

Stony Brook University



OFFICIAL COPY

The official electronic file of this thesis or dissertation is maintained by the University Libraries on behalf of The Graduate School at Stony Brook University.

© All Rights Reserved by Author.

Application and Development of Computational Tools in Drug Discovery

A Dissertation Presented

by

Trent Erik Balius

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Computational Biology)

Stony Brook University

May 2012

Copyright by
Trent Erik Balius
2012

Stony Brook University

The Graduate School

Trent Erik Balius

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Robert C. Rizzo – Dissertation Advisor
Associate Professor, Applied Mathematics and Statistics

David F. Green - Chairperson of Defense
Associate Professor, Applied Mathematics and Statistics

Carlos Simmerling,
Professor, Chemistry

W. Todd Miller,
Professor, Physiology and Biophysics

This dissertation is accepted by the Graduate School

Charles Taber

Interim Dean of the Graduate School

Abstract of the Dissertation

Application and Development of Computational Tools in Drug Discovery

by

Trent Erik Balius

Doctor of Philosophy

in

Applied Mathematics and Statistics

(Computational Biology)

Stony Brook University

2012

In this dissertation, I will discuss several interconnected projects motivated by drug development. These projects employ computational techniques to study molecular recognition of a ligand (drug) by a receptor (protein) characterized through use of structural and energetic analysis.

In Chapter 1, an introduction to computational drug discovery is presented, and methods used here are discussed. Epidermal Growth Factor Receptor (EGFR) is an important drug target for the treatment of cancer. In Chapter 2, we performed all atom molecular dynamics simulations of clinically relevant mutations of EGFR complexed with erlotinib (Tarceva) and other inhibitors. The per-residue decomposition of intermolecular van der Waals and electrostatic energies – termed here molecular footprints – are useful in characterizing mechanisms of drug resistance. For instance, the resistance to erlotinib and other inhibitors observed for the T790M mutation does not employ a steric clash mechanism as was discussed in the literature. In fact, our results show that favorable van der Waals interactions are increased at this position. Notably, water-mediated interactions were revealed to be highly important for explaining the resistance profiles.

Footprints are useful in understanding binding. We observed that a molecular footprint can be computed for any pose including conformers generated by docking. Thus, we developed a footprint-based rescoring function in DOCK 6.5, termed footprint similarity (FPS) score. The FPS scoring method is discussed in Chapter 3 and 4. This tool uses comparison methods (Euclidean distance and Pearson correlation) to quantify footprint similarities between a reference molecule and docked molecules. The FPS score enables users to rank-order virtual

screening results where the top scoring molecules have similar interactions as those of a reference. References may include known drugs, natural substrates, and low energy transition states, among other possibilities. This method was validated using pose reproduction, cross-docking, and enrichment studies. In addition, experimental collaborators have identify promising lead compounds from our virtual screening projects (using FPS re-ranking) including those targeting BotNT/A and HIV-gp41. The FPS rescoring method was generalized to use grids enabling footprint-guided docking. This grid-based FPS scoring method has been validated using pose reproduction experiments. Future directions and ongoing projects are also discussed including de novo design.

In Chapter 5, we conclude with a description of ongoing projects and ideas for future directions. Finally, we discuss two collaborative projects which are presented in the appendixes. The role of point mutations in resistance of the HIV fusion protein glycoprotein 41 (HIVgp41) to the binding of T20, an FDA approved therapeutic was characterized. Energetic error analysis and membrane contributions were also examined. For HIVgp41 application studies, molecular footprints were shown to be useful in understanding drug-target molecular recognition as well as drug resistance mutations. Motivated in part by participating in a docking symposium entitled “Docking and Scoring: A Review of Docking Programs” held at the American Chemical Society meeting (spring, 2011), DOCK 6.6 was evaluated as an enrichment tool using Receiver operating characteristic (ROC) curves and area under the curve (AUC) analysis on validation databases DUD and Wombat. Enrichment studies demonstrate above random global enrichment, as well as good early enrichments, revealing DOCK as a useful tool for virtual screening applications. All of the projects discussed here demonstrate the strength of computational techniques in drug discovery.

Table of Contents

List of Figures	vii
List of Tables	xiv
List of Abbreviations	xvi
Acknowledgments.....	xviii
Chapter 1. Introduction to Computational Structural Biology and Drug Discovery.	1
1.1 Importance of Computational Techniques for Drug Discovery.	1
1.2 Structures in Drug Discovery.....	5
1.3 Molecular Mechanics, Force Fields, and the Potential Energy Function.	6
1.4 Molecular Dynamics.....	9
1.5 DOCKing	13
1.6 Research Projects.....	16
Chapter 2. Quantitative Prediction of Fold Resistance for Inhibitors of EGFR.....	17
Abstract.....	17
2.1 Introduction.....	18
2.2 Methods.....	25
2.2.1 Binding Free Energies.....	25
2.2.2 Interaction Signatures: Molecular Footprints.....	27
2.2.3 Simulation Protocols.....	28
2.2.4 Analysis.....	29
2.3 Results and Discussion.....	30
2.3.1 Simulation Stability.....	30
2.3.2 Comparison with Crystallographic Structures.....	32
2.3.3 Correlation with Experimental Fold Resistance.....	34
2.3.4 Energetics of Binding: What Drives Association?.....	39
2.3.5 Origins of Resistance.....	43
2.3.6 Water Mediated Interactions.....	48
2.4 Conclusions.....	52
Chapter 3. Implementation and Evaluation of a Docking-Rescoring Method using Molecular Footprint Comparisons.....	56
Abstract.....	56
3.1 Introduction.....	57
3.2 Theoretical Methods.....	61
3.2.1 Footprint Comparisons.....	61
3.2.2 Pose Identification.....	64
3.2.3 Database Enrichment.....	66
3.3 Computational Details.....	67
3.3.2 Pose Identification Datasets.....	67

3.3.2 Database Enrichment Datasets	69
3.3.3 Footprint Reference Preparation.....	70
3.3.4 Footprint Rescoring Protocols.....	70
3.4 Results and Discussion.	71
3.4.1 Footprint Similarity (FPS) vs DOCK Cartesian Energy (DCE) Scores for Pose Identification.....	71
3.4.2 Functional Relationships between Methods used to Compute FPS Scores.	75
3.4.3 Approximate Mathematical Relationship.....	77
3.4.4 Predicted Positive and Predicted Negatives.	79
3.4.5 False Positive Examples.	82
3.4.6 False Negative Examples.	85
3.4.7 Database Enrichment.....	91
3.5 Conclusions.....	96
Chapter 4. Grid-based Molecular Footprint Comparison Method for Docking.	100
4.1. Introduction.....	100
4.2. Theoretical Methods	101
4.3 Computational Details.	103
4.4 Results and Discussion	105
4.4.1 Grid Generation.	105
4.4.2 Docking and Rescoring in Pose Reproduction.	109
4.4.2 FPS Behavior During Growth.	111
4.4.3 Conclusions.	114
Chapter 5. Dissertation Summary: Scientific Impact, Related Work, and Future Direction....	115
5.1 Structural and energetic analysis of EGFR simulations.	115
5.2 Development of docking methods and virtual screening protocols.....	118
5.3 Summary.....	120
Bibliography	121
Appendix A. Origins of Resistance to the HIVgp41 Viral Entry Inhibitor T20.....	134
A.1 Collaborative Project Summary.....	134
A.2 Binding Free Energy Convergence: Autocorrelation and Block Averaging.	135
A.3 Per-Residue Energetic Analysis.....	138
A.4 Membrane Interactions: The Importance of the WNWF Motif.....	139
Appendix B. Evaluation of DOCK 6 as a Database Enrichment Tool.	144
B.1 Collaborative Project Summary	144
B.2 Theoretical Methods.....	145
B.3 Results	151
B.4 Enrichment Study Conclusions.....	168

List of Figures

Chapter 1.

Figure 1-1. The drug discovery pathway shown in black. Techniques used at various stages of the pathway are shown in gray..... 2

Figure 1-2. Molecular recognition and the binding event. The ligand and protein are shown in their free state on the left and then they associate together (represented by the arrow) to form the complex shown on the right..... 3

Figure 1-3. Virtual screening schematic for DOCK. A crystallographic protein structure is prepared for docking by creation of a sphere set and grid. Virtual screening is performed by docking ca. 1,000,000 molecules. On the order of 100 compounds are chosen for purchase and experimental testing. Perform Experimental testing on selected molecules. Feed-back loop demonstrates the relationship between computation and experimentation..... 5

Figure 1-4. Molecular mechanics terms, divided into bonded and through-space interactions are presented. Each interaction type is labeled, illustrated and its equation is reported..... 7

Figure 1-5. Two-dimension depiction of molecular dynamics. The iterative schema for propagating motion is outlined where i denotes the atom and t , the time step. The force (using energy function), the acceleration, velocities, and then new positions are obtained, consecutively. Dark gray circles are the initial positions of the atoms, and light gray are the new positions..... 10

Figure 1-6. (a) Matrix interaction of all atom pairs, blue is the internal energy of the receptor, red is the internal energy of the ligand and purple is the through-space interactions. (b) Per-residue decomposition of the through-space interactions..... 13

Figure 1-7. (a) Structure of Erlotinib is shown with labeled rotatable bonds. (b) The molecule is broken up in to rigid segments. The definition of layers is defined: (c) for the first anchor (A1); or (d) for the second anchor (A2). (e) Cartoon of the growth tree is shown. 15

Chapter 2.

Figure 2-1. Ribbon diagram showing EGFR complexed with the ATP-competitive inhibitor erlotinib. Regions which change conformation (N-lobe helix and activation loop) upon receptor activation are shown in green. Locations of cancer causing mutations (deletion or point) which cause receptor activation are in red. The secondary T790M drug resistance mutation is shown in blue. Coordinates from pdb code 1M17. 20

Figure 2-2. Schematic representation of the thermodynamic cycle used to calculate free energies of binding (ΔG_b calcd) for comparison with experiment (ΔG_b exptl). The cycle highlights the relationship between ΔG_b exptl occurring in condensed phase with the free energy of interaction

in the gas-phase (ΔG_{gas}) modulated by three terms representing the free energy of hydration (ΔG_{hyd}) for the transfer from vacuum to water for each separate species (com=complex, rec=receptor, lig=ligand). 26

Figure 2-3. Fluctuations in computed free energies of binding (ΔG_b calcd) and root-mean-square-deviation (rmsd) for erlotinib with wild type EGFR (panels a and d), L858R (panels b and e) and L858R&T790M (panels c and f) vs time. Left panels show instantaneous energies (gray dots) and block-running averages over 100 frames (black line). Right panels show rmsds for receptor backbone (blue line), ligand (green line), and ligand quinazoline scaffold atoms (red line). 32

Figure 2-4. Representative snapshots from MD simulations of ligands with EGFR showing side-chain sampling of key residues (thin lines, N=10 each) vs crystallographic conformations (bold lines) for erlotinib (panel 4a), gefitinib (panel 4b), and AEE788 (panel 4c). Pdbcodes for 4a: red=1M17 (erlotinib with wildtype), blue=2ITY (gefitinib with wildtype), green=2J6M (AEE788 with wildtype). Pdbcodes for 4b: blue=2ITZ (gefitinib with L858R), green=2JIU (AEE with T790M). Pdbcodes for 4c: blue=2ITO (gefitinib with G719S), green=2ITP (AEE788 with G719S). 34

Figure 2-5. Predicted FR energies ($\Delta \Delta G_{\text{FR}}$ calcd) vs experimental FR energies ($\Delta \Delta G_{\text{FR}}$ exptl) for inhibitors with EGFR. Each point is the difference between results from two independent MD simulations (16 simulations total) from 5000 MD snapshots each. Data point for erlotinib with double mutant (>3.96) excluded from r^2 calculations given ambiguity in the experimental $\Delta \Delta G_{\text{FR}}$ measurement. 37

Figure 2-6. Predicted changes in Coulombic ($\Delta \Delta E_{\text{cou}}$ panel a) and van der Waals ($\Delta \Delta E_{\text{vdw}}$ panel b) energy components versus experimental fold resistance energy ($\Delta \Delta G_{\text{FR}}$ exptl) for inhibitors with EGFR. Data point for erlotinib with double mutant (>3.96) excluded from r^2 calculations given ambiguity in the experimental $\Delta \Delta G_{\text{FR}}$ measurement. 39

Figure 2-7. Correlation of the van der Waals energy (ΔE_{vdw}) component with ΔG_b exptl. Energies in kcal/mol. 41

Figure 2-8. Primary H-bonding (dashed lines) interactions for inhibitors with EGFR. 42

Figure 2-9. Comparison of per-residue H-bond footprints for erlotinib (red), gefitinib (blue), and AEE788 (green) with wildtype (panel a), L858R (panel b), L858R&T790M (panel c), and G719S (panel d) EGFR variants (N = 5000 MD snapshots). 42

Figure 2-10. Comparison of erlotinib binding poses (N=100) from EGFR simulations for L858R (left, red), L858R&T790M (middle, green), and overlaid (red vs green). Intermolecular H-bonds at position C797 shown in purple (N=5000). 44

Figure 2-11. Per-residue footprints for inhibitors with EGFR for cancer causing (L858R, panels a-b) and drug resistance (L858R&T790M, panels c-d) variants from a reduced set of amino acids in the contiguous range Q787-N808 (shaded region) or for which any ligand shows $\Delta E > 1$ kcal/mol..... 45

Figure 2-12. Water-mediate H-bonds for inhibitors with EGFR for L858R (left) and L858R&T790M (right). 12a visually shows population of waters at site 1 (S1 orange) and site 2 (S2 blue) over all 5000 simulation frames for representative erlotinib simulations. Site waters defined if water hydrogens are within 2.5 angstroms of each ligand at N* (S1) or residue Q791 at O (S2). 12b shows for all three inhibitors the average number (count/5000) of waters at S1 and S2. 12c shows for all three inhibitors the average pairwise Coulombic interaction energies between the two waters closest to each ligand at N* with residues T790 (or M790), Q791, T854, and the ligands. 49

Chapter 3.

Figure 3-1. Representative molecular footprints for (a) a single ligand, (b) a single ligand with two conformations, and (c) two different ligands derived from per-residue decomposition of the intermolecular van der Waals interactions as a function of primary sequence. For two footprints, similarity may be quantified using Pearson correlation coefficient (r), Euclidean distance (d), or related measures. For clarity only a portion of the footprints are shown. 59

Figure 3-2. Flow chart outlining footprint similarity calculation protocol..... 62

Figure 3-3. Schematic depiction of standard (thin) versus normalized (thick) footprint vectors (x , y). The maximum distance between normalized vectors on the unit circle is 2 while the distance between standard vectors can be infinite. 63

Figure 3-4. Partitioning of outcome space (positive or negative results, red region) as a function of prediction (predicted positive or predicted negative, green region) into four quadrants (blue region) representing (I) true positives, (II) false positives, (III) true negatives, and (IV) false negatives for (a) pose identification and (b) database enrichment definitions of success. Gray colored lines represent hypothetical data..... 65

Figure 3-5. Database preparation histograms. (a) Population of systems with a given number of clusterheads (max=50) derived from Cartesian space minimizations of grid-based results reported by Mukherjee et al.¹³ (b) Population of systems with a given rmsd using only the single lowest-rmsd pose found among the ensemble of poses retained. The portion to the left of the dashed line at 2 Å rmsd constitutes perfect sampling subsets for (RGD=775), fixed-anchor (FAD=748), and flexible (FLX=680) ligand sampling. (c) Population of ligand rmsds for reference poses after polar hydrogen optimizations using the energy grids (black line) and subsequent energy minimizations in Cartesian space (purple line) using a harmonic tether. 69

Figure 3-6. Functional relationships between footprint similarity (FPS) scores computed for van der Waals (VDW, top) and electrostatic (ES, bottom) interactions using (a, b) standard Pearson vs threshold Pearson, (c, d) standard Euclidean vs normalized Euclidean, and (e, f) standard Pearson vs normalized Euclidean. Population color ranges for green = [1, 50], blue = [51, 250], and red = [251, 500+] are derived from the total FLX ensemble of N = 26,830 footprints. 76

Figure 3-7. Two dimensional histograms of rmsd versus FPS_{VDW+ES} score for (a) the best scored poses (N=680) and (b) the entire ensemble derived from all poses (N=26,830). Population color ranges for green = [1, 5], blue = [6, 20], and red = [21, 30+].....	80
Figure 3-8. False positive examples type I. Excellent similarity scores ($FPS_{VDW+ES} < 0.3$) but classified as failures due to a close-to-medium geometric match ($rmsd > 2 \text{ \AA}$ and $< 5 \text{ \AA}$). The associated PDB code, rmsd in \AA , FPS score, and overlay of the predicted (green) versus crystallographic (red) pose are shown for each system.	83
Figure 3-9. False positive examples type II. Good similarity scores ($FPS_{VDW+ES} < 0.6$) but classified as failures due to a poor geometric match ($rmsd > 5 \text{ \AA}$). The associated PDB code, rmsd in \AA , FPS score, and overlay of the predicted (green) versus crystallographic (red) pose are shown for each system.	83
Figure 3-10. Pose and footprint comparisons for (a) 2QE4 and (b) 9AAT showing results for the reference pose in red, the docked pose in green, and per-residue differences as black bars.	87
Figure 3-11. Pose identification results for the carbonic anhydrase family using crossdocking ensembles from Mukherjee et al. ¹³ Blue, green, red, and white elements indicate successes, scoring failures, sampling failures, and incomplete growth, respectively. Three scoring methods were evaluated: (a) standard DCE_{VDW+ES} , (b) FPS_{VDW+ES} in which cognate ligands (diagonals) were used as the footprint-reference corresponding to each receptor, (c) FPS_{VDW+ES} in which footprint-references were derived by minimizing each ligand in each receptor and every matrix element used a unique reference. Note that in all cases the rmsd references employed the set of ligands minimized in each receptor.	89
Figure 3-12. Cognate protein-ligand footprints for the aligned carbonic anhydrase family. Residue X indicates a given residue is not conserved across all crystal structures from the PDB entries in terms of amino acid sequence or signifies a substitution or deletion.....	91
Figure 3-13. ROC enrichment curves for (a) neuraminidase, (b) trypsin, and (c) EGFR using different ranking methods.	92
Figure 3-14. Graphical representation of the 50 top and 50 bottom ranked poses obtained from docking the 475 active ligands from the DUD EGFR database and using (a) DCE_{VDW+ES} and (b) FPS_{VDW+ES} scoring functions. The reference (erlotinib) is shown in red surface with top ligands in green and bottom ligands in gray. On the bottom are corresponding histograms of molecular weight (MW) for the 100 top (best) and 100 bottom (worst) ranked molecules. Note that the large MW peak at ca. 340 for the 100 best scoring molecules using FPS_{VDW+ES} corresponds ca. to the MW of the erlotinib reference (393.44 g/mol).....	95
Figure 3-15. Number of molecules retained from a virtual screen of 906,914 molecules to EGFR using various FPS_{VDW+ES} score cutoff values. The graphic shows the 25 molecules identified (green) using a cutoff of 0.8 in comparison with the known drug erlotinib (red) which was used as the footprint reference.	96

Chapter 4.

Figure 4-1. Comparisons between grid-based (grey) and Cartesian (black) footprints. The van der Waals (a and b) and electrostatics (c and d), all-residues footprints (a and c), and the threshold-based residues plus remainder footprints (b and d) are shown for pdb code 2PK4.... 106

Figure 4-2. (a) comparison of Grid (DGE) score vs Cartesian (DCE) Score. (b) Comparison of Sum of the footprint grid-values vs standard grid score. (c) Comparison of the Grid-based footprint values vs (d) Histogram of the number of grids used for footprint-guided docking (# grids = # residues + 1). For panel A-C, blue dots are ES and red are VDW. 107

Figure 4-3. Shows six docking experiments using SB2010. Five random seeds were used to gauge variability. Success (blue), scoring failure (green) and sampling failure (red) add up to 100%. 111

Figure 4-4. FPS score for growth step for the 10 rotatable bond subset of SB2010 (N= 59) are shown. Left panels show FPS guided docking results. Right panels show rescoring results from a standard grid score docking experiment. 112

Figure 4-5. Fixed anchor docking branch converges to a low RMSD and low FPS score. Both structures before and after minimization are shown. 113

Chapter 5.

Figure 5- 1. Active vs inactive conformations. The grey ribbon indicates the backbone of EGFR. The colored regions represent the main differences between the two structures. An overlay of the N-lobe helix and the activation loop is also shown. 118

Appendix A.

Figure A-1. (a) Autocorrelation functions (ACF) of calculated binding energies (ΔG_{bind} calcd) for T20 with wild HIVgp41 and seven mutants. (b) Block average standard error of the mean (BASEM) in kcal/mol as a function of block size from 1 to 5000. Solid arrows indicate initial I37K results while dashed arrows indicate I37K results obtained using a different random seed. 137

Figure A-2. Intermolecular van der Waals interaction energy matrix (heat map) for all gp41 with T20 residues from the wildtype simulation. Footprint peak magnitudes represent summation ΔE_{vdw} (kcal/mol) along each row (T20) or column (gp41). Gray shaded regions indicate mapping of the gp41 mutation region(L33-L45) to a charged/polar patch (E136-E148) on T20 defined by the black box in the ca. center of the matrix. Heat map gradient colors indicate the magnitude of unfavorable (blue) or favorable (red) interactions with dark red being most favorable. 138

Figure A-3. Comparison of the per-residue van der Waals (a) and Coulombic (b) interaction energies between gp41 (dashed line, small squares) and lipid membrane (solid line, filled circles) for each T20 residue from wildtype simulations. Each datapoint represents the average value obtained from 10,000 MD snapshots saved every 1 ps. 140

Figure A-4. Interactions of the T20 WNWF motif with membrane. 3-4a shows radial distribution functions (rdfs) for the average distances (N=10,000 snapshots) between all lipid headgroup N atoms and (d1) the center of ring atoms at W159, (d2) the backbone O at N160, (d3) the center of ring atoms at W161, and (d4) the backbone O at F162 (d4). Summation of bins (panel a gray vertical lines) out to a distance $\leq 5 \text{ \AA}$ approximates the number of interactions occurring in the first T20-lipid solvation shell represented by the first peaks in the rdfs. 3-4b graphically illustrates representative favorable interactions (distances $\leq 5 \text{ \AA}$) for the T20-lipid interaction pairs defined as d2-d4 (magenta dashed lines). The graphic shows a single simulation snapshot of T20 (orange sticks) complexed with gp41 (gray surface) interacting with ten representative conformations of two nearby lipids (green sticks). Lipid head group carbons omitted for clarity. 142

Appendix B.

Figure B-1. Representative examples of ROC curves showing good (a), random (b), and poor (c) global enrichments. 146

Figure B-2. (a) Schematic showing possible enrichments at 0.1%, 1.0%, and 2.0% of the database scanned and (b)-(d) maximum Fold Enrichment (FE) values at each of these points. $FE = AUC / AUC_{ran}$ 147

Figure B-3. (a) Hypothetical ROC curves computed using two different values for the total number of molecules classified as active (P, positive) or decoy (N, negative). The gray curve was computed using $P_{initial}$ and $N_{initial}$ and the black curve was computed using P_{docked} and N_{docked} . (b) Missing data can be assumed as yielding perfect enrichment (blue upper line), no enrichment (lower red line), or random enrichment (purple middle line) to ensure the ROC curve will reach $TPR=1, FPR=1$. The dashed line is the random ROC curve. 151

Figure B-4. Bar plots showing AUC values sorted from high to low using SB/DUD PDB naive pairing results in panel a. 155

Figure B-5. Heatmaps showing enrichment study which employ SB/DUD PDB results and DUD actives and decoys. Receptors are labeled by the pdb code and family on the y-axis. Ligands are labeled by the DUD system name on the x-axis. Red = worse than random ($AUC < 0.5$), green = greater than or equal to random ($AUC \geq 0.5$), and white = non-paired systems. 157

Figure B-6. Heatmaps showing enrichment study which employ the DUD SUP receptor and DUD actives and decoys. Receptors are labeled by the pdb code and family on the y-axis. Ligands are labeled by the DUD system name on the x-axis. The area under the ROC curve for each receptor-ligand pairing is used as a measure of enrichment. No data available (white),

better than random enrichment (green), or worse than random enrichment red is indicated by the colored cells. 159

Figure B-7. Differences between DUD PDB (red) and DUD SUP (green) structure preps for (a) 1a8i and (b) 1hw8. Native ligands are shown in cyan. Binding site waters in the DUD SUP prep for 1a8i indicated by blue circles. 161

Figure B-8. (a) ROC curves for the 40 DUD families. (b) ROC curves for the 10 Wombat families (Wombat ligands + DUD decoys). ROC curves sorted from high to low according to total AUC using SB/DUD PDB data with SB/DUD PDB in black and SB/DUD SUP in gray. 164

Figure B-9. Semi-random pairings run with DOCK6.6. (a) ROC curves for the 40 DUD families. (b) ROC curves for the 10 WOMBAT families (WOMBAT ligands DUD decoys). Results are sorted by Stony Brook structures results in black and gray is the provided organizer structures. Receptors are labeled by the pdb code and family on the y-axis. Ligands are labeled by the DUD system name on the x-axis..... 165

Figure B-10. ROC curves for 3 pairs of DUD families (a) egfr and hsp90, (b) fxa and thrombin, (c) trypsin and hivpr. Top panels show the native pairing. Bottom panels show the semi-random pairing. 166

List of Tables

Chapter 2.

Table 2-1. Experimental fold resistance (FR) values for ATP-competitive inhibitors with EGFR.	22
Table 2-2. Experimental versus calculated Fold Resistance (FR) energies ($\Delta\Delta G_{FR}$) and energy components for ligands with EGFR.....	36
Table 2-3. Absolute free energies and component decomposition for inhibitors with EGFR.....	40

Chapter 3.

Table 3-1. Examples of possible reference types to derive molecular footprints.....	60
Table 3-2. Comparison methods and corresponding ranges for footprint similarity (FPS) scores.	62
Table 3-3. Pose identification success using Footprint similarity (FPS) vs DOCK Cartesian energy (DCE) methods to rescore rigid (RGD), fixed anchor (FAD) and flexible ligand (FLX) pose ensembles.....	73
Table 3-4. FLX results scored with FPS_{VDW+ES} for three differing footprint similarity score cutoffs using a 2 Å rmsd to separate positive from negative regions.	80
Table 3-5. False negative examples for the range defined by the range $rmsd < 1.0$ Å and $FPS_{VDW+ES} > 1.0$	86
Table 3-6. Area under the curve (AUC) and accompanying fold enrichment (FE) statistics from receiver operator characteristic (ROC) plots for three protein-ligand systems.	92

Chapter 4.

Table 4-1. Footprint Similarity Score docking coefficients.....	105
Table 4-2. Standard docking and rescoring. FPS scoring uses normalized Euclidian.....	108

Appendix A.

Table A- 1. Autocorrelation function percent uncorrelated data (ACF %) and block average standard error of the mean (BASEM) for ΔG_{bind} calc results from simulations of T20 with wildtype HIVgp41 and seven mutants for various block lengths.	137
---	-----

Appendix B.

Table B- 1. DUD systems and semi-random matching information.	149
Table B-2. Enrichment completion statistics using DUD PDB preparation.....	152
Table B-3. Enrichment completion statistics using DUD SUP preparation.	153
Table B-4. Global enrichment (total AUC) for native and semi-random pairings.	155
Table B-5. Global breakdown of DUD systems into three sets (better, equal or worse than random) using the AUC to define enrichment.	160
Table B-6. Averaged early enrichment at 0.1%, 1.0%, and 2.0% of the databases examined. .	168

List of Abbreviations

ACF, autocorrelation function

BASEM, block averaged standard errors of the mean

d , Euclidian distance

DCE, DOCK Cartesian Energy

DGE, DOCK Grid Energy

d_{norm} , normalized Euclidian distance

DOPC, dioleoylphosphatidylcholine

EGFR, epidermal growth factor receptor

EM, electron microscopy

ES, electrostatics

FP, fusion peptide

FPS, Footprint Similarity

FR, fold resistance

GB, Generalized Born

HIVgp41, human immunodeficiency virus glycoprotein 41

PB, Poisson-Boltzmann

PDB, Protein Data Bank

MD, Molecular Dynamics

MM-GBSA, Molecular Mechanics Generalized Born Surface Area

NMR, nuclear magnetic resonance

NSCLC, Non-small cell lung cancer

r , Pearson correlation

rmsd, root-mean-square-deviation

r_{thresh} , threshold Pearson correlation

TKD, tyrosine kinase domain

TLI, Trilinear interpolation

TI, Thermodynamic Integration

VDW, van der Waals

Acknowledgments

Foremost, I am grateful to my advisor Robert C. Rizzo for accepting me in his group, aiding me in writing and obtaining a NIH fellowship, giving me good advice and suggestions and being there for me during my time in his lab. Through his guidance I have become a better scientist. I feel great gratitude to my parents Julie L. Eiseman and Fredrick A. Balius, Jr. who have given me love, support and guidance throughout my life and academic career. Without their persistence and dedication my path through life may have been very different. Marie C. Gelato has given her love, and support especially during my graduate work and I am very indebted to her. My fiancée Natalie St. Fleur, who I love, has been a great support and I am grateful to her for sharing her life with me. My brother Stuart A. Balius and my extended family continue to be a source of personal strength. I am also, grateful to Pauline Wetzel and Ruthie Wetzel for treating me like family.

From my time at Stony Brook University, I am grateful to my other committee members, Drs. David F. Green, Carlos Simmerling, and W. Todd Miller, all of whom have given me a great deal of guidance throughout my graduate career. I have valued working in the collaborative environment of the Rizzo Group. I want to recognize the support of the past and present members who I have enjoyed getting to know. Sudipto Mukherjee worked with me closely on DOCK development and code writing. His work on SB2010, and SB2012 testset construction has greatly aided the work on DOCK. He is a great colleague and I am grateful to have worked with him. For the HIVgp41 project, Brian McGillick did good work and I enjoyed collaborating with him. Rashi Goyal and Patrick Holden developed and validated the docking protocol used in this work. Yulin Huang is a hard worker and I am glad that I was able to work

closely with her on the Kinase projects. I express gratitude to Jie Wu for help in code development during the beginning stages of footprint-similarity-scoring work, to Kenneth Ascher, William Berger, Rashi Goyal, Patrick Holden, and Yulin Huang for computational assistance in testing the footprint similarity code, and Kenneth Foreman for helpful discussions. I thank William J. Allen for giving me much advice and guidance with the writing of this dissertation and my postdoc job search during my final year of graduate school. I thank Disability Support Services at Stony Brook University for their assistance in test-taking. The Green lab and Simmerling lab, both former and present members, have also been great colleagues whom I feel privilege to know. I thank the DOCK development team, especially Scott Brozell, Demetri Moustakas and P. Therese Lang.

From my time at the University of Pittsburgh at Greensburg, I would like to thank and acknowledge the following people. I am very grateful to Dr. Mary Grace O'Donnell for her advice and support. It is because of her that I majored in mathematics. Dr. Lou Ann Sears helped organize my academic accommodations and was a source of support. Dr. Nancy Estrada is a wonderful person and teacher who encouraged my travels and personal growth.

At University of Pittsburgh, the late Dr. Merrill J. Egorin supported me and gave me the opportunity to perform summer research during my undergraduate career. Dr. Robert S. Parker co-advised me and I learned much during my time in his group interacting with Drs. John M. Harrold, Abhishek S. Soni and Jeffrey A Florian Jr.

From my time in the Frederick County Public School System, many educators helped me along the way and I am especially grateful to Ms. Carol Forman, Mr. Frederick Brainerd, and Ms. Jay H. Heidel.

I thank all those who have come before me, including the first Homo sapiens and their descendants; but especially the developers of the DOCK, AMBER, CHIMERA, NAMD, and VMD programs. How did we function without Google and Wikipedia? Both of these resources made my work much easier. The Protein Databank and Pubmed were indispensable.

Finally, I would like to thank the funding agencies involved including the Stony Brook University Office of the Vice President for Research, the School of Medicine (Carol M. Baldwin Breast Cancer Research Award), the New York State Office of Science Technology and Academic Research (NYSTAR), and NIH grants R01GM083669 (to RCR) and F31CA134201 (to TEB).

Chapter 1. Introduction to Computational Structural Biology and Drug Discovery.

This chapter provides background and introduction to the theory and computational methods used in this dissertation.

1.1 Importance of Computational Techniques for Drug Discovery.

The maintenance and further development of a strong and robust drug discovery pipeline is important for combating disease in a timely manner as new threats emerge. As is shown in Figure 1-1, there are many stages in the drug discovery pathway from target identification to FDA approval. Traversal of a molecule through the drug discovery pathway is a long, (on average 9-12 years) and expensive (upwards of 800 million dollars) process.¹

Increasingly, computational methods are being used to aid drug discovery and overall have made considerable impact on molecular medicine.²⁻⁵ The ability to model molecules *in silico* at various levels of theory, isolating the reasons for binding affinity and specificity of drugs to their respective targets, is a powerful tool. Importantly, computational techniques are designed to complement experimental techniques. In addition, they are proven time and money saving methods. In particular, they are used to aid both the discovery and optimization stages of

the pipeline.³ The primary computational techniques described in this dissertation are molecular dynamics and docking and their associated analyses. The goal is to increase our ability to capture the energetic and structural changes that accompany protein ligand binding.

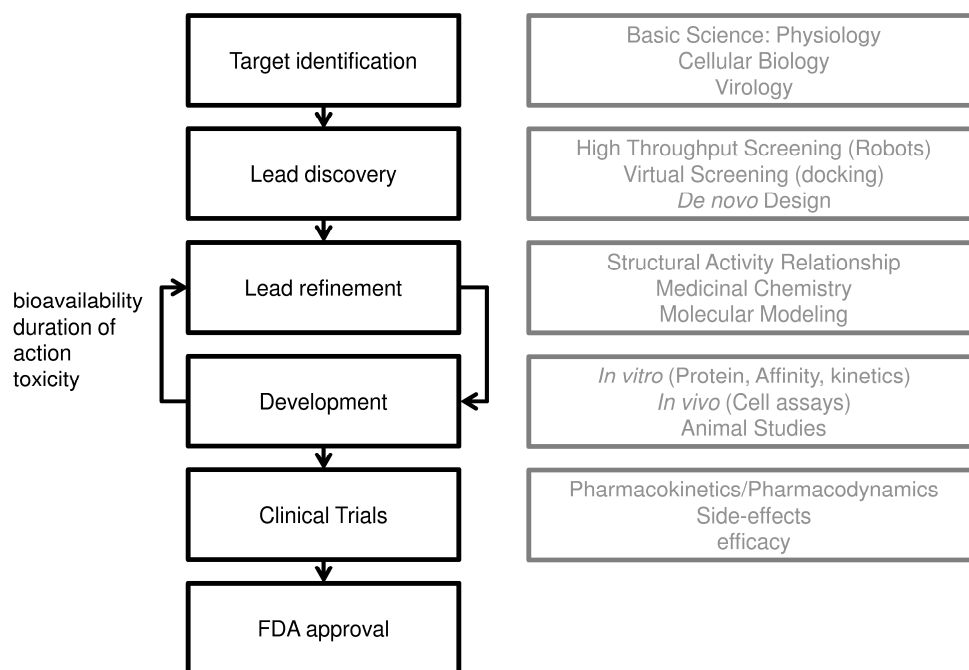


Figure 1-1. The drug discovery pathway shown in black. Techniques used at various stages of the pathway are shown in gray.

Molecular-targeted therapeutics are showing promise in combating disease including HIV/AIDS,^{6,7} influenza,⁸ and cancer.⁹ To focus on cancer, a leading cause of death,¹⁰ the very successful drug imatinib (Gleevec) used tools from structure-based drug design during development. Furthermore, imatinib was the first molecular-targeted therapeutic designed against a cancer target (the Bcr-Abl kinase in chronic myelogenous leukemia).¹¹ In addition, structure-based drug design including computational techniques have been highly useful in other kinase drug discovery projects.¹² In this work, the kinase Epidermal Growth Factor Receptor is studied in Chapter 2. It is important to note that computational techniques have also contributed

to the significant inroads made in combating HIV and its manifestation AIDS.⁶ Specifically, HIVgp41, an anti-fusion target, is studied in this dissertation also and is discussed in Appendix A.

An important goal is to understand the process of molecular recognition (Figure 1-2). For example, by calculating *in silico* the interactions of a small molecule (drug) with a biomolecule (protein) insights into binding may be obtained. By linking structural and energetic properties using physics-based energy functions (see subsection on molecular mechanics), it is possible to view the behavior of molecules at an atomic level of detail that is much more difficult experimentally, for example, the importance of specific interactions on an energetic basis, hydrogen-bonds, desolvation penalties, etc. Despite the approximations of classical molecular mechanics models, biological binding events can be captured remarkably well. For example, molecular dynamics simulations of small molecules bound with proteins (as described in Chapter 2) often show strong agreement between calculated and experimental values. Furthermore, docking calculations are often (ca. 70% of the time using DOCK) able to predicted a native-like binding geometry (within 2Å of a crystallographic pose) for a molecule in its active site.¹³ By comparing computational prediction with experimental results, confidence is gained that the underlying simulations are robust.

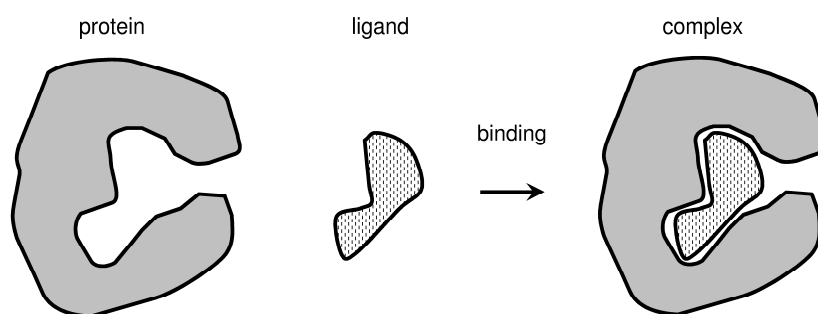


Figure 1-2. Molecular recognition and the binding event. The ligand and protein are shown in their free state on the left and then they associate together (represented by the arrow) to form the complex shown on the right.

Increasingly, an important drug discovery tool is virtual screening, the process of taking a large database of commercially available small molecule compounds and evaluating them in the computer to predict if they have activity. A work flow for virtual screening (using DOCK) is shown in Figure 1-3. A docking program (like DOCK), using a fast scoring method, is used to generate a rank ordered library for prioritization. Large vendor collections, such as those provided by the UCSF ZINC database,¹⁴ have made virtual screening easier for both academic and commercial labs. Often, compounds are rescored with more sophisticated functions. Top scoring compounds are then visually inspected and a subset is chosen for purchase and experimental testing for activity. As noted by “Tack” Kuntz a founder of the docking field,² the synergy between computation and experimentation should be emphasized (feedback loop in Figure 1-3). Current first line scoring functions in docking applications are very effective at enriching databases by discarding molecules that likely do not bind the target.¹⁵ However, first line scoring functions are often less effective at rank ordering, which highlights the need for more development. More accurate (expensive) methods typically use molecular dynamics simulations to generate ensemble-averaged energies. Nevertheless, even with their approximations, docking and virtual screening have a proven track record.^{15,16} See Chapter 3 and Appendix B for enrichment evaluation studies.

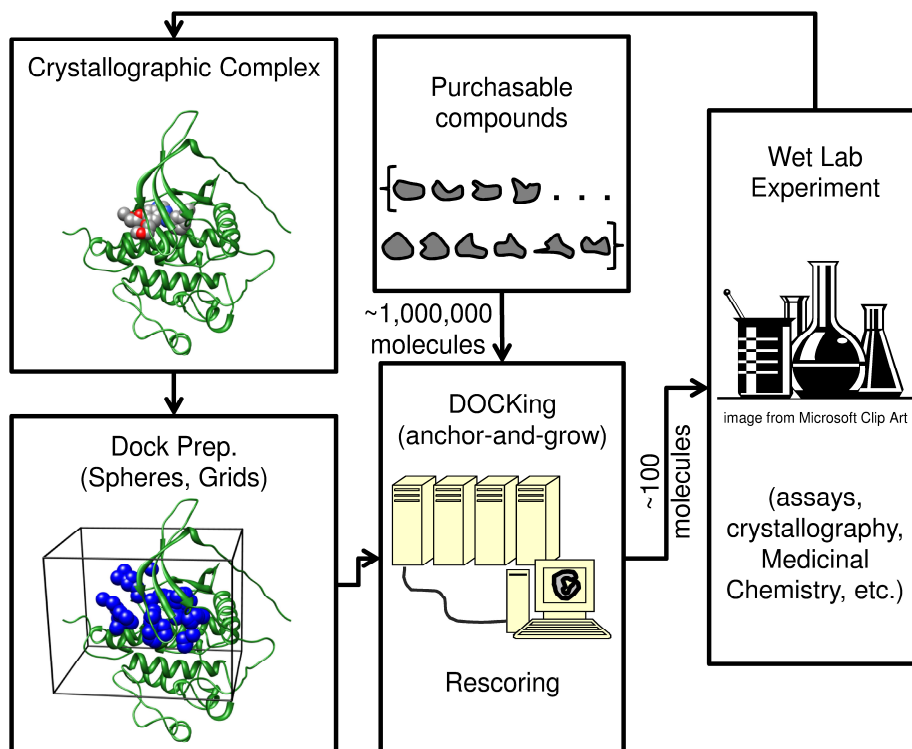


Figure 1-3. Virtual screening schematic for DOCK. A crystallographic protein structure is prepared for docking by creation of a sphere set and grid. Virtual screening is performed by docking ca. 1,000,000 molecules. On the order of 100 compounds are chosen for purchase and experimental testing. Perform Experimental testing on selected molecules. Feed-back loop demonstrates the relationship between computation and experimentation.

1.2 Structures in Drug Discovery

The study of biomolecules and their structure is an important aspect to understanding biology. For example, the mechanisms governing ligand binding, protein-protein interaction and enzymatic functions all may be elucidated through structural information. Although the experimental technique of X-ray crystallography provides static structures and the process of obtaining crystals may result in artifacts (e.g. crystal packing¹⁷), crystallography is an extremely powerful and important method to determine structures. More importantly, computational techniques, such as molecular modeling, link these structures with energy. Other techniques like molecular dynamics use the molecular mechanics energy function to simulate the motions of the molecule. To perform such calculations, a model of the molecules in the Cartesian (3-D)

coordinates with atom type and connectivity is needed. For proteins, the connectivity is known implicitly; however, for small molecules determining connectivity is more complicated and sometimes requires user input.

Often the initial structures are available from database repositories like the Protein Data Bank (PDB),¹⁸ which stores structures derived from experimental techniques such as X-ray crystallography and nuclear magnetic resonance (NMR), as well as electron microscopy (EM). Structural biology is very important to drug discovery and the pharmaceutical industry.¹⁸ Highlighting the continued relevance of structural biology, the PDB, since its founding in 1971,¹⁸ has grown rapidly and currently has over 78,000 structures.^{18,19} Among their many uses, these structures may serve as a starting point for homology modeling and molecular dynamics simulations. In addition they are used for target conformations in virtual screens, especially proteins crystalized with substrates. Furthermore, protein-ligand complexes are used in test set construction to evaluate docking protocols and programs.

1.3 Molecular Mechanics, Force Fields, and the Potential Energy Function.

The wealth of structural information can be used in combination with molecular mechanics to energetically describe the system. In molecular mechanics, a classical view of the world at the atomic level is imposed, and sets of parameters are applied to simple equations.²⁰⁻²² Figure 1-4 illustrates the 4 bonded terms and 2 through-space interactions. The total energy function is defined as follows:

$$E_{ff, top}(\vec{r}) = E_{\text{bonded}} + E_{\text{nonbonded}}$$

The potential energy E is dependent on the force field (ff) parameters and the topology (top), or connectivity, of the system and is a function of all of the coordinates.

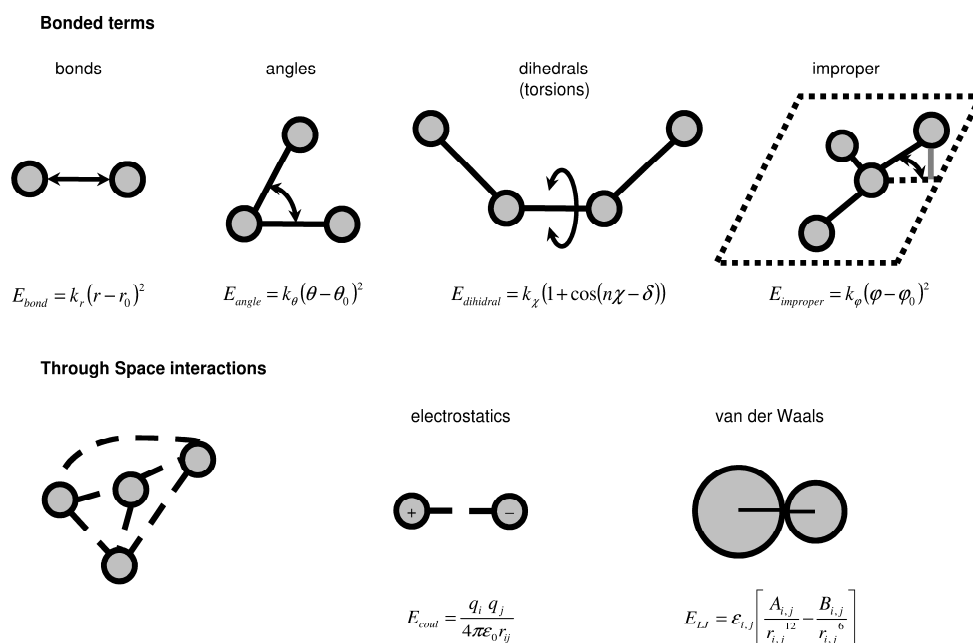


Figure 1-4. Molecular mechanics terms, divided into bonded and through-space interactions are presented. Each interaction type is labeled, illustrated and its equation is reported.

The bonded term equation is as follows:

$$E_{bonded} = \sum_{i \in \text{bonds}} k_r (r_i - r_0)^2 + \sum_{i \in \text{angles}} k_\theta (\theta_i - \theta_0)^2 + \sum_{i \in \text{dihed}} k_\chi (1 + \cos(n\chi_i - \delta)) + \sum_{i \in \text{improper}} k_\phi (\phi_i - \phi_0)^2$$

Bond (r) is a distance between two atoms; angle (θ) is calculated between 3 atoms; and the improper dihedral (ϕ) and dihedral angles (χ), are the angle between the two planes defined by four atoms in which two atoms are shared. The dihedral angle is defined by four consecutively bonded atoms; otherwise, it is an improper dihedral angle (Figure 1-4). In the AMBER force field (which is employed in this work),^{21,23,24} improper dihedral angles are treated the same as dihedral angles; however, other force fields model them differently as in the bonded equation above. This energy function consists of bonded terms where bonds, angles, and improper dihedral angles are modeled with Hooke's law that sets an ideal value (r_0 , θ_0 , and ϕ_0) and a force constant (k_r , k_θ , and k_ϕ). The dihedral angles follow a truncated Fourier series which allow for

multiple minima. The truncated Fourier series where k_γ is the height between maximum (top of peak) and minimum (bottom of well) points. The parameter n is used to define the period that is the length between the wells or peaks, by adjusting this n , more or fewer dihedral angles will be minima. The parameter δ is the offset this allows the shifting of ideal dihedral angles.

Non-bonded, or through-space, interactions are calculated with the following equation:

$$E_{\text{nonbonded}} = \sum_{i \in At} \left(\sum_{j \in At_i} \left(\frac{q_i q_j}{4\pi\epsilon_0 r_{i,j}} + \frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^6} \right) + \sum_{j \in At_{i,1-4}} s_e \left(\frac{q_i q_j}{4\pi\epsilon_0 r_{i,j}} \right) + s_v \left(\frac{A_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^6} \right) \right)$$

To account for through-space interactions, every atom is represented as a sphere with a radius and partial charge. In the above equation, At_i is the set of atom indices associated with atom i , such that At_i does not contain atom i or the 1-2 (atoms one bond away atom i), 1-3 (atoms two bonds away) or 1-4 (atoms three bonds away) atom indexes. These pairings are neglected due to proximity of the atoms. The 1-4 interactions are scaled (e.g. in AMBER: $s_e = 1 / 1.2$; $s_v = 1 / 2.0$).²¹ Often atoms beyond a cutoff are not included in the calculations to speed up the energy calculation. The variable q_i is the charge at atom i , the variable $r_{i,j}$ is the distance between atom i and atom j , $A_{i,j}$, $B_{i,j}$ are the van der Waals parameters which are defined by the well depth and radii of two atoms.

Next, we focus on sampling the conformational space of molecules, specifically protein-ligand complexes. In this dissertation, we primarily use molecular dynamics and docking (anchor-and-grow); other sampling approaches not discussed in this dissertation include Monte Carlo, and genetic algorithms.²⁵⁻²⁷

1.4 Molecular Dynamics.

Sampling. Molecular dynamics, employing a molecular mechanics force field, can be used to study the motion of molecules given a set of initial coordinates often obtained from the PDB. At room temperature molecules are in constant motion and therefore it is critical to look at an ensemble of structures in order to truly understand the energetics of binding. Let the vector $\vec{r} = (\vec{r}_1, \vec{r}_2, \dots, \vec{r}_N)$ where $\vec{r}_i = (x_i, y_i, z_i)$ specify the position of the atom i in Cartesian space. The forces on atom i can be calculated as follows:

$$\vec{F}_i = -\nabla_i E_{ff, top}(\vec{r})$$

$$\vec{a}_i = \vec{F}_i / m_i$$

$$\vec{a}_i = (a_{x,i}, a_{y,i}, a_{z,i})$$

Vector \vec{r} contains all atoms and has a length of $3N$. The force calculated here is only concerned with atom i . However, one can easily calculate the forces, $\vec{F} = (\vec{F}_1, \vec{F}_2, \dots, \vec{F}_N)$, and accelerations, $\vec{a} = (\vec{a}_1, \vec{a}_2, \dots, \vec{a}_N)$, on all atoms simultaneously.

$$v_{x,i}(t) = \int_0^t a_{x,i}(t') dt'$$

$$x_i(t) = \int_0^t v_{x,i}(t') dt'$$

The other y and z coordinates can likewise be calculated. The positions for all atoms are also calculated similarly. To illustrate the atom movement at each time step, we show a 2-D depiction (Figure 1-5). The dark gray circles move to light gray using the forces which are calculated for each atom based on the initial positions (dark gray circles) of all atoms. The light gray positions are then used as the initial positions for the next iteration.

