Published in final edited form as:
J Chem Theory Comput. 2012 October 9; 8(10): 3513–3525. doi:10.1021/ct300088r.

Site-Specific Fragment Identification Guided by Single-Step Free Energy Perturbation Calculations

E. Prabhu Raman, **Kenno Vanommeslaeghe**, and **Alexander D. MacKerell Jr.***
Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, 20 Penn Street HSF II, Baltimore MD 21201

Abstract

The *in-silico* Site Identification by Ligand Competitive Saturation (SILCS) approach identifies the binding sites of representative chemical entities on the entire protein surface, information that can be applied for computational fragment-based drug design. In this study, we report an efficient computational protocol that uses sampling of the protein-fragment conformational space obtained from the SILCS simulations and performs single step free energy perturbation (SSFEP) calculations to identify site-specific favorable chemical modifications of benzene involving substitutions of ring hydrogens with individual non-hydrogen atoms. The SSFEP method is able to capture the experimental trends in relative hydration free energies of benzene analogues and for two datasets of experimental relative binding free energies of congeneric series of ligands of the proteins α-thrombin and P38 MAP kinase. The approach includes a protocol in which data obtained from SILCS simulations of the proteins is first analyzed to identify favorable benzene binding sites following which an ensemble of benzene-protein conformations for that site is obtained. The SSFEP protocol applied to that ensemble results in good reproduction of experimental free energies of the α-thrombin ligands, but not for P38 MAP kinase ligands. Comparison with results from a P38 full-ligand simulation and analysis of conformations reveals the reason for the poor agreement being the connectivity with the remainder of the ligand, a limitation inherent in fragment-based methods. Since the SSFEP approach can identify favorable benzene modifications as well as identify the most favorable fragment conformations, the obtained information can be of value for fragment linking or structure-based optimization.

INTRODUCTION

Fragment based drug discovery (FBDD) has emerged as a promising alternative to high throughput screening (HTS) for the discovery of high affinity inhibitors.[1] Compared to HTS, by identifying compounds that can ultimately be modified or linked into higher affinity inhibitors, FBDD potentially provides more efficient coverage of chemical space while screening a smaller number of candidate molecules.[1] The first step in FBDD involves the detection of low molecular weight compounds (~ 150 Da) bound to the target protein surface.[2] The small compounds, or fragments, act as the starting point for the application of structure-based approaches to develop novel lead compounds. This may be achieved by either decorating the fragment with functional groups or linking fragments bound to neighboring sites on the target to improve the binding affinity. However, for any of these approaches, atomic detail information of the protein-fragment complex is required,

*Corresponding author: Alexander D. MacKerell, Jr., Room 629, HSF II, 20 Penn Street, Baltimore, MD 21201, Tel: 410-706-7442, Fax: 410-706-5017, alex@outerbanks.umaryland.edu.

Supporting Information Available
Supporting information contains the figures and tables referred to in the paper. This information is available free of charge via the Internet at http://pubs.acs.org/.

information that can be difficult to obtain due to weak binding affinity and inherent limitations of X-ray crystallography or NMR spectroscopy.[3]

Computational methods represent successful alternatives to experimental approaches to drug discovery and design. Docking based virtual screening has been used to effectively initiate a number of drug discovery campaigns, though it is limited in that it relies on pre-existing compounds. De novo drug design on the other hand involves the creation of novel chemical entities, with fragment-based methods representing the starting point for most de novo design strategies. In these approaches the type and location of fragments binding to the protein surface are detected,[3-7] followed by the linking of those fragments that bind to neighboring sites on a protein.[8,9] Towards this end computational fragment docking has shown great potential and recent developments[10] have moved beyond the traditional limitations of the method[11,12] associated with the use of a rigid protein and absence of aqueous solvation, among others.

The SILCS methodology[3,6] is an approach developed in our laboratory that involves MD simulations of the target protein in an aqueous solution of organic molecules representative of fragments of more complex drug-like molecules. In SILCS, flexibility of the protein and fragments is included explicitly as is the aqueous environment allowing exhaustive MD simulations to yield an ensemble of the distribution of the fragments and of water on the protein surface. This ensemble, in combination with control simulations in the absence of the target protein allows for generation of normalized 3D “FragMaps” that identify the favorable locations of different functional groups on the entire protein surface. Conversion of the FragMaps to free energies, based on a Boltzmann distribution, yields Grid Free Energies (GFEs)[13] that may be used to calculate free-energy contributions of fragments to ligand binding. The success of the approach was seen in the overlap of FragMaps/low energy regions of GFEs of fragments with crystallographic positions of functional groups of similar chemical type in both peptide-protein and inhibitor-protein complexes.

An inherent limitation in the SILCS methodology is the limited number of ligand types that can be included in the MD simulations. In published studies to date only propane and benzene have been included, along with water, limiting the information content from SILCS to aliphatic, aromatic, hydrogen bond donor and hydrogen bond acceptor functional classes. While ongoing studies in our laboratory indicate that a range of other small ligands may be used, and other ligands have been used in similar studies[13,14], this represents a significant limitation. Here we address this limitation by testing if the structural ensemble obtained from SILCS simulations of a protein in the presence of a limited set of fragments be used to allow for estimates of the change in the binding free energy associated with modifications of those fragments, such that the relative affinities of a wide range of fragments can be rapidly predicted. If this can be achieved then the number of possible fragments that can be predicted to bind favorably to a protein site can be significantly increased, thereby increasing the utility of SILCS in a *de-novo* FBDD strategy.

To achieve this we implemented a Single Step Free Energy Perturbation (SSFEP) method to identify site-specific favorable modifications to fragments thereby extending the SILCS methodology. The strategy is motivated by the “One Step Perturbation” method,[15,16] an approach that has been used for the calculation of relative hydration free energies[17] and relative binding free energies of drug-like molecules involving differences of many non-hydrogen atoms.[18,19] The present procedure involves obtaining a conformational ensemble of fragments from SILCS calculations, rather than using a fictitious reference compound that involves “soft-core” interactions,[15,16] and using that ensemble in conjunction with the free energy perturbation formula[20] to estimate the free energy change caused due to a chemical modification of the fragments used in the initial SILCS simulation. Target data for

validation include experimental relative hydration free energies of benzene analogues and relative binding free energies of drug-sized molecules containing a substituted phenyl ring.

METHODS

MD simulations

All simulations were performed using the CHARMM molecular simulation program,[21] the CHARMM protein force field[22] with CMAP backbone correction[23], and the TIP3P water model.[24] The CHARMM general force field (CGenFF)[25] version 2b5 was used for all ligands. The CGenFF program[26] version 0.9.1 beta (accessible through the ParamChem interface[27]) was used to obtain the topology, charges and initial guess parameters for the two parent inhibitors of thrombin and P38MK containing the unsubstituted phenyl group. These initial parameters were further optimized and validated per the CGenFF procedure which uses QM energies and geometries as target data.[25] For simulations involving proteins, any crystal water molecules present in the PDB coordinates were retained, as were any structurally important ions. The Reduce software[28] was used to choose optimal Asn and Gln side chain amide and His side chain ring orientations and CHARMM was used to add hydrogen atoms. Solvated orthorhombic periodic systems were generated by overlaying the crystal coordinates of the protein with a pre-equilibrated water box the dimensions of which were 10 Å longer than the maximum dimensions of the protein along each of the three orthogonal axes. All non-crystallographic water molecules with any atom within 2 Å of any protein atom were deleted. The net system charge was made neutral by replacing random water molecules with the appropriate number of sodium or chloride ions. For thrombin, the missing residues of the protein were built and the protocol for protein preparation was slightly different and it involved the deletion of crystal waters also based on the 2 Å cutoff as detailed in our previous work.[6] The present setup of the SILCS simulations is very similar to that reported in detail previously.[3,6] In short, the simulations involve NPT MD sampling of a protein in a solution of benzene and propane each present at 1M concentration. 10 trajectories, each of length 10-20 ns, are performed with each trajectory being different in the initial positioning of the fragments in the simulation box. To prevent the aggregation of hydrophobic molecules, an additional inter-fragment repulsive term is added to the potential, with that potential being linked to the nonbond cutoff of 8 Å due to the use of a particle-specific Lennard-Jones interaction to implement the respulsion.[3] To obtain the solvation free energy of benzene analogues, a 10ns simulation of benzene in a water box of dimensions 32 Å × 32 Å × 32 Å was performed with the benzene molecule restrained to the center of the box using a center of mass restraint of 0.5 kcal*mol^{-1}Å^{-2}.

