Sensitivity in predicted relative binding free energies from incremental ligand changes within a model binding site

Nathan Lim*

Department of Pharmaceutical Sciences

E-mail: limn1@uci.edu

Abstract

Despite innovations in sampling techniques for molecular dynamics (MD), reliable prediction of protein-ligand binding free energies from MD remains a challenging problem, even in well studied model binding sites like the apolar cavity of T4 Lysozyme L99A. In this study, we model recent experimental results that show the progressive opening of the binding pocket in response to a series of homologous ligands.[1] Even while using enhanced sampling techniques, we demonstrate that the predicted relative binding free energies (RBFE) are still highly sensitive to the initial protein conformational state. Particularly, we highlight the importance of sufficient sampling of protein conformational changes and possible techniques for addressing the issue.

*To whom correspondence should be addressed

1 Introduction

Proteins play a central role in many biological processes by modulating many key signalling pathways. It is unsurprising to find that proteins make up the vast majority of pharmaceutical drug targets. Many small-molecules on the market today induce their therapeutic effect by modulating the proteins biological activity through binding.[2–4] Thus, optimization of protein-ligand binding affinities is a central goal in early pharmaceutical drug design projects.

Recent advancements in technology and computational chemistry have led to increasing use of computer-aided drug design techniques for this lead optimization problem. Since accuracy and reliability of the approach being used is critical for success, it is here where the most rigorous of methods like free energy calculations can be applied in the prediction of protein-ligand binding affinities. Using methods like free energy perturbations (FEP) and λ-dynamics on molecular dynamics simulations, the difference in binding free energies between two ligands can be accurately computed through using an alchemical pathway. By computing relative binding free energies, much of the computational cost and difficulties of absolute binding free energy calculations is avoided.[5–11] Other advancements in forcefields, sampling algorithms, and emergence of GPU computing have led to the tremendous success of using FEP calculations across a wide range of protein targets with a variety of ligands.[12]

Here, we apply the same FEP protocol, which utilities replica exchange with solute tempering (FEP/REST),[13] to a very simple model binding site in an engineered mutant of T4 lysozyme (L99A). In this mutant of T4 lysozyme, the introduction of the L99A mutation creates a small apolar binding site that has been studied extensively experimentally[14–17] and computationally by docking[1,18–20] and free energy methods[13,21–27] Recent studies on this binding site have found that the protein adopts three discrete conformations in response to ligand binding. Through a series of eight congeneric ligands, that grow by a single methyl addition, the protein responds by a single helix rearrangement to accommodate the growing ligand.[1] Using this FEP/REST protocol and the aforementioned homologous ligand

series, we calculate relative protein-ligand binding affinities between ligands that occupy different discrete protein conformational states. In this study—while using the implemented default FEP/REST protocol—we demonstrate how the kinetically distinct protein states and structural rearrangement affects the accuracy and reliability of our predicted relative binding affinities. Further, we illustrate the importance of sufficient sampling of protein conformational changes and some modifications or possible techniques to address this issue.

2 Methods

Protein/Ligand preparation

All proteins were prepared and aligned in Maestro[28] using the 'Protein Preparation Wizard'[29–33] tool and with the following settings enabled (as they appear in the Maestro GUI menu):

- Preprocess: Assign bond orders, Add hydrogens, Create zero-order bonds to metals, Create disulfide bonds, Cap termini, Delete waters beyond 5Åfrom het groups
- Refine: Sample water orientations, Use PROPKA pH: 7.0, Remove waters with less than 3 H-bonds to non-waters, and restrained minimization.

Protein structures were taken from PDBs: 4W52, 4W53, 4W54, 4W55, 4W56, 4W57, 4W58, and 4W59 corresponding to protein-ligand bound structures of benzene, toluene, ethylbenzene, n-propylbenzene, sec-butylbenzene, n-butylbenzene, n-pentylbenzene, and n-hexylbenzene, respectively.[1] Each simulation starts from either the protein closed state (PDB:4W52) or the open state (PDF:4W59). When using the FEP+ protocol, ligand crystal structure positions were used as the starting position of the simulation. When using the LigandFEP protocol, a similar workflow to the tutorial[34] was followed. Generally, two options were taken:

(1a) If the simulation starts from the protein closed state, the benzene crystal position was

used as a reference for fragment building (PDB:4W52).
(1b) The corresponding ligand in the transformation was built by duplicating benzene in place and adding methyl groups.
(2a) If the simulation starts from the protein open state, the n-hexylbenzene crystal position was used as reference for fragment building (PDF:4W59).
(2b) The corresponding ligand in the transformation was built by duplicating n-hexylbenzene in place and deleting methyl groups.

Ligand tail fragments were added using the Build/Fragments toolbar in Maestro and were not overlaid or docked. As the ligand tails were built, bonds were manually rotated so that the tail was oriented in a similar manner as in their corresponding crystal structure. Following, the newly added atoms in the tail were locally minimized while leaving the core in its initial position. This was done in an attempt to correct bond angles and minimize the core RMSD, which LigandFEP uses to select the ligand heavy atoms to be included in the REST region.

Classification of alchemical transformations and color coding

Here, we classified ligands based on the primary protein conformation (closed, intermediate, or open) the ligand occupies from the experimental studies (Fig 2,Table 1). To be explicit, the set of closed ligands refers to benzene, toluene, ethylbenzene, and n-propylbenzene; n-/sec-butylbenzene for intermediate; and n-pentyl/n-hexylbezene for open ligands. The various protein states and ligands are then assigned a color accordingly: purple for closed, cyan for intermediate, and green for the open state.

Our sets of alchemical transformations consisted of 3 groups: 'closed-intermediate', 'closed-open', and an 'experimental' ligand set. The first two sets were classified based on the expected conformational change that would result from the alchemical transformation. For example, the alchemical transformation of benzene to n-butylbenzene falls into closed-intermediate

while benzene to n-hexylbenzene is a closed-open transformation. In our experimental set, we perform all possible combinations of transformations for ligands with available experimental binding affinities (Table 1). Ligands with available experimental binding affinities consisted of benzene, toluene, ethylbenzene, n-propylbenzene, and n-butylbenzene. This gives a total of 26 alchemical transformations in this study, 8 from 'closed-intermediate', 8 from 'closed-open', and 10 from the experimental set.

FEP protocols

Using the Schrödinger application suite (release 2015-3),[35] two FEP protocols were used in this project: FEP+[12] and LigandFEP.[34] FEP+ is a fully automated work flow that plans perturbation pathways based off the LOMAP[36] mapping algorithm which uses the maximum common substructure (MCS) between any pair of compounds. LigandFEP is an academic toolkit that generates the configuration files to perform the free energy calculation but is limited in the sense that the user must plan each perturbation path instead. Both FEP protocols use the default Desmond relaxation protocol and the FEP/REST methodology.[13,27,37,38]

Simulation details

Desmond[39–42] simulation protocols have been described previously in the supporting information[12] or can be found in greater detail in the Desmond User Manual.[43] The relaxation protocol beings with a simulation where solute molecules are restrained to their initial positions while minimizing using a Brownian dynamics NVT integrator for 100ps, followed by 12ps simulations at 10K with a NVT ensemble and then a NPT ensemble using the Langevin method.[44] Next is a 24ps simulation followed by a final 240ps simulation with solute molecules unrestrained, both are carried at room temperature with a NPT ensemble using Langevin. Production simulations were initially ran for the default setting of 5ns and then were carried out to a length of up to 55ns for closed-open transformations and up to 25ns for closed-intermediate transformations. Here, we use the final 15ns for closed-open simulations and

the final 10ns for closed-intermediate to calculate our final free energies, discarding the initial time as additional equilibration time. FEP/REST simulations were run on four GeForce GTX Titan Black GPUs using the Desmond/GPU engine with OPLS2005[45] and OPLS3[46] forcefield parameters. Here, results and discussion sections will primarily present information from using the LigandFEP protocol with OPLS3 forcefield parameters, additional data not discussed here can be found in the supplementary info.

REST region selection

In this study, by default, only heavy atoms in the ligand were included in the REST region unless specified otherwise. Further details on the temperature profile and how the REST region is normally selected can be found in previous studies[12,13] in the supporting information. Simulations that included protein heavy atoms in the REST region are referred to with the 'pREST' label, where selection of the particular residues is described as follows.

Based on visual inspection of our molecular dynamics simulations and considering the F-helix spans residues 107-115, we selected residues Glu108, Val111, and Gly113 to include into the REST region (Fig 3a). Glu108 sits near the start of the helix which appears as a hinge point for the opening and closing of the binding cavity (Fig 3b). Following, Val111 appears in the middle of the helix and was observed to undergo the largest motion during protein conformational changes (Fig 3c). Gly113 was included in order to collectively have hot regions approximately at the start, middle and end points of the helix.

Calculation of free energies and measurement of inconsistency

Throughout this study, we measure the inconsistency ($\Delta\Delta G_{\boldsymbol{\varepsilon_n}}$) between the final calculated free energies between simulations that start from the protein closed state ($\Delta\Delta G_{\mathbf{C_n}}$) versus the protein open state ($\Delta\Delta G_{\mathbf{O_n}}$) by simply taking the difference.

$$\boldsymbol{\Delta\Delta G_{\varepsilon_n}} = |\Delta\Delta G_{\mathbf{C_n}} - \Delta\Delta G_{\mathbf{O_n}}| \tag{1}$$

Then we compute the overall inconsistency—referred to as the 'Root-Mean-Square-Inconsistency'(RMSI)—for each set of alchemical transformations. The RMSI is calculated by using the differences ($\Delta\Delta G_{\boldsymbol{\varepsilon_n}}$) obtained from the comparisons between protein open and closed simulations.

$$\mathbf{RMSI} = \sqrt{\frac{\sum_n (\Delta\Delta G \varepsilon_n)^2}{n}} \tag{2}$$

Similarly, we compute the 'Root-Mean-Square-Error'(RMSE) when comparing with experimental free energies for both simulations starting from the protein closed and open state.

$$\mathbf{RMSE^O} = \sqrt{\frac{\sum_n (\Delta\Delta G_{\mathbf{O_n}} - \Delta\Delta G_{\mathbf{exp_n}})^2}{n}} \mathbf{RMSE^C} = \sqrt{\frac{\sum_n (\Delta\Delta G_{\mathbf{C_n}} - \Delta\Delta G_{\mathbf{exp_n}})^2}{n}} \tag{3}$$

Calculated free energies were determined using the Bennett acceptance ratio [47] (BAR) with error estimations using both bootstrapping and BAR analytical error prediction. [48] Hysteresis around closed thermodynamic cycles and best estimates of the free energies with their errors were calculated using the cycle closure algorithm discussed in a previous publication. [13]

Determining the protein conformation state using RMSD

In this study, we determine the state of the protein by computing the 'Root-Mean-Square-Deviation'(RMSD) of the protein backbone atoms spanning the F-helix relative to their positions found in the closed (PDB:4W53), intermediate (PDB:4W57), and open (PDB:4W59) crystal structures (Fig 1b). The set of RMSDs is computed over the course of the simulation, whereby we use the state with the lowest RMSD to color each time point correspondingly in our analyses of 'RMSD/time' and 'Color maps'. Here, we use VMD [49,50] to align and compute the RMSD of our Desmond trajectories relative to crystal structures. Further details on the procedure and the scripts used for these analyses are provided in the supplementary info.