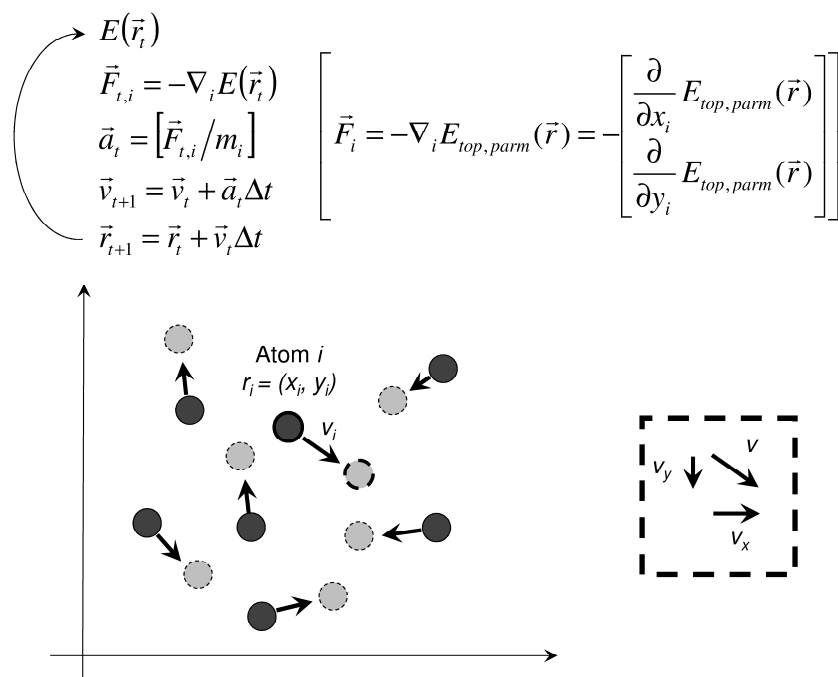


Figure 1-5. Two-dimension depiction of molecular dynamics. The iterative schema for propagating motion is outlined where i denotes the atom and t , the time step. The force (using energy function), the acceleration, velocities, and then new positions are obtained, consecutively. Dark gray circles are the initial positions of the atoms, and light gray are the new positions.

Velocity Verlet is a commonly-used numerical integration scheme for molecular dynamics.²⁸ In the AMBER software (in the sander module) the velocity Verlet leap frog implementation is used.²⁴

$$r(t + \Delta t) = r(t) + \Delta t v(t + 1/2\Delta t)$$

$$v(t + 1/2\Delta t) = v(t - 1/2\Delta t) + 1/2\Delta t a(t)$$

Here, Δt is the change in time or the time step. Typically a time step of 1-2 fs is used which is the vibration period of a bond. The velocity Verlet integration method has the property of conservation of energy,²⁸ which is very important for molecular dynamics and molecular modeling where researchers care about energetic properties.

In this dissertation, molecular dynamics is used to generate an ensemble of structures of a protein-ligand complex. The simulations are not meant to sample large conformational changes

but just sample about the starting coordinates typically initiated from X-ray structures to be able to better understand small-molecule protein interactions.

There are many post-processing methods which aid in the analysis of molecular dynamics trajectories including structural and energetic analysis. In this dissertation, root-mean squared deviation, block averaging standard errors of the mean, and autocorrelation functions are used to evaluate simulation stability and noise (Appendix A). When free energy calculations and experimental affinity data show agreement (Chapter 2 and Appendix A), decompositions of energies (including per-residue footprint analysis) may be used to explain the observed trends.

Calculating Binding Free Energies. Hydration (or solvation) is an incredibly important part of binding energy due to two important facts: (1) biology happens in an aqueous environment and (2) before something can bind an active site it must displace waters. One popular method for calculating free energy of binding is the Molecular Mechanics Generalized Born Solvent Accessible Surface Area (MM-GBSA) Method.^{29,30} Although an approximate method, MM-GBSA has many strengths. It is a relatively inexpensive method, topologically different structures can readily be compared and the simulations are easy to set up and run. As discussed below (Chapter 2 and Appendix A), molecular dynamics simulations are often run fully solvated in water and then the water are removed and the simulation is rescored with Generalized Born (GB) or alternatively Poisson-Boltzmann (PB) models.³⁰ Both GB and PB account for the desolvation effects and are both implicit water models. GB is an approximation of PB and is faster to calculate and is used in work discussed in Chapter 2. Both methods, however, are continuous, thus they can reproduce bulk-water interactions, but are not able to account for the discrete nature of water (See Chapter 2 for further discussion of MM-GBSA). Both GB and PB account for the entropy of solvent, but do not for solutes (protein and ligand).

The entropy of the solute is often ignored as is done in Chapters 2 and Appendix A, but the entropy term may be calculated using normal mode analysis or quasi-harmonic analysis.³¹ Alternative methods for calculating free energies of binding such as thermodynamic integration or free energy perturbation use an alchemical path to calculate free energy differences between two states. Although these alchemical methods are well grounded in theory and are proven accurate;³²⁻³⁴ they are more expensive, are difficult to set up and may only compare systems with small changes.

Per-residue energy decomposition. Molecular footprints are useful in gauging which residues are most important for binding (Chapter 2 and Appendix A). A footprint is the per-residue decomposition of the through space interaction energies between receptor and ligand. Decomposing the energetics in this way is possible with additive force fields used in this work. Footprints are used in all of the chapters below. Figure 1-6a shows a cartoon matrix of per-atom through-space interactions: The blue square, red square and the purple rectangles are the internal receptor interactions, internal ligand interactions and the intermolecular interactions respectively. The equation representing the summed energy are also color-coded. The matrix illustrates the ability to look at the decomposition of the interactions, one can look at interactions of pairs of atoms, or interaction of groups of atoms as is shown in the molecular-footprint cartoon in Figure 1-6b. One may also wish to look at an interaction matrix, the interaction of every residue with every residue as in Appendix A. In addition, footprints can be useful for identifying like binding ligands (Chapters 3 and 4). In the equation in Figure 1-4, *lig*, *rec*, and *comp* stand for the sets of ligand, receptor, and complex atoms, respectively. Likewise, E_{lig} , E_{rec} , and E_{comp} is the energy of the ligand, receptor, and complex atoms, respectively. $E_{lig,rec}$ is the through space or non-bonded interactions. ΔE approximates the change in energy from free state (Figure 1-2, left of arrow) to

bound state (Figure 1-2, right of arrow). This energy can be further decomposed into per residue components represented by vector E_{fp} . $resid(i)$ represents the set of atoms in residue i of the protein. See the method sections in Chapters 2, 3, and 4 for applications and more description of footprints.

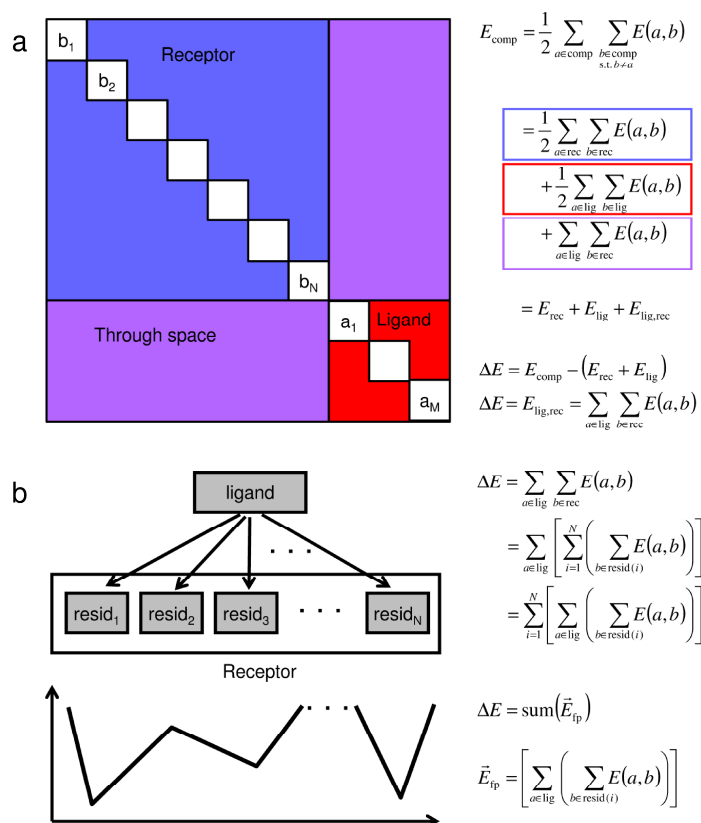


Figure 1-6. (a) Matrix interaction of all atom pairs, blue is the internal energy of the receptor, red is the internal energy of the ligand and purple is the through-space interactions. (b) Per-residue decomposition of the through-space interactions.

1.5 DOCKing.

Docking programs perform two tasks: (1) sample the correct pose and (2) score the poses correctly. Docking algorithms are used to place a small molecule in a protein pocket. Docking has been described using the lock-and-key metaphor; however, docking is more complicated due

to ligand and receptor flexibility. The program DOCK^{26,35-37} is the first widely used docking program and was developed in the group of Irwin D. Kuntz. Work described in Chapter 3 was released as DOCK v6.5, and work in Chapter 4 and Appendix B will be incorporated in a future release of DOCK.

Sampling. The DOCK 4²⁶ program introduced an algorithm called anchor-and-grow (A&G) for on-the-fly ligand growth in the context of a grid (calculated from a rigid receptor). DOCK 5,³⁶ and its descendent DOCK 6,³⁷ both continue to use A&G as their main ligand sampling tool. There are alternative sampling methods such as the hierarchical database approach,^{38,39} where ligand growth is pre-generated (i.e. just generated once) and can be docked rigidly to many different targets. The conformational pre-generation is preferred computationally if one docks to multiple targets otherwise there is no benefit over on-the-fly growth. On-the-fly growth is more convenient and may allow for sampling to be guided by the scoring functions, which may result in the ability to focus the sampling method at all levels of growth.

The A&G algorithm is as follows: first, rigid segments are identified (Figure 1-7a) and the molecule is broken up in to these segments (Figure 1-7b). Then, an anchor is chosen (Figure 1-7c). More than one anchor can be chosen (Figure 1-7d); the anchor list is sorted by number of heavy atoms and number of attachment points. A segment may be assigned to a different layer when an alternative anchor is chosen. For example, two different anchors are chosen in Figure 1-7 c and d, and the segments are arranged into layers differently for the most part. The anchor is then oriented into the pocket by using receptor spheres, which define a reverse image to the pocket to limit the search space. Growth from viable anchor orients then occurs. The A&G algorithm is a breadth-first method.²⁶ Growth proceeds by sampling torsions of each segment

one at a time. All nodes of the growth tree shown in Figure 1-7e placed within a grey stripe are enumerated before proceeding to the next step. In addition, segments are divided into layers, all segments in the layer are sampled before moving to the next layer. An exponential explosion of conformers is limited during growth by pruning and clustering as indicated by the red circle marked by an X in Figure 1-7e.

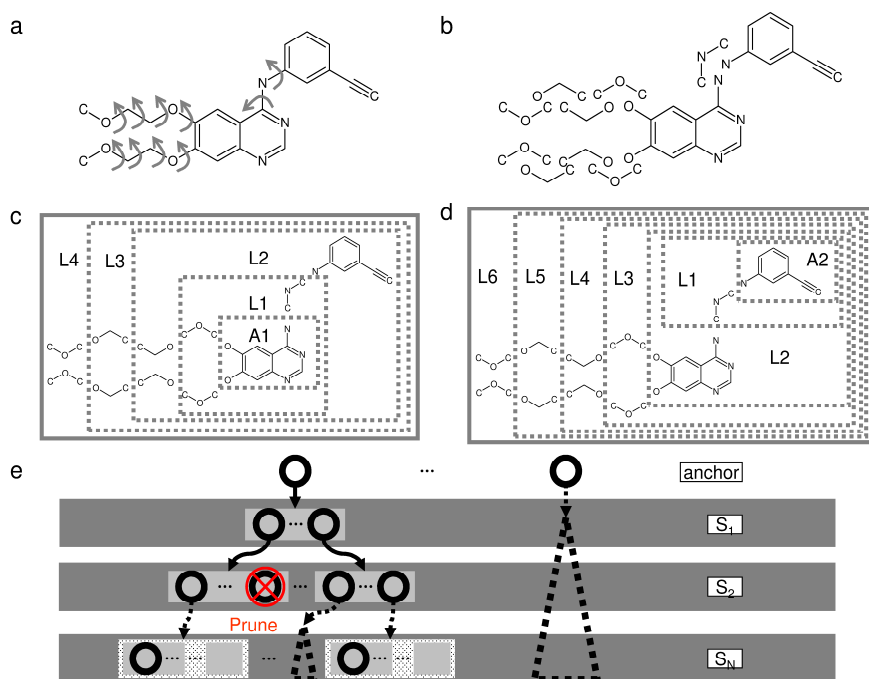


Figure 1-7. (a) Structure of Erlotinib is shown with labeled rotatable bonds. (b) The molecule is broken up into rigid segments. The definition of layers is defined: (c) for the first anchor (A1); or (d) for the second anchor (A2). (e) Cartoon of the growth tree is shown.

Scoring. DOCK uses a simplified molecular mechanics force field as the main scoring function, but only the intermolecular interaction energy is calculated.⁴⁰ DOCK does not calculate the bond, angle or dihedral terms; however, DOCK calculates an internal energy of the ligand, which consists of only the repulsive van der Waals term for DOCK 6.4,¹³ in order to prevent ligand internal clashes. The grid energy is used during standard docking to speed up the energy calculations.⁴⁰ The grid is generated by pre-computing interactions as is discussed in

more detail in Chapter 4. DOCK also has several alternative scoring and rescoring methods. Often it is useful to rescore docked results using more computationally expensive methods as is discussed in Chapter 3. Docking runs, like molecular dynamics, may also be post-processed. For example docking results might be rescored using alternative scoring functions to more accurately predict binding affinities of poses or molecules. An alternative rescoring method used to identify ligands that bind similarly to a reference molecule is described in Chapter 3. In addition chemical informatics techniques such as fingerprinting may be employed for clustering to identify chemically similar molecules.⁴¹

1.6 Research Projects.

This dissertation describes several research projects involving application and development of computational techniques with the purpose of aiding in the design of improved inhibitors for medically relevant targets (proteins). The project described in Chapter 2 involved simulations of the anti-cancer target Epidermal Growth Factor Receptor (EGFR) variants bound with three inhibitors. The purpose of the study is to understand origins of resistance.⁴² In the next project shown in Chapter 3, the Footprint Similarity (FPS) rescoring function was developed in DOCK and evaluated using pose reproduction, cross docking and enrichment experiments.⁴³ An extension of the FPS method to a Grid-based scoring function is shown in Chapter 4. In the final chapter (Chapter 5), impact of these studies, current and related work, and future directions are discussed. Appendixes A and B feature work from two collaborative projects: (A) analysis of binding of antifusion peptides targeting HIVgp41,⁴⁴ and (B) evaluation of DOCK6 as an enrichment tool for virtual screening.⁴⁵

Chapter 2. Quantitative Prediction of Fold Resistance for Inhibitors of EGFR.

This chapter has been published as **Balius, T. E.**; Rizzo, R. C. Quantitative Prediction of Fold Resistance for Inhibitors of EGFR. *Biochemistry*, **2009**, *48* (35), 8435-8448. Copyright © 2009 American Chemical Society. [doi:10.1021/bi900729a](https://doi.org/10.1021/bi900729a) PMID: 19627157

Author contributions. TEB and RCR designed research; TEB performed research, analyzed data, and wrote initial draft; TEB and RCR wrote the paper.

Abstract.

Clinical use of ATP-competitive inhibitors of the epidermal growth factor receptor (EGFR) kinase domain can lead to an acquired drug resistant mutant L858R&T790M which dramatically reduces binding affinity relative to a prevalent cancer causing mutation L858R. In this study, we have used molecular dynamics (MD) computer simulations, free energy calculations (MM-GBSA method), and per-residue footprint analysis to characterize binding of three inhibitors (erlotinib, gefitinib, and AEE788) with wildtype EGFR and three mutants. The goal is to characterize how variation in structure and energy correlate with changes in

experimental activities and to deduce origins of drug resistance. For seven fold resistance values, each computed from the difference of two independent computer simulations, excellent agreement was obtained with available experimental data ($r^2 = 0.84$). Importantly, the results correctly predict that affinity will increase as a result of L858R and decrease due to L858R&T790M. Per-residue analysis shows an increase in favorable packing at the site of the methionine mutation reaffirming a steric clash hypothesis is unlikely, however, large losses in van der Waals, Coulombic, and H-bond interactions strongly suggest that resistance is not due solely to changes in affinity for the native substrate ATP as recently proposed. Instead, the present results indicate that drug resistance more likely involves disruption of favorable interactions, including a water-mediated H-bond network between the ligands and residues T854, T790, and Q791, which could have important implication for guiding rational design of inhibitors with improved resistance profiles.

2.1 Introduction.

Cancer is the second highest cause of death within the United States led by lung and bronchial cancers for which an estimated 215,000 new cases and 161,000 deaths were reported in 2008.¹⁰ Non-small cell lung cancer (NSCLC) comprises the largest subset of lung cancers.⁴⁶ A major oncogene that drives tumorigenesis in NSCLC, as well as other types of cancer, is the membrane receptor tyrosine kinase Epidermal Growth Factor Receptor (EGFR). Overexpression of EGFR is observed in 62% of NSCLC tumors⁴⁷ and its role in mediating tumor cell growth and survival for NSCLC, as well as many other types of cancer, has been well described.⁴⁸⁻⁵⁰ The importance of EFGR has been clinically validated, and within the past several years, inhibitors of

EGFR have been approved for treatment of NSCLC, pancreatic, colorectal, head and neck, and breast cancers.^{48,50,51}

Structurally, EGFR can be divided into five distinct regions: an extra-cellular ligand binding domain, a trans-membrane domain, an intracellular juxtamembrane domain, an intracellular tyrosine kinase domain (TKD), and a C-tail region where phosphorylation occurs.^{52,53} Normally, EGFR is a monomer. However, extra-cellular ligand binding of endogenous EGF (epidermal growth factor) promotes dimerization with another protein from the ErbB family such as EGFR (ErbB1 or HER1), ErbB2 (HER2), ErbB3 (HER3), or ErbB4 (HER4).^{48,54} The homo- or heterodimerization event induces a conformational shift in the TKD from an inactive to active form.^{52,53,55} Activation results in binding of ATP, phosphorylation, and signal transduction through a number of downstream pathways.^{48,54,56} Normally, signaling activity is under tight regulatory control. However, cancer causing mutations can result in constitutive activation of EGFR.⁵³ ATP-competitive inhibitors have been described that preferentially bind the active or inactive conformation.^{57,58} The primary structural differences between the active and inactive forms is a conformational shift in the TKD activation loop and movement of the N-lobe helix, both of which are located near the ATP binding site (Figure 2-1).

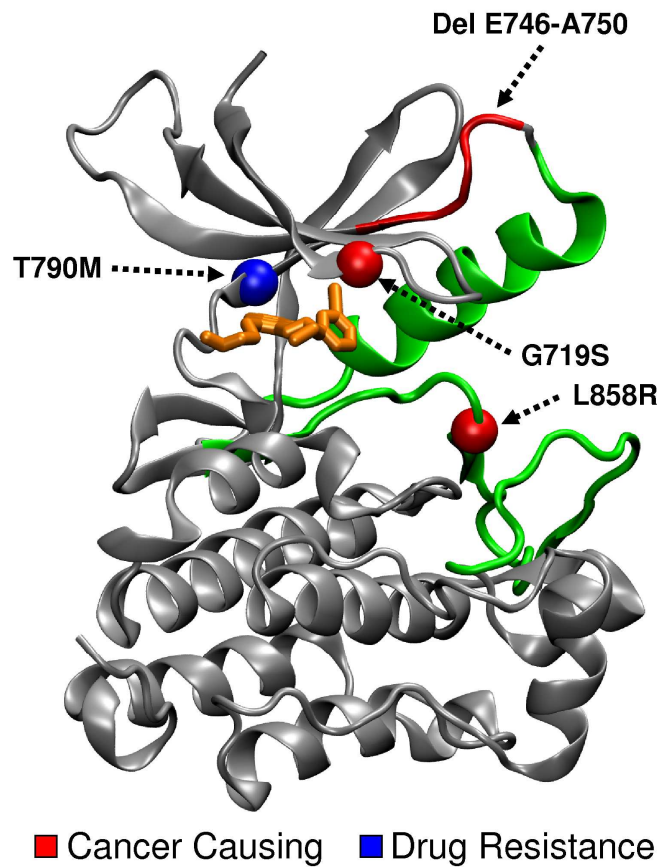
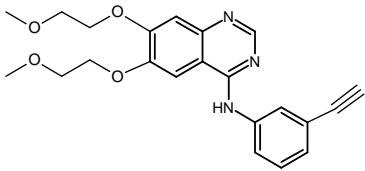
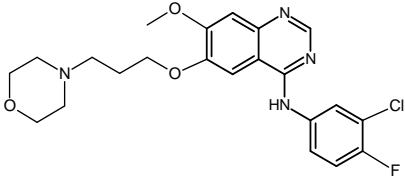
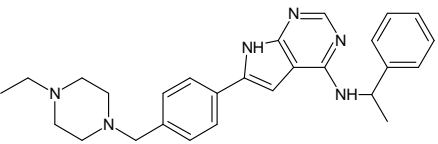


Figure 2-1. Ribbon diagram showing EGFR complexed with the ATP-competitive inhibitor erlotinib. Regions which change conformation (N-lobe helix and activation loop) upon receptor activation are shown in green. Locations of cancer causing mutations (deletion or point) which cause receptor activation are in red. The secondary T790M drug resistance mutation is shown in blue. Coordinates from pdb code 1M17.

There are two classes of inhibitors of EGFR: (i) monoclonal antibodies such as Cetuximab (IMC-C225) which target the extracellular domain and block binding of native EGF ligand to the receptor, and (ii) small molecules that compete with ATP in the intracellular TKD and block activity, regardless of endogenous ligand binding (Table 2-1).^{48-50,59} Focusing on ATP competitive inhibitors, approved small molecules of the TKD domain include erlotinib (Tarceva, OSI Pharmaceuticals), gefitinib (Iressa, AstraZeneca), and lapatinib (Tykerb, GlaxoSmithKline).⁵¹ A fourth compound called AEE788 (Novartis) is in development.⁶⁰ Although erlotinib and gefitinib primarily target EGFR, multi-receptor inhibition is possible

given the high structural homology of the TKD.^{48,57,60,61} Lapatinib is a dual inhibitor of EGFR and ErbB2^{57,59} and AEE788 binds EGFR, ErbB2, and the related VEGF receptor.^{48,60} Erlotinib is label-approved for use against NSCLC and pancreatic cancer while lapatinib is approved to treat patients with advanced or metastatic breast cancer whose tumors also overexpress HER2.⁵¹ Gefitinib was originally approved to treat NSCLC however the FDA has limited its usage given that no significant effect on patient survival was found.⁵¹ AEE788 is being evaluated as therapy for brain and central nervous system cancers.⁶⁰

Table 2-1. Experimental fold resistance (FR) values for ATP-competitive inhibitors with EGFR.

Inhibitor	Structure	Experimental Fold Resistance ^a		
		L858R / WT	L858R&T790M / L858R	G719S / WT
erlotinib		6.25 / 17.5 nM ^b 0.36 FR -0.61 $\Delta\Delta G_{FR}$	>10000 / 12.5 nM ^c >800 FR >3.96 $\Delta\Delta G_{FR}$	—
gefitinib		2.4 / 35.3 nM ^d 0.068 FR -1.59 $\Delta\Delta G_{FR}$	10.9 / 2.4 nM ^d 4.54 FR 0.90 $\Delta\Delta G_{FR}$	123.6 / 53.5 nM ^e 2.31 FR 0.50 $\Delta\Delta G_{FR}$
AEE788		1.1 / 5.3 nM ^d 0.21 FR -0.92 $\Delta\Delta G_{FR}$	18.6 / 1.1 nM ^d 16.9 FR 1.68 $\Delta\Delta G_{FR}$	11.3 / 10.9 nM ^e 1.04 FR 0.02 $\Delta\Delta G_{FR}$

^aFold Resistance (FR) = ratio of experimental activities. $\Delta\Delta G_{FR} \text{ exptl} \approx RT \ln(\text{FR})$ at 298.15 K in kcal/mol. ^bK_i values (nM) from Carey et al.⁶² ^cIC₅₀ values (nM) from Ji et al.⁶³ ^dK_d values (nM) from Yun et al.⁶⁴ ^eK_d values (nM) from Yun et al.⁶⁵

Several cancer causing mutations in EGFR have been reported which map to either the extracellular ligand binding domain (e.g. in glioblastoma)^{63,66} or the TKD region (e.g. in NSCLC),⁶⁷⁻⁷¹ which cause activation of EGFR independent of EGF-ligand binding. For the TKD domain, such cancer causing mutations can occur at positions L858R or G719S (point mutations), and E746-A750 or E746-S752 (referred to as exon 19 deletions).⁶⁷⁻⁷¹ L858R and exon 19 deletions are the most frequent mutations in NSCLC.^{72,73} Figure 2-1 shows the TKD domain cancer causing mutations mapped to EGFR in red. Although patients with wild type EGFR benefit from low molecular weight inhibitors, patients whose tumors harbor activating L858R or deletion mutations are especially responsive to erlotinib and gefitinib treatment.⁶⁹ Interestingly, in contrast to most systems in which mutations lead to a decrease in binding, studies have shown that affinity is enhanced for TKD ligands for L858R over wild type (Table 2-1). As these ligands bind preferentially to the active conformation this could explain their enhanced binding affinity toward the mutants. In contrast, decreases in affinity for gefitinib and AEE788 have been reported relative to wildtype for the G719S point mutant (Table 2-1).

As with many chemotherapeutics, acquired resistance to current EGFR inhibitors can occur with continued use.⁷⁴ A T790M resistance-mutation is commonly observed in patients treated with erlotinib and gefitinib for those tumors which also harbor the primary cancer causing point mutation at position L858R or exon 19 deletions.⁷⁴ The location of T790M is shown mapped in blue to the TKD site on EGFR in Figure 2-1. As shown in Table 2-1, the double mutant L858R&T790M shows marked decrease in inhibition when compared to the active L858R mutant alone for all three ligands. Erlotinib in particular shows a large >800 fold resistance (FR = ratio of activities). Similar secondary drug resistance mutations have been described previously for other molecular targeted therapeutics such as Gleevec. The well-known

BCR-Abl kinase resistance mutation at position T315I, referred to as the "gate keeper", arises from treatment with Gleevec.^{75,76} The gate keeper is important in modulating selectivity and affinity for BCR-Abl inhibitors and analogous here to the point mutant in EGFR at position T790M.

An improved understanding of the molecular determinates that drive ligand binding for EGFR is critical for development of improved inhibitors. Prior computational studies of this system have included use of homology and molecular modeling,⁷⁷ comparative molecular field analysis (CoMFA),⁷⁸ virtual screening,⁷⁹ and molecular dynamics.^{77,78,80,81} Use of MM-PBSA methods, similar in principle to the calculations employed in the present chapter, were reported by Hou et al.⁷⁸ for refinement of docked ligand poses, and by Liu et al.⁸⁰ to study the impact of point mutations on binding for gefitinib. Surprisingly, there have been few all-atom molecular dynamics studies reporting quantitative binding energy comparisons between theory and experiment for ligands with EGFR. In this report, we have carried out simulations of the TKD of EGFR in complex with three ATP-competitive inhibitors to investigate the effects of clinically relevant point mutations on ligand binding. Studies to address deletion mutations are in progress. Specifically, goals of the present project are threefold: (i) Development of robust quantitative computational models to study EGFR-ligand binding for wildtype, L858R, G719S, and the drug resistant double mutant L858R&T790M. (ii) Determine how variation in structural and energetic results correlate with variation in reported experimental activities. (iii) Deduce the origins of drug resistance. Characterization of FR at the molecular level will ultimately enable development of next generation compounds with improved resistance profiles.

2.2 Methods.

2.2.1 Binding Free Energies.

Accurate calculation of protein-ligand binding energies remains an important and challenging problem. In this report, we employ the molecular mechanics Generalized Born solvent accessible surface area (MM-GBSA) method^{29,30} to computationally estimate binding free energies (ΔG_b) for inhibitors with EGFR. Although considered to be an approximate free energy theory, the benefits of MM-GBSA include relative ease of set up, low computational overhead, and systems with dissimilar topologies can be more readily examined in comparison to other methods such as free energy perturbation. Projects in our laboratory employing similar protocols to that reported here, which successfully used MM-GBSA to study protein-ligand binding, include HIVgp41,⁸² neuraminidase,⁸³ and MMP-13.⁸⁴

The method relies on the thermodynamic relationship shown in Figure 2-2 to estimate the free energy of binding (ΔG_b calcd/exptl) which occurs in the condensed phase. The computed free energy of binding is estimated as the sum of nonbonded gas-phase (ΔG_{gas}) interactions modulated by the overall change in hydration free energy ($\Delta\Delta G_{\text{hyd}}$) for the complexation event.^{29,30} The hydration term accounts for important desolvation penalties, which include changes in entropy due to the hydrophobic effect, that occur as a result of unbound solvated species coming together to form a complex. Additional terms to include estimates for changes in solute entropy were not included in the present study.

A molecular dynamics trajectory of each protein-ligand complex is performed in explicit solvent with system energies, as well as root-mean-square-deviations (rmsd), being monitored for stability and convergence. For MM-GBSA analysis, the explicit solvent is stripped off and coordinates are separated into three individual species (complex, receptor, and ligand) with eqs

2-1 to 2-3 being used to compute the total binding affinity.^{29,30} The relevant individual energy terms include van der Waals (ΔE_{vdw}), Coulombic (ΔE_{coul}), polar (ΔG_{polar}), and non-polar ($\Delta G_{\text{nonpolar}}$) contributions. Generalized Born (ΔG_{polar}) and solvent accessible surface area ($\Delta G_{\text{nonpolar}}$) calculations are used to estimate ΔG_{hyd} for each individual species.⁸⁵

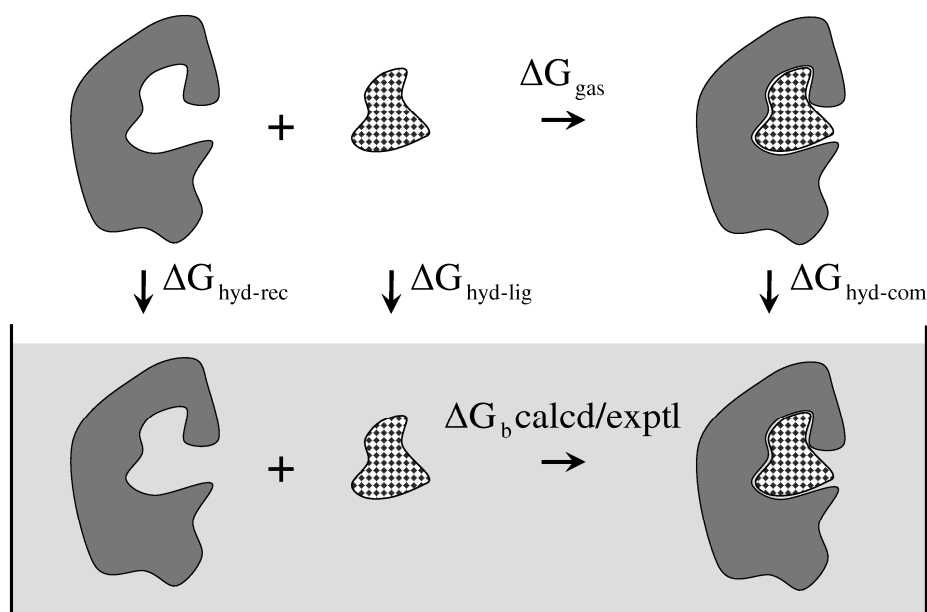


Figure 2-2. Schematic representation of the thermodynamic cycle used to calculate free energies of binding (ΔG_{b} calcd) for comparison with experiment (ΔG_{b} exptl). The cycle highlights the relationship between ΔG_{b} exptl occurring in condensed phase with the free energy of interaction in the gas-phase (ΔG_{gas}) modulated by three terms representing the free energy of hydration (ΔG_{hyd}) for the transfer from vacuum to water for each separate species (com=complex, rec=receptor, lig=ligand).

$$\Delta G_{\text{b}} \text{ exptl} \approx \Delta G_{\text{b}} \text{ calcd} = \Delta G_{\text{gas}} + \Delta G_{\text{hyd-com}} - (\Delta G_{\text{hyd-rec}} + \Delta G_{\text{hyd-lig}}) \quad (2-1)$$

$$\Delta G_{\text{gas}} = \Delta E_{\text{vdw}} + \Delta E_{\text{coul}} \quad (2-2)$$

$$\Delta G_{\text{hyd-species}} = \Delta G_{\text{polar}} + \Delta G_{\text{nonpolar}} \quad (2-3)$$

2.2.2 Interaction Signatures: Molecular Footprints.

To identify important binding site residues and characterize how interactions may change as a result of mutation, structural and energetic molecular "footprints" were computed for each MD trajectory. Footprints represent the per-residue decomposition of interactions, averaged over the production simulations, between each EGFR residue and the inhibitors. Our laboratory has successfully used such footprints to deduce origins of resistance conferred by a R292K mutation for sialic acid-based inhibitors of neuraminidase,⁸³ and to show that the hydrophobic pocket region on HIVgp41 is an important drug target site for modulating binding affinity.⁸² Separate footprints for Coulombic and van der Waals energy, as well as hydrogen bonds were computed. Difference footprints were also computed, using results from the L858R&T790M – L858R simulations, and represent the change in energy (or H-bonds) at each residue due to mutation.

System Setups: A single set of receptor coordinates (pdb code 1M17),⁵⁸ of EGFR in the active form, was used as the basis for construction of all simulation setups. All solvent was removed from the 1M17 structure and only the TKD of EGFR (defined as a.a. numbers 710-983) were retained. Initial geometries for ligands erlotinib, gefitinib, and AEE788, were obtained from 1M17,⁵⁸ 2ITY,⁶⁵ and 2J6M⁶⁵ pdb codes, respectively and placed into the master 1M17 reference frame through alignment of receptor backbone atoms in common with each pdb. Mutant forms of EGFR (L858R, L858R&T790M, and G719S) were obtained through manual modification of 1M17 to the desired residue(s). Starting rotameric states for modified side chains were made using energetic packing consideration subject to visual inspection to ensure there were no intermolecular clashes as a result of model building. The MOE⁸⁶ program was used for initial preparation of ligand (mol2 format) and receptor (pdb format) files for subsequent processing. The AMBER⁸⁷ program modules *leap* and *antechamber* were used to assemble,

solvate (10 Å buffer), and assign force-field parameters for each complex consisting of FF99SB²³ (protein), TIP3P⁸⁸ (solvent), and GAFF⁸⁹ (ligand). For the ligands, partial atomic charges were obtained at the HF/6-31G**//HF/6-31G* level of theory via the ChelpG⁹⁰ method using Gaussian98.⁹¹ All ligands were modeled as having a net zero charge. Unless otherwise stated system setups employed default input parameters for each program. The size of the complete model was 274 receptor residues plus one ligand residue solvated in a TIP3P periodic solvent box of ca. size $77 \times 89 \times 75 \text{ \AA}^3$ containing ca. 14,050 waters.

2.2.3 Simulation Protocols.

A nine step equilibration protocol, consisting of short energy minimizations and molecular dynamics (MD), was used to eliminate any unfavorable interactions which may have occurred as a result of model building and to gently adjust the starting structure to the molecular mechanics force field prior to production MD. Heavy atoms of the complex were initially restrained to the crystallographic coordinates using a harmonic restraint force constant of 5 kcal/mol Å² with water molecules and hydrogen atoms free to move during 1000 steps of steepest decent energy minimization (step 1). Each subsequent equilibration step used the last set of coordinates from the previous step as the restraint reference structure. Next, the same restraint mask and coefficient were used for 50ps of MD in which waters and hydrogens were further able to adjust (step 2). This was followed by three minimizations of 1000 steps each in which the restraints were reduced from 2, to 0.1, to 0.05 kcal/mol Å² respectively (steps 3-5). Three additional MD runs of 50ps were run where weights were reduced from 1 to 0.5 (steps 6-7) followed by 0.1 kcal/mol Å² (step 8) with only backbone receptor atoms at C α , C, and N being restrained in the latter step. The final 50ps MD equilibration step used the 0.1 kcal/mol Å² receptor backbone weight, but only the last three residues on each N- and C-terminus were

restrained (step 9). Production MD employed the same weak restraints as the final equilibration step and was extended to 5ns with the restraint reference updated every ns. A time step of 1fs was used for equilibration and increased to 2fs for production MD which concurrently required use of the SHAKE⁹² algorithm. Coordinates for post processing were saved every 1ps. Long range electrostatics were computed using the particle mesh Ewald (PME)⁹³ with a real-space cut off of 8 Å. A constant temperature of 298.15 K and pressure of 1 bar was maintained during the simulations through Berendsen schemes⁹⁴ with heat bath coupling and pressure relaxation time constants of 1.0 ps. These calculations employed the AMBER8 *sander* module.

2.2.4 Analysis.

Binding free energies, fold resistance, and molecular footprints were obtained from post-processing of each protein-ligand complex MD trajectory. All waters were removed unless otherwise stated. Individual snapshots (N=5000) were split into coordinates representing separate ligand, receptor and complex and single point calculations using *sander* were performed to obtain the energy terms (ΔE_{coul} , ΔE_{vdw} , ΔG_{polar} , $\Delta G_{\text{nonpolar}}$) used to compute free energies of binding (Figure 2-2, eqs 2-1 to 2-3). As in previously reported studies from our laboratory⁸²⁻⁸⁴, the GB model implemented into AMBER8 described by Onufriev et al.⁹⁵ (type igb=5) was used to estimate polar energies (ΔG_{polar}) with mbondi2 radii and dielectric constants of 1 and 78.5. Nonpolar energies ($\Delta G_{\text{nonpolar}}$) were obtained from solvent accessible surface area calculations via $\Delta G_{\text{nonpolar}} = \gamma \text{SASA} + \beta$ using standard constants of $\gamma = 0.00542 \text{ kcal/mol } \text{Å}^2$ and $\beta = 0.92 \text{ kcal/mol}$.^{30,96} GBSA calculations for molecules containing fluorine and chlorine employed radii of 1.50 Å (F) and 1.70 Å (Cl) which required modification to the AMBER8 distribution file src/sander/mdread.f. In-house scripts were used to compute the per-residue decomposition

(molecular footprints) for intermolecular H-bonds, Coulombic, and van der Waals interactions for which the sum over all the EGFR residues is equivalent to the total value (i.e. ΔE_{coul} , ΔE_{vdw} , H-bond). Hydrogen bonds were defined as a structural interaction between three atoms X_D - H_D -- X_A with a distance less than or equal to 2.5 Å and angle between 120 and 180 degrees. The NAMD program⁹⁷ was used to compute and gauge the importance of two highly populated waters involved in a network of water-mediated ligand H-bonds through calculation of the pairwise Coulombic and van der Waals energies between the waters ($2H_2O$ = species 1) and the interactions partners (T790/M790, Q791, T854, or ligand = species 2).

For the analysis of fold resistance (FR), the experimental free energies are estimated as $\Delta\Delta G_{\text{FR}}^{\text{exptl}} \approx RT \ln(\text{FR})$ at 298.15 K using FR ratios in Table 2-1. It should be emphasized that FR is defined as the ratio of two activities (mutant/wildtype) thus it is important that both measurements be made under the same conditions which usually implies that data be obtained from the same laboratory to minimize artifacts resulting from different experimental protocols (i.e. assay conditions). Computationally, fold resistance energies ($\Delta\Delta G_{\text{FR}}^{\text{calcd}}$) are defined as the difference in predicted free energies of binding from two independent simulations (e.g. $\Delta G_{\text{L858R}} - \Delta G_{\text{WT}}$) as obtained in each case via eqs 2-1 to 2-3.

2.3 Results and Discussion.

2.3.1 Simulation Stability.

To assess the behavior of the MD simulations, and gauge the robustness of results, structural root-mean-square-deviations (rmsds) and system energies were examined as a function of time. As demonstrated in Figure 2-3 which is representative, simulations of erlotinib complexed with EGFR for wild type, L858R, and L858R&T790M show reasonable stability in

plots for the estimated free energies of binding (ΔG_b calcd) and for rmsds. Block-smoothed energies (black line) indicate good behavior when computed from the running average of the previous 100 instantaneous snapshot energies (gray points) as shown in Figure 2-3 left. On the right in Figure 2-3 are shown instantaneous rmsds values for the EGFR backbone at C α , C, N, and O (blue line), erlotinib heavy atoms (green line), and erlotinib quinazoline ring atoms (red line). Here, the reported rmsd values are obtained after each individual snapshot is fit to the first frame of the production MD runs using receptor C α backbone atoms as the match criteria. Thus, rmsd values for the ligands reflect variation in both internal geometry as well as rigid body movement relative to the protein. At some points in the trajectories, erlotinib reveals larger than expected rmsd values ($> 2 \text{ \AA}$) which could be of concern (Figure 2-3, green lines). However, examination of rmsds for only the central fused-ring quinazoline scaffold (see Table 2-1) shows much lower values (Figure 2-3, red lines) which indicates erlotinib remains anchored in the binding pocket and it is only the solvent exposed flexible r-groups extending off the quinazoline ring which fluctuate significantly. Other simulations behave similarly, with gefitinib (4-aminoquinazoline) and AEE788 (pyrrolopyrimidine) scaffolds showing less movement than the overall ligand. In all cases, rmsds results for the receptor backbone are well behaved and low ($< 2 \text{ \AA}$) which additionally indicate robust simulation behavior (Figure 2-3, blue lines).

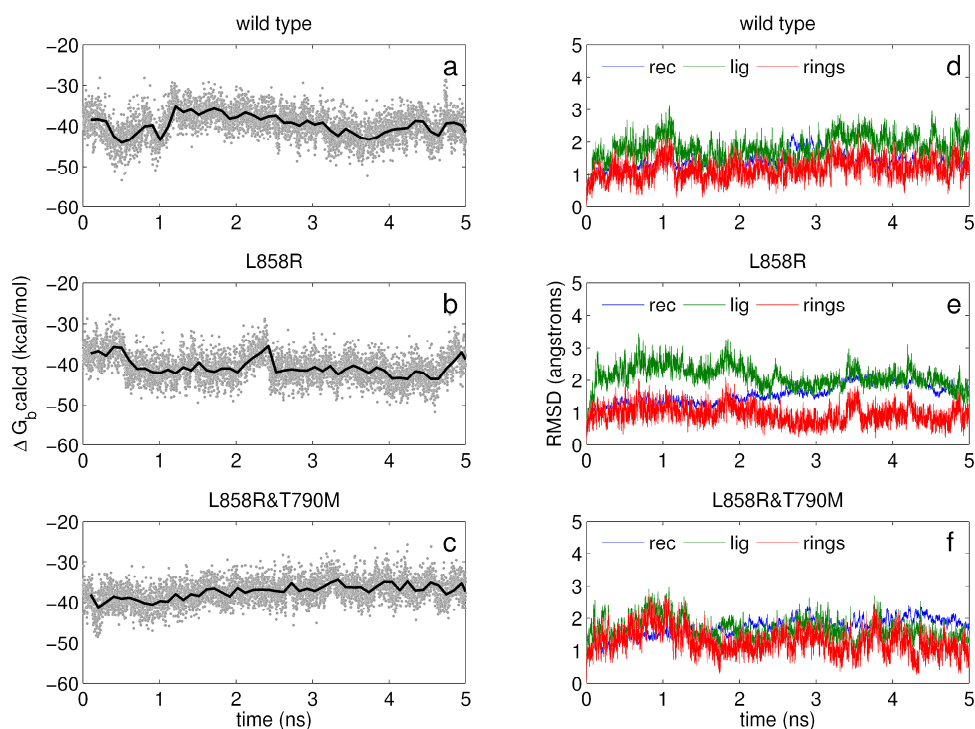


Figure 2-3. Fluctuations in computed free energies of binding (ΔG_b calcd) and root-mean-square-deviation (rmsd) for erlotinib with wild type EGFR (panels a and d), L858R (panels b and e) and L858R&T790M (panels c and f) vs time. Left panels show instantaneous energies (gray dots) and block-running averages over 100 frames (black line). Right panels show rmsds for receptor backbone (blue line), ligand (green line), and ligand quinazoline scaffold atoms (red line).

2.3.2 Comparison with Crystallographic Structures.

Only a single crystallographic structure⁵⁸ of active form EGFR (pdb code 1M17) was available at the time of our initial mutant setups for erlotinib. All subsequent simulations employed the same set of protein coordinates originally derived for this ligand. However, other EGFR structures, including those with several of the mutations studied here, have been reported.^{64,65} To structurally compare the theoretical and experimental results, as well as assess computer sampling during the simulations, evenly spaced MD snapshots were individually fit to available crystallographic structures again using C α backbone atoms as the match criteria. As shown in Figure 2-4 for three representative simulations, sampling of ligand positions is consistent with the experimentally observed conformations. And as expected, solvent exposed

regions of the ligands visually show greater movement than for scaffolds which is in agreement with the rmsd results plotted in Figure 2-3 (panels d to f). Notably, Figure 2-4 highlights how key crystallographic positions of important sidechains including G719 (or S719 mutant), T854, L858, T790 (or M790 mutant), and M793 are well-sampled during the simulations and consistent with experiment. An anomaly is a difference in the rotameric states sampled for L858R vs the crystal structure (Figure 2-4b, blue line). Here, the MD simulations sample a solvent exposed Arg conformation as opposed to an intramolecular H-bond as seen in the crystal structure. Although longer MD simulations might be required to sample the experimental L858R rotamer, since both conformations appear to be physically reasonable, an alternative would be to begin simulations using crystallographic coordinates of L858R instead of those based on the 1M17 models. However, available EGFR structures of L858R mutants show disorder in the residue range spanning 867-875. Additionally, the activation loop region in these structures (defined as 855 to 876)⁵⁸ adopts a unique conformation which is dependent on the choice of crystallographic buffer conditions.⁶⁵ Thus, given the considerable ambiguity in how to model nine missing residues (a.a. 867-875) into an unknown activation loop conformation, we have elected to retain the models originally constructed using 1M17 containing the complete loop. As described below, the good agreement between computational and experimental activities obtained using the 1M17-derived coordinates suggests these are reasonable models to study EGFR-ligand binding in the kinase active form.

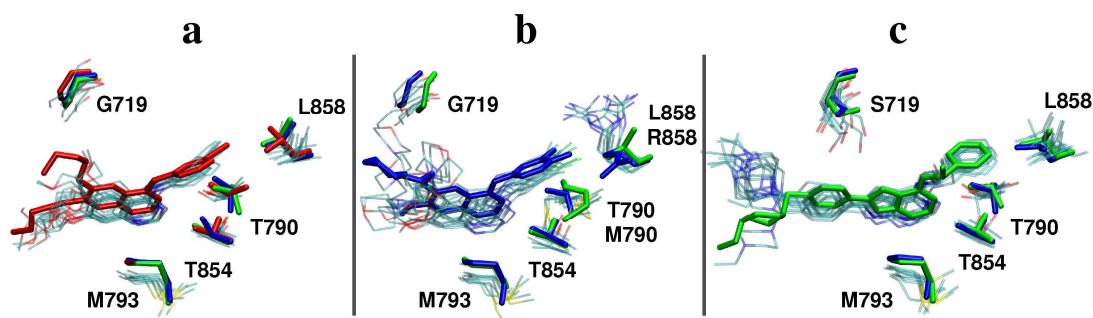


Figure 2-4. Representative snapshots from MD simulations of ligands with EGFR showing side-chain sampling of key residues (thin lines, N=10 each) vs crystallographic conformations (bold lines) for erlotinib (panel 4a), gefitinib (panel 4b), and AEE788 (panel 4c). Pdbcodes for 4a: red=1M17 (erlotinib with wildtype), blue=2ITY (gefitinib with wildtype), green=2J6M (AEE788 with wildtype). Pdbcodes for 4b: blue=2ITZ (gefitinib with L858R), green=2JIU (AEE with T790M). Pdbcodes for 4c: blue=2ITO (gefitinib with G719S), green=2ITP (AEE788 with G719S).

2.3.3 Correlation with Experimental Fold Resistance.

Overall, the computational results are strongly correlated with the experimental fold resistance values as shown in Table 2-2 and graphically plotted in Figure 2-5. Calculated values represent average quantities obtained over 5000 MD snapshots. Low standard errors of the mean (sem) indicate the energetic results are converged. Notably, the computational results correctly predict that affinity is always enhanced (negative $\Delta\Delta G_{FR}$ values) for all three ligands with the cancer causing L858R EGFR mutation relative to wildtype (Table 2-2 columns E vs F). Further, results for the drug resistant double mutant (L858R&T790M) correctly predict that decreases (positive $\Delta\Delta G_{FR}$ values) will occur in binding relative to L858R alone (Table 2-2 columns E vs F). Compellingly, the magnitudes for the energetic changes which occur across the inhibitor series in Table 2-2 are in excellent agreement with experiment. For example, results for erlotinib ($\Delta\Delta G_{FR}$ calcd = 3.30 vs $\Delta\Delta G_{FR}$ exptl > 3.96 kcal/mol) and AEE788 ($\Delta\Delta G_{FR}$ calcd = 2.40 vs $\Delta\Delta G_{FR}$ exptl = 1.68 kcal/mol) both show much larger computational and experimental FR values for the double mutant relative to gefitinib ($\Delta\Delta G_{FR}$ calcd = 0.27 vs $\Delta\Delta G_{FR}$ exptl = 0.90 kcal/mol) which is less affected. Despite the fact that the simulations correctly predict AEE788 to bind

more tightly to L858R, a minor discrepancy is the improper rank ordering for L858R–WT relative to gefitinib. In terms of sign, the sole outlier in Table 2-2 is for AEE788 for which the G719S/WT fold resistance yields essentially no energetic change experimentally but our calculations show enhanced affinity. Interestingly, a prediction for the effect of G719S on binding of erlotinib also shows enhanced affinity (Table 2-2). FR calculations for gefitinib with G719S yield the correct experimental trend. Despite the one outlier, there is excellent accord overall, and a linear fit between the data points shows a strong correlation coefficient of $r^2 = 0.84$ (Figure 2-5, Table 2-2) which indicates the simulations well reproduce trends in the experimental FR energies.

