Graduate School Form 30
(Updated 11/20/2014)

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Bo Wang

Entitled
STRUCTURE-BASED COMPUTATIONAL STUDIES OF PROTEIN-LIGAND INTERACTIONS

For the degree of Master of Science

Is approved by the final examining committee:

Samy O. Meroueh

Jingzhi Pu

Donald B. Boyd

Christoph A. Naumann

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Approved by Major Professor(s): Samy O. Meroueh

Approved by: Eric C. Long 11/24/2014

Head of the Department Graduate Program Date

STRUCTURE-BASED COMPUTATIONAL STUDIES OF PROTEIN-LIGAND INTERACTIONS

A Thesis

Submitted to the Faculty

of

Purdue University

by

Bo Wang

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science

December 2014

Purdue University

Indianapolis, Indiana

TABLE OF CONTENTS

Page

LIST OF TABLES iv

LIST OF FIGURES v

ABSTRACT vi

CHAPTER 1. MOLECULAR RECOGNITION IN A DIVERSE SET OF PROTEIN-LIGAND INTERACTIONS STUDIED WITH MOLECULAR DYNAMICS SIMULATIONS AND END-POINT FREE ENERGY CALCULATIONS 1

1.1 Introduction 1

1.2 Materials and Methods 3

1.2.1 Scoring Protein-Ligands Complexes 4

1.2.2 MD Simulations 4

1.2.3 MD-Based Free Energy Calculations 6

1.2.4 Correlation Analysis 9

1.3 Results 10

1.3.1 Calculations of Binding Free Energies and Comparison to Isothermal Titration Calorimetry Data 10

1.3.2 Rank-Ordering Protein-Ligand Complexes 19

1.4 Discussion 29

CHAPTER 2. ENRICHMENT OF CHEMICAL LIBRARIES DOCKED TO PROTEIN CONFORMATIONAL ENSEMBLES AND APPLICATION TO ALDEHYDE DEHYDROGENASE 2 33

2.1 Introduction 33

2.2 Materials and Methods 35

Page

2.2.1 Data Set Preparation 35
2.2.2 MD Simulations 37
2.2.3 Scoring Protein-Ligand Complexes 37
2.2.4 Compound Docking 37
2.2.5 Receiver Operating Characteristic Plot and Statistical Analysis 38
2.2.6 ALDH2 Virtual Screening 38
2.2.7 ALDH2 Inhibition Assay 39
2.3 Results 40
2.3.1 Enrichment in the Conformational Ensemble 40
2.3.2 Enrichment for Individual Snapshots 43
2.3.3 Training Support Vector Machine Target Specific with MD Structure 46
2.3.4 *A Priori* Selection of MD Snapshots with High Enrichment Power 49
2.3.5 Rank-Ordering in Crystal and MD Structures 51
2.3.6 Virtual Screening Chemical Library against MD Structures Leads to ALDH2 Small-Molecule Inhibitors 54
2.4 Discussion 56
REFERENCES 60

APPENDICES

Appendix A Supporting Information: Tables 75
Appendix B Supporting Information: Figures 78

LIST OF TABLES

Table Page

1.1 Calculated Free Energies of Selected Protein-Ligand Complexes (Set 1) 5

1.2 Calculated Free Energies of Selected Protein-Ligand Complexes (Set 2) 14

1.3 Correlation Coefficients for Free Energy Calculations 17

1.4 Correlation Coefficients for MM-PBSA Calculations with Different Dielectric Constants... 21

1.5 Components of Free Energy Calculations 23

1.6 Correlation Coefficients for Components of Free Energy Calculations..................... 26

2.1 Validation Set for Enrichment Studies 36

2.2 SVMSP Enrichment Performance of Different Cluster Size..................................... 42

A.1 Atom Types Used in Developing Knowledge-Based Pair Potentials........................ 75

A.2 Parameters Used in MM-PBSA/MM-GBSA Calculations (AMBER9).................... 76

A.3 Enrichment Performance of Different Cluster Size on Test Set Using ROC-AUC Score for Other Scoring Functions.. 77

LIST OF FIGURES

Figure Page

1.1 Chemical Structure of Bound Ligands in Protein-Ligand Complexes 12
1.2 Stereoview of Three-Dimensional Structures for Apo Proteins and Protein-Ligand Complexes 13
1.3 Calculated Free Energy Deviation from Experimental Energy 18
1.4 Regression Plots between Experimental Free Energy and Calculated Free Energy 24
1.5 Regression Plots between Experimental Free Energy and Components of Calculated Free Energy 25
1.6 Performance on Pearson's *R*p, Spearman ρ and Kendall τ Correlation Coefficients 28
2.1 ROC-AUC Scores for Different Clusters of MD Structures 43
2.2 ROC-AUC for Individual Snapshots 44
2.3 ROC-AUC for Individual Snapshots Using SVMSP_{MD} 48
2.4 Filtering MD Snapshots for *A Priori* Identification of High Enrichment Structures 50
2.5 Correlation in Rank-Ordering of Compounds between Different Structures 53
2.6 SVMSP Rescoring of MD Snapshots Identifies ALDH2 Inhibitors 55
B.1 Comparison of Dynamics of Free Ligand and Protein to the Protein-Ligand Complex 78
B.2 Regression Plots between Mean SVMSP Score of DUD Compounds and ROC-AUC of Correspond Snapshots 79

ABSTRACT

Wang, Bo M.S., Purdue University, December 2014. Structure-Based Computational Studies of Protein-Ligand Interactions. Major Professor: Samy Meroueh.

Molecular recognition plays an important role in biological systems. The purpose of this study was get better understanding of the process by incorporating computational tools. Molecular Mechanics-Generalized Born Surface Area (MM-GBSA) method and Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA) method, the end-point free energy calculations provide the binding free energy the can be used to rank-order protein–ligand structures in virtual screening for compound or target identification. Free energy calculations were performed on a diverse set of 11 proteins bound to 14 small molecules was carried out for. A direct comparison was taken between the calculated free energy and the experimental isothermal titration calorimetry (ITC) data. Four and three systems in MM-GBSA and MM-PBSA calculations, respectively, reproduced the ITC free energy within 1 kcal·mol^{-1}. MM-GBSA exhibited better rank-ordering with a Spearman ρ of 0.68 compared to 0.40 for MM-PBSA with dielectric constant ($\varepsilon = 1$). The rank-ordering performance of MM-PBSA improved with increasing ε ($\rho = 0.91$ for $\varepsilon = 10$), but the contributions of electrostatics became significantly lower at larger ε level, suggesting that the only nonpolar and entropy components contribute to the improved results. Our previously developed scoring function, Support Vector

Regression Knowledge-Based (SVRKB), resulted in excellent rank-ordering ($\rho = 0.81$) when applied into MD simulations. Filtering MD snapshots by prescoring protein–ligand complexes with a machine learning-based approach (SVMSP) resulted in a significant improvement in the MM-PBSA results ($\varepsilon = 1$) from $\rho = 0.40$ to $\rho = 0.81$. Finally, the nonpolar components in the free energy calculations showed strong correlation to the ITC free energy while the electrostatic components did not; the computed entropies did not correlate with the ITC entropy.

Explicit-solvent molecular dynamics (MD) simulations offer an opportunity to sample multiple conformational states of a protein-ligand system in molecular recognition. SVMSP is a target-specific rescoring method that combines machine learning with statistical potentials. We evaluate the performance of SVMSP in its ability to enrich chemical libraries docked to MD structures. Seven proteins from the Directory of Useful Decoys (DUD) were involved in the study. We followed an innovative approach by training SVMSP scoring models using MD structures (SVMSP_{MD}). The resulting models remarkably improved enrichment in two cases. We also explored approaches for *a priori* identification of MD snapshots with high enrichment power from an MD simulation in the absence of active compounds. SVMSP rescoring of protein–compound MD structures was applied for the search of small-molecule inhibitors of the mitochondrial enzyme aldehyde dehydrogenase 2 (ALDH2). Rank-ordering of a commercial library of 50,000 compounds docked to MD optimized structures of ALDH2 led to five small-molecule inhibitors. Four compounds had IC_{50}s below 5 μM. These compounds serve as leads for the design and synthesis of more potent and selective ALDH2 inhibitors.

CHAPTER 1. MOLECULAR RECOGNITION IN A DIVERSE SET OF PROTEIN-LIGAND INTERACTIONS STUDIED WITH MOLECULAR DYNAMICS SIMULATIONS AND END-POINT FREE ENERGY CALCULATIONS

1.1 Introduction

Molecular Dynamics (MD) simulation-based free energy calculations have been extensively used to predict the strength of protein-ligand interactions. Every step of drug discovery, from hit identification to lead optimization, can benefit from precise prediction of small molecules bound to protein structures. Free energy calculations can be used for target discovery when applied to a compound docked to the human proteome.[1] Several rigorous methods such as free energy perturbation and thermodynamic integration have been developed for accurate free energy calculations.[2-8] However, in the virtual screening of large chemical or combinatorial libraries, these methods cannot easily be incorporated.[9] Molecular dynamics (MD)-based MM-GBSA or MM-PBSA[10], typical end-point methods, offer an alternative to carry out rigorous free energy calculations. The calculations can consider structurally diverse molecules.

The MM-GBSA or MM-PBSA free energy consists of several of several terms that include a potential energy, a polar and non-polar solvation energy, and an entropy. These components that can be determined independently. More than one approach exists for the calculation of these components. For example, there are different force fields that can obtain the potential energy, which typically includes electrostatic and van der Waals energies.[11] The solvation energy can be calculated using either Poisson-Boltzmann[12] (PB) or Generalized-Born (GB) models.[13] Two commonly used approaches, namely normal mode analysis and quasiharmonic approximation, can be applied for entropy estimation. [14, 15] Finally, the calculations are performed on multiple snapshots collected from MD simulations.[16-18] The selection of different collections of structures is expected to affect the predicted free energy.[19]

Here, MM-GBSA and MM-PBSA calculations were applied to determine the free energy of binding and rank-order a diverse set of protein-ligand complexes. The diversity in the structures of the ligand and targets distinguishes this work from previous efforts that have typically been limited to calculations on congeneric series of compounds on the same target protein. In addition, the experimental isothermal titration calorimetry (ITC) data was used in the comparisons with predicted energy to reduce the uncertainties in the comparisons between predicted and experimental data. A set of 14 protein-ligand structures obtained from the PDBcal database with high quality structural and thermodynamic binding data.[20] Extensive explicit-solvent MD simulations were performed, and various implementations of MM-GBSA and MM-PBSA were used to study the binding of these complexes. Our previously-developed scoring functions were also tested for their ability to rank-order the complexes by scoring MD structures. The effect of induced-fit conformational changes on rank-ordering these complexes was studied by performing separate simulations for ligand, protein and protein-ligand complexes. Components of the MM-GBSA and MM-PBSA free energy were compared with the ITC free energy, enthalpy and entropy.

1.2 Materials and Methods

1.2.1 Scoring Protein-Ligands Complexes

We previously reported the Support Vector Machine Target SPecific (SVMSP) model[21] for enriching databases and Support Vector Regression Knowledge-Based (SVRKB) scoring[21, 22] for rank-ordering protein-compound complexes based on their binding affinity. SVMSP was specific by developed for each individual target protein and SVRKB was a generalized scoring function for predicting the binding affinity of protein-ligand interactions. The SVMSP model was developed by using protein-ligand crystal structures from the sc-PDB database (v2010)[23] for the positive set (active compounds classification) and randomly selected compounds docked to the target of interest as the negative set (inactive compounds classification). An improvement was made from the previous working removing crystal structures in which the ligand contains highly charged moieties such as sulfate or phosphate groups to refine the positive set, which resulted in a final set of 4,677 complexes. Random selected 5,000 compounds from the ChemDiv library[24] were docked to the pocket on the corresponding target to build the negative set of the model.

In the development of SVMSP and SVRKB models, we extended our previous knowledge-based descriptors by using 14 distinct protein atom types and 16 ligand atom types (Appendix A, Table A.1).[21] This resulted in 224 atom-pairs based potentials. We used 76 pair potentials for the vectors of SVMSP and SVRKB. A higher SVMSP score corresponds to a higher probability that the compound is an active one to the target. The higher SVRKB score indicate a higher binding affinity.

1.2.2 MD Simulations

Explicit-solvent MD simulations were carried out for 14 complexes of small molecules bound to a protein, which were selected from the PDBcal database (Table 1.1).[20] Crystal structures of the target proteins were obtained from RCSB Protein Data Bank.[25] Preparation of the structures was performed by adding hydrogen atoms and modeling missing gaps with BIOPOLYMER module in SYBYL 8.0 (Tripos International, St. Louis, Missouri, USA). Optimization of the hydrogen bonding network was processed by using the REDUCE[26] program in adjusting residue orientation and protonation states. All the ligand structures were extracted from crystal structures and visually checked and prepared in SYBYL. The compound was assigned AM1-BCC[27] charges using the *antechamber* program from the AMBER9 package.[28] Water molecules from crystal structures within 5 Å to any atoms on the protein or compound were retained. No atom on the protein was within 14 Å from any side of the box. The solvated box was further neutralized with Na^+ or Cl^- counterions using the *leap* program from the AMBER9 package.[28]

Simulations were carried out using the *pmemd* program in AMBER9 with ff03 force field[29] in periodic boundary conditions. All bonds involving hydrogen atoms were constrained by using the SHAKE algorithm.[30] The simulations were carried out using a 2 fs time step. The particle mesh Ewald (PME) method was used to treat long-range electrostatics. Simulations were performed at the conditions of 298 K under 1 atm in NPT ensemble employing Langevin thermostat[31] and Berendsen barostat.[32] Water molecules were first energy-minimized and equilibrated by running a short simulation with the complex fixed using Cartesian restraints.

Table 1.1 Calculated Free Energies of Selected Protein-Ligand Complexes (Set 1)

		Methods							
Target	Complex PDB	$\Delta G_{MM\text{-}GBSA}$ (kcal•mol^{-1})	$\Delta G_{MM\text{-}PBSA}$ (dielc = 1) (kcal•mol^{-1})	$\Delta G_{MM\text{-}PBSA}$ (dielc = 2) (kcal•mol^{-1})	$\Delta G_{MM\text{-}SVRKB}$ (kcal•mol^{-1})	$\Delta G_{SVMSP//MM\text{-}PBSA}$ (dielc = 1) (kcal•mol^{-1})	$\Delta G_{SVMSP//MM\text{-}PBSA}$ (dielc = 2) (kcal•mol^{-1})	ΔG_{ITC} (kcal•mol^{-1})	Ligand No.
human cyclophilin A	1CWA	-17.4±0.7	-13.1±0.8	-36.6±0.7	-7.3	-17.0±0.4	-41.0±0.7	-10.9±0.03	**1**
HIV-1 protease	1HPV	-18.9±0.8	-8.2±0.8	-38.5±0.8	-10.1	-15.6±0.3	-43.7±0.8	-12.6	**2**
HIV-1 protease	1HPX	-31.1±0.8	-12.5±0.8	-49.0±0.8	-11.5	-21.1±0.4	-58.9±0.8	-13.3	**3**
HIV-1 protease	1HXW	-32.4±0.8	-9.9±0.8	-49.3±0.8	-10.1	-24.6±0.4	-65.8±0.8	-13.63±0.07	**4**
leukocyte function-associated antigen-1	1RD4	-14.9±0.8	-9.8±0.8	-30.9±0.8	-11.1	-17.9±0.2	-39.5±0.8	-10.73	**5**
porcine odorant-binding protein	1DZK	-6.2±0.6	-5.9±0.6	-13.2±0.6	-5.8	-6.0±0.2	-13.0±0.6	-9.2	**6**
mouse major urinary protein	1I06	-4.4±0.7	-4.0±0.7	-9.1±0.7	-7.2	-3.2±0.2	-9.0±0.7	-8.38±0.52	**7**
mouse major urinary protein	1QY1	-7.5±0.7	-7.7±0.7	-13.9±0.7	-6.3	-8.8±0.3	-14.9±0.7	-8.1±0.07	**8**
DNA gyrase subunit B	1KZN	-24.1±0.9	-4.7±1.0	-41.7±0.9	-8.8	-14.0±0.6	-49.8±0.9	-9.785	**9**
hen lysozyme C	1LZB	-7.67±0.7	3.6±0.7	-28.5±0.8	-9.0	-5.7±0.3	-39.9±0.7	-7±0.01	**10**
human galectin-3	1KJL	-13.7±0.6	-8.1±0.7	-22.3±0.6	-5.2	-12.1±0.3	-25.1±0.6	-5.6	**11**
purine nucleoside receptor A	2FQY	-19.4±1.0	-13.1±1.1	-33.8±1.0	-6.9	-11.3±0.6	-35.3±1.0	-8.81±0.09	**12**
bovine pancreatic trypsin	1S0R	-6.6±0.8	-11.6±0.9	4.9±0.8	-5.5	-10.9±0.5	5.2±0.8	-6.35±0.07	**13**
human brain fatty acid-binding protein	1FDQ	-10.4±0.7	-12.8±0.9	-40.3±0.8	-9.8	-14.9±0.6	-41.7±0.7	-10.1	**14**

A series of energy minimizations followed up in which the Cartesian restraints were gradually relaxed from 500 kcal·Å^{-2} to 0 kcal·Å^{-2}, and the system was subsequently gradually heated to 298 K via a 48 ps MD run. For each target, 6 independent simulations in length of 4 ns were performed by assigning different initial velocities. MD snapshots were saved every 1 ps yielding 6,000 structures per trajectory. The first 2 ns in each trajectory were discarded for equilibration.

1.2.3 MD-Based Free Energy Calculations

MM-PBSA and MM-GBSA free energy calculations combine internal energy, solvation energy based on electrostatic and nonpolar contributions, and the entropy. These calculations are carried out on snapshots collected from MD simulations. The binding free energy is expressed as:

$$\Delta G_{MM-PBSA} = \Delta E_{PBTOT} - T\Delta S_{NM/QHA}$$

$$\Delta G_{MM-GBSA} = \Delta E_{GBTOT} - T\Delta S_{NM}$$

where $\Delta G_{MM-PBSA}$ and $\Delta G_{MM-GBSA}$ are binding free energies calculated by MM-PBSA and MM-GBSA method, ΔE_{PBTOT} and ΔE_{GBTOT} are the combined internal and solvation energies, T is system temperature. $\Delta S_{NM/QHA}$ is entropy determined by normal mode calculation or quasiharmonic analysis. The internal energy is determined using the Lennard-Jones and Coulomb potentials[33] in the Amber force-field (ΔE_{GAS}). The solvation energy is determined using Poisson-Boltzmann or Generalized-Born solvation models (ΔE_{PBSOL} or ΔE_{GBSOL}):

$$\Delta E_{PBTOT} = \Delta E_{PBSOL} + \Delta E_{GAS}$$

$$\Delta E_{GBTOT} = \Delta E_{GBSOL} + \Delta E_{GAS}$$

where ΔE_{PBSOL} and ΔE_{GBSOL} are the solvation free energies calculated with PB or GB model, and ΔE_{GAS} is the molecular mechanical energies. The molecular mechanical energies are composed of three components:

$$\Delta E_{GAS} = \Delta E_{ELE} + \Delta E_{VDW} + \Delta E_{INT}$$

where ΔE_{ELE} is the non-bonded electrostatic energy, ΔE_{VDW} is non-bonded van der Waals energy, and ΔE_{INT} is the internal energies composed of bond, angle, and dihedral energies.

The solvation free energies can be calculated using PB or GB model, expressed respectively by:

$$\Delta E_{PBSOL} = \Delta E_{PBSUR} + \Delta E_{PBCAL}$$

$$\Delta E_{GBSOL} = \Delta E_{GBSUR} + \Delta E_{GB}$$

where ΔE_{PBSUR} and ΔE_{GBSUR} are hydrophobic contribution to desolvation energy, ΔE_{PBCAL} and ΔE_{GB} are reaction field energies.[34]

All the binding energies are determined by:

$$\Delta E = E^{PL} - E^{P} - E^{L}$$

where E^{PL}, E^{P} and E^{L} are total energies corresponding to protein-ligand complex (PL), protein (P) and ligand (L), respectively.

The molecular mechanical gas phase energies were calculated by the *sander* program from the AMBER9 package, including the internal energy, van der Waals and electrostatic interactions. The dielectric constant for electrostatic interactions was set to 1.0. The polar contributions of the solvation free energy were calculated with Poisson-Boltzmann (PB) method using the *pbsa* program[12] and generalized Born (GB) method implemented in *sander*. The nonpolar contributions of the desolvation energy were determined with solvent-accessible-surface-area (SASA) dependent terms.[35] The surface area was calculated by *molsurf* program.[36] The surface tension used to calculate the nonpolar contribution to the free energy of solvation is 0.0072. In the PB method, the reaction field energy was calculated with the dielectric constants for protein and solvent as 1.0 and 80.0, respectively. In the test of the contribution of dielectric constant, we use various dielectric constant for the solute from 1 to 10, 15 and 20. The default value of the dielectric constant is 1. The solvent probe radius was set to 1.6 Å, which was optimized by Tan and Luo.[37] Atomic radii were values optimized by Tan and Luo.[37] The calculation based on the GB method was performed with the Onufriev's GB model.[38, 39] The SASA calculation was switched to the Icosahedra (ICOSA) method; surface area was computed by recursively approximating a sphere around an atom, starting from an icosahedra. Two different methods were applied for the calculation of entropies of the protein-ligand complexes. The quasiharmonic approximation was analyzed using the *ptraj* program in AMBER. Normal mode conformational entropies were estimated with the *nmode* module from AMBER. The distance-dependent dielectric constant was set to 4. Maximum number of cycles of minimization was set to 10,000. The convergence criterion for the energy gradient to stop minimization was 0.0001. All the detailed parameters of MM-PBSA and MM-GBSA free energy calculation are shown in Appendix A, Table A.2.

