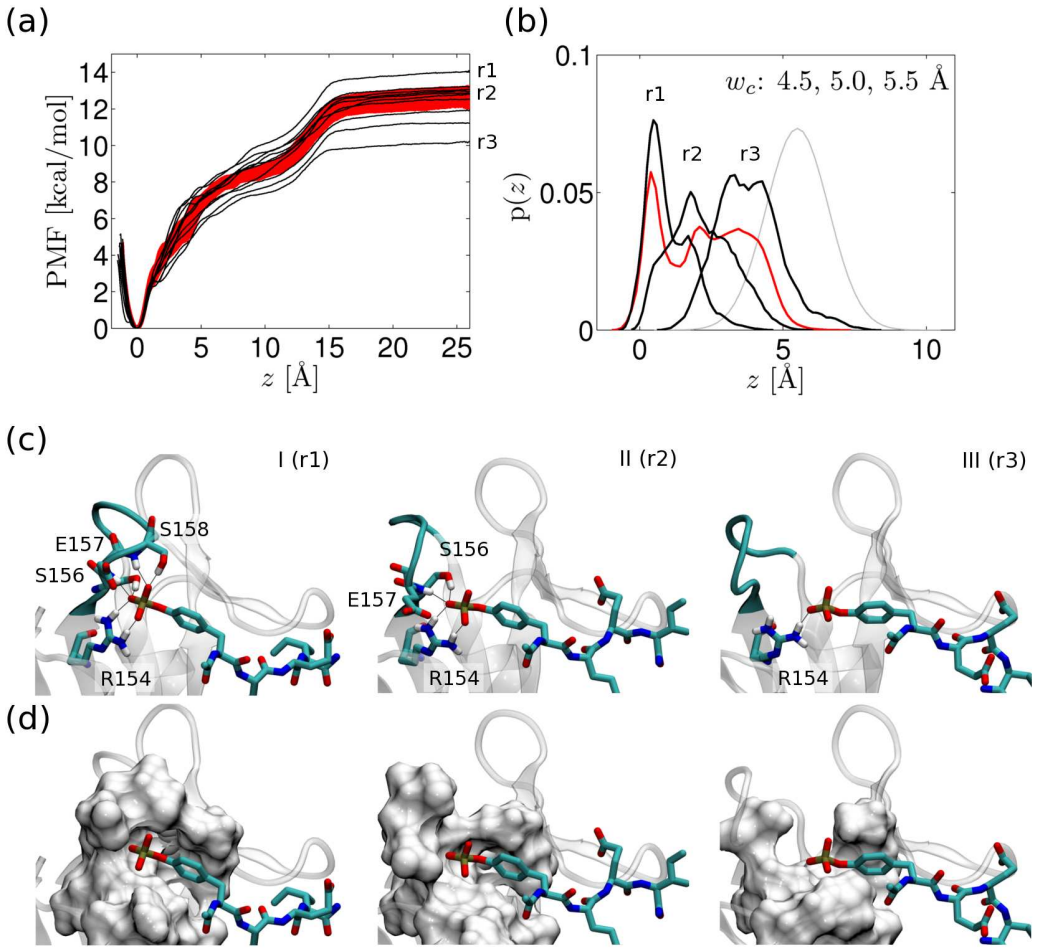

Figure 6: (a) PMF of all members of the Set 4 ensemble (black lines) against the PMF range of Set 3 (red band). Notable individual members corresponding to excessive, accurate and underestimated binding free energies are denoted *r*1, *r*2 and *r*3 respectively. (b) Integrated probability distribution of *r*1, *r*2 and *r*3 for the differential sampling region exhibited in the PMF across windows centered at 4.5, 5, and 5.5 Å. The aggregate distribution of Set 3 is also shown (red) as well as the theoretical distribution for a system acting only under the restraining potential (gray). (c) Principal structural correlates, corresponding to the three sampling peaks in (b), showing pY interaction with R184, S156, E157 and S158 of SH2. The differential conformation of the flexible loop region (cyan) corresponds to the degree of hydrogen bonding exhibited. (d) Surface representation of the corresponding conformations.