Table 2-2. Experimental versus calculated Fold Resistance (FR) energies ($\Delta\Delta G_{FR}$) and energy components for ligands with EGFR.

inhibitor	$\Delta\Delta E_{vdw}$ A	$\Delta\Delta E_{coul}$ B	$\Delta\Delta G_{polar}$ C	$\Delta\Delta G_{nonpolar}$ D	$\Delta\Delta G_{FR}$ calcd E =(A+B+C+D)	$\Delta\Delta G_{FR}$ exptl F
L858R – WT						
erlotinib	-0.86 ± 0.06	-0.34 ± 0.13	0.21 ± 0.11	0.06 ± 0.003	-0.97 ± 0.07	-0.61
gefitinib	-0.99 ± 0.06	-0.72 ± 0.07	-0.58 ± 0.06	-0.01 ± 0.004	-2.30 ± 0.07	-1.59
AEE788	-2.41 ± 0.06	-0.48 ± 0.07	0.36 ± 0.06	-0.30 ± 0.005	-2.84 ± 0.06	-0.92
L858R&T790M – L858R						
erlotinib	2.30 ± 0.06	7.42 ± 0.11	-6.56 ± 0.10	0.09 ± 0.003	3.30 ± 0.06	>3.96
gefitinib	-0.10 ± 0.05	-0.06 ± 0.07	0.49 ± 0.06	-0.06 ± 0.004	0.27 ± 0.06	0.90
AEE788	3.39 ± 0.07	3.15 ± 0.09	-4.33 ± 0.07	0.20 ± 0.004	2.40 ± 0.08	1.68
G719S – WT						
erlotinib	-2.08 ± 0.06	-0.05 ± 0.12	-0.24 ± 0.11	0.04 ± 0.003	-2.38 ± 0.07	not reported
gefitinib	0.74 ± 0.07	-0.85 ± 0.07	1.59 ± 0.07	0.04 ± 0.004	1.50 ± 0.08	0.50
AEE788	-0.65 ± 0.06	-0.78 ± 0.06	0.55 ± 0.05	0.08 ± 0.005	-0.81 ± 0.07	0.02
$r^2 =$	0.70	0.47	0.19	0.30	0.84	7 data points ^c

^a $\Delta\Delta G_{FR}$ calcd derived from the difference of two independent simulations (eg L858R – WT) computed using eqs 2-1 to 2-3. ^b $\Delta\Delta G_{FR}$ exptl values from Table 2-1. Correlations coefficients (r^2 values) obtained from fitting the change in each energy component to $\Delta\Delta G_{FR}$ exptl. All energies in kcal/mol \pm standard errors of the mean from 5000 MD snapshots. ^cData point for erlotinib with double mutant (>3.96) excluded from r^2 calculations given ambiguity in the experimental $\Delta\Delta G_{FR}$ measurement.

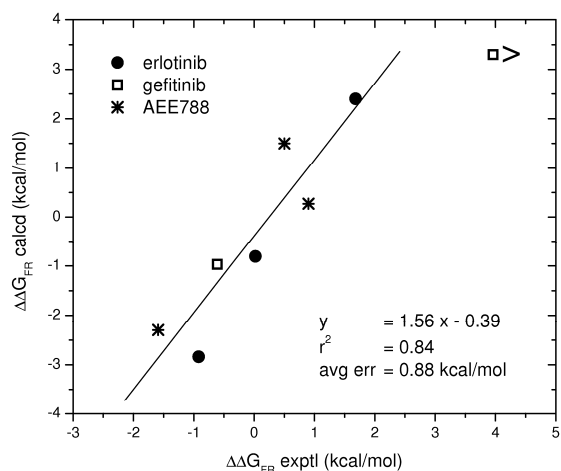


Figure 2-5. Predicted FR energies ($\Delta\Delta G_{FR} \text{ calcd}$) vs experimental FR energies ($\Delta\Delta G_{FR} \text{ exptl}$) for inhibitors with EGFR. Each point is the difference between results from two independent MD simulations (16 simulations total) from 5000 MD snapshots each. Data point for erlotinib with double mutant (>3.96) excluded from r^2 calculations given ambiguity in the experimental $\Delta\Delta G_{FR}$ measurement.

Examination of the individual terms which comprise $\Delta\Delta G_{FR} \text{ calcd}$ along with calculation of correlation coefficients (r^2 values) for each term with $\Delta\Delta G_{FR} \text{ exptl}$ was done to pinpoint which term(s) best explain experimental variation and thus resistance. It should be noted that due to ambiguities in the experimental FR measurement for erlotinib with the double mutant (>3.96 kcal/mol, Table 2-2) all fittings excluded this data point. For L858R relative to wildtype EGFR, all three inhibitors show more favorable van der Waals and Coulombic interactions which lead to an overall stronger computed $\Delta\Delta G_{FR}$ in agreement with experiment (Table 2-2 columns A and B). For the drug resistant mutant (L858R&T790M – L858R), the most dramatic losses observed experimentally correlate with the large computed losses in van der Waals and Coulombic energy for erlotinib ($\Delta\Delta G_{FR} > 3.96$, $\Delta\Delta E_{vdw} = 2.30$, $\Delta\Delta E_{coul} = 7.42$ kcal/mol) and AEE788 ($\Delta\Delta G_{FR} = 1.68$, $\Delta\Delta E_{vdw} = 3.39$, $\Delta\Delta E_{coul} = 3.15$ kcal/mol). For gefitinib with the double mutant the less deleterious effect on binding ($\Delta\Delta G_{FR} = 0.90$ kcal/mol) appears to be solely from changes in

desolvation ($\Delta\Delta G_{\text{polar}} = 0.49$ kcal/mol) given the minor changes computed in the other terms ($\Delta\Delta E_{\text{coul}} = -0.06$, $\Delta\Delta E_{\text{vdw}} = -0.10$ kcal/mol). For the G719S mutation relative to wildtype, binding losses for gefitinib again appear to be a result of increased desolvation ($\Delta\Delta G_{\text{polar}} = 1.59$ kcal/mol) as any gains computed in Coulombic energy are offset by reduction in steric packing ($\Delta\Delta E_{\text{coul}} = -0.85$ vs $\Delta\Delta E_{\text{vdw}} = 0.74$ kcal/mol). For AEE788 with G719S, the previously noted disagreement between computed and experimental affinities for this data point renders component analysis here indeterminate. Affinity for erlotinib with G719S is predicted to be enhanced primarily as a result of increased van der Waals interactions.

Notably, the most correlated term in Table 2-2 with experiment is for $\Delta\Delta G_{\text{FR}}$ calcd ($r^2 = 0.84$) indicating that for these systems a balance of energetic terms is most important for describing changes in FR. Of the individual components, changes in van der Waals energy ($\Delta\Delta E_{\text{vdw}}$ $r^2 = 0.70$) show the largest r^2 value followed by Coulombic ($\Delta\Delta E_{\text{coul}}$ $r^2 = 0.47$), nonpolar $\Delta\Delta G_{\text{nonpolar}}$ ($r^2 = 0.30$), and polar desolvation energies ($\Delta\Delta G_{\text{polar}} = 0.19$). The low r^2 value of 0.04 obtained for the sum of $\Delta\Delta E_{\text{coul}}$ and $\Delta\Delta G_{\text{polar}}$ vs experiment suggests that steric packing probably contributes more to variation in FR as opposed to changes in solvent mediated electrostatics. Interestingly, visually plotting changes in energy components vs $\Delta\Delta G_{\text{FR}}$ exptl reveals grouped data in the $\Delta\Delta E_{\text{coul}}$ plot which do not appear to lie on the trend line. A fit of this cluster alone leads to an even poorer correlation (Figure 2-6a dashed line). In contrast, Figure 2-6b shows how changes in $\Delta\Delta E_{\text{vdw}}$ are more closely associated with changes in $\Delta\Delta G_{\text{FR}}$ across the entire dataset.

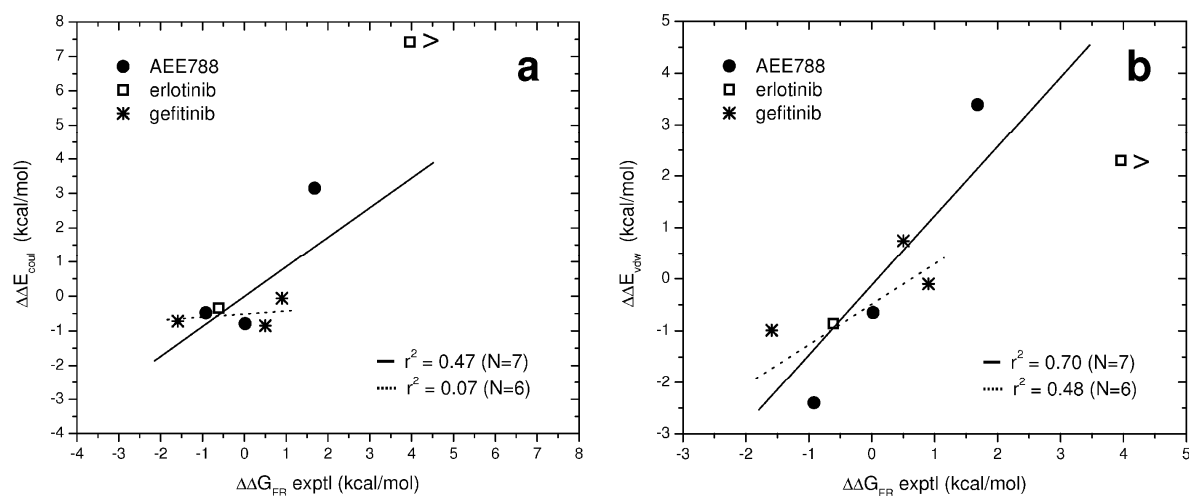


Figure 2-6. Predicted changes in Coulombic ($\Delta\Delta E_{\text{coul}}$ panel a) and van der Waals ($\Delta\Delta E_{\text{vdw}}$ panel b) energy components versus experimental fold resistance energy ($\Delta\Delta G_{\text{FR}} \text{ exptl}$) for inhibitors with EGFR. Data point for erlotinib with double mutant (>3.96) excluded from r^2 calculations given ambiguity in the experimental $\Delta\Delta G_{\text{FR}}$ measurement.

2.3.4 Energetics of Binding: What Drives Association?

To further characterize how terms contribute to molecular recognition, results from the underlying free energy of binding (ΔG_b) used to determine $\Delta\Delta G_{\text{FR}}$ were examined (Table 2-3). Overall, inhibitor binding appears to be most strongly driven by van der Waals interactions. Values for ΔE_{coul} are always less favorable than ΔE_{vdw} and not sufficient to overcome the competing unfavorable polar desolvation terms (ΔG_{polar}) which suggests steric packing dominates association. For the EGFR variants studied, gefitinib shows stronger ΔE_{vdw} interactions relative to either erlotinib or AEE788. A plot of ΔE_{vdw} vs $\Delta G_b \text{ exptl}$ highlights the separation between gefitinib and AEE788 and additionally shows how changes in van der Waals interactions may track for individual ligands (Figure 2-7). Although the combined correlation with $\Delta G_b \text{ exptl}$ is poor ($\Delta E_{\text{vdw}} r^2 = 0.004$), van der Waals energies for gefitinib ($r^2 = 0.83$) or AEE788 ($r^2 = 0.82$) when plotted separately show strong correlation with experiment (Figure 2-7).

Table 2-3. Absolute free energies and component decomposition for inhibitors with EGFR.

system	ΔE_{vdw} A	ΔE_{coul} B	ΔG_{polar} C	$\Delta G_{\text{nonpolar}}$ D	$\Delta G_{\text{b, calcd}}$ E=A+B+C+D	$\Delta G_{\text{b, exptl}}^{\text{a}}$ F	Hbond G
erlotinib							
wildtype	-49.01 ± 0.04	-24.71 ± 0.09	39.73 ± 0.08	-6.05 ± 0.002	-39.69 ± 0.05	-10.58^{b}	1.82
L858R	-49.86 ± 0.04	-25.04 ± 0.09	39.94 ± 0.07	-5.99 ± 0.002	-40.66 ± 0.05	-11.19^{b}	2.17
L858R&T790M	-47.57 ± 0.05	-17.62 ± 0.07	33.38 ± 0.06	-5.89 ± 0.002	-37.36 ± 0.04	$> -6.82^{\text{c}}$	0.99
G719S	-51.09 ± 0.04	-24.76 ± 0.08	39.49 ± 0.07	-6.01 ± 0.003	-42.07 ± 0.05	not reported	1.95
gefitinib							
wildtype	-53.50 ± 0.05	-14.02 ± 0.05	28.80 ± 0.04	-6.30 ± 0.003	-45.01 ± 0.06	-10.17^{d}	1.16
L858R	-54.49 ± 0.04	-14.74 ± 0.04	28.22 ± 0.04	-6.31 ± 0.003	-47.32 ± 0.05	-11.76^{d}	1.24
L858R&T790M	-54.59 ± 0.04	-14.80 ± 0.05	28.71 ± 0.05	-6.37 ± 0.003	-47.05 ± 0.05	-10.86^{d}	1.05
G719S	-52.76 ± 0.04	-14.87 ± 0.06	30.39 ± 0.05	-6.26 ± 0.002	-43.51 ± 0.05	-9.42^{e}	1.08
AEE788							
wildtype	-50.08 ± 0.05	-21.77 ± 0.04	31.97 ± 0.03	-5.93 ± 0.004	-45.81 ± 0.05	-11.29^{d}	2.02
L858R	-52.49 ± 0.04	-22.26 ± 0.06	32.33 ± 0.05	-6.24 ± 0.003	-48.65 ± 0.04	-12.22^{d}	2.19
L858R&T790M	-49.10 ± 0.06	-19.11 ± 0.07	28.00 ± 0.05	-6.03 ± 0.003	-46.25 ± 0.07	-10.55^{d}	2.48
G719S	-50.73 ± 0.04	-22.56 ± 0.04	32.52 ± 0.03	-5.85 ± 0.003	-46.62 ± 0.04	-10.86^{e}	1.99

^a $\Delta G_{\text{b, exptl}} \approx RT \ln(\text{activities})$ at 298.15 K in kcal/mol. ^bKi values (nM) from Carey et al.⁶² ^cIC₅₀ values (nM) from Ji et al.⁶³
^dKd values (nM) from Yun et al.⁶⁴ ^eKd values (nM) from Yun et al.⁶⁵

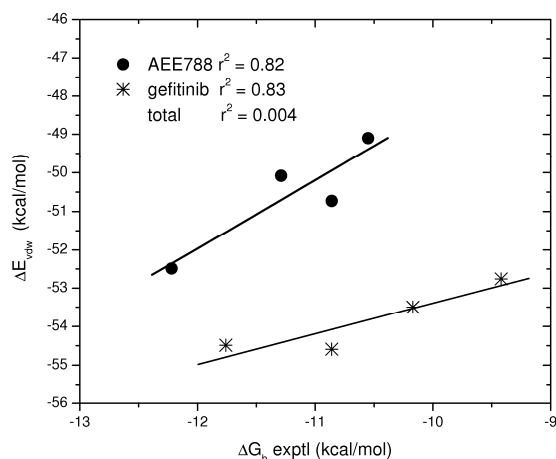


Figure 2-7. Correlation of the van der Waals energy (ΔE_{vdw}) component with ΔG_b exptl. Energies in kcal/mol.

Despite the importance of steric packing, electrostatics in this system appear to play critical roles in mediating affinity. For example, differences in intermolecular H-bonding, as illustrated graphically in Figure 2-8, likely contribute to enhanced Coulombic interactions for AEE788 and erlotinib relative to gefitinib (Table 2-3 column B). Average number of H-bonds (Table 2-3 column G) shows 2.02 interactions for AEE788 with wildtype EGFR followed by erlotinib at 1.82 and gefitinib at 1.16. All the inhibitors show highly populated and significant H-bonding with the backbone amide hydrogen at position M793. A second interaction at M793 for AEE788 largely accounts for the greater number of H-bond relative to the other inhibitors (Figure 2-8). For erlotinib, an additional significant H-bond is observed between the backbone at C797 and a terminal O atom for which the other inhibitors have no spatial equivalent (Figure 2-8). A less populated yet quantifiable interaction for erlotinib includes a unique pi-type H-bond made between the ligand's para-alkyne and T790@OH. Pi-type interactions for erlotinib were counted by simply defining the centroid of the alkyne $C \equiv C$ bond as an H-bond acceptor. Interestingly, the unique H-bond acceptor in erlotinib is replaced by a spatially analogous interaction in gefitinib between the meta-chlorine and T790@OH. AEE788 also shows a weak

H-bond at position T790 although this was only observed in the simulation of L858R. Here, a slightly different positioning of AEE788 in the binding pocket relative to the other inhibitors allows for a third H-bond with the pyrrolopyrimidine scaffold (Figure 2-8).

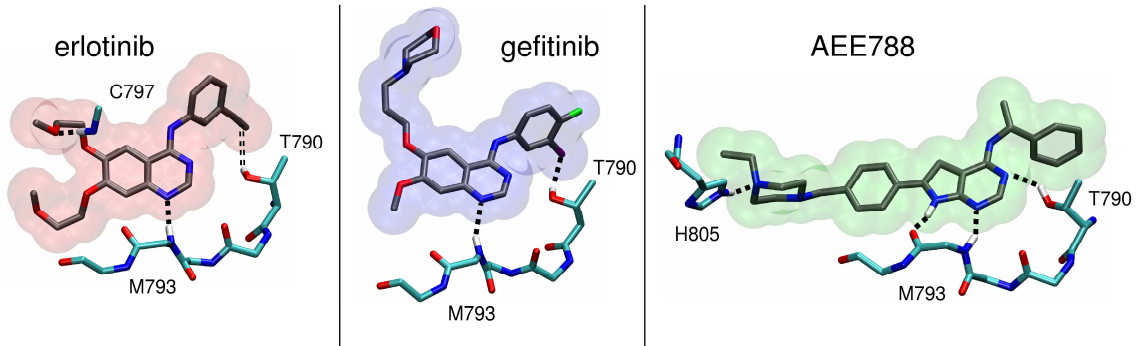


Figure 2-8. Primary H-bonding (dashed lines) interactions for inhibitors with EGFR.

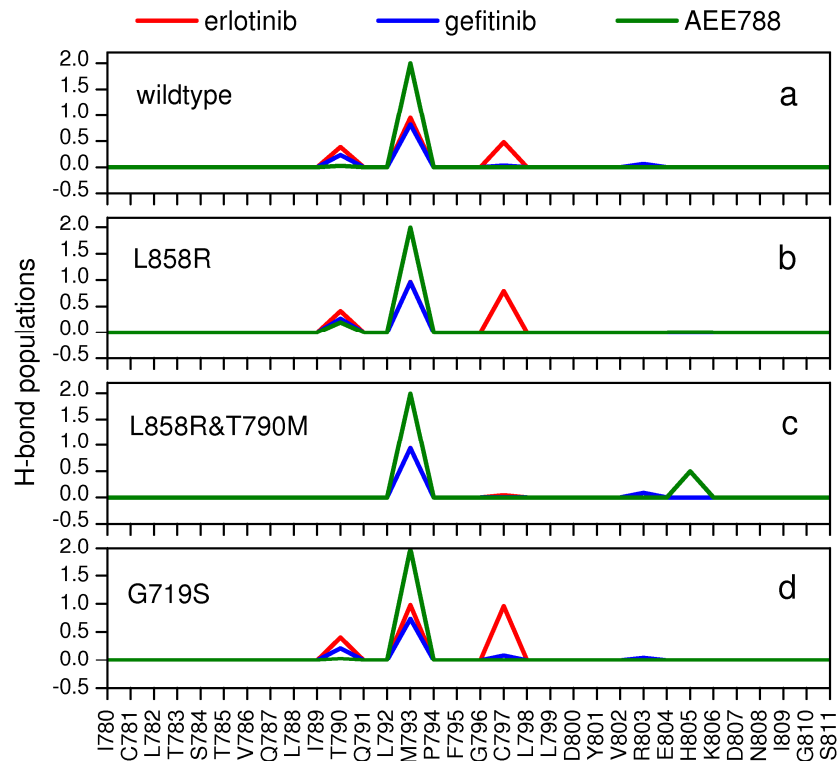


Figure 2-9. Comparison of per-residue H-bond footprints for erlotinib (red), gefitinib (blue), and AEE788 (green) with wildtype (panel a), L858R (panel b), L858R&T790M (panel c), and G719S (panel d) EGFR variants (N = 5000 MD snapshots).

2.3.5 Origins of Resistance.

In order to gauge the relative importance that specific amino acids may contribute to binding, the number of intermolecular H-bonds, van der Waals energy, and Coulombic energy were computed on a per-residue basis. Examinations of H-bond footprint plots (Figure 2-9) show consistency in overall shape from simulation to simulation which provides additional support that results obtained from averaging 5000 MD frames are converged and well-behaved. As an example, greater number of H-bonds are consistently obtained for AEE788 (2 key scaffold H-bonds) versus other inhibitors (1 key scaffold H-bond) across the various simulations (Figures 2-8 and 2-9). Although L858R&T790 does not appear to affect the number of H-bonds at this key backbone position, the resistant mutant clearly results in abolishment of weaker H-bond interactions for all inhibitors at the site of the T790 mutation relative to L858R or wildtype alone (Figure 2-9c vs 2-9a,b). In addition, for erlotinib, the more significant H-bond at position C797 is also lost as a result of the double mutant (Figure 2-9c vs 2-9b, red line). Here, the loss at C797 is the result of only a slight shift in the binding pocket, otherwise, erlotinib appears well accommodated in the double mutant (Figure 2-10). The similarity in binding obtained here between L858R&T790M vs L858R suggests a steric clash mechanism of resistance is unlikely and consistent with recent crystallographic evidence from Yun et al.⁶⁴ (see discussion below). Although no clear reason was identified, the slight increases in H-bonding computed at position C797 for erlotinib with the single mutants in Figure 2-9 relative to wildtype may contribute to both the experimental (L858R) and predicted (G719S) increase in affinity for this compound (Table 2-2).

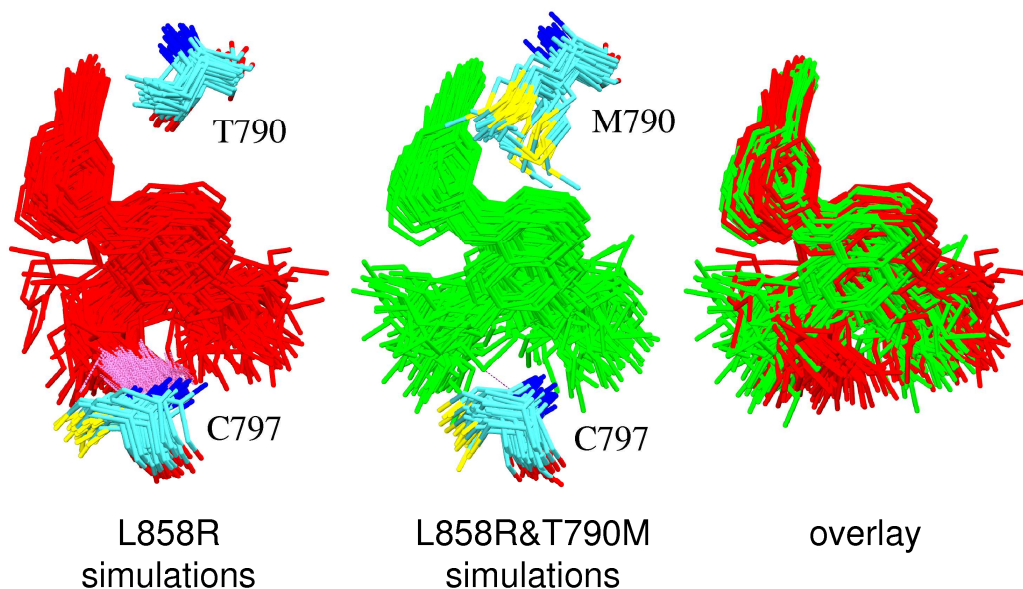


Figure 2-10. Comparison of erlotinib binding poses (N=100) from EGFR simulations for L858R (left, red), L858R&T790M (middle, green), and overlaid (red vs green). Intermolecular H-bonds at position C797 shown in purple (N=5000).

Energetic footprints representing van der Waals and Coulombic per-residue contributions were also plotted to quantify changes as a result of drug resistant EGFR relative to L858R. Focusing in on the key residues, Figure 2-11 shows the contiguous region between Q787 and N808 (shaded) and residues for which favorable interactions are computed to be ca. > 1 kcal/mol. Again, the strong similarity in the general shape of the footprints, as well as the similar magnitudes in ΔE_{vdw} at specific positions (i.e. L718, A743, K745, T790, L792, L845 (Figure 2-11a) suggest that the computational results are sensitive enough to highlight both regions with conserved interaction as well as reflect differences which may prove useful in understanding affinity.

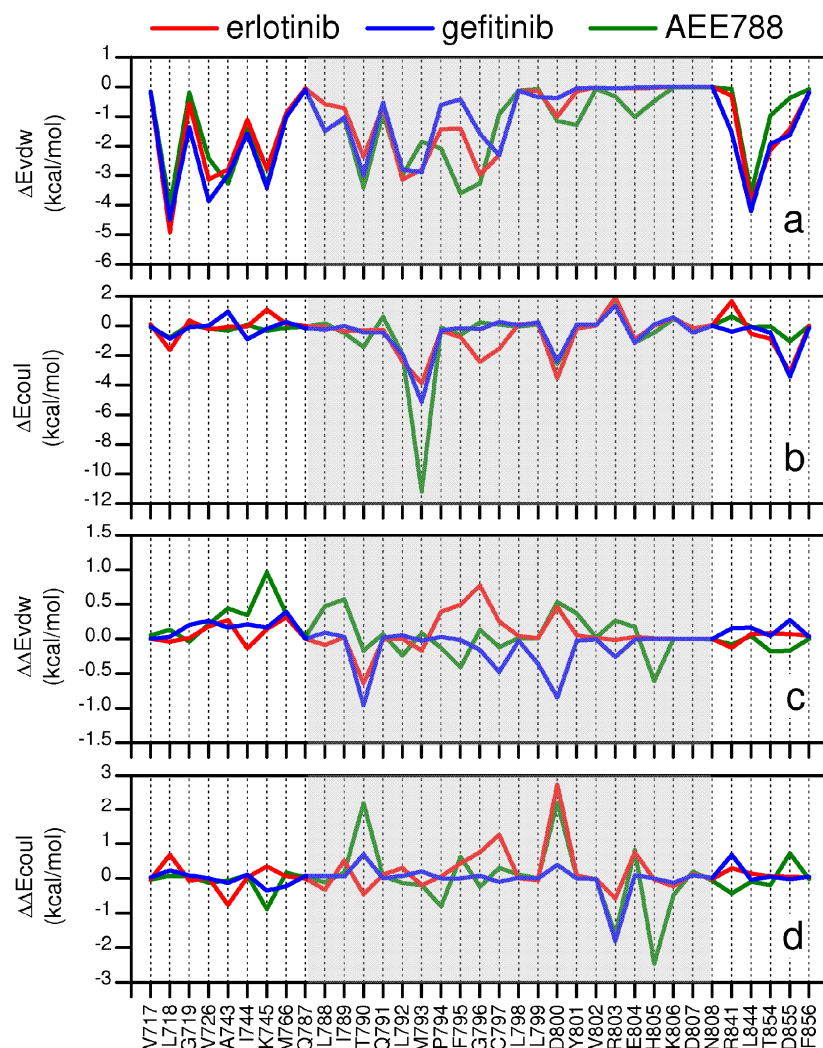


Figure 2-11. Per-residue footprints for inhibitors with EGFR for cancer causing (L858R, panels a-b) and drug resistance (L858R&T790M, panels c-d) variants from a reduced set of amino acids in the contiguous range Q787-N808 (shaded region) or for which any ligand shows $\Delta E > 1$ kcal/mol.

Consistent with the H-bond patterns described in Figures 2-8 and 2-9, in which ligands show high population of H-bonds between M793 and the central scaffolds, the most favorable ΔE_{coul} interactions for all ligands occur with residue M793 (Figure 2-11b). As before, the strongest interactions are computed for AEE788 (-11.5 kcal/mol, green line) versus gefitinib (-5.1 kcal/mol, blue line) or erlotinib (-3.9 kcal/mol, red line) which mirrors the fact that AEE788's scaffold makes two H-bonds versus one for the other inhibitors (Figure 2-8). Less

populated, but "standard" H-bonds between T790 and AEE788, and C797 and erlotinib are also visible in the ΔE_{coul} footprints but as expected are weaker than those with M793 (ca. -1 to -1.5 kcal/mol). The more unique erlotinib (pi-type), or gefitinib (chlorine-type) interactions with T790 depicted in Figure 8 are not readily apparent in the ΔE_{coul} footprints but instead are presumably reflected in the favorable ΔE_{vdw} energies which occur at this position (Figure 2-11a vs 2-11b).

Examination of difference footprints ($\Delta\Delta E_{\text{vdw}}$ and $\Delta\Delta E_{\text{coul}}$) computed from the L858R&T790M – L858R breakdowns show that erlotinib and AEE788 lose significant interactions, on a residue-by-residue basis, as a result of the deleterious mutation relative to L858R (Figure 2-11c-d). In contrast, and in agreement with the fact that gefitinib is experimentally the least affected by the resistance mutation, the $\Delta\Delta E_{\text{coul}}$ footprint is flatter, shows no overall reduction in total Coulombic energy (Table 2-2), and changes on a per-residue basis show negligible losses at all positions (Figure 2-11d, blue line). A prior study from our laboratory of neuraminidase inhibitors also revealed that the most robust compound had an overall flatter $\Delta\Delta E_{\text{coul}}$ and ΔH -bond profile.⁸³ The most significant $\Delta\Delta E_{\text{coul}}$ energy losses (ca $> +1$ kcal/mol) occur for erlotinib (Figure 2-11d, red line) at positions C797 and D800, and for AEE788 at positions T790M and D800 (Figure 2-11d, green line). Losses in $\Delta\Delta E_{\text{coul}}$ for erlotinib at position C797 are expected to be a result of the previously described H-bond disruption (Figures 2-8 to 2-10). For AEE788, the significant loss in $\Delta\Delta E_{\text{coul}}$ at T790M is likely due in part to disruption of the third H-bond with the pyrrolopyrimidine scaffold as a result of the resistance mutation (Figures 2-8 and 2-9). No straightforward structural explanation was found to explain reductions in $\Delta\Delta E_{\text{coul}}$ at position D800 for erlotinib or AEE788 but as a charged residue this could be a long-range and nonspecific effect. For AEE788 with the double mutant,

increases in $\Delta\Delta E_{\text{coul}}$ at position H805 (Figure 2-11c) are traceably to the formation of a new piperazine ring H-bond (Figure 2-8, 2-9c). However, despite the fact that gains in interaction energy occur at this position, changes overall in $\Delta\Delta E_{\text{coul}}$ and $\Delta\Delta E_{\text{vdw}}$ for AEE788 are still unfavorable (Table 2-2).

Prior studies by Daub et al.⁹⁸ and Kobayashi et al.⁹⁹ hypothesized that a steric clash was the likely mechanism of drug resistance for T790M. And Liu et al.⁸⁰ reported MD simulation results of gefitinib with either T790M or L858R&T790M which led to "ligand escape from the binding pocket" which could also be consistent with a steric clash. However, a recent study reported by Yun et al.⁶⁴ suggests this is not a likely mechanism as a co-crystal structure of AEE788 with a T790M single mutant shows essentially the same binding pose as wildtype. Results from the present study similarly suggest that a threonine / methionine swap in the double mutant will not result in a steric clash given that wildtype, L858R, and L858R&T790M simulations show an overall consistent binding pose (Figure 2-10). In addition, an examination of the van der Waals difference footprints ($\Delta\Delta E_{\text{vdw}}$) shows that for all inhibitors a methionine at position 790 is energetically accommodated in the pocket and steric packing interactions localized to this position in fact become more favorable as a result of the double mutation (Figure 2-11c). Increased packing as a result of T790M is physically reasonable and occurs as a result of the hydrophilic to hydrophobic substitution. Although other van der Waals changes are less readily explained, the H805 increase with AEE788 coincides with the previously noted piperazine H-bond. Compellingly, erlotinib (positions F795, G796, D800) and AEE788 (positions K745, I789, D800) show significant losses in $\Delta\Delta E_{\text{vdw}}$ in contrast to gefitinib, which likely contributes to these compounds being more affected by the double mutations (Figure 2-11c).

2.3.6 Water Mediated Interactions.

Examination of the underlying explicit solvent TIP3P-MD trajectories, used subsequently for continuum-based free energy calculations, revealed water molecules which appear to be important for positioning of ligands in the binding pocket. High water occupancy is observed at two primary positions, termed site 1 (S1) and site 2 (S2), as shown in Figure 2-12a for erlotinib with L858R and the double mutant, which are representative. Figure 2-12b quantifies S1 and S2 populations for all six inhibitor simulations with averages = total count/5000 frames. Site waters were defined as present if a water hydrogen was within 2.5 Å of each ligand's relevant nitrogen acceptor (S1) or residue Q791 at O (S2). Importantly, the MD simulations reproduce the crystallographically observed water at S1 for all ligands.^{58,65} The water at both sites are observed in the crystal structure of AEE788 with EGFR (2J6M).⁶⁵ For all ligands with L858R (Figure 2-12b left), waters are present 50-90% at S1 and >80% at S2 which indicates these are long lived significant interactions. As shown in Figure 2-12a, these waters are involved in a quadrifurcated H-bonding network involving the ligands with three nearby residues (T790, Q791, and T854), including the site of the known drug resistance mutation T790M. Notably, in all cases, occupancy at S1 and S2 is reduced as a result of L858R&T790 (Figure 2-12b right).

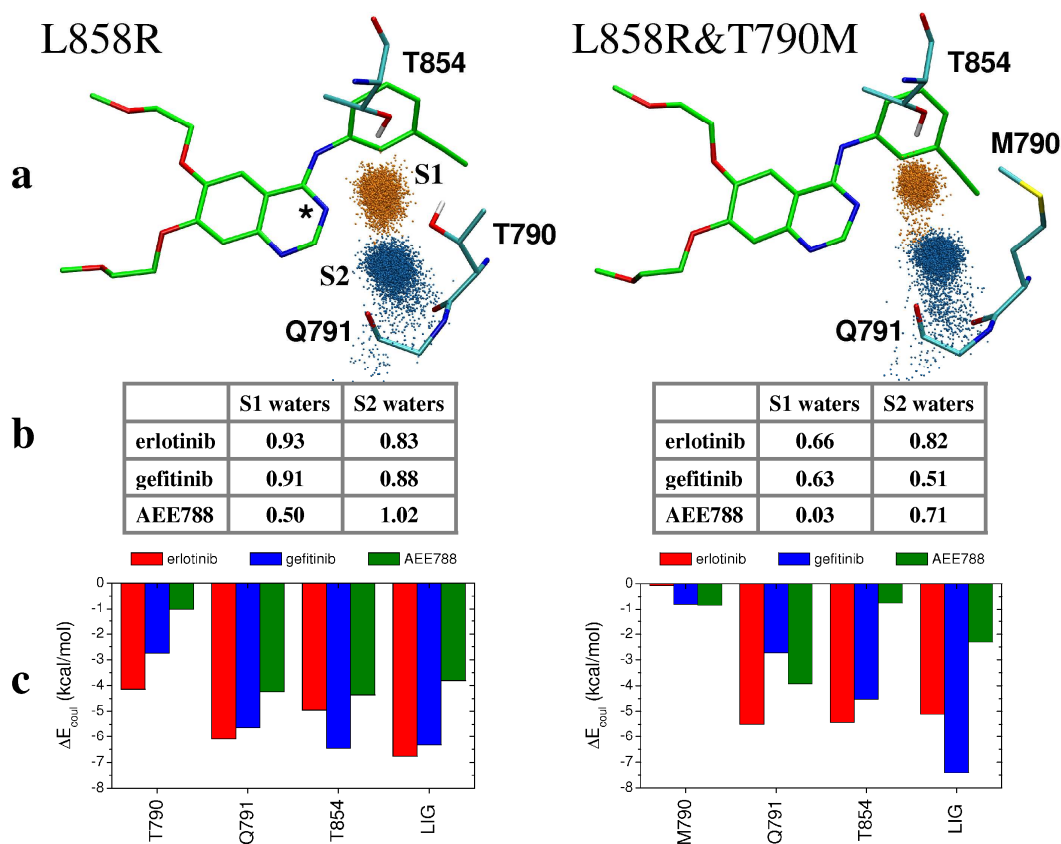


Figure 2-12. Water-mediate H-bonds for inhibitors with EGFR for L858R (left) and L858R&T790M (right). 12a visually shows population of waters at site 1 (S1 orange) and site 2 (S2 blue) over all 5000 simulation frames for representative erlotinib simulations. Site waters defined if water hydrogens are within 2.5 angstroms of each ligand at N* (S1) or residue Q791 at O (S2). 12b shows for all three inhibitors the average number (count/5000) of waters at S1 and S2. 12c shows for all three inhibitors the average pairwise Coulombic interaction energies between the two waters closest to each ligand at N* with residues T790 (or M790), Q791, T854, and the ligands.

As an alternative metric, energy calculations reveal favorable Coulombic interactions between pocket waters and amino acids in the H-bond network including the ligands (Figure 2-12c left). Here, the two waters closest to each ligand at N* (Figure 2-12a left) were used define key pocket waters. Interestingly the L858R&T790M mutant leads to changes in bridging water interactions with each ligand that roughly mirror trends in the experimental FR data with erlotinib ($\Delta\Delta E_{\text{coul}} = +1.7$ kcal/mol) and AEE788 ($\Delta\Delta E_{\text{coul}} = +1.5$ kcal/mol) both being adversely affected compared to gefitinib ($\Delta\Delta E_{\text{coul}} = -1.1$ kcal/mol). Favorable electrostatic interactions

between these waters and residue 790 (Figure 2-12c right) are similarly reduced as a result of the double mutant, particularly for erlotinib (red bar), and thus expected to lead to weaker protein-ligand binding. Further, despite the fact that some water-mediated H-bonding with M790 is observed, an overall weaker network would be expected due to the fact that sulfur is a weaker H-bond acceptor than oxygen.¹⁰⁰ Overall, the energetic description (Figure 2-12c) is consistent with the reduced population counts (Figure 2-12b) suggesting weaker interactions in the drug resistant mutant.

H-bonding between quinazoline-based inhibitors and binding site waters were previously predicted by Wissner et al.⁷⁷ and Hou et al.⁷⁸ although interestingly the two studies came to different conclusions as to whether residue T854 or T790 was involved. Here, calculations indicate that both T854 and T790 residues make significant water-mediated ligand interactions. Stamos et al.⁵⁸ noted the T790 bridging water in the erlotinib-EGFR crystal structure, but suggested it was not significant citing data reported by Rewcastle et al.¹⁰¹ in which only a minor effect on affinity was seen for related ligands where the H-bond acceptor was substituted for carbon. However, examination of the original activities (see Rewcastle et al., Table 2-1, compounds 15 vs 20) show > 5000 fold loss between compounds that differ only by a nitrogen at the T790 acceptor position which suggests the water is in fact important.¹⁰¹ And, a recent docking study by Cavasotto et al.⁷⁹ notes that inclusion of this bridging water was necessary to correctly reproduce the binding pose of the EGFR inhibitor AG1478.

In conjunction with their proposed steric clash mechanism, Kobayashi et al.⁹⁹ also hypothesized that disruption of water-mediated binding would be a factor in resistance. For the water-mediated interactions at Q791 (Figure 2-12), the H-bonds primarily involve the backbone carbonyl oxygen thus any alteration of sidechains at this site would be expected to be less

detrimental, particularly since there is little direct van der Waals contact or favorable Coulombic interactions with the ligand at Q791. However, the simulation results strongly suggest that a mutation at position T854 would disrupt the quadrifurcated network and, in a manner analogous to T790, disrupt water-mediated ligand binding. This hypothesis is consistent with results recently reported by Bean et al.¹⁰² in which a novel T854A resistance mutation was identified from a patient with reduced affinity for erlotinib. A combination of mutations involving T790 and T854, if biologically viable, would likely lead to further disruption of the H-bond network involving inhibitors and an increase in unfavorable fold resistance. While our current studies cannot rule out the recent hypothesis by Yun et al.⁶⁴ that T790M resistance is caused primarily by increased affinity for ATP, based on the present simulations, it is reasonable to propose that disruption of water-mediated H-bond networks involving the inhibitors (Figure 2-12) is a contributing factor. Additionally, given the fact that our calculations yield quantitative energetic agreement with experiment, yet involve only inhibitors and EGFR (and not ATP), strongly suggests that differences in affinity for ATP are not the sole cause of experimentally observed drug resistance. Additional studies are needed to more fully address this issue.

An examination of the network shown in Figure 2-12 indicates the possibility of designing alternative H-bonding involving residues T854, T790, and Q791. EGFR inhibitors based on a 4,6-dianilinopyrimidine scaffold have been reported¹⁰³ which are proposed to make direct H-bonds with both M793 and T790. However, as expected, the T790M mutant showed resistance against a representative compound in the series presumably due to the loss of a direct H-bond between the pyrimidine N3 and the OH at position 790. Wissner et al.⁷⁷ has reported an inhibitor in which the bridging nitrogen atom was replaced by a cyano group and proposed to displace the site 1 water. However, the cyano compound was also proposed to make a direct H-

bond with T854 thus the recently reported T854A mutation would likely lead to a loss in binding. An alternative strategy to address resistance, provided that sufficient specificity could be achieved, would be the design of inhibitors with additional protein backbone H-bonds (direct or water mediated). In any event, due to the entropically favorable process of displacing bound waters,¹⁰⁴ analogs which replace the water-mediated interactions seen here may show enhanced affinity. Alternative binding patterns are likely to result in unique resistance profiles which may prove useful.

2.4 Conclusions.

In this study, all-atom explicit solvent molecular dynamics followed by free energy calculations were employed to compute fold resistance energies for three ATP-competitive inhibitors (erlotinib, gefitinib, and AEE788) with epidermal growth factor receptor (EGFR) for wildtype, and L858R, G719S, and L858R&T790M mutants. The primary purpose of this study was development of robust quantitative computational models to compute EGFR-ligand binding, characterize how variation in structural and energetic results correlate with variation in reported experimental activities, and determine origins of drug resistance. System stability and overall convergence of results was carefully monitored through comparisons with crystallographic structures (Figure 2-4), and by plotting instantaneous and running block averages for free energies of binding and root-mean-square deviations (Figure 2-3). Fluctuations in energy and structure show the simulations are well-behaved, comparable with other studies from our laboratory,⁸²⁻⁸⁴ and low standard errors of the mean (Tables 2-2 and 2-3) indicate the results are reasonably converged.

Notably, computed fold resistance energies, which represent a ratio of activities and are obtained from the difference in results of two independent MD simulations, show excellent agreement with available experimental data ($r^2 = 0.84$). Importantly, the magnitudes of the experimental and theoretical FR results are similar (Table 2-2, Figure 2-5). For all inhibitors, the simulations correctly predict that affinity for EGFR will increase as a result of the cancer causing L858R mutation relative to wildtype and decrease as a result of a drug resistant double mutant (L858R&T790M) relative to L858R (Table 2-2 columns E vs F). Affinity predictions for gefitinib with a second cancer causing mutation at position G719S also yield the correct experimental trend. The sole outlier in the study is for AEE788 with G719S in which the computational results incorrectly predict the mutation to be slightly favorable (Table 2-2).

Decomposition of the contributing components to $\Delta\Delta G_{FR}$, and the underlying absolute ΔG_b values used to compute FR, reveal modest gains in favorable van der Waals and Coulombic energies for all three inhibitors as a result of the cancer causing mutation L858R and large losses for erlotinib and AEE788 for the drug resistance double mutant L858R&T790M (Table 2-2). Losses for gefitinib appear to be a result of increased desolvation penalties. Values for ΔE_{vdw} are computed to be more favorable than other ΔG_b calcd terms (Table 2-3), which suggest that steric packing is the dominant driving force for association. In general, van der Waals interactions are stronger for gefitinib relative to other compounds and changes in ΔE_{vdw} track especially well with ΔG_b exptl for gefitinib and AEE788 (Figure 2-7). Relative $\Delta\Delta E_{vdw}$ energies are modestly correlated with $\Delta\Delta G_{FR}$ across the series (Figure 2-6b).

Despite the obvious importance of van der Waals terms, the simulation results indicate that electrostatic interactions are critical for specificity and correct positioning of ligands in the ATP binding pocket (Figures 2-8, 2-9, and 2-11). Examination of per-residue H-bonding and

Coulombic energy reveal changes at key amino acids which are important for understanding origins of fold resistance (Figures 2-9 to 2-11). Two highly populated H-bonds for AEE788, and one for erlotinib and gefitinib, are observed between inhibitors and the EGFR backbone at position M793 (Figures 2-8 and 2-9). Coulombic energy footprints (Figures 2-11) mirror the H-bond trends with M793 showing stronger interaction energies formed with AEE788 (-11.5 kcal/mol) versus gefitinib (-5.1 kcal/mol) or erlotinib (-3.9 kcal/mol). The resistance mutation L858R&T790 does not change interactions localized to M793, however, all inhibitors lose a less populated H-bond at the site of the T790 mutation (Figure 2-9 panels b vs c). Losses at 790 are traced to unique H-bonds (Figures 2-8 and 2-9) involving the acetylene group of erlotinib (pi-type), and a chlorine atom in the case of gefitinib (chlorine-type). For erlotinib, the loss of an additional H-bond at position C797 leads to an overall reduction (-1.18 H-bonds) which likely contributes to the larger FR energy compared with the other inhibitors (gefininb -0.19 H-bonds, AEE688 +0.29 H-bonds).

The simulations additionally reveal a significant network of water-mediated H-bonds involving a spatially equivalent nitrogen atom on each inhibitor, residues T854, T790, Q791, and two bridging waters which become disrupted as a result of the L858R&T790M drug resistance mutation (Figure 2-12a,b). The bridging waters interact favorably with residues in the binding pocket and the double mutation leads to reduced Coulombic energies, especially for erlotinib and gefitinib (Figure 2-12c), and reduced overall occupancy (Figure 2-12b). The calculations suggest that resistance likely involves changes in water-mediated H-bonds, in contrast to prior reports, which hypothesize that EGFR resistance is primarily a function of either a steric clash^{98,99} involving methionine 790 or due to increased affinity for the native substrate ATP.⁶⁴ In agreement with recent crystallographic evidence,⁶⁴ per-residue footprint calculations (Figure 2-

11c) and structural analysis (Figure 2-10) reveal favorable van der Waals energies with T790M which indicates a steric clash mechanism of resistance is unlikely. Finally, while increased affinity for ATP⁶⁴ may be a contributor to resistance, the present results suggest that disruption of favorable interactions, including changes in H-bonding, are likely to be as important and thus should be considered when designing next-generation compounds.

The growing problem of drug resistance, arising from clinical use of EGFR molecular targeted therapeutics, highlights the need for continued studies to elucidate how binding affinity is modulated by mutations and how ligands could be modified to circumvent deleterious changes. The present study has participated in these aims through calculation of fold resistance energies for inhibitors of EGFR which show quantitative agreement with experiment thereby providing a framework to probe origins of resistance. The simulations correctly predict the effects of the L858R cancer causing mutation and the L858R&T790 drug resistance mutant for three inhibitors. Residue-based structural and energetic analysis was used to identify how key sidechains are involved in binding, how water molecules mediate affinity through an intricate network of H-bonding, and how interactions change as a result of the mutations. Similar to the growing arsenal of antivirals required to effectively combat HIV, design of multiple kinase inhibitors is likely to be an important long term strategy to address issues of drug resistance.

Chapter 3. Implementation and Evaluation of a Docking-Rescoring Method using Molecular Footprint Comparisons.

This chapter has been published as **Balius, T. E.**; Mukherjee, S.; Rizzo, R. C. Implementation and Evaluation of a Docking-Rescoring Method using Molecular Footprint Comparisons, *J. Comput. Chem.*, **2011**, 32 (10), 2273–2289. Copyright © 2011, Wiley Periodicals, Inc. [doi: 10.1002/jcc.21814](https://doi.org/10.1002/jcc.21814) PMID: 21541962

Author contributions. TEB, SM, and RCR designed research; TEB contributed new computational tool, analyzed data and wrote initial draft; TEB and RCR wrote the paper with assistance from SM.

Abstract.

A docking-rescoring method, based on per-residue van der Waals (VDW), electrostatic (ES), or hydrogen bond (HB) energies has been developed to aid discovery of ligands that have interaction signatures with a target (footprints) similar to that of a reference. Biologically useful references could include known drugs, inhibitors, substrates, transition states, or side-chains that mediate protein-protein interactions. Termed footprint similarity (FPS) score, the method, as

implemented in the program DOCK, was validated and characterized using: (1) pose identification, (2) crossdocking, (3) enrichment, and (4) virtual screening. Improvements in pose identification (6-12%) were obtained using footprint-based (FPS_{VDW+ES}) vs standard DOCK (DCE_{VDW+ES}) scoring as evaluated on three large datasets (680-775 systems) from the SB2010 database. Enhanced pose identification was also observed using FPS (45.4% or 70.9%) compared with DCE (17.8%) methods to rank challenging crossdocking ensembles from carbonic anhydrase. Enrichment tests, for three representative systems, revealed FPS_{VDW+ES} scoring yields significant early fold enrichment in the top 10% of ranked databases. For EGFR, top FPS poses are nicely accommodated in the molecular envelope defined by the reference in comparison with DCE which yields distinct molecular weight bias towards larger molecules. Results from a representative virtual screen of ca. 1 million compounds additionally illustrate how ligands with footprints similar to a known inhibitor can readily be identified from within large commercially available databases. By providing an alternative way to rank ligand poses in a simple yet directed manner we anticipate that FPS scoring will be a useful tool for docking and structure-based design.