For the 'RMSD/time' analysis, see Figure 4b) for reference. Here, we plot the RMSD to the closed helix, represented by the black line, where each time point is colored according to the lowest RMSD state. We apply the RMSD/time analysis only to the simulation

corresponding to the end state ligand of interest (Λ_{11}). By tracking the RMSD relative to the closed helix, we can monitor if the protein opens (by high RMSD with green points) or closes (by low RMSD with purpe points). Additionally, we gain some insight on the time required to capture the opening or closing of the binding cavity.

It is important to note that by limiting our view to only the end-state replica, we encounter time points that may appear contradictory, at a glance. For example in figure 4b, between 3-4ns there are two frames where the protein is closed and the RMSD is slightly greater than 2.0Å. Following, there is one frame where the protein is open but the RMSD is lower than 2.0Å. This appears to be an artifact of coordinate swapping during replica exchanges from other intermediate lambda windows, resulting in apparently contradictory time points.

We address this limitation by viewing all replicas in what we call 'Color maps', see Figure 7b for an example. Essentially, our 'Color map' analysis is the same as our 'RMSD/time' plots but without the RMSD line plot. In other words, we color time points according to the protein state of lowest RMSD and do this for all replicas but do not track the RMSD relative to the closed state. Through a collective view of all replicas, we gain a better perspective of the overall protein conformational sampling and the states they occupy over each replica's separate trajectories. Using color maps, it becomes visually easy to see if intermediate—higher temperature—lambda windows are able to sample, say the open state, and if this leads to an enhancement in sampling at the end states via replica exchange.

3 Results

Calculated free energies depend strongly on starting protein conformation

Using the default FEP/REST methodology,[13] we find calculated free energies significantly depend on the protein starting conformation, especially for large perturbations (i.e. opening

the cavity from the closed state). To illustrate this, we begin our molecular dynamics simulations both from the protein closed and open conformations then perform alchemical transformations to ligands that occupy another protein conformational state. For example, in the alchemical transformation of benzene to n-hexylbenzene—starting from the protein closed state—we expect to see opening of the binding cavity when the ligand is in the fully interacting n-hexylbenzene state. In this study, we demonstrate using the default Schrödinger FEP protocol settings of a 5ns simulation time and REST region selection does not facilitate adequate sampling of the motion in the F-helix and does not eliminate the dependence on the initial protein state.

Closed-Open Ligand Transformations

An examination of the largest alchemical transformation, benzene to n-hexylbenzene, clearly highlights the sampling challenges faced when using the standard FEP/REST protocol. From experimental data of ligand occupancies (Table 1), we expect in our simulations of n-hexylbenzene to see the protein primarily in the open state over the closed state. Instead, we find the protein remains trapped in its initial conformational state whether we start from closed (Fig 4a) or open (Fig 4b) over the course of the 5ns simulation. From the protein closed simulation, the protein only briefly samples the intermediate state around 3ns but never enters the open conformation. As the protein tries to accommodate n-hexylbenzene and enter its preferred open state, protein-ligand strain results, yielding a positive value for $\Delta\Delta G_{calc}$(+4.13 kcal/mol). On the other hand, in the protein open simulations, the protein already begins in its preferred state for n-hexylbenzene and stays only in this open state. As expected, the $\Delta\Delta G_{calc}$ is negative(-0.61 kcal/mol) as there is no occurrence of large protein-ligand strain in order to open the cavity. By remaining trapped in the initial state, we under-sample the open state if we begin from the closed state or over-sample it if we begin from the open state. Ultimately, we arrive at two very different relative free energies values, where the inconsistency is as large as +4.74 kcal/mol for the same transformation of

benzene to n-hexylbenzene.

In the overall set, we similarly observe protein closed simulations to yield positive free energies and negative for protein open simulations. In turn, we find the overall inconsistency to be very high with a RMSI of +4 kcal/mol (Fig 8a,Table 5) Clearly, despite the use of FEP/REST, we are unable to get sufficient sampling in the protein within the standard 5ns time frame. Instead, we encounter sampling problems as the protein remains in its initial conformational state throughout the simulation. As a result, our calculated free energies exhibit high dependence on the initial protein configuration which is reflected by the large RMSI.

Closed-Intermediate Ligand Transformations

In the case of closed-intermediate alchemical transformations, we find that the calculated free energies still have some (albeit much smaller) dependence on the initial protein conformation, using the default protocol. For this set of alchemical transformations, we find the RMSI to be +0.60 kcal/mol (Fig 8c,Table 2). Considering this set involves a smaller protein conformational change and smaller perturbations to the ligand, it is unsurprising to find the RMSI to be much smaller than our closed-open transformation set.

Although, the collective RMSI for closed-intermediate transformations falls below the acceptable range of less than 1 kcal/mol, we can still see a dependence on the initial protein configuration by viewing transformations involving n-butylbenzene. For these cases in particular, we observe the same pattern of protein closed simulations yielding positive free energies and negative for protein open simulations. We do not see this pattern for transformations with sec-butylbenzene as it does not partially occupy the open state, unlike n-butylbenzene (Table 1). Through this observation, we demonstrate further that with the default protocol does not eliminate the free energy dependence on the protein starting conformation, even for smaller perturbations.

Experimental Ligand Transformations

Now, when we compare $\Delta\Delta G_{calc}$ against $\Delta\Delta G_{exp}$, we find that simulations starting from the protein closed conformation are further from converging to $\Delta\Delta G_{exp}$ than when starting from the protein open conformation. Here, we calculate the RMS-'Error' with experiment and find the RMSE for protein closed simulations to be +1.0 kcal/mol and +0.58 kcal/mol with protein open (Fig 8e,Table 8). By the fact that protein open simulations are much closer to $\Delta\Delta G_{exp}$, once again demonstrates our calculated free energies depend on the initial protein state. Unsurprisingly, the large RMSE seen for protein closed simulations primarily comes from transformations involving n-butylbenzene. Evidently, we find the simulations involving n-butylbenzene remain trapped in their respective starting conformations, resulting in inadequate sampling in the protein closed simulations (Fig. 5a) versus the protein open simulations (Fig. 5b). Despite performing much smaller alchemical transformations, this shows we still encounter sampling problems that result in $\Delta\Delta G_{calc}$ that depend on the initial protein conformation, evident when comparing to $\Delta\Delta G_{exp}$.

Including protein residues into the REST region (pREST) improves sampling

Primarily, we encounter major sampling problems when we begin our simulations from the protein closed state and attempt a mutation which should result in the binding cavity opening. In order to facilitate protein motion, thereby enhancing protein sampling, we included 3 key residues spanning the F-helix region into the REST region, which we will denote simulations using this with 'pREST' (Fig 3a). By expanding the REST region, we are able to drive the F-helix out its initial state trap by locally heating up key regions and thereby reduce our sampling problem.

To demonstrate the REST improvement over the default protocol, we return to the case of benzene to n-hexylbenzene. Here, we show the facilitation of the helix motion by first

referring to Figure 4a which shows that there is no sampling of the open state for the default protocol. Now with the pREST, we see a few open state points around 3ns and even a single open point before closing again after our initial step (Fig 6a). Alternatively, we can further illustrate the enhancement of protein sampling by viewing all replicas collectively, using the color maps. In reference to Figure 7a and Figure 7b, we illustrate that there is far less sampling of the intermediate or open protein states in default simulations versus the pREST simulations.

Collectively, we find only some minor improvements in the RMSI for all our closed-open and closed-intermediate transformations while using pREST. For closed-open transformations (Table 6), the RMSI reduces to +2.78 kcal/mol (previously +4 kcal/mol). On the other hand, for closed-intermediate, the RMSI raises slightly to +0.78 kcal/mol (previously +0.60 kcal/mol), which may be due to statistical noise(?) (Table 3). Generally, simulations starting from the closed state had $\Delta\Delta G_{calc}$ values that moved towards favorability (i.e. more negative $\Delta\Delta G_{calc}$), while for protein open simulations $\Delta\Delta G_{calc}$ values tended towards unfavorability (i.e. more positive $\Delta\Delta G_{calc}$). This is indicative of the fact that pREST is indeed improving sampling, but it is evident that our $\Delta\Delta G_{calc}$ are still far from convergence, given the RMSI is still large, especially for closed/open transformations.

Long simulations enhances protein conformational sampling from more exchanges

Although we see improvements in sampling with pREST, the standard implemented time frame of 5ns clearly is not long enough to gain adequate sampling, particularly if we start from the protein closed state. By running longer, we allow our simulations to perform more exchanges across replicas and thereby allow for better sampling of all conformational states at the relevant end state replicas.

Returning to our most extreme transformation, benzene to n-hexylbenzene, we have shown pREST alone does not facilitate adequate sampling of the open state (Fig 6a). Now,

when we run much longer we see far more sampling of the open protein conformational state in the final 10ns window (Fig 6b).[1]In viewing all the replicas (Fig 7c), we illustrate the dramatic increase in protein conformational sampling in stark contrast to our previous 5ns simulations (Fig 7b).

By simulating longer with pREST we dramatically increase our sampling of the intermediate and open protein states and almost entirely eliminate the dependence on the initial protein conformational state. For the set of closed-open transformations the RMSI dramatically falls to +0.57 kcal/mol (Fig. 8b,Table 7) and a RMSI of +0.42 kcal/mol for the closed-intermediate (Fig. 8d,Table 4). Similarly, for experimental ligand transformations, our RMSE for protein closed simulations falls to +0.54 kcal/mol (Fig. 8f,Table 9). Now, all our inconsistencies in the final calculated free energies and error from experiment fall within a much more reasonable range of less than +1 kcal/mol.

4 Discussion

In this study, we find that relative free energy calculations can suffer from substantial convergence problems resulting from relatively modest protein conformational changes. Although, the protein conformational changes in T4 lysozyme (L99A) are extremely localized to a rearrangement of a single helix (Fig 1a,Fig 1b), we still encounter challenges in sampling. These problems have profound implications for the accuracy of computed relative free energies in these cases. Particularly, we find that calculated relative free energies can depend on the initial protein conformational state by up to 4 kcal/mol.

By looking at alchemical transformations that involve a conformation change in the protein, we show the $\Delta\Delta Gs_{calc}$ are sensitive to the initial protein conformational state when utilizing the default implemented FEP protocol. Such cases are when the alchemical transformation involves mutating ligands that primarily occupy the closed state (i.e. benzene to

[1]We remind the reader that for closed-open transformations we use the final 15ns to compute the final free energies. We only show the final 10ns in our RMSD/time analysis to avoid overcrowding data points.

n-propyl) into ligands that occupy the intermediate state (i.e. sec-/n-butyl) or, especially, the open state (n-pentyl/n-hexyl) (Table 1). By our RMSD analyses, we show the protein remains trapped in its initial state throughout the simulation when using the implemented default protocol. Through remaining trapped, we are unable to adequately sample the necessary protein conformational states and thereby obtain inconsistent $\Delta\Delta Gs_{calc}$ depending on whether we start simulations from the open or closed protein configuration.