Unless otherwise mentioned, all systems in the presence of periodic boundary conditions were minimized for 500 steps with the steepest descent algorithm[29] while employing harmonic positional restraints with a force constant of 1 kcal*mol^{-1}Å^{-2} per atomic mass unit on protein non-hydrogen atoms. The leapfrog variant of the Verlet integrator[30] with a time step of 2 fs was used for molecular dynamics (MD) simulations. Water geometries and bonds involving hydrogen atoms were constrained using the SHAKE algorithm.[31] Long-range electrostatic interactions were handled with the particle-mesh Ewald method[32] with a real space cutoff of 8 Å. For PME a kappa value of 0.32 was used and the order of B-spline interpolation was 6. The grid spacing was set to ≈ 1 Å. A switching function[33] was applied to the Lennard-Jones interactions in the range of 5 to 8 Å, and a long-range isotropic correction[30] was applied to the energy and pressure for Lennard-Jones interactions beyond the cutoff length. Following minimization the system was heated with the same positional restraints over 10 ps to 298 K by periodic reassignment of velocities,[34] followed by an equilibration for 10 ps using velocity reassignment. In the production simulations that followed, unless otherwise indicated, the positional restraints were replaced by weak restraints on only the protein backbone Cα atoms with a force constant of 0.01

kcal*mol^{-1}A^{-2} per atomic mass unit to prevent the rotation of the protein in the simulation box. Temperature and pressure were maintained at 298 K and 1 atm with a Nosé-Hoover thermostat[35,36] and Langevin piston barostat,[37] respectively. Snapshots were output every 2ps for analysis for the protein-ligand MD simulations. For the SILCS simulations, a larger number of trajectories was obtained and the snapshot output frequency was 5ps. For the benzene+water system, a force-switching function was applied to the Lennard-Jones interactions in the range of 10 to 12 Å and the real space cutoff for PME electrostatics was 12 Å. The system was minimized for 5000 steps with the steepest descent algorithm[29] while employing harmonic positional restraints with a force constant of 1 kcal*mol^{-1}Å^{-2} on benzene atoms. Following minimization the system was equilibrated with the same positional restraints for 1ns using velocity reassignment followed by a 10ns production simulation in the NPT ensemble with snapshots output for analysis every 2 ps.

Identification of benzene binding sites

From the SILCS simulation of α-thrombin, benzene carbon atoms less than 5 Å from any protein atom were binned into a 3D grid or "FragMap" composed of 1Å × 1Å × 1Å volume elements and the FragMap probability grid was Boltzmann transformed into the grid free energy (GFE).[3,6] The centers of grid elements having a GFE value lower than −1.2 kcal/mol were clustered to identify binding sites of benzene on the protein surface using the following algorithm. An arbitrarily chosen grid center point was assigned to the first cluster and thereafter, each grid element was either assigned to an existing cluster if its center was located closer in Euclidian space than the cluster radius value of 5Å to that cluster or a new cluster was created otherwise. After the inclusion of each element in a cluster, the cluster center was recomputed as the mean of the coordinates of the members. Following the initial assignment, an iterative loop was run, which would redo the cluster assignment based on the distance from the existing cluster centers. The iteration was terminated once no more updates of the cluster assignment occurred; typically only one or two iterations were required.

Single step perturbation calculations

The alchemical free energy difference of transforming a ligand *L1* to *L2* in environment *E* for each of the sites identified using the clustering algorithm is computed per the perturbation formula[20] as follows:

$$\Delta G^{E}_{L1\rightarrow L2} = -RT \quad \ln\left\langle \exp^{-\left(E_{L2}-E_{L1}\right)/RT} \right\rangle_{L1} \quad (1)$$

where RT (*=0.592* kcal/mol) is the product of the ideal gas constant and the absolute temperature and E_{L1} and E_{L2} are the ligand energies. The average is computed over the ensemble of conformations obtained from the simulation of ligand L1 in environment E. The energy of a ligand X and its environment is decomposed into the following terms:

$$E_X = E_{XX} + E_{XE} + E_{EE} \quad (2)$$

where E_{XE} is the nonbonded interaction energy between ligand X and environment E and E_{XX} is the internal energy of ligand X. The self-energy of the environment E_{EE} cancels when computing the energy difference between two ligands as the precalculated ensemble of conformations of the protein and solution from the SILCS simulations are identical. The relative solvation and binding free energies computed in this work are given as follows in Eqns 3 and 4 respectively:

$$\Delta\Delta G^{solv}_{L1\rightarrow L2}=\Delta G^{water}_{L1\rightarrow L2}-\Delta G^{vacuum}_{L1\rightarrow L2} \quad (3)$$

$$\Delta\Delta G^{bind}_{L1\rightarrow L2}=\Delta G^{protein}_{L1\rightarrow L2}-\Delta G^{water}_{L1\rightarrow L2} \quad (4)$$

Where, $\Delta\Delta G^{E}_{L1\rightarrow L2}$ is the alchemical free energy difference computed in environment E per Equation 1. In the present work, $L1$ is always benzene and $L2$ is one of the 8 monosubstituted benzene analogues. The test set included several ligands in which the phenyl ring could assume two possible orientations in the binding pocket due to the rotation about the bond linking the phenyl ring to the rest of the ligand. Since the SSFEP calculations do not allow for rotation of the phenyl ring, the relative free energies of binding were combined using the following equation.[38]

$$\Delta\Delta G=-RT\ln\lfloor\exp^{-\Delta\Delta G^{bind}_{O1}/RT}+\exp^{-\Delta\Delta G^{bind}_{O2}/RT}\rfloor+ \quad (5)$$

In equation 5 the subscripts $O1$ and $O2$ indicate the two different ring orientations. For the ligands in the test set that involved two substitutions on the phenyl ring, the free energy difference was obtained by summing the relative free energies computed for the individual single substitution analogues. Strictly speaking, this is an approximation because the contributions are not additive, but its utility is demonstrated by the observation that it reproduces the experimental trend, consistent with previous studies.[39]

Simulations to evaluate the free energy difference between benzene and its analogues using SSFEP were set up and carried out as follows. In order to mimic the phenyl ring on a larger inhibitor, where the ring is not free to rotate, in the anisotropic protein environment, it is necessary to distinguish the 6 possible substitution positions on the benzene ring. This was accomplished by first choosing a reference conformation of benzene in the environment. In the case of the two studied proteins, results are reported with the reference conformation being the crystal conformation of the phenyl ring of the corresponding parent inhibitor (ATI and MKI). In the results section, we show that the choice of reference conformation does not influence the results significantly. For each snapshot, the rotation of the benzene ring was neglected and the carbon atoms were renamed (without altering the coordinates themselves) so as to have the minimum possible RMSD with respect to the reference conformation, where the RMSD is sensitive to the label of each carbon atom. This results in orientation #1 of the substituted benzene, which is “aligned” to the reference conformation. The 5 additional orientations (i.e. with the substituent at positions 2 through 6) are subsequently generated resulting in a total of 6 orientations for each snapshot. This approach is necessary because if a given position (e.g. position 1) of the benzene ring was assigned a new atom type at the beginning of the trajectory and maintained throughout the trajectory, it is highly likely that the benzene ring would rotate such that position 1 on the ring would now occupy the location on the protein surface previously occupied by one of the other 5 positions, which cannot occur with a phenyl group that is part of a larger bound ligand. It is worth restating that the coordinates are not in any way altered in generating these orientations, only the label of each atom is changed so that in the subsequent alignment step the six possible orientations obtained. Only the ring carbon atoms are considered in the RMSD computation. Following this step, the precomputed energy-minimized conformations of the respective benzene analogs are aligned to the benzene conformation from the SILCS simulation or from the protein-ligand simulation. For aniline, the planar nitrogen conformation was chosen instead of the slightly pyramidized conformation (which is slightly more stable). Figure 1 illustrates the alignment procedure by displaying the generated conformations of fluorobenzene from the analysis of the benzene conformational distribution obtained from

the SILCS simulations of α-thrombin. The 20 most favorable conformations in each of the six orientations of the ligand are depicted with the fluorine substituent colored differently in each orientation. As expected, a broad distribution of the substituents is observed, which is centered approximately at the six substitution positions on the phenyl ring and partially overlaps with the neighboring substituent distributions. In the case of phenol, two conformations that differ in the position of the alcohol hydrogen atom were generated for each of the 6 orientations, resulting in 12 geometries to be evaluated. By using previously energy-minimized analogues, one does not consider contributions from slight deviations from planarity of the benzene observed in MD simulations to the calculated free energy differences. We assume that contributions from such minor deformations cancel out when calculating free energy differences.