3.1 Introduction.

A primary role of a docking program is as a virtual screening tool to help identify biologically active compounds.²⁻⁴ Binding geometries (termed poses) are predicted for candidate ligands with a target and metrics such as intermolecular interaction energy are used to identify (via rank-ordering) the best scoring molecules. Thus, docking programs can be thought of as filters, through which large databases (on the order of millions) may be passed, to isolate property-enriched subsets for further evaluation.^{105,106} To evaluate the accuracy of programs and

protocols,¹⁰⁷ two main experiments termed pose identification¹³ and database enrichment¹⁰⁸ are used. To assess pose identification accuracy, crystallographic ligand-receptor complexes are used as controls to determine if the docking program can reproduce the correct ligand geometry (typically ≤ 2 Å rmsd) and whether that pose is ranked best.¹³ To assess database enrichment (Appendix B), a group of active ligands (affinity to the target has been confirmed) is seeded into a large group of decoy molecules (no affinity to the target is presumed) to determine if the rank-ordered list of molecules (active and inactives) will contain, with high probability, the known binders among the more favorably scored list elements.¹⁰⁸

Programs such as DOCK,^{35,37} often use a physics-based energy function consisting of electrostatic (Coulombic) and steric (van der Waals) terms with the total sum of pairwise intermolecular interactions being used as the basis for rank-ordering. Alternatively, rank-ordering methods could employ known binding determinates (i.e. pharmacophores) to help identify compounds that interact with a target in a specific way which is not solely based on an energetic sum. This study explores the utility of using residue-based decompositions of electrostatic, steric, and hydrogen bonding interactions to derive 2-D pharmacophores (termed here as molecular footprints) as shown schematically in Figure 3-1. In general, a footprint may be thought of as a unique interaction signature between any two species. Further, as the sum of the residue-based contributions is equal to the overall total interaction energy, the breakdown enables identification of the amino acids which are likely to be most important for molecular recognition.

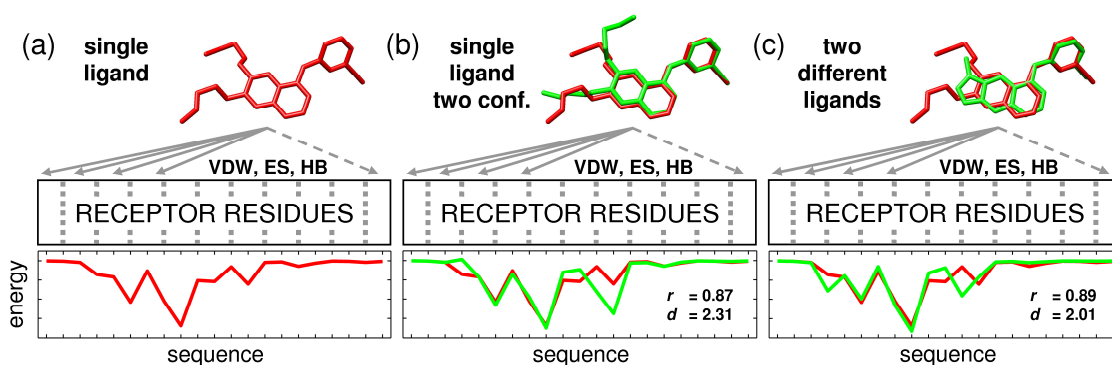


Figure 3-1. Representative molecular footprints for (a) a single ligand, (b) a single ligand with two conformations, and (c) two different ligands derived from per-residue decomposition of the intermolecular van der Waals interactions as a function of primary sequence. For two footprints, similarity may be quantified using Pearson correlation coefficient (r), Euclidean distance (d), or related measures. For clarity only a portion of the footprints are shown.

Footprints consist of a string of residue numbers, each with an associated intensity (Figure 3-1a), thus the correspondence between any two strings can easily be quantified (Figure 3-1b-c) using familiar metrics such as Pearson correlation or Euclidean distance. Comparisons can be between two conformations of the same molecule (Figure 3-1b) or between two different ligands (Figure 3-1c). Termed here footprint similarity (FPS) score, several potentially useful applications for virtual screening are envisioned with the general focus being identification of small organic molecules that score highly in comparison to a known reference compound. The footprint comparison shown in Figure 3-1c is between the FDA approved drug erlotinib (red pose) and an experimental kinase inhibitor (green pose). Table 3-1 lists possible sources of reference footprints including those derived from a known drug or inhibitor, a native substrate or cofactor, a transition state, or key side chains which mediate protein-protein interactions. Footprints can be manually modified, to decrease the importance a given side-chain prone to mutate may have in molecular recognition, which could assist identification of compounds with enhanced resistance profiles. Finally, use of ensemble or average-weighted footprints, derived

from multiple crystal structures or molecular dynamics/Monte Carlo simulations, could be used to account for receptor flexibility.

Table 3-1. Examples of possible reference types to derive molecular footprints.

Reference Types	Description
Known inhibitor	FDA-approved drug or experimental inhibitor validated to bind
Natural substrate	Native peptide or cofactor
Transition state	Predicted transition state geometry for a chemical reaction
Modified structure	Key functionality/substructure (side-chain mediating protein-protein interactions)
Text file footprint	Modified entries to increase/decrease importance of select residues (resistance mutations)
Ensemble-weighted	Averaged footprints derived from MD/MC simulations

Our laboratory^{42,44,82,83} and others¹⁰⁹⁻¹¹² have successfully used footprint-like methods in the context of molecular dynamics and Monte Carlo simulations to help identify key residues involved in molecular recognition (also discussed in Chapters 2 and Appendix A). Related approaches have also been reported for use in docking with the distinction that these have typically employed binary bit-string representations (termed interaction fingerprints)¹¹³⁻¹²³ instead of energy-based decompositions as used here. Specifically, Deng et al.¹¹³ introduced the SIFt method which employs a Tanimoto metric to compute the similarity between two bit strings derived from the presence or absence of seven interaction types occurring at a given residue. The SIFt method and various extensions^{115,116,121} have been shown to be useful for identifying native ligand poses, protein-family clustering, database enrichment, and library design. Other bit-string related procedures have also been reported.^{114,117-120,122} For example, Pfeffer et al.¹¹⁹ has recently reported a method based on a per-atom partitioning of the scoring function DrugScore^{CSD}, which was shown to yield improved results for pose identification and enrichment compared with several other methods tested.

A long term goal of our laboratory is the development of method and protocols to increase the accuracy of docking methods used in virtual screening. The primary objectives of this work are to: (i) introduce and test methods to compute footprint similarity (FPS) scores as implemented into the program DOCK,³⁷ (ii) evaluate *pose identification* accuracy using the recently reported SB2010¹³ database developed in our laboratory, (iii) and characterize *database enrichment* properties using representative systems from the DUD¹⁰⁸ database. It should be emphasized that this is a post-processing technique for molecules already docked and is simply an alternative method which facilitates re-ranking by footprint similarity. It is also important to note that FPS scoring makes use of the underlying physics-based energy function in DOCK and involves no additional parameterization beyond that used in any standard molecular mechanics force field.

3.2 Theoretical Methods.

3.2.1 Footprint Comparisons.

Footprint similarity (FPS) scores in this work are built from three scoring descriptors: van der Waals energies (VDW), Coulombic energies scaled by a distance dependent dielectric constant (ES), and hydrogen bond energies (HB). Consensus scores based on two ($FPS_{VDW+ES} = FPS_{VDW} + FPS_{ES}$) or three ($FPS_{VDW+ES+HB} = FPS_{VDW} + FPS_{ES} + FPS_{HB}$) terms were also evaluated. The general schematic for computing FPS scores is shown in Figure 3-2. The procedure involves setting up the system for DOCK calculations, preparation of a reference molecule, and generation of candidate poses (see Computational Details section). It is important to note footprints are decompositions in Cartesian space, thus Cartesian energy minimizations are recommended for both the crystallographic reference and candidate poses. A footprint is

defined as a vector $\vec{x} = [x_1, \dots, x_N]$ where N is the number of residues in the receptor and x_i is the interaction energy between the i^{th} residue and the ligand. To quantify the likeness between two footprint vectors \vec{x} and \vec{y} , four different methods for computing similarity were evaluated: standard Euclidean Distance, normalized Euclidean Distance, standard Pearson Correlation, and threshold Pearson Correlation. It should be emphasized that the different comparison methods and combinations produce FPS scores with different ranges as summarized in Table 3-2 and described further below.

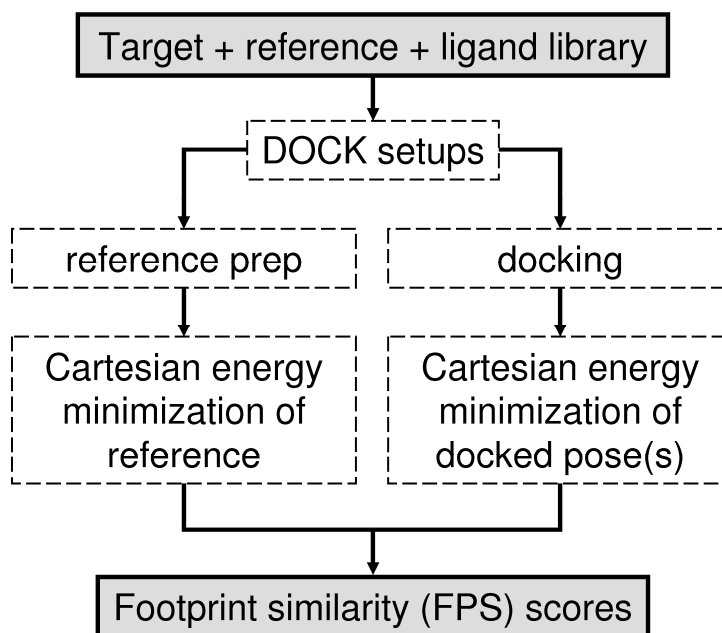


Figure 3-2. Flow chart outlining footprint similarity calculation protocol.

Table 3-2. Comparison methods and corresponding ranges for footprint similarity (FPS) scores.

Comparison Method	Ranges ^a		
	FPS _{VDW} , FPS _{ES} , FPS _{HB}	FPS _{VDW+ES}	FPS _{VDW+ES+HB}
Standard Euclidean (d)	[<u>0</u> , ∞)	[<u>0</u> , ∞)	[<u>0</u> , ∞)
Normalized Euclidean (d_{norm})	[<u>0</u> , 2]	[<u>0</u> , 4]	[<u>0</u> , 6]
Standard Pearson (r)	[-1, <u>1</u>]	[-2, <u>2</u>]	[-3, <u>3</u>]
Threshold Pearson (r_{thresh})	[-1, <u>1</u>]	[-2, <u>2</u>]	[-3, <u>3</u>]

^aThe most favorable score possible for each method is underlined.

Standard Euclidean distance (d) makes use of the distance formula ($d = \sqrt{\sum (x_i - y_i)^2}$) to quantify differences between two footprint vectors. The metric compares interaction signatures in terms of the absolute magnitudes occurring at each residue position. Alternatively, a normalized Euclidean distance (d_{norm}) may be computed by using normalized footprint vectors ($\vec{X} = \vec{x}/\|\vec{x}\|$, $\vec{Y} = \vec{y}/\|\vec{y}\|$). Here, the normalization procedure yields unit footprint vectors resulting in relative, as opposed to absolute, magnitudes being compared. Thus, normalized Euclidean distance may be thought of as a general measure of shape overlap. As illustrated in Figure 3-3, for a single type of footprint (VDW, ES, or HB), standard Euclidean distance maps from 0 to infinity while normalized Euclidean distance maps from 0 to 2 (Table 3-2).

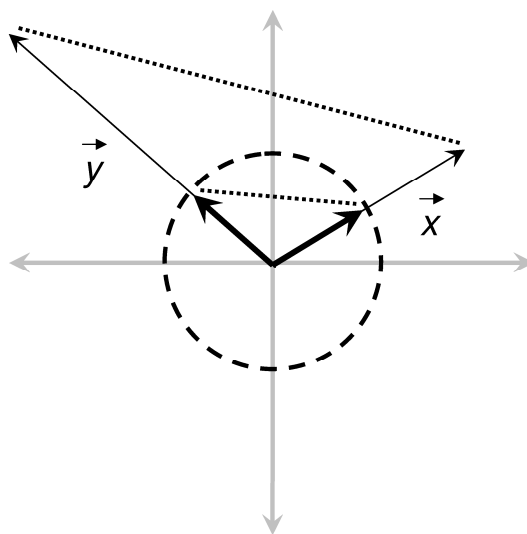


Figure 3-3. Schematic depiction of standard (thin) versus normalized (thick) footprint vectors (x , y). The maximum distance between normalized vectors on the unit circle is 2 while the distance between standard vectors can be infinite.

Similarity measures based on the standard Pearson correlation coefficient ($r = \text{cov}(\vec{x}, \vec{y}) / \sqrt{\text{var}(\vec{x}) \text{var}(\vec{y})}$) were also evaluated. Somewhat similar to normalized Euclidean distance, the standard Pearson metric quantifies similarity based on the relative magnitudes of each interaction. As a fourth alternative, threshold-based Pearson correlation coefficients (r_{thresh}) were also computed using a reduced set of residues consisting of only the most significant interactions. In this case, a user-defined threshold is employed and the union of the two footprints (reference and candidate pose) is used to enforce that an identical set of residues is used in the calculation. Interactions here were included when the absolute value \geq threshold which for VDW, ES, and HB footprints was set to 1.0, 0.1, and 0.5 kcal/mol, respectively. For a single type of footprint, both standard and threshold Pearson coefficients map from -1 to 1 (Table 3-2). It is important to note the distinction in nomenclature between "threshold" which may be used to determine which residues are most important, and therefore to be included in a threshold-based footprint, from a "score cutoff" (as described below) which may be used to identify footprint(s) with strong similarity.

3.2.2 Pose Identification.

As illustrated in Figure 3-4, two key tests were employed to characterize the utility of using footprint-based methods for structure-based drug design. The first test (Figure 3-4a) involves examining the ability of footprint methods to correctly identify crystallographically determined binding geometries out of a set of decoys. Here, a score cutoff (i.e. correlation or distance value) is employed to classify whether a given pose will be predicted positive or predicted negative (Figure 3-4a, green region). To determine if predictions are actually positive or negative (Figure 3-4a, red region) the commonly employed $\leq 2.0 \text{ \AA}$ rmsd criteria is used to

assess if the pose is similar to that in the experimental complex. The results can also be classified into four quadrants (Figure 3-4a, blue region) representing (I) true positives (predicted positive && positive), (II) false positives (predicted positive && negative), (III) true negatives (predicted negative && negative), and (IV) false negatives (predicted negative && positive). The sum of the components in each of the different colored regions will be equal (positive + negative results = predicted positive + predicted negative results = true positive + false positive + true negative + false negative results).

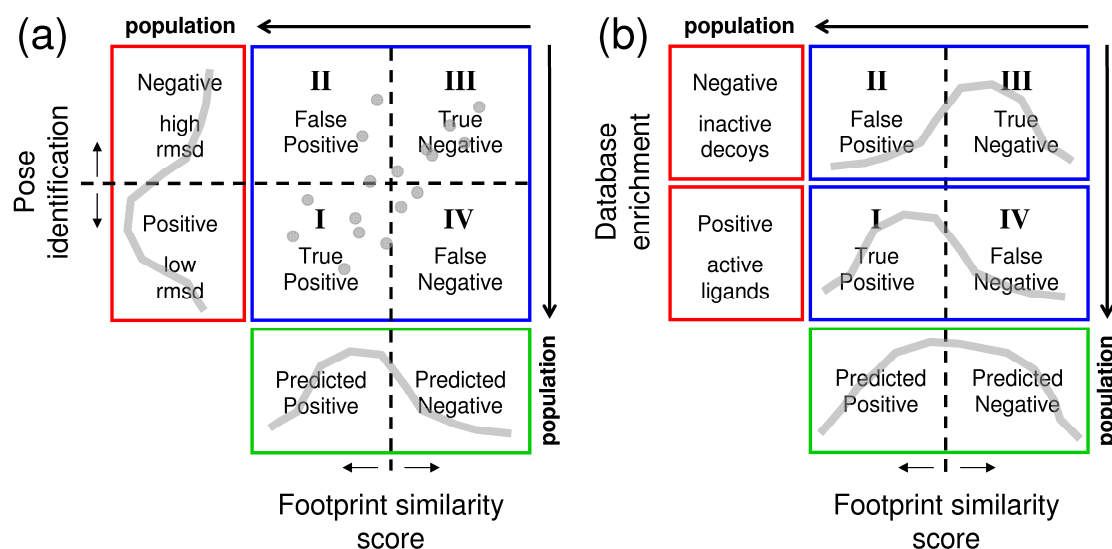


Figure 3-4. Partitioning of outcome space (positive or negative results, red region) as a function of prediction (predicted positive or predicted negative, green region) into four quadrants (blue region) representing (I) true positives, (II) false positives, (III) true negatives, and (IV) false negatives for (a) pose identification and (b) database enrichment definitions of success. Gray colored lines represent hypothetical data.

As a specific example, if a Euclidean-based footprint score cutoff of ≤ 0.3 was employed to make a classification, a molecule with a similarity score which equals 0.2 would be predicted as positive. Although the choice of score cutoff used to make prediction is somewhat arbitrary it should also be chosen with care. For example, although a more generous cutoff could be used to improve the number of true positives, as described further below, there is the risk that the

number of false positives may also increase. As a general rule the goal is to maximize the number of true positives and true negatives while minimizing the false positives and false negatives as is illustrated by the hypothetical data in the graphic (Figure 3-4a, gray data).

3.2.3 Database Enrichment.

The second key test (Figure 3-4b) involves assessing the ability of footprint-based scoring to predict whether a given compound will have biological activity (yes or no definition). From a virtual screening standpoint, if active ligands can be statistically scored better than inactive ligands (termed database enrichment), then rank-ordering of candidate ligands based on score provides a mechanism for focusing on only the most promising compounds. Using databases such as DUD,^{106,108} consisting of known active ligands seeded into a large group of decoys, scoring accuracy (Figure 3-4b, red region) is gauged by comparing the number of active and inactive compounds (Figure 3-4b green region) predicted to be in a given percentage of the database. As illustrated by the hypothetical data in the graphic, (Figure 3-4b gray lines) the scoring function should ideally separate active vs inactive molecules when viewed as histograms. As before, if a score cutoff is applied, the results can be classified into four quadrants (Figure 3-4b, blue region). However, as the actual positive and negative regions (Figure 3-3b, red region) are binary (yes/no activity); each sub-region contains only a single value representing the number of actual actives or decoys.

The amount of enrichment a given method provides versus random prediction is often gauged through use of receiver operator characteristic (ROC) curves¹²⁴ which plot the true positive rate (true positives / positives) versus the false positive rate (false positives / negatives). In conjunction with calculation of the area under the curves (AUC), both ROC and AUC metrics

can be used to identify which classifiers are significantly better than random. For example, a truly random classifier will have an ROC with slope = 1 and an AUC of 0.5 while a scoring function which is good at separating positives from negatives will yield a steep early rise in the ROC curve with a corresponding AUC much closer to 1.0. The amount of fold enrichment (FE) in any given region of the database (typically the first 1-10%) may also be of interest and is defined here by taking the AUC for the range of interest normalized by the area expected from a random classifier ($FE = AUC / AUC_{\text{random}}$). See Appendix B for additional discussion of enrichment studies as an evaluation tool.

3.3 Computational Details.

3.3.2 Pose Identification Datasets.

Candidate binding geometries to quantify pose identification success rates were derived from the SB2010 database recently reported by Mukherjee et al.¹³ and interested readers should consult the manuscript for specifics regarding receptor and ligand structure preparation steps and docking protocols. Briefly, three distinct sampling methods were used to generate ensembles of poses, for each of the 780 protein-ligand complexes in SB2010, containing potentially correct ligand binding geometries as well as numerous low-energy decoys. The rigid (RGD) protocol attempts to rigidly place and optimize the known experimental pose back into the binding site through sampling the six degrees of rigid body translation and rotation. The fixed anchor (FAD) protocol tests re-growth of a molecule starting from crystallographic ligand scaffold positions. The flexible (FLX) protocol employs the DOCK anchor-and-grow algorithm,^{26,36} which involves orienting of ligand scaffolds (anchors) into the binding site followed by flexible conformer growth. A top-first clustering procedure^{26,36} was used to prune away redundant structures

during orientation and growth (FAD, FLX) and at the final stage of ranking (RGD, FAD, FLX). The retained group, termed clusterheads, each represent the lowest energy pose identified among geometrically related structures ($< 2 \text{ \AA}$) sampled during the docking. Use of clusterheads helps to ensure diversity in the ensemble when retaining a reduced set of top-scoring poses.

For the three datasets (RGD, FAD, FLX) the 50 top-ranked clusterheads for each system were energy minimized and rank-ordered on the protein Cartesian coordinates using 1000 iterations of simplex optimization with DOCK6.4. To help enforce that the original grid-based and subsequent Cartesian space poses would remain similar after an energy minimization, an rmsd-based harmonic tether was used to restrain each pose to the original input coordinates (force constant $k = 10 \text{ kcal / mol \AA}^2$). As shown in Figure 3-5a, most, but not all SB2010 systems have at least 50 clusterheads. Figure 3-5b plots the lowest-rmsd pose identified, relative to the experimental ligand geometry, after energy minimization of the original grid-based ensembles (clusterheads). The number of systems in Figure 3-5b to the left of the 2 \AA rmsd line for each sampling protocol (RGD=775, FAD=748, FLX=680) constitutes perfect sampling subsets, with associated ensembles (RGD = 38,569, FAD = 19,073, and FLX = 26,830), and these subsets were employed in the pose identification experiments described below. Importantly, use of perfect sampling subsets ensures that at least one pose for each system is close to the experimental pose which is an appropriate data group to use for tests designed to evaluate scoring (not sampling) accuracy.

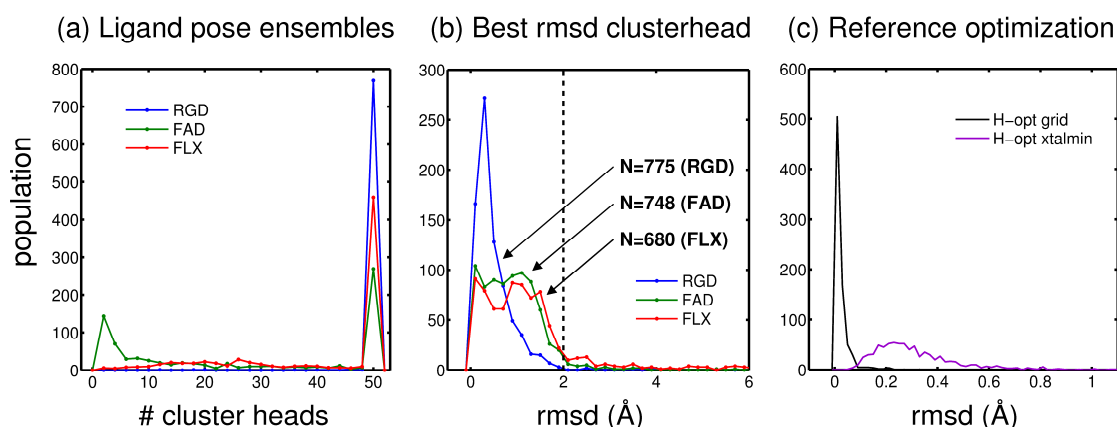


Figure 3-5. Database preparation histograms. (a) Population of systems with a given number of clusterheads (max=50) derived from Cartesian space minimizations of grid-based results reported by Mukherjee et al.¹³ (b) Population of systems with a given rmsd using only the single lowest-rmsd pose found among the ensemble of poses retained. The portion to the left of the dashed line at 2 Å rmsd constitutes perfect sampling subsets for (RGD=775), fixed-anchor (FAD=748), and flexible (FLX=680) ligand sampling. (c) Population of ligand rmsds for reference poses after polar hydrogen optimizations using the energy grids (black line) and subsequent energy minimizations in Cartesian space (purple line) using a harmonic tether.

3.3.2 Database Enrichment Datasets.

For the enrichment tests, systems were taken from the directory of useful decoys (DUD) database.^{106,108} Three systems were evaluated, (i) neuraminidase (pdb code 1A4G)¹²⁵ consisting of 1,874 decoys and 49 actives, (ii) trypsin (pdb code 1BJU)¹²⁶ consisting of 1,664 decoys and 49 actives, and (iii) EGFR (pdb code 1M17)⁵⁸ consisting of 15,996 decoys and 475 actives. Decoy and active ligands were used as originally downloaded from DUD (default protonation states and partial atomic charges). Docking setups (receptor preparation, energy grids, docking spheres, etc) were taken from SB2010¹³ with the native cognate ligands from each pdb entry used as the footprint reference (zanamivir from 1A4G, benzamidine derivative from 1BJU, and erlotinib from 1M17). Docking calculations employed identical grid-based FLX protocols described by Mukherjee et al.¹³ with the exception that the single best scoring pose was retained for subsequent Cartesian-based energy minimization (as described above) followed by footprint rescaling. Enrichment was evaluated by plotting standard ROC curves.

3.3.3 Footprint Reference Preparation.

Initial testing revealed that in some cases footprints, and thus FPS scores, could be sensitive to placement of hydrogen atoms. Perhaps not surprisingly, sensitivity appeared to be most pronounced for electrostatic (ES) and hydrogen-bond (HB) interactions involving charged moieties. To reduce variability as a result of sub-optimal hydrogen rotamers in molecules used as the reference, an optimization procedure was developed in which growth routines in the DOCK6.4 program were co-opted for sampling polar –OH, –SH and –NH groups deemed most susceptible. The procedure uses a modified DOCK flexible definition file (flex.defn) with six angle steps sampled for each torsion at 0°, 60°, 120°, 180°, 240°, and 300° followed by minimization. Sampling is performed using standard DOCK energy grids to achieve quick optimization and a stiff harmonic restraint ($k = 1000 \text{ kcal / mol \AA}^2$) is used on ligand heavy atoms to insure only hydrogen atoms move. Following sampling the most favorable pose is minimized on the Cartesian coordinates (restraint $k = 10 \text{ kcal / mol \AA}^2$) so that footprints may be computed. It should be noted that additional hydrogen optimization is not generally necessary for poses generated using FAD or FLX protocols as the –OH, –SH and –NH polar groups are sampled during ligand growth procedures. Thus, hydrogen optimization was only done for molecules used as a reference. As shown in Figure 3-5c the hydrogen optimization and subsequent minimization process minimally alters the experimental binding poses (rmsds typically $\leq 0.2 \text{ \AA}$) yet these structures result in better behaved reference footprints.

3.3.4 Footprint Rescoring Protocols.

The modular nature of the DOCK program lends itself to be easily extended with new scoring functions.^{36,37} The ability to compute footprints and footprint similarity scores was

implemented into an inhouse version of DOCK6.4 as a new scoring function termed "descriptor score". Code modifications will be made available to registered users of DOCK through the official UCSF distribution site (<http://dock.compbio.ucsf.edu>) in the near future. FPS scores (FPS_{VDW} , FPS_{ES} , FPS_{HB} or any combination thereof) may be calculated with any of the four comparison methods described above (standard Pearson, threshold Pearson, standard Euclidean, normalized Euclidean) using a user supplied reference. If desired, users can also output a comma separated text file consisting of a list of residue numbers with associated energies for the reference and candidate poses which facilitates graphical plotting of the footprints. Importantly, the FPS rescoring procedure is relatively fast. As an example, grid-based docking of 15,996 molecules to EGFR using the DUD subset with FLX protocols takes ca 159 seconds per molecule on single 3.2Mhz Pentium IV cpu. Energy minimization in Cartesian space takes an additional ca 17 seconds per molecule followed by FPS scoring which takes ca 0.13 seconds per molecule. Thus, compared to the time required for flexible docking the additional costs to obtain FPS scores are minimal.

3.4 Results and Discussion.

3.4.1 Footprint Similarity (FPS) vs DOCK Cartesian Energy (DCE) Scores for Pose Identification.

Table 3-3 shows pose identification results using FPS or DCE scoring criteria to choose a "top pose" from among the RGD (N =775), FAD (N=748), and FLX (N=680) perfect sampling subsets (Figure 3-5b) from the SB2010 database¹³ (see Methods). Ideally this top pose should match the crystal structure with a low heavy atom rmsd. Here, use of perfect sampling subsets ensure that at least one pose for each system is in fact within 2 Å rmsd of the crystal structure.

Other poses ($> 2 \text{ \AA}$ rmsd) in each system ensemble (Figure 3-5b) may be thought of as decoys. For a given protocol, percent success is the ratio between the number of systems with top poses correctly identified and the total number in the perfect sampling subset (e.g. in Table 3-3 the top right most entry is $80.9\% = 627 \text{ identified} / 775 \text{ possible} \times 100$). For each of the three subsets (RGD, FAD, and FLX) the standard Pearson, standard Euclidean, normalized Euclidean, and threshold Pearson methods were used to compute footprint similarities (FPS) scores using footprints representing VDW, ES, VDW+ES, or VDW+ES+HB terms. It is important to note that no scoring cutoff (i.e. above/below a certain FPS value) was employed in choosing top scoring poses for the results presented in Table 3-3. For each system, the best scoring pose was always retained even if the FPS score relative to the reference was poor.

Table 3-3. Pose identification success using Footprint similarity (FPS) vs DOCK Cartesian energy (DCE) methods to rescore rigid (RGD), fixed anchor (FAD) and flexible ligand (FLX) pose ensembles.

Row	Ligand Ensemble	FPS Standard	FPS Standard	FPS Threshold	FPS Normalized	DCE
		Pearson A	Euclidean B	Pearson C	Euclidean D	E
VDW+ES						
1	RGD	691 (89.2%) ^a	718 (92.6%)	683 (88.1%)	707 (91.2%)	627 (80.9%)
2	FAD	642 (85.8%)	638 (85.3%)	644 (86.1%)	652 (87.2%)	606 (81.0%)
3	FLX	563 (82.8%)	565 (83.1%)	556 (81.8%)	574 (84.4%)	489 (71.9%)
VDW						
4	RGD	687 (88.6%)	684 (88.3%)	662 (85.4%)	687 (88.6%)	445 (57.4%)
5	FAD	638 (85.3%)	630 (84.2%)	621 (83.0%)	638 (85.3%)	464 (62.0%)
6	FLX	545 (80.1%)	539 (79.3%)	525 (77.2%)	545 (80.1%)	309 (45.4%)
ES						
7	RGD	579 (74.7%)	583 (75.2%)	576 (74.3%)	579 (74.7%)	398 (51.4%)
8	FAD	601 (80.3%)	573 (76.6%)	598 (79.9%)	603 (80.6%)	460 (61.5%)
9	FLX	521 (76.6%)	505 (74.3%)	513 (75.4%)	522 (76.8%)	314 (46.2%)
VDW+ES+HB						
10	RGD	670 (86.5%)	726 (93.7%)	590 (76.1%)	685 (88.4%)	633 (81.7%)
11	FAD	621 (83.0%)	643 (86.0%)	590 (78.9%)	632 (84.5%)	606 (81.0%)
12	FLX	557 (81.9%)	564 (82.9%)	501 (73.7%)	561 (82.5%)	492 (72.4%)

^aNumber of molecules in which the pose identified was ≤ 2 Å from the x-tal structure pose followed by success rates in parenthesis. Pose ensembles (RGD = 775, FAD = 748, FLX = 680) derived from docking runs reported by Mukherjee et al.¹³

It should be emphasized that the results in Table 3-3 only test scoring and not sampling. Thus, the VDW+ES values (rows 1 to 3 column E) for DCE_{VDW+ES} (RGD=80.9%, FAD=81.0%, FLX=71.9%) are representative of the accuracy of the standard DOCK scoring function. Importantly, these results are similar to those reported by Mukherjee et al.¹³ (RGD=83.5%, FAD=81.6%, FLX=72.6%) for the analogous perfect-sampling subsets suggesting excellent correspondence between grid-based and Cartesian-based results. With one exception (threshold Pearson with $FPS_{VDW+ES+HB}$, rows 10-12 column C) all methods and protocols in Table 3-3 for computing footprint similarity scores yield higher success rates than the comparable DCE scores. Although the values for many of the tests in Table 3-3 yield similar results, overall, use of the FPS_{VDW+ES} footprint classifier with normalized Euclidean distance (rows 1 to 3 column D) appears best at identifying correct poses from within the various ligand ensembles (RGD=91.2%, FAD=87.2%, and FLX=84.4%). Specifically, FPS_{VDW+ES} increases success over comparable DCE_{VDW+ES} scores by 10.3 %, 6.2 %, and 12.5 % for RGD, FAD, and FLX respectively. Although not directly comparable, due to differences in dataset size and/or analysis, prior studies have also reported improvements in identification of native-like poses, relative to using a standard scoring function, using bit-string representations and related methods. Interested readers should consult studies by Singh and coworkers,^{113,115,116,121} Kelly et al.,¹¹⁴ Marcou et al.,¹¹⁷ Mpamhanga et al.,¹¹⁸ Pfeffer et al.,¹¹⁹ Renner et al.,¹²⁰ and Pérez-Nueno et al.¹²²

Interestingly, use of a single energetic descriptor in DCE scores yields severely degraded results compared to using the corresponding footprint for FPS scores (Table 3-3 columns D vs E rows 4 to 9). For example, normalized Euclidean FPS_{VDW} (80-89%) and FPS_{ES} (75-81%) show much higher success rates versus analogous DCE_{VDW} (45-62%) and DCE_{ES} (46-62%) methods. Thus, the information encoded by a single VDW or ES footprint vector appears to be sufficient

to identify native-like poses. However, tests using a single HB footprint revealed that there is insufficient information encoded due to the discrete nature of hydrogen bonds employed in the present implementation (yes or no geometric definition with one hydrogen bond = -1 kcal/mol). In addition, the fact that numerous poses (real or decoy) make only a few, or even no hydrogen bond interactions with a target, precludes rank-ordering using HB footprints alone. In any event, use of HB in conjunction with VDW and ES footprints is not problematic however the addition generally decreases the success rates (Table 3-3 VDW+ES vs VDW+ES+HB). Surprisingly, modifying the standard DCE scoring function to energetically account for intermolecular hydrogen bonds yields no degradation and in fact shows a slight improvement (Table 3-3 $DCE_{VDW+ES+HB}$ vs DCE_{VDW+ES}).

3.4.2 Functional Relationships between Methods used to Compute FPS Scores.

To more closely examine how results using two different comparison methods may be related, Figure 3-6 shows three functional relationships (standard vs threshold Pearson, standard vs normalized Euclidean, and standard Pearson vs normalized Euclidean) derived from plotting clusterhead ensembles for all FLX systems ($N = 680$ structures \times ca. 39.445 average # of clusterheads each = 26,830 footprints). Data derived from both ES and VDW footprint similarity scores are shown and the results are colored by population. Across all datapoints, both Pearson methods yield results which are quantitatively similar especially when FPS scores are highly correlated (r -values near 1) or fall within the 0 to 1 range (Figure 3-6a,b blue and red populations). Interestingly, when the FPS_{VDW} scores themselves become anti-correlated (r -values ≤ 0) there is significantly less agreement between the two comparison methods but only for the VDW results. In contrast, results using both Euclidean methods also show a strong linear

relationship when FPS scores are nearest 0 and therefore most correlated (Figure 3-6c,d blue and red populations) but as the scores themselves become less-correlated (values $\gg 0$) the ca linear relationship is lost for both VDW and ES results. Finally, the strong relationship ($d_{norm} \approx \sqrt{2(1-r)}$, see derivation in section 4.4.3) between standard Pearson and normalized Euclidean methods (Figure 3-6e,f) suggests both comparison metrics will yield very similar results across the entire range. As noted above, FPS_{VDW+ES} scoring in combination with normalized Euclidean distance appears marginally best at pose identification. Therefore, unless otherwise stated, and to simplify discussion in the remainder of the text, normalized Euclidean methods in combination with FLX results will be emphasized.

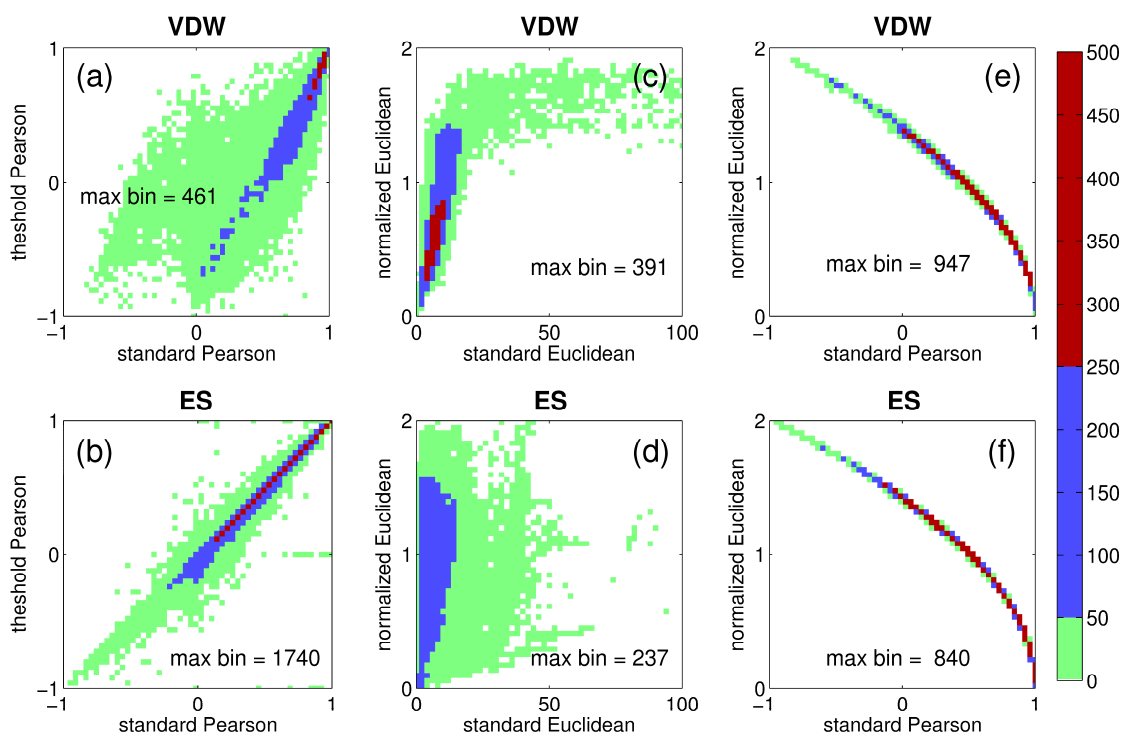


Figure 3-6. Functional relationships between footprint similarity (FPS) scores computed for van der Waals (VDW, top) and electrostatic (ES, bottom) interactions using (a, b) standard Pearson vs threshold Pearson, (c, d) standard Euclidean vs normalized Euclidean, and (e, f) standard Pearson vs normalized Euclidean. Population color ranges for green = [1, 50], blue = [51, 250], and red = [251, 500+] are derived from the total FLX ensemble of $N = 26,830$ footprints.

3.4.3 Approximate Mathematical Relationship.

The following corollary and theorem prove the approximate relationship that exists between normalized Euclidean distance and Pearson Correlation Coefficient (two methods employed to calculate the similarity for interaction footprint vectors). The strong relationship is demonstrated in Figure 3-6 panels e-f. This relationship is strong because the mean of the footprint vectors is usually close to zero; however, if the mean is not then this relationship will be much weaker. For example for threshold based footprints the mean will likely not be close to zero.

Corollary 1: if \vec{u} and \vec{v} are unit vectors, then $\|\vec{u} - \vec{v}\| = \sqrt{2(1 - \cos(\theta))}$, where θ is the angle between \vec{u} and \vec{v} .

Proof:

Let \vec{u} and \vec{v} be unit vectors. Then,

$$\begin{aligned} \cos(\theta) &= \frac{\vec{u} \cdot \vec{v}}{\|\vec{u}\| \|\vec{v}\|} = \vec{u} \cdot \vec{v} \quad (\because \|\vec{u}\| = 1) \\ \|\vec{u} - \vec{v}\| &= \sqrt{\sum (u_i - v_i)^2} = \sqrt{\sum u_i^2 + \sum v_i^2 - 2\sum u_i v_i} \\ &= \sqrt{1+1-2\sum u_i v_i} \quad (\because \|\vec{u}\|^2 = \sum u_i^2 = 1) \\ &= \sqrt{2(1 - \vec{u} \cdot \vec{v})} \end{aligned}$$

Therefore, $\|\vec{u} - \vec{v}\| = \sqrt{2(1 - \cos(\theta))}$

Theorem: $d_{norm} \approx \sqrt{2(1-r)}$ for two vectors x and y , whose means are close to zero, where d_{norm} is the normalized Euclidean distance and r is the Pearson Correlation Coefficient (both calculated between the two vectors).

Proof:

$$r = \frac{\text{cov}(\bar{x}, \bar{y})}{\sqrt{\text{var}(\bar{x})}\sqrt{\text{var}(\bar{y})}}$$

The correlation coefficient can also be thought of as the cosine of the angle formed between the mean-modulated vectors x^μ and y^μ where $\bar{x}^\mu = [x_i - \mu_x]$ and $\bar{y}^\mu = [y_i - \mu_y]$, and μ represents the mean of each vector.

$$r = \frac{\bar{x}^\mu \cdot \bar{y}^\mu}{\|\bar{x}^\mu\| \|\bar{y}^\mu\|} = \cos(\theta)$$

Note that

$$r = \cos(\theta) = \frac{\bar{x}^\mu \cdot \bar{y}^\mu}{\|\bar{x}^\mu\| \|\bar{y}^\mu\|} \approx \frac{\bar{x} \cdot \bar{y}}{\|\bar{x}\| \|\bar{y}\|} = \cos(\theta^*)$$

The mean of a footprint is normally close to zero since most footprint entries are close to zero so this is a reasonable approximation. The normalized Euclidean distance (d_{norm}) is defined as:

$$d_{norm} = \|\bar{\chi} - \bar{\gamma}\|$$

where $\bar{\chi} = \frac{\bar{x}}{\|\bar{x}\|}$, and $\bar{\gamma} = \frac{\bar{y}}{\|\bar{y}\|}$.

$$d_{norm} = \sqrt{2(1 - \cos(\theta^*))} \approx \sqrt{2(1-r)}$$

There is a relationship between the $\cos(\theta^*)$ and normalized Euclidean (d_{norm}) by **Corollary 1** and because the angle between two vectors is the same as that between their unit vectors. Therefore, the approximate relationship is demonstrated.

3.4.4 Predicted Positive and Predicted Negatives.

Figure 3-7 and Table 3-4 show normalized Euclidean results using the FLX-derived dataset in terms of the three areas of partitioning described in Methods (see Figure 3-4a) comprised of (1) positive and negative regions, (2) predicted positive and predicted negative regions, and (3) true positive, false positive, true negative, and false negative quadrants. Positive and negative regions in Figure 3-7 are shown below and above, respectively, the horizontal dashed line at 2 Å in the rmsd histograms (left of each panel). From a prediction standpoint, the predicted positives and predicted negatives in Figure 3-7 are to the left and right, respectively, of the vertical dashed line representing a 0.6 score cutoff in the FPS score histograms (bottom of each panel). It is important to note that the choice of a specific FPS score cutoff choice for prediction is user defined. Table 3-4 lists results using a 0.3, 0.6, or 0.9 score cutoff which under these conditions appear to be reasonable choices. Results in each of the four quadrants in Figure 3-7a,b indicate populations which follow the color ranges for green = [1, 5], blue = [6, 20], and red = [21, 30+]. For completeness, Figure 3-7 and Table 3-4 show results both when keeping only the *best scored* pose identified for each of the 680 FLX systems as well as for *all poses* in the total ensemble of FLX-derived clusterheads (N=26,830).

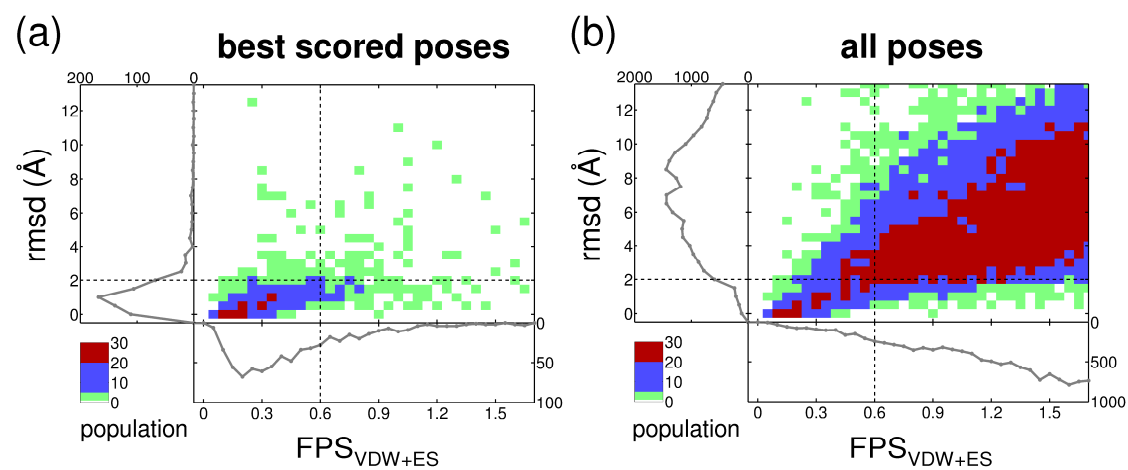


Figure 3-7. Two dimensional histograms of rmsd versus $\text{FPS}_{\text{VDW+ES}}$ score for (a) the best scored poses ($N=680$) and (b) the entire ensemble derived from all poses ($N=26,830$). Population color ranges for green = [1, 5], blue = [6, 20], and red = [21, 30+].

Table 3-4. FLX results scored with $\text{FPS}_{\text{VDW+ES}}$ for three differing footprint similarity score cutoffs using a 2 Å rmsd to separate positive from negative regions.

Set	Cutoff	Positive	Negative	Predicted Positive	Predicted Negative	True Positive	False Positive	True Negative	False Negative
best scored ^a	0.3			251	429	240	11	95	334
	0.6	574	106	507	173	458	49	57	116
	0.9			618	62	537	81	25	37
all poses ^b	0.3			295	26,535	261	34	25,831	704
	0.6	965	25,865	1,185	25,645	577	608	25,257	388
	0.9			3,026	23,804	759	2267	23,598	206

^a $N = 680$. ^b $N = 26,830$.

Generally good separation is observed in Figure 3-7 with higher populations appearing in true positive and true negative quadrants relative to false positive and false negative quadrants (population legend follows red > blue > green). Ideally, the number of true positives and true negatives should be near 100% while the number of false positives and false negatives should be near 0%. Quantitatively, the percent values of each quadrant, computed from the raw numbers in Table 3-4, suggest useful predictive ability. For example, the best scored poses dataset using a FPS cutoff of 0.6 yields a strong true positive rate = 79.8% ($458 / 574 \times 100$) and a relatively strong true negative rate = 53.8%. The corresponding false positive (46.2%) and false negative (20.2%) rates are smaller as desired. At the looser 0.9 cutoff the true positive rate substantially increases to 93.6% however the corresponding false positive rate also increases (76.4%) which is not desirable. As expected, the true negative (23.6%) and false negative (6.4%) rates show a corresponding decrease. Importantly, as discussed further below, a substantial number of poses labeled here as false positive appear to be miscategorized. Roughly similar trends (true positive and true negative quadrants > false positive and false negative quadrants) are seen using the dataset derived from all ligand poses (Figure 3-7b). At the 0.6 cutoff the true positive rate = 59.8%, the true negative rate = 97.6%, the false positive rate = 2.3%, and the false negative rate = 40.2%. Here, the large numbers of decoys (negatives) present in the all poses dataset (N = 25,865) yields excellent statistics for both true negative and false positive rates. For comparison, similar analysis based on a quadrant partitioning of rmsd vs score was reported by Marco et al.¹¹⁷ using binary fingerprinting.

3.4.5 False Positive Examples.

Focusing on the best scored dataset, although changing the FPS score cutoff from 0.6 \rightarrow 0.9 increases the number of true positives (79.8% \rightarrow 93.6%) the number of false positives also increases (46.2% \rightarrow 76.4%). In general, as the vertical dashed line representing footprint similarity in Figure 3-7 is shifted from left to right, greater numbers of false positives will occur. However, while it may be acceptable in a virtual screen to discard molecules that could bind (false negatives) as long as a sufficient number of true positives are retained, it is extremely undesirable to retain non-active molecules (false positives) because molecules without activity may be passed onto more costly testing (i.e. purchase or synthesis).

Figures 3-8 and 3-9 graphically illustrates how poses classified as false positive may in fact be geometrically and chemically correct in terms of binding. Overlays of predicted (green) versus crystallographic (red) poses are shown along with corresponding FPS score, rmsd in Å, and potential sources of misclassification which primarily involves: (i) symmetry issues with rmsd calculations, and (ii) solvent-exposed moieties which do not interact with the binding site. False positives not belonging to either of these two categories appear to arise from the potentially useful phenomena (i.e. in virtual screening) that poses can yield similar footprints despite poor geometric overlap and are here labeled promiscuous. The group in Figure 3-8, termed type I false positives, represents those from the best pose dataset with excellent footprint overlap ($\text{FPS}_{\text{VDW+ES}} < 0.3$) but were classified as failures in terms of a close-to-medium geometric match ($\text{rmsd} > 2 \text{ \AA}$ and $< 5 \text{ \AA}$). The group in Figure 3-9, termed type II false positives, shows more extreme cases in which ligand poses have reasonable footprint overlap ($\text{FPS}_{\text{VDW+ES}} < 0.6$) but very poor geometric matches ($\text{rmsd} > 5 \text{ \AA}$). It is important to note the categories used here in

Figures 3-8 and 3-9 defining systems as solvent-exposed, promiscuous, and to a lesser extent symmetry-related, are subject to interpretation.

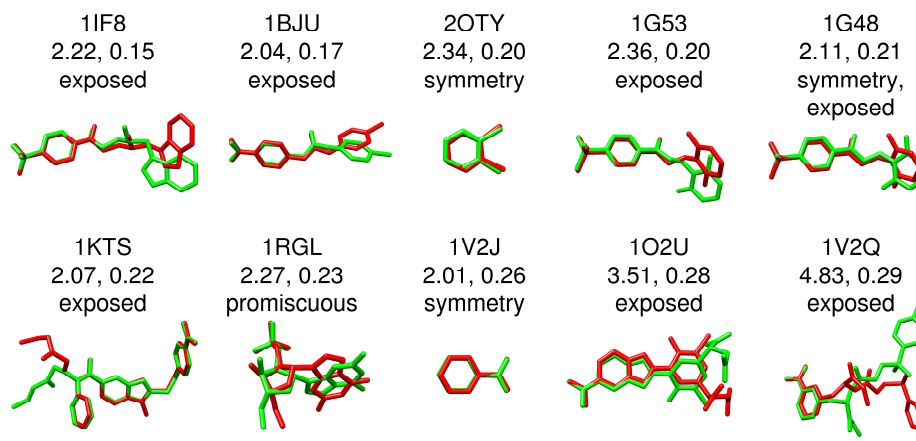


Figure 3-8. False positive examples type I. Excellent similarity scores ($FPS_{VDW+ES} < 0.3$) but classified as failures due to a close-to-medium geometric match ($rmsd > 2 \text{ \AA}$ and $< 5 \text{ \AA}$). The associated PDB code, rmsd in \AA , FPS score, and overlay of the predicted (green) versus crystallographic (red) pose are shown for each system.