Without prior knowledge of preferred protein conformational states on ligand binding, we can arrive at very different binding affinity predictions based on the initial protein state being used in the simulation. By starting from the protein close state and growing the ligand we obtain $\Delta\Delta G_{calc}$ values that appear overly positive or unfavorable due to high protein-ligand energy strain and inability to sample the open state. However, when we begin with the protein open state our $\Delta\Delta G_{calc}$ values appear overly negative or favorable, a result from inability to sample the closed state and not encountering protein-ligand strain. If we only had the crystal structure of the closed protein-ligand complexes, we would blindly conclude that the much larger, open-ligands bind to T4 lysozyme worse than smaller ones. On the other hand, if only the open protein-ligand complexes were available, the opposite would be concluded in that larger ligands are better binders than smaller ligands.

By including key residues into the REST region and simulating longer, we reduce the $\Delta\Delta G_{calc}$ dependence on the initial protein configuration to a more reasonable range of less than 1 kcal/mol. Through expanding the REST region, intermediate lambda windows are able to more easily access the intermediate and open conformations by effectively heating key residues that facilitated protein motion, illustrated in Fig 7b. Further, by simulating longer we allow for more exchanges between replicas, which in turn enhances sampling at our physically relevant end state replica (Fig 7c). With these modifications to the default protocol, we almost completely converge our $\Delta\Delta G_{calc}$ to the same value regardless of the starting protein conformation.

Generally, our brute-force approach of simulating longer and multiple trials with varied

protein structures is not a desirable or a even a feasible approach, especially in early drug discovery phases. At the industrial level, ligand libraries can be large—driving computational cost exponentially if we simulate longer—or experimental structures can be sparse for new therapeutic protein targets. For future studies, approaches using Markov State Models (MSMs)[51] can potentially be of great use for identifying discrete protein conformations. MSMs build a representation of the conformational space from batches of short molecular simulations, whereby the discrete states and transition rates between them can be determined in an efficient manner. Utilizing MSMs can thereby provide useful insight on the various protein conformational states before running free energy predictions.

5 Conclusions

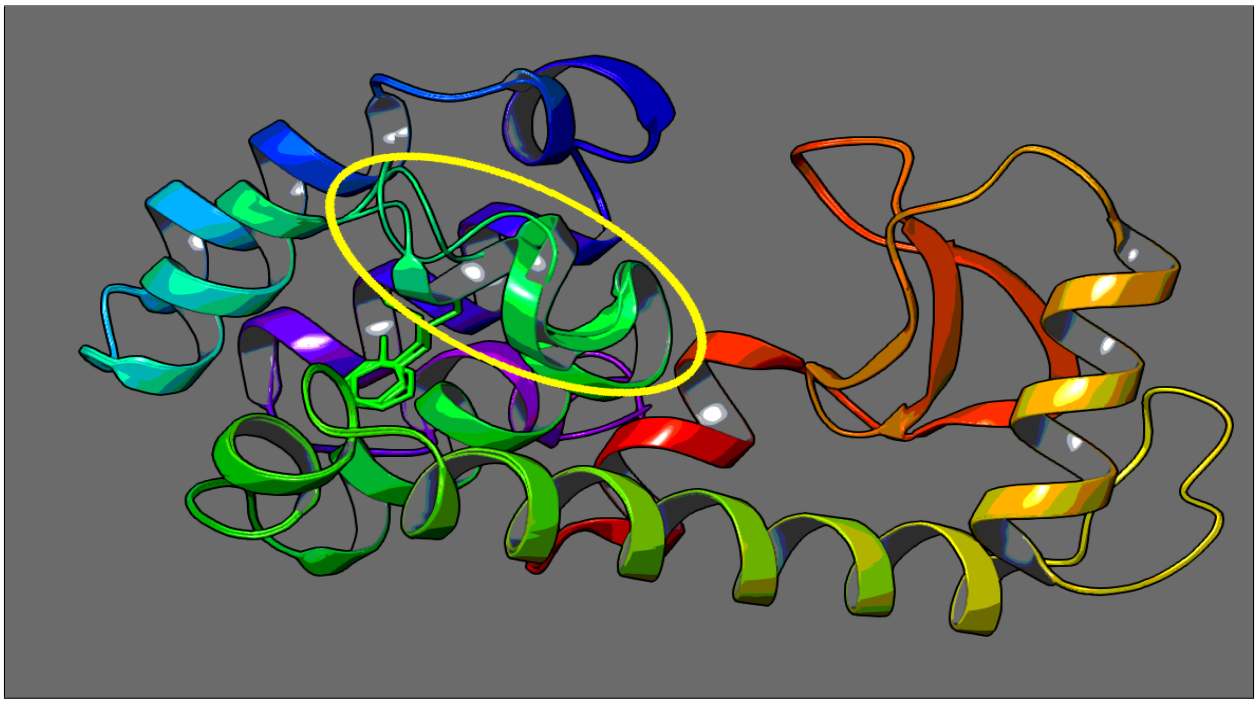
Overall, we have shown that the presence of kinetically distinct protein conformational states can dramatically impact the accuracy free energy calculations. In a worse-case scenario for this study, if only the apo protein structure was available and with no prior knowledge of the discrete states, identifying the bias in the free energies would have been challenging, if at all possible. It would have been especially challenging as there would essentially be no indicators that the final free energies were sensitive to the initial protein configuration. Only from prior knowledge of the discrete states and by our tedious systematic trials were we able to identify and address the bias in our final calculated free energies.

An alarming issue to be raised here is that our worst case scenario—only having the apo protein structure—is a challenge medicinal chemists face often in early stages of drug discovery. For example, a medicinal chemist could be tasked with docking a ligand library to the apo protein structure of some new exciting potential therapeutic target. In order to determine the most suitable candidates to pursue further, the chemist is then asked to run binding free energy predictions. Dangerously, the chemist would then discard ligands with apparently low binding affinities, where these ligands only appear unfavorable because

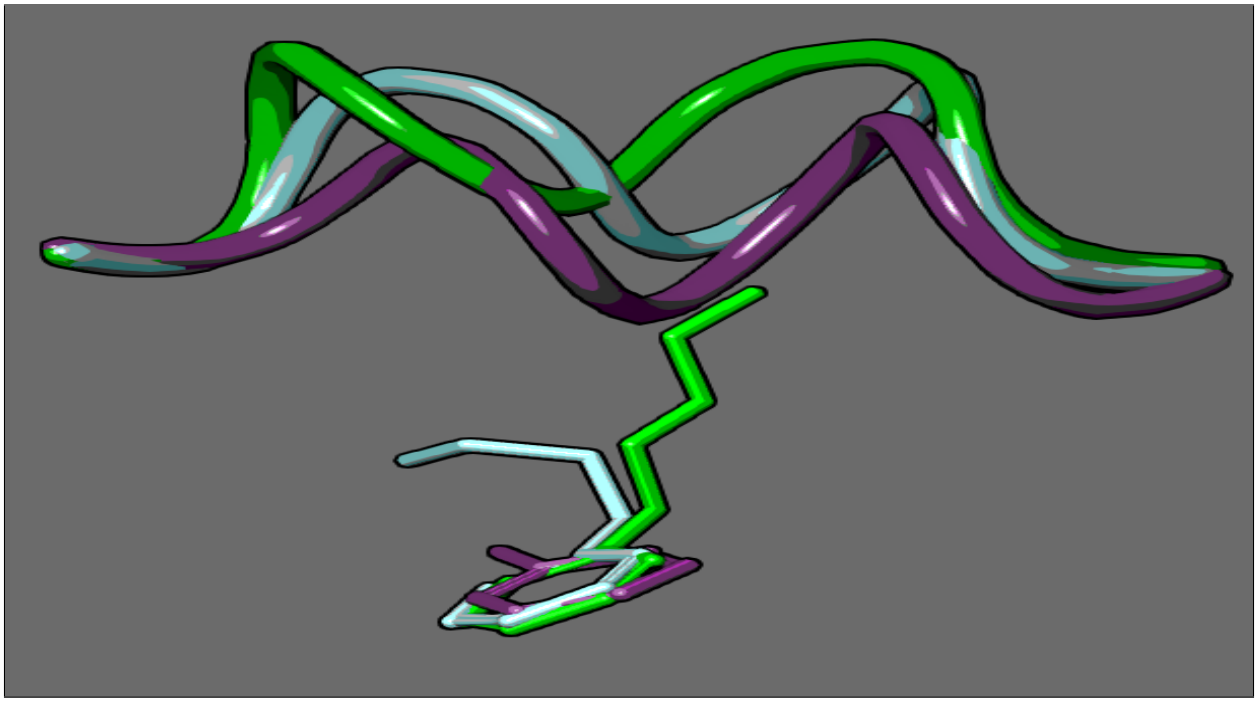
of unsampled protein conformational changes. Without prior knowledge of the existence of other protein conformational states, the chemist would never know that the calculations were actually incorrect due to insufficient sampling of the other possible states.

Although FEP calculations have shown tremendous recent successes on a variety of protein targets,[12] we demonstrate challenges in protein sampling remain. Using T4 lysozyme (L99A) as our simple model system, we highlight sampling problems even from a relatively small (1-3.5Å) and localized single helix rearrangement in response to a series of growing ligands. Through this study, we show using a typical 5ns simulation with only ligand atoms in the REST region, yields free energies that are sensitive to the initial protein conformation. Only by longer simulation times and expansion of the REST region to include key protein residues were we able to eliminate the dependence on the starting conformation. This study demonstrates that special attention and care should be exercised when performing FEP calculations where regions of flexibility surround the binding site. More importantly, prior to performing binding free energy calculations, we present strong evidence on the importance of identifying the occurrence of protein conformational changes upon ligand binding,