For the MM-PBSA or MM-GBSA free energy calculations, a set of 500 structures for each protein-ligand complex was extracted from trajectories of MD simulations at regular intervals. For $\Delta G_{\text{SVMSP//MMPBSA}}$ and $\Delta G_{\text{SVMSP//MM-GBSA}}$, all snapshots from MD simulations were first scored by SVMSP. The top scoring 500 structures were selected for free energy calculation. For $\Delta G_{\text{MM-SVRKB}}$, all snapshots were scored by SVRKB first; the mean value of SVRKB score of all snapshots was used for calculated binding affinity (pK_d) of the complex using:

$$\Delta G = -2.303\ RT(pK_d)$$

where R is the gas constant, T is room temperature (298.15 K).

1.2.4 Correlation Analysis

Three correlation metrics, Pearson's correlation coefficient R_p, Spearman correlation coefficient ρ, and Kendall tau τ, were used in model parameterization and performance assessment. The correlation analysis was done using packages in R (version 1.12.1). The 95% confidence interval was calculated using the 5,000 replicate bootstrap sampling.

The Pearson product-moment correlation coefficient R_p is a measure of linear dependence between two variables *x* and *y*, giving a value between +1 and −1 inclusive. It was given by:

$$R_p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \cdot \sum_i (y_i - \bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the mean value for x_i and y_i respectively. The Spearman's rank correlation coefficient ρ assesses how well the association of two variables can be described using a monotonic function. It was given by

$$\rho = 1 - \frac{6 \sum_i (x_i' - y_i')^2}{n(n^2 - 1)}$$

where x_i' and y_i' denote the ranks of x_i and y_i, n is the total number of x-y pairs. A perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other. Kendall tau rank correlation coefficient τ is a measure of the association between two measured quantities. It was given by

$$\tau = \frac{\sum_{i<j} sign(x_j - x_i) \cdot sign(y_j - y_i)}{\frac{1}{2}n(n-1)}$$

when the values of x_i and y_i are unique.

1.3 Results

1.3.1 Calculations of Binding Free Energies and Comparison to Isothermal Titration Calorimetry Data

Free energy calculations were carried out for a set of 14 protein-ligand interactions using MM-GBSA and MM-PBSA (Figure 1.2). The structure of these complexes was previously solved by crystallography and characterization of the binding was done by ITC. The set contains 11 unique proteins and 14 structurally different ligands. The ligands include a cyclic peptide (**1**), peptidomimetics (**2**- **4**), small organic molecules (**5**-**10**, and **13**), carbohydrates (**10** and **11**), a nucleoside (**12**) and a fatty acid (**14**) (Figure 1.1). Among the small organic molecules, four were fragment-like (**6**-**8**, and **13**) with molecular weight less than 200 Da. Calculations were carried out using the MM-GBSA and MM-PBSA approach on multiple MD structures collected from 12 ns of simulation. The computed MM-GBSA or MM-PBSA free energies were compared to experimental binding affinity data ΔG_{ITC} (Table 1.1, Figure 1.3A). Among the 14 complexes, the predicted $\Delta G_{\text{MM-PBSA}}$ were excellent (less than 1 kcal•mol^{-1}) for three of the ligands, namely for (i) **3** binding to HIV-1 protease (PDB code: 1HPX; $|\Delta\Delta G|$ = 0.8); (ii) **8** binding to mouse major urinary protein 1 (PDB code: 1QY1; |$\Delta\Delta G$| = 0.4); and (iii) binding of **5** to human leukocyte function-associated antigen-1 (PDB code: 1RD4; |$\Delta\Delta G$| = 0.9). The predicted binding affinities for another five ligands were between 2 and 4 kcal•mol^{-1}, namely for (i) **1** binding to human cyclophilin A (PDB code: 1CWA; |$\Delta\Delta G$| = 2.2); (ii) **6** binding to porcine odorant-binding protein (PDB code: 1DZK; |$\Delta\Delta G$| = 2.2); (iii) **14** binding to human brain fatty acid-binding protein (PDB code: 1FDQ; |$\Delta\Delta G$| = 2.7); (iv) **4** binding to HIV-1 protease (PDB code: 1HXW; |$\Delta\Delta G$| = 3.7); and (v) **11** bound to human galectin-3 (PDB code: 1KJL). The remaining predicted affinities for compounds **2**, **7**, **9**, **10**, and **13** were larger than 4 kcal•mol^{-1}. An overall measure of the deviation of the MM-PBSA free energy from the ITC free energy is provided by the root-mean-square of the calculated free energy deviation from experimental energy $\Delta\Delta \mathrm{G}_{\mathrm{RMS}}$, which was 4.4 kcal•mol^{-1}. The median $\Delta\Delta G$ ($\Delta\Delta G_{\mathrm{MED}}$) for MM-PBSA is 3.5. The effect of the dielectric constant on the MM-PBSA calculations was also investigated (Table 1.4). Doubling the dielectric constant from 1 to 2 resulted in a significantly worse agreement between the MM-PBSA and ITC free energy as evidenced by a 5-fold increase in $\Delta\Delta G_{\mathrm{RMS}}$ and a 7-fold increase in ($\Delta\Delta G_{\mathrm{MED}}$). This was also observed for calculations performed with larger dielectric constants (Table 1.4).

The above calculations are repeated using a GB model for the electrostatic solvation free energy (MM-GBSA). MM-GBSA free energies were significantly larger than MM-PBSA free energies. In some cases, $\Delta\Delta G$ of MM-GBSA energies exceeded 18 kcal•mol^{-1}. Seven of the MM-GBSA free energies deviated from the ITC free energies by 5 kcal•mol^{-1} compared with only two for MM-PBSA. Overall the MM-GBSA free energy showed greater deviation from the ITC free energy ($\Delta\Delta G_{\mathrm{RMS}}$ = 9.2 kcal•mol^{-1}) compared with MM-PBSA ($\Delta\Delta G_{\mathrm{RMS}}$ = 4.4 kcal•mol^{-1}). The median $\Delta\Delta G$ for MM-GBSA is 5.2 kcal•mol^{-1}.

Despite the large absolute values, MM-GBSA reproduced the free energy of binding remarkably well in four cases with ΔΔ*G* less than 1 kcal•mol^{-1}: (i) **7** binding to the mouse major urinary protein 1 (PDB code: 1QY1; |ΔΔ*G*| = 0.6); (ii) **10** binding to hen lysozyme C (PDB code: 1LZB; |ΔΔ*G*| = 0.7); (iii) **13** binding to the bovine pancreatic trypsin (PDB code: 1S0R; |ΔΔ*G*| = 0.3); and finally (iv) **14** bound to human brain fatty acid-binding protein (PDB code: 1FDQ; |ΔΔ*G*| = 0.3).

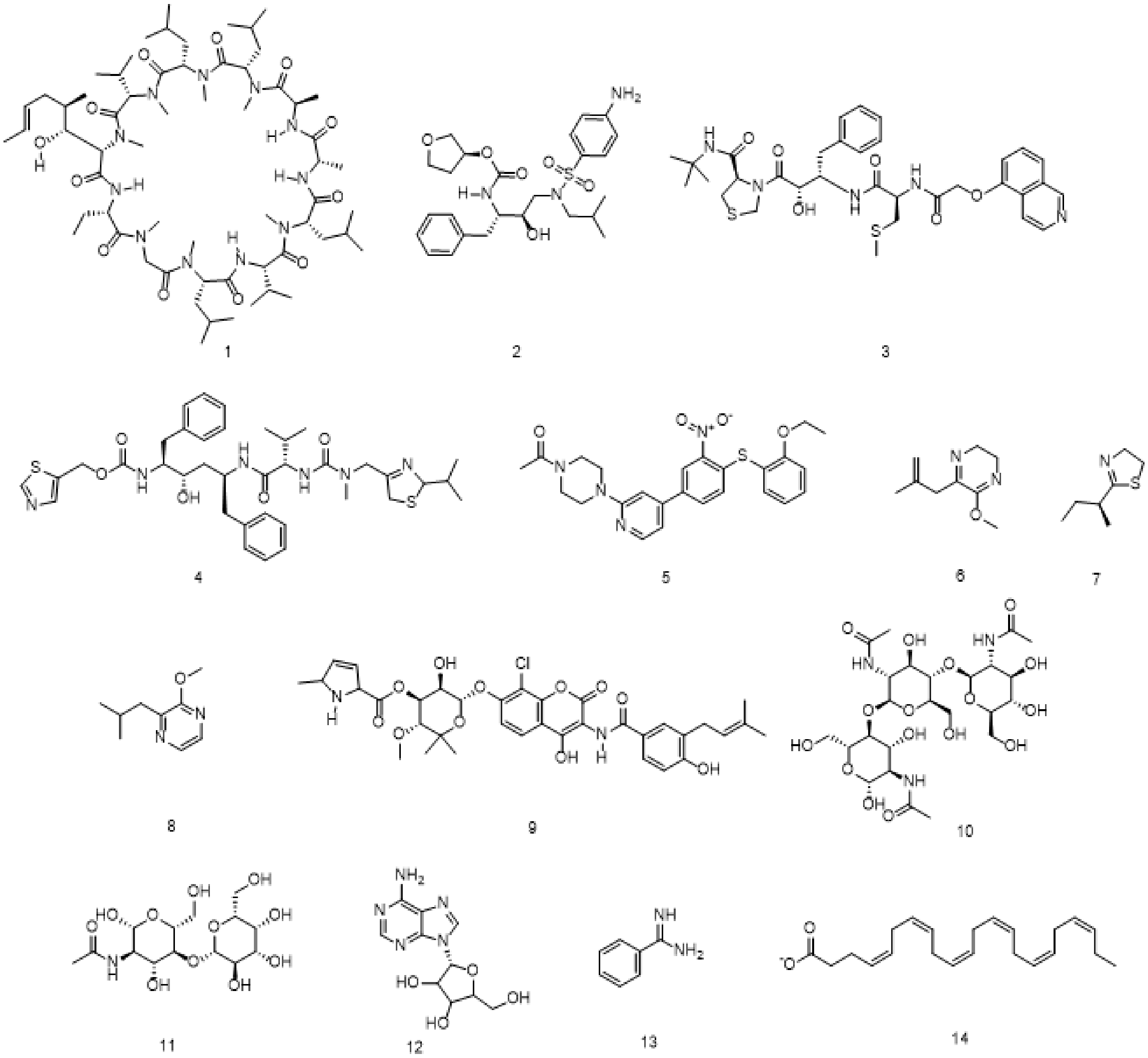


Figure 1.1 Chemical Structure of Bound Ligands in Protein-Ligand Complexes

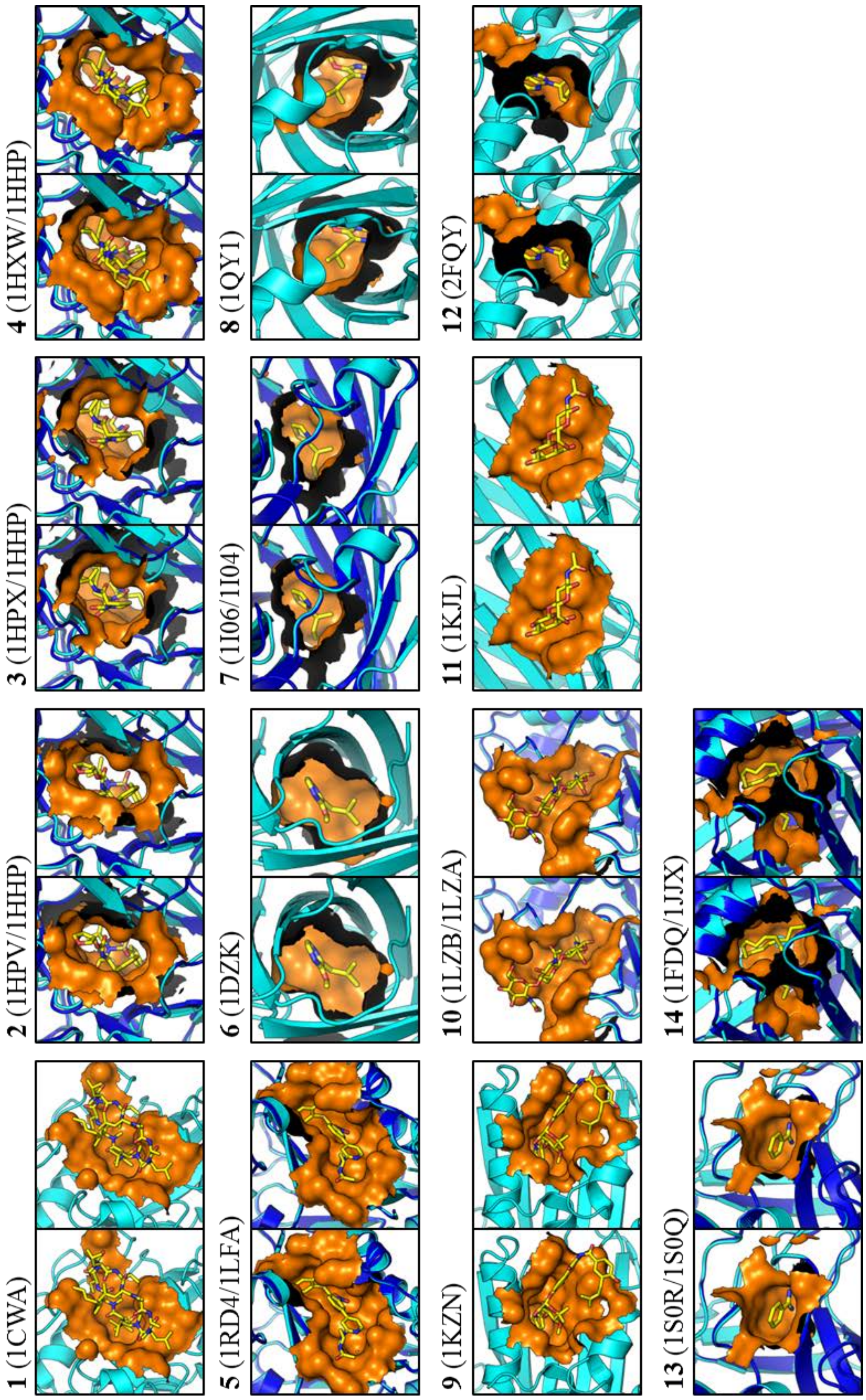


Figure 1.2 Stereoview of Three-Dimensional Structures for Apo Proteins and Protein-Ligand Complexes

Table 1.2 Calculated Free Energies of Selected Protein-Ligand Complexes (Set 2)

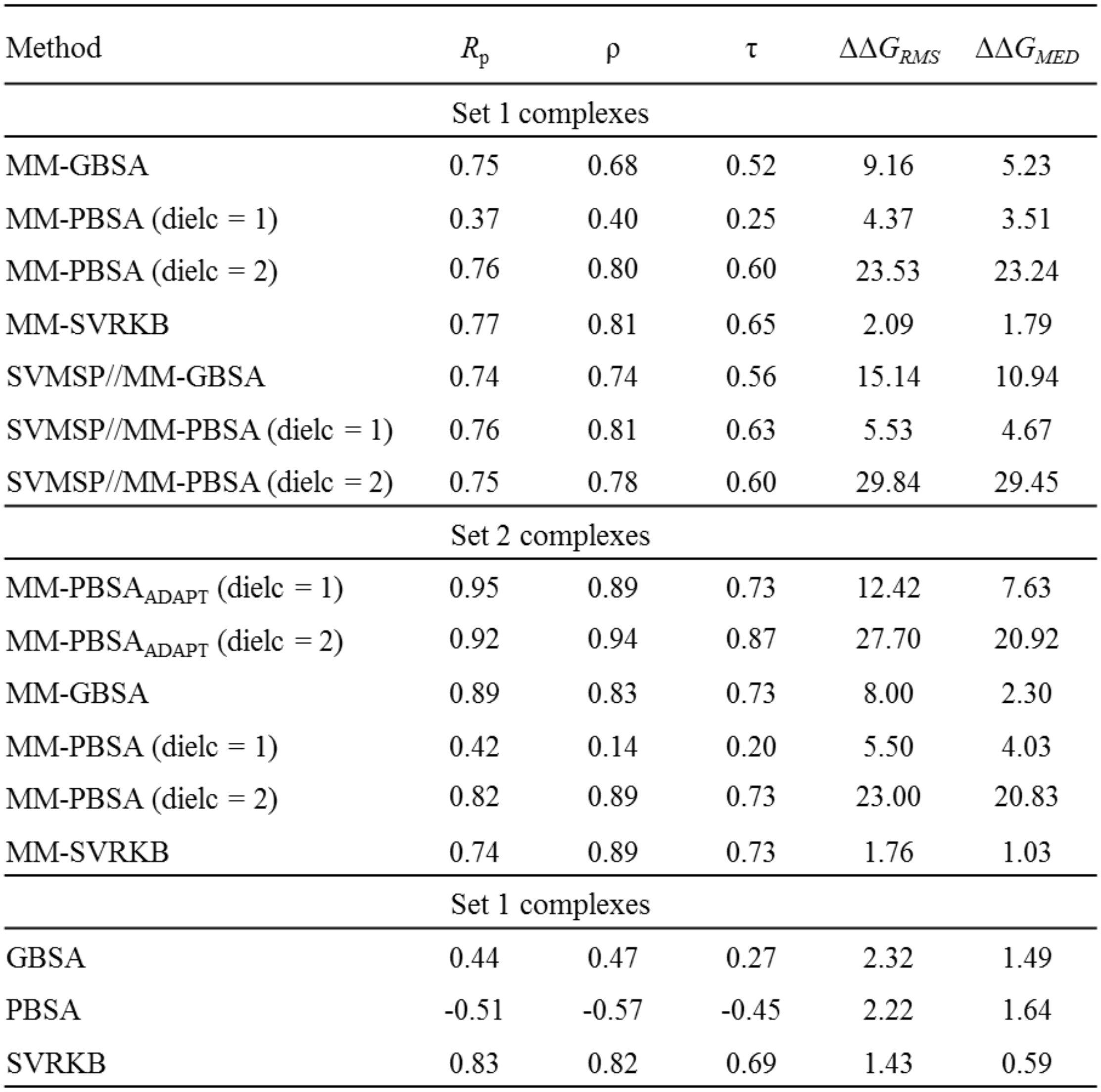
			Methods					
Target	Complex PDB	Apo PDB	$\Delta G_{PB\text{-}ADAPT}$ (dielc = 1) (kcal•mol^{-1})	$\Delta G_{PB\text{-}ADAPT}$ (dielc = 2) (kcal•mol^{-1})	$\Delta G_{MM\text{-}PBSA}$ (dielc = 1) (kcal•mol^{-1})	$\Delta G_{MM\text{-}PBSA}$ (dielc = 2) (kcal•mol^{-1})	ΔG_{ITC} (kcal•mol^{-1})	Ligand No.
HIV-1 protease	1HXW	1HHP	-29.4±1.1	-67.1±2.8	-9.9±0.8	-49.3±0.8	-13.63±0.07	**4**
leukocyte function-associated antigen-1	1RD4	1LFA	-8.1±1.1	-35.8±3.1	-9.8±0.8	-30.9±0.8	-10.73	**5**
mouse major urinary protein	1I06	1I04	-2.4±0.9	-9.7±3.1	-4.0±0.7	-9.1±0.7	-8.38±0.52	**7**
hen lysozyme C	1LZB	1LZA	13.0±0.9	-22.6±2.3	3.6±0.7	-28.5±0.8	-7±0.01	**10**
bovine pancreatic trypsin	1S0R	1S0Q	6.7±1.1	15.0±2.7	-11.6±0.9	4.9±0.8	-6.35±0.07	**13**
human brain fatty acid-binding protein	1FDQ	1JJX	-2.2±1.1	-30.6±2.8	-12.8±0.9	-40.3±0.8	-10.1	**14**

Typically, MM-GBSA calculations are carried out by running a single simulation for the complex. Implicit in this approach is that the ligand will only select conformations of the apo protein that are similar to those that are sampled by the protein in the protein-ligand complex. However, there are numerous examples of ligand binding that leads to conformational change of the protein. The free energy of this conformational change, also known as adaptation energy, contributes to the overall free energy of binding.[40] We investigate the role of this adaptation energy for 6 of the 14 complexes (Table 1.2 and Figure 1.3B) for which the crystal structure of the apo was solved independently from the complex structure. Starting with the structure of complex, apo and ligand, three separate MD simulations were carried out. The root-mean-square deviation (RMSD) of the free protein and ligand were determined with respect to the crystal structure of the protein and ligand in the complex crystal structure (Appendix B, Figure B.1). The protein and ligand sampled different structures in the free-state compared to the bound state.