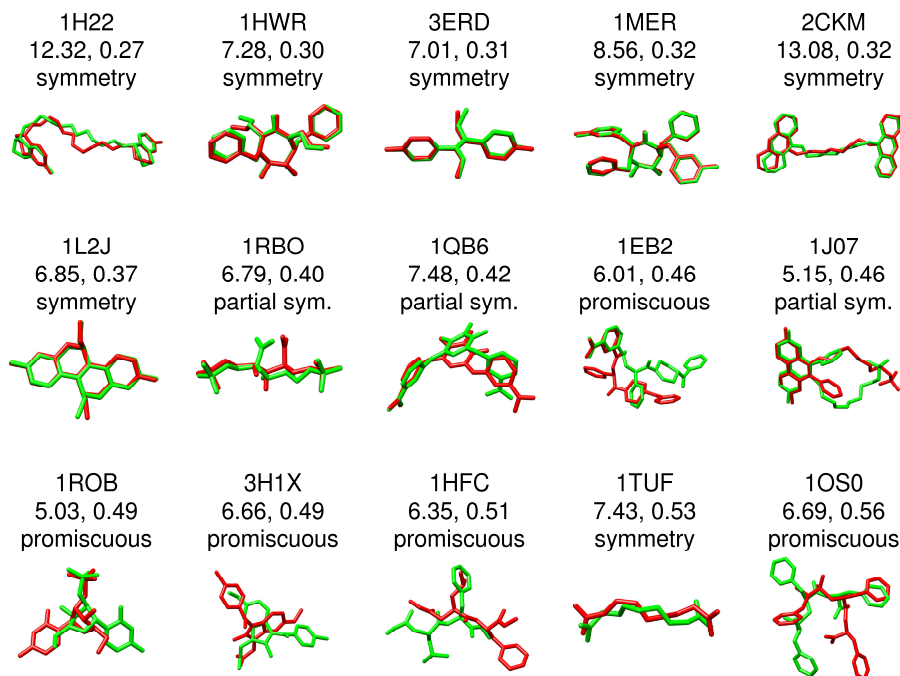


Figure 3-9. False positive examples type II. Good similarity scores ($FPS_{VDW+ES} < 0.6$) but classified as failures due to a poor geometric match ($rmsd > 5 \text{ \AA}$). The associated PDB code, rmsd in \AA , FPS score, and overlay of the predicted (green) versus crystallographic (red) pose are shown for each system.

Symmetry: Symmetric molecules or molecules containing moieties with symmetry often produce docked candidates in which poses or functional groups (i.e. aromatic rings) are flipped about an axis of symmetry. At this time, DOCK does not correct for symmetry and several of the best scored poses in Figures 3-8 and 3-9, based on visual examination, have higher than expected rmsd values. Particularly dramatic examples are the two symmetric HIV protease inhibitors 1HWR (7.28 Å) and 1MER (8.56 Å) shown in Figure 3-9 which have essentially perfect overlap with the reference but high rmsds. More extreme examples include the long symmetric inhibitors 2CKM (13.08 Å) and 1H22 (12.32 Å) for which the lowest energy docked poses are flipped by ca 180 degrees resulting in high rmsd. Marcou et al.¹¹⁷ similarly found that many false positives also turned out to be molecules containing symmetry. As a possible alternative to traditional rmsd-based methods, Kroemer et al.¹²⁷ has described the interaction-based accuracy classification (IBAC) method which judges the correctness of a docked pose by manually comparing key receptor-ligand interactions identified in the crystal structure. IBAC however, as the authors note, is not easily automated. More recently, Trott and Olson¹²⁸ have introduced an alternative definition for computing rmsd in the program AutoDock Vina which the authors indicate accounts for symmetry, partial symmetry, and near symmetry. Efforts to incorporate symmetry-corrected rmsd calculations into DOCK are under evaluation. It should be emphasized however that although accounting for symmetry may affect pose identification accuracy it will not directly impact virtual screening.

Solvent-exposed: Solvent exposed moieties of a bound ligand may not interact strongly with the receptor. In such cases, it is not unexpected that exposed groups could adopt multiple conformations while the bulk of the molecule, and therefore the footprint, remains unchanged. As the rmsd metric takes into account all ligand atoms such systems could be unfairly penalized

by using rmsd to evaluate potential FPS scoring accuracy. Interestingly, most of the false positive errors in the type I classification (close to medium rmsds) appear to fall into the "exposed" category. 1KTS¹²⁹ provides a clear example. Here, only the solvent exposed ethyl ester substituent is not well overlaid (Figure 3-8) although the rest of the molecule shows almost perfect overlap. Perhaps not surprisingly, as discussed by Nar et al.¹³⁰ for the same ligand bound to Factor Xa, the electron density is not as well defined in this region.

Promiscuous: Importantly, the available conformational space for a given ligand, even for ligands with no (or imperfect) symmetry, can yield reasonable FPS scores despite poor geometric and/or chemical overlap as shown in Figures 3-8 and 3-9 for 1RGL, 1ROB, 3H1K, 1HFC, and 1OS0. Labeled here as promiscuous, of the three categories (symmetry, solvent exposed, or promiscuous), these could be considered as bona fide failures of the pose identification tests. On the other hand, the misidentified conformations also suggest, conceptually, that compounds with high footprint overlap can be structurally diverse. Additional crossdocking and database enrichment studies presented below strongly suggests this hypothesis to be true. From a virtual screening standpoint, promiscuity may in fact be desirable by allowing for identification of new molecules with chemotypes, scaffolds, and/or functionality different from known inhibitors.

3.4.6 False Negative Examples.

Although false negatives are generally considered less problematic than false positives an examination of systems which fall into this category was undertaken to more fully characterize the method. Table 3-5 shows representative examples in which good geometric overlap (low rmsds) is observed for correspondingly poor footprint scores (high FPS) defined by the ranges

rmsd < 1.0 Å and $FPS_{VDW+ES} > 1.0$. Interestingly, the poor FPS_{VDW+ES} scores in these false negative examples arise because only one of the two terms, FPS_{VDW} or FPS_{ES} is sub-optimal (Table 3-5 underline entries).

Table 3-5. False negative examples for the range defined by the range rmsd < 1.0 Å and $FPS_{VDW+ES} > 1.0$.

Code	rmsd (Å)	FPS_{VDW+ES}	FPS_{VDW}^a	FPS_{ES}^a
2QE4	0.35	1.34	0.14	<u>1.20</u>
1HSH	0.77	1.31	0.16	<u>1.15</u>
2F80	0.70	1.47	0.47	<u>0.99</u>
1CPI	0.56	1.05	0.13	<u>0.92</u>
1TNL	0.69	1.11	<u>0.92</u>	0.19
2JJ3	0.44	1.18	<u>0.81</u>	0.37
9AAT	1.00	1.01	<u>0.79</u>	0.22

^aPoor scores for individual terms are underlined.

In many instances, close inspection reveals the source of the poor footprint term as illustrated in Figure 3-10 for two representative systems, 2QE4 (estrogen receptor) and 9AAT (aspartate aminotransferase). As before, results for the reference and candidate poses are shown in red and green respectively. Only the most significant footprint interactions are shown with energetic differences indicated in black. Figure 3-10a dramatically highlights how a poor electrostatic footprint overlap may be a result of variation in intermolecular hydrogen bonding. Specifically, the positioning of a key ligand hydrogen atom (indicated as spheres), on the left side of the reference molecule (red) in Figure 3-10a, results in favorable ES interactions with Glu43 but unfavorable ES interactions with Arg84. However, for the candidate pose (green) with an alternate polar hydrogen rotamer, both interactions are reduced significantly in magnitude and a new favorable ES interaction is observed with the backbone carbonyl at position Leu77. In contrast, both poses show the same rotameric state for hydrogen bonding with His203 (overlapping spheres on the right side of molecules) and the accompanying

energetic difference at this position is zero. Overall, the observed correspondence between changes in geometry with energy is physically reasonable.

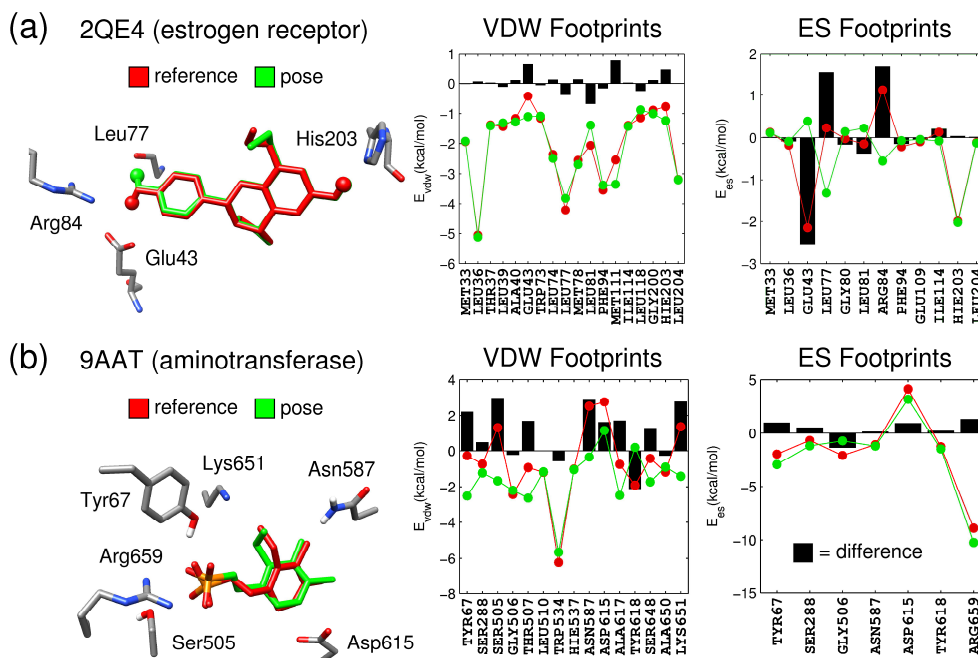


Figure 3-10. Pose and footprint comparisons for (a) 2QE4 and (b) 9AAT showing results for the reference pose in red, the docked pose in green, and per-residue differences as black bars.

The second example (Figure 3-10b) is representative of cases in Table 3-5 in which VDW footprints are dissimilar despite well-overlapped FPS_{ES} profiles. Here, the significant ES attraction between the ligand sulfonate and Arg659 (~ 10 kcal/mol), in concert with interactions at Tyr67 (favorable) and Asp615 (unfavorable) which are in general greater in magnitude than any individual VDW energy, likely impacts the fact steric packing differences show greater variation. Interestingly, the candidate pose in Figure 3-10b (green line), in comparison to the reference (red line), yields a somewhat more satisfying VDW footprint in that most active site per-residue terms become favorable while at the same time the ES footprint remains unchanged. In contrast, the reference pose (red line), shows unfavorable energies (e.g. at Ser205, Asn587, Asp615, and Lys651) which could indicate sub-optimal X-ray refinement of the ligand. In any

event, the strong interaction with Trp534 is well preserved by both poses. Importantly, both examples in Figure 3-10 provide evidence that molecular footprints capture interactions which make physical sense but additionally highlight the need for care when preparing reference poses, especially for ligands containing polar hydrogens. In a more general sense, the results also indicate the importance of using intermolecular energy minimization, prior to computing FPS scores, for all binding geometries being considered, including references. Although for 9AAT, energy minimization alone was not sufficient to alleviate all unfavorable steric packing in the original crystallographic pose (Figure 3-10b red VDW footprints).

Crossdocking Rescoring: As recently reported by Mukherjee et al.¹³ carbonic anhydrase provides a good system on which to test new scoring functions given that crossdocking experiments, despite high scoring failures, yield few sampling failures. Crossdocking employs a related family of proteins, aligned into a common "master" coordinate frame, thus enabling docking of all ligands into all receptors. Importantly, the alignments provide, in addition to cognate protein-ligand pairs that lie on the diagonal matrix entries, off-diagonal elements for which a hypothetical reference pose can be established for all possible combinations. Figure 3-11 shows results using the aligned carbonic anhydrase family from the SB2010 testset.¹³ Here, pose identification accuracy was determined across the 29 x 29 matrix using two FPS_{VDW+ES} scoring schemes (Figure 3-11b,c), to rerank ensembles of poses generated by docking each ligand into each receptor, for comparison with the standard DCE_{VDW+ES} method (Figure 3-11a). It is important to note that in these experiments only the number of scoring failures (green), and thus actual success rates (blue), will be affected. Sampling failures and/or incomplete growth (red and white elements) do not change depending on which function is used

corresponding element for footprint-based scoring to the receptor contained within that element. Not surprisingly, this protocol yields the highest overall success rates (Figure 3-11c, blue entries), which serves to confirm the overall robustness of the footprint procedure, although in practice using a unique reference for each ligand is somewhat artificial. Nevertheless, the progressive increase in total matrix success (% coverage of blue squares) in going from DCE_{VDW+ES} , (17.8%), to FPS_{VDW+ES} using diagonal references (45.4%), to FPS_{VDW+ES} using unique references (70.9%), demonstrates utility of the method for identification of specific binding patterns. Additional virtual screening tests as described below provide further support.

As an additional visual point of reference, Figure 3-12 shows the molecular footprints for the cognate diagonal entries of the carbonic anhydrase family which for clarity consist of only the most significant (favorable or unfavorable) interactions. Notably, the significant commonalities in the overlaid cognate footprints emphasize the similar types of interactions made by this group of inhibitors (sulfonamides and related compounds) in the carbonic anhydrase binding site. In particular, the strong interactions between zinc (residue Z), both positive (VDW) or negative (ES), are well-conserved across all inhibitors. Importantly, the plot derived from these crystallographic references provides strong evidence that FPS pharmacophoric patterns are a reproducible property and are thus encoding potentially useful information. Deng et al.¹¹³ came to a similar conclusion that bit-strings generated with the SIFt method encode useful patterns based on an analysis of 89 kinase-inhibitor complexes. In Figure 3-12, it should also be emphasized that each diagonal matrix entry represents a separate structure deposited with the PDB thus the receptor length and/or sequences may not be identical (insertions, deletions, missing residues, etc) despite the fact they are all the same protein. To facilitate visualization of multiple receptors together, a protocol incorporating ClustalW¹³¹ multi-

sequence alignments was developed and the positions labeled X in Figure 3-12 represent amino acids not conserved across the 29 PDB entries. Besides visualization, the alignment protocol also provides a convenient way to generate multi-receptor (i.e. average) footprints which could also be used for FPS scoring.

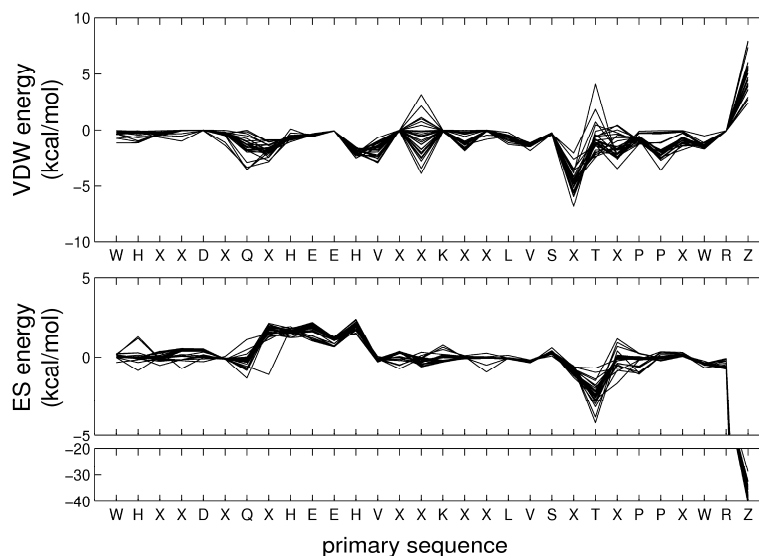


Figure 3-12. Cognate protein-ligand footprints for the aligned carbonic anhydrase family. Residue X indicates a given residue is not conserved across all crystal structures from the PDB entries in terms of amino acid sequence or signifies a substitution or deletion.

3.4.7 Database Enrichment.

The last group of experiments to characterize FPS scoring involves database enrichment. Figure 3-13a-c and Table 3-6 shows enrichment results for three representative systems, neuraminidase, trypsin, and EGFR, taken from the DUD database.^{106,108} Here, docking was first performed using the grid-based DOCK protocol described in Methods prior to rank-ordering using DCE_{VDW+ES} , FPS_{VDW+ES} , FPS_{VDW} , and FPS_{ES} functions. In the present studies, 100% of actives and >96% of decoys produced a viable docked pose. Figure 3-13 shows standard receiver operator characteristic (ROC) enrichment curves while Table 3-6 lists corresponding

area under the curve (AUC) results along with fold enrichment (FE) values computed from the total (FE_{tot}), top 10% (FE_{top}), and bottom 10% (FE_{bot}) of the ROC curves. It is important to note that unlike other analysis, ROC curves inherently include use of the entire range of FPS score cutoffs. A rank-ordered list (FPS scores with associated molecules) is analyzed by continuously varying the score cutoff from best (zero or few molecules retained) to worst (all molecules retained) score.

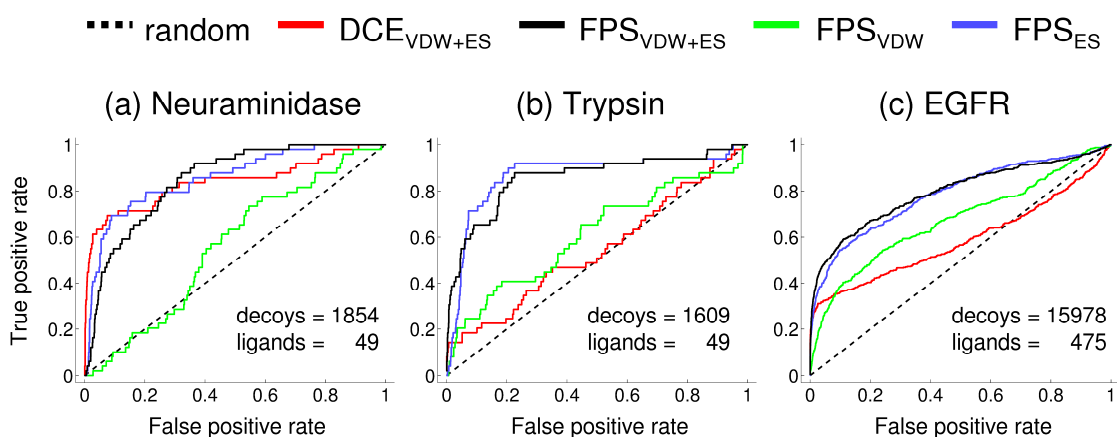


Figure 3-13. ROC enrichment curves for (a) neuraminidase, (b) trypsin, and (c) EGFR using different ranking methods.

Table 3-6. Area under the curve (AUC) and accompanying fold enrichment (FE) statistics from receiver operator characteristic (ROC) plots for three protein-ligand systems.

		AUC_{tot}^a	$FE_{tot} = AUC_{tot} / AUC_{tot, rand}$	$FE_{top} = AUC_{top} / AUC_{top, rand}$	$FE_{bot} = AUC_{bot} / AUC_{bot, rand}$
Neuraminidase 1A4G	Random	0.50	1.00	1.00	1.00
	DCE _{VDW+ES}	0.84	1.68	11.88	1.05
	FPS _{VDW+ES}	0.85	1.69	6.32	1.06
	FPS _{VDW}	0.56	1.12	0.64	1.04
	FPS _{ES}	0.86	1.71	9.04	1.06
Trypsin 1BJU	DCE _{VDW+ES}	0.55	1.09	3.18	1.01
	FPS _{VDW+ES}	0.86	1.71	<u>9.65</u>	1.04
	FPS _{VDW}	0.61	1.22	3.50	0.96
	FPS _{ES}	0.87	1.73	8.23	1.03
EGFR 1M17	DCE _{VDW+ES}	0.59	1.18	6.29	0.97
	FPS _{VDW+ES}	0.79	1.59	<u>9.21</u>	1.03
	FPS _{VDW}	0.67	1.35	4.89	1.03
	FPS _{ES}	0.78	1.57	8.20	1.03

^a AUC_{tot} is 100% of the database, AUC_{top} is the top 10%, and AUC_{bot} is the bottom 10%

Visually, the standard DCE_{VDW+ES} (Figure 3-13 red lines) and FPS_{VDW+ES} (Figure 3-13 black lines) rankings yield initial steep upwardly sloping ROC curves for all systems which is an indication of "early enrichment" compared to random (Figure 3-13a, dashed line). The original DUD paper employing DOCK3.5¹⁰⁸ similarly obtained early strong enrichments for neuraminidase and trypsin although differences in sampling and scoring protocols between the two studies make a direct comparison here difficult. Interestingly, the FPS_{VDW+ES} ROC curves show enrichment is maintained throughout the entire database ranking (Figure 3-13 black lines) in contrast to DCE_{VDW+ES} which show degradation, in the case of trypsin and EGFR, as increasingly larger percentages of each database are examined (Figure 3-13 red lines). ROC curves derived using FPS_{VDW} or FPS_{ES} methods suggest in some systems better enrichment may be obtained using only a single descriptor. For example, ES-based rankings alone show strong enrichment for neuraminidase in comparison to VDW which is essentially random (Figure 3-13a blue vs green lines). This finding is physically reasonable considering the highly-charged neuraminidase binding site and consistent with an earlier study from our laboratory in which the best correlation with experimental binding free energies was obtained using the electrostatic component from MM-GBSA calculations.⁸³

From a more quantitative standpoint, fold enrichment statistics using FPS_{VDW+ES} rankings reveal > 9-fold enrichment over random for trypsin ($FE_{top} = 9.65$) and EGFR ($FE_{top} = 9.21$) in the critically important top 10% region of the ROC curve space (Table 3-6 underlined entries). For neuraminidase in this region, DCE_{VDW+ES} performs best ($FE_{top} = 11.88$) followed by the previously mentioned electrostatic term FPS_{ES} ($FE_{top} = 9.04$) and finally FPS_{VDW+ES} ($FE_{top} = 6.32$). With one exception, FPS_{vdw} for neuraminidase ($FE_{top} = 0.64$), footprint similarity rankings always lead to significant early fold enrichment versus random (1.00). Good

enrichment has also been reported by the groups using related computational methods which encode binding interaction patterns.^{113,115,117,118,122} For example, Deng et al.¹¹³ (see Table 1 in the reference) reported use of the SIFt method led to better enrichment, than two other scoring methods considered, for the identification of 16 known p38 inhibitors out of a database of 1000 decoys. Likewise, ROC curves reported by Marcou et al.¹¹⁷ (see Figure 8 in the reference) revealed that use of interaction fingerprints led to stronger enrichment, than other tested scoring functions, using a database of 19 actives and 22,230 decoys.

Focusing on the EGFR system, Figure 3-14 shows differences in the ensemble of docked compounds chosen using either DCE_{VDW+ES} (14a) or FPS_{VDW+ES} (14b) scoring. The top panel in Figure 3-14 shows overlaid poses representing the top 50 (green = best) or bottom 50 (gray = worst) ranked compounds in relationship to the molecular surface envelope derived from the crystallographic pose of the known drug erlotinib (red surface). The bottom panel in Figure 3-14 shows corresponding molecular weight (MW) histograms for top (green) and bottom (red) ranked ensembles with the number of compounds increased to 100. It is immediately apparent that DCE_{VDW+ES} scoring leads to MW bias due to the fact that the molecular mechanics-based additive function increases proportionally with ligand size.¹³² The ensemble of top-ranked compounds in Figure 3-14a yield significantly larger ligands (green molecules and MW curve) which, in this example, do not appear to fit as well in the molecular surface envelope of erlotinib as bottom ranked compounds which are smaller (gray molecules and MW curve). In sharp contrast, when FPS score rankings are employed using erlotinib as a reference, the 50 top ligands fit the molecular surface envelope almost perfectly (Figure 3-14b top panel). Further, MW bias of top-ranked ligands here does not favor size but instead favors MW similar to that of the reference (erlotinib = 393.44 g/mol) as shown by the large green MW peak in Figure 3-14b

(bottom panel). Interestingly, the top-ranked molecules using FPS are somewhat smaller on average than erlotinib (ca 340 vs 393 g/mol) which, for this example, is likely a function of the composition of the DUD database. Bottom ranked ligands show no particular bias and are spread throughout the entire MW range (Figure 3-14b gray line). Overall, the current FPS implementation appears to yield targeted, understandable, and robust enrichments.

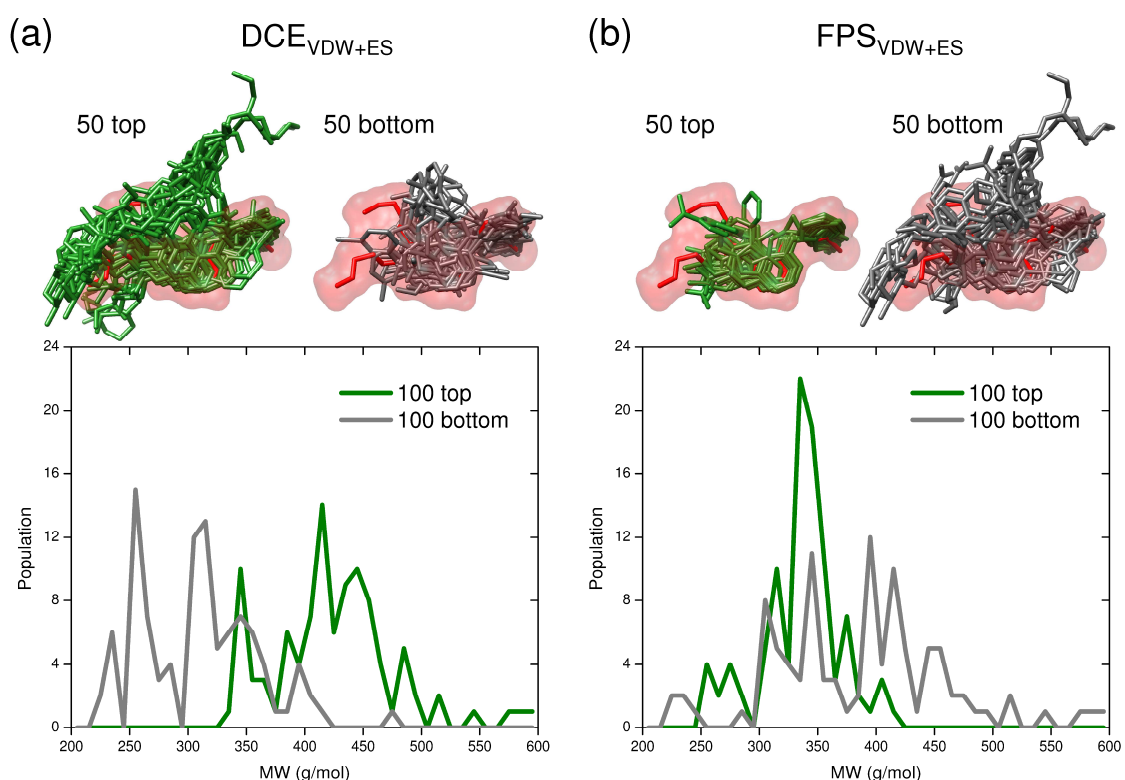


Figure 3-14. Graphical representation of the 50 top and 50 bottom ranked poses obtained from docking the 475 active ligands from the DUD EGFR database and using (a) DCE_{VDW+ES} and (b) FPS_{VDW+ES} scoring functions. The reference (erlotinib) is shown in red surface with top ligands in green and bottom ligands in gray. On the bottom are corresponding histograms of molecular weight (MW) for the 100 top (best) and 100 bottom (worst) ranked molecules. Note that the large MW peak at ca. 340 for the 100 best scoring molecules using FPS_{VDW+ES} corresponds ca. to the MW of the erlotinib reference (393.44 g/mol).

As a final example of the potential utility of FPS scoring, Figure 3-15 shows representative virtual screening results for EGFR, derived from docking and rescoring of 906,914 commercially available compounds from the ZINC database¹⁴ (Chemdiv vendor), to

ascertain how many compounds would be identified which make interactions similar to that of erlotinib at a given FPS score cutoff. Although the number of compounds identified at any score cutoff is likely to be system dependent, and a function of which database is screened and which reference molecule is employed, the results in Figure 3-15 suggest a reasonable number of molecules (i.e. 25 to 201 molecules) can readily be identified out of ca. 1 million compounds using a score cutoff range of 0.8-0.9. Similar to the results obtained in the DUD example above (Figure 3-14b), the graphic in Figure 3-15 highlights significant pose overlap in docked geometries for the 25 compounds obtained using a 0.8 cutoff from the virtual screen, which fit well into the molecular surface envelope defined by the reference erlotinib (red).

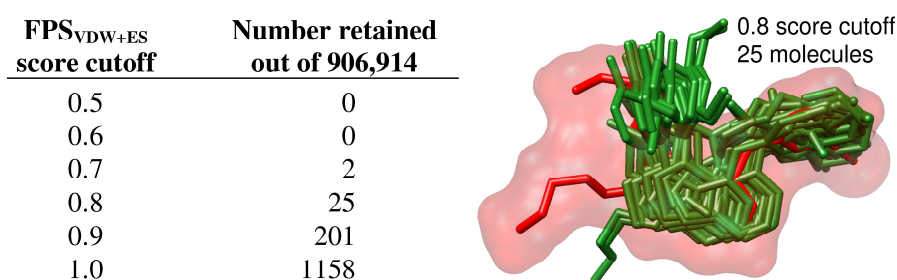


Figure 3-15. Number of molecules retained from a virtual screen of 906,914 molecules to EGFR using various FPS_{VDW+ES} score cutoff values. The graphic shows the 25 molecules identified (green) using a cutoff of 0.8 in comparison with the known drug erlotinib (red) which was used as the footprint reference.

3.5 Conclusions

The primary goal of this study was to introduce and evaluate a new DOCK scoring function, termed footprint similarity (FPS) score which employs per-residue interaction maps (footprints) to derive a binding site comparison metric between any two molecules. From a practical standpoint, FPS scoring facilitates rapid identification of ligands whose binding interaction patterns resemble that of a reference molecule used as an input query. Thus, the method may find utility in a variety of structure-based drug design scenarios. Potentially useful

outcomes include identification of ligands which make footprints similar to known drugs or inhibitors, native substrates or cofactors, transition states, or side-chains which mediate protein-protein interactions (Table 3-1). Identification of ligands with footprints similar to a known reference but based on novel chemotypes could facilitate scaffold hopping. And, identification of ligands having footprints which do not rely on residues that are prone to mutation could enable development of inhibitors with enhanced resistance profiles.

Several FPS score types were evaluated in this study (Table 3-2) which employed footprints based on intermolecular van der Waals energies (FPS_{VDW}), Coulombic energies scaled by a distance dependent dielectric constant (FPS_{ES}), and hydrogen bond energies (FPS_{HB}). Combination scores constructed from two (FPS_{VDW+ES}) or three ($FPS_{VDW+ES+HB}$) footprint types were also evaluated. Footprint similarities were quantified using standard Euclidean Distance, normalized Euclidean Distance, standard Pearson Correlation, and threshold Pearson Correlation metrics (Tables 3-2 and 3-3) and functional relationships between these methods were examined (Figure 3-6). Results using the different FPS protocols were compared with those obtained using the standard DOCK Cartesian energy scoring function (DCE_{VDW+ES}) on tests designed to primarily assess accuracy of (1) pose identification and (2) database enrichment using cognate ligands from crystallographic complexes deposited in the PDB as references. To facilitate comparison with the work presented here, should other groups wish to evaluate their interaction-based functions and/or docking codes, the datasets for pose identification and crossdocking are available from the SB2010¹³ website (<http://rizzolab.org>) and for enrichment from the DUD¹⁰⁸ website (<http://dud.docking.org>).

With one exception, all FPS protocols yielded improved pose identification success, using three large datasets (680-775 systems) to assess accuracy, relative to using comparable

DCE methods (Table 3-3). Overall, the FPS_{VDW+ES} function in combination with normalized Euclidian distance yielded the best results (Table 3-3). Success using FPS_{VDW+ES} , defined as the pose being $\leq 2.0 \text{ \AA}$ from experiment, showed increases over analogous DCE_{VDW+ES} scores by ca 10%, 6%, and 12% using rigid (RGD = 775), fixed anchor (FAD = 748), and flexible (FLX=680) perfect sampling subsets derived from the SB2010¹³ database. Additional tests, using ensembles derived from crossdocking (29 ligands to 29 receptors) showed significantly greater success (matrix coverage) using two different FPS protocols (45.4% and 70.9%) compared with DCE (17.8%) for a challenging carbonic anhydrase family (Figure 3-11).

A close examination of results (Table 3-4, Figure 3-7) classified as false positive (good FPS score and bad rmsd) revealed in many cases, poses that were both geometrically and chemically correct (Figures 3-8 and 3-9) and that misclassifications can result due to deficiencies with how current DOCK pair-wise rmsd routines handle symmetry. The results indicate the reported success rates for pose identification are in fact a lower bound on the potential accuracy of the calculations. A related issue, in which otherwise well-overlaid ligands showed rmsds $> 2.0 \text{ \AA}$ rmsd was traced to differences only in solvent exposed moieties (Figure 3-8) which for many cases would not reasonably be considered a failure. Examination of false negatives (bad FPS score and good rmsd, Table 3-5) revealed in some cases that small variations in pose geometry can yield larger than expected differences in energy (Figure 3-10), especially for interactions involving charged groups and/or polar hydrogens, which highlights the need for care when preparing a reference.

Area under the curve (AUC) and fold enrichment (FE) statistics (Table 3-6) derived from receiver operator characteristic (ROC) curves (Figure 3-13), for three representative systems from the DUD database,¹⁰⁸ reveal significant fold enrichment using FPS_{VDW+ES} (neuraminidase =

6.32, trypsin = 9.65, and EGFR = 9.21) compared to random (1.00) in the most critical early regime (top 10%) of the ranked databases. In two out of three cases, the FPS_{VDW+ES} enrichment exceed those obtained using the standard DOCK DCE_{VDW+ES} scores (Table 3-6). Close inspection of EGFR results reveals DCE_{VDW+ES} scoring leads to top-ranked molecules not well-accommodated in the molecular surface envelope defined by the cognate ligand erlotinib and have a distinct MW bias towards larger molecules (Figure 3-14a). In sharp contrast, top-ranked molecules using FPS_{VDW+ES} using erlotinib as the footprint reference lead to poses which nicely fit within the binding envelope and have a MW biased towards the reference (Figure 3-14b). Finally, the potential utility of the method for identification of novel compounds was demonstrated by a representative virtual screen to EGFR. On-the-fly flexible ligand docking of ca 1 million compounds obtained from ZINC,¹⁰⁸ followed by FPS_{VDW+ES} re-ranking using erlotinib as a reference (Figure 3-15), yielded a reasonable number of compounds (25-201) with good FPS scores (0.8-0.9) available for purchase. Taken together, the results of this comprehensive study strongly suggest the implementation of footprint-based comparison methods into DOCK will have utility for structure-based design. A future goal, based on studies in progress, is to incorporate molecular footprints with de novo design methods to bias construction of new ligands from scratch towards that of a reference.

Chapter 4. Grid-based Molecular Footprint Comparison

Method for Docking.

This chapter contains results which are to be incorporated in a manuscript currently in preparation.

TEB, SM and RCR designed research plan. TEB performed research, analysis and is writing the initial draft.

4.1. Introduction

In drug design, it is often convenient to employ prior knowledge such as using a known drug or substrate as a reference. In Chapter 3, we showed that Footprint Similarity (FPS) score, when used as a post-docking rescoring tool, is a powerful approach to identify ligands that bind similarly to a reference molecule. However, applying FPS score, which was previously calculated using Cartesian space receptors, to *de novo* design techniques or to standard docking was computationally restrictive due to the time cost. Many docking programs, including DOCK, use grids to speed up energy calculation from $O(M*N)$ time depending on the size of the receptor (M) and the ligand(N) to $O(N)$ time which only depends on the size of ligand.^{40,133}

We envision that footprint-guided sampling, using a reference, can be used in at least three important ways: (1) pose identification; (2) virtual screening and (3) *de novo* design. For pose identification, chemically similar ligands may be docked to determine the binding mode using a reference molecule with a known pose. For virtual screening (molecule libraries) and *de novo* design (fragment libraries) the objective is to identify new molecules that bind to a receptor like known molecules, but are chemically different. In this work, we demonstrate that grids can be used to speed up footprint calculations, we describe a generalization of the FPS scoring method that utilizes grids, and we show a proof-of-concept that grid-based footprints can be used to guide docking using the SB2010 pose identification test set.¹³

4.2. Theoretical Methods

Grid-based Footprints. Here we show that we can extend Cartesian-footprints described in Chapter 3 to grid-based footprints. The standard molecular mechanics energy function is pair-wise additive. Therefore, we can separate the non-bonded interaction energy into per-residue components. Grids for each residue are calculated using the standard energy function. The through space interactions between receptor and ligand can be written as follows:

$$E_c = \sum_{i \in L} \left(\sqrt{A_i} \sum_{j \in R} \frac{\sqrt{A_j}}{r_{i,j}^a} - \sqrt{B_i} \sum_{j \in R} \frac{\sqrt{B_j}}{r_{i,j}^b} + 332q_i \sum_{j \in R} \frac{q_j}{Dr_{i,j}} \right) \quad (4-1)$$

Here, i and j are indexes for the ligand L and receptor R , respectively. E_c is the Cartesian energy, D is the dielectric function (often a distance dependent function is used, $D = 4r$), $r_{i,j}$ is the distance between atoms i and j , and q_i is the charge at atom i . The value 332.0 converts the electrostatic (ES) energy into kcal/mol. The van der Waals (VDW) terms are generalized to work with any VDW exponents (a and b):

$$A_i = \varepsilon \left(\frac{b}{a-b} \right) (2R_i)^a, \quad B_i = \varepsilon \left(\frac{b}{a-b} \right) (2R_i)^b$$

A_i and B_i are the VDW parameters where ε is well depth and R is the minimum distance r in the well. Fast implementation of energy calculations during docking can be achieved by pre-computing the interactions and storing the potential energy on a grid. Every point p on the grid (gd) has three values: attractive VDW (gd_{a_vdw}), repulsive VDW (gd_{r_vdw}) and ES energies (gd_{es}):

$$gd_{a_vdw}(p) = \sum_{j \in R} \frac{\sqrt{A_j}}{r_{p,j}^a}, \quad gd_{r_vdw}(p) = \sum_{j \in R} \frac{\sqrt{B_j}}{r_{p,j}^b}, \quad \text{and} \quad gd_{es}(p) = 332 \sum_{j \in R} \frac{q_j}{Dr_{p,j}}$$

Grids are energy potentials; one can think of each grid point (p) as the interaction of the receptor with a dummy atom (where $A_i = 1$, $B_i = 1$ and $q_i = 1$). The through-space interactions can be approximated by interpolating grid values onto the ligand atoms (a_i) located at some point between grid points. The grid energy (E_g) calculated by Eq. 2 is an approximation of the Cartesian energy (E_c) shown in Eq. 1 ($E_c \approx E_g$). The finer the grid spacing, the better the approximation of the Cartesian energy.

$$E_g = \sum_{i \in L} \left(\sqrt{A_i} TLI(\vec{g}_{a_vdw}, a_i) - \sqrt{B_i} TLI(\vec{g}_{r_vdw}, a_i) + 332 q_i TLI(\vec{g}_{es}, a_i) \right) \quad (4-2)$$

Here, the function TLI performs a tri-linear interpolation of each of vectors:

$$\vec{g}_{a_vdw} = [gd_{a_vdw}(p_1), \dots, gd_{a_vdw}(p_8)]$$

$$\vec{g}_{r_vdw} = [gd_{r_vdw}(p_1), \dots, gd_{r_vdw}(p_8)]$$

$$\vec{g}_{es} = [gd_{es}(p_1), \dots, gd_{es}(p_8)]$$

The TLI is a linear function of the form:

$$TLI(\vec{g}, a) = TLI(\vec{g}_1 + \vec{g}_2, a) = TLI(\vec{g}_1, a) + TLI(\vec{g}_2, a) \quad (4-3)$$

The TLI combines the potentials stored using the eight closest grid points (p_1, \dots, p_8) and the position of atom i to approximate the Cartesian-based energy at atom i .

We can obtain a per-residue decomposition of the grid calculation by generating a grid for each residue as is shown for a single grid point (p) here:

$$gd(p) = \sum_{l \in R} E_c(p, l) = \sum_{i \in [1, M]} \sum_{l \in S_i} E_c(p, l) = \sum_{i \in [1, M]} gd_{S_i}(p) \quad (4-4)$$

Here R is the set of receptor atoms and S_i is the set of atoms in residue i . The receptor has N atoms and M residues. The union of the residue atom sets is the set of receptor atoms:

$$\bigcup_{i=1}^M S_i = R$$

And the sum of the size (number of atoms) of the residues is the size of the receptor:

$$\sum_{i=1}^M |S_i| = N$$

We calculate the per-residue components ($E_{g,S(i)}$) by using the grids generated for each receptor residue ($G_{S(i)}$). Moreover, by Equations 4-2 to 4-4, the sum of these components add up to the standard grid energy (Eq. 4-5).

$$E_g = \sum_{i \in [1, M]} E_{g,S_i} \quad (4-5)$$

4.3 Computational Details.

Residue selection and grid generation. A threshold based criteria was used to identify important residues, based on their standard DOCK Cartesian energy (DCE) with the optimized (H-opt) ligand crystallographic pose, unless otherwise stated. H-opt structure generation is described in Chapter 3. Residues with absolute interaction energies exceeding 1.0 kcal/mol for VDW or 0.5 kcal/mol for ES were selected for residue grid generation. All other residues were grouped together to create a remainder grid. For each receptor, grids were generated for the selected residues and for the remainder. An in-house python script iteratively calls the grid

program for each selected residue and the remainder. For this study, all grids are generated at 0.4 Å resolution. The grids were calculated using 6-9 VDW exponents for a softer receptor to somewhat mimic receptor flexibility.¹³⁴

Pose reproduction experiments. The SB2010 dataset was used.¹³ Six docking runs were performed using different random seeds. The success rates were calculated for each run and the averages were used as a means to better gauge docking success. As previously discussed,¹³ docking experiments may have one of three outcomes: (1) docking success is when the program selected a correct posed (within 2 Å of native pose), (2) scoring failure occurs when the correct pose was sampled but it was not scored at the top of the list, and (3) sampling failure is when docking did not sample the correct pose.

Standard Docking and Rescoring. Docking here uses the same protocol as discussed in Mukherjee et al.¹³ One modification was that final clustering parameter used was changed from 2.0 Å to 0.5 Å; this affects sampling and rescoring. DOCK uses a best-first clustering method. The rescoring with DOCK Cartesian energy (DCE) and FPS used the 6-9 VDW exponents as were used for the grid calculations. This makes comparisons easier between the DOCK grid energy (DGE) and DCE scores.

Grid-based Footprint Similarity Score docking. As described in Chapter 3, standard Euclidean, normalized Euclidean, and Pearson correlation are three methods employed to calculate Grid-based FPS score. *Footprint-guided docking* experiments employs solely Grid-based FPS score (using the three footprint comparison methods) to “guide” growth, meaning that anchor-and-grow prunes, clusters, and minimizes growth using FPS score. *Footprint-nudged docking* experiments use the standard grid score “nudged” by the FPS score, meaning that anchor-and-grow prunes, clusters, and minimizes growth using FPS score combined with grid

score. Both “guide” (coefficients C1 and C2 are zero) and “nudged” (coefficients C1 and C2 are equal to one) uses eq 4-6.

$$\text{Total Score} = C1 * DGE_{ES} + C2 * DGE_{VDW} + C3 * FPS_{ES} + C4 * FPS_{VDW} \quad (4-6)$$

The coefficients used by the score (eq 4-6) are list in Table 4-1 for the different Grid-based FPS methods.

Table 4-1. Footprint Similarity Score docking coefficients.

	Footprint-guided docking	Footprint-nudged docking
Standard Euclidean ^a	C1 = C2 = 0 C3 = C4 = 1	C1 = C2 = 1 C3 = C4 = 1
Normalized Euclidean	C1 = C2 = 0 C3 = C4 = 1	C1 = C2 = 1 C3 = C4 = 20
Pearson correlation	C1 = C2 = 0 C3 = C4 = -1	C1 = C2 = 1 C3 = C4 = -20

^a for equations 4-6

4.4 Results and Discussion

4.4.1 Grid Generation.

Grid vs Cartesian energies. As discussed above, Cartesian per-residue energy decomposition can be approximated by using multiple energy grids. Per-residue energy decomposition is accomplished by the creation of a grid for each protein residue. For example the smallest receptor in SB2010, consisting of 80 residues, is pdb code 2PK4, plasminogen kringle-4 protein which is an important protein in blood clotting,¹³⁵ and a comparison of grid-based (grey filled circles) and Cartesian (black squares) footprints are shown in Figure 4-1. We see that the two energy footprints are very similar. However, for the grid-based method, for most systems, it is infeasible to consider all residues in the FPS comparison calculation; the grids take up a large amount of memory during runs. In addition, the time needed to perform the docking

calculations increases as more grids are used. As is shown in Figure 4-2a, the 0.4Å resolution grids well approximate the electrostatic interactions ($r = 1.000$) but does a slightly poorer job for the VDW footprint ($r = 0.994$). Interestingly, preliminary studies showed that finer grids do not in fact result in better docking success. This observation may be a consequence of the fact that coarser grids smooth the underlying energy landscape allowing better sampling.

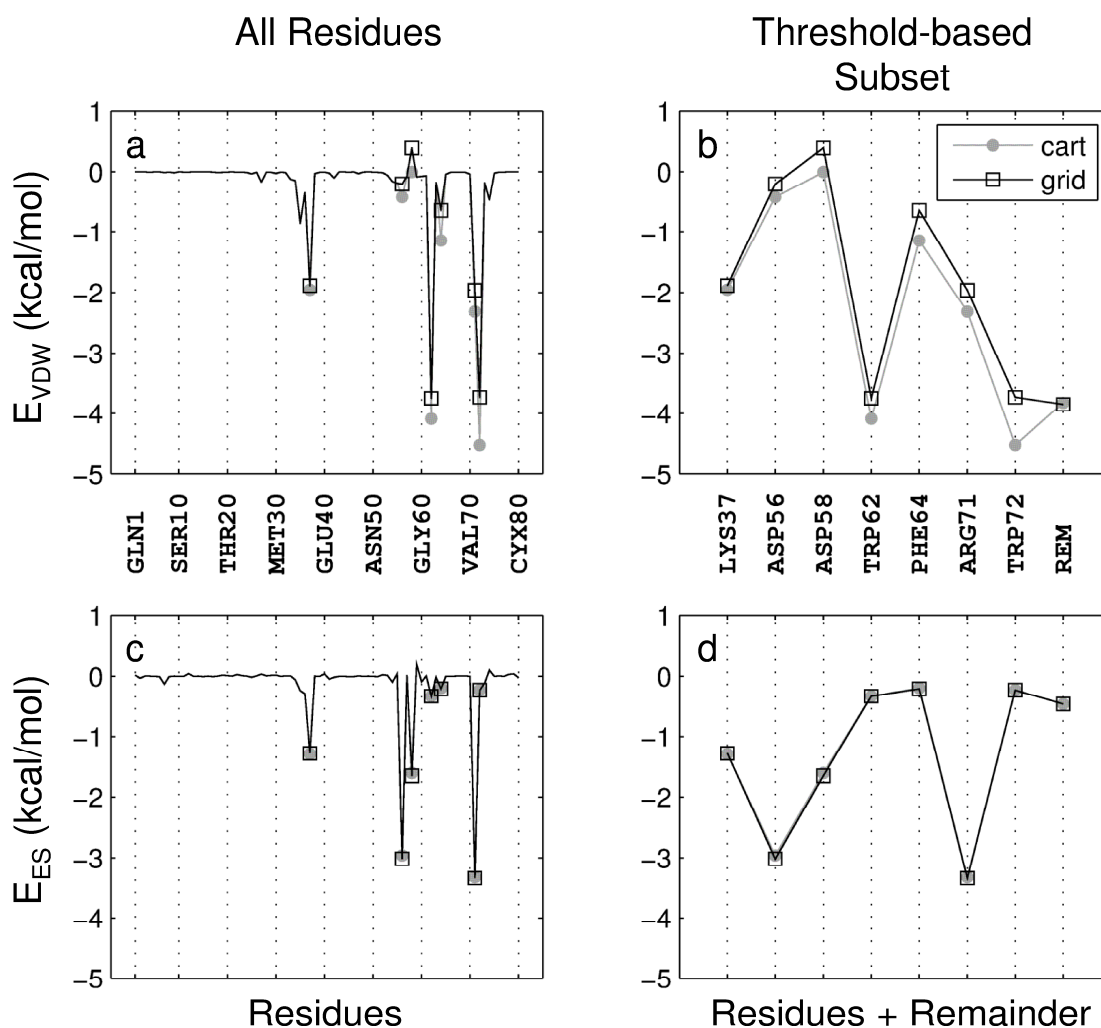


Figure 4-1. Comparisons between grid-based (grey) and Cartesian (black) footprints. The van der Waals (a and b) and electrostatics (c and d), all-residues footprints (a and c), and the threshold-based residues plus remainder footprints (b and d) are shown for pdb code 2PK4.