All energy computations on the composite ligand-environment snapshots were performed using an in-house post-processing routine involving CHARMM.[21] The nonbonded interaction energy between the ligand and the environment was computed with a cutoff of 29 Å. In the range of 28 to 29 Å, a force-switching function was applied to the electrostatic and the Lennard-Jones interactions. Periodic images were re-built in the post-processing routine and were included in the calculations. Other than the explicit calculation of the pairwise non-bonded interactions, long range electrostatic or Lennard-Jones correction terms were not considered. This treatment of the nonbond interactions was applied for all the SSFEP energy calculations although the truncation schemes in the MD simulations to generate the ensembles in protein and in solution were different (see above). As a check, we performed a second set of simulation of benzene in water that used the same non-bonded truncation scheme as in the protein simulations and found no significant difference in the free energy values (data not shown).

Thermodynamic integration calculations

Thermodynamic Integration (TI) calculations were performed using the PERT module in CHARMM[21] to obtain the relative hydration free energies of benzene analogues in order to check for any dependence of the results on the force field. The system setup for the alchemical transformation in solution involved the same dynamics parameters as used for the benzene-water MD simulation described above. For each transformation, both forward and backward perturbations were performed using 22 λ-windows, each being 100 ps long including a 50ps equilibration period. All solute and water bonds were held fixed using the SHAKE algorithm.[31] Transformations in vacuum were performed with infinite nonbonded cutoffs and involved 22 λ-windows, each being 20ps long including a 4ps equilibration period.

Analyses

Computed $\Delta\Delta G$ values are compared with experimental values using correlation plots and computing R^2 values of linear regression. In order to quantify the ability of the method to rank order ligands by binding free energy, we use the Predictive Index (PI) metric developed by Pearlman and Charifson.[40]

$$\begin{aligned} PI &= \frac{\sum_{j>i}\sum_{i} W_{ij}C_{ij}}{\sum_{j>i}\sum_{i} W_{ij}} \\ W_{ij} &= |E(j) - E(i)| \\ C_{ij} &= \begin{cases} -1 & \text{if } [E(j) - E(i)] / [P(j) - P(i)] < 0 \\ 1 & \text{if } [E(j) - E(i)] / [P(j) - P(i)] > 0 \\ 0 & \text{if } [P(j) - P(i)] = 0 \end{cases} \end{aligned} \quad (6)$$

where *E(i)* and *P(i)* are the experimental and computed values of relative free energies of data point *i*, respectively. By definition PI can assume values between −1 and 1. A value of 1 implies all data points were ranked correctly pairwise, −1 implies all pairs were incorrectly ranked and 0 implies totally random predictions.

RESULTS

To validate the presented method, three systems were selected. The first dataset involves the relative hydration free energy of benzene analogues. To investigate the approach in the presence of a protein, two systems were selected; α-thrombin and P38 MAP kinase (P38MK) for which relative free energies have been measured for a large set of compounds (14 and 16, respectively) that involve single and double phenyl ring substitutions. The choice of full drug-sized molecules that incorporate the fragment is made keeping in mind the ultimate use of the protocol, which is to link the modified fragments into drug-sized molecules. In addition, the choice of model systems is limited due to the lack of a large dataset of experimental values for benzene analogs themselves. The only case to our knowledge where experimental data is available for individual substituted benzenes is that of T4-Lysozyme.[41] Unfortunately, only a few ligands in these studies feature a single heavy atom substitution, for which our protocol is designed so that the useable fraction of the T4-Lysozyme dataset is too small for our purpose. It should be emphasized that in the present study, the relative binding free energy calculations are made separately for each of the six positions on benzene. This allows for direct comparison with the specific substitutions on the phenyl ring on the drug-like molecules, where reorientation of the benzene is restricted due to its connectivity to the remainder of the compound.

Relative solvation free energies of benzene derivatives

A 10 ns production MD simulation of a single benzene molecule in a cubic box of 1100 water molecules was performed. The SSFEP protocol was then applied to the resulting 5000 snapshots and the relative solvation free energy $\Delta\Delta G^{solv}_{benzene \rightarrow analogue}$ of 8 benzene analogues computed using equations 1-3. Even though benzene is in an isotropic environment in the present system, the six possible transformations were generated for each analogue and $\Delta\Delta G^{solv}_{benzene \rightarrow analogue}$ were computed separately for each transformation in order to check the convergence of the results. Figure 2a shows predicted values ± 1 standard deviation averaged over the 6 substitutions vs. the experimental data.[42] A high R^2 value of *0.95* and a PI of 0.99 shows that the experimental trend is reproduced. The small error bars in the figure also show that the values computed separately for each orientation are in reasonable agreement with each other and therefore show that the 10 ns simulation is satisfactory to obtain converged free energies. Relative solvation free energies for the polar compounds are underestimated, but nevertheless the trend in the relative values is captured. TI calculations to compute solvation free energy were performed to check for any force-field dependence in the results. An average of the TI calculation performed in the forward and negative of the value in backward direction was evaluated to yield the TI relative free energy. Figure S1a in the supporting information shows a satisfactory level of agreement between the forward and the backward direction calculations in vacuum and in solution. Figure 2b shows the SSFEP computed values vs. those computed using TI. Tabulated values of the three data sets show that some of the deviations from the experimental values are due to the force field but most are not. In general, the SSFEP calculations predict the hydration free energies of non-polar analogues to be more favorable than experiment. For fluorobenzene, the TI calculations also predict more favorable free energy. However, for chloro-, bromo- and iodobenzene, the TI calculations do not overestimate the free energy as the SSFEP calculations do. For the polar molecules phenol and aniline, the TI values better match the experiment, whereas for

pyridine and toluene this trend is not observed. Figure S1b in the supporting information plots the TI computed free energies vs. experiment. While the R^2 of 0.91 and PI of 0.91 are slightly lower than those obtained from SSFEP calculations, the slope of the regression line at 0.93 is closer to 1 than obtained from the SSFEP results at 0.65, showing that the systematic underestimation of polar and overestimation of non-polar compound free energies in SSFEP is not present in TI results. Overall, the TI calculations better reproduce the experimental data, but the SSFEP calculations also show reasonable correlation with the latter. Having observed a close correspondence between TI and experimental results, further TI calculations for the binding free energies were deemed unnecessary.

Relative binding free energies of α-thrombin ligands

As a first test case for the prediction of relative binding free energies of a series of substituted phenyl rings using the SSFEP method we chose the protein α-thrombin. Baum et al.[43] have measured the binding affinities of a congeneric series of thrombin inhibitors, which differ mainly in substitutions on the phenyl ring that occupies the specificity pocket of the protein. 14 ligands that have one or more single heavy atom substitutions on the phenyl ring of the inhibitor were chosen and these are shown in Figures 3a and b, along with the parent ligand (compound 5), referred to as ATI, short for α-thrombin inhibitor. For each analogue, ΔG^{bind} was calculated as $RT\ln K_i$ and the ΔG^{bind} of the unsubstituted compound was subtracted from this value in order to obtain the experimental $\Delta\Delta G^{bind}_{benzene\rightarrow analogue}$. There can be two possible conformations that the substituted phenyl ring may occupy in the binding pocket for many of these ligands. For example, for ligand 1a, the fluorine substitutions on positions R2 and R4 are equivalent. However, since the SSFEP calculations separate out the free energy values at distinct substitution positions and conversions between these alternate conformations cannot occur, the contributions from two orientations are combined using equation 5.[38]

The ensemble of benzene conformations on which SSFEP calculations were performed was obtained from two independent 10ns simulations of ATI in the binding pocket of thrombin. The apo-structure of thrombin (PDB 3D49) was used in all calculations reported in this paper. The initial conformation of ATI was obtained from the crystal structure (PDB 2ZFF) of the thrombin-ATI complex[43] and was overlaid with the apo structure of the protein (PDB 3D49) based on optimal alignment of the protein conformations followed by the deletion of overlapping crystallographic water molecules using a 2 Å cutoff. Two 10ns NPT MD simulations of the complex resulted in 5000 × 2 conformations of the phenyl ring that were extracted from the MD snapshots and these were subject to the SSFEP protocol. The initial phenyl ring conformation in ATI was chosen as a reference and the 6 possible transformations were generated for each snapshot. These transformations could thus be mapped to the ligands for which experimental data is available. It must be noted that the SSFEP energy evaluations were performed with only the benzene ring and not the whole ligand. Therefore the same protocol was applied to calculate the free energy differences associated with the benzene substitutions when the ensemble of benzene orientations is generated from a simulation of the full inhibitor-protein complex or from SILCS simulations (see below). This includes removal of rotation of the ring based on optimal alignment of ring atoms with the reference conformation. Due to the phenyl ring being attached to the rest of the ligand, this rotation is minimal for the inhibitor-protein complex simulation and results in nearly identical predictions with or without the rotation removal (see below). The cumulative 20ns sampling was divided into four 5ns segments and SSFEP calculations were performed separately on the four ensembles. Averages of the resulting values vs. the experimental binding free energies are listed in Figure 3a and plotted in Figure 3c for the 14 ligands, where the length of the error bars is equal to twice the standard deviation of the four values. Overall, the experimental trend is well reproduced with a reasonable correlation (R^2

= 0.53). Predictive index computed per equation 6 is 0.78, which indicates good rank ordering ability of the method in accordance with experimental binding free energy. Compound 6d, which has a double Cl and OH substitution, is an outlier. Removal of this compound from the dataset causes the R^2 and PI values to improve to 0.67 and 0.80, respectively. Figure S2a in the supporting information shows the nearly identical results obtained without removal of rotation, as expected given the constrained orientation of the phenyl ring due to the remainder of the ligand.