The snapshots from the three separate simulations of complex, apo and ligand are used to carry out MM-PBSA free energy calculations ($\Delta G_{\text{PB-ADAPT}}$) (Table 1.2). These are compared with the standard MM-PBSA free energies ($\Delta G_{\text{MM-PBSA}}$) (Table 1.2, Figure 1.3B). Overall, the RMSD of $\Delta G_{PB\text{-}ADAPT}$ from the ITC free energies is $\Delta\Delta G_{\text{RMS}}$ = 12.4 kcal•mol^{-1} with a median $\Delta\Delta G$ of 7.6 kcal•mol^{-1} (Table 1.3). Hence, $\Delta G_{\text{PB-ADAPT}}$ resulted in overall greater deviation from the experimental free energy than both MM-GBSA ($\Delta G_{\text{MM-GBSA}}$) and MM-PBSA ($\Delta G_{\text{MM-PBSA}}$). Only one out of the 6 complexes, namely **5** in complex with human leukocyte function-associated antigen-1 (PDB code: 1RD4; $|\Delta\Delta G|$ = 2.6), showed reasonable agreement with experiment (<3 kcal•mol^{-1}). The remaining five exhibited binding free energies that were substantially different from the ITC data.

A question of interest is whether scoring functions can generate reliable binding affinities when carried out on multiple structures sampled from MD simulations instead of crystal structures. To address this question, we applied our recently-developed scoring function, SVRKB,[21] to snapshots from MD simulations. The empirical scoring function is trained on three-dimensional protein-ligand crystal structures and experimentally-measured binding affinity data. SVRKB is used to score MD snapshots of the 14 complexes considered for MM-GBSA and MM-PBSA calculations (Table 1.1 and 1.2). We refer to this approach as MM-SVRKB to emphasize the use of multiple MD structures in the scoring. MM-SVRKB ($\Delta\Delta G_{\mathrm{RMS}}$ = 2.1 kcal•mol^{-1}) showed better agreement with the experimental free energies than MM-PBSA ($\Delta\Delta G_{\mathrm{RMS}}$ = 4.4 kcal•mol^{-1}). In fact, $|\Delta\Delta G_{\mathrm{MM\text{-}SVRKB}}|$ was less than 2 kcal•mol^{-1} for 10 of the targets, compared with three for the MM-PBSA calculations. None of the predicted MM-SVRKB binding affinities was greater than 5 kcal·mol^{-1} than the experimentally-measured affinity.

Table 1.3 Correlation Coefficients for Free Energy Calculations



Method	R_p	ρ	τ	$\Delta\Delta G_{RMS}$	$\Delta\Delta G_{MED}$
		Set 1 complexes			
MM-GBSA	0.75	0.68	0.52	9.16	5.23
MM-PBSA (dielc = 1)	0.37	0.40	0.25	4.37	3.51
MM-PBSA (dielc = 2)	0.76	0.80	0.60	23.53	23.24
MM-SVRKB	0.77	0.81	0.65	2.09	1.79
SVMSP//MM-GBSA	0.74	0.74	0.56	15.14	10.94
SVMSP//MM-PBSA (dielc = 1)	0.76	0.81	0.63	5.53	4.67
SVMSP//MM-PBSA (dielc = 2)	0.75	0.78	0.60	29.84	29.45
		Set 2 complexes			
MM-PBSA$_{ADAPT}$ (dielc = 1)	0.95	0.89	0.73	12.42	7.63
MM-PBSA$_{ADAPT}$ (dielc = 2)	0.92	0.94	0.87	27.70	20.92
MM-GBSA	0.89	0.83	0.73	8.00	2.30
MM-PBSA (dielc = 1)	0.42	0.14	0.20	5.50	4.03
MM-PBSA (dielc = 2)	0.82	0.89	0.73	23.00	20.83
MM-SVRKB	0.74	0.89	0.73	1.76	1.03
		Set 1 complexes			
GBSA	0.44	0.47	0.27	2.32	1.49
PBSA	-0.51	-0.57	-0.45	2.22	1.64
SVRKB	0.83	0.82	0.69	1.43	0.59

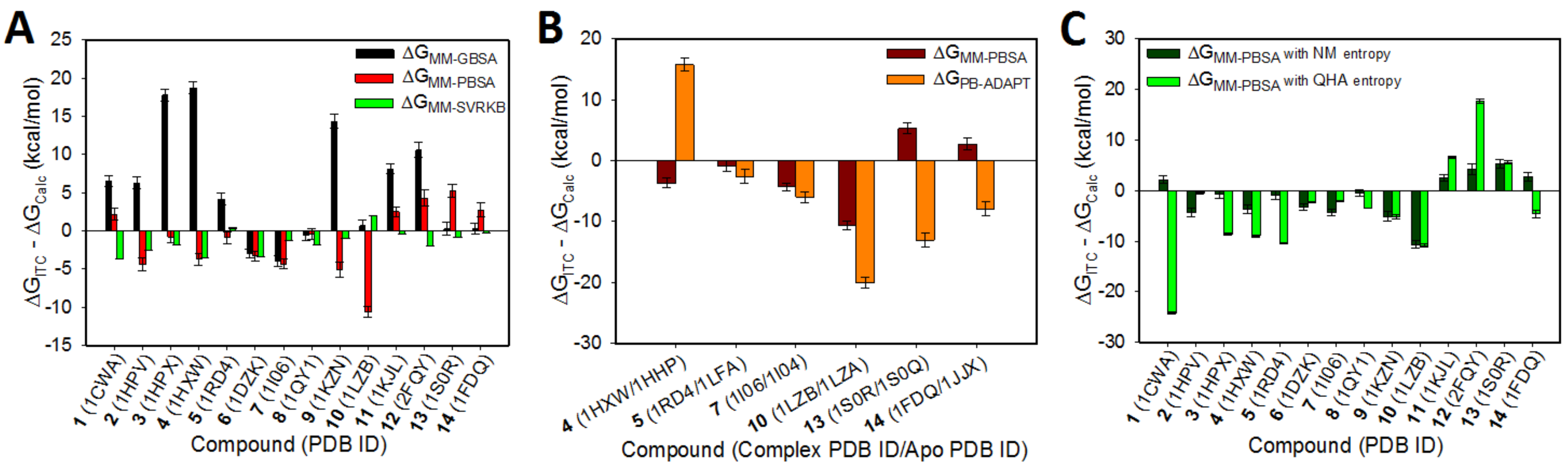


Figure 1.3 Calculated Free Energy Deviation from Experimental Energy

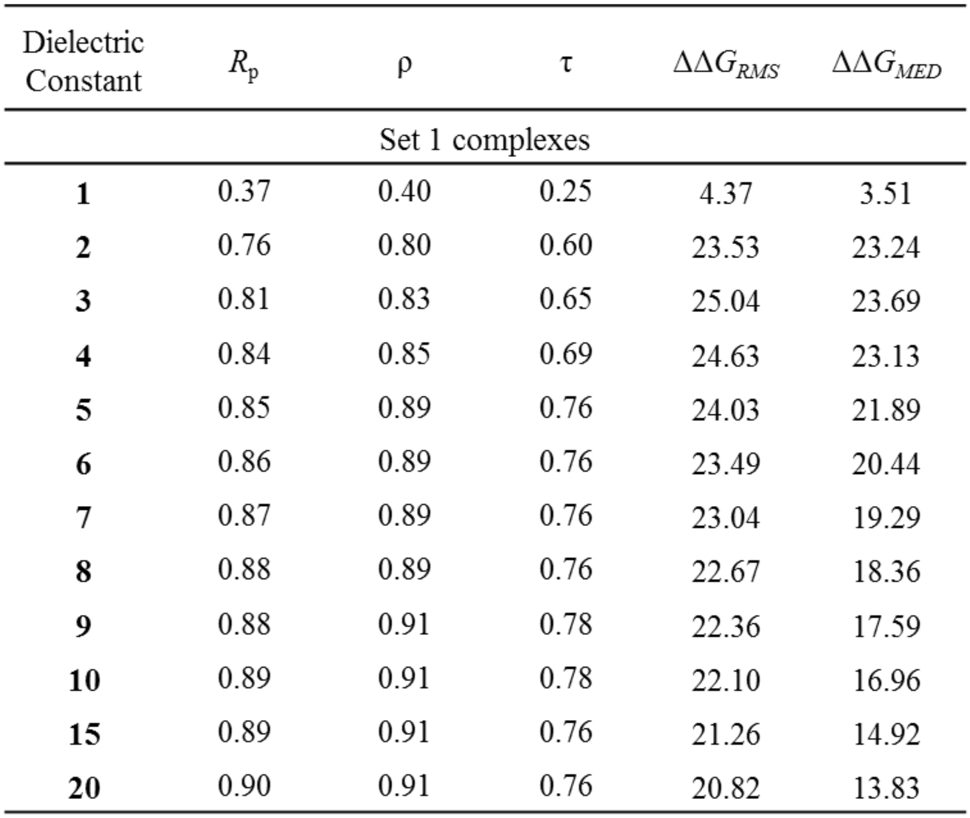
Finally, we compared calculations performed using harmonic versus quasiharmonic approaches for the entropy of binding. Two approaches were considered, namely normal mode analysis, or the use of a quasiharmonic approach where the entropies are determined by a covariance analysis of the fluctuations obtained from the MD simulations. The MM-PBSA free energies obtained with the normal mode analysis resulted in a $\Delta\Delta G_{\mathrm{RMS}}$ = 4.4 kcal•mol^{-1} when compared with the ITC free energy, and a median of 3.5 kcal•mol^{-1} for $\Delta\Delta G$ (Figure 1.3C). On the other hand, the MM-PBSA free energies for the quasiharmonic approach led to a $\Delta\Delta G_{\mathrm{RMS}}$ of 10.1 kcal•mol^{-1} and a median value of 6.1 kcal•mol^{-1}.

1.3.2 Rank-Ordering Protein-Ligand Complexes

Performance to rank-order complexes was evaluated using three correlation metrics, namely the Pearson's correlation coefficient (R_{p}), Spearman's rho (ρ), and Kendall's tau (τ). Pearson's coefficient is the more traditional metric used to measure the correlation between observed and predicted affinities. Spearman's rho is a non-parametric measure of the correlation between the *ranked lists* of the experimental binding affinities and the scores. It ranges between -1 and 1. A negative value corresponds to inverse correlation while a positive value suggests correlation between the variables. Kendall's tau (τ) was also considered to assess rank-ordered correlation as suggested by Jain and Nicholls.[41] τ has the advantage of being more robust and can be more easily interpreted. It corresponds to the probability of having the same trend between two rank-ordered lists.

It is interesting that despite the better performance of MM-PBSA in predicting the absolute free energy, the opposite is observed for rank-ordering. All three correlation coefficients metrics were significantly higher for MM-GBSA ($R_p = 0.75$; $\rho = 0.68$; $\tau = 0.52$) compared with MM-PBSA ($R_p = 0.37$; $\rho = 0.40$; $\tau = 0.25$) (Table 1.3, Figure 1.4A). At higher dielectric constants, the correlations for MM-PBSA significantly improves (Table 1.4). A mere doubling of the dielectric constant from 1 to 2 led to a similar increase in the correlation factors ($R_p = 0.77$; $\rho = 0.81$; $\tau = 0.65$). Further increase of the dielectric beyond two results in smaller increases in performance, as illustrated by the correlations at a dielectric constant of 20 ($R_p = 0.90$; $\rho = 0.91$; $\tau = 0.76$). But inspection of the components of the free energy (Table 1.6) reveals that this increase in performance is not due to more accurate representation of the electrostatic component of the free energy. Instead, it is attributed to the significantly smaller contributions of the electrostatic energy at higher dielectric constants. An increase in the dielectric constant reduced ΔE_{ELE} and ΔE_{PB} by a factor of $1/\epsilon$ and $1/\epsilon^2$, respectively, where ϵ is the dielectric constant. As a results, the lower contributions from the electrostatic component results in a free energy component that is dominated by the non-polar and entropy terms. SVRKB applied to MD structures (MM-SVRKB) showed better performance than MM-GBSA ($R_p = 0.77$; $\rho = 0.81$; $\tau = 0.65$) (Figure 1.4C, Figure 1.6A). Interestingly, free energies that included the adaptation energy ($\Delta G_{PB\text{-}ADAPT}$) exhibited dramatic improvement over MM-PBSA ($R_p = 0.95$; $\rho = 0.89$; $\tau = 0.73$) (Table 1.3, Set 2, Figure 1.4B). $\Delta G_{\mathrm{PB\text{-}ADAPT}}$ correlations are also better than MM-SVRKB ($R_p = 0.74$; $\rho = 0.89$; $\tau = 0.73$).

Table 1.4 Correlation Coefficients for MM-PBSA Calculations with Different Dielectric Constants



Dielectric Constant	R_p	ρ	τ	$\Delta\Delta G_{RMS}$	$\Delta\Delta G_{MED}$
			Set 1 complexes		
1	0.37	0.40	0.25	4.37	3.51
2	0.76	0.80	0.60	23.53	23.24
3	0.81	0.83	0.65	25.04	23.69
4	0.84	0.85	0.69	24.63	23.13
5	0.85	0.89	0.76	24.03	21.89
6	0.86	0.89	0.76	23.49	20.44
7	0.87	0.89	0.76	23.04	19.29
8	0.88	0.89	0.76	22.67	18.36
9	0.88	0.91	0.78	22.36	17.59
10	0.89	0.91	0.78	22.10	16.96
15	0.89	0.91	0.76	21.26	14.92
20	0.90	0.91	0.76	20.82	13.83

MM-GBSA and MM-PBSA calculations are performed on multiple structures collected from MD simulations. Typically, snapshots are selected at regular intervals. We wondered how MD snapshots can be pre-scored to improve the MM-PBSA or MM-SVRKB results. We had previously developed a scoring approach (SVMSP) to distinguish between native and non-native binding modes.[21] Scoring of MD snapshots with SVMSP is expected to enrich these structures for native-like complexes. SVMSP was used to score all snapshots from MD simulations for each of the 14 targets considered in this work. A total of 500 complexes with the top SVMSP scores were selected for MM-GBSA calculations. The combined SVMSP//MM-GBSA scoring did not improve the predictive abilities of MM-GBSA suggesting that the GB method is less sensitive to the structure used in the calculation (Table 1.3, Figure 1.6B). However, a dramatic boost in performance is observed for SVMSP//MM-PBSA *(*Table 1.3, Figure 1.6B*)*. In fact, an increase of 0.39, 0.41, and 0.38 is seen for R_p, ρ and τ, respectively. In Set 2, SVMSP//MMPBSA's prediction of the binding affinity trend is as good as MM-PBSA$_{\mathrm{ADAPT}}$. Components of the MM-PBSA or MM-GBSA calculations are insightful as they provide insight into the free energy of binding (Table 1.6). But an important question is whether these components correlate with the experimentally-determined thermodynamic parameters provided by ITC. Each component of the MM-GBSA and MM-PBSA calculations is plotted against the ITC free energy. It was interesting, but not completely surprising,[42, 43] that the non-polar components of the binding affinity correlated with the ITC free energy (Table 1.5 and Figure 1.5). The correlation coefficients were R_p = 0.89, ρ = 0.90, τ = 0.76 and $\mathrm{R_p}$ = 0.88, ρ = 0.89, τ = 0.74 for the van der Waals energy (ΔE_{VDW}) and the non-polar component of the solvation free energy (ΔE_{NP}), respectively. There was no correlation between the electrostatic components of the free energy (ΔE_{ELE}) and the ITC free energy. There was also no correlation between the reaction field energy calculated by PB (ΔE_{PB}) and the ITC free energy. This is consistent with previous results that showed that the non-polar component of the free energy was a significantly better predictor of the stability of protein-protein complexes than the electrostatic component.[42, 43] Finally, there was no correlation between molecular weight of ligand and binding affinity (R_p = -0.51, ρ = -0.65, τ = -0.51).

Table 1.5 Components of Free Energy Calculations

PDB	Ligand No.	ΔE_{ELE}		ΔE_{VDW}	ΔE_{NP}	ΔE_{PB}		$T\Delta S_{NM}$	$T\Delta S_{QHA}$
		dielc = 1	dielc =2			dielc = 1	dielc =2		
1CWA	**1**	-34.5±0.3	-17.2±0.2	-58.1±0.1	-6.7±0.01	54.9±0.3	14.4±0.1	-31.3±0.7	-54.7
1HPV	**2**	-35.2±0.3	-17.4±0.1	-60.4±0.2	-6.8±0.01	63.4±0.2	15.2±0.0	-30.9±0.8	-14.9
1HPX	**3**	-43.9±0.4	-21.3±0.2	-70.7±0.2	-8.1±0.01	77.6±0.4	18.0±0.1	-32.7±0.8	-35.8
1HXW	**4**	-41.8±0.3	-20.9±0.2	-75.2±0.2	-8.5±0.01	79.4±0.3	19.0±0.1	-36.2±0.7	-30.9
1RD4	**5**	-7.4±0.2	-4.2±0.1	-53.9±0.2	-6.4±0.01	33.3±0.2	8.8±0.1	-24.6±0.8	-41.3
1DZK	**6**	-3.3±0.1	-1.6±0.0	-27.5±0.1	-3.9±0.01	13.0±0.1	4.0±0.0	-15.8±0.6	-43.1
1I06	**7**	-1.3±0.1	-0.6±0.0	-21.4±0.1	-3.6±0.01	7.7±0.1	2.2±0.0	-14.6±0.7	-13.4
1QY1	**8**	-3.5±0.1	-1.7±0.0	-26.9±0.1	-3.9±0.01	10.8±0.1	3.0±0.0	-15.9±0.7	-22.1
1KZN	**9**	-38.1±0.4	-18.5±0.2	-63.5±0.2	-6.8±0.02	74.2±0.4	17.8±0.1	-29.4±0.9	-43.0
1LZB	**10**	-57.3±0.9	-28.9±0.4	-38.5±0.3	-5.3±0.02	78.4±0.9	18.7±0.2	-26.5±0.6	-26.9
1KJL	**11**	-61.6±0.3	-31.2±0.2	-21.2±0.1	-3.7±0.01	57.7±0.3	13.0±0.1	-20.7±0.6	-18.5
2FQY	**12**	-65.1±0.3	-32.4±0.1	-34.7±0.1	-4.3±0.01	69.7±0.4	16.6±0.1	-21.2±0.9	-35.3
1S0R	**13**	42.7±0.6	21.0±0.3	-19.3±0.1	-3.0±0.01	-48.9±0.5	-11.2±0.1	-16.9±0.8	-15.2
1FDQ	**14**	-93.8±1.0	-45.4±0.5	-41.2±0.2	-5.8±0.01	100.7±0.6	24.5±0.1	-27.3±0.6	-24.0

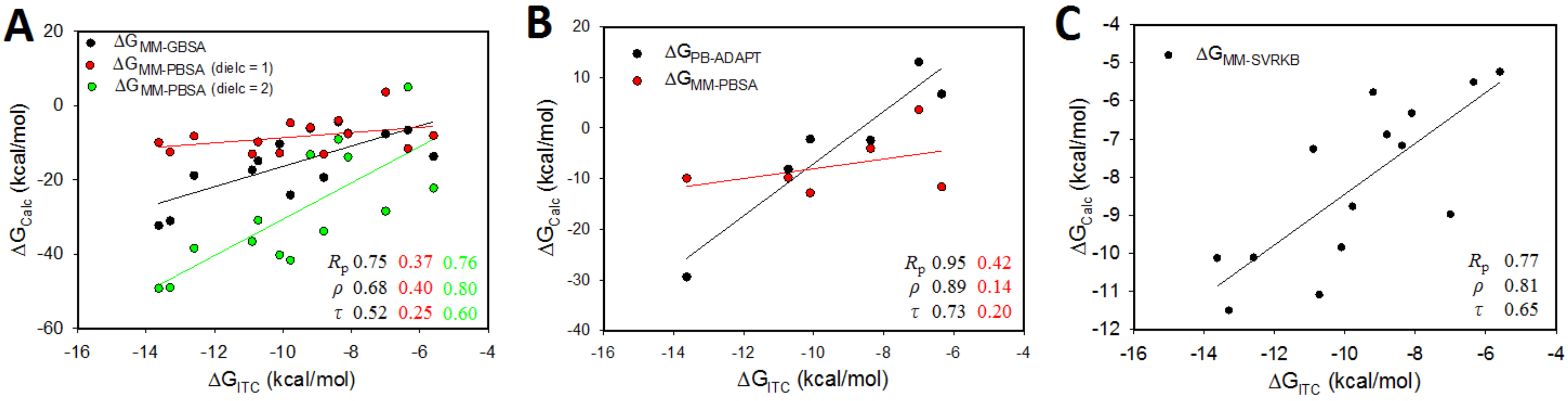


Figure 1.4 Regression Plots between Experimental Free Energy and Calculated Free Energy