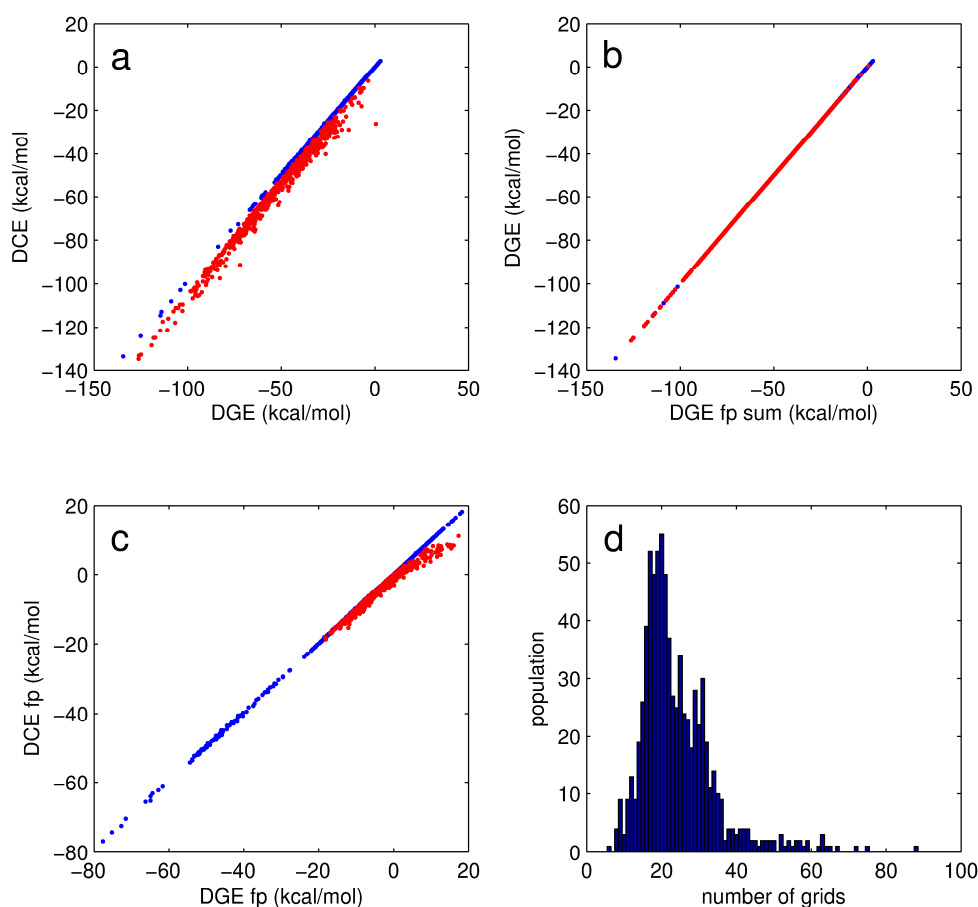


Figure 4-2. (a) comparison of Grid (DGE) score vs Cartesian (DCE) Score. (b) Comparison of Sum of the footprint grid-values vs standard grid score. (c) Comparison of the Grid-based footprint values vs (d) Histogram of the number of grids used for footprint-guided docking (# grids = # residues + 1). For panel A-C, blue dots are ES and red are VDW.

Remainder vs no remainder. In Chapter 3, we described a threshold-based method footprint correlation comparison which used no remainder where we used the union of the two sets (pose and reference) and the residues selection is calculated for every pose on the fly. For the grid-based footprint method, we must pre-compute the grids used to describe the footprint using only the residue selection for the reference since the candidate pose interactions will only be for those residues originally deemed important for the reference. A problem could arise if neglected residues would interact strongly with the candidate pose. For example, if a reference makes strong interactions with 12 residues and a generated pose interacts with 11 and the

intersection between these sets is 9 residues then the docked pose makes interactions with two residues not considered in the FPS calculation. To alleviate this drawback, we will include a "remainder-grid" consisting of all the remaining residues (those without their own grids). Although these could be considered non-important residues, their inclusion ensures that their sum adds up to the total grid energy of the protein. This is demonstrated by the essentially perfect correlation shown in Figure 4-2b, a poorer correlation would be expected if remainders were not included. Importantly, excellent agreement is seen between the DGE and DCE footprints ($r = 1.000$, $r = 0.988$) as is shown in Figure 4-2c. As is the case in the DGE and DCE scores, there is also poorer agreement between the VDW footprints compared with ES footprints. On average 20 grids per receptor are used Figure 4-2d.

Table 4-2. Standard docking and rescoring. FPS scoring uses normalized Euclidian.

	docking	function	residues	FPS type	success ^a (%)	failures (%)		time (min)
						scoring	sampling	
a	Stand.	Grid	all	--	63.6	25.9	10.5	9.97
b	rescore ^b	Grid-FPS	sel+rem	Eucl	79.4	10.1	10.5	0.23
c	rescore ^b	Grid-FPS	sel+rem	Norm	79.0	10.4	10.5	0.21
d	rescore ^b	Grid-FPS	sel+rem	Corr	78.5	11.0	10.5	0.18
e	min ^b	Cart	all	--	63.4	21.5	15.2	15.22
f	rescore ^c	Cart-FPS	all resid	Norm	79.0	5.9	15.2	0.09
g	rescore ^c	Cart-FPS	thres	Norm	77.9	7.0	15.2	0.11
h	rescore ^c	Cart-FPS	sel	Norm	78.5	6.3	15.2	0.10
i	rescore ^c	Cart-FPS	sel+rem	Norm	78.8	6.0	15.2	0.10
j	Stand.	Grid-Sum	sel+rem		64.3	25.6	10.1	41.16
k	Guided	Grid-FPS	sel+rem	Eucl	77.5	9.4	13.1	38.72
l	Guided	Grid-FPS	sel+rem	Norm	30.2	13.5	56.3	38.13
m	Guided	Grid-FPS	sel+rem	Corr	20.5	12.2	67.3	37.77
n	Nudged	Grid-FPS	sel+rem	Eucl	81.0	11.1	7.9	41.49
o	Nudged	Grid-FPS	sel+rem	Norm	79.8	11.0	9.3	43.18
p	Nudged	Grid-FPS	sel+rem	Corr	72.2	17.5	10.3	42.25

^a N = 780, average of 5 runs. ^b standard docking results (row a) results rescored or minimized with alternative scoring function. ^c rescoring the Cartesian minimization results (row e).

4.4.2 Docking and Rescoring in Pose Reproduction.

Table 4-2 shows statistics (docking success, scoring failure, and sampling failure) for numerous experiments (SB2010, N = 780) comparing Grid-FPS results with their Cartesian counterparts (Cart-FPS). Comparisons with the standard docking scoring function are also reported. The average outcome of six docking runs are reported for standard DGE docking (Table 4-2, rows a, j); rescoring the standard docking runs with Grid-FPS (Table 4-2, rows b to d), standard DCE (Table 4-2, row e), and Cart-FPS (Table 4-2, rows f to i); and docking calculations in which their sampling is influenced by Grid-FPS scores (row k to p). The grid-FPS is a decomposition of DGE score which is the scoring function used during docking, thus no sampling failures were added (rows b to d). Docking success is comparable between Cartesian minimization, 63.4% (row e) and grid docking to 63.6% (row a), however, a decrease in sampling success (4.7%) occurs upon minimization with sampling failures going from 10.5% (grid, row a) to 15.2% (Cartesian, row e). The FPS all residues method does the best with an average success of 79.0% (row f). Interestingly, when comparing FPS threshold and FPS selection with no remainder, the success increases from 77.9% (row g) to 78.5% (row h), this is most likely due to the inclusion of more electrostatic residues for consideration. There is a further increase in success when using the FPS selection remainder 78.8% (row i); however, FPS selection remainder is still not as good as the FPS for all residues. The sampling failures are the same among the Cartesian results. The “Grid FPS selection remainder” does as well as the Cartesian best method (FPS all residues), with a dock success of 79.0% (row c, f). Because we can use the grid based dock results without minimization the grid based rescoring has no additional sampling failures.

Overall in Table 4-2 the combination of FPS with grid score (termed FPS-nudged) does a better job at pose identification than the standard docking scoring function, or FPS-guided method alone across 3 different footprint comparison methods: Euclidian (row n), normalized Euclidean (row o), Pearson correlation (row p). Figure 4-3 shows the results graphically with each vertical bar representing a result using a different random seed. Importantly, the nudged methods appear to be most accurate in combination with the Euclidian comparison metric (81.0% success rate, row n). Furthermore, Euclidian comparison metric also does very well for Guided docking (77.5% row k) which is due to both magnitude and shape matching. For the FPS-guided results, the poor performance of normalized Euclidean (30.2%, row l) and Pearson correlation (20.5%, row m) is most likely due to the fact that magnitudes are not matched. A lack of magnitude information may be acceptable when only used for rescoring because only the best solutions (lowest energy) would be considered. However, since the FPS scoring function affects sampling during growth (FPS-guided), without consideration of the magnitude, energetically unfavorable poses may be scored favorably in FPS space which can be problematic. However, when grid energy is used in combination with the normalized Euclidean (79.8%) and Pearson correlation (72.2%) both methods do very well and thus the combination methods would be recommended. Importantly, there are only on average 5.84 systems (or 0.7%) that differ between the standard grid (row a) vs standard Grid-Sum (row j) results showing the robustness of a multi-grid energy decompositions strategy. However, since the footprint-guided and footprint-nudged docking methods are for example ~4 times slower than standard docking (when comparing rows j-p with row a, Table 4-2) care must be used in their application.

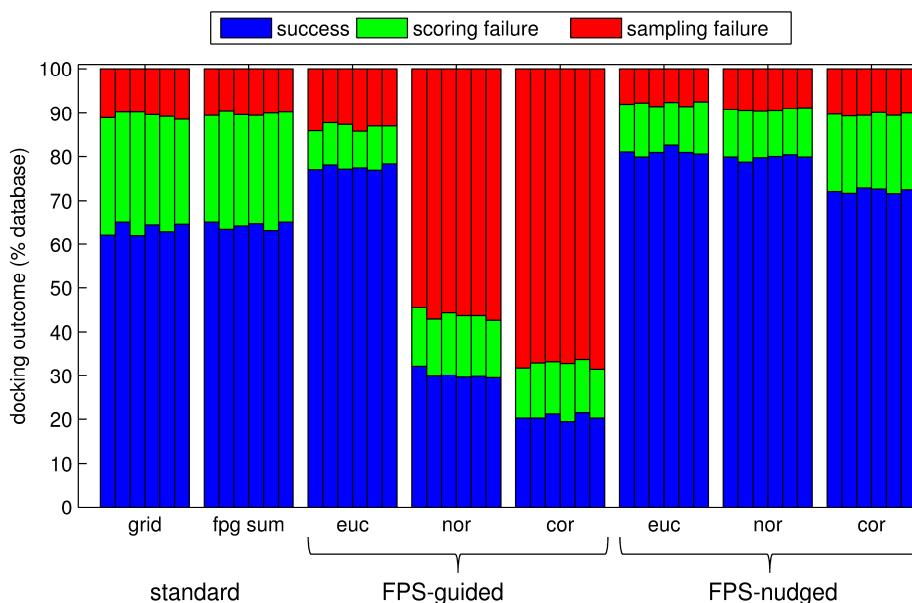


Figure 4-3. Shows six docking experiments using SB2010. Five random seeds were used to gauge variability. Success (blue), scoring failure (green) and sampling failure (red) add up to 100%.

Finally, It should be noted that the approximate functional relationship between normalized Euclidean and Pearson correlation described in Chapter 3 will break down because the mean of the footprints will not be expected to be close to zero since only the peaks that should make strong interactions are included. For this reason normalized Euclidean and Pearson correlation will likely give users different rank orderings.

4.4.2 FPS Behavior During Growth.

To evaluate behavior of FPS (using Euclidian comparison method) as a function of growth step, fixed anchor docking (which begins growth starting from an anchor in the crystallographic position) using a restraint [$k = 10 \text{ kcal}/(\text{mol } \text{Å}^2)$] was performed on the 10 rotatable bond subset ($N = 59$) of the SB2010 testset. Figure 4-4 shows box plots for the 10 RB subset (top) and for a single system, pdb code 1M17 (middle). Both FPS-guided docking (Figure 4-4, left) and rescored standard docking (Figure 4-4, right) was performed. In general, the FPS-

guided results, overall yield better footprint values at the end of the runs than the corresponding rescoring results. Although many of the branches do not converge to a good FPS score, there is a clear downward trend of the median values (Figure 4.4, top panel). In fact, this shows that on average the FPS scoring method is guiding growth toward better footprint similarities. The bottom panel of Figure 4-4 and Figure 4-5 show one branch in which structures and energetics converges to a reasonable native-like answer. The representative 1M17 example shows the system converges to a lower FPS value than the all-system set (N=59). In Figure 4-5, one branch of the growth tree is shown for erlotinib, the structures and the corresponding footprints.

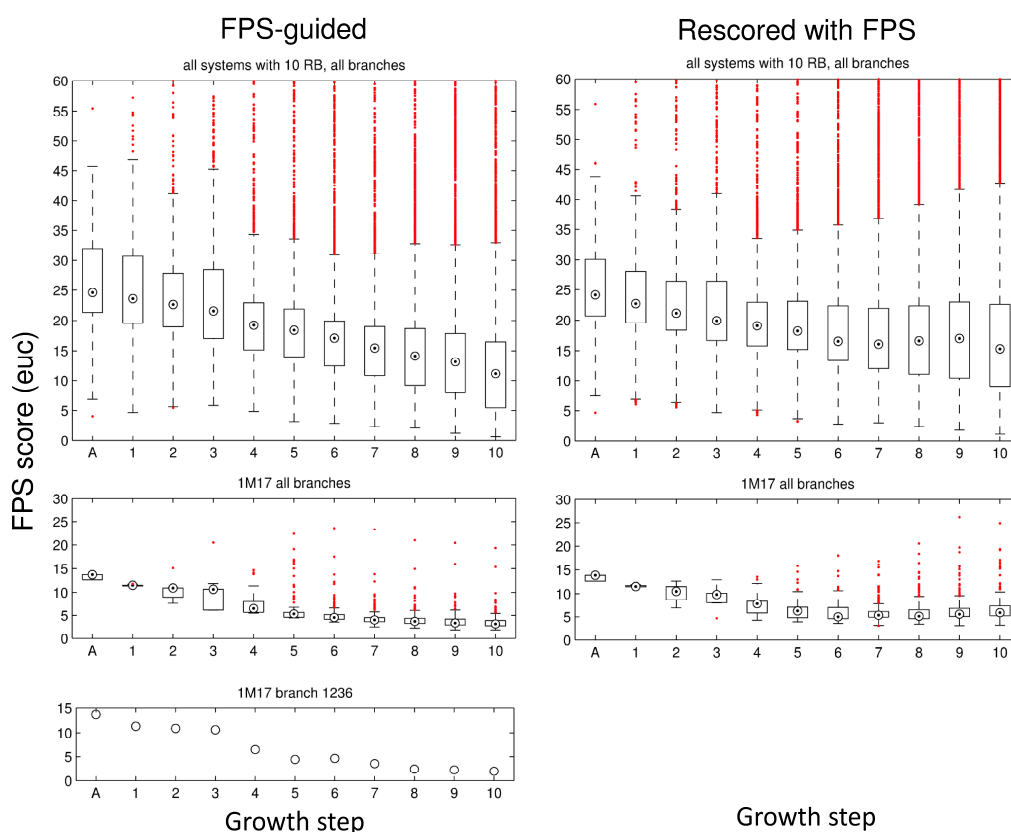


Figure 4-4. FPS score for growth step for the 10 rotatable bond subset of SB2010 (N= 59) are shown. Left panels show FPS guided docking results. Right panels show rescoring results from a standard grid score docking experiment.

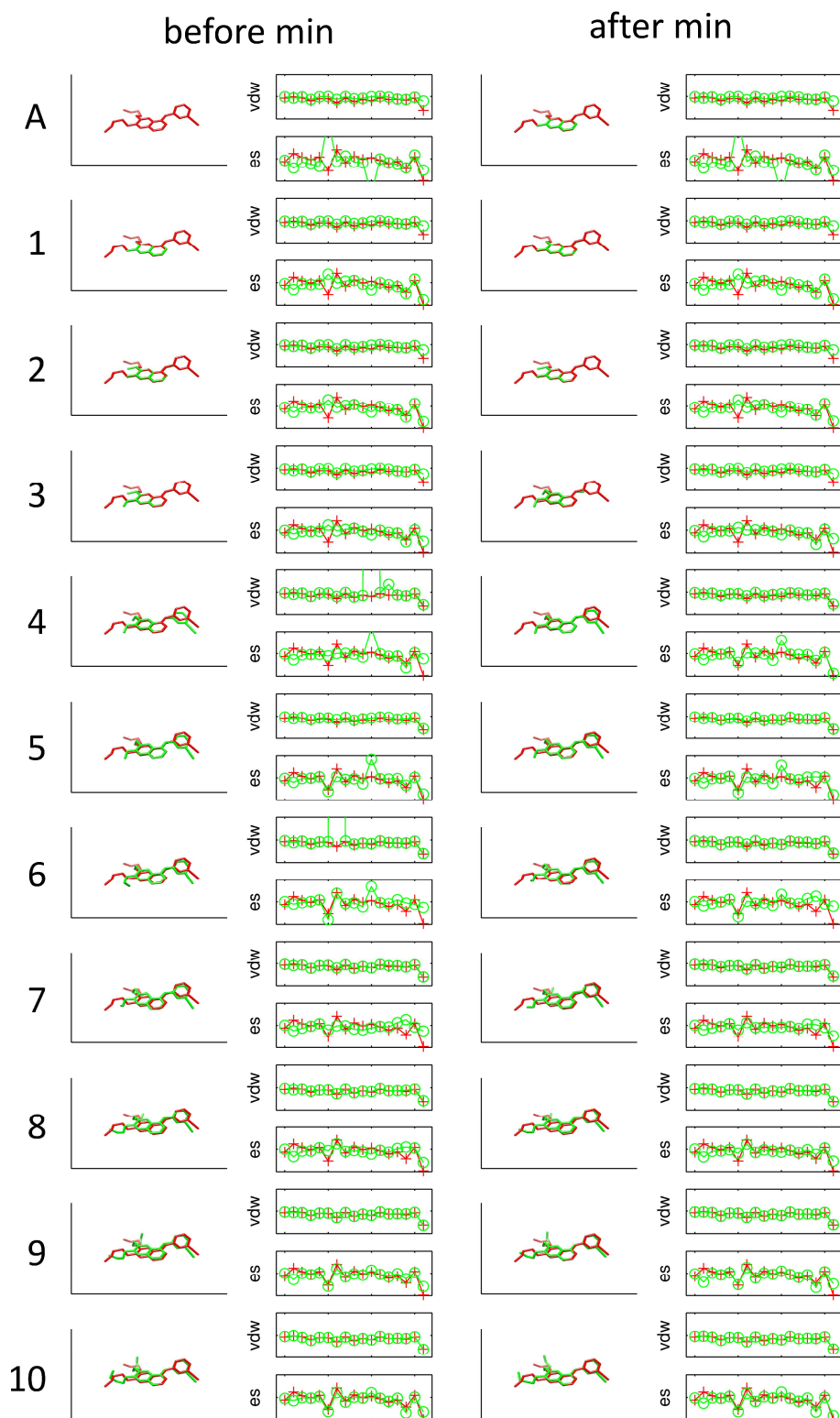


Figure 4-5. Fixed anchor docking branch converges to a low RMSD and low FPS score. Both structures before and after minimization are shown.

Here, growth steps before and after minimization are shown. We observe that after-minimization values converge to a favorable FPS score (Figure 4-4, bottom panel) which is also evident by the final well-overlaid footprint at the bottom left panel of Figure 4-5 (step 10).

4.4.3 Conclusions.

This chapter describes work in progress to implement and validate a grid-based footprint comparison method into the program DOCK. Our results have shown that a grid-based method produces footprints similar to those for Cartesian decompositions (Figures 4-1 and 4-2). The method does about 1% better than standard docking with rescoring but is on average about ca. 4 times slower (Table 4-2). Nevertheless, this may be an acceptable tradeoff for specific docking scenarios. FPS-guided docking, which uses only FPS score to drive sampling, does very well using standard Euclidean but very poorly for normalized Euclidean and standard Pearson which is most likely do to the need to account for the magnitude of the interactions (Figure 4-3 and Table 4-2). Standard Euclidean matches both the magnitude and the shape of the footprint spectra while normalized Euclidean and standard Pearson only match the shape. All comparison methods perform well in the FPS-nudged docking experiments (Figure 4-3 and Table 4-2). However, for virtual screening or *de novo* design application it may be better to use FPS-nudged docking with a shape matching method to find molecules that may have larger magnitudes than the reference. The FPS-guided results clearly show that use of FPS during growth yields molecular interactions which are more similar to the reference than standard grid sampling. The FPS scores are lower for the guided results than for the rescoring results (Figure 4-4 and Figure 4-5). These results indicate that the grid-based footprint may be a useful scoring function to drive sampling for *de novo* design applications.

Chapter 5. Dissertation Summary: Scientific Impact, Related Work, and Future Direction.

As discussed in the introduction (Chapter 1) computational modeling is an increasingly important part of the overall drug discovery process. The studies presented in this dissertation provide additional examples of well validated computational work in which discoveries and advances have been made regarding the anticancer target EGFR and progress towards improved virtual screening protocols. This final chapter summarizes key findings of the projects in Chapters 2, 3, and 4 in terms of *scientific impact*, and *related work and future direction*.

5.1 Structural and energetic analysis of EGFR simulations.

Scientific impact.

As described in Chapter 2, through all-atom simulations, we obtained structural and energetic data which allowed us to characterize the mechanisms of resistance of EGFR to inhibitors. A key finding was the reduction of the water mediated interactions when comparing a baseline system (L858R) to a drug resistant mutant (L858R&T790M). The impact of this study is a better understanding of (1) the binding of current inhibitors to EGFR variants and (2) the role mutations in the receptor play in ligand binding in part by perturbing hydration patterns. Further

motivations of the work are to develop new and more robust inhibitors of EGFR and/or IGF-1R to combat cancer.

Related work and future direction.

Building on our understanding of the role of key water molecules to binding, Huang and Rizzo have examined the role of selectivity and mutation in the binding of lapatinib to the inactive conformation of HER family members including EGFR.¹³⁶ Ligand changes were also proposed by the group to probe the importance of specific water mediated interactions.

A future direction might be to use the data from Rewcastle et al.¹⁰¹ to probe the structure activity relationship of molecules similar to erlotinib and gefitinib (Chapter 2), using Thermodynamic Integration (TI) of ligands with the active conformation of wild-type EGFR. Simulations could be used to predict effects of small differences in ligands on binding to the mutant receptor. By probing the relationships between activity and structure, we can understand which interactions are desirable for the design of next generation inhibitors that will target resistant receptors. Extending footprint-like analysis to TI calculations (as briefly outlined in eqs. 5-1 to 5-4) would also be worthwhile.

$$V_{\text{sys}} = V_{\text{rec}}^{\text{internal}} + V_{\text{lig}}^{\text{internal}} + V_{\text{wat}}^{\text{internal}} + V_{\text{lig,rec}} + V_{\text{lig,wat}} + V_{\text{rec,wat}} \quad 5-1$$

$$\frac{\partial V_{\text{sys}}}{\partial \lambda} = \frac{\partial V_{\text{lig}}^{\text{internal}}}{\partial \lambda} + \frac{\partial V_{\text{lig,rec}}}{\partial \lambda} + \frac{\partial V_{\text{lig,wat}}}{\partial \lambda} \quad 5-2$$

$$\frac{\partial V_{\text{lig,rec}}}{\partial \lambda} = \sum_{i \in \text{rec}} \frac{\partial V_{\text{lig,rec}(i)}}{\partial \lambda} \quad 5-3$$

$$\Delta \Delta G_{\text{bind}}^{\text{com}} = \int_{\lambda=0}^1 \frac{\partial}{\partial \lambda} V_{\text{lig}}^{\text{internal}} + \sum_{i \in \text{rec}} \int_{\lambda=0}^1 \frac{\partial}{\partial \lambda} V_{\text{lig,rec}(i)} + \int_{\lambda=0}^1 \frac{\partial}{\partial \lambda} V_{\text{lig,wat}} \quad 5-4$$

Here, the solvated complex system energy (V_{sys}) may be broken down into internal and multi-species components eqs. 5-1 and 5-2. In the case where only part of the ligand is contained in the lambda mask, only three of the components will change with lambda, the rest will be zero. Since summation, integration and partial derivatives are all linear operators, we may decompose the $\partial V_{lig,rec}/\partial\lambda$ into per-residue components and then integrate over these components (eqs. 5-3 and 5-4).

As discussed in the literature, the L858R&T790M mutation increases binding towards ATP.⁶⁴ Part of the reason for resistance may not be just the weakening of interactions with inhibitors but also the strengthening of the binding of ATP.⁶⁴ Interestingly, our preliminary unpublished simulations, using the same protocol described in Chapter 2, indicate that the ligand AMPPNP (an ATP analog) may preferentially bind to the mutant, as hypothesized by Yun et al.⁶⁴ Future computational studies may be undertaken to more completely explore this result.

Identification of alternative binding sites on EGFR is an important strategy for development of new inhibitors. It would be interesting to construct and validate how the carboxyl tail of EGFR docks into the kinase domain to allow phosphorylation of its tyrosine and identify or design small molecules to block this event. Since MIG6 is a naturally produced peptide that inhibits the dimerization of EGFR,¹³⁷ it could be advantageous to discover small molecule inhibitors that will disrupt the dimerization event by mimicking MIG6 interactions.

Additional studies using molecular dynamics, umbrella sampling and nudged elastic band methods could be performed to probe how ligand binding, mutations and dimerization shift the equilibrium of EGFR from inactive to active conformations. In addition, an allosteric pocket may be identified by examining the pathway EGFR traverses from active to inactive states. There are two main areas of differences between the active and inactive conformations in the N-

lobe helix and the activation-loop as shown in Figure 5-1. Through a preliminary study, a path using structural linear interpolation with 20 windows between active and inactive conformations was identified. A cavity near the N-lobe helix could be targeted to possibly trap the receptor in an intermediate state. Another observation is that the residue K860 appears to impede the conformational change between active and inactive conformations. Additional studies employing umbrella sampling and/or nudged elastic band are warranted to better determine the transition path between these two conformations.

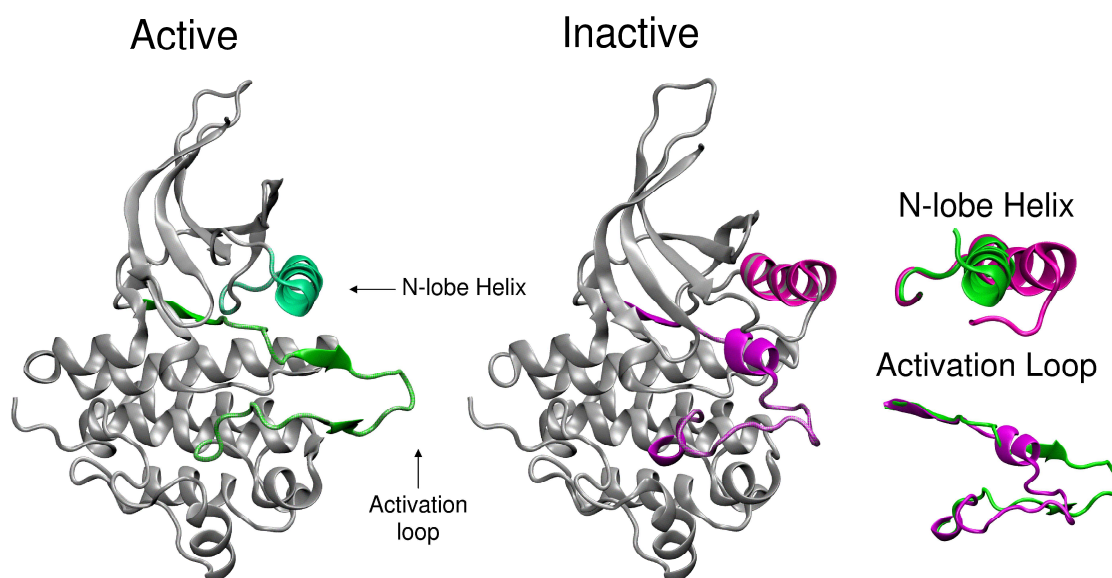


Figure 5- 1. Active vs inactive conformations. The grey ribbon indicates the backbone of EGFR. The colored regions represent the main differences between the two structures. An overlay of the N-lobe helix and the activation loop is also shown.

5.2 Development of docking methods and virtual screening protocols.

Scientific impact.

A major component of this work has involved DOCK development to aid the drug discovery efforts in our laboratory and also the community. Key improvements both to sampling and scoring routines have been made as discussed below. Importantly, compounds targeting HIV gp41¹³⁸ and botulinum neurotoxin, identified in computational virtual screens using DOCK

with the new footprint-based scoring function (Chapter 3), provided leads with experimental activity. Such leads form the basis for additional studies and development for these therapeutic targets. All of the work on DOCK has been released to the community in DOCK v6.4 and v6.5 or will soon be made available in future releases. Scientific contributions have included: (1) assistance with modifying growth routines to include internal energy, (2) adding growth tree protocol analysis feature, (3) adding RMSD tether, (4) implementing of Descriptor Score (FPS scoring function and hydrogen bonding code, Chapter 3), (5) adding anchor selection, (6) implementing Multi Grid code (Grid-based footprints, Chapter 4), and (7) assistance in the implementation of a SASA-based scoring function to estimate ligand burial. Collaborative work also involved assisting with the construction of the SB2010 pose reproduction database.¹³

Related work and future direction.

Accounting for receptor flexibility during docking may be useful in drug discovery applications.^{139,140} To incorporate a method into DOCK, we have developed a preliminary implementation of a Multi-Grid scoring function in addition to the scoring functions discussed in Chapter 3 and 4. There are two schemes for combining the grids: One is termed Multiple Average Receptor (MAR) and the other is Multiple Independent Receptors (MIR). Ongoing work in the lab is focused on further development and robust validation using crossdocking and enrichment studies.

The grid-based footprint scoring function described in Chapter 4 is hypothesized to be useful in *de novo* design by facilitating construction of molecules with similar per-residue interactions as those of a known reference. A new *de novo* design method is being implemented into DOCK using the anchor-and-grow infrastructure to construct new molecules from a library

of scaffolds, linkers, and side-chains. An important feature of the *de novo* protocol is that fragment libraries will be generated from synthesizable molecules contained in databases such as ZINC.

Finally, the DOCK development work discussed above was in part motivated to apply the improved methods to lead discovery applications targeting EGFR and other kinases. Virtual screens to target IGF-1R are planned for the near future. An additional goal is to identify leads that will target both T790M EGFR and IGF-1R using multi-grid docking methods. Since IGF-1R has a methionine at the gate keeper residue position, it is hypothesized virtual screening can identify lead compounds with reasonable affinity to both proteins.

5.3 Summary.

As demonstrated in this dissertation, computational tools are useful in aiding drug discovery efforts with the long term goal of treating human disease. The most striking results of the studies described in this dissertation are as follows. (1) The role of the reduction in water mediated interactions between EGFR resistance mutant and erlotinib was identified. (2) The development of a new footprint-based scoring method was validated. And (3) third-party virtual screening applications which employed the footprint-based methods have resulted in identification of new lead compounds for biologically relevant targets.

Bibliography

- (1) Dickson, M.; Gagnon, J. P. Key factors in the rising cost of new drug discovery and development. *Nature Reviews. Drug Discovery* **2004**, *3*, 417-29.
- (2) Kuntz, I. D. Structure-based strategies for drug design and discovery. *Science* **1992**, *257*, 1078-82.
- (3) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813-8.
- (4) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862-5.
- (5) Jorgensen, W. L. Efficient drug lead discovery and optimization. *Acc Chem Res* **2009**, *42*, 724-33.
- (6) Structure-Based Drug Design Fact Sheet.
http://www.nigms.nih.gov/Education/structure_drugs.htm (accessed March 25, 2012).
- (7) Tomasselli, A. G.; Heinrikson, R. L. Targeting the HIV-protease in AIDS therapy: a current clinical perspective. *Biochimica et Biophysica Acta* **2000**, *1477*, 189-214.
- (8) von Itzstein, M. The war against influenza: discovery and development of sialidase inhibitors. *Nature Reviews. Drug Discovery* **2007**, *6*, 967-74.
- (9) Gibbs, J. B. Mechanism-based target identification and drug discovery in cancer research. *Science* **2000**, *287*, 1969-73.
- (10) Jemal, A.; Siegel, R.; Ward, E.; Hao, Y.; Xu, J.; Murray, T.; Thun, M. J. Cancer statistics, 2008. *CA: a Cancer Journal for Clinicians* **2008**, *58*, 71-96.
- (11) Capdeville, R.; Buchdunger, E.; Zimmermann, J.; Matter, A. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nature Reviews. Drug Discovery* **2002**, *1*, 493-502.
- (12) Congreve, M.; Murray, C. W.; Blundell, T. L. Keynote review: Structural biology and drug discovery. *Drug Discovery Today* **2005**, *10*, 895-907.
- (13) Mukherjee, S.; Balias, T. E.; Rizzo, R. C. Docking validation resources: protein family and ligand flexibility experiments. *Journal of Chemical Information and Modeling* **2010**, *50*, 1986-2000.

- (14) Irwin, J. J.; Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling* **2005**, *45*, 177-82.
- (15) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *Journal of Medicinal Chemistry* **2006**, *49*, 5851-5.
- (16) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews. Drug Discovery* **2004**, *3*, 935-49.
- (17) Layten, M.; Hornak, V.; Simmerling, C. The open structure of a multi-drug-resistant HIV-1 protease is stabilized by crystal packing contacts. *J Am Chem Soc* **2006**, *128*, 13360-1.
- (18) Celebrating structural biology. *Nature Structural & Molecular Biology* **2011**, *18*, 1304-16.
- (19) PDB website. <http://www.rcsb.org/pdb/statistics/holdings.do> (accessed Jan 17, 2011)
- (20) Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J. Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids. *Journal of the American Chemical Society* **1996**, *118*, 11225-36.
- (21) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *Journal of the American Chemical Society* **1995**, *117*, 5179-97.
- (22) Mackerell, A. D., Jr. Empirical force fields for biological macromolecules: overview and issues. *Journal of Computational Chemistry* **2004**, *25*, 1584-604.
- (23) Hornak, V.; Abel, R.; Okur, A.; Strockbine, B.; Roitberg, A.; Simmerling, C. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* **2006**, *65*, 712-25.
- (24) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham Iii, T. E.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* **1995**, *91*, 1-41.
- (25) Jorgensen, W. L.; Tirado-Rives, J. Monte Carlo vs Molecular Dynamics for Conformational Sampling. *The Journal of Physical Chemistry* **1996**, *100*, 14508-13.
- (26) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design* **2001**, *15*, 411-28.

- (27) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry* **1998**, *19*, 1639-62.
- (28) Leach, A. R. *Molecular Modelling Principles and Applications*; second edition ed.; Pearson Education: Harlow, England, 2007.
- (29) Massova, I.; Kollman, P. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspectives in Drug Discovery and Design* **2000**, *18*, 113-35.
- (30) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., 3rd. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of Chemical Research* **2000**, *33*, 889-97.
- (31) Baron, R.; Hunenberger, P. H.; McCammon, J. A. Absolute Single-Molecule Entropies from Quasi-Harmonic Analysis of Microsecond Molecular Dynamics: Correction Terms and Convergence Properties. *Journal of Chemical Theory and Computation* **2009**, *5*, 3150-60.
- (32) Bash, P. A.; Singh, U. C.; Brown, F. K.; Langridge, R.; Kollman, P. A. Calculation of the relative change in binding free energy of a protein-inhibitor complex. *Science* **1987**, *235*, 574-6.
- (33) Jorgensen, W. L. Free energy calculations: a breakthrough for modeling organic chemistry in solution. *Accounts of Chemical Research* **1989**, *22*, 184-9.
- (34) Kollman, P. Free energy calculations: Applications to chemical and biochemical phenomena. *Chemical Reviews* **1993**, *93*, 2395-417.
- (35) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *Journal of Molecular Biology* **1982**, *161*, 269-88.
- (36) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *Journal of Computer-Aided Molecular Design* **2006**, *20*, 601-19.
- (37) Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D. DOCK 6: combining techniques to model RNA-small molecule complexes. *RNA* **2009**, *15*, 1219-30.
- (38) Lorber, D. M.; Shoichet, B. K. Flexible ligand docking using conformational ensembles. *Protein Science* **1998**, *7*, 938-50.
- (39) Lorber, D. M.; Shoichet, B. K. Hierarchical docking of databases of multiple ligand conformations. *Current Topics in Medicinal Chemistry* **2005**, *5*, 739-49.
- (40) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *Journal of Computational Chemistry* **1992**, *13*, 505-24.

- (41) Brown, R. D.; Martin, Y. C. Use of Structure–Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection. *Journal of Chemical Information and Computer Sciences* **1996**, *36*, 572-84.
- (42) Balias, T. E.; Rizzo, R. C. Quantitative prediction of fold resistance for inhibitors of EGFR. *Biochemistry* **2009**, *48*, 8435-48.
- (43) Balias, T. E.; Mukherjee, S.; Rizzo, R. C. Implementation and evaluation of a docking-rescoring method using molecular footprint comparisons. *Journal of Computational Chemistry* **2011**, *32*, 2273–89,.
- (44) McGillick, B. E.; Balias, T. E.; Mukherjee, S.; Rizzo, R. C. Origins of resistance to the HIVgp41 viral entry inhibitor T20. *Biochemistry* **2010**, *49*, 3575-92.
- (45) Brozell, S. R.; Mukherjee, S.; Balias, T. E.; Roe, D. R.; Case, D. A.; Rizzo, R. C. Evaluation of DOCK 6 as a pose generation and database enrichment tool. *Journal of Computer-Aided Molecular Design* **2012**, In press. doi:10.1007/s10822-012-9565-y.
- (46) Travis, W. D.; Travis, L. B.; Devesa, S. S. Lung cancer. *Cancer* **1995**, *75*, 191-202.
- (47) Hirsch, F. R.; Varella-Garcia, M.; Bunn, P. A., Jr.; Di Maria, M. V.; Veve, R.; Bremmes, R. M.; Baron, A. E.; Zeng, C.; Franklin, W. A. Epidermal growth factor receptor in non-small-cell lung carcinomas: correlation between gene copy number and protein expression and impact on prognosis. *Journal of Clinical Oncology* **2003**, *21*, 3798-807.
- (48) Hynes, N. E.; Lane, H. A. ERBB receptors and cancer: the complexity of targeted inhibitors. *Nature Reviews. Cancer* **2005**, *5*, 341-54.
- (49) Mendelsohn, J.; Baselga, J. Status of epidermal growth factor receptor antagonists in the biology and treatment of cancer. *Journal of Clinical Oncology* **2003**, *21*, 2787-99.
- (50) Rocha-Lima, C. M.; Soares, H. P.; Raez, L. E.; Singal, R. EGFR targeting of solid tumors. *Cancer Control* **2007**, *14*, 295-304.
- (51) Drugs@FDA website. <http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm> (accessed Jan 22, 2009)
- (52) Riese, D. J., 2nd; Gallo, R. M.; Settleman, J. Mutational activation of ErbB family receptor tyrosine kinases: insights into mechanisms of signal transduction and tumorigenesis. *Bioessays* **2007**, *29*, 558-65.
- (53) Zhang, X.; Gureasko, J.; Shen, K.; Cole, P. A.; Kuriyan, J. An allosteric mechanism for activation of the kinase domain of epidermal growth factor receptor. *Cell* **2006**, *125*, 1137-49.
- (54) Yarden, Y.; Sliwkowski, M. X. Untangling the ErbB signalling network. *Nature Reviews. Molecular Cell Biology* **2001**, *2*, 127-37.

- (55) Hubbard, S. R.; Miller, W. T. Receptor tyrosine kinases: mechanisms of activation and signaling. *Current Opinion in Cell Biology* **2007**, *19*, 117-23.
- (56) Scaltriti, M.; Baselga, J. The epidermal growth factor receptor pathway: a model for targeted therapy. *Clinical Cancer Research* **2006**, *12*, 5268-72.
- (57) Wood, E. R.; Truesdale, A. T.; McDonald, O. B.; Yuan, D.; Hassell, A.; Dickerson, S. H.; Ellis, B.; Pennisi, C.; Horne, E.; Lackey, K.; Alligood, K. J.; Rusnak, D. W.; Gilmer, T. M.; Shewchuk, L. A unique structure for epidermal growth factor receptor bound to GW572016 (Lapatinib): relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. *Cancer Research* **2004**, *64*, 6652-9.
- (58) Stamos, J.; Sliwkowski, M. X.; Eigenbrot, C. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *The Journal of Biological Chemistry* **2002**, *277*, 46265-72.
- (59) Fry, D. W. Mechanism of action of erbB tyrosine kinase inhibitors. *Experimental Cell Research* **2003**, *284*, 131-9.
- (60) Traxler, P.; Allegrini, P. R.; Brandt, R.; Brueggen, J.; Cozens, R.; Fabbro, D.; Grosios, K.; Lane, H. A.; McSheehy, P.; Mestan, J.; Meyer, T.; Tang, C.; Wartmann, M.; Wood, J.; Caravatti, G. AEE788: a dual family epidermal growth factor receptor/ErbB2 and vascular endothelial growth factor receptor tyrosine kinase inhibitor with antitumor and antiangiogenic activity. *Cancer Research* **2004**, *64*, 4931-41.
- (61) Karaman, M. W.; Herrgard, S.; Treiber, D. K.; Gallant, P.; Atteridge, C. E.; Campbell, B. T.; Chan, K. W.; Ciceri, P.; Davis, M. I.; Edeen, P. T.; Faraoni, R.; Floyd, M.; Hunt, J. P.; Lockhart, D. J.; Milanov, Z. V.; Morrison, M. J.; Pallares, G.; Patel, H. K.; Pritchard, S.; Wodicka, L. M.; Zarrinkar, P. P. A quantitative analysis of kinase inhibitor selectivity. *Nature Biotechnology* **2008**, *26*, 127-32.
- (62) Carey, K. D.; Garton, A. J.; Romero, M. S.; Kahler, J.; Thomson, S.; Ross, S.; Park, F.; Haley, J. D.; Gibson, N.; Sliwkowski, M. X. Kinetic analysis of epidermal growth factor receptor somatic mutant proteins shows increased sensitivity to the epidermal growth factor receptor tyrosine kinase inhibitor, erlotinib. *Cancer Research* **2006**, *66*, 8163-71.
- (63) Ji, H.; Zhao, X.; Yuza, Y.; Shimamura, T.; Li, D.; Protopopov, A.; Jung, B. L.; McNamara, K.; Xia, H.; Glatt, K. A.; Thomas, R. K.; Sasaki, H.; Horner, J. W.; Eck, M.; Mitchell, A.; Sun, Y.; Al-Hashem, R.; Bronson, R. T.; Rabindran, S. K.; Discafani, C. M.; Maher, E.; Shapiro, G. I.; Meyerson, M.; Wong, K. K. Epidermal growth factor receptor variant III mutations in lung tumorigenesis and sensitivity to tyrosine kinase inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 7817-22.
- (64) Yun, C. H.; Mengwasser, K. E.; Toms, A. V.; Woo, M. S.; Greulich, H.; Wong, K. K.; Meyerson, M.; Eck, M. J. The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105*, 2070-5.

- (65) Yun, C. H.; Boggon, T. J.; Li, Y.; Woo, M. S.; Greulich, H.; Meyerson, M.; Eck, M. J. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell* **2007**, *11*, 217-27.
- (66) Wikstrand, C. J.; McLendon, R. E.; Friedman, A. H.; Bigner, D. D. Cell surface localization and density of the tumor-associated variant of the epidermal growth factor receptor, EGFRvIII. *Cancer Research* **1997**, *57*, 4130-40.
- (67) Lynch, T. J.; Bell, D. W.; Sordella, R.; Gurubhagavatula, S.; Okimoto, R. A.; Brannigan, B. W.; Harris, P. L.; Haserlat, S. M.; Supko, J. G.; Haluska, F. G.; Louis, D. N.; Christiani, D. C.; Settleman, J.; Haber, D. A. Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. *The New England Journal of Medicine* **2004**, *350*, 2129-39.
- (68) Paez, J. G.; Janne, P. A.; Lee, J. C.; Tracy, S.; Greulich, H.; Gabriel, S.; Herman, P.; Kaye, F. J.; Lindeman, N.; Boggon, T. J.; Naoki, K.; Sasaki, H.; Fujii, Y.; Eck, M. J.; Sellers, W. R.; Johnson, B. E.; Meyerson, M. EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy. *Science* **2004**, *304*, 1497-500.
- (69) Pao, W.; Miller, V.; Zakowski, M.; Doherty, J.; Politi, K.; Sarkaria, I.; Singh, B.; Heelan, R.; Rusch, V.; Fulton, L.; Mardis, E.; Kupfer, D.; Wilson, R.; Kris, M.; Varmus, H. EGF receptor gene mutations are common in lung cancers from "never smokers" and are associated with sensitivity of tumors to gefitinib and erlotinib. *Proceedings of the National Academy of Sciences of the United States of America* **2004**, *101*, 13306-11.
- (70) Sordella, R.; Bell, D. W.; Haber, D. A.; Settleman, J. Gefitinib-sensitizing EGFR mutations in lung cancer activate anti-apoptotic pathways. *Science* **2004**, *305*, 1163-7.
- (71) Riely, G. J.; Pao, W.; Pham, D.; Li, A. R.; Rizvi, N.; Venkatraman, E. S.; Zakowski, M. F.; Kris, M. G.; Ladanyi, M.; Miller, V. A. Clinical course of patients with non-small cell lung cancer and epidermal growth factor receptor exon 19 and exon 21 mutations treated with gefitinib or erlotinib. *Clinical Cancer Research* **2006**, *12*, 839-44.
- (72) Sequist, L. V.; Bell, D. W.; Lynch, T. J.; Haber, D. A. Molecular predictors of response to epidermal growth factor receptor antagonists in non-small-cell lung cancer. *Journal of Clinical Oncology* **2007**, *25*, 587-95.
- (73) Pao, W.; Miller, V. A. Epidermal growth factor receptor mutations, small-molecule kinase inhibitors, and non-small-cell lung cancer: current knowledge and future directions. *Journal of Clinical Oncology* **2005**, *23*, 2556-68.
- (74) Pao, W.; Miller, V. A.; Politi, K. A.; Riely, G. J.; Somwar, R.; Zakowski, M. F.; Kris, M. G.; Varmus, H. Acquired resistance of lung adenocarcinomas to gefitinib or erlotinib is associated with a second mutation in the EGFR kinase domain. *PLoS Medicine* **2005**, *2*, e73.
- (75) Branford, S.; Rudzki, Z.; Walsh, S.; Grigg, A.; Arthur, C.; Taylor, K.; Herrmann, R.; Lynch, K. P.; Hughes, T. P. High frequency of point mutations clustered within the adenosine triphosphate-binding region of BCR/ABL in patients with chronic myeloid leukemia or Ph-

positive acute lymphoblastic leukemia who develop imatinib (STI571) resistance. *Blood* **2002**, *99*, 3472-5.

(76) Gorre, M. E.; Mohammed, M.; Ellwood, K.; Hsu, N.; Paquette, R.; Rao, P. N.; Sawyers, C. L. Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **2001**, *293*, 876-80.

(77) Wissner, A.; Berger, D. M.; Boschelli, D. H.; Floyd, M. B., Jr.; Greenberger, L. M.; Gruber, B. C.; Johnson, B. D.; Mamuya, N.; Nilakantan, R.; Reich, M. F.; Shen, R.; Tsou, H. R.; Upeslakis, E.; Wang, Y. F.; Wu, B.; Ye, F.; Zhang, N. 4-Anilino-6,7-dialkoxyquinoline-3-carbonitrile inhibitors of epidermal growth factor receptor kinase and their bioisosteric relationship to the 4-anilino-6,7-dialkoxyquinazoline inhibitors. *Journal of Medicinal Chemistry* **2000**, *43*, 3244-56.

(78) Hou, T.; Zhu, L.; Chen, L.; Xu, X. Mapping the binding site of a large set of quinazoline type EGF-R inhibitors using molecular field analyses and molecular docking studies. *Journal of Chemical Information and Computer Sciences* **2003**, *43*, 273-87.

(79) Cavasotto, C. N.; Ortiz, M. A.; Abagyan, R. A.; Piedrafita, F. J. In silico identification of novel EGFR inhibitors with antiproliferative activity against cancer cells. *Bioorganic & Medicinal Chemistry Letters* **2006**, *16*, 1969-74.

(80) Liu, B.; Bernard, B.; Wu, J. H. Impact of EGFR point mutations on the sensitivity to gefitinib: insights from comparative structural analyses and molecular dynamics simulations. *Proteins* **2006**, *65*, 331-46.

(81) Liu, Y.; Purvis, J.; Shih, A.; Weinstein, J.; Agrawal, N.; Radhakrishnan, R. A multiscale computational approach to dissect early events in the Erb family receptor mediated activation, differential signaling, and relevance to oncogenic transformations. *Annals of Biomedical Engineering* **2007**, *35*, 1012-25.

(82) Strockbine, B.; Rizzo, R. C. Binding of antifusion peptides with HIVgp41 from molecular dynamics simulations: quantitative correlation with experiment. *Proteins* **2007**, *67*, 630-42.

(83) Chachra, R.; Rizzo, R. C. Origins of Resistance Conferred by the R292K Neuraminidase Mutation via Molecular Dynamics and Free Energy Calculations. *Journal of Chemical Theory and Computation* **2008**, *4*, 1526-40.

(84) Carrascal, N.; Rizzo, R. C. Calculation of binding free energies for non-zinc chelating pyrimidine dicarboxamide inhibitors with MMP-13. *Bioorganic & Medicinal Chemistry Letters* **2009**, *19*, 47-50.

(85) Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *Journal of the American Chemical Society* **1990**, *112*, 6127-9.

(86) Chemical Computing Group: Montreal, Canada, 2007.

- (87) Case, D. A.; Darden, T. A.; Cheatham, T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Merz, K. M.; Wang, B.; Pearlman, D. A.; Crowley, M.; Brozell, S.; Tsui, V.; Gohlke, H.; Mongan, J.; Hornak, V.; Cui, G.; Beroza, P.; Schafmeister, C.; Caldwell, J. W.; Ross, W. S.; Kollman, P. A.; AMBER 8, University of California, San Francisco: 2004.
- (88) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics* **1983**, *79*, 926-35.
- (89) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157-74.
- (90) Breneman, C. M.; Wiberg, K. B. Determining atom-centered monopoles from molecular electrostatic potentials. The need for high sampling density in formamide conformational analysis. *Journal of Computational Chemistry* **1990**, *11*, 361-73.
- (91) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A., Jr.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle, E. S.; Pople, J. A.; Gaussian Inc.: Pittsburgh PA, 1998.
- (92) 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. *Journal of Computational Physics* **1977**, *23*, 327-41.
- (93) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *The Journal of Chemical Physics* **1993**, *98*, 10089-92.
- (94) 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. *The Journal of Chemical Physics* **1984**, *81*, 3684-90.
- (95) Onufriev, A.; Bashford, D.; Case, D. A. Exploring protein native states and large-scale conformational changes with a modified generalized born model. *Proteins* **2004**, *55*, 383-94.
- (96) Sitkoff, D.; Sharp, K. A.; Honig, B. Accurate Calculation of Hydration Free Energies Using Macroscopic Solvent Models. *The Journal of Physical Chemistry* **1994**, *98*, 1978-88.
- (97) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K. Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* **2005**, *26*, 1781-802.