Relative binding free energies of p38-MAP kinase (P38MK) ligands

As our second test case for the prediction of relative binding free energies, we chose a set of 16 p38 MAP kinase inhibitors from a congeneric series for which experimental pIC_{50} data are available.[40] In this system, two sets of simulations were performed from which the SSFEP free energy estimates were made. In the first only the Cα restraints on the protein were included, as with α-thrombin, and in the second, larger harmonic restraints on all protein atoms were included. Figure 4a and b show the parent inhibitor, referred to as MKI (short for MAP kinase inhibitor), and the list of modifications that differ by substitutions on a phenyl ring (R1 to R5). The experimental $\Delta\Delta G^{bind}_{benzene\rightarrow analogue}$ for each analogue was computed by taking the difference between $RT\ln 10^{-pIC50}$ values computed for the substituted and unsubstituted analogue. As with thrombin, there exist contributions to the free energy from multiple possible orientations that the phenyl ring can assume in the binding pocket and therefore, equation 5 was used to compute the SSFEP free energy values corresponding to those ligands. Following the same protocol as for thrombin, two independent 10ns MD simulations of MKI in complex with P38MK were performed. The initial coordinates were obtained from the co-crystal structure of the protein in complex with a very similar inhibitor (PDB 1OUY). The phenyl ring conformation in the crystal structure was used as the reference conformation for generating the 6 possible transformations for the benzene analogues. Figure 4c displays SSFEP predictions averaged over the 4 5ns windows vs. experimental values of the relative binding free energies of the ligand computed from the protein-unrestrained simulation. Poor correlation is observed with respect to the experimental values; however a PI of 0.51, lower than obtained with thrombin, still shows that satisfactory rank ordering is obtained.

Two previous computational studies have sought to reproduce the P38MK experimental data as a test of the accuracy of thermodynamic integration calculations.[38,40] Results from those studies highlight difficulties faced in calculating relative free energies in this system, even by highly precise methods. Pearlman and Charifson[40] performed thermodynamic integration calculations to reproduce the relative binding free energies of the same set of ligands and found poor predictability due to the protein pocket being very flexible. They could only get a reasonable prediction when using a harmonic restraint of 0.5 kcal*mol^{-1} $Å^{-2}$ on protein atoms. Accordingly, following their approach, we performed a second set of simulations referred to as the “restrained” simulations in which restraints of 0.5 kcal*mol^{-1} $Å^{-2}$ were applied to all protein atoms. Figure 4b lists the predicted values from the restrained simulation and Figure 4d plots them vs. experimental data. The correlation with experimental data is improved over that of the initial simulation predictions, and the PI value has increased to 0.74. Additionally, the variance in the calculated values also is seen to be higher in the predicted values from the unrestrained simulation, confirming that the flexibility of this pocket may indeed be the cause of the relatively poor predictability. These results are in line with the previously published study on this protein indicating it to be a particularly difficult challenge due to its inherent flexibility.[40]

As with thrombin, an ensemble of benzene conformations was generated from the inhibitor-protein simulation, with the removal of rotation of the phenyl ring not performed. Figure

S2b in the supporting information shows the results obtained without this modification for the protein-restrained simulation. A relatively worse correspondence with the experimental data is obtained with a PI of 0.53, indicating the importance of rotation removal in this protein, which appears to be associated with the flexibility issues discussed above.

In addition we note, as done previously,[38,40] that the conversion of pIC_{50} values into ΔG is approximate as opposed to conversion from K_i values, which is another potential source of error. The PI values obtained for the same dataset using two studies that involved precise TI calculations were 0.62[38] and 0.84[40], indicating that the results obtained using the rapid SSFEP protocol are reasonable.

Application of the SSFEP protocol to SILCS simulation data of thrombin

In the previous sections we showed that by applying the SSFEP protocol on the phenyl ring snapshots generated from MD simulations of protein-ligand complexes it is possible to reproduce the experimental relative binding free energy values of simple substitutions of the ring. In this section, an approach is applied that extracts conformations from SILCS simulation trajectories and applies the SSFEP protocol to the resultant ensemble. SILCS simulations involve MD simulations of a protein in aqueous solution of small molecules. In a recent publication[6] we reported SILCS simulations of thrombin in a solution of benzene and propane molecules in which the benzene FragMaps correctly located the S1-specificity pocket where the phenyl group of the α-thrombin inhibitor ATI is located. This suggested the possibility that the ensemble of benzene generated from the SILCS simulations itself may be of utility in combination with SSFEP calculations to predict the relative binding of the ATI analogs.

First, the benzene FragMap in the grid free energy (GFE) representation was created using the last 5ns of the 10 published 20ns long SILCS trajectories. Grid centers having a free energy value below a threshold of −1.2 kcal/mol were clustered and the cluster centers identified. Figure 5a, where the cluster centers are represented by green spheres, shows that the FragMaps identify the S1-pocket which coincides with the location of the substituted phenyl group in ATI. From the SILCS simulation trajectories, we select all benzene (and the corresponding environment) conformations for which any benzene carbon atom is closer than 3 Å from the cluster center in the S1-pocket. This leads to selection of benzene molecules in the S1-pocket, while still sampling a relatively broad ensemble of conformations required for the SSFEP calculations. Applying this procedure resulted in 605 snapshots being selected from the SILCS simulations, for which the respective benzene conformations are displayed in Figure 5b. The SSFEP protocol was applied to this ensemble. The reference benzene conformation used to generate different rotations of each ligand was the same as above, and the predicted changes in the free energy of binding of ATI were subsequently estimated using the SSFEP approach. Figure 5c and d show that trends in the experimental relative binding free energies are well reproduced with an R^2 value of 0.74. The relatively high predictive index of 0.87 indicates that the predictions rank most pairs correctly.

The utility of the SSFEP method lies not just in identifying favorable chemical modifications but also geometries as noted previously in the one-step perturbation implementation.[16] There exist X-ray co-crystal structures of thrombin in complex with three inhibitors of the fourteen analyzed above, namely 1a, 1b and 3a (defined in Figure 3), which correspond to flurobenzene, chlorobenzene and toluene, respectively. As discussed above, for these ligands the location of the substitution can be at two distinct positions R2 or R4. The SSFEP calculations based on the SILCS simulations for all three analogues predict the R2 position to be more favorable than R4. From the R2 position SSFEP calculations, the top 20 most favorable conformations, as quantified by most negative $\Delta E_{analogue\text{-}benzene}$ values,

were selected and are shown for flurobenzene, chlorobenzene and toluene in Figure 6a, b and c, respectively. Overlaid on each panel are the crystal conformations[43] of the corresponding ligand. The agreement with the crystal structures shows that the SSFEP calculations correctly identified the R2 substitution location. In fact, the R2 substitution is the most favorable of the six, with $\Delta\Delta G^{bind}_{benzene \rightarrow analogue}$ values of −2.67 and −1.54 kcal/mol for chlorobenzene and toluene, respectively. For flurobenzene, the substitution at the R2 position is less favorable (−1.03 kcal/mol) than the R5 position (−1.80 kcal/mol), but nevertheless still favorable. Somewhat expectedly, the next most favorable position for chlorine substitution after R2 is R5 at −1.68 kcal/mol. In agreement with this prediction, compound 6a has two chlorine substitutions at positions R2 and R5 and is the highest affinity ligand in the dataset. There exists a crystal structure (PDB 1TA2[44]) of a ligand very similar to ATI, which has a double chlorine substitution at R2 and R5 position in agreement with our prediction.