Figures



(a) T4-L99A



(b) T4-L99A

Figure 1: T4-L99A

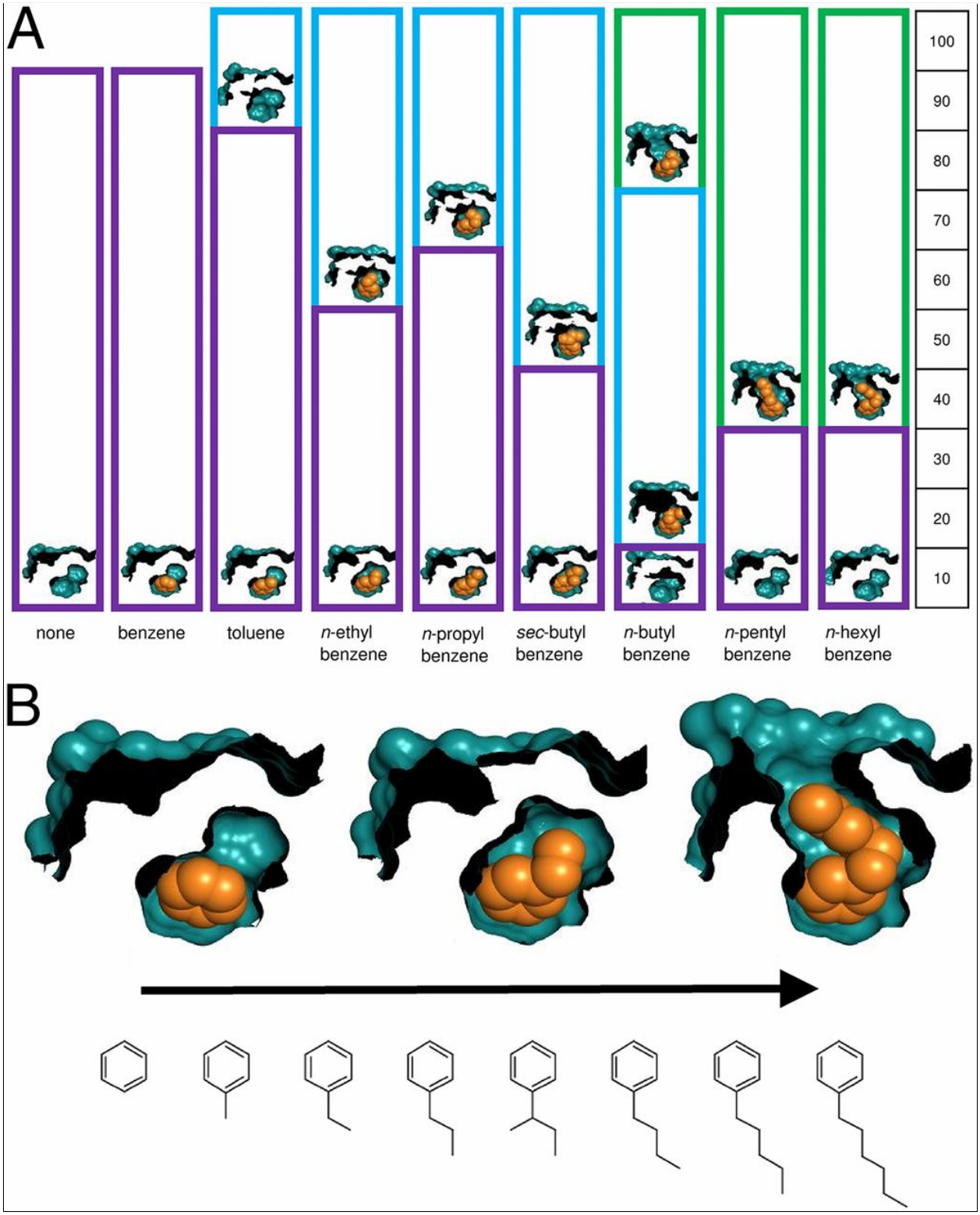
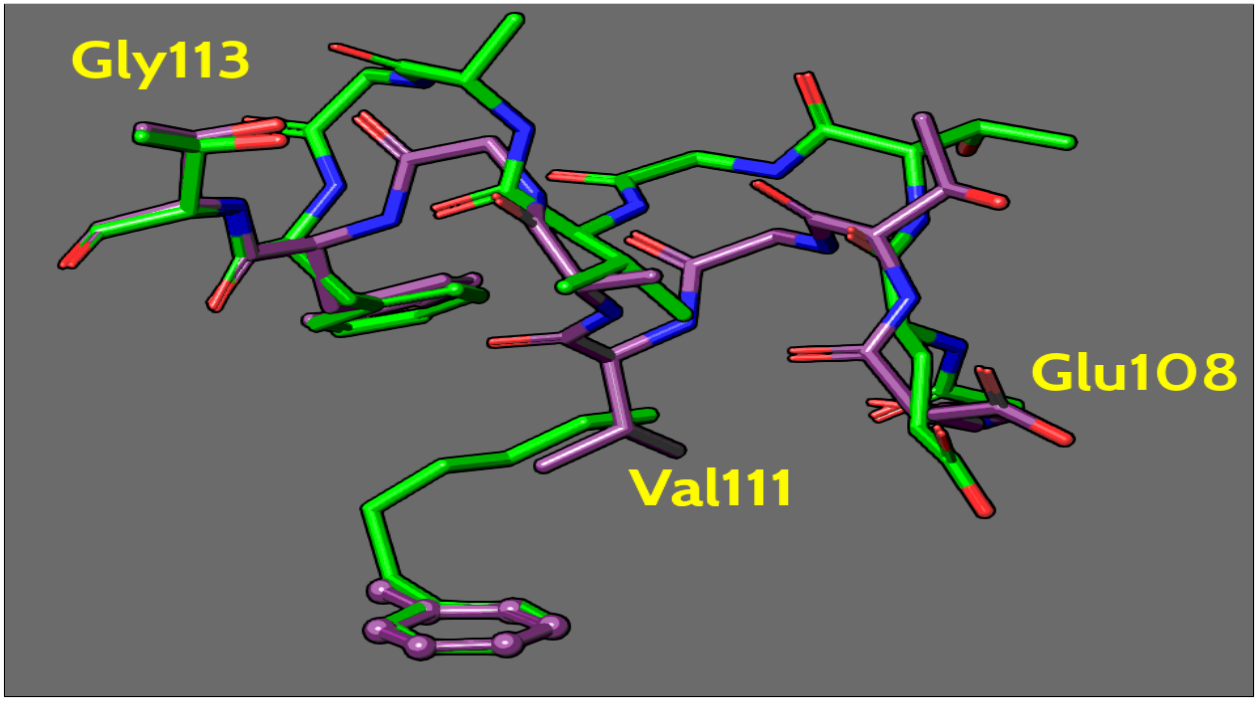
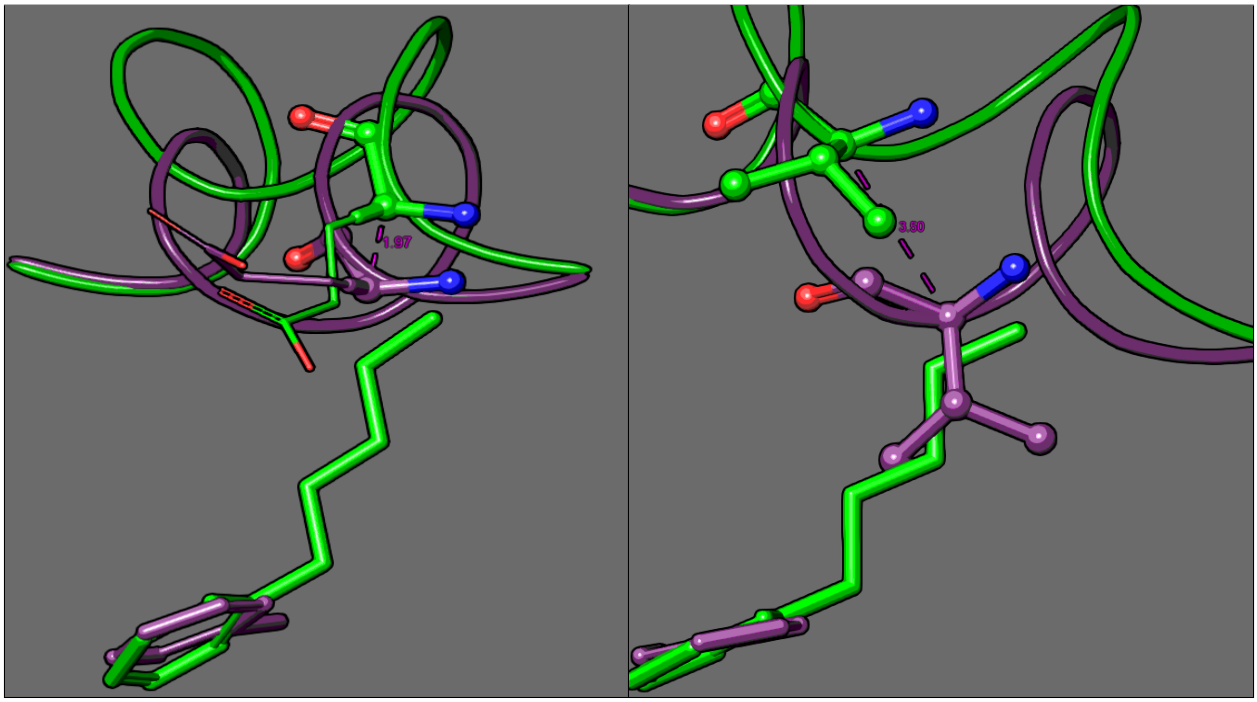


Figure 2: "Congeneric ligands are accommodated in L99A with conformational changes. (A) In the L99A cavity, the ligand poses were assigned to their respective protein conformations by matching the ligand occupancy with that of the F-helix conformation, which was typically unambiguous. (B) Molecular surface of the cavity, cut away to reveal the ligand (orange space-filling model), in examples of the closed (benzene complex), intermediate (ethylbenzene complex), and open (n-hexylbenzene complex) conformations. The full congeneric series is shown." Adapted figure[1]