- (98) Daub, H.; Specht, K.; Ullrich, A. Strategies to overcome resistance to targeted protein kinase inhibitors. *Nature Reviews. Drug Discovery* **2004**, *3*, 1001-10.
- (99) Kobayashi, S.; Boggon, T. J.; Dayaram, T.; Janne, P. A.; Kocher, O.; Meyerson, M.; Johnson, B. E.; Eck, M. J.; Tenen, D. G.; Halmos, B. EGFR mutation and resistance of non-small-cell lung cancer to gefitinib. *The New England Journal of Medicine* **2005**, *352*, 786-92.
- (100) Rablen, P. R.; Lockman, J. W.; Jorgensen, W. L. Ab Initio Study of Hydrogen-Bonded Complexes of Small Organic Molecules with Water. *The Journal of Physical Chemistry A* **1998**, *102*, 3782-97.
- (101) Rewcastle, G. W.; Denny, W. A.; Bridges, A. J.; Zhou, H.; Cody, D. R.; McMichael, A.; Fry, D. W. Tyrosine kinase inhibitors. 5. Synthesis and structure-activity relationships for 4-[(phenylmethyl)amino]- and 4-(phenylamino)quinazolines as potent adenosine 5'-triphosphate binding site inhibitors of the tyrosine kinase domain of the epidermal growth factor receptor. *Journal of Medicinal Chemistry* **1995**, *38*, 3482-7.
- (102) Bean, J.; Riely, G. J.; Balak, M.; Marks, J. L.; Ladanyi, M.; Miller, V. A.; Pao, W. Acquired resistance to epidermal growth factor receptor kinase inhibitors associated with a novel T854A mutation in a patient with EGFR-mutant lung adenocarcinoma. *Clinical Cancer Research* **2008**, *14*, 7519-25.
- (103) Zhang, Q.; Liu, Y.; Gao, F.; Ding, Q.; Cho, C.; Hur, W.; Jin, Y.; Uno, T.; Joazeiro, C. A.; Gray, N. Discovery of EGFR selective 4,6-disubstituted pyrimidines from a combinatorial kinase-directed heterocycle library. *Journal of the American Chemical Society* **2006**, *128*, 2182-3.
- (104) Dunitz, J. D. The entropic cost of bound water in crystals and biomolecules. *Science* **1994**, *264*, 670.
- (105) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kretsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *Journal of Chemical Information and Modeling* **2007**, *47*, 1504-19.
- (106) Irwin, J. J. Community benchmarks for virtual screening. *Journal of Computer-Aided Molecular Design* **2008**, *22*, 193-9.
- (107) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *Journal of Computer-Aided Molecular Design* **2008**, *22*, 133-9.
- (108) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry* **2006**, *49*, 6789-801.
- (109) Tominaga, Y.; Jorgensen, W. L. General model for estimation of the inhibition of protein kinases using Monte Carlo simulations. *Journal of Medicinal Chemistry* **2004**, *47*, 2534-49.

- (110) Gohlke, H.; Kiel, C.; Case, D. A. Insights into protein-protein binding by binding free energy calculation and free energy decomposition for the Ras-Raf and Ras-RalGDS complexes. *Journal of Molecular Biology* **2003**, *330*, 891-913.
- (111) Zou, H.; Luo, C.; Zheng, S.; Luo, X.; Zhu, W.; Chen, K.; Shen, J.; Jiang, H. Molecular insight into the interaction between IFABP and PA by using MM-PBSA and alanine scanning methods. *The Journal of Physical Chemistry B* **2007**, *111*, 9104-13.
- (112) Carrascal, N.; Green, D. F. Energetic decomposition with the generalized-born and Poisson-Boltzmann solvent models: lessons from association of G-protein components. *The Journal of Physical Chemistry B* **2010**, *114*, 5096-116.
- (113) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *Journal of Medicinal Chemistry* **2004**, *47*, 337-44.
- (114) Kelly, M. D.; Mancera, R. L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 1942-51.
- (115) Chuaqui, C.; Deng, Z.; Singh, J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *Journal of Medicinal Chemistry* **2005**, *48*, 121-33.
- (116) Deng, Z.; Chuaqui, C.; Singh, J. Knowledge-based design of target-focused libraries using protein-ligand interaction constraints. *Journal of Medicinal Chemistry* **2006**, *49*, 490-500.
- (117) Marcou, G.; Rognan, D. Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *Journal of Chemical Information and Modeling* **2007**, *47*, 195-207.
- (118) Mpamhanga, C. P.; Chen, B.; McLay, I. M.; Willett, P. Knowledge-based interaction fingerprint scoring: a simple method for improving the effectiveness of fast scoring functions. *Journal of Chemical Information and Modeling* **2006**, *46*, 686-98.
- (119) Pfeffer, P.; Neudert, G.; Klebe, G. DrugScoreFP: profiling protein-ligand interactions using fingerprint simplicity paired with knowledge-based potential fields. *Chemistry Central Journal* **2008**, *2*, S16.
- (120) Renner, S.; Derksen, S.; Radestock, S.; Morchen, F. Maximum common binding modes (MCBM): consensus docking scoring using multiple ligand information and interaction fingerprints. *Journal of Chemical Information and Modeling* **2008**, *48*, 319-32.
- (121) Nandigam, R. K.; Kim, S.; Singh, J.; Chuaqui, C. Position specific interaction dependent scoring technique for virtual screening based on weighted protein-ligand interaction fingerprint profiles. *Journal of Chemical Information and Modeling* **2009**, *49*, 1185-92.

- (122) Perez-Nueno, V. I.; Rabal, O.; Borrell, J. I.; Teixido, J. APIF: a new interaction fingerprint based on atom pairs and its application to virtual screening. *Journal of Chemical Information and Modeling* **2009**, *49*, 1245-60.
- (123) Brewerton, S. C. The use of protein-ligand interaction fingerprints in docking. *Current Opinion in Drug Discovery & Development* **2008**, *11*, 356-64.
- (124) 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. *Journal of Medicinal Chemistry* **2005**, *48*, 2534-47.
- (125) Taylor, N. R.; Cleasby, A.; Singh, O.; Skarzynski, T.; Wonacott, A. J.; Smith, P. W.; Sollis, S. L.; Howes, P. D.; Cherry, P. C.; Bethell, R.; Colman, P.; Varghese, J. Dihydropyranocarboxamides related to zanamivir: a new series of inhibitors of influenza virus sialidases. 2. Crystallographic and molecular modeling study of complexes of 4-amino-4H-pyran-6-carboxamides and sialidase from influenza virus types A and B. *Journal of Medicinal Chemistry* **1998**, *41*, 798-807.
- (126) Presnell, S. R.; Patil, G. S.; Mura, C.; Jude, K. M.; Conley, J. M.; Bertrand, J. A.; Kam, C. M.; Powers, J. C.; Williams, L. D. Oxyanion-mediated inhibition of serine proteases. *Biochemistry* **1998**, *37*, 17068-81.
- (127) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F. Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. *Journal of Chemical Information and Computer Sciences* **2004**, *44*, 871-81.
- (128) Trott, O.; Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **2010**, *31*, 455-61.
- (129) Huel, N. H.; Nar, H.; Priepke, H.; Ries, U.; Stassen, J. M.; Wienen, W. Structure-based design of novel potent nonpeptide thrombin inhibitors. *Journal of Medicinal Chemistry* **2002**, *45*, 1757-66.
- (130) Nar, H.; Bauer, M.; Schmid, A.; Stassen, J. M.; Wienen, W.; Priepke, H. W.; Kauffmann, I. K.; Ries, U. J.; Huel, N. H. Structural basis for inhibition promiscuity of dual specific thrombin and factor Xa blood coagulation inhibitors. *Structure* **2001**, *9*, 29-37.
- (131) Larkin, M. A.; Blackshields, G.; Brown, N. P.; Chenna, R.; McGettigan, P. A.; McWilliam, H.; Valentin, F.; Wallace, I. M.; Wilm, A.; Lopez, R.; Thompson, J. D.; Gibson, T. J.; Higgins, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics* **2007**, *23*, 2947-8.
- (132) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, *96*, 9997-10002.

- (133) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* **1985**, *28*, 849-57.
- (134) Ferrari, A. M.; Wei, B. Q.; Costantino, L.; Shoichet, B. K. Soft docking and multiple receptor conformations in virtual screening. *Journal of Medicinal Chemistry* **2004**, *47*, 5076-84.
- (135) Wu, T. P.; Padmanabhan, K.; Tulinsky, A.; Mulichak, A. M. The refined structure of the epsilon-aminocaproic acid complex of human plasminogen kringle 4. *Biochemistry*. **1991**, *30*, 10589-94.
- (136) Huang, Y.; Rizzo, R. C. A Water-Based Mechanism of Specificity and Resistance for Lapatinib with ErbB Family Kinases. *Biochemistry* **2012**, In press, doi:10.1021/bi2016553.
- (137) Zhang, X.; Pickin, K. A.; Bose, R.; Jura, N.; Cole, P. A.; Kuriyan, J. Inhibition of the EGF receptor by binding of MIG6 to an activating kinase domain interface. *Nature* **2007**, *450*, 741-4.
- (138) Holden, P. M.; Kaur, H.; Goyal, R.; Gochin, M.; Rizzo, R. C. Footprint-based identification of viral entry inhibitors targeting HIVgp41. *Bioorganic & Medicinal Chemistry Letters* **2012**, in press, doi:10.1016/j.bmcl.2012.02.017.
- (139) Amaro, R. E.; Baron, R.; McCammon, J. A. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *Journal of Computer-Aided Molecular Design* **2008**, *22*, 693-705.
- (140) Wei, B. Q.; Weaver, L. H.; Ferrari, A. M.; Matthews, B. W.; Shoichet, B. K. Testing a flexible-receptor docking algorithm in a model binding site. *Journal of Molecular Biology* **2004**, *337*, 1161-82.
- (141) Eggink, D.; Langedijk, J. P.; Bonvin, A. M.; Deng, Y.; Lu, M.; Berkhout, B.; Sanders, R. W. Detailed mechanistic insights into HIV-1 sensitivity to three generations of fusion inhibitors. *The Journal of Biological Chemistry* **2009**, *284*, 26941-50.
- (142) Gochin, M.; Cai, L. The role of amphiphilicity and negative charge in glycoprotein 41 interactions in the hydrophobic pocket. *Journal of Medicinal Chemistry* **2009**, *52*, 4338-44.
- (143) Lawless, M. K.; Barney, S.; Guthrie, K. I.; Bucy, T. B.; Petteway, S. R., Jr.; Merutka, G. HIV-1 membrane fusion mechanism: structural studies of the interactions between biologically-active peptides from gp41. *Biochemistry* **1996**, *35*, 13697-708.
- (144) Hildinger, M.; Dittmar, M. T.; Schult-Dietrich, P.; Fehse, B.; Schnierle, B. S.; Thaler, S.; Stiegler, G.; Welker, R.; von Laer, D. Membrane-anchored peptide inhibits human immunodeficiency virus entry. *Journal of Virology* **2001**, *75*, 3038-42.
- (145) Peisajovich, S. G.; Gallo, S. A.; Blumenthal, R.; Shai, Y. C-terminal octylation rescues an inactive T20 mutant: implications for the mechanism of HIV/SIMIAN immunodeficiency virus-induced membrane fusion. *The Journal of Biological Chemistry* **2003**, *278*, 21012-7.

- (146) Wexler-Cohen, Y.; Johnson, B. T.; Puri, A.; Blumenthal, R.; Shai, Y. Structurally altered peptides reveal an important role for N-terminal heptad repeat binding and stability in the inhibitory action of HIV-1 peptide DP178. *The Journal of Biological Chemistry* **2006**, *281*, 9005-10.
- (147) Liu, S.; Jing, W.; Cheung, B.; Lu, H.; Sun, J.; Yan, X.; Niu, J.; Farmar, J.; Wu, S.; Jiang, S. HIV gp41 C-terminal heptad repeat contains multifunctional domains. Relation to mechanisms of action of anti-HIV peptides. *The Journal of Biological Chemistry* **2007**, *282*, 9612-20.
- (148) Champagne, K.; Shishido, A.; Root, M. J. Interactions of HIV-1 inhibitory peptide T20 with the gp41 N-HR coiled coil. *The Journal of Biological Chemistry* **2009**, *284*, 3619-27.
- (149) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *Journal of Medicinal Chemistry* **2007**, *50*, 726-41.
- (150) Good, A. C.; Oprea, T. I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *Journal of Computer-Aided Molecular Design* **2008**, *22*, 169-78.
- (151) Sheridan, R. P.; McGaughey, G. B.; Cornell, W. D. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *J Comput Aided Mol Des* **2008**, *22*, 257-65.
- (152) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J Chem Inf Model* **2007**, *47*, 488-508.
- (153) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219-28.
- (154) Maignan, S.; Guilloteau, J. P.; Pouzieux, S.; Choi-Sledeski, Y. M.; Becker, M. R.; Klein, S. I.; Ewing, W. R.; Pauls, H. W.; Spada, A. P.; Mikol, V. Crystal structures of human factor Xa complexed with potent inhibitors. *Journal of Medicinal Chemistry* **2000**, *43*, 3226-32.
- (155) DesJarlais, R. L.; Seibel, G. L.; Kuntz, I. D.; Furth, P. S.; Alvarez, J. C.; Ortiz de Montellano, P. R.; DeCamp, D. L.; Babe, L. M.; Craik, C. S. Structure-based design of nonpeptide inhibitors specific for the human immunodeficiency virus 1 protease. *Proceedings of the National Academy of Sciences of the United States of America* **1990**, *87*, 6644-8.
- (156) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today* **2006**, *11*, 580-94.

Appendix A. Origins of Resistance to the HIVgp41 Viral Entry Inhibitor T20.

Collaborative work described in this chapter has been published in McGillick, B. E.; **Balius, T. E.**; Mukherjee, S.; Rizzo, R. C. Origins of Resistance to the HIVgp41 Viral Entry Inhibitor T20. *Biochemistry*, **2010**, *49* (17), 3575-3592. Copyright © 2010 American Chemical Society. [doi:10.1021/bi901915g](https://doi.org/10.1021/bi901915g) PMID: 20230061

Author contributions: BEM, TEB, SM, and RCR designed research; BEM, TEB, SM performed research; BEM, TEB, and RCR wrote paper with assistance from SM.

A.1 Collaborative Project Summary.

Fuzeon (T20), which targets HIVgp41, is the first FDA approved drug for inhibition of viral-host membrane fusion. In this work, a model of T20 bound to HIVgp41 was constructed and solvated in explicit water and lipid to elucidate the molecular reasons for clinically relevant drug resistance. Molecular dynamics simulations (run for 10 ns) were post-processed to yield detailed structural and energetic results. Good agreement was observed between residues involved in resistance and those shown to be energetically important through per-residue energy

decomposition (footprint analysis). Six out of seven HIVgp41 point mutations (L33Q, G36V, I37K, V38E, Q40H, Q40K and outlier L33S) which deleteriously affect binding of T20 showed good overall agreement with experiment ($r^2 = 0.72$, $N=6$).

My main contributions to this work include: (1) detailed convergence evaluation; (2) energy decomposition heatmap calculations; and (3) peptide-lipid interaction analysis. These three key points, are discussed in the following subsections which are taken from the published work.⁴⁴

A.2 Binding Free Energy Convergence: Autocorrelation and Block Averaging.

An important part of any computer simulation analysis is estimation of uncertainty. In this work, simulation convergence was assessed (Figure A-1) by plotting autocorrelation functions (ACF) and block averaged standard errors of the mean (BASEM) for the ΔG_{bind} calcd time-series from T20 with wildtype gp41, as well as seven gp41 mutants (L33Q, L33S, G36V, I37K, V38E, Q40H, Q40K). For a given time series, ACF functions provide a means to assess how data are correlated and BASEM analysis allows error convergence to be estimated by allowing block length to vary. Ideally, reasonably uncorrelated data should be used for block averaging. Figure A-1a reveals that all ACF curves for ΔG_{bind} calcd drop quickly from 1 (100% correlated) at lag time = 0 ps, to less than ca. 25% correlated at 250 ps, before showing fluctuations which oscillate about 0% which is indicative of largely uncorrelated data.

Interestingly, the ACF results for T20 with I37K show a distinctly different trend (green solid arrow), in comparison to other trajectories, which is an indication this one simulation is not in reasonable equilibrium or as well-behaved. BASEM curves for similarly reveal I37K as an outlier in terms of poorly converged error estimates. An additional MD simulation was run for I37K which employed a different random seed number in an attempt to yield a trajectory with

improved equilibrium/convergence statistics. As shown in Figure A-1 (dashed green arrows), ACS and BASEM curves for the new MD run yielded results more consistent with the other converged simulations and this latter trajectory was used for all further analysis.

Error estimates for ΔG_{bind} calcd were obtained using the BASEM plots in Figure A-1b. Here, block averaged standard errors of the mean in kcal/mol were computed as a function of block size which ranges from 1 ps to 5000 ps (1 to 5000 snapshots). The BASEM results show the expected exponential increase as block averaging size becomes larger that begins to reach a plateau which is indicative of reasonable equilibrium. BASEM errors computed at $N=1$ likely underestimate the true error in the simulations while at $N=5000$ (1/2 the simulation length) the errors may not be reliable due to the fact that only two blocks are used. For the present 10 ns trajectories, Table A-1 plots BASEM errors, along with the % of uncorrelated data from ACF curves, for the eight simulations at block sizes of 1 ps, 100 ps, 250 ps, 500 ps, and 1000 ps. Individual system results and the overall average for these five block sizes are both reported. The regime between 250 and 500 ps yields data which is largely uncorrelated (ca 75 - 85%), and having a sufficiently large numbers of independent blocks (250 ps = 40 blocks and 500 ps = 20 blocks), thus the errors of 0.68 - 0.91 kcal/mol may be taken as a reasonable estimate of the statistical noise for ΔG_{bind} calcd for these simulations. Although the overall structural, energetic, and convergence metrics monitored over the course of the MD trajectories (Figure A-1 and Table A-1) indicate the T20-gp41 models are physically reasonable and well-behaved, ideally, longer time-series for each simulation would be desirable.

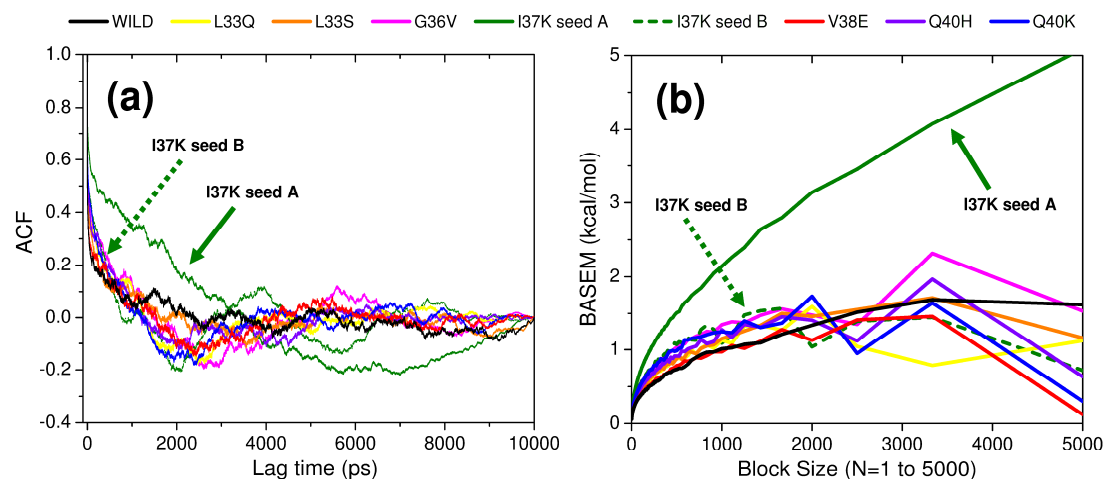


Figure A-1. (a) Autocorrelation functions (ACF) of calculated binding energies ($\Delta G_{\text{bind}} \text{ calc}$) for T20 with wild HIVgp41 and seven mutants. (b) Block average standard error of the mean (BASEM) in kcal/mol as a function of block size from 1 to 5000. Solid arrows indicate initial I37K results while dashed arrows indicate I37K results obtained using a different random seed.

Table A- 1. Autocorrelation function percent uncorrelated data (ACF %) and block average standard error of the mean (BASEM) for $\Delta G_{\text{bind}} \text{ calc}$ results from simulations of T20 with wildtype HIVgp41 and seven mutants for various block lengths.

mutation	N = 1 ps		N = 100 ps		N = 250 ps		N = 500 ps		N = 1000 ps	
	ACF % ^a	BASEM ^b	ACF %	BASEM	ACF %	BASEM	ACF %	BASEM	ACF %	BASEM
WT	46.13	0.06	79.44	0.38	82.93	0.56	87.91	0.73	95.10	1.01
L33Q	41.45	0.07	65.54	0.44	75.20	0.65	84.34	0.88	88.44	1.13
L33S	39.96	0.07	69.46	0.45	81.34	0.67	88.43	0.88	90.89	1.15
G36V	34.54	0.08	64.29	0.52	71.87	0.76	78.96	1.03	86.10	1.34
I37K	33.64	0.08	59.43	0.54	69.77	0.80	85.24	1.07	100.39	1.09
V38E	44.79	0.07	74.14	0.39	79.68	0.56	86.96	0.78	95.13	0.97
Q40H	36.55	0.07	66.64	0.48	70.09	0.68	81.25	0.91	94.81	1.27
Q40K	32.04	0.07	61.20	0.53	75.26	0.79	83.38	0.99	96.64	1.24
average	38.64	0.07	67.52	0.47	75.77	0.68	84.56	0.91	93.44	1.15

^aACF in % of uncorrelated data. ^bBASEM energies in kcal/mol.

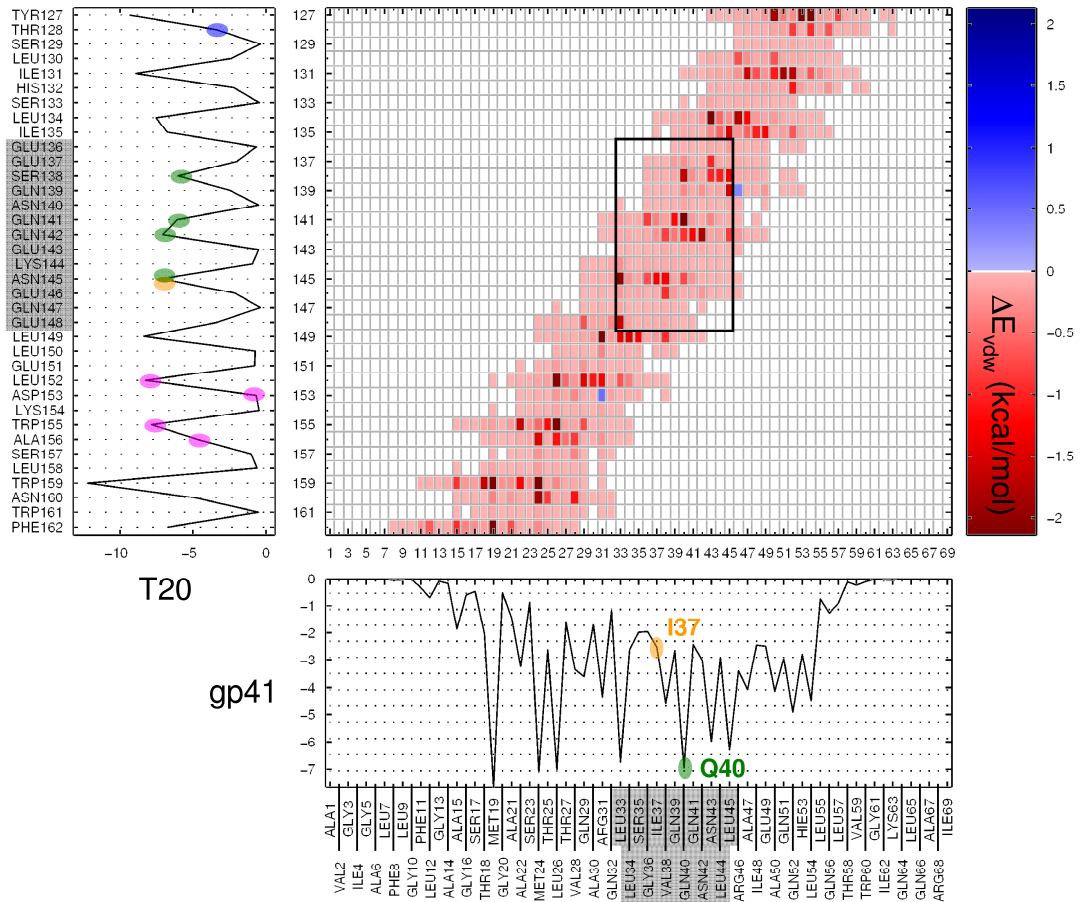


Figure A-2. Intermolecular van der Waals interaction energy matrix (heat map) for all gp41 with T20 residues from the wildtype simulation. Footprint peak magnitudes represent summation ΔE_{vdw} (kcal/mol) along each row (T20) or column (gp41). Gray shaded regions indicate mapping of the gp41 mutation region(L33-L45) to a charged/polar patch (E136-E148) on T20 defined by the black box in the ca. center of the matrix. Heat map gradient colors indicate the magnitude of unfavorable (blue) or favorable (red) interactions with dark red being most favorable.

A.3 Per-Residue Energetic Analysis.

As discussed in Chapter 1, per-residue energy decompositions are very useful at identifying important interactions. For example, favorable Coulombic attractions in the T20 system were observed to lead in some cases to unfavorable van der Waals energies at the site of the interaction as shown by the two matrix entries colored blue in Figure A-2 which represent an intermolecular R31-D153 salt bridge and a R46-Q139 H-bond between gp41 and T20. For V38E, intermolecular repulsions which occurs as a result of the change to a negative residue

leads to large losses in favorable $\Delta\Delta E_{\text{coul}}$ and a corresponding unfavorable solvent-mediated electrostatic energy $\Delta\Delta G_{\text{elec}}$.⁴⁴ Overall, the simulation results suggest that $\Delta\Delta G_{\text{FR}}$ is dominated by losses in favorable steric packing for I37K and Q40K vs losses in favorable electrostatic energy for V38E as a result of large changes in Coulombic attraction or repulsion. This analysis is consistent with a recent study by Eggink et al.¹⁴¹ who grouped charged gp41 mutations into similar mechanistic classes. In a related study, Gochin et al.¹⁴² found long-range electrostatic forces to be important for C-peptide peptide binding.

A.4 Membrane Interactions: The Importance of the WNWF Motif.

A growing body of experimental evidence¹⁴³⁻¹⁴⁸ suggests that T20 activity requires both N-helical binding and membrane binding. In particular, the four C-terminal residues on T20 termed the WNWF motif (residues 159-162, Figure 1 in paper⁴⁴) are thought to interact with the host cell membrane during fusion.^{146,147} A side-by-side comparison of van der Waals and Coulombic energy footprints (Figure A-3) clearly indicate that in addition to gp41 (dashed lines), the C-terminal end of T20 also makes significant interactions with the lipid bilayer (black lines). T20 packing interactions with membrane first appear to be important starting around W155, followed by L158, W159, W161, and F162. Importantly, the two terminal T20 residues (Trp161, and Phe162) show strikingly enhanced lipid-packing compared with other residues and in a relative sense the interactions are much stronger than those which occur with gp41 (Figure A-3a solid vs dashed lines). L158 also makes stronger van der Waals interactions with the membrane than with gp41. Interestingly, the side-by-side comparison reveals that when C-terminal T20 residues make strong packing interactions with gp41, weaker interactions are observed with the lipid and vice versa.

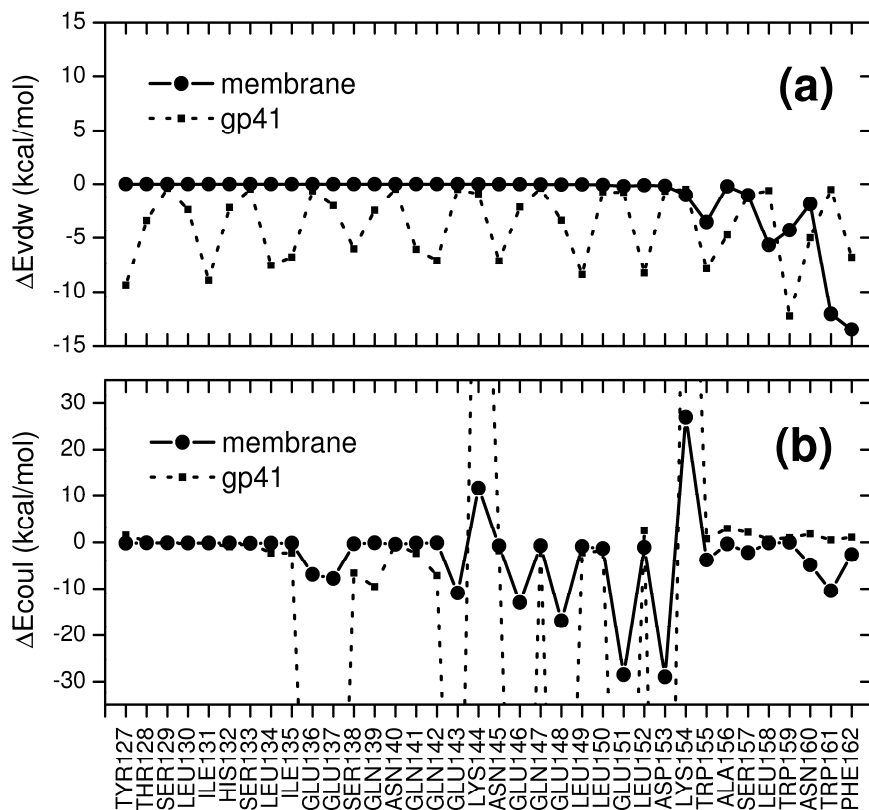


Figure A-3. Comparison of the per-residue van der Waals (a) and Coulombic (b) interaction energies between gp41 (dashed line, small squares) and lipid membrane (solid line, filled circles) for each T20 residue from wildtype simulations. Each datapoint represents the average value obtained from 10,000 MD snapshots saved every 1 ps.

The Coulombic plot reveals that the seven charged T20 residues interact with gp41 and the membrane (Figure A-3b dashed vs black lines) in an overall similar manner (i.e. peak location) which results in a net favorable interaction energy with the membrane. Interestingly, despite the fact that the T20 is bound along the full length of the inner coiled-coil, more favorable Coulombic energies are observed to occur with the membrane, as opposed to gp41, for three out of the four residues in the C-terminal WNWF motif (Figure A-3b dashed vs black lines). The regularly repeating pattern in the van der Waals footprint indicates which T20 residues are in close contact with gp41 (Figure A-3a dashed lines) and include W159, N160, and F162 in the WNWF motif. In contrast, prior experimental studies^{146,148} have suggested T20 C-

terminal residues may not interact with gp41. Champagne et al.¹⁴⁸ recently reported that the mutation WNWF→ANAA led to no change in T20 binding affinity using a gp41 construct termed 5H-ex (N-helix a.a. 19-71). And, Wexler-Cohen et al.¹⁴⁶ has reported that D-configuration substitutions on T20 at positions L158 and W159 did not significantly change binding to a construct termed N54 (N-helix a.a. 17-70). However, the van der Waals heat map results (Figure A-2 x-axis) clearly indicate that T20 also makes favorable contacts with the gp41 inner coiled-coil down to residues 16, 15, and 12. Thus, the shorter truncated N-helical sites employed in these prior experimental studies^{146,148} may not have provided a complete binding interface for T20. Consistent with this explanation, Champagne et al.¹⁴⁸ noted that, compared to the peptide C37, substantial differences in affinity for 5H-ex were observed with T20 which might suggest the construct does not fully mimic a complete gp41 binding site.

From a structural perspective, several types of interesting interactions were identified with the aid of stereo 3-D visualization between residues in the WNWF motif and the membrane which could contribute to stability of the overall complex (Figure A-4). Specifically, lipid head groups appeared to be interacting favorably with aromatic rings and with the T20 backbone. In an effort to quantify such interactions, radial distribution functions (rdfs) were computed for several types of interactions identified visually as being potentially long-lived. Distinct structural features, representative of a "first solvation shell" for T20 with lipid were observed in several rdfs as shown in Figure A-4a. Representative examples of favorable interactions (d2-d4 distances ≤ 5 Å) are visually illustrated in the accompanying graphic (Figure A-4b).

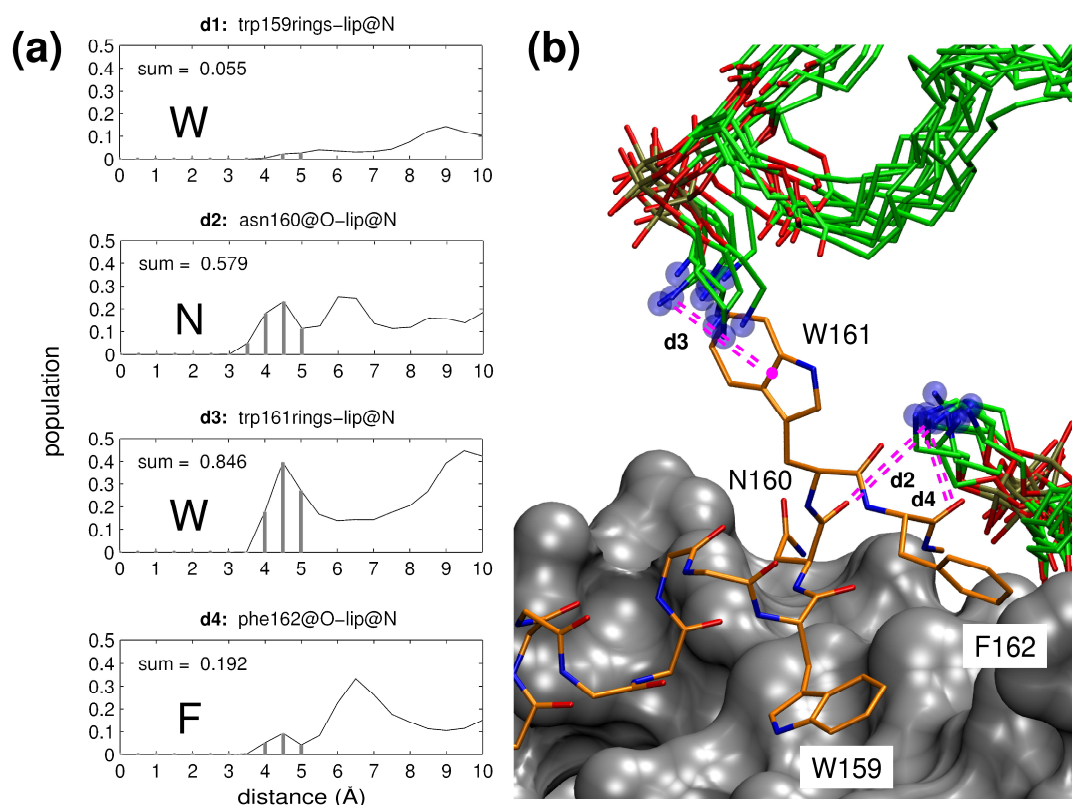


Figure A-4. Interactions of the T20 WNWF motif with membrane. 3-4a shows radial distribution functions (rdfs) for the average distances ($N=10,000$ snapshots) between all lipid headgroup N atoms and (d1) the center of ring atoms at W159, (d2) the backbone O at N160, (d3) the center of ring atoms at W161, and (d4) the backbone O at F162 (d4). Summation of bins (panel a gray vertical lines) out to a distance ≤ 5 Å approximates the number of interactions occurring in the first T20-lipid solvation shell represented by the first peaks in the rdfs. 3-4b graphically illustrates representative favorable interactions (distances ≤ 5 Å) for the T20-lipid interaction pairs defined as d2-d4 (magenta dashed lines). The graphic shows a single simulation snapshot of T20 (orange sticks) complexed with gp41 (gray surface) interacting with ten representative conformations of two nearby lipids (green sticks). Lipid head group carbons omitted for clarity.

Plots in Figure A-4a show the rdf which yielded the largest number of interactions (distance ≤ 5 Å) out of several atom selections evaluated for each WNWF residue with the lipid head groups. Although numerous distance definitions (and thus rdfs) are possible, summations up to the peaks at around 5 Å strongly suggest that the aromatic ring of W161 is particularly well solvated by lipid head groups with 0.85 interactions followed by the backbone O of N160 with 0.58 interactions (Figure A-4a). For W161, this highly populated pi-type interaction likely corresponds with the distinct Coulombic energy peak seen at this position in the membrane vs

gp41 footprint plots for T20 shown in Figure A-3. In sharp contrast, the rdf for the aromatic ring of W159 shows essentially no first solvation shell peak (Figure A-4a top) indicating an absence of specific interactions with the polar lipid head groups. This result is consistent with W159 being particularly well packed on the gp41 surface in comparison with other residues in the WNWF region (Figure A-3 dashed vs solid lines). For F162, although the backbone O does show weak lipid interactions (Figure A-4, d4), surprisingly, the side chain is not solvated by head groups as seen with W161, despite the fact that the aromatic ring might also have participated in pi-type bonding. Examination of the MD trajectories reveals that the phenyl group on F162 is too well buried in the lipid bilayer to make direct contact with polar head groups. Supporting this explanation, results in Figure A-4a show that F162, the terminal residue on T20, makes stronger more favorable van der Waals interactions with the membrane than any other residue in the WNWF motif.

Appendix B. Evaluation of DOCK 6 as a Database

Enrichment Tool.

Collaborative work described in this chapter has been published in Brozell, S. R.; Mukherjee, S.; **Balius, T. E.**; Roe, D. R.; Case, D. A.; Rizzo, R. C., Evaluation of DOCK 6 as a Pose Generation and Database Enrichment Tool. *J. Comput. Aided Mol. Des.* 2012, in press. Copyright © 2012 Springer Science+Business Media B.V. [doi:10.1007/s10822-012-9565-y](https://doi.org/10.1007/s10822-012-9565-y)

Author contributions. SRB, SM, TEB, DRR, DAC, and RCR designed research; SRB, SM, TEB, DRR performed research; SRB, SM, TEB, DRR, DAC, and RCR wrote the paper.

B.1 Collaborative Project Summary

We participated in a special symposium titled "Docking and Scoring: A Review of Docking Programs" which took place at the 241th American Chemical Society National Meeting held in Anaheim, CA (2011). The meeting led to our participation in a special issue in the *Journal of Computer-Aided Molecular Design* as a collaboration between two DOCK development groups Stony Brook University (Rizzo) and Rutgers University (Case) groups. As a requirement we were asked to evaluate DOCK6 using two main evaluation methods: pose

reproduction using the ASTEX Database¹⁴⁹ and enrichment experiments using the DUD database. My contributions, which focuses on docking enrichment studies, is described below and is taken from our manuscript which has been accepted.⁴⁵

B.2 Theoretical Methods

Enrichment Metrics. For accessing the accuracy of docking programs for virtual screening, Receiver Operating Characteristic (ROC) curves are used to evaluate how well methods favorably rank known active molecules compared with a large number of decoys.¹²⁴ In ROC curves, the True Positive Rate ($TPR=TP/P$) is plotted vs the False Positive Rate ($FPR=FP/N$), where TP is the number of True Positives, P is the total number of Positives (actives), FP is the number of False Positives, and N is the total number of Negatives (decoys). Figure B-1 shows ROC curve examples representative of good enrichment (panel a), random enrichment (panel b), and poor enrichment (panel c). Quantitatively, the total area under the curve (AUC) of a ROC plot provides a measure of global enrichment. In a practical sense however, good early enrichment is reasonably expected to be more important for prioritizing compounds identified by virtual screening of large ligand libraries. The example in Figure B-1c illustrates ROC curve behavior with poor total enrichment but strong early enrichment. Several methods for assessing early enrichments have been reported.¹⁰⁷ In this report, we use four very specific definitions (%TPR, %FPR, %AUC, and FE) to assess early enrichment as outlined below.

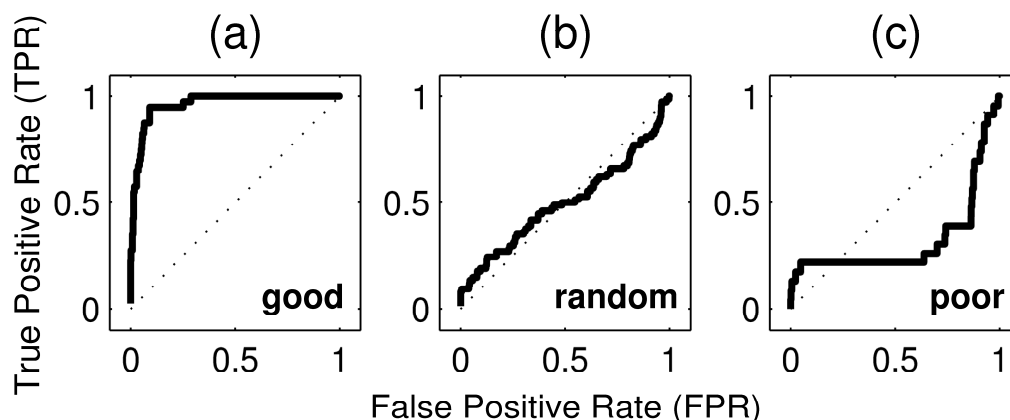


Figure B-1. Representative examples of ROC curves showing good (a), random (b), and poor (c) global enrichments.

For early enrichment, we report %TPR, and %FPR for different percentages (0.1, 1.0, 2.0) of the ranked database. Transforming to percentages yield scaled values which are much more readable for very early enrichments (i.e. not near zero). Percentages yield %FPR and %TPR in the range [0, 100]. We also report %AUC for early enrichment which is in the range [0, 10000]. We feel %AUC is a more meaningful metric to gauge early enrichment since the values involve the area and not a single point on the ROC curve. When reporting total AUC we report unscaled values. The panel in Figure B-2a illustrates the relationship between %FPR (X_ϵ), %TPR (Y_ϵ), and the % of database ranked ($\epsilon = 0.1, 1.0, 2.0$) for three different values. It should be noted that, the %FPR and the % of database screened are not necessarily equal. However, under most conditions, when $P \ll N$ this is in fact a reasonable assumption ($X_\epsilon \approx \epsilon$).

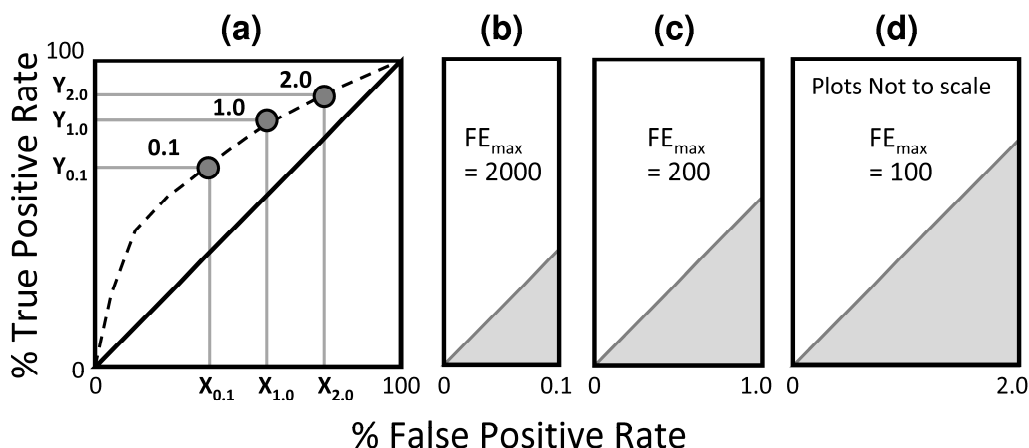


Figure B-2. (a) Schematic showing possible enrichments at 0.1%, 1.0%, and 2.0% of the database scanned and (b)-(d) maximum Fold Enrichment (FE) values at each of these points. $FE = AUC / AUC_{ran}$.

The fourth early enrichment metric used here is termed fold enrichment, recently employed to assess the footprint-based scoring function,⁴³ and defined as $FE = AUC / AUC_{ran}$. This quantity facilitates comparison with random behavior and is a very intuitive measure. Using the approximation $X_\epsilon \approx \epsilon$ when 0.1%, 1.0%, and 2.0% of the database are kept then 2000, 200, and 100 are the best possible (FE_{max}) fold enrichment values (Figure B-2b to d). It is important to note that for FE it is only meaningful to compare values calculated at the same percentage of the database but not between different percentages.

Enrichment Datasets. DUD¹⁰⁸ consists of forty protein structures (39 from the PDB and one homology model, denoted here with the name MODL) divided into six families as shown in Table B-1: Metalloenzyme (N=4), Nuclear Hormone Receptor (N=8), Kinase (N=9), Folate Enzyme (N=2), Serine Protease (N=2), and finally a miscellaneous family called “Other Enzymes” (N=14) the same breakdown as in the DUD paper.¹⁰⁸ Table B-1 lists for each entry, the associated pdb code, the DUD system name, number of active ligands, number of decoys, number of WOMBAT active ligands (if applicable), and the so-called semi-random pairings for which the DUD receptors were combined with actives and decoys developed for a different

system. The WOMBAT set¹⁵⁰ consists of active small molecules for 13 of the DUD systems. Two WOMBAT sets did not correspond to any DUD receptor and were not used. For another set, active ligands for Estrogen Receptor alpha were a mix of both agonist and antagonists. However, DUD distinguishes between agonist and antagonist receptors thus WOMBAT runs for Estrogen Receptor are not included. In summary, the DUD SUP datasets consist of 40 DUD and 10 WOMBAT sets of active compounds which were used with the 40 supplied receptors. For DUD PDB, two receptors 1L2I and 1AH3 were excluded owing to structural problems in the PDB coordinates. Therefore, the DUD PDB datasets consist of 38 DUD and 9 WOMBAT sets of active compounds, with corresponding DUD decoy ligand sets, with 38 corresponding receptors.

Table B- 1. DUD systems and semi-random matching information.

Protein Family	PDB name	DUD name	No. ligands	No. decoys	No. WOMBAT	Semi-random Pairing	Pairing Description
Metallo-enzyme ^a (N = 4)	1O86	ace	49	1797		ada	
	1NDW	ada	39	927		ace	
	1H1D	comt ^b	11	468		pde5	
	1XP0	pde5	88	1978	101	comt	
Nuclear Hormone Receptor (N = 8)	2AO6	ar	79	2854	56	rxr	
	1L2I ^d	er_agon	67	2570	83 ^c	mr	
	3ERT	er_antag	39	1448	83 ^c	ppar	
	1M2Z	gr	78	2947		pr	
	2AA2	mr ^b	15	636		er_agon	
	1FM9	ppar	85	3127	43	er_antag	
	1SR7	pr	27	1041		gr	
	1MVC	rxr	20	750		ar	
Kinase (N = 9)	1CKP	cdk2	72	2074	190	pdgfrb	
	1M17	egfr	475	15996	81	hsp90	
	1AGW	fgfr1	120	4550		src	
	1UY6	hsp90	37	979		egfr	
	1KV2	p38	454	9141	60	vegfr2	
	MODL	pdgfrb	170	5980		cdk2	
	2SRC	src	159	6319		fgfr1	
	1KIM	tk ^b	22	891		pnp	unrelated
	1VR2	vegfr2	88	2906		p38	
Folate Enzyme (N = 2)	3DFR	dhfr	410	8367		gart	
	1C2T	gart	40	879		dhfr	
Serine Protease (N = 3)	1F0R	fxa	146	5745	125	thrombin	
	1BA8	thrombin	72	2456		fxa	
	1BJU	trypsin	49	1664		hivpr	unrelated
Other Enzymes ^a (N = 14)	1EVE	ache	107	3892		hmga	
	1AH3 ^d	alr2	26	995	42	ampc	
	1XGJ	ampc	21	786		alr2	
	1Q4G	cox1	25	911		sahh	
	1CX2	cox2	426	13289	88	na	
	1A8I	gpb	52	2140		hivrt	
	1HPX	hivpr	62	2038		trypsin	unrelated
	1RT1	hivrt	43	1519	120	gpb	
	1HW8	hmga	35	1480		ache	
	1P44	inha	86	3266		parp	
	1A4G	na	49	1874		cox2	
	1EFY	parp	35	1351		inha	
	1B8O	pnp	50	1036		tk	unrelated
	1A7A	sahh	33	1346		cox1	

^aProteins in these groups are unrelated, and not expected to cross-enrich. ^bThe very small number of ligands makes evaluating enrichment statistics at 0.1% problematic (< 1 ligand). ^cER α ligands not used for WOMBAT studies. ^dDUD systems not used in DUD PDB prep.

For each PDB code in DUD, the organizers also prepared alternative non-native pairings (termed semi-random pairings) which was described as a way to perform "null hypothesis"

testing. Unrelated proteins would be expected to yield enrichments no better than random (null hypothesis). However, due to the fact that the semi-random pairings actually involve structurally related proteins (Table B-1), good enrichment may not be unexpected for certain pairings. Metalloenzyme and Other Enzymes groups represent collections of unrelated proteins.

Enrichment Calculations. For a given molecule (active or decoy) that is docked to a target it is important to note that a final answer may not always be obtained.^{151,152} Such occurrences necessitate a decision as to how subsequent ROC curve analysis should be performed. For the current study, this was generally not problematic as on average a viable 3-D pose was returned for 99% of the molecules (Tables B-2 and B-3). From a practical standpoint, there are at least three possible ways of generating ROC curves (see Figure B-3 for a graphical representation) when not all molecules yield a ranked answer: (1) Generate ROC curves by ignoring molecules for which an answer was not obtained (Figure B-3 a black line). In this case the number of actives (P, positives) and decoys (N, negatives) employed become P_{docked} and N_{docked} . (2) Generate ROC curves by using the initial number of actives and decoys (P_{initial} and N_{initial}). In this instance ROC curves may not always reach TPR=1, FPR=1 (Figure B-3a gray line). (3) Generate ROC curves by assuming perfect, none, or random enrichment for the molecules which do not yield a final answer (Figure B-3b gray and purple lines). In this scenario ROC curves always reach TPR=1, FPR=1. In the current manuscript, ROC curves (and accompanying AUC values) were generated using method 3 and using the reasonable assumption of random enrichment for missing molecules.