Since this protocol is designed for use in an exploratory context, which does not assume the availability of an existing crystal structure to serve as a reference, the sensitivity of the results to the choice of reference benzene conformation (used to assign the six possible rotational states to the benzene analogues) was tested. We arbitrarily selected two reference conformations from the 605 benzene snapshots and named them ref2 and ref3, respectively. In addition, a fourth conformation, ref4, was selected, which shows the best overlap with the benzene FragMap that was constructed from the SILCS simulation data. Figure S3 in the supporting information shows these conformations, which have RMSDs of 0.98, 1.25 and 1.30 Å, respectively, with respect to the original reference conformation; i.e. the conformation of the phenyl ring in ATI (named ref1). Figure S3 shows that there is good agreement between the four different sets of the predicted 42 values (6 orientations × 7 ligands) of $\Delta\Delta G^{bind}_{benzene \rightarrow analogue}$ as evidenced by the correlation plots between ref1-ref2, ref1-ref3 and ref1-ref4. Few differences are seen, mostly for unfavorably predicted values, which will not be of potential interest in the subsequent drug design process. Thus, the predictions are not highly sensitive to the choice of the reference conformation and the method can therefore be used in an exploratory context.

Application of the SSFEP protocol to SILCS simulation data of P38MK

The protocol as applied above to thrombin was followed for P38MK. Starting from the crystal conformation, ten trajectories of SILCS simulations were performed for 10ns each. The last 5ns segment of each trajectory was used for benzene FragMap construction. Figure 7a displays the overlay of the crystal conformation of MKI with the benzene FragMaps, which correctly identify the substituted phenyl ring of the inhibitor, in addition to the other di-chloro substituted phenyl ring. The low free energy grid centers with GFE < −1.2 kcal/mol were clustered, with the centers of the clusters shown as green spheres in the figure. From the cluster corresponding to the inhibitor, 1000 snapshots (shown in Figure 7b) were obtained and were subject to the SSFEP protocol with the same reference benzene conformation as used before. Figure 7c shows that this results in poor predictability of the experimental data. The reason for this may be the flexibility issues associated with this system as discussed above, in combination with the lack of the intra-ligand constraints caused by the SILCS-based sampling having been performed with a benzene molecule instead of the full ligand. These factors would combine to lead to the conformational ensemble of the benzene molecule in the binding pocket not being representative of that of the phenyl ring in inhibitor MKI, thereby leading to poor agreement with the experimental data.

To test the consistency of benzene conformational distribution from the SILCS simulation with that of the phenyl ring in the full inhibitor, the following analysis was performed. From

the first MD simulation trajectory of the MKI-P38MK complex with the protein restrained, the phenyl ring atoms from the snapshots of the simulation were binned into 1 Å^3 cubic volume elements, forming a 3D probability grid of the ring carbon atoms in the binding pocket. Next, we selected the 50 top conformations of 7 singly substituted analogues in each orientation separately that contribute most favorably to the relative binding free energy and computed the overlap of these conformations with the full-ligand phenyl carbon probability grid. For some ligands there were less than 50 conformations that have a negative (ie. favorable) $\Delta E_{analogue\text{-}benzene}$, such that only the favorable conformations were included. To quantify the extent of overlap, we define an overlap coefficient as follows.

$$OC = \frac{1}{N}\sum_{i=1}^{N} \min\left\{grid\left(x_{i,j}, y_{i,j}, z_{i,j}\right)\right\}_{j=1}^{6} \quad (7)$$

In equation 7, for any conformation i out of $N (\leq 50)$, the $grid()$ function returns the grid occupancy value at the x_{ij}, y_{ij}, z_{ij} position of each ring carbon atom j. A minimum of the occupancy is considered over the six ring atoms j, as this measure is more sensitive to the level of inconsistency between the probability grid and conformations analyzed than any other measures such as the average of the occupancy values. In Figure 7d, we plot the OC (normalized to % values) computed for 7 analogues involving a single substitution vs. absolute errors in the prediction of $\Delta\Delta G^{bind}_{benzene \rightarrow analogue}$. For analogues that involve two alternate conformations in the binding pocket, OC was computed only for the more favorable conformation as judged by the $\Delta\Delta G^{bind}_{benzene \rightarrow analogue}$ value. The red squares show the OC values for the 7 P38MK ligands. As a comparison, similar analysis was performed for 8 thrombin ligands involving a single substitution. OC values of the thrombin ligands are shown as blue squares. In general, the OC values are higher for thrombin, indicating that the benzene spatial distributions overlap better with those of the phenyl moiety from the ligand ATI as compared to P38MK. Correspondingly, the errors in the predicted free energies are lower for thrombin. Moreover, based on the general distribution of the ATI and P38MK data points taken together, it is apparent that the spread in prediction becomes higher as the OC becomes lower; i.e. the data points are roughly evenly distributed below an imaginary line going from the top left to the bottom right of the figure. This analysis supports the hypothesis that the inconsistency of the SILCS simulation ensemble of benzene orientations for P38MK with that of the phenyl moiety in the full-ligand simulation is the reason for the poor prediction. Indeed, this limitation is inherent in fragment based drug discovery in general, as discussed below.

DISCUSSION

The goal of the present protocol is to rapidly identify favorable modifications to fragments that are explicitly sampled in SILCS simulations by using SSFEP calculations applied to a selected conformational subspace. Several approximations and assumptions are made. The first approximation is that of alchemical free energy perturbations performed in a single step, which have the potential to lead to non-overlapping phase space of the two end states. The agreement obtained with experimental hydration and binding free energies suggests that despite this approximation, the method can rank ligands reasonably correctly in the case of single heavy atom modifications. In previous studies it has been noted that it may be difficult to obtain accurate results using the SSFEP method when the end states differ significantly in polar character due to differing environmental responses.[18,19] In the hydration free energy test case, there was a tendency for the SSFEP method to underestimate the free energy decrease upon addition of polar groups to benzene, though the addition of

polar groups was properly predicted to lead to more favorable free energies of solvation vs. non-polar substituents, leading to the relatively high R^2 and PI values (Figure 2).

The second approximation involved the removal of rotation and the separate free energy evaluations of each of the 6 orientations. The underlying assumption in this approximation is that as fragment complexity increases; i.e. as the symmetric benzene molecule is transformed to a substituted analogue, the binding orientation is expected to become specific. Free energy evaluations of orientations separately could neglect enthalpic and entropic contributions arising from other orientations. However, in the context in which this method is to be used, the subsequent linking of fragments into drug-like molecules, where free rotation of the phenyl ring will not be possible or at least restricted due to linkage, this approximation is necessary. Indeed, having separate free energy values for different orientations is exactly what is desired in a subsequent fragment-linking step, which is not straightforward to obtain from traditional free energy methods such as TI, unless additional restraints are applied. For P38MK, the SSFEP results involving the full-ligand simulations were seen to be in much better agreement with the experiment when the calculation was performed after rotation removal. This again shows the importance of this step to account for the lack of specificity that the unsubstituted phenyl ring would have, which may not yield an ensemble consistent with the substitution.

The third approximation is that the effect of multiple simultaneous substitutions is treated in an additive manner.[39,45] For the thrombin dataset, we obtained a significantly higher correlation ($R^2 = 0.91$) when only considering the 9 singly substituted analogues (data not shown) showing the limitations of this approximation. Instead of using the additive assumption, we initially attempted to introduce the simultaneous substitutions in the SSFEP calculations itself, but failed to obtain a close correspondence with the experimental results. This is suggested to be due to the failure to find simultaneously favorable benzene-environment conformations for the multiple substitutions in the solution and/or in protein environment within the time scale of the unperturbed simulation. Similar observations have been made before in the soft-core based one-step perturbation method.[16] Thus, the methodology in the present protocol is expected to be most applicable to single heavy atom substitutions. Further investigation into sampling is required to extend it to predictions of multiple simultaneous heavy atom substitutions.