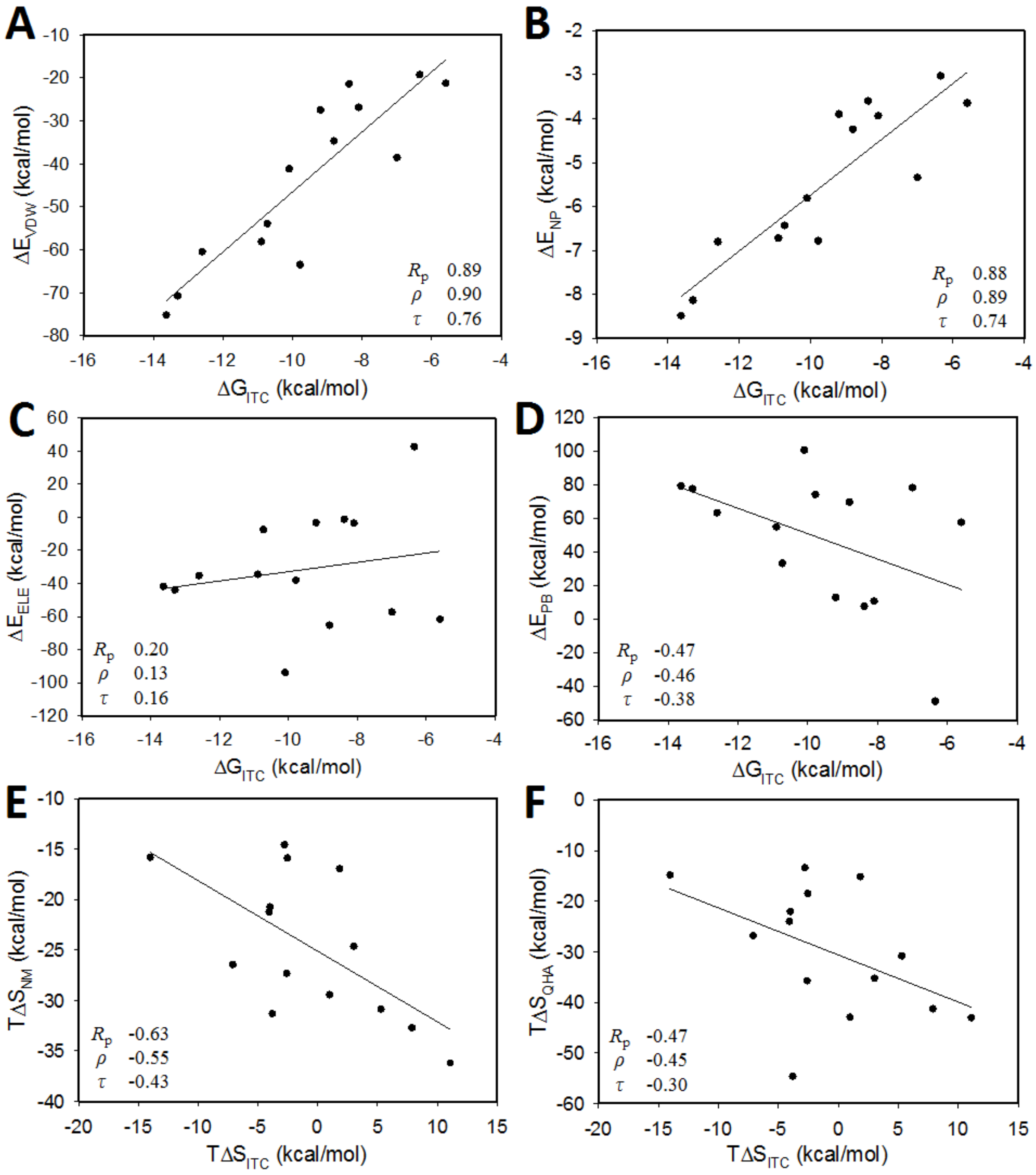
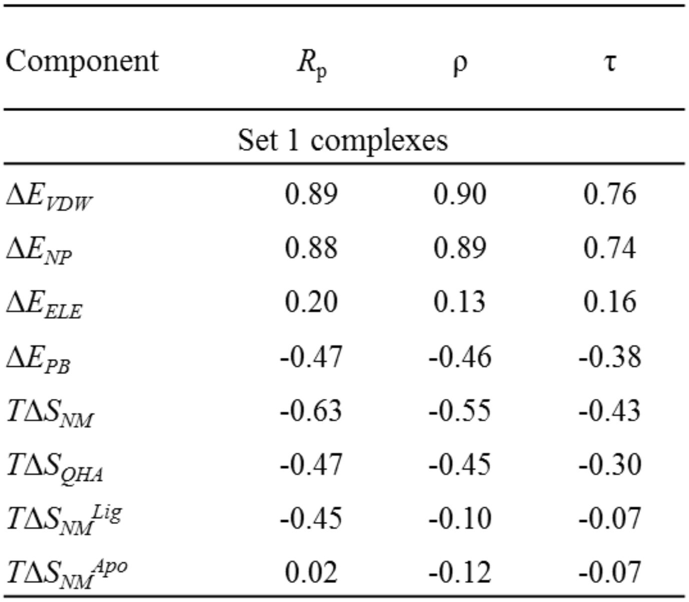


Figure 1.5 Regression Plots between Experimental Free Energy and Components of Calculated Free Energy

Table 1.6 Correlation Coefficients for Components of Free Energy Calculations



Component	R_p	ρ	τ
Set 1 complexes			
ΔE_{VDW}	0.89	0.90	0.76
ΔE_{NP}	0.88	0.89	0.74
ΔE_{ELE}	0.20	0.13	0.16
ΔE_{PB}	-0.47	-0.46	-0.38
$T\Delta S_{NM}$	-0.63	-0.55	-0.43
$T\Delta S_{QHA}$	-0.47	-0.45	-0.30
$T\Delta S_{NM}{}^{Lig}$	-0.45	-0.10	-0.07
$T\Delta S_{NM}{}^{Apo}$	0.02	-0.12	-0.07

The entropy component of the MM-GBSA and MM-PBSA calculations follows a similar trend to the true entropy of binding. The availability of ITC data for each of our systems provides an opportunity to compare computed versus experimental entropy. For MM-PBSA and MM-GBSA, the entropy is typically determined using either normal modes or a quasiharmonic analysis. Figure 1.3C shows that these two approaches result in different free energies with overall better agreement for the free energy from the normal mode analysis. The correlations MM-PBSA free energies using normal mode was R_p = 0.37, ρ = 0.40, τ = 0.25, compared with R_p = -0.20, ρ = -0.30, τ = -0.25 for the quasiharmonic analysis. The normal mode and quasiharmonic entropies are compared to the experimental entropy. A plot of $T\Delta S_{ITC}$ versus $T\Delta S_{NM}$ or $T\Delta S_{QHA}$ shows that computed and experimental entropies are inversely correlated with correlation coefficients of (R_p = -0.63; ρ = -0.55; τ = -0.43) and (R_p = -0.47; ρ = -0.45; τ = -0.30), respectively (Figure 1.5E and 1.5F). No change is observed when the entropy change of ligand only ($T\Delta S_{\mathrm{NM}}^{\mathrm{Lig}}$), or receptor only ($T\Delta S_{\mathrm{NM}}^{\mathrm{Apo}}$) are compared to the ITC entropy (Table 1.5).

The performance of MM-GBSA and MM-PBSA is compared to GBSA and PBSA, which correspond to calculations performed on a single crystal structure for each of the complexes in Table 1.1. Correlation coefficients reveal that both GBSA and PBSA perform poorly in rank-ordering complexes when a single crystal structure is used (Figure 1.6C). For GBSA all three correlation factors were smaller than 0.5 (R_{p} = 0.44; ρ = 0.47; τ = 0.27), and for PBSA, predicted and experimental data were inversely correlated (R_{p} = -0.51; ρ = -0.57; τ = -0.45). SVRKB, on the other hand, performed well consistent with our previous study [44] (R_{p} = 0.83; ρ = 0.82; τ = 0.69).

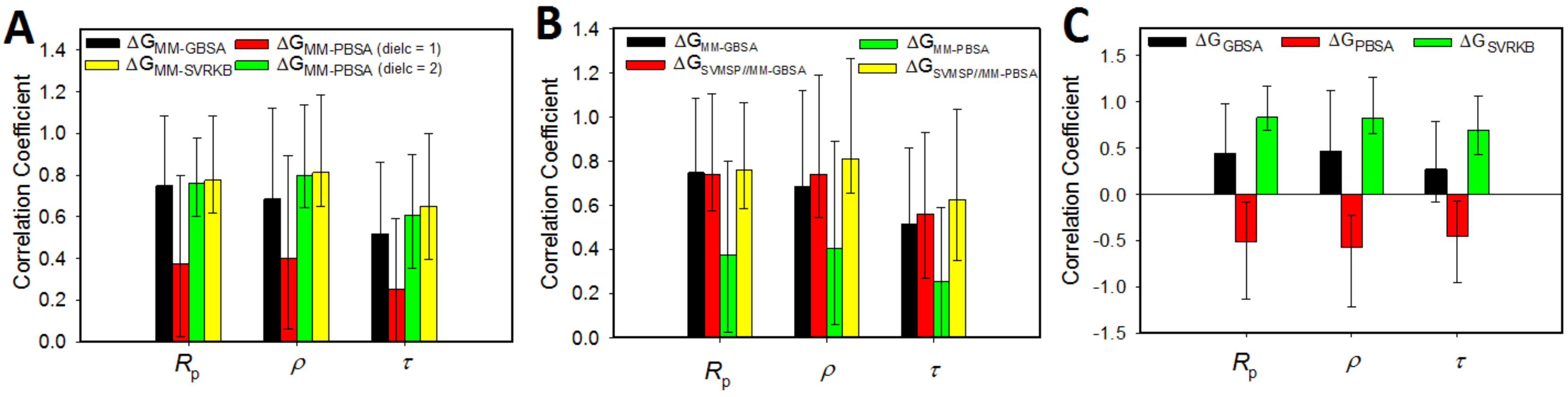


Figure 1.6 Performance on Pearson’s Rp, Spearman ρ and Kendall τ Correlation Coefficients

1.4 Discussion

MM-GBSA and MM-PBSA calculations are applied to a diverse set of 14 ligands bound to 11 different proteins. A unique aspect of this work is that (i) a diverse set of proteins and ligands are used in contrast to most studies that compare ligands bound to the same protein; (ii) all complexes were previously solved by x-ray crystallography and binding was characterized by ITC. The ligands included small organic compounds, cyclic and linear peptides, fragment-like small molecules, and carbohydrates. Most free energy calculations did not accurately reproduce the ITC free energy. But there were several cases that were in excellent agreement with ITC: three complexes for MM-PBSA calculations, and four for MM-GBSA. Overall, MM-PBSA resulted in less deviation from the experimental data than MM-GBSA. But the opposite was observed for rank-ordering. MM-GBSA correlated significantly better with the ITC free energy (R_p = 0.75; ρ = 0.68; τ = 0.52) when compared to MM-PBSA (R_p = 0.37; ρ = 0.40; τ = 0.25). The non-polar terms (ΔG_{VDW} and ΔG_{NP}) showed strong correlation with the experimental free energy (R_p = 0.89; ρ = 0.90; τ = 0.76 and R_p = 0.88; ρ = 0.89; τ = 0.74 for ΔG_{VDW} and ΔG_{NP}, respectively). An increase in the dielectric constant for the MM-PBSA calculations worsened agreement of the computed and experimental free energies. However, rank-ordering appeared to significantly improve upon increase of the dielectric constant. But close inspection of the components of the free energy reveals that this increase is attributed to the lower contribution of electrostatics as a results of an increase of the dielectric constant. The Coulomb and electrostatic terms are inversely proportional to the dielectric constant and to the square of the dielectric constant, respectively. Less contribution from electrostatics leads to a free energy that is dominated by the non-polar and entropy components resulting to better performance. There was no correlation between the electrostatic components of the MM-GBSA and MM-PBSA free energy and the ITC free energy.

Two models for the entropy were considered, normal mode and quasiharmonic. Normal mode analysis assumes that each structure is at a potential energy minimum. Quasiharmonic analysis is based on a covariance analysis of the atomic fluctuation. Our data showed that the free energies using normal mode analysis correlated significantly better than free energies using quasiharmonic analysis. A possible explanation is that the simulations used in this study may not have been sufficiently long to ensure convergence of quasiharmonic analysis. Neither the normal mode entropy nor the quasiharmonic entropies correlated with the ITC entropy. This is likely due to the fact that the ITC entropy includes both solvation and configurational entropy,[45] while the computed entropy only includes the configurational entropy. The solvation entropies may be indirectly captured by the other terms of the MM-GBSA or MM-PBSA free energy.

Small-molecule binding often induces conformational change to the target protein. This adaptation energy is often ignored in MM-GBSA or MM-PBSA calculations as a single simulation is carried out starting with the complex structure. The structure of the apo protein is extracted from the complex. We studied the effect of this adaptation energy by running a separate simulation for ligand, apo and complex structures. We did this for 6 of the 14 complexes whose apo structure was solved independently by x-ray crystallography. Overall, this resulted in poorer agreement with the ITC data when comparing the absolute values of the free energies. However, the adaptation energy resulted in a significant boost in rank-ordering. The $\Delta G_{\text{PB-ADAPT}}$ resulted in a Pearson (Spearman) correlation of 0.95

(0.89) compared with 0.89 (0.83) for $\Delta G_{MM\text{-}GBSA}$ and 0.42 (0.14) for $\Delta G_{\mathrm{MM\text{-}PBSA}}$. The $\Delta G_{\mathrm{PB\text{-}ADAPT}}$ showed the best rank-ordering among all methods that were tested in this work.

Typically snapshots for MM-GBSA or MM-PBSA calculations are selected at regular intervals in an MD simulation. We wondered whether different approaches for selecting structures will influence the free energy of binding. We used a recently-developed machine learning-based scoring approach (SVMSP) to pre-score all the snapshots in a trajectory. SVMSP is trained from crystallography and docked protein-decoy structures to classify protein-ligand complexes.[21] It is therefore expected that the method will enrich MD snapshots for native-like structures. Rank-ordering of snapshots (Table 1.3) had little influence on the MM-GBSA free energies (Table 1.3). However, rank-ordering with MM-PBSA calculations improved significantly from $R_{\mathrm{p}} = 0.37$, $\rho = 0.40$, $\tau = 0.25$ for snapshots selected at regular intervals to $R_{\mathrm{p}} = 0.76$, $\rho = 0.81$, $\tau = 0.63$ for SVMSP-selected snapshots. These results indicate that the Poisson-Boltzmann calculations are more sensitive to the quality of the structure than MM-GBSA.

In sum, MM-GBSA and MM-PBSA methods come short in reliably reproducing the free energy of binding. However, these methods can perform remarkably well for rank-ordering diverse set of compounds. MM-GBSA can perform well by merely using snapshots from an MD simulation of the complex, while MM-PBSA is significantly more sensitive to the structures used. Filtering MD structures with scoring functions to enrich for native-like complexes results in excellent rank-ordering by MM-PBSA. In addition, running separate simulations of the receptor also improves the rank-ordering abilities of MM-PBSA. While previous studies have found that rank-ordering performance for MM-PBSA improves with increasing dielectric constant, we found that this is mainly due to the smaller contributions of electrostatics as a result of increasing the dielectric constant (at ε = 5, for example, the Coulomb energy is reduced by a factor of 1/5 and the PB solvation energy by 1/25). It was remarkable that the non-polar components correlated very well with the free energy. The combination of non-polar and entropy also correlated very well with the free energy, which is why overall correlation improved at higher dielectric constants for MM-PBSA. Finally, the MM-PBSA entropy does not correlation with the ITC entropy.

CHAPTER 2. ENRICHMENT OF CHEMICAL LIBRARIES DOCKED TO PROTEIN CONFORMATIONAL ENSEMBLES AND APPLICATION TO ALDEHYDE DEHYDROGENASE 2

2.1 Introduction

Structure-based virtual screening is widely used in the search for small molecules to probe the function of proteins and nucleic acids in chemical biology and drug discovery.[46,47] Typically, a chemical library is docked to a pocket on a target structure, followed by the ranking of the resulting protein-compound complexes in a process known as scoring. The top candidates are acquired or prepared for experimental validation. Several scoring methods have been developed over the years; these include empirical,[48-56], knowledge-based,[16, 57-65] and force field-based.[66-73] We recently developed a new scoring approach that combines machine learning and statistical knowledge-based potentials for rank-ordering Support Vector Regression Knowledge-Based (SVRKB)[74] and database enrichment Support Vector Machine SPecific (SVMSP).[21] The former is regression-based and trained on crystal structures using corresponding experimental binding affinities, while the latter is based on classification and is trained strictly on three-dimensional structures of protein-ligand complexes using both actives and decoys.

Part of the challenge with the use of structure-based virtual screening is protein flexibility.[75-77] It is ignored in the majority of cases by docking compounds strictly to a crystal structure,[78-80] although there are examples that have used multiple crystal structures,[81-83] NMR structures,[84-87] or a combination of the two.[11, 12] Albeit less common, the use of molecular dynamics (MD) simulations to generate an ensemble of structures has also been reported in virtual screening efforts that have led to active compounds.[88, 89] Our recent study led to the discovery of small-molecule inhibitors of a tight protein-protein interaction by docking a chemical library to protein structures collected from explicit-solvent MD simulations.[90] Several studies have attempted to gain a deeper understanding of the role of MD structures on chemical database enrichment.[76-78]

Here, we conduct an in-depth study to investigate the SVMSP scoring approach in chemical database enrichment using structures collected from explicit-solvent MD simulations. We explore enrichment for individual and ensemble of snapshots. In addition, we follow an innovative approach that explores the use of MD structures for the development of scoring functions for virtual screening. Also, we investigate the *a priori* identification of MD snapshots with high enrichment power from an MD simulation. Finally, SVMSP scoring of protein-compound MD structures is applied in the virtual screening of commercial libraries against the mitochondrial aldehyde dehydrogenase 2 enzyme (ALDH2) enzyme. ALDH2 catalyzes the NAD$^+$-dependent oxidation of a broad spectrum of endogenous and biogenic aldehydes to their corresponding carboxylic acids. ALDH2 is commonly associated with its role in alcohol metabolism, but it has been suggested as a potential target for a variety of diseases that include addiction and cancer. Top candidates that emerged from virtual screening were acquired and tested for inhibition of enzyme activity.

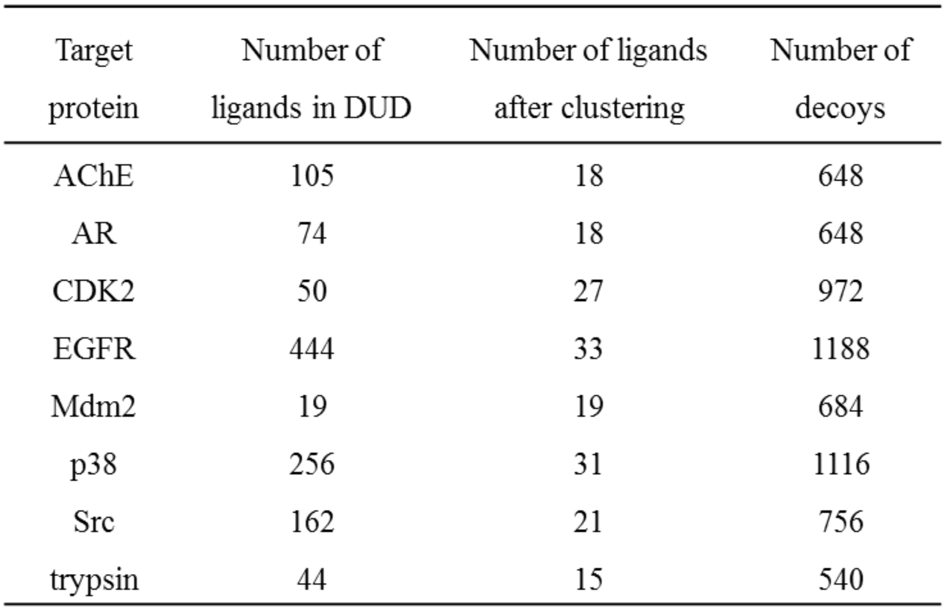
2.2 Materials and Methods

2.2.1 Data Set Preparation

For the enrichment study, 7 protein structures from the Directory of Useful Decoys (DUD)[92] and one from our in-house validation set, namely MDM2 (mouse double minute 2 homolog) (PDB code: 1RV1), were used to assess the performance of scoring functions. The DUD proteins include acetycholinesterase AChE (PDB code: 1EVE), human androgen receptor AR (PDB code: 1XQ2), human cyclin-dependent kinase 2 CDK2 (PDB code: 1CKP), human epidermal growth factor receptor EGFR (PDB code: 1M17), human mitogen-activated protein kinase 14 known as p38 (PDB code 1KV2), human proto-oncogene tyrosine-protein kinase Src (PDB code: 2SRC), and cationic trypsin (PDB code: 1BJU).

To ensure diversity among the active compounds in DUD, the compounds were clustered by chemical similarity. FP3 fingerprints were generated for every ligand with Open Babel.[93] A Tanimoto coefficient matrix was calculated for each target by Open Babel. Hierarchical clustering method was applied with the *cluster* package in python2.6 to cluster compounds. The *getlevel* threshold in the *cluster* package was set to 0.1, which means that any two compounds with Tanimoto coefficient deviation less than 0.1 will be included into the same cluster. The number of compounds after clustering for each target is shown in Table 2.1. The ratio of active ligands to decoys ($N_{ligands}/N_{decoys}$) was kept to 1:36 following the convention adopted in DUD.