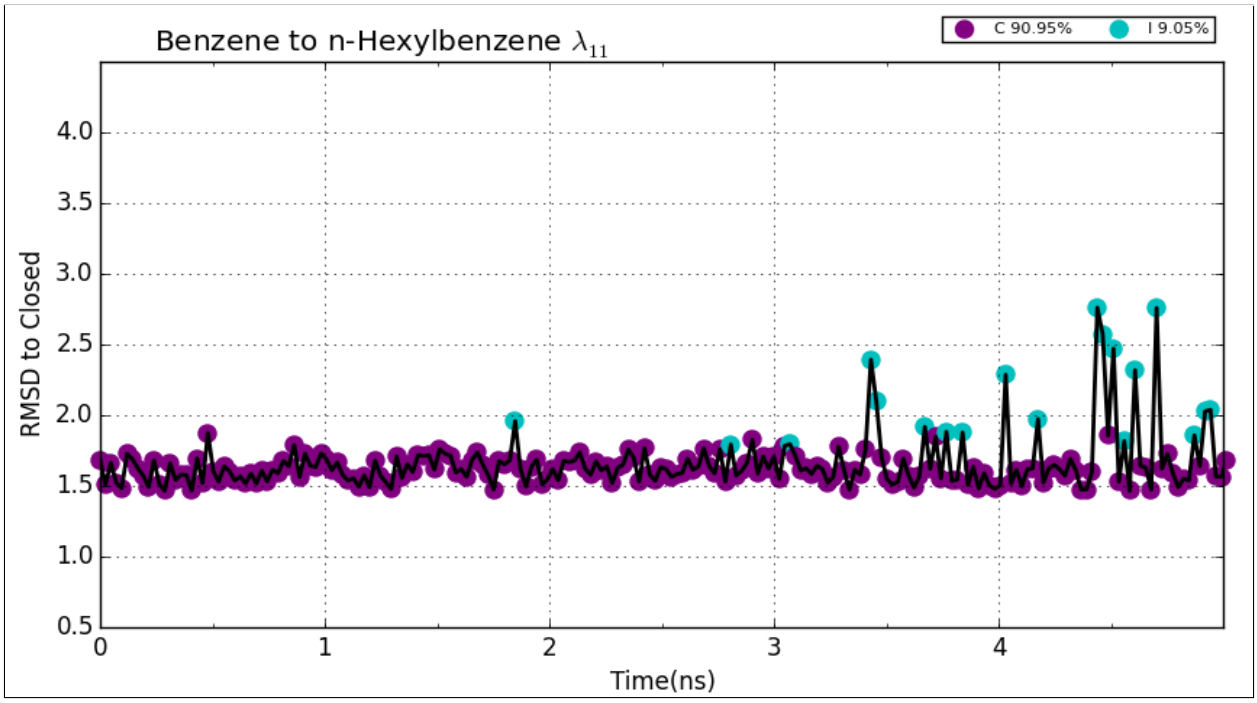
(a) Selected pREST residues



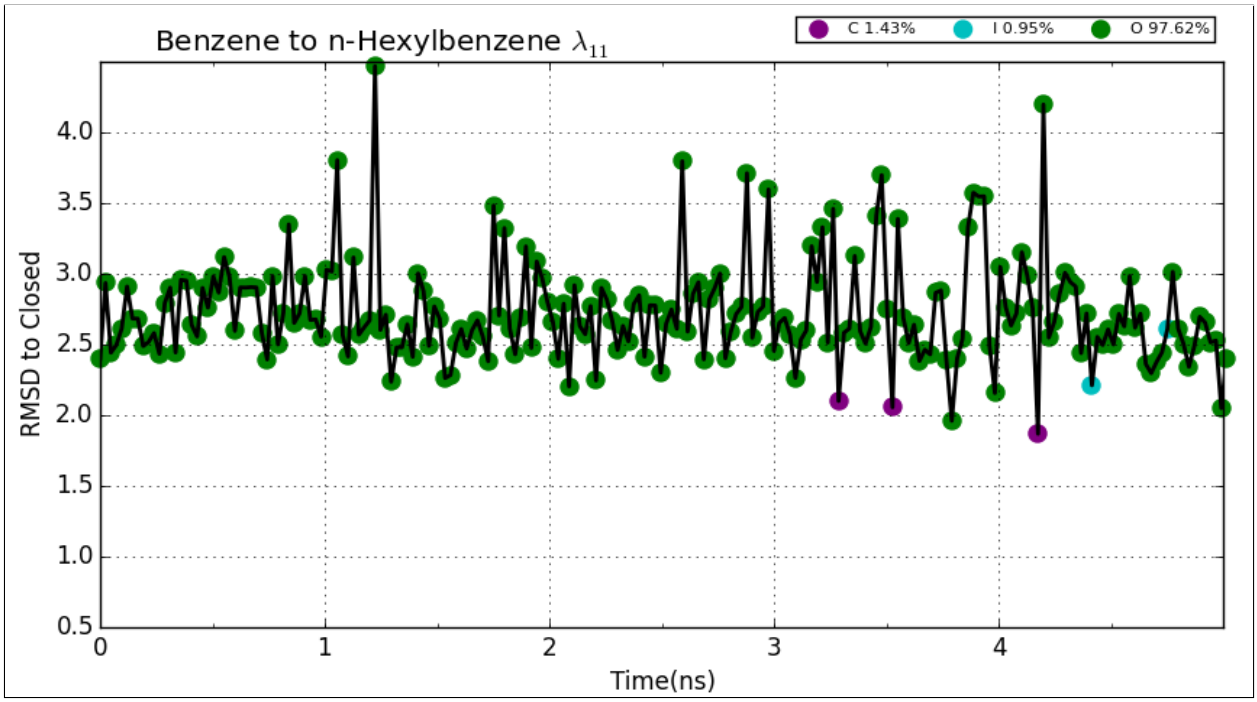
(b) Residue Glu108

(c) Residue Val111

Figure 3: pREST residues

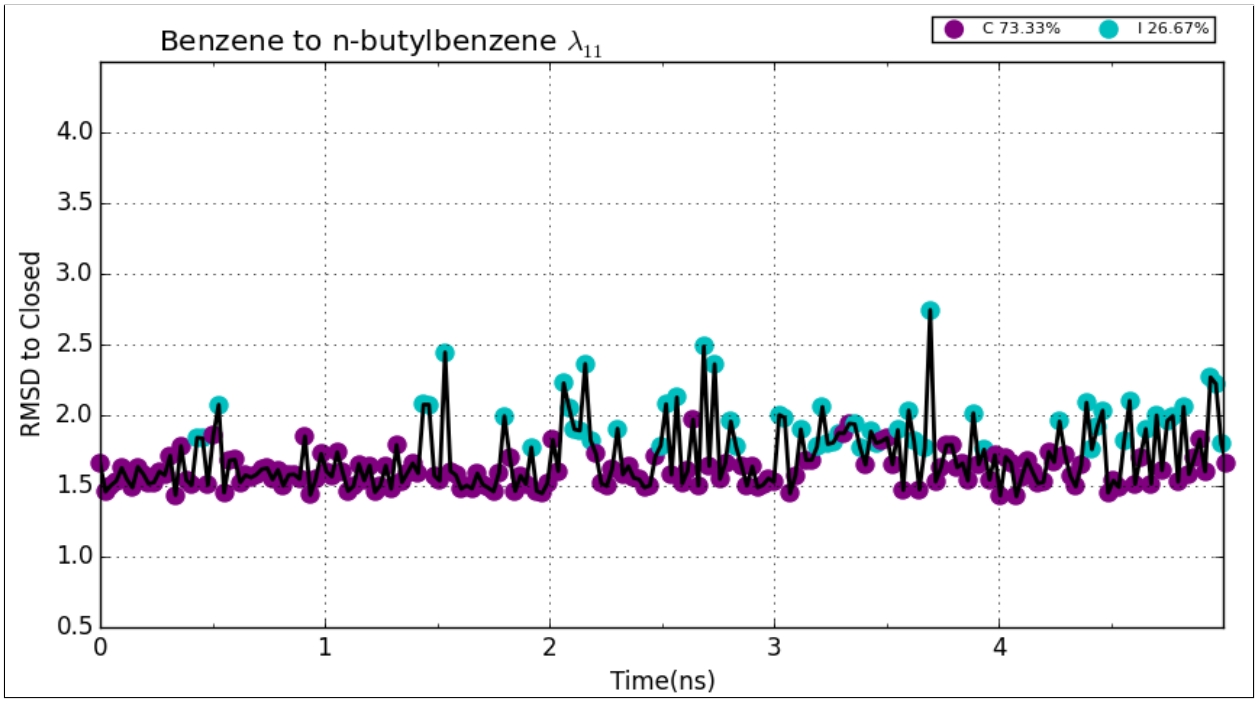


(a) Closed - Benzene to n-Hexylbenzene 0-5ns RMSD Replica11

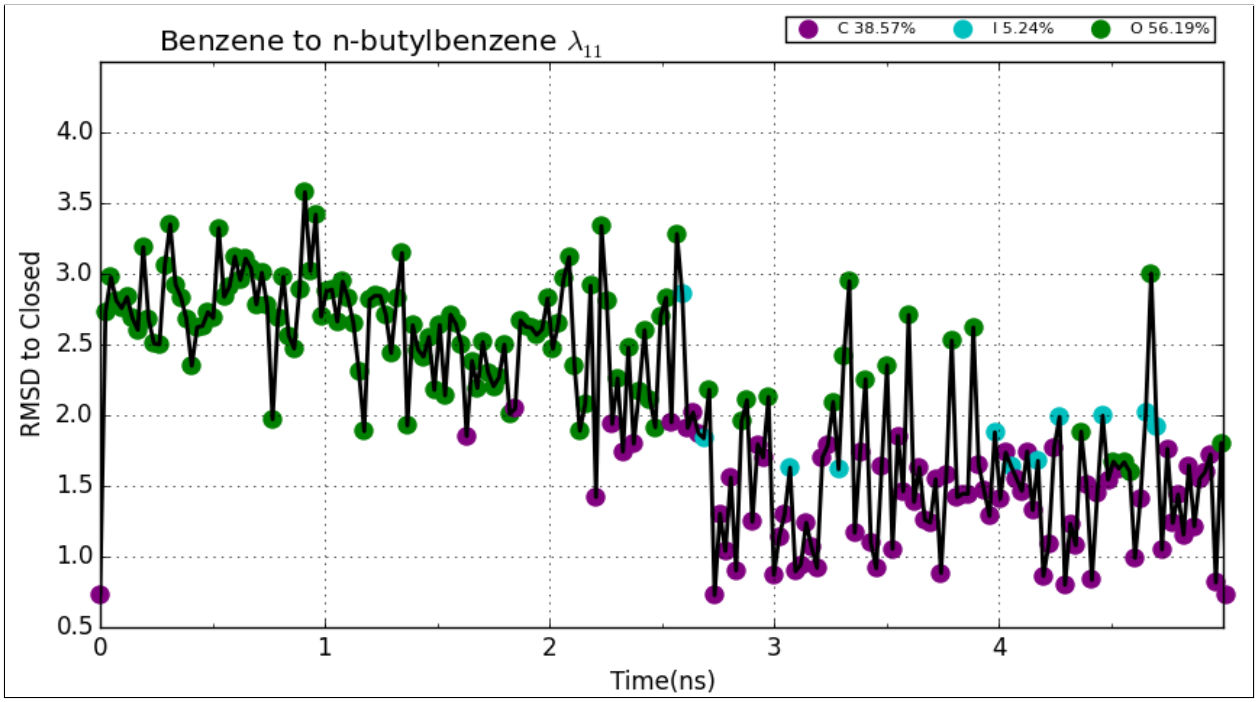


(b) Open - Benzene to n-Hexylbenzene 0-5ns RMSD Replica11

Figure 4: Benzene To n-Hexylbenzene (Default)