Finally, even though the method aims to identify fragments, the test set used for validating binding free energy predictions involved large drug-sized molecules due to availability of data and also keeping in mind the fact that the fragment detection step is followed by linking fragments into drug-sized molecules. SSFEP calculations on thrombin SILCS simulation results reproduced the experimental data of the full α-thrombin ligands to a reasonable extent. The predictions are much more accurate than those made from the SSFEP calculations applied to the full-ligand simulation. This is likely due to the optimal benzene-environment conformations generated in the SILCS simulations, which may also yield more representative water distributions on the protein surface as the removal of overlapping crystal water molecules during the generation of thrombin-ATI complex structure has the potential to lead to inaccuracies.[38] For P38MK, the SSFEP calculations performed on the SILCS simulation data did not predict the experimental data. This appears to be due to the non-overlapping conformational spaces of benzene from the SILCS simulation with that of the phenyl ring in the P38MK binding pocket due to the presence of the remainder of the ligand. Thus, the disagreement with experiment is due to contributions arising from linkage with other fragments – an inherent limitation of both experimental and theoretical fragment based methods. Simply, if the linking of fragments in a full ligand does not significantly perturb the conformational space sampled by the individual fragments, predictions made based on the individual fragments will more likely be valid. Accordingly, contributions

arising from fragment linkage need to be accounted for in fragment linking methods, which must carefully use geometries and energetic contributions only from those conformations which are consistent with the linkage.[9]

The key advantage of the SSFEP method in combination with SILCS is efficiency. SILCS calculations require about a week on 10 × 8 processors to obtain ten 10 ns simulations of a system with ~23,500 atoms, from which the FragMaps and GFEs for hydrogen bond donors and acceptors, aliphatics and aromatics were obtained. These data can be used in manifold ways towards drug design as detailed previously.[6] Extending this dataset to a range of substituted benzene analogs required 1.5 hours on 20 single cores of a typical commodity cluster, a process that involved the use of 1000 conformations to evaluate the SSFEP free energy changes of 8 ligands in 6 orientations. It is anticipated that the protocol should be applicable with minor modifications to fragments other than benzene that involve single heavy atom substitutions, though this remains to be tested. This would lead to rapid expansion of chemical space of fragments while requiring explicit sampling only for a few and at minimal additional computational costs.

CONCLUSIONS

Presented is a method that identifies favorable fragment binding sites by analyzing protein-fragment SILCS MD simulations, followed by selecting the relevant conformational subspace pertaining to a protein site of interest. Single step free energy perturbation (SSFEP) calculations performed on the resultant ensemble identify chemical modifications to the bound fragments and corresponding orientations that are predicted to result in a gain in binding free energy. The SSFEP calculations were first validated using experimental hydration free energies of benzene analogues as target data. Relative binding free energies were computed for two sets of ligands of the proteins α-thrombin and P38MK differing only in phenyl ring substitutions. The SSFEP protocol applied to the ensemble obtained from protein-ligand complex MD simulations showed modest ability in rank ordering ligands based on affinity. The protocol was then applied to thrombin SILCS simulation data and the calculated relative free energies of the phenyl analogues show good agreement with experimental data. For P38MK, it was shown that the results of benzene analogues cannot be compared to experimental data of the full drug sized ligand due to the conformational distributions of the benzene ring in these two contexts being different, a problem not observed with thrombin. Contributions due to fragment linkage, an important problem in fragment based methods, need to be carefully considered in the subsequent fragment-linking algorithm. It is expected that with minor modifications, the methodology can be applicable to other rigid fragments that can be sampled in SILCS simulations, though this remains to be tested. As the present protocol is a post-processing method, it allows for site-specific favorable modifications of fragments to be rapidly identified, thus enhancing the utility of SILCS simulations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank all members of the MacKerell group for helpful discussions. We are grateful to Dr. Olgun Guvench for suggestions related to SILCS simulation setup. This work was supported by NIH grants, MH092940, CA120215 and CA107331, NSF grant CHE-0823198 and the Samuel Waxman Cancer Research Foundation. The authors acknowledge computer time and resources from the Computer Aided Drug Design (CADD) Center at the University of Maryland, Baltimore.

REFERENCES

(1). Erlanson DA, McDowell RS, O' Brien T. J. Med. Chem. 2004; 47:3463–3482. [PubMed: 15214773]

(2). Murray CW, Rees DC. Nat. Chem. 2009; 1:187–192. [PubMed: 21378847]

(3). Guvench O, MacKerell AD Jr. PLoS Comput. Biol. 2009; 5:e1000435. [PubMed: 19593374]

(4). Miranker A, Karplus M. Proteins. 1991; 11:29–34. [PubMed: 1961699]

(5). Majeux N, Scarsi M, Apostolakis J, Ehrhardt C, Caflisch A. Proteins. 1999; 37:88–105. [PubMed: 10451553]

(6). Raman EP, Yu W, Guvench O, MacKerell AD Jr. J. Chem. Inf. Model. 2011; 51:877–896. [PubMed: 21456594]

(7). Kulp JL, Pompliano DL, Guarnieri F. J. Am. Chem. Soc. 133:10740–10743. [PubMed: 21682273]

(8). Dey F, Caflisch A. J. Chem. Inf. Model. 2008; 48:679–690. [PubMed: 18307332]

(9). Clark M, Meshkat S, Talbot GT, Carnevali P, Wiseman JS. J. Chem. Inf. Model. 2009; 49:1901–1913. [PubMed: 19610599]

(10). Leis S, Zacharias M. J. Comput. Chem. 2011; 32:3433–3439. [PubMed: 21919015]

(11). Huang D, Caflisch A. J. Mol. Recognit. 2010; 23:183–193. [PubMed: 19718684]

(12). Genheden S, Mikulskis P, Hu L, Kongsted J, Soderhjelm P, Ryde U. J. Am. Chem. Soc. 133:13081–13092. [PubMed: 21728337]

(13). Wang S, Yang C-Y. ACS Med. Chem. Lett. 2011; 2:280–284.

(14). Lexa KW, Carlson HA. J. Am. Chem. Soc. 2011; 133:200–202. [PubMed: 21158470]

(15). Liu H, Mark AE, Gunsteren W. F. v. J. Phys. Chem. 1996; 100:9485–9494.

(16). Oostenbrink C, van Gunsteren WF. Proteins. 2004; 54:237–246. [PubMed: 14696186]

(17). Mordasini TZ, McCammon JA. J. Phys. Chem. B. 2000; 104:360–367.

(18). Oostenbrink C, van Gunsteren WF. Proc. Natl. Acad. Sci. U S A. 2005; 102:6750–6754. [PubMed: 15767587]

(19). Oostenbrink C, van Gunsteren WF. J. Comput. Chem. 2003; 24:1730–1739. [PubMed: 12964191]

(20). Zwanzig RW. J. Chem. Phys. 1954; 22:1420–1426.

(21). Brooks BR, Brooks CL III, MacKerell AD Jr. Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M. J. Comput. Chem. 2009; 30:1545–1614. [PubMed: 19444816]

(22). MacKerell AD Jr. Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiórkiewicz-Kuczera J, Yin D, Karplus M. J. Phys. Chem. B. 1998; 102:3586–3616.

(23). MacKerell AD Jr. Feig M, Brooks CL 3rd. J. Comput. Chem. 2004; 25:1400–15. [PubMed: 15185334]

(24). Durell SR, Brooks BR, Ben-Naim A. J. Phys. Chem. 1994; 98:2198–2202.

(25). Vanommeslaeghe K, Hatcher E, Acharya C, Kundu S, Zhong S, Shim J, Darian E, Guvench O, Lopes P, Vorobyov I, MacKerell AD Jr. J. Comput. Chem. 2010; 31:671–690. [PubMed: 19575467]

(26). Vanommeslaeghe K, Raman EP, MacKerell AD Jr. in preparation. 2012

(27). 2011. http://www.paramchem.org

(28). Word JM, Lovell SC, Richardson JS, Richardson DC. J. Mol. Biol. 1999; 285:1735–1747. [PubMed: 9917408]

(29). Levitt M, Lifson S. J. Mol. Biol. 1969; 46:269–279. [PubMed: 5360040]

(30). Allen, MP.; Tildesley, DJ. Computer Simulation of Liquids. Oxford University Press; Oxford: 1987.