Table 2.1 Validation Set for Enrichment Studies



Target protein	Number of ligands in DUD	Number of ligands after clustering	Number of decoys
AChE	105	18	648
AR	74	18	648
CDK2	50	27	972
EGFR	444	33	1188
Mdm2	19	19	684
p38	256	31	1116
Src	162	21	756
trypsin	44	15	540

2.2.2 MD Simulations

All the criteria of MD simulations were set to be the same as described in section 1.2.2.

For each target, 4 independent 6 ns simulations were performed. MD snapshots were collected every 1 ps yielding 6,000 structures per trajectory, or 24,000 structures in total.

The first 1 ns in each trajectory was discarded for equilibration. A set of 500 snapshots was extracted at regular intervals from the resulting 20 000 snapshots for each protein. Atoms within 5 Å around ligand in crystal structure were considered as pocket atoms. The 500 trajectory frames were further clustered into groups based on pairwise similarity measured by root-mean-square deviation (RMSD) of pocket atoms with *ptraj* program in AMBER. The hierarchical clustering algorithm was used to cluster all 500 structures into sets of 5, 10, 20, 30, 50, 100, and 250 structures.

2.2.3 Scoring Protein-Ligand Complexes

SVMSP models were built in the same manner as described in section 1.2.1. The $SVMSP_{KINASE}$ model was developed for kinase targets only. The positive set included only kinase structures refined sc-PDB database, consisting of 763 crystal structures. The negative set for $SVMSP_{KINASE}$ was the same as SVMSP model. $SVMSP_{MD}$ models were created by using decoy compounds docked to MD snapshots for the negative training set. A total of 5,000 randomly selected compounds were docked to each MD snapshot. The positive set consisted of the same structures as were used to develop SVMSP. When the positive set employed kinase only positive set, the model was called $SVMSP_{KINASE\text{-}MD}$.

2.2.4 Compound Docking

All the molecular docking reported in this work was done using AutoDock Vina.[94] The *exhaustiveness* parameter of Vina program was set to default value of 8. A maximum number of 9 binding modes were generated, with maximum energy difference between the best and the worst binding mode set to 3 kcal·mol^{-1}. The docking pose with the lowest energy estimated by Vina was selected as the best binding pose for further scoring. The box size was 19 Å.

2.2.5 Receiver Operating Characteristic Plot and Statistical Analysis

A tool that is commonly used to assess the performance of a scoring function is the receiver operating characteristic (ROC) plot.[95] An ROC curve is constructed by ranking the docked complexes, selecting a set of compounds starting from the highest scoring compounds, and counting the number of active compounds. This process is repeated a number of times for a gradually increasing set of compounds selected from the ranked list. In an ROC plot, the farther away the curve is from the diagonal, the better the performance of the scoring function. The area under the ROC curve, which we refer to as ROC-AUC, can also be used as a representation of the performance of the scoring function. A perfect scoring function will result in an area under the curve of 1, while a random classification will have an ROC-AUC of 0.5.

2.2.6 ALDH2 Virtual Screening

The initial coordinates of ALDH2 taken into the molecular dynamic (MD) simulations were obtained from RCSB Protein Data Bank (PDB code: 1O04). The PDB file was imported into Maestro (version 9.3, Schrödinger, LLC, New York, NY, 2012), prepared using the Protein Preparation Wizard.[96] Bond orders were assigned, hydrogen atoms were added, disulfide bonds were created, and selenomethionines were converted to methionines. Crystal water molecules were kept. MD simulations were carried out as described above. By assigning different initial velocities, 5 independent 7 ns simulations were carried out for a total length of 35 ns simulation. The first 2 ns of each trajectory were considered as part of the equilibration process and discarded. MD snapshots were saved every 1 ps yielding 5,000 structures per trajectory. In total, 25,000 snapshots were collected. The snapshots were clustered into 75 sets by the *ptraj* program using atoms around the active site pocket. The hierarchical clustering algorithm was used for the clustering. Among 75 clusters, the top 50 clusters that had the most snapshots were selected. A representative snapshot was chosen for virtual screening from each of the 50 clusters. Around 50,000 compounds from the ChemDiv80 [24] library were docked to each of the 50 snapshots using Vina.

Docked receptor-ligand complexes were rescored using the SVMSP scoring function. The 5,000 randomly picked compounds from ChemDiv library docked to ALDH2 crystal structure were used as negative set to build the SVMSP model. For each compound, the highest score among all the snapshots within the cluster was used to rank all the compounds. The top scoring 5,000 compounds from the ChemDiv80 library were selected. Canvas similarity and clustering script[97, 98] in Maestro program were applied to cluster the top compounds. Atom triplet fingerprint type with 32-bit precision was used. Atom typing scheme was Daylight invariant atom types. The single linkage method was used to generate 150 clusters. The compounds representing the 150 cluster center were selected for further experiments.

2.2.7 ALDH2 Inhibition Assay

Compounds were first screened using a high-throughput dehydrogenase assay to measure the production of NADH via fluorescence (excitation λ = 340 nm, emission λ = 465 nm) on an Ultra384 plate reader over a 10 min period. The screening assay used 20 nM ALDH2, 30 μM propionaldehyde, 100 μM NAD$^+$, and 50 μM compound in 25 mM BES, pH 7.5 with 2% (v/v) DMSO in a 96-well black plate with a final volume of 200 μL. The compounds that showed inhibition in this assay were then tested for their effect on ALDH2 dehydrogenase activity at 50 μM concentration using a Beckman DU-640. The dehydrogenase assay used 150 nM ALDH2, 100 μM propionaldehyde, and 200 μM NAD^+ in 25 mM BES, pH 7.5 with 1% (v/v) DMSO. The assays were monitored at 340 nm for the increase in NADH production (extinction coefficient = 6.22 mM^{-1} cm^{-1}). If compounds showed inhibition at this concentration, assays to determine the concentration dependence of inhibition (IC_{50}) were performed. IC_{50} values toward inhibition of ALDH2 activity were measured with compound concentrations ranging from 0 to 100 μM. All IC_{50} values were determined by fitting to the 4-parameter logistics function in SigmaPlot (v12).

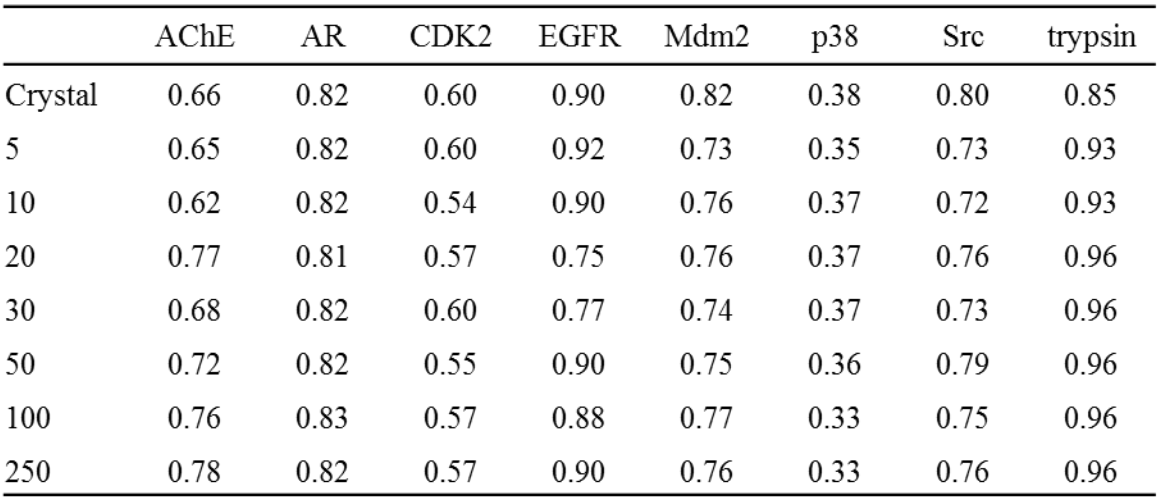
2.3 Results

2.3.1 Enrichment in the Conformational Ensemble

We were particularly interested in assessing how our scoring approach, SVMSP, affects enrichment of compound libraries docked to MD structures collected from explicit-solvent MD simulations. To that end, MD simulations were carried out for 8 proteins that included 7 proteins from the Directory of Decoys (DUD), namely androgen receptor (AR), acetylcholinesterase (AChE), trypsin, cyclin-dependent kinase 2 (CDK2), epidermal growth factor receptor (EGFR), mitogen-activated protein kinase (p38) and proto-oncogene protein tyrosine kinase (Src). One additional protein, MDM2, which is involved in a protein-protein interaction with p53, was added to the list (Table 2.1). A total of 24 ns of simulation was carried out for each protein. In each case, a set of 500 structures was collected at regular intervals from 20 000 snapshots generated by the simulations. Decoy and active compounds obtained from DUD (or generated for MDM2) were docked to the 500 MD snapshots with AutoDock Vina. The resulting complexes were scored with SVMSP,[21] ChemScore,[99] GoldScore,[100] and GBSA.[10]

For the crystal structures, the ROC-AUC ranged from 0.38 for p38 to 0.90 for EGFR. The ROC-AUC for SVMSP was larger than 0.8 for five out of the eight proteins. For MD structures, there were two cases, trypsin and AChE, which showed a gradual increase in the ROC-AUC as the size of the cluster became larger (Figure 2.1). In the case of trypsin, the ROC-AUC nearly reached a value of 1 for the cluster for 250 structures. For AChE, an improvement of nearly 0.1 in ROC-AUC was observed when compared to the crystal structure. In the case of p38 kinase, CDK2, and AR, the performance remained constant at 0.33, 0.57, and 0.82, respectively. Src reveals an initial drop in ROC-AUC of about 0.1 units to 0.7, which does not change as the number of structures is increased. For EGFR, the ROC-AUC was constant for 5 and 10 snapshots but dropped by 0.2 units for 20 and 30 structures only to show an increase back to 0.9 for 50, 100, and 250 structures. The results suggest that a cluster of 50 snapshots is likely to result in the best performance across a set of diverse proteins for SVMSP. All data presented below uses the 50 MD snapshots unless otherwise stated.

Table 2.2 SVMSP Enrichment Performance of Different Cluster Size



	AChE	AR	CDK2	EGFR	Mdm2	p38	Src	trypsin
Crystal	0.66	0.82	0.60	0.90	0.82	0.38	0.80	0.85
5	0.65	0.82	0.60	0.92	0.73	0.35	0.73	0.93
10	0.62	0.82	0.54	0.90	0.76	0.37	0.72	0.93
20	0.77	0.81	0.57	0.75	0.76	0.37	0.76	0.96
30	0.68	0.82	0.60	0.77	0.74	0.37	0.73	0.96
50	0.72	0.82	0.55	0.90	0.75	0.36	0.79	0.96
100	0.76	0.83	0.57	0.88	0.77	0.33	0.75	0.96
250	0.78	0.82	0.57	0.90	0.76	0.33	0.76	0.96

Enrichment performance for the other three scoring approaches, namely GoldScore, ChemScore, and GBSA, were poor in all systems when the crystal structure was used. ROC-AUCs do not change significantly in all four scoring functions with respect to the size of the cluster (Figure 2.1 and Appendix A, Table A.3). For ChemScore, GoldScore, and GBSA rescoring, performance is similar to the crystal structure in each cluster. An exception is for GBSA in AR where a drop from 0.7 to 0.4 is observed.

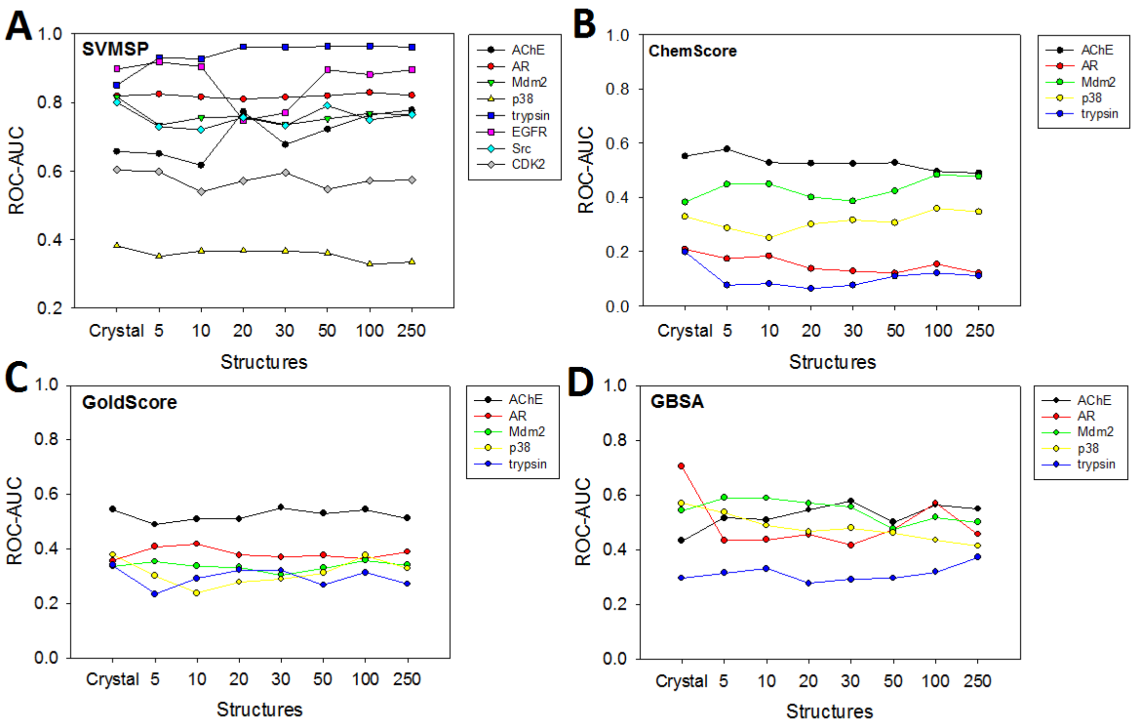


Figure 2.1 ROC-AUC Scores for Different Clusters of MD Structures

2.3.2 Enrichment for Individual Snapshots

The docking of all actives and decoys to 500 snapshots collected from the MD simulations of each protein in Table 2.1 provided an opportunity to explore enrichment for individual MD structures. ROC-AUC scores were determined for all 500 snapshots collected for each of the 8 target proteins in Table 2.1. The ROC-AUC for each snapshot was plotted against its structural deviation from the crystal structure measured by the root-mean-squared derivation (RMSD) (Figure 2.2). No direct correlation between ROC-AUC and RMSD is observed. This suggests that greater overall structural deviation from the crystal structure does not translate into lower or higher enrichment performance (Figure 2.2).

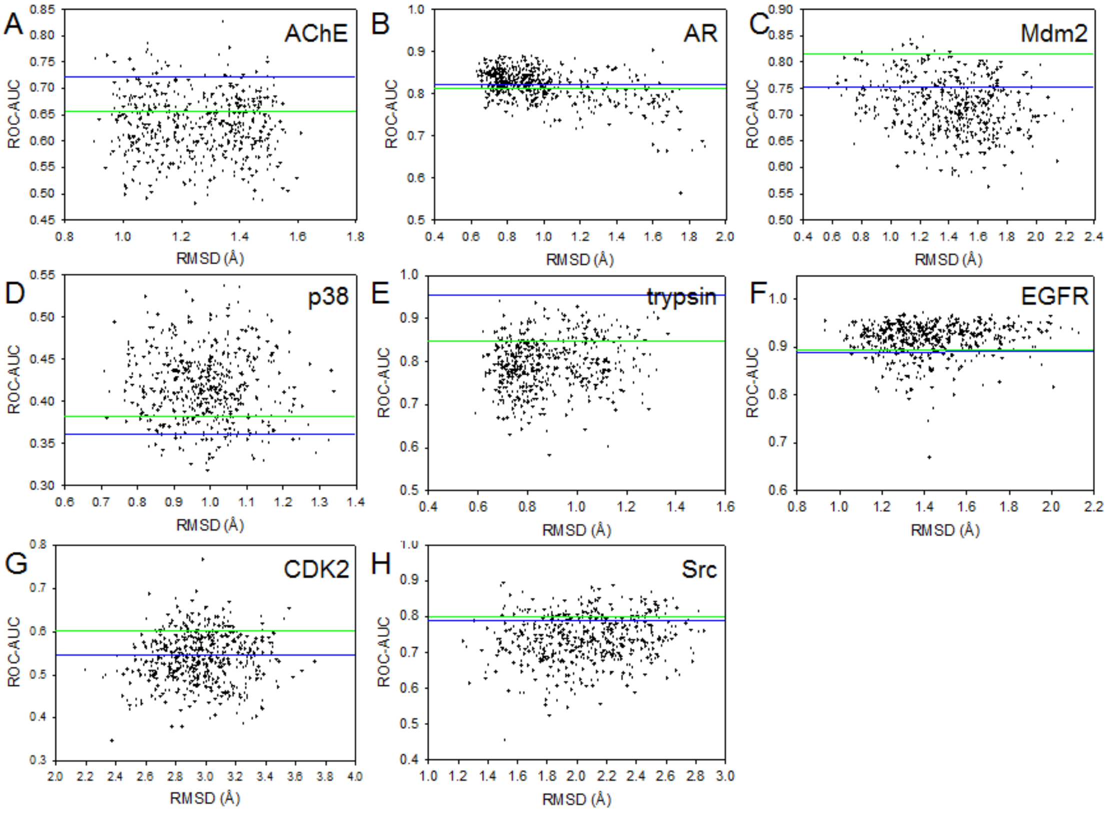
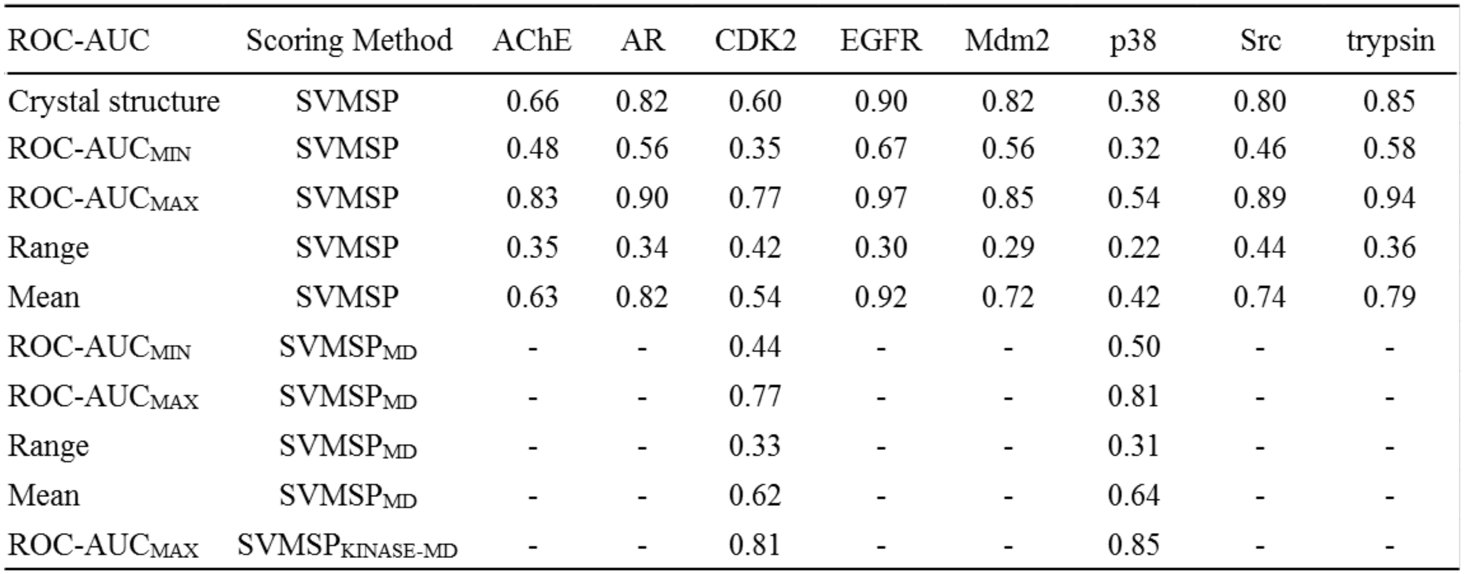


Figure 2.2 ROC-AUC for Individual Snapshots

Table 2.3 SVMSP Enrichment Performance for 500 MD Snapshots



ROC-AUC	Scoring Method	AChE	AR	CDK2	EGFR	Mdm2	p38	Src	trypsin
Crystal structure	SVMSP	0.66	0.82	0.60	0.90	0.82	0.38	0.80	0.85
ROC-AUC$_{MIN}$	SVMSP	0.48	0.56	0.35	0.67	0.56	0.32	0.46	0.58
ROC-AUC$_{MAX}$	SVMSP	0.83	0.90	0.77	0.97	0.85	0.54	0.89	0.94
Range	SVMSP	0.35	0.34	0.42	0.30	0.29	0.22	0.44	0.36
Mean	SVMSP	0.63	0.82	0.54	0.92	0.72	0.42	0.74	0.79
ROC-AUC$_{MIN}$	SVMSP$_{MD}$	-	-	0.44	-	-	0.50	-	-
ROC-AUC$_{MAX}$	SVMSP$_{MD}$	-	-	0.77	-	-	0.81	-	-
Range	SVMSP$_{MD}$	-	-	0.33	-	-	0.31	-	-
Mean	SVMSP$_{MD}$	-	-	0.62	-	-	0.64	-	-
ROC-AUC$_{MAX}$	SVMSP$_{KINASE-MD}$	-	-	0.81	-	-	0.85	-	-

What is notable from this data is the large fluctuation in the ROC-AUC among the 500 snapshots. Enrichment in several MD snapshots exceeded that of the corresponding crystal structure (Table 2.3). A total of 32, 55, 2, 81, 17, 76, 13, and 18% percent of the snapshots for AChE, AR, MDM2, p38, trypsin, EGFR, CDK2 and Src, respectively, exhibited better performance than the crystal structure. In some cases, there exists MD snapshots that significantly exceeded the enrichment power of the crystal structure. For example, for AChE, the snapshot with the maximum ROC-AUC (ROC-AUC$_{\mathrm{MAX}}$) was 0.83, nearly 0.2 higher than the crystal structure. A similar snapshot was identified for AR (ROC-AUC$_{\mathrm{MAX}}$ = 0.90), CDK2 (ROC-AUC$_{\mathrm{MAX}}$ = 0.77), EGFR (ROC-AUC$_{\mathrm{MAX}}$ = 0.97), MDM2 (ROC-AUC$_{\mathrm{MAX}}$ = 0.85), trypsin (ROC-AUC$_{\mathrm{MAX}}$ = 0.94), and Src (ROC-AUC$_{\mathrm{MAX}}$ = 0.74). Two proteins had poor enrichment both in the crystal (ROC-AUC of 0.38 and 0.60, respectively) and MD structures (ROC-AUC of 0.36 and 0.55, respectively). For these two proteins, ROC-AUC$_{\mathrm{MAX}}$ was 0.54 and 0.77, respectively.