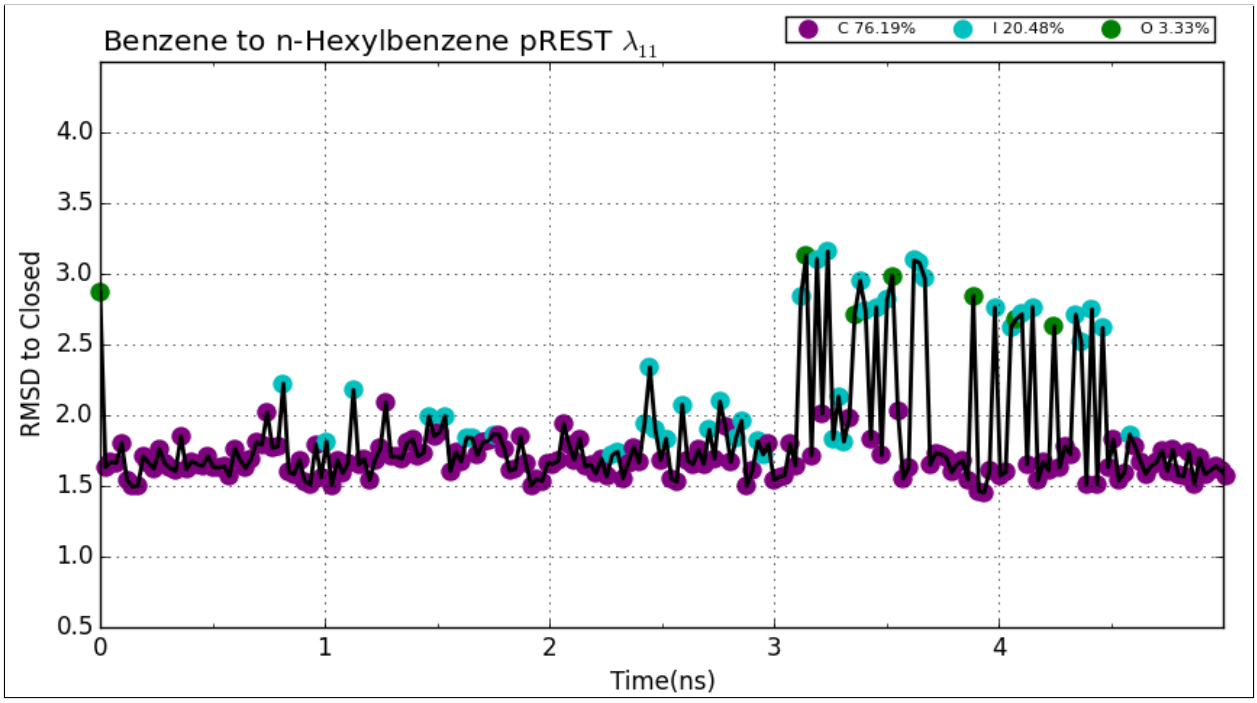


(a) Closed - Benzene to n-butylbenzene 0-5ns RMSD Replica11

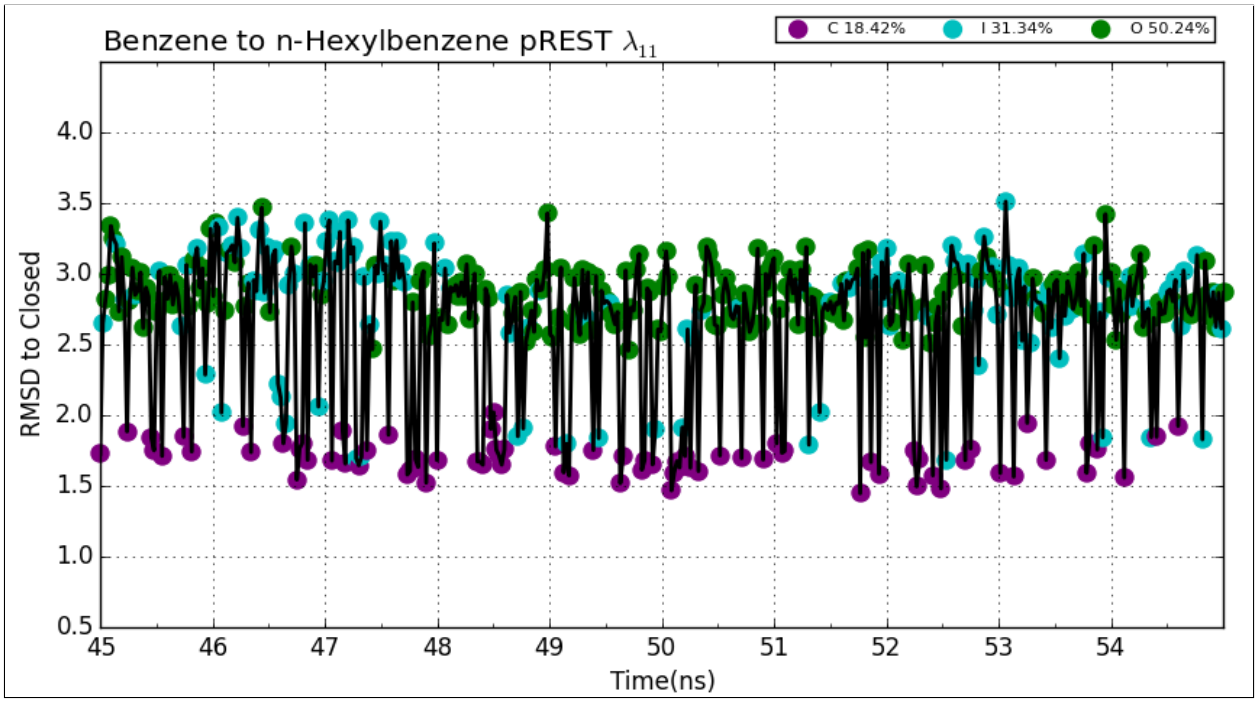


(b) Open - Benzene to n-butylbenzene 0-5ns RMSD Replica11

Figure 5: Benzene To n-Butylbenzenebenzene (Default)

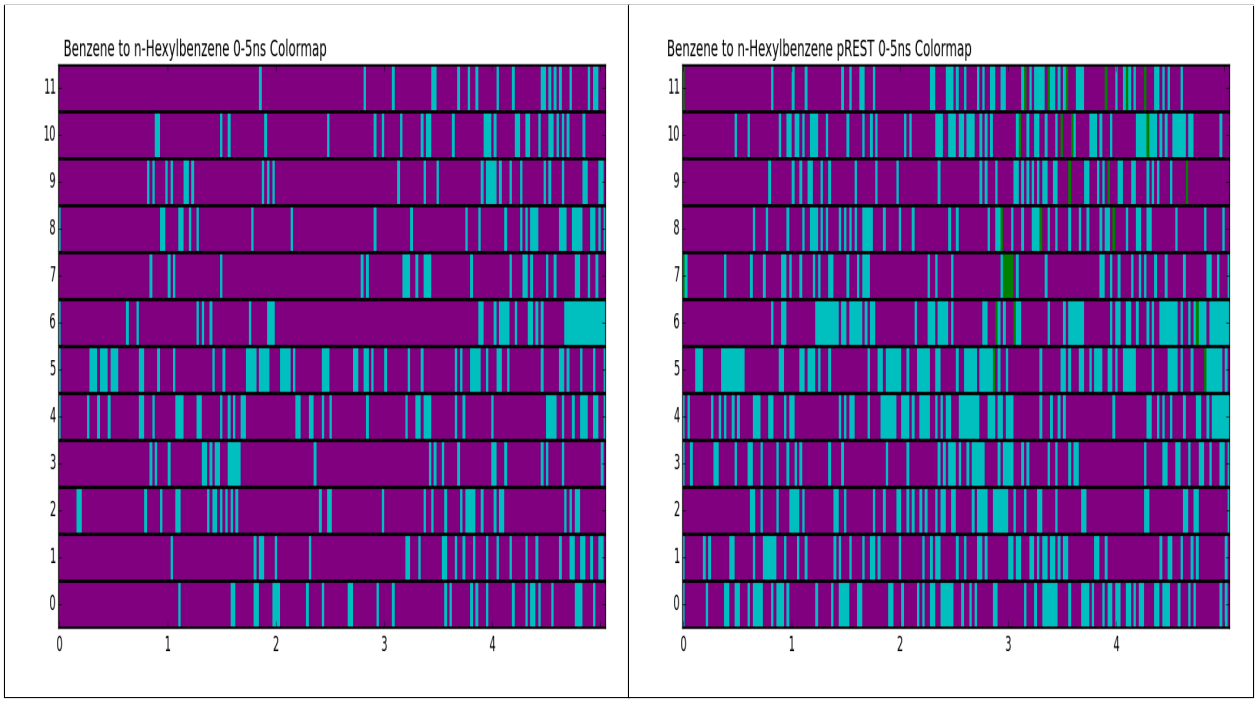


(a) Closed - Benzene to n-Hexylbenzene 0-5ns RMSD Replica11 pREST

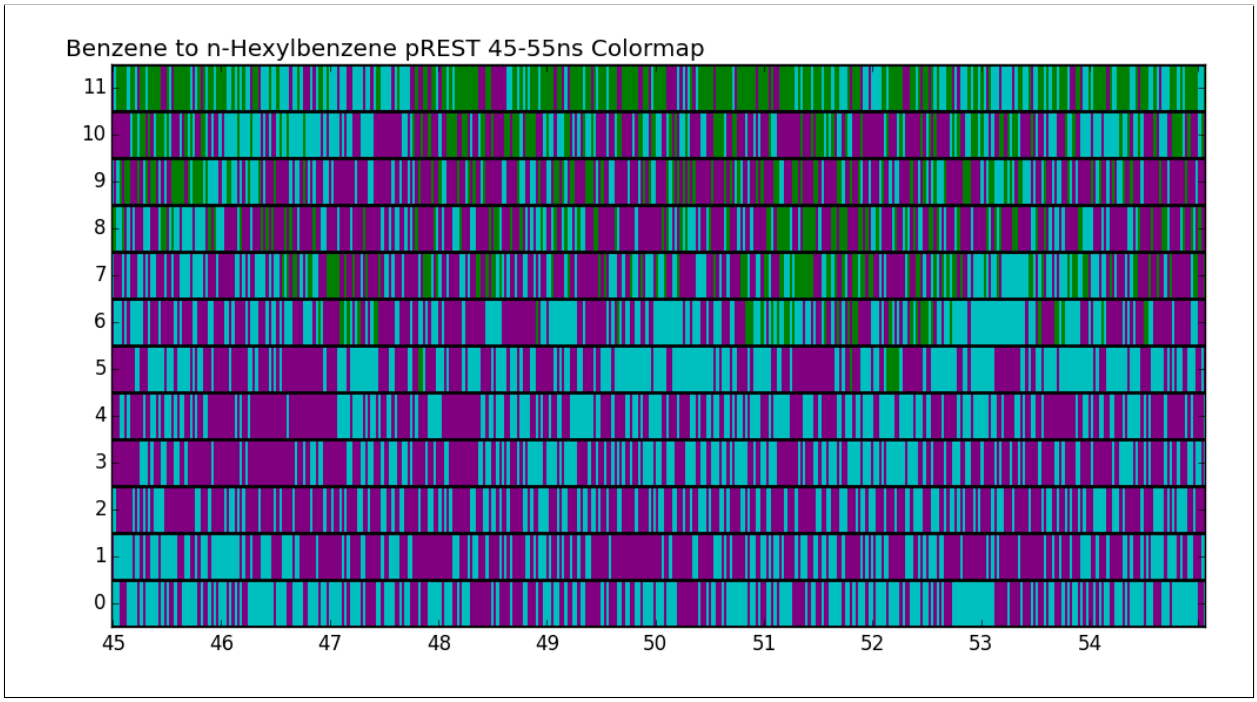


(b) Closed - Benzene to n-Hexylbenzene 45-55ns RMSD Replica11 pREST

Figure 6: Benzene To n-Hexylbenzene (pREST)