(31). Ryckaert JP, Ciccotti G, Berendsen HJC. J. Comput. Phys. 1977; 23:327–341.
(32). Darden T, York D, Pedersen L. J. Chem. Phys. 1993; 98:10089–10092.
(33). Steinbach PJ, Brooks BR. J. Comput. Chem. 1994; 15:667–683.
(34). Andersen HC. J. Chem. Phys. 1980; 72:2384–2393.
(35). Nosé S. Mol. Phys. 1984; 52:255–268.
(36). Hoover WG. Phys. Rev. A. 1985; 31:1695–1697. [PubMed: 9895674]
(37). Feller SE, Zhang YH, Pastor RW, Brooks BR. J. Chem. Phys. 1995; 103:4613–4621.
(38). Luccarelli J, Michel J, Tirado-Rives J, Jorgensen WL. J. Chem. Theor. Comput. 2010; 6:3850–3856.
(39). Jorissen RN, Reddy GS, Ali A, Altman MD, Chellappan S, Anjum SG, Tidor B, Schiffer CA, Rana TM, Gilson MK. J. Med. Chem. 2009; 52:737–754. [PubMed: 19193159]
(40). Pearlman DA, Charifson PS. J. Med. Chem. 2001; 44:3417–3423. [PubMed: 11585447]
(41). Morton A, Matthews BW. Biochemistry. 1995; 34:8576–8588. [PubMed: 7612599]
(42). Mobley DL, Bayly CI, Cooper MD, Shirts MR, Dill KA. J. Chem. Theor. Comput. 2009; 5:350–358.
(43). Baum B, Mohamed M, Zayed M, Gerlach C, Heine A, Hangauer D, Klebe G. J. Mol. Biol. 2009; 390:56–69. [PubMed: 19409395]
(44). Tucker TJ, Brady SF, Lumma WC, Lewis SD, Gardell SJ, Naylor-Olsen AM, Yan Y, Sisko JT, Stauffer KJ, Lucas BJ, Lynch JJ, Cook JJ, Stranieri MT, Holahan MA, Lyle EA, Baskin EP, Chen IW, Dancheck KB, Krueger JA, Cooper CM, Vacca JP. J. Med. Chem. 1998; 41:3210–3219. [PubMed: 9703466]
(45). Free SM Jr. Wilson JW. J. Med. Chem. 1964; 7:395–399. [PubMed: 14221113]

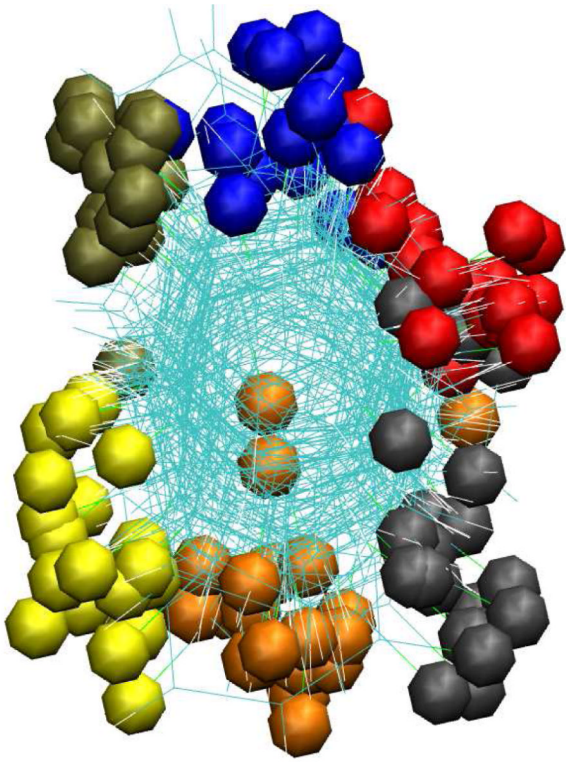
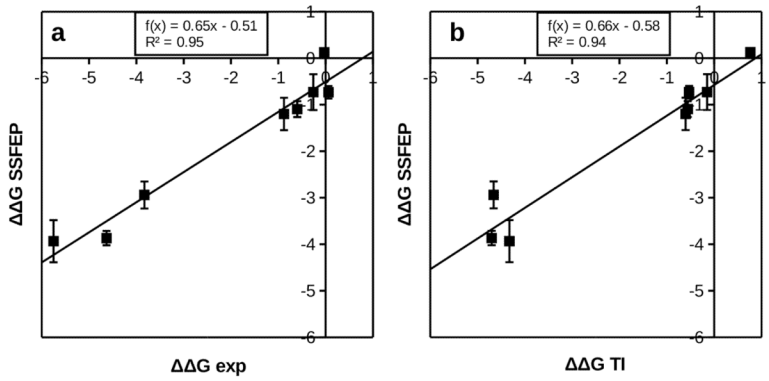


Figure 1.
An illustration of the six different orientations generated for fluorobenzene in the binding pocket of α-thrombin obtained by applying applying the SSFEP protocol to benzene conformations from the α-thrombin SILCS simulations. 20 most favorable conformations in each orientation are depicted with the fluorine atom colored differently for each orientation.



molecule	ΔΔG exp	ΔΔG TI	ΔΔG SSFEP
chlorobenzene	-0.26	-0.15	-0.73
fluorobenzene	0.06	-0.53	-0.73
bromobenzene	-0.60	-0.56	-1.10
iodobenzene	-0.88	-0.61	-1.20
phenol	-5.75	-4.33	-3.93
toluene	-0.03	0.76	0.13
aniline	-4.63	-4.70	-3.87
pyridine	-3.83	-4.66	-2.94

Figure 2.

Relative hydration free energies of benzene analogues with respect to benzene computed using the SSFEP protocol versus (a) experimental data (from supporting information of Mobley et al.[42]) and (b) thermodynamic integration (TI) data. The length of the error bars in the computed values is equal to twice the standard deviation in the six different set of calculations corresponding to the six orientations of the benzene analogue. The same data are shown in the table below. The units are kcal/mol.

a

ID	substituti-on	Ki (nM)	ΔΔG exp	ΔΔG SSFE-P
5	none	11210	0	0
1a	2-F	3990	-0.61	-1.78
1b	2-Cl	180	-2.45	-3.27
1c	2-Br	560	-1.77	-3.78
1d	2-I	540	-1.80	-4.24
2c	2-OH	4910	-0.49	-1.20
2d	2-NH2	14730	0.16	-0.73
3a	2-Me	1210	-1.32	-1.55
6a	1,4-Cl	60	-3.10	-4.20
6b	1-Me,4-Cl	260	-2.23	-3.99
6d	1-OH,4-Cl	433	-1.93	-7.02
6e	1-F,4-Cl	500	-1.84	-4.69
7a	1,4-Me	2780	-0.83	-2.55
7b	3-Cl	3990	-0.61	-0.49
7c	1-Cl	1460	-1.21	-1.28

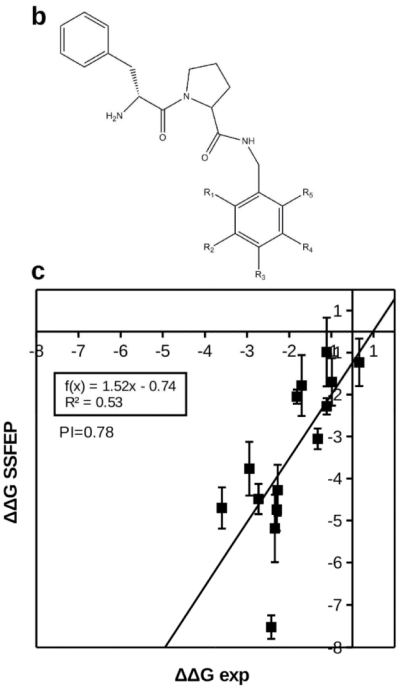
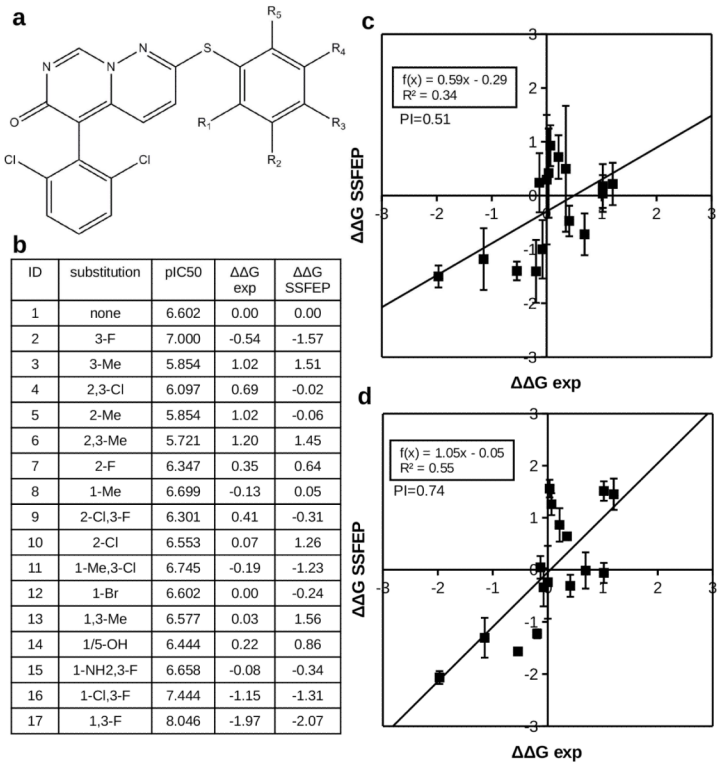


Figure 3.