2.3.3 Training Support Vector Machine Target Specific with MD Structure

Our SVMSP models have been developed entirely using protein-compound co-crystal structures (positive set) and compounds docked to the target crystal structure (negative set). We explored the possibility of using MD structures to develop SVMSP scoring models. To accomplish this, we followed the same protocol for developing the SVMSP models except that compounds in the negative set were docked to MD snapshots of the target of interest. We continue to use cocrystal structures for the positive set. The resulting SVMSP models (SVMSP$_{\mathrm{MD}}$) are tested on all 500 snapshots for two targets, namely p38 and CDK2. These targets were selected because of the poor enrichment that was observed in both X-ray and MD structures. A remarkable increase in the ROC-AUC for SVMSP$_{MD}$ is observed for p38 from 0.42 to 0.64 (Figure 2.3A and Table 2.3). ROC-AUC$_{MAX}$ was 0.81, compared to 0.54 using the crystal structure. A similar increase in performance was observed for CDK2 by 0.10. The mean ROC-AUC is 0.62, compared with 0.54 for SVMSP trained strictly with crystal structures. In fact, more than 29 snapshots were found to have an ROC-AUC greater than 0.7 for SVMSP$_{MD}$ in contrast to only one snapshot with the standard SVMSP approach. We also developed SVMSP$_{KINASE-MD}$ model applied using strictly kinase cocrystal structures for the positive set. An improvement in the mean ROC-AUC is observed in both cases by nearly 0.05 over SVMSP$_{MD}$ (Figure 2.3). In addition, a significantly greater number of snapshots with ROC-AUC greater than 0.70 were identified (6% for SVMSP$_{MD}$ versus 20% for SVMSP$_{KINASE-MD}$). The maximum ROC-AUC also increased by 0.05 relative to SVMSP$_{MD}$ (Table 3).

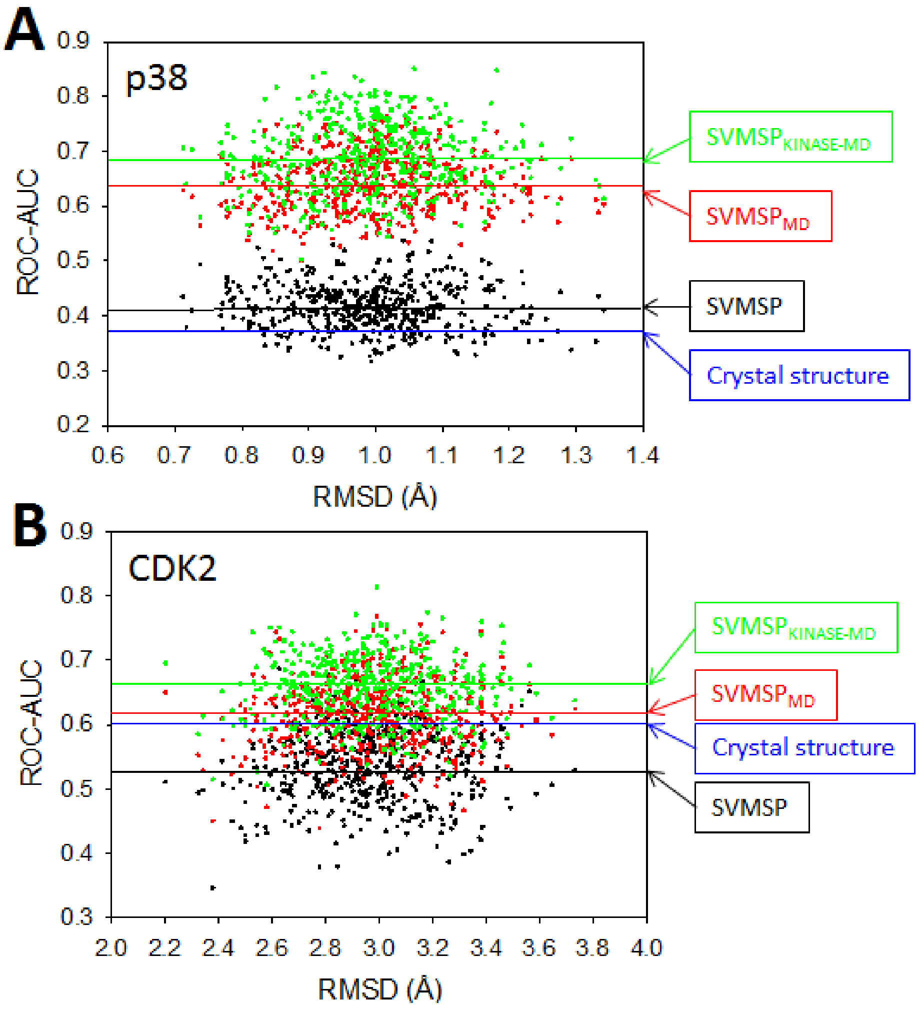


Figure 2.3 ROC-AUC for Individual Snapshots Using SVMSP$_{MD}$

2.3.4 *A Priori* Selection of MD Snapshots with High Enrichment Power

The aforementioned results show that a subset of MD snapshots possess greater enrichment power than the crystal structure. A question of interest is whether one can preselect these MD structures from an MD simulation of the apo structure in the absence of actives. Our set of protein-compound structures enable us to address this question since we know the enrichment power of each snapshot. The ROC-AUC can be used as a score to rank-order the snapshots.

A plot of ROC-AUC for each snapshot of the proteins in Table 2.1 reveals that MD snapshots with higher enrichment power have a tendency to have lower average SVMSP scores for decoys (Appendix B, Figure B.2). Hence, to identify MD snapshots with high enrichment power, one could dock randomly selected compounds to the snapshots and rank the snapshots with SVMSP. The snapshots with the lowest SVMSP scores are likely to have the highest enrichment power (least likely to bind to the random compounds). To test this we docked a set of randomly selected compounds to each of the 500 snapshots of EGFR and Src. These compounds were scored with SVMSP, and a median decoy score is determined for SVMSP. In each case, snapshots were ranked with the median SVMSP score. To determine how effectively we are filtering these MD snapshots for structures with high enrichment power, we defined ROC-AUC thresholds of 50, 60, 70, 80 and 90% of the ROC-AUC range (ROC-AUC$_{MAX}$ - ROC-AUC$_{MIN}$) score (Figure 2.4). So a 50% threshold means that if an MD snapshot has an ROC-AUC that is greater than 50% of the value of the maximum ROC-AUC minus the minimum ROC-AUC of the crystal structure, it is considered a true positive (high enrichment structure). This threshold enabled us to construct ROC curves to test how effectively we are enriching for snapshots that exceed this threshold. In the case of EGFR, assuming a 50% threshold, the ability to *a priori* identify high enrichment structures using strictly decoy compounds is high as evidenced by an ROC-AUC of 0.90 (Figure 2.4). When a more stringent definition is used for high enrichment power (90% of the ROC-AUC of the crystal structure), the *a priori* identification of high enrichment power MD structures becomes more challenging as evidenced by a decrease in the ROC-AUC to 0.63. For Src, a similar performance is found with ROC-AUC of 0.71 for a 50% threshold but less significant enrichment is obtained (0.76) using a 90% threshold.

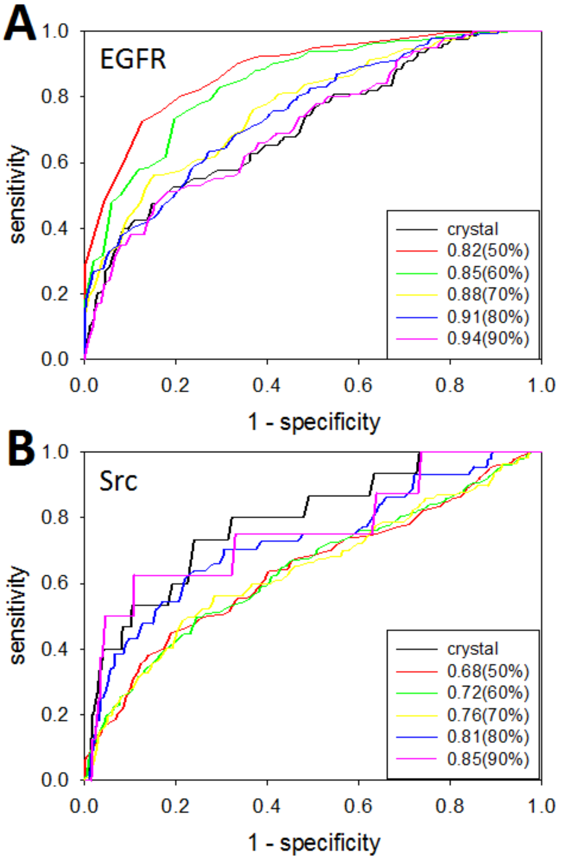


Figure 2.4 Filtering MD Snapshots for *A Priori* Identification of High Enrichment Structures

2.3.5 Rank-Ordering in Crystal and MD Structures

While ROC-AUC data gives a measure of enrichment, it did not provide insight into the rank-ordering of compounds among MD snapshots. Rank-ordering was compared for MD structures using Kendall's τ. The correlation metric is a measure of rank correlation, which provides insight into the similarity of the ordering of the data. The correlation coefficient ranges from -1 (inversely correlated) to 1 (correlated). We used τ to compare the rank-ordering of the X-ray and 50 MD snapshot to each other. The data is illustrated in a 2D color-coded map in Figure 2.5. The maps reveal that changes in the rank-ordering among structures can vary substantially from one protein to the other. In the case of AChE and trypsin, for example, there was little similarity in the ordering of the compounds from snapshot to snapshot as evidenced by the relatively low τ values (Figure 2.5). In fact, there was a higher tendency for the rank-ordering to be inversely correlated. Src, CDK2, and MDM2, on the other hand, showed less inverse correlation than AChE and trypsin. But the three proteins had more pronounced fluctuation in their rank-ordering. Two targets, p38 and EGFR, revealed an even higher τ values (greater than 0.5), suggesting less effect of conformational change on the binding of compounds. Finally, rank-ordering of AR was the least sensitive to changes in the structure of the protein as evidenced by τ values exceeding 0.6 in the majority of structures. Figure 2.5I shows τ comparing the rank-ordering in the crystal structure versus all the 50 snapshots. Interestingly, the correlation trends show similarity with correlation among MD snapshots. Interestingly, AR was the only case that showed a strong correlation between the ordering of compounds in the MD and X-ray structures. MDM2 snapshots showed the highest similarity in the ordering of compounds with one snapshots exhibit very similar correlation with the X-ray structure of the protein. AChE snapshots were the least similar to the X-ray structure of the protein.

Overall, it was interesting that the ordering of compounds among MD structures did not correlate with enrichment performance. For example, the ROC-AUC in p38 was relatively poor compared to EGFR (0.36 and 0.90, respectively), but they both showed similar 2D maps in Figure 2.5. Conversely, ROC-AUC values were relatively similar in AChE and MDM2 (0.72 and 0.75, respectively), but their similarity maps were dramatically different.

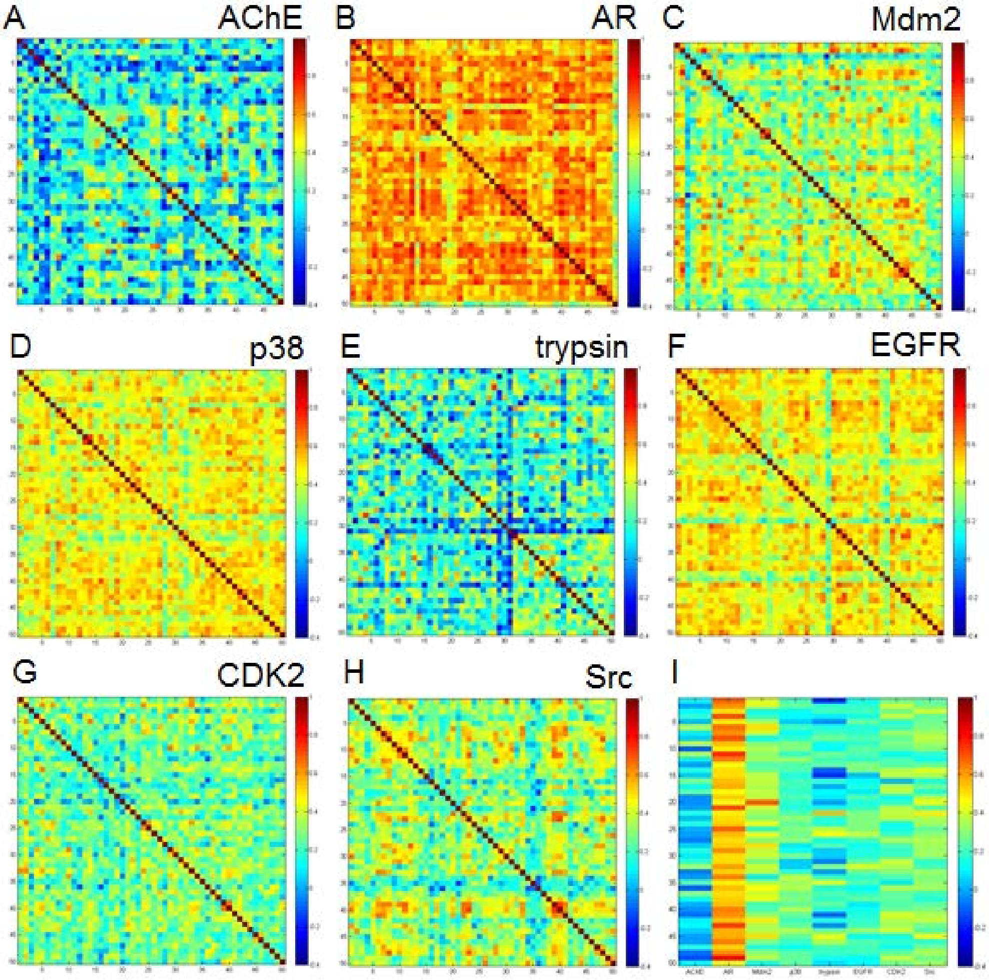


Figure 2.5 Correlation in Rank-Ordering of Compounds between Different Structures

2.3.6 Virtual Screening Chemical Library against MD Structures Leads to ALDH2 Small-Molecule Inhibitors

We applied SVMSP restoring of MD structures to the aldehyde dehydrogenase 2 (ALDH2) enzyme using SVMSP as the scoring approach. The crystal structure of ALDH2 in its apo form (PDB code: 1O04) was used to carry out explicit-solvent unbiased MD simulations.[101-104] Five independent simulations with 7 ns in length (35 ns total) yielded 25,000 snapshots. These were clustered by RMSD *ptraj*[105] as described above. A set of 50 representative snapshots were selected from the clusters. A focused set of the ChemDiv commercial library[24] containing 50,000 compounds were docked to each of the 50 snapshots by AutoDock Vina.[94] Docked receptor-ligand complexes were rescored with SVMSP. For each of the 50,000 compounds, the 50 MD snapshots to which they were docked were ranked and the top score was selected. The scores were used to rank the 50,000 compounds. The top 1,000 compounds were clustered into 150 sets that led to the selection of a representative compound from each set. Among the 150 compounds, 111 were commercially available and purchased for screening. A dehydrogenase assay that we have previously developed[106] was used to screen all 111 compounds at an initial concentration of 50 µM (Figure 2.6A). Compounds that inhibited ALDH2 dehydrogenase activity by more than 50% were selected for a follow-up concentration dependent study. Among them, five compounds inhibited the enzyme in a concentration-dependent manner (Figure 2.6B). The IC50s were 2.32, ~23, 0.62, 1.58, and 3.51 for ALDH400, ALDH417, ALDH423, ALDH427, and ALDH440, respectively (Figure 2.6C).

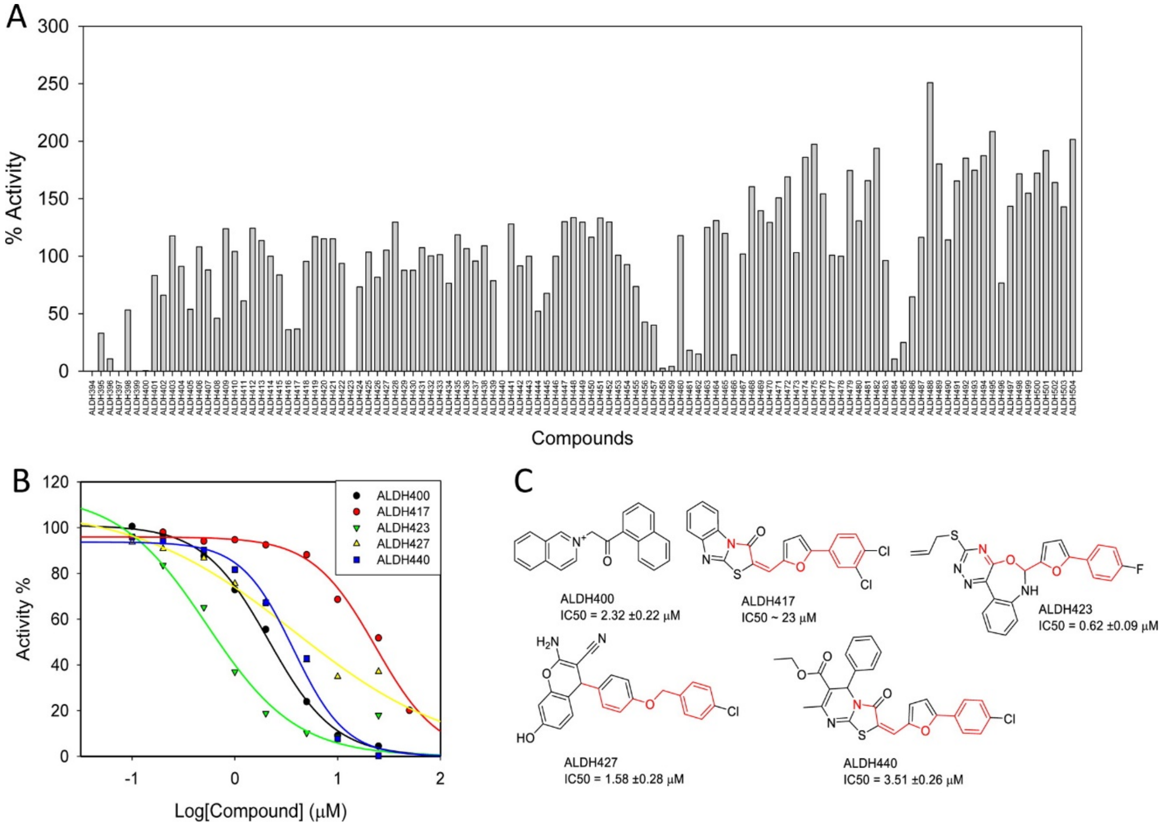


Figure 2.6 SVMSP Rescoring of MD Snapshots Identifies ALDH2 Inhibitors

Three compounds had some structural similarity as highlighted in red in Figure 2.6C. ALDH417, ALDH423, and ALDH440 contain a phenylfuran moiety. ALDH417 and ALDH440 exhibited even greater similarity that includes a similar thiazolidinone ring. The benzene ring of the phenylfuran is disubstituted in ALDH417 with two chlorine atoms at the *meta* and *para* positions, while ALDH423 and ALDH440 possess a fluorine and chlorine atom at the *para* position, respectively (Figure 2.6C). Inspection of ALDH427 reveals that the compound possess a benzyloxy group that mimics the phenylfuran of ALDH417. The position of the oxygen atom and benzene rings of the benzyloxy moiety mimic the oxygen atom of the furan and benzene ring of ALDH417, respectively. In fact, comparison of ALDH417, ALDH427, and ALDH440 reveals that there are five bonds between the oxygen atom and the pyrazole, pyrimidine, and pyran rings of ALDH417, ALDH440 and ALDH427 suggesting that the rings occupy a similar position within the binding pocket of ALDH2.