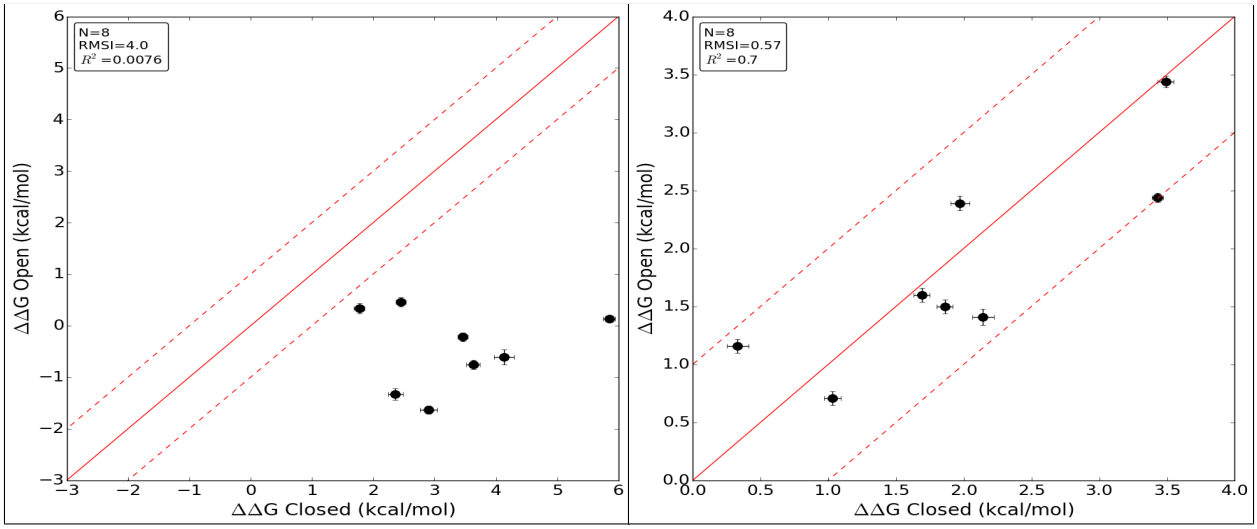


(a) Default (b) pREST



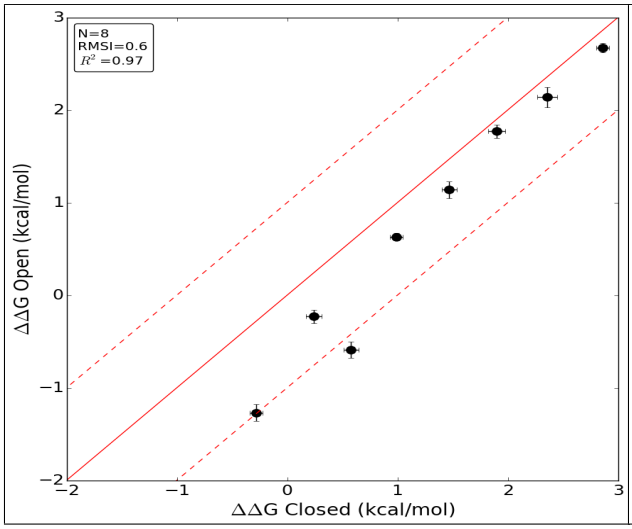
(c) Closed - Benzene to n-Hexylbenzene 45-55ns Colormap pREST

Figure 7: Benzene To n-Hexylbenzene (pREST) Colormap

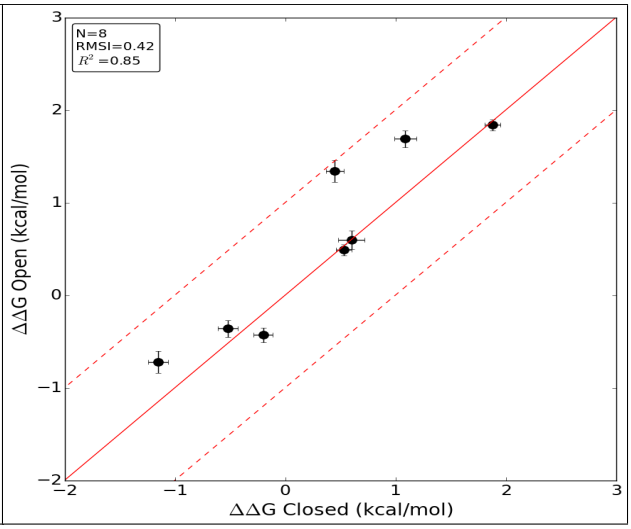


(a) Closed-Open XYplot (Default)

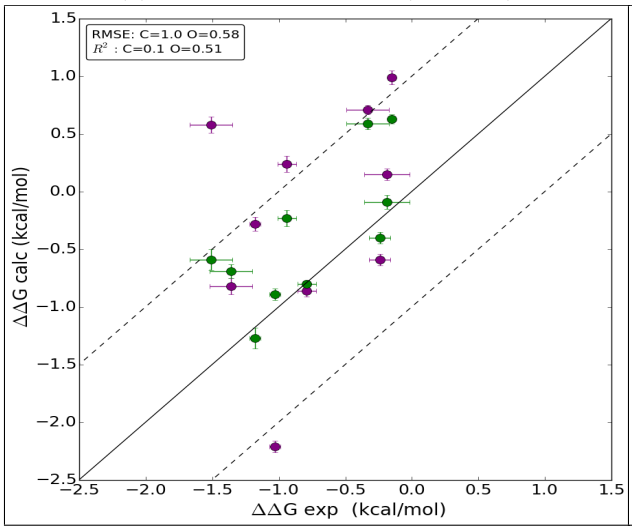
(b) Closed-Open XYplot (pREST)



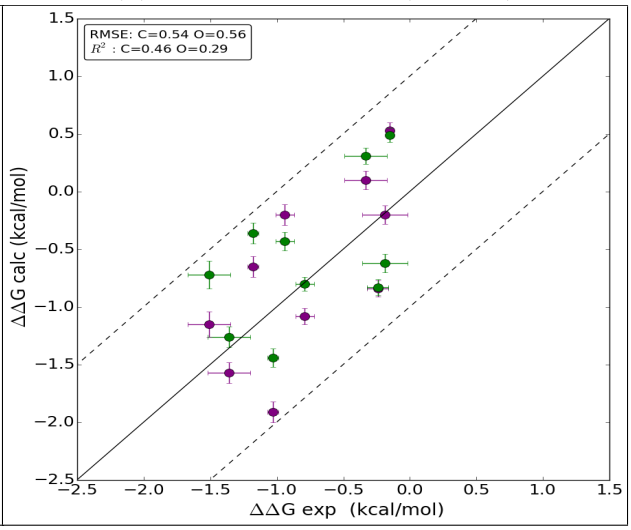
(c) Closed-Int XYplot (Default)



(d) Closed-Int XYplot (pREST)



(e) Exp XYplot (Default)



(f) Closed-Int XYplot (pREST)

Figure 8: Conf-XY

Tables

Table 1: Loop Occupancies & Binding Affinities[1,16]

PDB	Ligand	C	I	O	ΔG_{exp}	σ_{exp}
4W52	benzene	0.9	-	-	-5.19	0.16
4W53	toluene	0.8	0.2	-	-5.52	0.04
4W54	ethylbenzene	0.5	0.5	-	-5.76	0.07
4W55	n-propylbenzene	0.6	0.4	-	-6.55	0.02
4W56	sec-butylbenzene	0.4	0.6	-	N/A	-
4W57	n-butylbenzene	0.1	0.6	0.3	-6.70	0.02
4W58	n-pentylbenzene	0.3	-	0.7	N/A	-
4W59	n-hexylbenzene	0.3	-	0.7	N/A	-

Table 2: Closed-Intermediate Transformations

Ligand 1	Ligand 2	$\Delta\Delta G_C$	σ_C	$\Delta\Delta G_O$	σ_O	$\Delta\Delta G_\varepsilon$
benzene	n-butylbenzene	0.58	*0.07*	-0.59	*0.09*	1.17
toluene	n-butylbenzene	-0.28	*0.06*	-1.27	*0.09*	0.99
ethylbenzene	n-butylbenzene	0.24	*0.07*	-0.23	*0.07*	0.47
n-propylbenzene	n-butylbenzene	0.99	*0.06*	0.63	*0.04*	0.36
benzene	sec-butylbenzene	2.36	*0.09*	2.14	*0.11*	0.22
toluene	sec-butylbenzene	1.47	*0.07*	1.14	*0.09*	0.33
ethylbenzene	sec-butylbenzene	1.90	*0.08*	1.77	*0.07*	0.13
n-propylbenzene	sec-butylbenzene	2.86	*0.06*	2.67	*0.05*	0.19

Table 3: Closed-Intermediate Transformations pREST

Ligand 1	Ligand 2	$\Delta\Delta G_C$	σ_C	$\Delta\Delta G_O$	σ_O	$\Delta\Delta G_\varepsilon$
benzene	n-butylbenzene	-0.10	*0.11*	-0.72	*0.12*	0.62
toluene	n-butylbenzene	0.90	*0.09*	-0.36	*0.09*	1.26
ethylbenzene	n-butylbenzene	-0.20	*0.09*	-0.43	*0.08*	0.23
n-propylbenzene	n-butylbenzene	1.00	*0.07*	0.49	*0.06*	0.51
benzene	sec-butylbenzene	0.45	*0.08*	1.34	*0.12*	0.89
toluene	sec-butylbenzene	0.60	*0.12*	0.60	*0.10*	0.0
ethylbenzene	sec-butylbenzene	1.09	*0.10*	1.69	*0.09*	0.60
n-propylbenzene	sec-butylbenzene	1.88	*0.07*	3.05	*0.08*	1.17

Table 4: Closed-Intermediate Transformations pREST 15-25ns

Ligand 1	Ligand 2	$\Delta\Delta G_C$	σ_C	$\Delta\Delta G_O$	σ_O	$\Delta\Delta G_\epsilon$
benzene	n-butylbenzene	-1.15	*0.09*	-0.72	*0.12*	0.43
toluene	n-butylbenzene	-0.52	*0.09*	-0.36	*0.09*	0.16
ethylbenzene	n-butylbenzene	-0.20	*0.09*	-0.43	*0.08*	0.23
n-propylbenzene	n-butylbenzene	0.53	*0.07*	0.49	*0.06*	0.04
benzene	sec-butylbenzene	0.45	*0.08*	1.34	*0.12*	0.89
toluene	sec-butylbenzene	0.60	*0.12*	0.60	*0.10*	0.0
ethylbenzene	sec-butylbenzene	1.09	*0.10*	1.69	*0.09*	0.60
n-propylbenzene	sec-butylbenzene	1.88	*0.07*	1.84	*0.06*	0.04

Table 5: Closed-Open Transformations

Ligand 1	Ligand 2	$\Delta\Delta G_C$	σ_C	$\Delta\Delta G_O$	σ_O	$\Delta\Delta G_\epsilon$
benzene	n-pentylbenzene	2.36	*0.12*	-1.33	*0.11*	3.69
toluene	n-pentylbenzene	1.77	*0.09*	0.34	*0.10*	1.43
ethylbenzene	n-pentylbenzene	2.45	*0.08*	0.46	*0.09*	1.99
n-propylbenzene	n-pentylbenzene	3.46	*0.08*	-0.22	*0.08*	3.68
benzene	n-hexylbenzene	4.13	*0.16*	-0.61	*0.15*	4.74
toluene	n-hexylbenzene	2.90	*0.14*	-1.63	*0.08*	4.53
ethylbenzene	n-hexylbenzene	3.63	*0.11*	-0.76	*0.09*	4.39
n-propylbenzene	n-hexylbenzene	5.85	*0.10*	0.13	*0.06*	5.72

[a] Some text; [b] Some more text.

Table 6: Closed-Open Transformations pREST

Ligand 1	Ligand 2	$\Delta\Delta G_C$	σ_C	$\Delta\Delta G_O$	σ_O	$\Delta\Delta G_\epsilon$
benzene	n-pentylbenzene	1.45	*0.13*	0.15	*0.10*	1.30
toluene	n-pentylbenzene	1.40	*0.13*	0.82	*0.11*	0.58
ethylbenzene	n-pentylbenzene	2.89	*0.10*	1.32	*0.10*	1.57
n-propylbenzene	n-pentylbenzene	4.40	*0.12*	1.06	*0.09*	3.34
benzene	n-hexylbenzene	2.74	*0.19*	1.37	*0.13*	1.37
toluene	n-hexylbenzene	3.21	*0.15*	-1.08	*0.09*	4.29
ethylbenzene	n-hexylbenzene	3.39	*0.11*	-0.14	*0.10*	3.53
n-propylbenzene	n-hexylbenzene	4.93	*0.12*	1.28	*0.10*	3.65

[a] Some text; [b] Some more text.