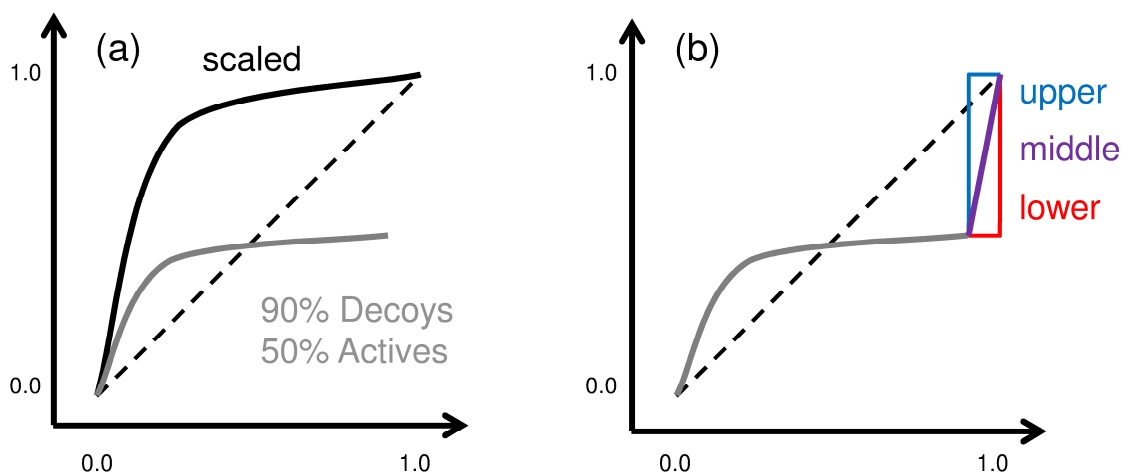


Figure B-3. (a) Hypothetical ROC curves computed using two different values for the total number of molecules classified as active (P, positive) or decoy (N, negative). The gray curve was computed using P_{initial} and N_{initial} and the black curve was computed using P_{docked} and N_{docked} . (b) Missing data can be assumed as yielding perfect enrichment (blue upper line), no enrichment (lower red line), or random enrichment (purple middle line) to ensure the ROC curve will reach TPR=1, FPR=1. The dashed line is the random ROC curve.

B.3 Results

DOCK Completion Rates for DUD Enrichment Experiments. DOCK completion statistics are presented below for both the DUD PDB (Table B-2) and DUD SUP (Table B-3) experiments. In cases where a DOCK answer was not generated possible reasons include an exceedingly high energy score (i.e. $> +1000$ kcal/mol), incompatibility with the binding site (i.e. large molecule in a small cavity), or insufficient torsion sampling. On average, a DOCK pose was returned in 99% of the runs. Shaded entries indicate the cases where less than 90% of the initial (total) molecules yielded a final (dockd) pose. On the average, 99% of molecules yielded a final answer.

Table B-2. Enrichment completion statistics using DUD PDB preparation.

PDB	DUD pairing		Normal				Semi-random			
	Normal	Semi-rand	decoys		actives		decoys		actives	
			initial	dockd	initial	dockd	initial	dockd	initial	dockd
1O86	ace	ada	1797	1787	49	49	927	925	39	39
1NDW	ada	ace	927	925	39	39	1797	1787	49	49
1H1D	comt	pde5	468	467	11	11	1978	1970	88	88
1XP0	pde5	comt	1978	1970	88	88	468	467	11	11
2AO6	ar	rxr	2854	2812	79	74	750	714	20	20
1L2I	er_agon	mr	NA	NA	NA	NA	NA	NA	NA	NA
3ERT	er_antag	ppar	1448	1437	39	38	3127	3116	85	85
1M2Z	gr	pr	2947	2694	78 ^a	58	1041	1015	27	22
2AA2	mr	er_agon	636	601	15	8	2570	2560	67	65
1FM9	ppar	er_antag	3127	3116	85	85	1448	1437	39	38
1SR7	pr	gr	1041	1023	27	23	2947	2845	78	64
1MVC	rxr	ar	750	714	20	20	2854	2815	79	72
1CKP	cdk2	pdgfrb	2074	2074	72	71	5980	5977	170	170
1M17	egfr	hsp90	15996	15990	475	475	979	978	37	37
1AGW	fgfr1	src	4550	4537	120	120	6319	6311	159	159
1UY6	hsp90	egfr	979	978	37	37	15996	15990	475	475
1KV2	p38	vegfr2	9141	9127	454	454	2906	2903	88	88
MODL	pdgfrb	cdk2	5980	5977	170	170	2074	2074	72	71
2SRC	src	fgfr1	6319	6311	159	159	4550	4537	120	120
1KIM	tk	pnp	891	889	22	22	1036	1033	50	50
1VR2	vegfr2	p38	2906	2903	88	88	9141	9128	454	454
3DFR	dhfr	gart	8367	8354	410	410	879	877	40	40
1C2T	gart	dhfr	879	877	40	40	8367	8354	410	409
1F0R	fxa	thrombin	5745	5728	146	146	2456	2404	72	72
1BA8	thrombin	fxa	2456	2404	72	72	5745	5728	146	146
1BJU	trypsin	hivpr	1664	1615	49	49	2038	2029	62	62
1EVE	ache	hmga	3892	3868	107	107	1480	1475	35	23
1AH3	alr2	ampc	NA	NA	NA	NA	NA	NA	NA	NA
1XGJ	ampc	alr2	786	786	21	21	995	992	26	26
1Q4G	cox1	sahh	911	910	25	25	1346	1345	33	33
1CX2	cox2	na	13289	13259	426	423	1874	1862	49	49
1A8I	gpb	hivrt	2140	2136	52	52	1519	1513	43	42
1HPX	hivpr	trypsin	2038	2029	62	62	1664	1615	49	49
1RT1	hivrt	gpb	1519	1513	43	43	2140	2136	52	52
1HW8	hmga	ache	1480	1475	35	23	3892	3867	107	107
1P44	inha	parp	3266	3255	86	86	1351	1350	35	35
1A4G	na	cox2	1874	1862	49	49	13289	13265	426	423
1EFY	parp	inha	1351	1350	35	35	3266	3256	86	86
1B8O	pnp	tk	1036	1035	50	50	891	890	22	22
1A7A	sahh	cox1	1346	1345	33	33	911	908	25	25

^aShaded entries indicate DUD sets in which less than 90% of the initial (total) ligands yielded a final (dockd) pose.

Table B-3. Enrichment completion statistics using DUD SUP preparation.

PDB	DUD pairing		Normal				Semi-random			
			decoys		actives		decoys		actives	
	Normal	Semi-rand	initial	dockd	initial	dockd	initial	dockd	initial	dockd
1O86	ace	ada	1797	1787	49	49	927	925	39	39
1NDW	ada	ace	927	925	39	39	1797	1787	49	49
1H1D	comt	pde5	468	467	11	11	1978	1970	88	88
1XP0	pde5	comt	1978	1969	88	88	468	467	11	11
2AO6	ar	rxr	2854	2807	79 ^a	71	750	677	20	16
1L2I	er_agon	mr	2570	2560	67	65	636	608	15	10
3ERT	er_antag	ppar	1448	1437	39	39	3127	3116	85	85
1M2Z	gr	pr	2947	2663	78	56	1041	1017	27	22
2AA2	mr	er_agon	636	599	15	9	2570	2561	67	65
1FM9	ppar	er_antag	3127	3116	85	85	1448	1437	39	38
1SR7	pr	gr	1041	1014	27	24	2947	2839	78	65
1MVC	rxr	ar	750	708	20	20	2854	2813	79	73
1CKP	cdk2	pdgfrb	2074	2074	72	71	5980	5977	170	170
1M17	egfr	hsp90	15996	15990	475	475	979	978	37	37
1AGW	fgfr1	src	4550	4537	120	120	6319	6311	159	159
1UY6	hsp90	egfr	979	978	37	37	15996	15990	475	475
1KV2	p38	vegfr2	9141	9127	454	454	2906	2903	88	88
MODL	pdgfrb	cdk2	5980	5977	170	170	2074	2074	72	71
2SRC	src	fgfr1	6319	6311	159	159	4550	4537	120	120
1KIM	tk	pnp	891	890	22	22	1036	1035	50	50
1VR2	vegfr2	p38	2906	2903	88	88	9141	9128	454	454
3DFR	dhfr	gart	8367	8354	410	410	879	878	40	40
1C2T	gart	dhfr	879	878	40	40	8367	8354	410	410
1F0R	fxa	thrombin	5745	5728	146	146	2456	2404	72	72
1BA8	thrombin	fxa	2456	2404	72	72	5745	5728	146	146
1BJU	trypsin	hivpr	1664	1615	49	49	2038	2029	62	62
1EVE	ache	hmga	3892	3868	107	107	1480	1475	35	23
1AH3	alr2	ampc	995	992	26	26	786	786	21	21
1XGJ	ampc	alr2	786	786	21	21	995	991	26	26
1Q4G	cox1	sahh	911	910	25	25	1346	1344	33	33
1CX2	cox2	na	13289	13259	426	423	1874	1862	49	49
1A8I	gpb	hivrt	2140	2135	52	50	1519	1504	43	38
1HPX	hivpr	trypsin	2038	2029	62	62	1664	1615	49	49
1RT1	hivrt	gpb	1519	1513	43	42	2140	2136	52	52
1HW8	hmga	ache	1480	1475	35	23	3892	3868	107	107
1P44	inha	parp	3266	3256	86	86	1351	1350	35	35
1A4G	na	cox2	1874	1862	49	49	13289	13263	426	423
1EFY	parp	inha	1351	1350	35	35	3266	3256	86	86
1B8O	pnp	tk	1036	1017	50	48	891	887	22	21
1A7A	sahh	cox1	1346	1329	33	33	911	787	25	23

^aShaded entries indicate DUD sets in which less than 90% of the initial (total) ligands yielded a final (dockd) pose.

Global Enrichment Statistics (Total AUC). Table B-4 shows the overall global enrichment results (AUC values) for both the native and semi-random pairing using receptor structure derived from the protein databank (PDB) or supplied by the organizers (SUP) using actives and decoys from DUD or actives from WOMBAT. AUC values for individual systems are shown as bar plots in Figure B-4. The results for the native pairings yield, in the case of the DUD PDB runs (Figure B-4a, Table B-4), individual AUC values which range from 0.29 (bad enrichment) to 0.96 (good enrichment). For the DUD SUP results the max AUC value at 0.90 is not as high (Figure B-4b, Table B-4). Interestingly, both receptor preps using DUD actives and decoys yield nearly identical average AUC values of 0.60 and 0.59. This is a somewhat surprising result as the different preps use different partial charge assignments (FF99SB²³ vs Gasteiger¹⁵³), hydrogen orientation, and protonation states of the receptor (as discussed further below). For the WOMBAT results, average enrichment for the native pairings is worse than random (DUD SUP = 0.42, DUD PDB = 0.45). And, min AUC values for are significantly lower (0.13 and 0.16) than the corresponding DUD values (0.29 and 0.21). This is likely a function of the fact that two of the WOMBAT ROC curves (pde5 and ppar) perform significantly worse than random, a phenomenon not generally seen in the other ROC curves, as is discussed in more detail in the subsequent subsection (see ROC Curve Analysis). In addition, the much smaller WOMBAT dataset size (~10) vs DUD (~40) exacerbates this difference.

Table B-4. Global enrichment (total AUC) for native and semi-random pairings.

	Native pairing				Semi-random pairing			
	DUD lig		WOMBAT lig		DUD lig		WOMBAT lig	
AUC ^a	PDB	SUP	PDB	SUP	PDB	SUP	PDB	SUP
avg	0.60	0.59	0.42	0.45	0.48	0.46	0.39	0.40
std	0.17	0.17	0.17	0.16	0.18	0.17	0.15	0.15
median	0.56	0.56	0.41	0.50	0.48	0.46	0.41	0.44
max	0.96	0.90	0.60	0.61	0.84	0.85	0.61	0.62
min	0.29	0.21	0.13	0.16	0.15	0.10	0.08	0.09

^aTotal AUC calculated using FPR and TPR [0, 1]

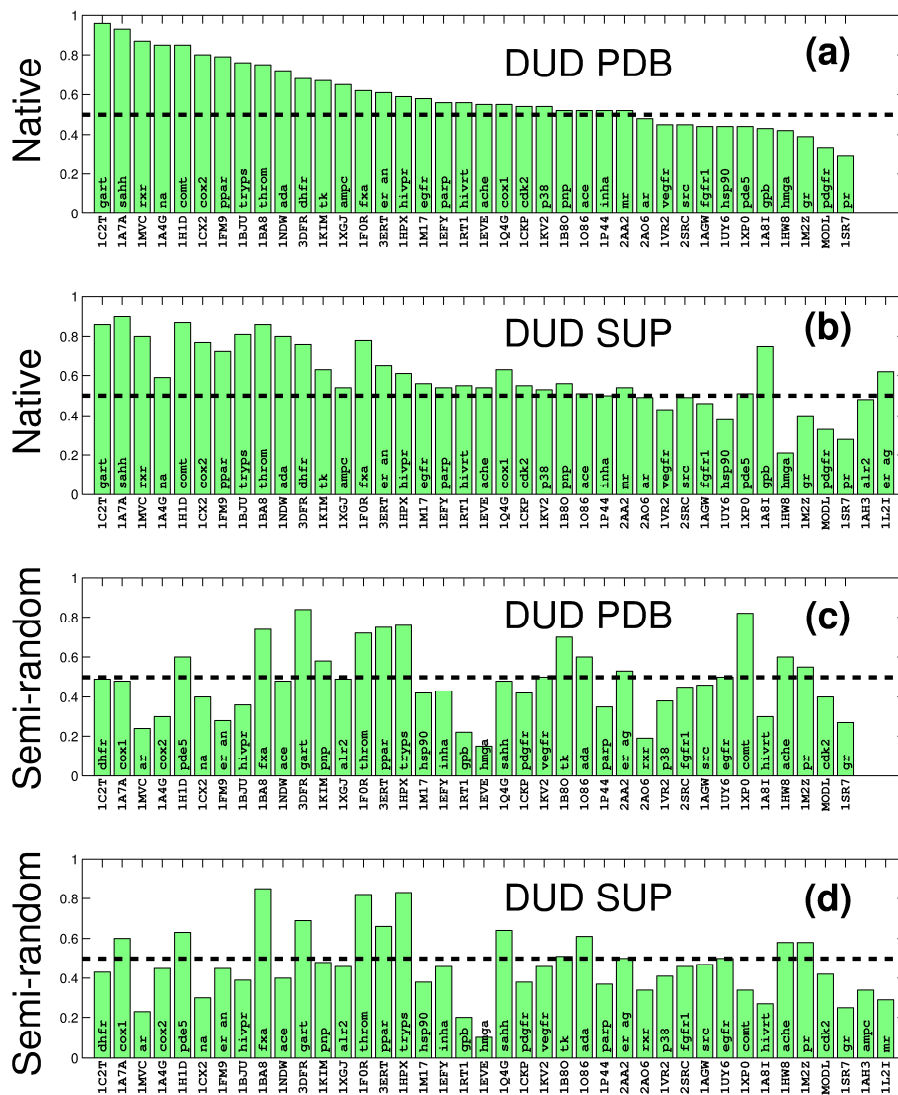


Figure B-4. Bar plots showing AUC values sorted from high to low using SB/DUD PDB naive pairing results in panel a.

As expected, results from semi-random experiments (Table B-4, Figure B-4c-d) show average AUC values which are consistently worse than random (avg AUC < 0.50). However, as discussed further below, for many semi-random pairings good cross-enrichment is observed, especially for receptors within the same protein family. This is reflected in the relatively high max AUC values which are in the range 0.84-0.85 (Table B-4) although they are lower than the native pairing (0.90-0.96). For DUD actives, average AUC drops from the higher 0.59-0.60 for the native pairing to 0.46-0.48 for the semi-random pairing. However, for WOMBAT actives, average AUCs are only marginally higher for the native pairing (0.42-0.45), compared to the semi-random pairing (0.39-0.40). This could be due to the fact that 7/10 of WOMBAT semi-random pairings are between proteins in the same family. As with the native pairings, there is striking agreement for the semi-random set, between DUD PDB and DUD SUP results, despite differences in the two different structure preparations.

Individual AUC Values and Group-based Analysis (Heatmaps). Figure B-4 shows bar plots of individual AUC values, comprising the total values in Table B-4, for all pairings for both structure preparations. Here, the plots are sorted by descending total AUC using the DUD PDB native pairing results. Figure B-5 shows a matrix representation of the data (heatmap) for the DUD PDB results which facilitates, for example, determining if a related group (or family) yields good or poor enrichment. Similar results are seen for the DUD SUP runs (Figure B-6). In Figure B-5, receptors are labeled according to the pdb code (and group) on the y-axis and ligands are labeled according to the DUD system name on the x-axis. The entries in Figure B-5 indicate native pairings (diagonal elements) and semi-random pairings (off diagonal elements). Red = worse than random (AUC < 0.5), green = greater than or equal to random (AUC ≥ 0.5), and white = non-paired systems. See Table B-5 for a detailed breakdown of systems which fall into

the different categories (better, equal, or worse than random). Figure B-5 groupings highlight the fact that many of the so called semi-random pairings are in fact not random but are pairings between related protein structures. Thus, for many semi-random pairings, good enrichment would not be unexpected. Exceptions would likely include the "other enzyme" group, and to a lesser extent the metalloenzyme group. Ideally, experiments using all receptors paired with all active-decoys sets should be performed (entire matrix) to derive better statistics.

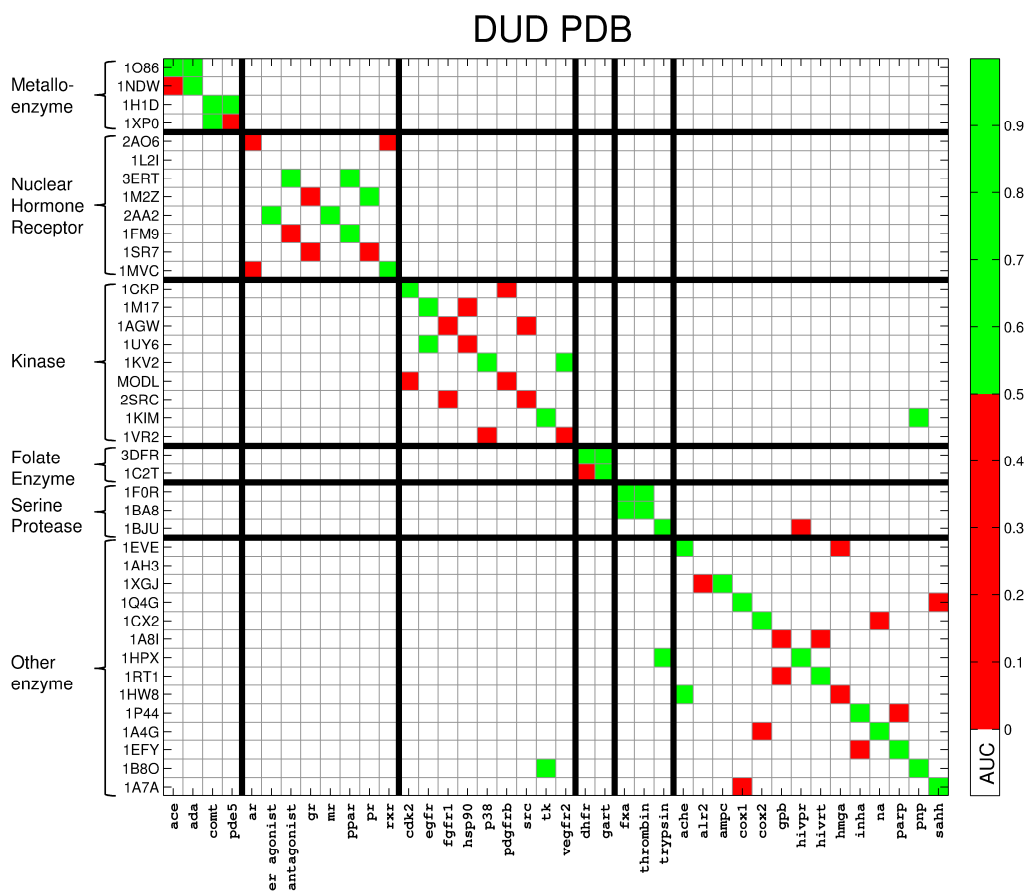


Figure B-5. Heatmaps showing enrichment study which employ SB/DUD PDB results and DUD actives and decoys. Receptors are labeled by the pdb code and family on the y-axis. Ligands are labeled by the DUD system name on the x-axis. Red = worse than random ($AUC < 0.5$), green = greater than or equal to random ($AUC \geq 0.5$), and white = non-paired systems.

As evident by the median value of 0.56 (27/38 have $AUC \geq 0.5$) for the DUD PDB results in Table B-4, most systems yield better than random enrichment (green squares) for the

native pairings occurring on the diagonal: metalloenzymes = 3/4, nuclear hormone receptor = 4/7, kinase = 4/9, folate enzyme = 2/2, serine protease = 3/3, other enzymes = 11/13. For the analogous off-diagonal experiments the median AUC value 0.48 (15/38 have $AUC \geq 0.5$) indicates a roughly even split between good and bad enrichment: metalloenzymes = 3/4, nuclear hormone receptor = 3/7, kinase = 2/8, folate enzyme = 1/2, serine protease = 2/2, other enzymes = 1/11, misc pairings = 3/4. Importantly, off-diagonal elements for the "other enzymes" group, consisting of mostly unrelated proteins, yields poor enrichment which is to be expected. This group is probably the most useful overall as an actual null hypothesis test set for evaluating virtual screening.

Of all the groups evaluated, the serine protease group yields the best overall enrichment (5/5) for the five pairings evaluated (3 diagonal, 2 off-diagonal). Cross-enrichment was also observed for serine proteases in the original DUD manuscript (Table 2 in Ref.¹⁰⁸). This is not surprising given that trypsin, factor Xa and thrombin are very similar proteins. In fact, trypsin has been successfully used as a template for developing thrombin and factor Xa inhibitors.^{130,154} Interestingly, while the trypsin receptor (1bjv) with HIV protease actives (hivpr) shows no enrichment (1bjv-hivpr pairing, Figures B-4 c and d, B-5) the use of the HIV protease receptor (1hpx) with trypsin actives (trypsin) does show enrichment (1hpx-trypsin pairing, Figures B-4 c and d, B-5).

Global Enrichment for DUD SUP set. Figure B-6 shows enrichment results for the native and semi-random pairing for the DUD SUP set. This figure corresponds with the Figure B-5 showing the same data for the DUD PDB set. Despite differences between the two setups, the heatmaps yield remarkably similar results for both preps. Table B-5 shows the three counts (better, equal or worse than random) for results obtained using DUD PDB and DUD SUP

preparations. As is noted in the heatmaps discussion section, the native results yield more better than random results (27/38, 29/40) compared to semi-random (13/38, 12/40).

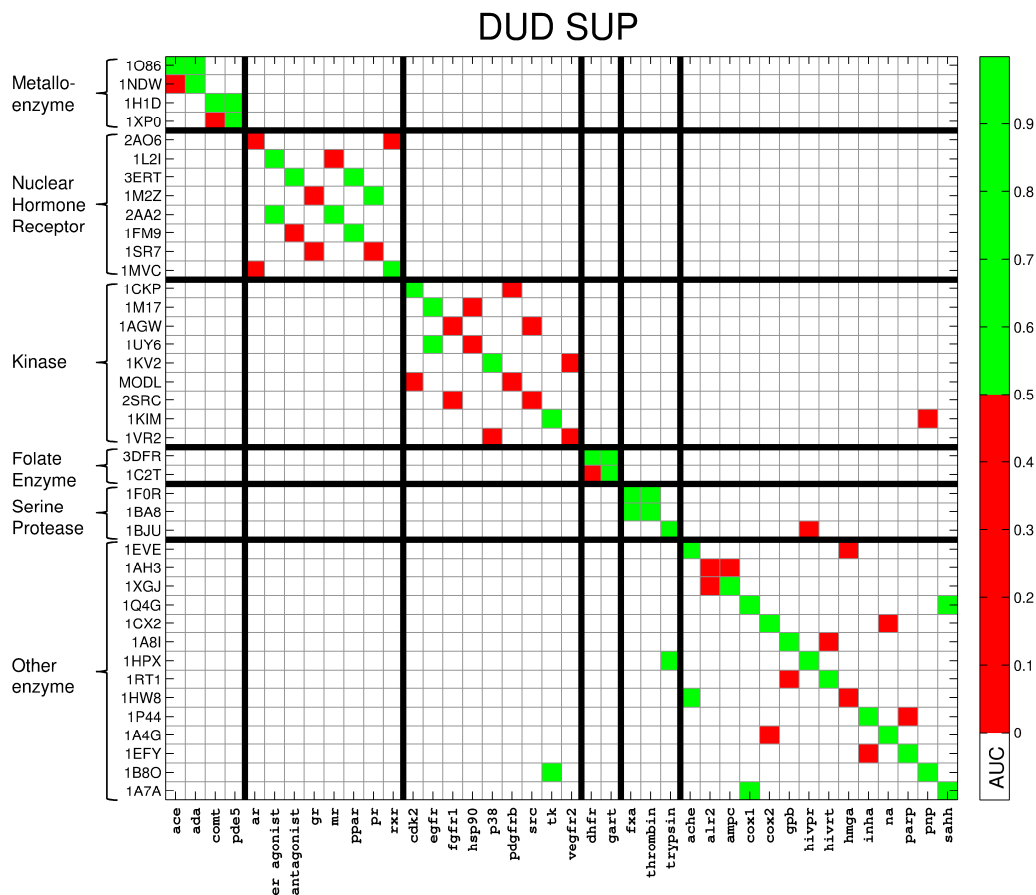


Figure B-6. Heatmaps showing enrichment study which employ the DUD SUP receptor and DUD actives and decoys. Receptors are labeled by the pdb code and family on the y-axis. Ligands are labeled by the DUD system name on the x-axis. The area under the ROC curve for each receptor-ligand pairing is used as a measure of enrichment. No data available (white), better than random enrichment (green), or worse than random enrichment red is indicated by the colored cells.

Table B-5. Global breakdown of DUD systems into three sets (better, equal or worse than random) using the AUC to define enrichment.

type	better	equal	worse
DUD PDB, Native	27	0	11
DUD PDB, Semi-rand	13	2	23
DUD SUP, Native	29	1	10
DUD SUP, Semi-rand	12	2	26

In contrast to serine protease, the kinase group shows poor native and semi-random enrichment. An interesting observation is that epidermal growth factor receptor (egfr) ligands show enrichment with the heat shock protein (1uy6) receptor while the converse pairing (1m17-hsp90) is worse than random enrichment for both preparations (Figure B-4a-d, B-5, B-6). However, since the 1uy6-hsp90 native pairing also had sub-random enrichment (Figures B-4a-b, B-5) this result may not be unexpected if the set of actives and decoys dominate enrichment behavior as discussed below. Surprisingly, when thymidine kinase is paired with purine nucleoside phosphorylase actives, and vice versa, (1kim-pnp and 1b8o-tk pairings, Figure B-4c-d, B-5), these two unrelated systems enrich one another. The authors of the original DUD paper also noted that these two enzymes yielded cross-enrichment (Table 2 in DUD paper¹⁰⁸) in addition to thymidine kinase being a promiscuous target.

System-Specific Analysis: DUD PDB vs DUD SUP Preparations. Although the overall average AUC (0.59, 0.60) for native pairings is essentially the same, examination of the individual results in Figure B-4a-b reveals that there can be differences in enrichment depending on which structure preparation (DUD PDB vs DUD SUP) is used. The DUD PDB prep starts with raw pdb files from scratch (see Methods) while the DUD SUP structures, with the exception of hydrogens added by the organizers, were from the original DUD database.¹⁰⁸ Examples include the good enrichment seen for system 1a8i but only when using the DUD SUP receptor

prep (Figure B-4a vs B-4b, 0.43 vs 0.75). Examination of the glycogen phosphorylase receptor in this system show that the DUD SUP prep contains active site waters in the binding site (Figure B-7b) which are absent in the DUD PDB prep. The known actives may use the water-mediated interactions to their advantage although a more detailed study should be performed to determine the actual importance. For certain systems, binding site water is known to play critically important roles. Another, more subtle example is shown in Figure B-7a for 1hw8. Here, the native pairings appear to favor the DUD PDB receptor prep (Figure B-4a vs B-4b, 0.42 vs 0.21) although the average AUC values are both below random. The hmg coa reductase receptor in this system has four chains, labeled A-D in the original pdb structure, with four occupied binding sites. For the DUD PDB prep, the binding site used was at the interface of chains A-B in contrast to the DUD SUP prep which was at the interface of chains C-D. Differences between the two sites (see Figure B-7a) involve a conformational change of methionine and alternate orientations of two cysteine thiol hydrogens which appear to favor the DUD PDB prep. These structural differences are likely a result of a nearby co-factor originally present in the C-D site.

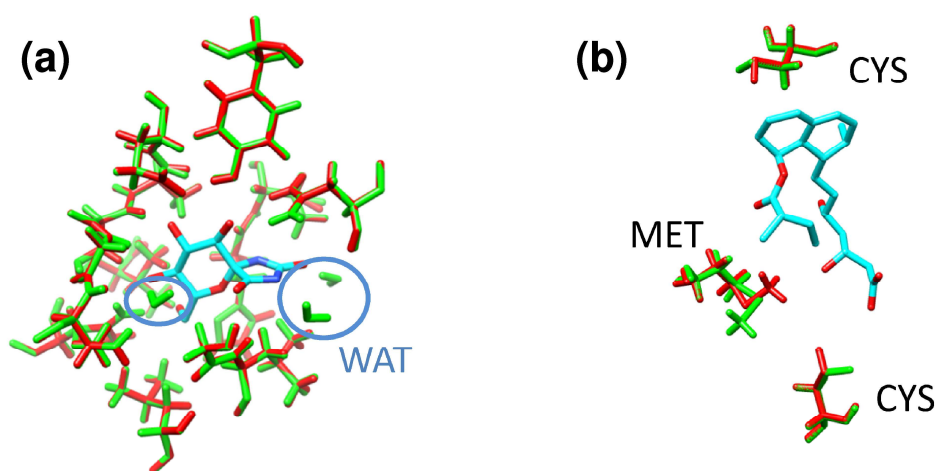


Figure B-7. Differences between DUD PDB (red) and DUD SUP (green) structure preps for (a) 1a8i and (b) 1hw8. Native ligands are shown in cyan. Binding site waters in the DUD SUP prep for 1a8i indicated by blue circles.

Differences are also observed for systems 1f0r and 1a4g. For factor Xa (1f0r), protonation state differences of a histidine near the binding site could influence the computed enrichments (DUD SUP = 0.78, DUD PDB = 0.62). For neuraminidase (1a4g), electrostatics are known to be especially important⁸³, thus the use of dramatically different charge models (F99SB²³ vs Gasteiger¹⁵³) would likely influence the results (DUD PDB = 0.85, DUD = SUP = 0.59). Finally, although differences in a binding site environment would normally be expected to affect enrichment, for some systems this is not always observed. For example, for catechol o-methyl transferase in 1h1d, high enrichment is observed using both preps (AUC 0.85, 0.87) despite the fact the DUD SUP prep is missing part of the co-factor (S-adenyl methionine, SAM) included in the DUD PDB prep. Overall, the results highlight how use of different receptor sites or structure preparations, and by analogy alternative crystallographic coordinates of the same receptor can influence enrichment.

ROC Curve Analysis. Figure B-8a shows standard ROC curves for the native DUD pairings (see Figures B-9 and B-10 for semi-random pairings). The curves are sorted from high to low according to total AUC using SB/DUD PDB data with SB/DUD PDB shown in black and SB/DUD SUP shown in gray. These comparisons allow, at a glance, which systems yield overall good, reasonable, poor, or early enrichments. Most of the curves have strikingly similar shapes despite the fact that two different structure preps were evaluated. In contrast, 1hw8 and 1a8i show dramatically different shaped ROC plots which ultimately lead to the larger differences discussed above. Systems with poor overall enrichment may still have good early enrichment which is characterized by relatively steep upward sloping curves starting on the left right hand corner of the ROC plots which then decrease as more and more of the database is covered. Although the left part of the curve will be above the random line the right part of the curve may

approach the random (1m17, 1kv2) or even dip below (1hpx, 1o86). This phenomena, interestingly, seems especially true for kinases 2src, 1vr2, 1xp0, and 1agw. Indirectly, the inherent plasticity of kinase binding sites could adversely affect enrichment due to known induced fit effects which can be ligand dependent. For example, the few actives which might favor a particular kinase conformation could be scored favorably but in absence of receptor flexibility the remaining pool of actives might not yield favorable scores thus accounting for the observed early, but not global, enrichment. Other systems with short lasting early enrichment include MODL (based on a homology model) and 1sr7. For DUD vs WOMBAT native pairings two systems in particular, 1xp0 and 1fm9, stand out as having significantly different ROC curves (Figure B-8b). The ROC curve for 1fm9 shows good enrichment behavior when using DUD but a sub-random curve with WOMBAT. For 1xp0, although the overall DUD results are not much better than random the WOMBAT results are always sub-random. Additional analysis would be required to determine the cause of these enrichment differences.

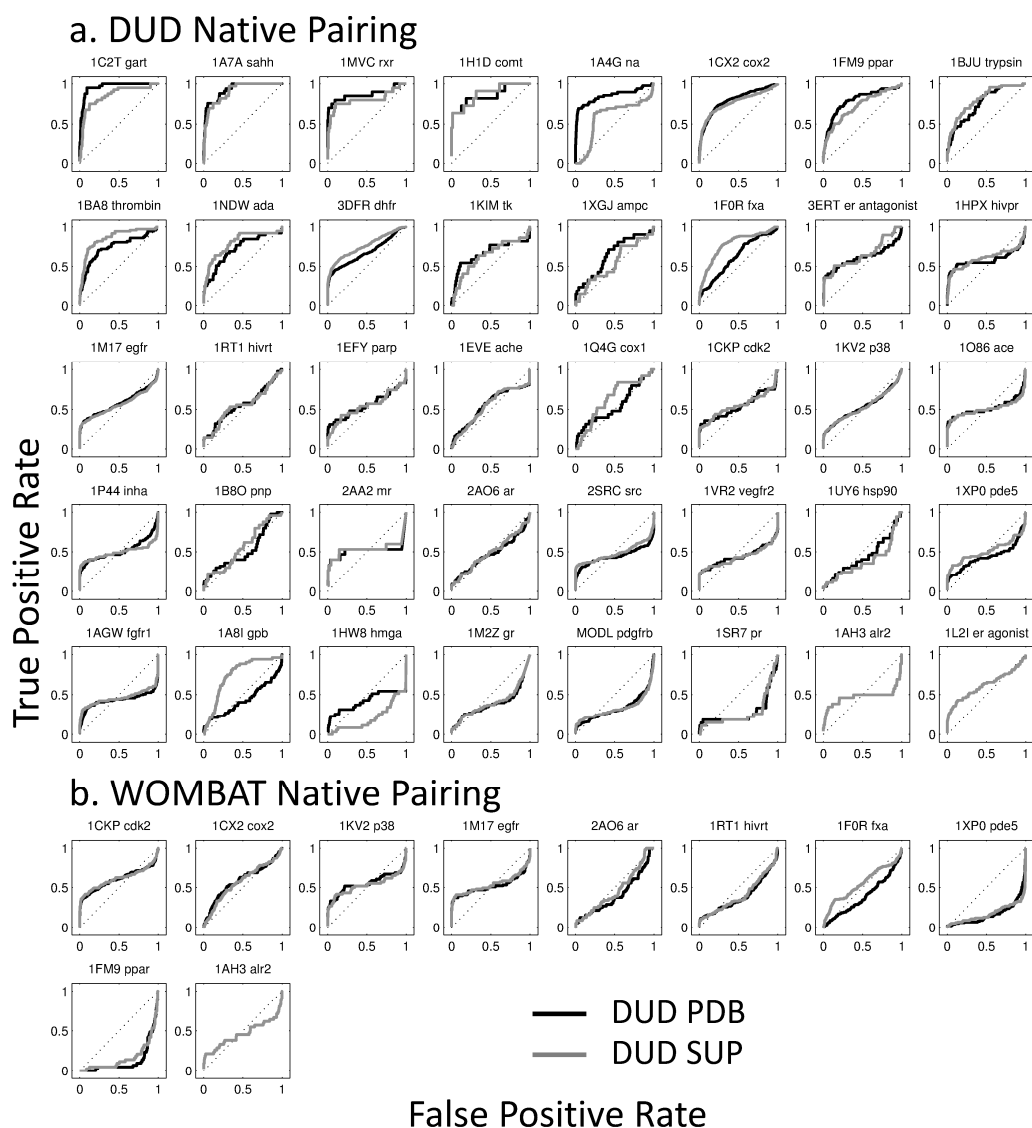


Figure B-8. (a) ROC curves for the 40 DUD families. (b) ROC curves for the 10 Wombat families (Wombat ligands + DUD decoys). ROC curves sorted from high to low according to total AUC using SB/DUD PDB data with SB/DUD PDB in black and SB/DUD SUP in gray.

ROC Curves for Semi-Random Pairing. Overall, much poorer enrichment is observed, as expected for the semi-random pairing (Figure B-9), compared to Figure B-8. For each curve, the results for both DUD SUP and DUD PDB are shown. Good enrichment is observed for pairings between related enzyme families (serine proteases, nuclear hormone receptors, and folate enzymes).

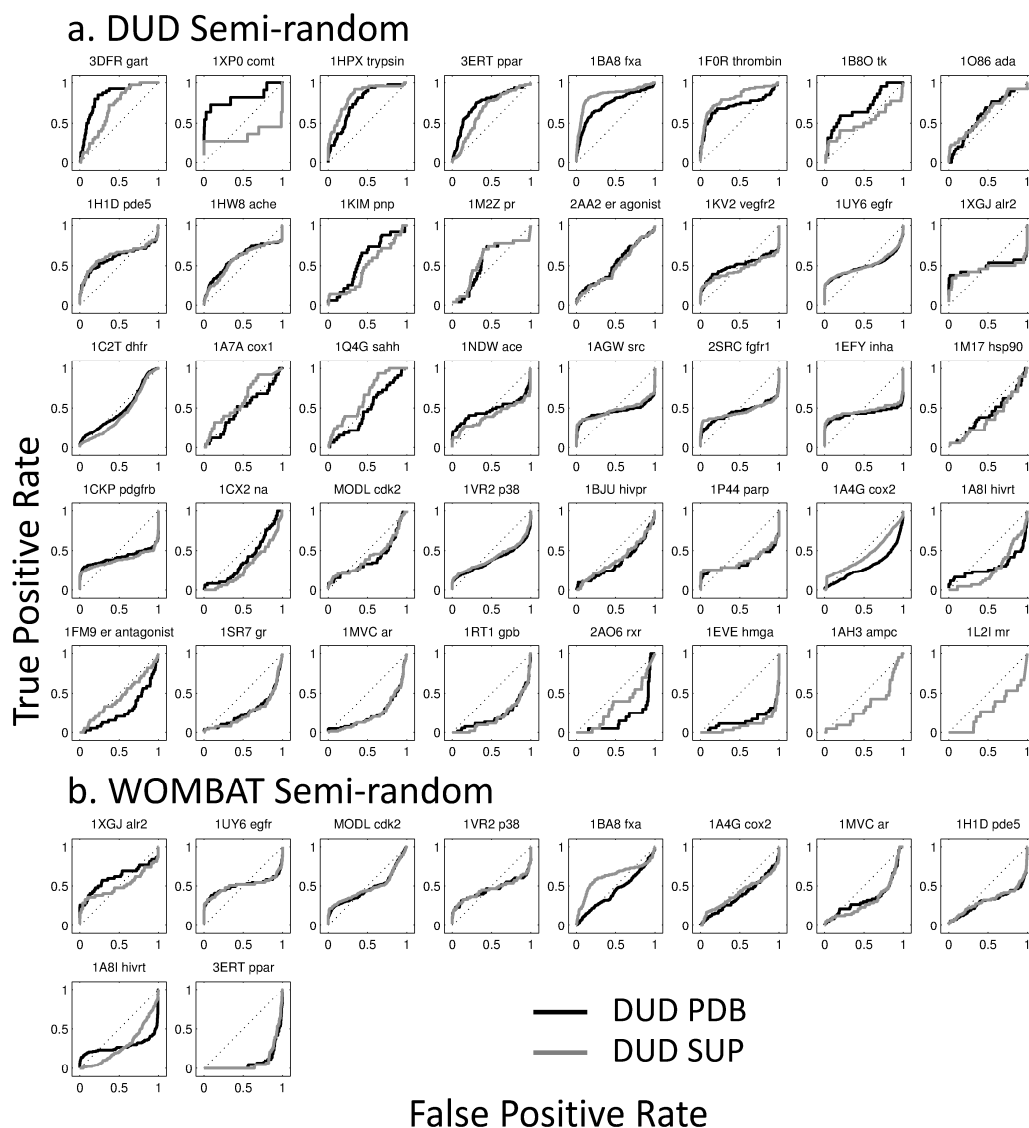


Figure B-9. Semi-random pairings run with DOCK6.6. (a) ROC curves for the 40 DUD families. (b) ROC curves for the 10 WOMBAT families (WOMBAT ligands DUD decoys). Results are sorted by Stony Brook structures results in black and gray is the provided organizer structures. Receptors are labeled by the pdb code and family on the y-axis. Ligands are labeled by the DUD system name on the x-axis.

Comparing Native vs Semi-random. An interesting observation is that a given ligand set (actives + decoys) can yield similarly shaped ROC curves in two receptors suggesting that shape, in some instances, can be driven by properties inherent in the ligand set. Two examples of this phenomenon are the kinase pair egfr and hsp90 (Figure B-10a) and the serine proteases pair fxa and thrombin (Figure B-10b). The egfr ligand set shows strong early enrichment with

both receptors. Conversely, the hsp90 ligand set yields curves that are slightly below random. As previously mentioned, fxa and thrombin have strong enrichment with the DUD SUP performing better than DUD PDB for all four combinations. On the other hand, this observation does not appear to hold when the receptors are unrelated. As an example, hivpr (1hpx) more strongly enriches the trypsin ligands than its cognate ligands (Figure B-10c). Conversely, the trypsin receptor (1bjv) performs worse than random for the hivpr ligands as noted above in the heatmap discussion. More studies are needed to explore these issues in greater detail. For completeness, Figure B-9 contains ROC curves for all semi-random pairings and may be used to aid in making additional comparisons.

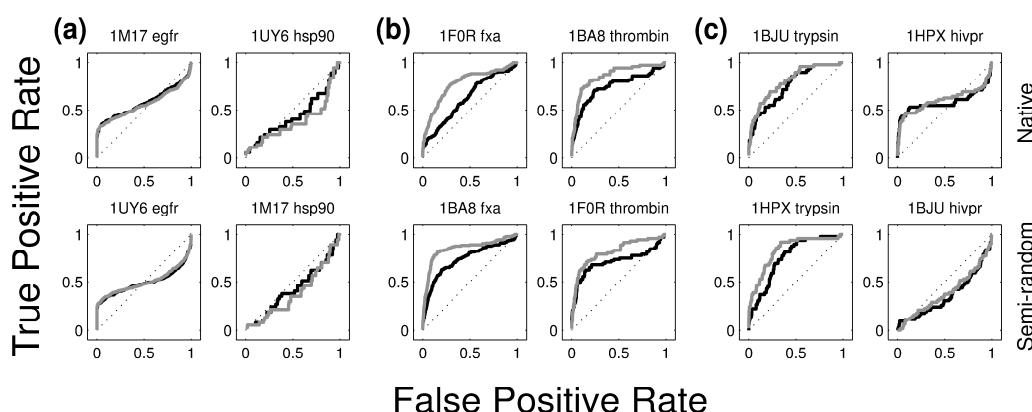


Figure B-10. ROC curves for 3 pairs of DUD families (a) egfr and hsp90, (b) fxa and thrombin, (c) trypsin and hivpr. Top panels show the native pairing. Bottom panels show the semi-random pairing.

Early Enrichment. Table B-6 lists average early enrichment statistics, in terms of 0.1%, 1.0%, and 2.0% of the database examined, using several metrics including %TPR, %FPR, %AUC, and FE (see Methods for definitions). Which is the best metric for quantifying early enrichment is still an active area of research. Here, to help gauge enrichment, values for the expected random and best possible cases are also reported. Results from the 0.1% bracket are reported at the request of the organizers but in some cases may be too

small to be meaningful considering the small number of molecules contained in some DUD datasets. Analysis presented below is focused on results obtained using DUD at 1.0% and 2.0% of the database. WOMBAT data is provided for completeness.

Importantly, average early enrichment is always significantly better than the hypothetical random results. For example, comparison of FE values using DUD PDB native pairings (Table B-6) yields: 9.99 (1.0%), and 10.09 (2.0%). This is an order of magnitude improvement over random FE (1.0) regardless of the theoretical maximum which will decrease as larger percentages of the database are examined (Figure B-2, Table B-6). The much better than random FE results in these early regions mirror the visual trends seen in the ROC curves shown above (Figure B-8) and should favorably benefit virtual screening efforts. The same trends are observed for the other three metrics (%TPR, %FPR, %AUC) with the computed enrichments being consistently better than random.

A comparison of results using the two different structure preps similarly shows consistently improved early enrichments for the DUD PDB prep. For example, %TPR values for DUD PDB vs DUD SUP yield: 15.14 vs 13.02 (1.0%) and 20.40 vs 17.33 (2.0%). Similar trends are observed for %AUC: 4.99 vs 4.17 (1.0%) and 20.19 vs 17.45 (2.0%) as well as FE: 9.99 vs 8.34 (1.0%) and 10.09 vs 8.72 (2.0%). In addition, %FPR values are lower, which is desirable: 0.56 vs 0.62 (1.0%), and 1.43 vs 1.52 (2.0%). Overall, use of the DUD PDB prep appears to yield better early enrichment results despite the overall average AUCs being the same at 0.60 and 0.59 (Table B-4). Finally, the native pairings yield higher early enrichment values compared to the semi-random pairings which is to be expected. For example, native vs semi-random DUD PDB %AUC results yield 4.99 vs 3.53 (1.0%) and 20.19 vs 12.49 (2.0%). The %TPR results for these runs yield the same trend: 15.14 vs 8.58 (1.0%) and 20.40 vs 11.47 (2.0%).

Table B-6. Averaged early enrichment at 0.1%, 1.0%, and 2.0% of the databases examined.

Data- base	Metric ^a	Best ^b	Ran ^c	Native pairing				Semi-random pairing			
				DUD		WOMBAT		DUD		WOMBAT	
				PDB	SUP	PDB	SUP	PDB	SUP	PDB	SUP
0.1%	%TPR	100	0.1	2.57	2.34	2.59	2.29	1.90	1.40	2.90	2.36
0.1%	%FPR	0	0.1	0.03	0.03	0.06	0.06	0.04	0.06	0.05	0.06
0.1%	%AUC	10	0.005	0.01	0.01	0.10	0.07	0.01	0.00	0.07	0.06
0.1%	FE	2000	1.0	2.00	2.88	19.51	13.50	1.73	0.83	13.00	11.79
1.0%	%TPR	100	1.0	15.14	13.02	9.69	9.29	8.58	7.71	8.43	7.65
1.0%	%FPR	0	1.0	0.56	0.62	0.78	0.75	0.77	0.79	0.76	0.82
1.0%	%AUC	100	0.5	4.99	4.17	5.14	4.49	3.53	3.17	4.10	3.91
1.0%	FE	200	1.0	9.99	8.34	10.29	8.97	7.06	6.35	8.20	7.83
2.0%	%TPR	100	2.0	20.40	17.33	11.86	12.38	11.47	10.68	10.65	9.98
2.0%	%FPR	0	2.0	1.43	1.52	1.67	1.66	1.71	1.71	1.69	1.77
2.0%	%AUC	200	2.0	20.19	17.45	14.26	13.74	12.49	11.63	13.12	12.46
2.0%	FE	100	1.0	10.09	8.72	7.13	6.87	6.24	5.82	6.56	6.23

^aMetrics for %TPR and %FPR indicate percentages on [0,100] while %AUC is on [0, 10000] calculated using %FPR and %TPR, FE is unitless. ^bBest = best possible enrichment. ^cRan = random enrichment. Best and Ran values are estimations used for comparison purposes.

B.4 Enrichment Study Conclusions.

To evaluate the ability of DOCK6 to enrich actives vs decoy molecules, the standard DOCK grid scoring function was used to screen 38-40 systems (Table B-4, Figure B-4 and B-5) contained in the DUD database using either organizer supplied receptor coordinates (DUD SUP) or the protein data bank (DUD PDB). In addition to the standard DUD sets a subset of 10 systems employing WOMBAT actives was performed. DUD PDB results (Figure B-4a, Table B-4) using native pairings yield AUC values obtained from ROC curve analysis (Figure B-8) ranging from 0.29 (bad enrichment) to 0.96 (good enrichment) with an average AUC of 0.60 (27/38 have $AUC \geq 0.5$). ROC curve analysis visually indicates good early enrichment for most systems which was quantified using several metrics including, %TPR, %FPR, %AUC, and FE, at 0.1%, 1.0%, and 2.0% of the database examined (Table B-6). In particular, at 1.0%, and 2.0% an order of magnitude improvement over random FE (1.0) is observed for the DUD PDB native pairing results. Overall, the early enrichment trends are encouraging and consistent with

previous studies^{2,4,16,155,156} in which DOCK was used to successfully identify active lead molecules through virtual screening.

As expected, analogous DUD PDB results for semi-random pairings show a lower average AUC (0.48). However, appreciable enrichment is observed among groups of related receptors (Figure B-5). In contrast, the miscellaneous other enzyme group, comprised of unrelated receptors, shows only 1/11 systems with AUC > 0.5. Some system specific analysis revealed possible contributors to differential enrichments according to which of the two receptor preparation protocols was used (DUD PDB vs DUD SUP protocols). Observed differences include alternate side chain conformations, numbers of binding site waters, partial atomic charges, and protonation states.