2.4 Discussion

We conducted a study to explore how enrichment in virtual screening of chemical libraries is affected by scoring MD structures of protein-compound complexes using a combined machine learning and statistical potential approach that we recently developed (SVMSP). We found that using an ensemble of MD structures showed similar enrichments to the crystal structure even as the size of the ensemble grew to 250 structures. It is worth mentioning that performance for the crystal structure was already good for most structures with four of the eight structures exhibiting ROC-AUC greater than 0.8. Interestingly, analysis of individual MD structures showed that there is a large number of snapshots that led to enrichment that significantly exceeded that of the crystal structure. Further probing revealed that enrichment was not correlated with structural deviation of the MD snapshots from the crystal structure. In addition, different MD snapshots resulted in different rank-ordering of compounds, suggesting that MD snapshots may also enhance diversity of the compounds identified in virtual screening.

These results prompted us to wonder whether using MD structures in the training of SVMSP may further improve enrichment. To test this, we picked two particularly challenging systems, namely, p38 and CDK2, for which enrichment was not better than random in the ensemble. In fact, none of the individual snapshots in p38 exhibited ROC-

AUC values better than random, and for CDK2, the majority of the snapshots had ROC-AUC lower than 0.6. When SVMSP was trained using compounds docked to MD snapshots ($SVMSP_{MD}$) for the negative set, we found a substantial increase in the enrichment performance, particularly for p38. The ROC-AUC in the ensemble increased from 0.42 to 0.64 for $SVMSP_{MD}$ and $ROC\text{-}AUC_{MAX}$ increased to 0.81 from 0.54 when negative set compounds docked to the crystal structure were used. Even greater enhancement was obtained when the positive set was strictly limited to kinases ($SVMSP_{KINASE-MD}$), with $ROC\text{-}AUC_{MAX}$ reaching 0.85. In CDK2, similar, but less pronounced, increases were observed.

We applied SVMSP scoring to MD snapshots to the mitochondrial aldehyde dehydrogenase 2 (ALDH2), which catalyzes the NAD^+-dependent oxidation of a broad spectrum of endogenous and biogenic aldehydes to their corresponding carboxylic acids. In humans, aldehyde dehydrogenases comprise a diverse gene family with approximately 20 members in the human genome sequence. ALDH2 may be an important drug target that has been implicated in drug addiction and other neurological disorders. We applied SVMSP to rank-order compounds docked to MD structures of ALDH2. The purpose of this exercise was not only to put SVMSP scoring of MD snapshots to the test but also for the discovery of small molecule ALDH2 inhibitors that can be pursued in future drug discovery efforts for this important class of enzyme family. The screening of 50 000 commercially available compounds against 50 MD snapshots of ALDH2 led to five compounds that inhibited the enzyme's dehydrogenase activity in a concentration-dependent manner. One compound (ALDH423) had submicromolar activity, while another three (ALDH400, ALDH427, and ALDH440) inhibited with IC_{50}s lower than 5 μM.

Interestingly, three compounds showed structural similarity. These compounds offer an opportunity to develop small-molecule inhibitors of the ALDH2 with higher affinity and selectivity across members of the ALDH family. The discovery of inhibitors does not validate SVMSP scoring of MD structures, but, combined with the extensive studies using validation sets that we have conducted, this work demonstrates that this approach can result in effective library enrichment.

In summary, we applied our SVMSP scoring approach to rank-order small molecules docked to conformational ensembles of proteins collected from explicit-solvent MD simulations. We found that a larger number of MD structures does not affect enrichment. But MD structures lead to greater diversity in the conformation of small molecules identified in virtual screening. Overall, the performance of SVMSP was better than other scoring functions for X-ray and MD structures. It is worth mentioning that we did not assess whether the docking methods generated accurate poses. This would be difficult to test particularly for the MD snapshots. However, SVMSP is trained using high quality protein−ligand crystal structures as positive set, and we expect that the scoring approach will favor native-like structures. In our previous work, we have shown that filtering protein−ligand MD snapshots with SVMSP resulted in significantly better rank-ordering of these complexes based on the binding affinity. Interestingly, MD simulations generated individual MD snapshots that showed significantly better enrichment than the X-ray structure. Two proteins were particularly challenging, and both X-ray and MD structures exhibited random enrichment. To overcome this challenge, we used MD snapshots to train SVMSP models and discovered a remarkable increase in performance in enrichment. We also embarked on an effort to identify high-performance MD structures *a* *priori* from an MD simulation of the apo protein. We found that it was possible to enrich apo protein MD structures by scoring randomly selected compounds docked to these structures using SVMSP. Finally, we put SVMSP rescoring to the test by rescoring a commercially available chemical library docked to the ALDH2 enzyme. Enzymology studies for the top candidates that emerged from a set of 50,000 compounds led to four compounds that had IC_{50}s below 5 μM. These compounds serve as leads for the design and synthesis of more potent and selective ALDH2 inhibitors.

REFERENCES

REFERENCES

1. Li, L.; Li, J.; Khanna, M.; Jo, I.; Baird, J. P.; Meroueh, S. O., Docking Small Molecules to Predicted Off-Targets of the Cancer Drug Erlotinib Leads to Inhibitors of Lung Cancer Cell Proliferation with Suitable In vitro Pharmacokinetic Properties. *ACS. Med. Chem. Lett.* **2010**, 1, 229-233.

2. de Ruiter, A.; Oostenbrink, C., Efficient and Accurate Free Energy Calculations on Trypsin Inhibitors. *J. Chem. Theory Comput.* **2012**.

3. Wan, S.; Coveney, P. V.; Flower, D. R., Peptide recognition by the T cell receptor: comparison of binding free energies from thermodynamic integration, Poisson–Boltzmann and linear interaction energy approximations. *Philos. Transact. A Math. Phys. Eng. Sci.* **2005**, 363, 2037-2053.

4. Golemi-Kotra, D.; Meroueh, S. O.; Kim, C.; Vakulenko, S. B.; Bulychev, A.; Stemmler, A. J.; Stemmler, T. L.; Mobashery, S., The Importance of a Critical Protonation State and the Fate of the Catalytic Steps in Class A β-Lactamases and Penicillin-binding Proteins. *J. Biol. Chem.* **2004**, 279, 34665-34673.

5. Oostenbrink, B. C.; Pitera, J. W.; van Lipzig, M. M. H.; Meerman, J. H. N.; van Gunsteren, W. F., Simulations of the Estrogen Receptor Ligand-Binding Domain: Affinity of Natural Ligands and Xenoestrogens. *J. Med. Chem.* **2000**, 43, 4594-4605.

6. Lawrenz, M.; Wereszczynski, J.; Ortiz-Sánchez, J.; Nichols, S.; McCammon, J. A., Thermodynamic integration to predict host-guest binding affinities. *J. Comput. Aided Mol. Des.* **2012**, 26, 569-576.

7. Steinbrecher, T.; Case, D. A.; Labahn, A., Free energy calculations on the binding of novel thiolactomycin derivatives to E. coli fatty acid synthase I. *Biorg. Med. Chem.* **2012**, 20, 3446-3453.

8. Lawrenz, M.; Wereszczynski, J.; Amaro, R.; Walker, R.; Roitberg, A.; McCammon, J. A., Impact of calcium on N1 influenza neuraminidase dynamics and binding free energy. *Proteins: Struct. Funct. Bioinform.* **2010**, 78, 2523-2532.

9. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**.

10. Chong, L. T.; Dempster, S. E.; Hendsch, Z. S.; Lee, L.-P.; Tidor, B., Computation of electrostatic complements to proteins: A case of charge stabilized binding. *Protein Sci.* **1998**, 7, 206-210.

11. Ponder, J. W.; Case, D. A., Force fields for protein simulations. *Adv. Protein Chem.* **2003**, 66, 27-85.

12. Luo, R.; David, L.; Gilson, M. K., Accelerated Poisson–Boltzmann calculations for static and dynamic systems. *J. Comput. Chem.* **2002**, 23, 1244-1253.

13. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, 112, 6127-6129.

14. Karplus, M.; Kushick, J. N., Method for estimating the configurational entropy of macromolecules. *Macromolecules* **1981**, 14, 325-332.

15. Wang, J.; Morin, P.; Wang, W.; Kollman, P. A., Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *J. Am. Chem. Soc.* **2001**, 123, 5221-5230.

16. Gohlke, H.; Hendlich, M.; Klebe, G., Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, 295, 337-356.

17. Basdevant, N.; Weinstein, H.; Ceruso, M., Thermodynamic Basis for Promiscuity and Selectivity in Protein−Protein Interactions: PDZ Domains, a Case Study. *J. Am. Chem. Soc.* **2006**, 128, 12766-12777.

18. Grünberg, R.; Nilges, M.; Leckner, J., Flexibility and Conformational Entropy in Protein-Protein Binding. *Structure* **2006**, 14, 683-693.

19. Lill, M. A.; Thompson, J. J., Solvent Interaction Energy Calculations on Molecular Dynamics Trajectories: Increasing the Efficiency Using Systematic Frame Selection. *J. Chem. Inf. Model.* **2011**, 51, 2680-2689.

20. Li, L.; Dantzer, J. J.; Nowacki, J.; O'Callaghan, B. J.; Meroueh, S. O., PDBcal: A Comprehensive Dataset for Receptor–Ligand Interactions with Three-dimensional Structures and Binding Thermodynamics from Isothermal Titration Calorimetry. *Chem. Biol. Drug Des.* **2008**, 71, 529-532.

21. Li, L.; Khanna, M.; Jo, I.; Wang, F.; Ashpole, N. M.; Hudmon, A.; Meroueh, S. O., Target-Specific Support Vector Machine Scoring in Structure-Based Virtual Screening: Computational Validation, In Vitro Testing in Kinases, and Effects on Lung Cancer Cell Proliferation. *J. Chem. Inf. Model.* **2011**, 51, 755-759.

22. Li, L.; Wang, B.; Meroueh, S. O., Support Vector Machine Scoring of Receptor-Ligand Complexes for Virtual Screening of Chemical Libraries. *J. Chem. Inf. Model.* **2011**, 51, 2132-2138.

23. Kellenberger, E.; Muller, P.; Schalon, C.; Bret, G.; Foata, N.; Rognan, D., sc-PDB: an Annotated Database of Druggable Binding Sites from the Protein Data Bank. *J. Chem. Inf. Model.* **2006**, 46, 717-727.

24. Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G., ZINC: a free tool to discover chemistry for biology. *Journal of chemical information and modeling* **2012**, 52, 1757-68.

25. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Res.* **2000**, 28, 235-242.

26. Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C., Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **1999**, 285, 1735-1747.

27. Jakalian, A.; Jack, D. B.; Bayly, C. I., Fast, efficient generation of high-quality atomic charges. AM1-BCC model: II. Parameterization and validation. *J. Comput. Chem.* **2002**, 23, 1623-1641.

28. Case, D. A.; Darden, T. A.; T.E. Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Pearlman, D. A.; Crowley, M.; Walker, R. C.; Zhang, W.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Wu, X.; Brozell, S. R.; Tsui, V.; Schafmeister, H.; Ross, W. S.; Kollman, P. A. *AMBER9*, University of California, San Fransico, 2006.

29. Duan, Y.; Wu, C.; Chowdhury, S.; Lee, M. C.; Xiong, G.; Zhang, W.; Yang, R.; Cieplak, P.; Luo, R.; Lee, T.; Caldwell, J.; Wang, J.; Kollman, P., A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **2003**, 24, 1999-2012.

30. Ryckaert, J.-P.; Ciccotti, G.; Berendsen, H. J. C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* **1977**, 23, 327-341.

31. Loncharich, R. J.; Brooks, B. R.; Pastor, R. W., Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **1992**, 32, 523-35.

32. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **1984**, 81, 3684-3690.

33. Jones, J. E., On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *P. Roy. Soc. Lond. A Mat.* **1924**, 106, 463-477.

34. Honig, B.; Nicholls, A., Classical electrostatics in biology and chemistry. *Science* **1995**, 268, 1144-1149.

35. Sitkoff, D.; Sharp, K. A.; Honig, B., Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *J. Phys. Chem.* **1994**, 98, 1978-1988.

36. Connolly, M., Analytical molecular surface calculation. *J. Appl. Crystallogr.* **1983**, 16, 548-558.

37. Tan, C.; Yang, L.; Luo, R., How Well Does Poisson−Boltzmann Implicit Solvent Agree with Explicit Solvent? A Quantitative Analysis. *J. Phys. Chem. B* **2006**, 110, 18680-18687.

38. Onufriev, A.; Bashford, D.; Case, D. A., Exploring protein native states and large-scale conformational changes with a modified Generalized Born model. *Proteins: Struct. Funct. Bioinform.* **2004**, 55, 383-394.

39. Onufriev, A.; Bashford, D.; Case, D. A., Modification of the Generalized Born model suitable for macromolecules. *J. Phys. Chem. B* **2000**, 104, 3712-3720.

40. Liang, S.; Li, L.; Hsu, W.-L.; Pilcher, M. N.; Uversky, V.; Zhou, Y.; Dunker, A. K.; Meroueh, S. O., Exploring the Molecular Design of Protein Interaction Sites with Molecular Dynamics Simulations and Free Energy Calculations. *Biochemistry* **2008**, 48, 399-414.

41. Jain, A.; Nicholls, A., Recommendations for evaluation of computational methods. *J. Comput. Aided Mol. Des.* **2008**, 22, 133-139.

42. Liang, S.; Li, L.; Hsu, W.-L.; Pilcher, M. N.; Uversky, V.; Zhou, Y.; Dunker, K. A.; Meroueh, S. O., Exploring the molecular design of protein interaction sites with molecular dynamics simulations and free energy calculations. *Biochemistry* **2008**, 48, 399-414.

43. Li, L.; Liang, S.; Pilcher, M. M.; Meroueh, S. O., Incorporating receptor flexibility in the molecular design of protein interfaces. *Protein Eng. Des. Sel.* **2009**, 22, 575-586.

44. Li, L.; Wang, B.; Meroueh, S. O., Support vector regression scoring of receptor–ligand complexes for rank-ordering and virtual screening of chemical libraries. *J. Chem. Inf. Model.* **2011**, 51, 2132-2138.

45. Lee, K. H.; Xie, D.; Freire, E.; Amzel, M. L., Estimation of changes in side chain configurational entropy in binding and folding: General methods and application to helix formation. *Proteins: Struct. Funct. Bioinform.* **1994**, 20, 68-84.

46. Shoichet, B. K.; Kobilka, B. K., Structure-based drug screening for G-protein-coupled receptors. *Trends Pharmacol. Sci.* **2012**, 33, 268-272.

47. Tomašić, T.; Kovač, A.; Klebe, G.; Blanot, D.; Gobec, S.; Kikelj, D.; Mašič, L., Virtual screening for potential inhibitors of bacterial MurC and MurD ligases. *J. Mol. Model.* **2012**, 18, 1063-1072.

48. Zheng, Z.; Merz, K. M., Ligand Identification Scoring Algorithm (LISA). *J. Chem. Inf. Model.* **2011**, 51, 1296-1306.

49. Böhm, H.-J., The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput. Aided Mol. Des.* **1994**, 8, 243-256.

50. Korb, O.; Stützle, T.; Exner, T. E., Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS. *J. Chem. Inf. Model.* **2009**, 49, 84-96.

51. Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S., Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, 47, 1739-1749.

52. Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G., A Fast Flexible Docking Method using an Incremental Construction Algorithm. *J. Mol. Biol.* **1996**, 261, 470-489.

53. Zilian, D.; Sotriffer, C. A., SFCscoreRF: A Random Forest-Based Scoring Function for Improved Affinity Prediction of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2013**, 53, 1923-1933.

54. Li, G.-B.; Yang, L.-L.; Wang, W.-J.; Li, L.-L.; Yang, S.-Y., ID-Score: A New Empirical Scoring Function Based on a Comprehensive Set of Descriptors Related to Protein–Ligand Interactions. *J. Chem. Inf. Model.* **2013**, 53, 592-600.

55. Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R., Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, 267, 727-748.

56. Wang, R.; Liu, L.; Lai, L.; Tang, Y., SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-Ligand Complex. *J. Mol. Model.* **1998**, 4, 379-394.

57. Muegge, I.; Martin, Y. C., A General and Fast Scoring Function for Protein−Ligand Interactions: A Simplified Potential Approach. *J. Med. Chem.* **1999**, 42, 791-804.

58. Muegge, I., A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspect. Drug Discov. Des.* **2000**, 20, 99-114.

59. Muegge, I., Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, 22, 418-425.

60. Ishchenko, A. V.; Shakhnovich, E. I., SMall Molecule Growth 2001 (SMoG2001): An Improved Knowledge-Based Scoring Function for Protein−Ligand Interactions. *J. Med. Chem.* **2002**, 45, 2770-2780.

61. Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M., BLEEP—potential of mean force describing protein–ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, 20, 1165-1176.

62. Velec, H. F. G.; Gohlke, H.; Klebe, G., DrugScoreCSDKnowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction. *J. Med. Chem.* **2005**, 48, 6296-6303.

63. Huang, S.-Y.; Zou, X., An iterative knowledge-based scoring function to predict protein–ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* **2006**, 27, 1876-1882.

64. Huang, S.-Y.; Zou, X., Inclusion of Solvation and Entropy in the Knowledge-Based Scoring Function for Protein−Ligand Interactions. *J. Chem. Inf. Model.* **2010**, 50, 262-273.

65. Neudert, G.; Klebe, G., DSX: A Knowledge-Based Scoring Function for the Assessment of Protein–Ligand Complexes. *J. Chem. Inf. Model.* **2011**, 51, 2731-2745.

66. Ewing, T. A.; Makino, S.; Skillman, A. G.; Kuntz, I., DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput. Aided Mol. Des.* **2001**, 15, 411-428.

67. Yin, S.; Biedermannova, L.; Vondrasek, J.; Dokholyan, N. V., MedusaScore: An Accurate Force Field-Based Scoring Function for Virtual Drug Screening. *J. Chem. Inf. Model.* **2008**, 48, 1656-1662.

68. Lee, M. C.; Duan, Y., Distinguish protein decoys by Using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. *Proteins: Struct. Funct. Bioinform.* **2004**, 55, 620-634.

69. Pencheva, T.; Lagorce, D.; Pajeva, I.; Villoutreix, B.; Miteva, M., AMMOS: Automated Molecular Mechanics Optimization tool for in silico Screening. *BMC Bioinformatics* **2008**, 9, 438.

70. Raha, K.; Merz, K. M., Large-Scale Validation of a Quantum Mechanics Based Scoring Function: Predicting the Binding Affinity and the Binding Mode of a Diverse Set of Protein−Ligand Complexes. *J. Med. Chem.* **2005**, 48, 4558-4575.

71. McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K., Gaussian docking functions. *Biopolymers* **2003**, 68, 76-90.

72. Huang, N.; Kalyanaraman, C.; Irwin, J. J.; Jacobson, M. P., Physics-Based Scoring of Protein−Ligand Complexes: Enrichment of Known Inhibitors in Large-Scale Virtual Screening. *J. Chem. Inf. Model.* **2005**, 46, 243-253.

73. Ferrara, P.; Curioni, A.; Vangrevelinghe, E.; Meyer, T.; Mordasini, T.; Andreoni, W.; Acklin, P.; Jacoby, E., New Scoring Functions for Virtual Screening from Molecular Dynamics Simulations with a Quantum-Refined Force-Field (QRFF-MD). Application to Cyclin-Dependent Kinase 2. *J. Chem. Inf. Model.* **2005**, 46, 254-263.

74. Li, L.; Wang, B.; Meroueh, S. O., Support Vector Machine Scoring of Receptor-Ligand Complexes for Virtual Screening of Chemical Libraries. *J. Chem. Inf. Model.* **2011**, 51, 2132-2138.

75. Feixas, F.; Lindert, S.; Sinko, W.; McCammon, J. A., Exploring the role of receptor flexibility in structure-based drug discovery. *Biophys. Chem.* **2014**, 186, 31-45.

76. Tarcsay, Á.; Paragi, G.; Vass, M.; Jójárt, B.; Bogár, F.; Keserű, G. M., The Impact of Molecular Dynamics Sampling on the Performance of Virtual Screening against GPCRs. *J. Chem. Inf. Model.* **2013**, 53, 2990-2999.

77. Nichols, S. E.; Baron, R.; Ivetac, A.; McCammon, J. A., Predictive Power of Molecular Dynamics Receptor Structures in Virtual Screening. *J. Chem. Inf. Model.* **2011**, 51, 1439-1446.

78. Wang, L.; Yang, C.; Lu, W.; Liu, L.; Gao, R.; Liao, S.; Zhao, Z.; Zhu, L.; Xu, Y.; Li, H.; Huang, J.; Zhu, W., Discovery of new potent inhibitors for carbonic anhydrase IX by structure-based virtual screening. *Bioorg. Med. Chem. Lett.* **2013**, 23, 3496-3499.

79. Lu, J.; Xin, S.; Meng, H.; Veldman, M.; Schoenfeld, D.; Che, C.; Yan, R.; Zhong, H.; Li, S.; Lin, S., A Novel Anti-Tumor Inhibitor Identified by Virtual Screen with PLK1 Structure and Zebrafish Assay. *PLoS One* **2013**, 8, e53317.