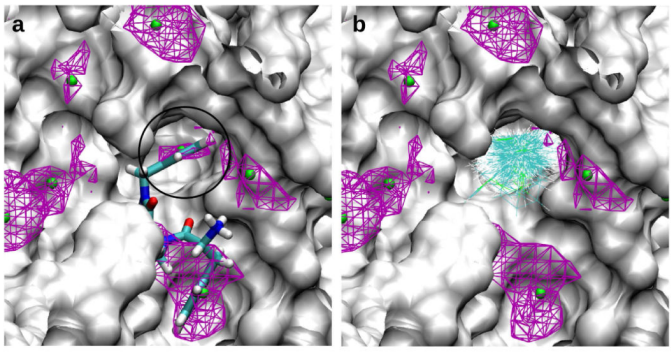
(a) 14 substitutions on the phenyl ring of α-thrombin inhibitor ATI shown in **(b)**. (a) also lists the experimental K_i values[43], converted experimental ΔΔG and computed ΔΔG values. Experimental ΔΔG value for each analogue was obtained as the difference between $RT\ln K_i$ values of the analogue and the unsubstituted compound 5. Computed values were obtained using the SSFEP protocol applied to thrombin-ATI MD simulations and are averaged over the 4 5ns blocks. **(c)** plots computed vs. experimental values. Error bars indicate ±1 standard deviation resulting from the 4 blocks of data used in averaging. The units are kcal/mol.



ID	substitution	pIC50	ΔΔG exp	ΔΔG SSFEP
1	none	6.602	0.00	0.00
2	3-F	7.000	-0.54	-1.57
3	3-Me	5.854	1.02	1.51
4	2,3-Cl	6.097	0.69	-0.02
5	2-Me	5.854	1.02	-0.06
6	2,3-Me	5.721	1.20	1.45
7	2-F	6.347	0.35	0.64
8	1-Me	6.699	-0.13	0.05
9	2-Cl,3-F	6.301	0.41	-0.31
10	2-Cl	6.553	0.07	1.26
11	1-Me,3-Cl	6.745	-0.19	-1.23
12	1-Br	6.602	0.00	-0.24
13	1,3-Me	6.577	0.03	1.56
14	1/5-OH	6.444	0.22	0.86
15	1-NH2,3-F	6.658	-0.08	-0.34
16	1-Cl,3-F	7.444	-1.15	-1.31
17	1,3-F	8.046	-1.97	-2.07

Figure 4.

(a) Parent MAP kinase inhibitor MKI, **(b)** 16 substitutions forming the congeneric series with their experimental pIC50[40], converted experimental and computed ΔΔG values using protein restrained simulation. Experimental ΔΔG values for each analogue were obtained as the difference of $RT\ln 10^{-\mathrm{pIC50}}$ transformed values of the analogue and that of the unsubstituted compound 1. **(c)** Computed vs. experimental ΔΔG values with the computed values obtained by averaging 4 5ns blocks from the SSFEP protocol applied to a phenyl ring conformation from the simulations involving the full ligand. Error bars indicate ±1 standard deviation resulting from the 4 blocks of data used in averaging. **(d)** same as (**c**), but with protein restraints. The data plotted are the same as that listed in (a). The units are kcal/mol.



c

ID	substituti-on	ΔΔG exp	ΔΔG SSFEP
5	none	0.00	0.00
1a	2-F	-0.61	-0.69
1b	2-Cl	-2.45	-2.27
1c	2-Br	-1.77	-2.23
1d	2-I	-1.80	-2.56
2c	2-OH	-0.49	-0.40
2d	2-NH2	0.16	0.58
3a	2-Me	-1.32	-1.16
6a	1,4-Cl	-3.10	-3.94
6b	1-Me,4-Cl	-2.23	-3.27
6d	1-OH,4-Cl	-1.93	-4.03
6e	1-F,4-Cl	-1.84	-4.06
7a	1,4-Me	-0.83	-2.16
7b	3-Cl	-0.61	-0.54
7c	1-Cl	-1.21	-1.29

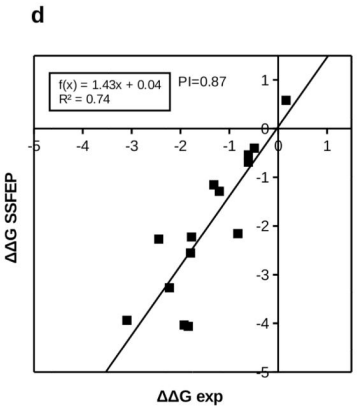


Figure 5.

(a) Crystal structure of apo α-thrombin (PDB 3D49) overlaid with benzene FragMap displayed at a grid free energy cutoff of −1.2 kcal/mol in purple wireframe representation. Green spheres show the cluster centers of the favorable benzene binding regions. The encircled region shows the S1-pocket. **(b)** the conformations selected from the SILCS simulations for SSFEP calculations. **(c)** and **(d)** Relative binding free energies of benzene analogues computed using the SSFEP protocol applied to SILCS trajectories versus experimental data. The units are kcal/mol.

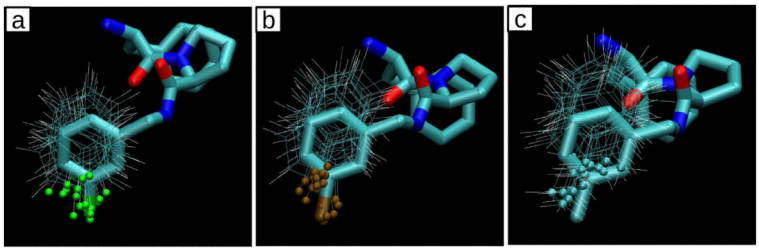


Figure 6.
20 most favorable conformations of fluorobenzene, chlorobenzene and toluene obtained from the SSFEP calculations corresponding to the appropriate orientations overlaid on the crystal conformations 2ZDV, 2ZC9 and 2ZF0 in panels **(a)**, **(b)** and **(c)**, respectively.

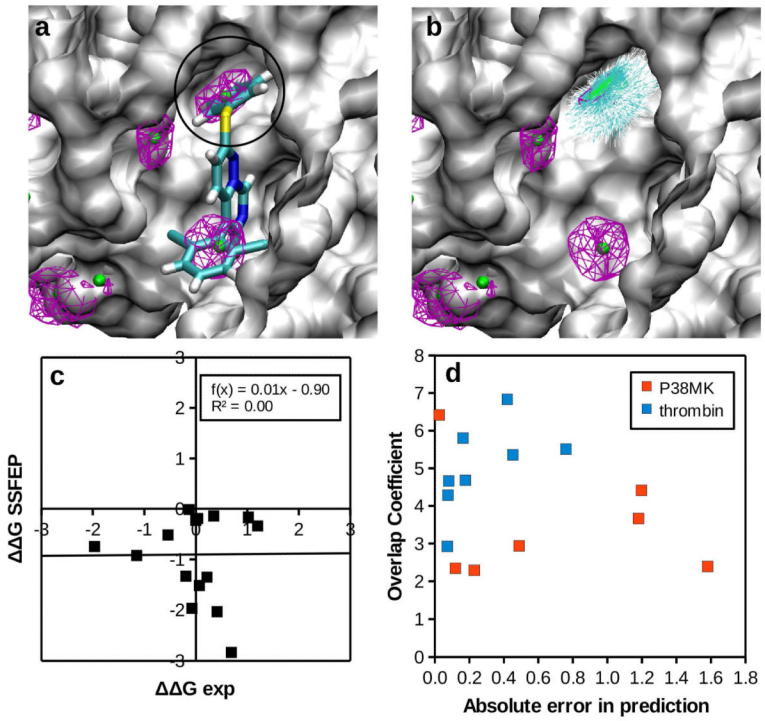


Figure 7.

(a) Crystal structure of P38 MAP kinase overlaid with benzene FragMap displayed at a grid free energy cutoff of −1.2 kcal/mol in purple wireframe representation. Green spheres show the cluster centers of the favorable benzene binding regions. The encircled region shows the binding pocket. **(b)** Conformations selected from the SILCS simulations for SSFEP calculations. **(c)** SSFEP computed relative binding free energies from SILCS simulation data vs. experimental data. **(d)** Overlap coefficient computed per Eqn 7 for 7 and 8 singly substituted benzene analogues of P38MK and thrombin vs. absolute error in prediction. The units of energy are kcal/mol.