Table 7: Closed-Open Transformations pREST 40-55ns

Ligand 1	Ligand 2	$\Delta\Delta G_C$	σ_C	$\Delta\Delta G_O$	σ_O	$\Delta\Delta G_\varepsilon$
benzene	n-pentylbenzene	1.86	*0.06*	1.50	*0.06*	0.36
toluene	n-pentylbenzene	1.03	*0.06*	0.71	*0.06*	0.32
ethylbenzene	n-pentylbenzene	1.69	*0.06*	1.60	*0.06*	0.09
n-propylbenzene	n-pentylbenzene	3.43	*0.04*	2.44	*0.04*	0.99
benzene	n-hexylbenzene	2.14	*0.08*	1.41	*0.07*	0.73
toluene	n-hexylbenzene	0.33	*0.08*	1.16	*0.06*	0.84
ethylbenzene	n-hexylbenzene	1.97	*0.07*	2.39	*0.06*	0.42
n-propylbenzene	n-hexylbenzene	3.49	*0.06*	3.44	*0.05*	0.05

[a] Some text; [b] Some more text.

Table 8: Exp Set

Ligand 1	Ligand 2	ΔG_{exp}	σ_{exp}	$\Delta\Delta G_C$	σ_C	$\Delta\Delta G_O$	σ_O
benzene	toluene	-0.33	0.16	0.71	0.04	0.59	0.05
benzene	ethylbenzene	-0.19	0.17	0.15	0.05	-0.09	0.06
benzene	n-propylbenzene	-1.36	0.16	-0.82	0.07	0.06	0.06
toluene	ethylbenzene	-0.24	0.08	-0.59	0.05	0.05	0.05
toluene	n-propylbenzene	-1.03	0.04	-2.21	0.05	0.05	0.05
ethylbenzene	n-propylbenzene	-0.79	0.07	-0.86	0.05	0.03	0.03
benzene	n-butylbenzene	-1.51	0.16	0.58	0.07	-0.59	0.09
toluene	n-butylbenzene	-1.18	0.04	-0.28	0.06	0.09	0.09
ethylbenzene	n-butylbenzene	-0.94	0.07	0.24	0.07	-0.23	0.07
n-propylbenzene	n-butylbenzene	-0.15	0.03	0.99	0.06	0.63	0.04

Table 9: exp set pREST

Ligand 1	Ligand 2	ΔG_{exp}	σ_{exp}	$\Delta\Delta G_C$	σ_C	$\Delta\Delta G_O$	σ_O
benzene	toluene	-0.33	0.16	0.10	0.08	0.31	0.07
benzene	ethylbenzene	-0.19	0.17	-0.20	0.08	-0.62	0.08
benzene	n-propylbenzene	-1.36	0.16	-1.57	0.09	-1.26	0.09
toluene	ethylbenzene	-0.24	0.08	-0.84	0.07	-0.83	0.07
toluene	n-propylbenzene	-1.03	0.04	-1.91	0.09	-1.44	0.08
ethylbenzene	n-propylbenzene	-0.79	0.07	-1.08	0.07	-0.80	0.06
benzene	n-butylbenzene	-1.51	0.16	-1.15	0.11	-0.72	0.12
toluene	n-butylbenzene	-1.18	0.04	-0.65	0.09	-0.36	0.09
ethylbenzene	n-butylbenzene	-0.94	0.07	-0.20	0.09	-0.43	0.08
n-propylbenzene	n-butylbenzene	-0.15	0.03	0.53	0.07	0.49	0.06

References

(1) Merski, M.; Fischer, M.; Balius, T. E.; Eidam, O.; Shoichet, B. K. *Proceedings of the National Academy of Sciences* **2015**, *112*, 5039–5044.

(2) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. *Nature reviews Drug discovery* **2006**, *5*, 993–996.

(3) Landry, Y.; Gies, J.-P. *Fundamental & Clinical Pharmacology* **2008**, *22*, 1–18.

(4) Lundstrom, K. In *G Protein-Coupled Receptors in Drug Discovery*; Leifert, R. W., Ed.; Humana Press: Totowa, NJ, 2009; Chapter An Overview on GPCRs and Drug Discovery: Structure-Based Drug Design and Structural Biology on GPCRs, pp 51–66.

(5) Homeyer, N.; Stoll, F.; Hillisch, A.; Gohlke, H. *Journal of Chemical Theory and Computation* **2014**, *10*, 3331–3344, PMID: 26588302.

(6) Chipot, C.; Pohorille, A. *Free energy calculations*; Springer, 2007.

(7) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. *Current opinion in structural biology* **2011**, *21*, 150–160.

(8) Knight, J. L.; Brooks, C. L. *Journal of computational chemistry* **2009**, *30*, 1692–1700.

(9) Zheng, L.; Chen, M.; Yang, W. *Proceedings of the National Academy of Sciences* **2008**, *105*, 20227–20232.

(10) Gallicchio, E.; Levy, R. M. *Current opinion in structural biology* **2011**, *21*, 161–166.

(11) Hansen, N.; van Gunsteren, W. F. *Journal of Chemical Theory and Computation* **2014**, *10*, 2632–2647, PMID: 26586503.

(12) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.;

Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. *Journal of the American Chemical Society* **2015**, *137*, 2695–2703, PMID: 25625324.

(13) Wang, L.; Berne, B. J.; Friesner, R. A. *Proceedings of the National Academy of Sciences* **2012**, *109*, 1937–1942.

(14) Eriksson, A. E.; Baase, W. A.; Zhang, X.-J.; Heinz, D. W.; Blaber, M.; Baldwin, E. P.; Matthews, B. W. *Science* **1992**, *255*, 178–183.

(15) Eriksson, A.; Baase, W. A.; Matthews, B. W. *Journal of molecular biology* **1993**, *229*, 747–769.

(16) Morton, A.; Baase, W. A.; Matthews, B. W. *Biochemistry* **1995**, *34*, 8564–8575, PMID: 7612598.

(17) Morton, A.; Matthews, B. W. *Biochemistry* **1995**, *34*, 8576–8588, PMID: 7612599.

(18) Wei, B. Q.; Baase, W. A.; Weaver, L. H.; Matthews, B. W.; Shoichet, B. K. *Journal of molecular biology* **2002**, *322*, 339–355.

(19) Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. *Journal of molecular biology* **2004**, *337*, 1161–1182.

(20) Graves, A. P.; Brenk, R.; Shoichet, B. K. *Journal of medicinal chemistry* **2005**, *48*, 3714–3728.

(21) Mobley, D. L.; Graves, A. P.; Chodera, J. D.; McReynolds, A. C.; Shoichet, B. K.; Dill, K. A. *Journal of Molecular Biology* **2007**, *371*, 1118 – 1134.

(22) Hermans, J.; Wang, L. *Journal of the American Chemical Society* **1997**, *119*, 2707–2714.

(23) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. *The Journal of Physical Chemistry B* **2003**, *107*, 9535–9551.

(24) Deng, Y.; Roux, B. *Journal of Chemical Theory and Computation* **2006**, *2*, 1255–1273.

(25) Mann, G.; Hermans, J. *Journal of molecular biology* **2000**, *302*, 979–989.

(26) Boyce, S. E.; Mobley, D. L.; Rocklin, G. J.; Graves, A. P.; Dill, K. A.; Shoichet, B. K. *Journal of Molecular Biology* **2009**, *394*, 747 – 763.

(27) Wang, L.; Deng, Y.; Knight, J. L.; Wu, Y.; Kim, B.; Sherman, W.; Shelley, J. C.; Lin, T.; Abel, R. *Journal of Chemical Theory and Computation* **2013**, *9*, 1282–1293, PMID: 26588769.

(28) Schrödinger, LLC, Maestro, version 10.3. 2015.

(29) Schrödinger, LLC, Schrödinger Suite 2015-3 Protein Preparation Wizard. 2015.

(30) Schrödinger, LLC, Epik, version 3.3. 2015.

(31) Schrödinger, LLC, Impact, version 6.8. 2015.

(32) Schrödinger, LLC, Prime, version 4.1. 2015.

(33) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 221–234.

(34) Lim, N. M. Schrödinger Academy Molecular Dynamics Ligand FEP Tutorial. Schrödinger: New York, NY, 2015.

(35) Schrödinger, LLC, Maestro-Desmond Interoperability Tools, version 4.3. 2015.

(36) Liu, S.; Wu, Y.; Lin, T.; Abel, R.; Redmann, J. P.; Summa, C. M.; Jaber, V. R.; Lim, N. M.; Mobley, D. L. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 755–770.

(37) Liu, P.; Kim, B.; Friesner, R. A.; Berne, B. J. *Proceedings of the National Academy of Sciences of the United States of America* **2005**, *102*, 13749–13754.

(38) Wang, L.; Friesner, R. A.; Berne, B. J. *The Journal of Physical Chemistry B* **2011**, *115*, 9431–9438, PMID: 21714551.

(39) D.E. Shaw Research, Desmond Molecular Dynamics System, version 4.3. 2015.

(40) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. Proceedings of the 2006 ACM/IEEE Conference on Supercomputing. New York, NY, USA, 2006.

(41) Shivakumar, D.; Williams, J.; Wu, Y.; Damm, W.; Shelley, J.; Sherman, W. *Journal of Chemical Theory and Computation* **2010**, *6*, 1509–1519, PMID: 26615687.

(42) Guo, Z.; Mohanty, U.; Noehre, J.; Sawyer, T. K.; Sherman, W.; Krilov, G. *Chemical Biology & Drug Design* **2010**, *75*, 348–359.

(43) D.E. Shaw Research, Desmond Users Guide. version 3.6.1.1/0.8.

(44) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. *The Journal of Chemical Physics* **1995**, *103*, 4613–4621.

(45) Banks, J. L.; Beard, H. S.; Cao, Y.; Cho, A. E.; Damm, W.; Farid, R.; Felts, A. K.; Halgren, T. A.; Mainz, D. T.; Maple, J. R.; Murphy, R.; Philipp, D. M.; Repasky, M. P.; Zhang, L. Y.; Berne, B. J.; Friesner, R. A.; Gallicchio, E.; Levy, R. M. *Journal of Computational Chemistry* **2005**, *26*, 1752–1780.

(46) Harder, E.; Damm, W.; Maple, J.; Wu, C.; Reboul, M.; Xiang, J. Y.; Wang, L.; Lupyan, D.; Dahlgren, M. K.; Knight, J. L.; Kaus, J. W.; Cerutti, D. S.; Krilov, G.; Jor-

gensen, W. L.; Abel, R.; Friesner, R. A. *Journal of Chemical Theory and Computation* **2016**, *12*, 281–296, PMID: 26584231.

(47) Bennett, C. H. *Journal of Computational Physics* **1976**, *22*, 245–268.

(48) Hahn, A. M.; Then, H. *Physical Review E* **2009**, *80*, 031111.

(49) Humphrey, W.; Dalke, A.; Schulten, K. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(50) Eargle, J.; Wright, D.; Luthey-Schulten, Z. *Bioinformatics* **2006**, *22*, 504–506.

(51) Bowman, G. R.; Pande, V. S.; Noé, F. *An introduction to markov state models and their application to long timescale molecular simulation*; Springer Science & Business Media, 2013; Vol. 797.