80. Amaning, K.; Lowinski, M.; Vallee, F.; Steier, V.; Marcireau, C.; Ugolini, A.; Delorme, C.; Foucalt, F.; McCort, G.; Derimay, N.; Andouche, C.; Vougier, S.; Llopart, S.; Halland, N.; Rak, A., The use of virtual screening and differential scanning fluorimetry for the rapid identification of fragments active against MEK1. *Bioorg. Med. Chem. Lett.* **2013**, 23, 3620-3626.

81. Dixit, A.; Verkhivker, G. M., Integrating Ligand-Based and Protein-Centric Virtual Screening of Kinase Inhibitors Using Ensembles of Multiple Protein Kinase Genes and Conformations. *J. Chem. Inf. Model.* **2012**, 52, 2501-2515.

82. Zhou, S.; Li, Y.; Hou, T., Feasibility of Using Molecular Docking-Based Virtual Screening for Searching Dual Target Kinase Inhibitors. *J. Chem. Inf. Model.* **2013**, 53, 982-996.

83. Stigliani, J.-L.; Bernardes-Genisson, V.; Bernadou, J.; Pratviel, G., Cross-docking study on InhA inhibitors: a combination of Autodock Vina and PM6-DH2 simulations to retrieve bio-active conformations. *Org. Biomol. Chem.* **2012**, 10, 6341-6349.

84. Knegtel, R. M. A.; Kuntz, I. D.; Oshiro, C. M., Molecular docking to ensembles of protein structures. *J. Mol. Biol.* **1997**, 266, 424-440.

85. Kim, D.; Lee, Y. H.; Hwang, H. Y.; Kim, K. K.; Park, H. J., Z-DNA binding proteins as targets for structure-based virtual screening. *Curr. Drug Targets* **2010**, 11, 335-44.

86. Huang, S.-Y.; Zou, X., Efficient molecular docking of NMR structures: Application to HIV-1 protease. *Protein Sci.* **2007**, 16, 43-51.

87. Isvoran, A.; Badel, A.; Craescu, C.; Miron, S.; Miteva, M., Exploring NMR ensembles of calcium binding proteins: Perspectives to design inhibitors of protein-protein interactions. *BMC Struct. Biol.* **2011**, 11, 24.

88. Zhao, H.; Huang, D.; Caflisch, A., Discovery of Tyrosine Kinase Inhibitors by Docking into an Inactive Kinase Conformation Generated by Molecular Dynamics. *ChemMedChem* **2012**, 7, 1983-1990.

89. Ekonomiuk, D.; Su, X.-C.; Ozawa, K.; Bodenreider, C.; Lim, S. P.; Otting, G.; Huang, D.; Caflisch, A., Flaviviral Protease Inhibitors Identified by Fragment-Based Library Docking into a Structure Generated by Molecular Dynamics. *J. Med. Chem.* **2009**, 52, 4860-4868.

90. Khanna, M.; Wang, F.; Jo, I.; Knabe, W. E.; Wilson, S. M.; Li, L.; Bum-Erdene, K.; Li, J.; W. Sledge, G.; Khanna, R.; Meroueh, S. O., Targeting Multiple Conformations Leads to Small Molecule Inhibitors of the uPAR·uPA Protein–Protein Interaction That Block Cancer Cell Invasion. *ACS Chem. Biol.* **2011**, 6, 1232-1243.

91. Cala, O.; Remy, M.-H.; Guillet, V.; Merdes, A.; Mourey, L.; Milon, A.; Czaplicki, G., Virtual and Biophysical Screening Targeting the γ-Tubulin Complex – A New Target for the Inhibition of Microtubule Nucleation. *PLoS One* **2013**, 8, e63908.

92. Huang, N.; Shoichet, B. K.; Irwin, J. J., Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, 49, 6789-6801.

93. O'Boyle, N.; Banck, M.; James, C.; Morley, C.; Vandermeersch, T.; Hutchison, G., Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, 3, 33.

94. Trott, O.; Olson, A. J., AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, 31, 455-461.

95. Triballeau, N.; Acher, F.; Brabet, I.; Pin, J.-P.; Bertrand, H.-O., Virtual Screening Workflow Development Guided by the "Receiver Operating Characteristic" Curve Approach. Application to High-Throughput Docking on Metabotropic Glutamate Receptor Subtype 4. *J. Med. Chem.* **2005**, 48, 2534-2547.

96. Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W., Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **2013**, 27, 221-234.

97. Duan, J.; Dixon, S. L.; Lowrie, J. F.; Sherman, W., Analysis and comparison of 2D fingerprints: Insights into database screening performance using eight fingerprint methods. *J. Mol. Graph. Model.* **2010**, 29, 157-170.

98. Sastry, M.; Lowrie, J. F.; Dixon, S. L.; Sherman, W., Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *J. Chem. Inf. Model.* **2010**, 50, 771-784.

99. Eldridge, M.; Murray, C.; Auton, T.; Paolini, G.; Mee, R., Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **1997**, 11, 425-445.

100. Jones, G.; Willett, P.; Glen, R. C., Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, 245, 43-53.

101. Si, J.; Mueller, L.; Collins, S. J., CaMKII regulates retinoic acid receptor transcriptional activity and the differentiation of myeloid leukemia cells. *J. Clin. Invest.* **2007**, 117, 1412-21.

102. Ang, E. S.; Zhang, P.; Steer, J. H.; Tan, J. W.; Yip, K.; Zheng, M. H.; Joyce, D. A.; Xu, J., Calcium/calmodulin-dependent kinase activity is required for efficient induction of osteoclast differentiation and bone resorption by receptor activator of nuclear factor kappa B ligand (RANKL). *J. Cell. Physiol.* **2007**, 212, 787-95.

103. Marganski, W. A.; Gangopadhyay, S. S.; Je, H. D.; Gallant, C.; Morgan, K. G., Targeting of a novel Ca+2/calmodulin-dependent protein kinase II is essential for extracellular signal-regulated kinase-mediated signaling in differentiated smooth muscle cells. *Circ. Res.* **2005**, 97, 541-9.

104. Bouallegue, A.; Pandey, N. R.; Srivastava, A. K., CaMKII knockdown attenuates H2O2-induced phosphorylation of ERK1/2, PKB/Akt, and IGF-1R in vascular smooth muscle cells. *Free Radic. Biol. Med.* **2009**, 47, 858-66.

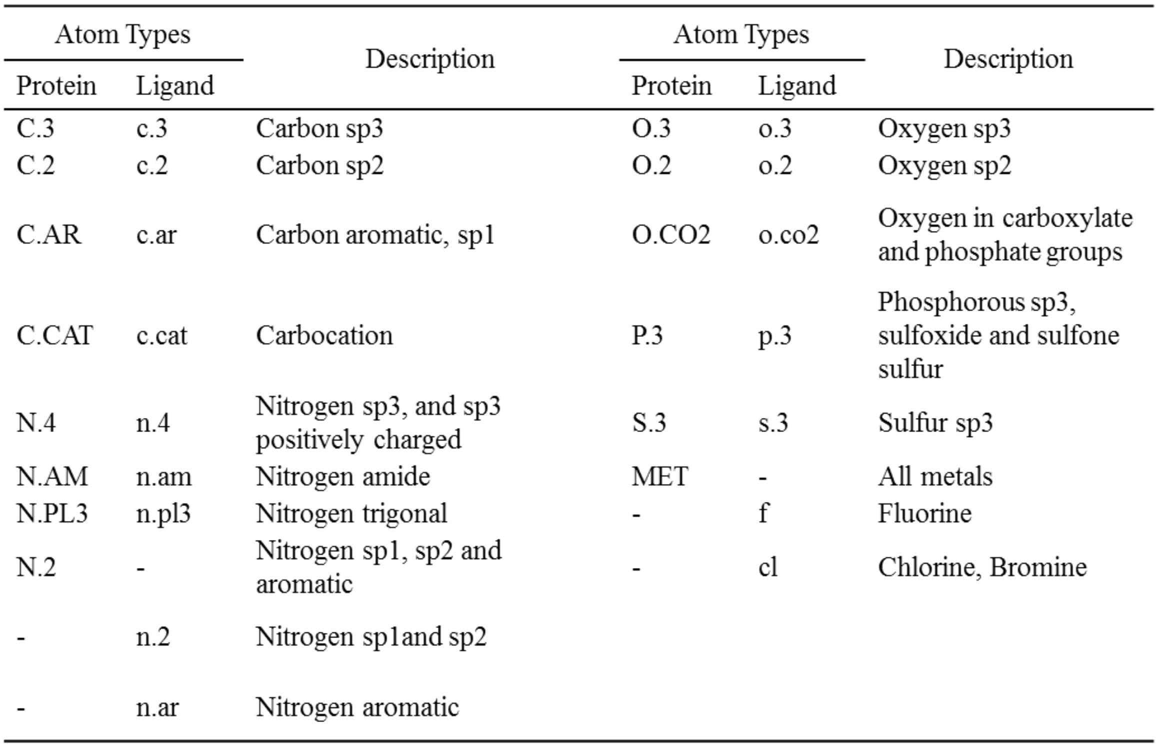
105. Case, D. A.; Darden, T. A.; T.E. Cheatham, I.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Swails, J.; Goetz, A. W.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Wolf, R. M.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Salomon-Ferrer, R.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A. *AMBER 12*, University of California, San Fransico, 2012.

106. Parajuli, B.; Kimble-Hill, A. C.; Khanna, M.; Ivanova, Y.; Meroueh, S.; Hurley, T. D., Discovery of novel regulators of aldehyde dehydrogenase isoenzymes. *Chem. Biol. Interact.* **2011**, 191, 153-8.

APPENDICES

Appendix A Supporting Information: Tables

Table A.1 Atom Types Used in Developing Knowledge-Based Pair Potentials



Atom Types		Description	Atom Types		Description
Protein	Ligand		Protein	Ligand	
C.3	c.3	Carbon sp3	O.3	o.3	Oxygen sp3
C.2	c.2	Carbon sp2	O.2	o.2	Oxygen sp2
C.AR	c.ar	Carbon aromatic, sp1	O.CO2	o.co2	Oxygen in carboxylate and phosphate groups
C.CAT	c.cat	Carbocation	P.3	p.3	Phosphorous sp3, sulfoxide and sulfone sulfur
N.4	n.4	Nitrogen sp3, and sp3 positively charged	S.3	s.3	Sulfur sp3
N.AM	n.am	Nitrogen amide	MET	-	All metals
N.PL3	n.pl3	Nitrogen trigonal	-	f	Fluorine
N.2	-	Nitrogen sp1, sp2 and aromatic	-	cl	Chlorine, Bromine
-	n.2	Nitrogen sp1and sp2			
-	n.ar	Nitrogen aromatic			

Table A.2 Parameters Used in MM-PBSA/MM-GBSA Calculations (AMBER9)

Parameters	**Values**
PB (free energies calculation using *pbsa*)	
Method used for solving the PB equation. (PROC)	2 (*pbsa* program)
Reference state taken for PB calculation. (REFE)	0
Dielectric constant for the solute. (INDI)	1.0 (2-10,15,20, agreed with DIELC)
Dielectric constant for the surrounding solvent. (EXDI)	80.0
Ionic strength (in mM) for the Poisson-Boltzmann solvent. (ISTRING)	0.0
Solvent probe radius in Angstrom. (PRBRAD)	1.6
Option to set up radii for PB calculation. (RADIOPT)	1
Lattice spacing in number of grids per Angstrom. (SCALE)	2
Number of iterations with linear PB equation. (LINIT)	1000
Values used to compute the nonpolar contribution G_{NP} to the desolvation according to G_{NP} = SURFTEN * SASA + SURFOFF (SURFEN/SURFOFF)	0.0072/0.00
GB (free energies calculation using GB model in *sander*)	
GB model. (IGB)	2 (Onufriev's GB)
Method used for SASA calculation. (GBSA)	2 (ICOSA)
Concentration (in M) of 1-1 mobile counterions in solution. (SALTCON)	0.00
Dielectricity constant for the surrounding solvent. (EXTDIEL)	80.0
Dielectricity constant for the solute. (INTDIEL)	1.0
Values used to compute the nonpolar contribution G_{NP} to the desolvation according to G_{NP} = SURFTEN * SASA + SURFOFF (SURFEN/SURFOFF)	0.0072/0.00
MM (gas phase energies calculation using *sander*)	
Dielectric constant for electrostatic interactions. (DIELC)	1 (2-10,15,20,)
MS (nonpolar contributions calculation using *molsurf*)	
Radius of the probe sphere used to calculate the SAS. (PROBE)	0.0
NM (entropies calculation with *nmode*)	
Distance-dependent dielectric constant. (DIELC)	4
Maximum number of cycles of minimization. (MAXCYC)	10000
Convergence criterion fro the energy gradient. (DRMS)	0.0001

Table A.3 Enrichment Performance of Different Cluster Size on Test Set Using ROC-AUC Score for Other Scoring

	ChemScore					GoldScore					GBSA				
	AChE	AR	Mdm2	p38	trypsin	AChE	AR	Mdm2	p38	trypsin	AChE	AR	Mdm2	p38	trypsin
Crystal	0.55	0.21	0.38	0.33	0.20	0.54	0.36	0.34	0.38	0.34	0.43	0.70	0.54	0.57	0.30
5	0.58	0.17	0.45	0.29	0.08	0.49	0.41	0.35	0.30	0.23	0.52	0.43	0.59	0.54	0.32
10	0.53	0.18	0.45	0.25	0.08	0.51	0.42	0.34	0.24	0.29	0.51	0.44	0.59	0.49	0.33
20	0.53	0.14	0.40	0.30	0.06	0.51	0.38	0.33	0.28	0.32	0.55	0.46	0.57	0.47	0.28
30	0.52	0.13	0.39	0.32	0.08	0.55	0.37	0.30	0.29	0.32	0.58	0.42	0.56	0.48	0.29
50	0.53	0.12	0.42	0.31	0.11	0.53	0.38	0.33	0.31	0.27	0.50	0.47	0.48	0.46	0.30
100	0.50	0.15	0.48	0.36	0.12	0.55	0.36	0.36	0.38	0.31	0.56	0.57	0.52	0.44	0.32
250	0.49	0.12	0.48	0.35	0.11	0.51	0.39	0.34	0.33	0.27	0.55	0.46	0.50	0.41	0.37

Appendix B Supporting Information: Figures

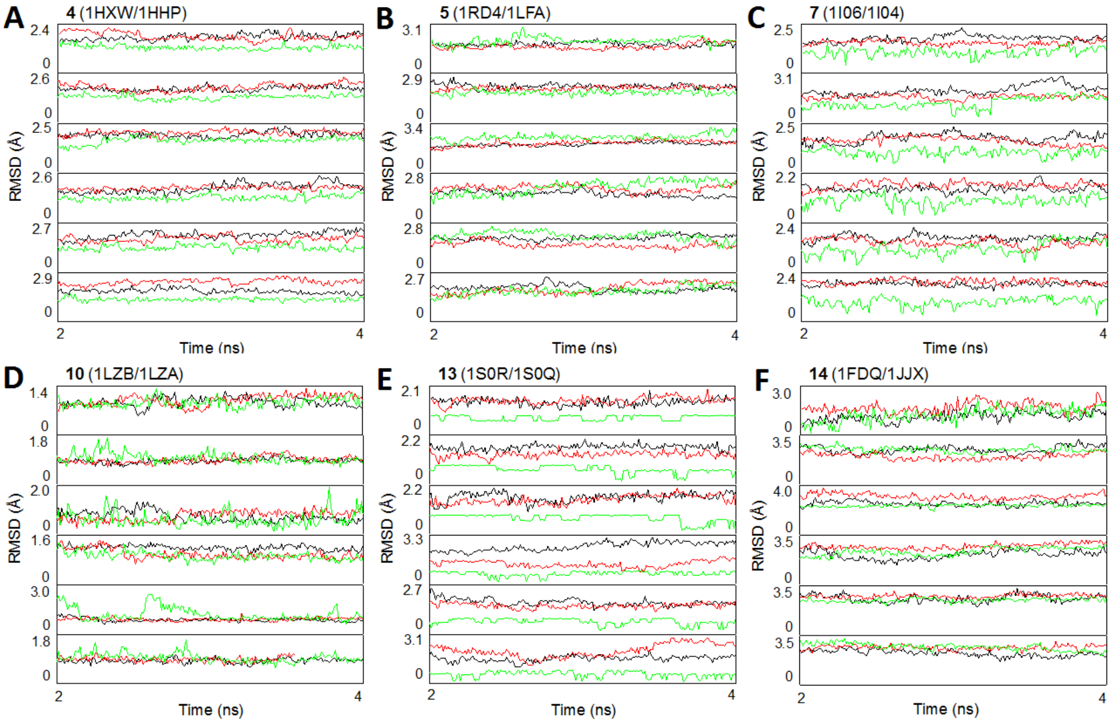


Figure B.1 Comparison of Dynamics of Free Ligand and Protein to the Protein-Ligand Complex

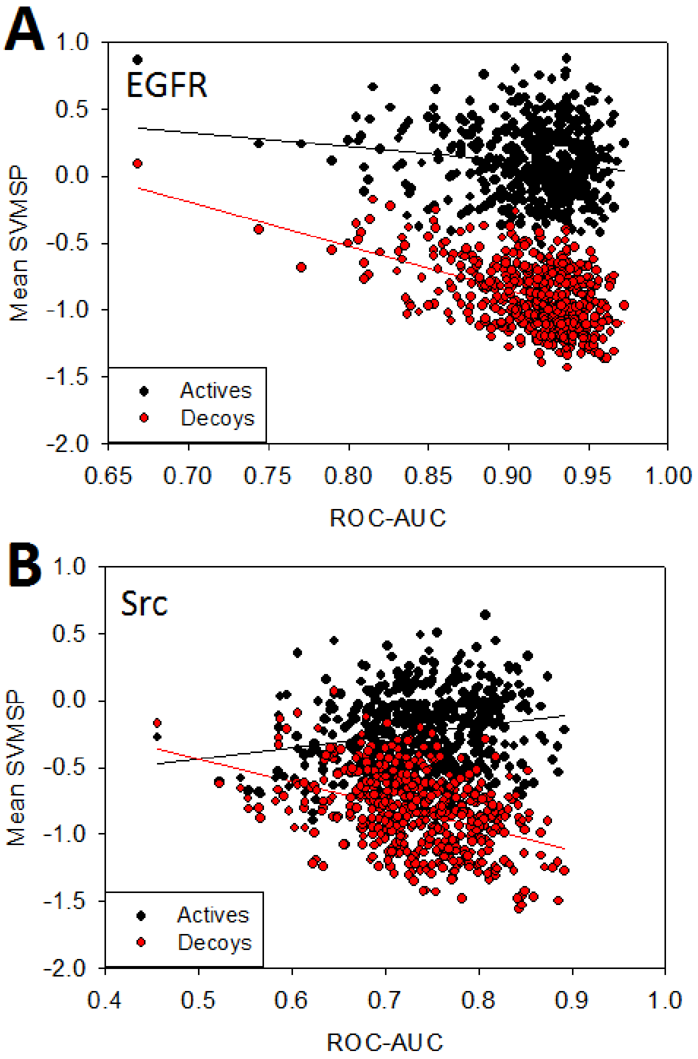


Figure B.2 Regression Plots between Mean SVMSP Score of DUD Compounds and ROC-AUC of Correspond Snapshots

Copyright Clearance Center RightsLink®

Home | Account Info | Help | Live Chat

ACS Publications
Most Trusted. Most Cited. Most Read.

Title:	Molecular Recognition in a Diverse Set of Protein–Ligand Interactions Studied with Molecular Dynamics Simulations and End-Point Free Energy Calculations
Author:	Bo Wang, Liwei Li, Thomas D. Hurley, et al
Publication:	Journal of Chemical Information and Modeling
Publisher:	American Chemical Society
Date:	Oct 1, 2013

Logged in as:
Bo Wang
Account #:
3000863387

LOGOUT

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK | CLOSE WINDOW

Copyright Clearance Center RightsLink®

Home | Account Info | Help | Live Chat

ACS Publications
Most Trusted. Most Cited. Most Read.

Title:	Enrichment of Chemical Libraries Docked to Protein Conformational Ensembles and Application to Aldehyde Dehydrogenase 2
Author:	Bo Wang, Cameron D. Buchman, Liwei Li, et al
Publication:	Journal of Chemical Information and Modeling
Publisher:	American Chemical Society
Date:	Jul 1, 2014

Logged in as:
Bo Wang
Account #:
3000863387

LOGOUT

PERMISSION/LICENSE IS GRANTED FOR YOUR ORDER AT NO CHARGE

This type of permission/license, instead of the standard Terms & Conditions, is sent to you because no fee is being charged for your order. Please note the following:

- Permission is granted for your request in both print and electronic formats, and translations.
- If figures and/or tables were requested, they may be adapted or used in part.
- Please print this page for your records and send a copy of it to your publisher/graduate school.
- Appropriate credit for the requested material should be given as follows: "Reprinted (adapted) with permission from (COMPLETE REFERENCE CITATION). Copyright (YEAR) American Chemical Society." Insert appropriate information in place of the capitalized words.
- One-time permission is granted only for the use specified in your request. No additional uses are granted (such as derivative works or other editions). For any other uses, please submit a new request.

BACK | CLOSE WINDOW