

# **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.**

**Structure-Based Drug Design Targeting HIVgp41**

A Dissertation Presented

by

**Lingling Jiang**

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

**August 2015**

**Stony Brook University**

The Graduate School

**Lingling Jiang**

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**  
**Professor, Applied Mathematics and Statistics**

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

**Yuefan Deng**  
**Professor, Applied Mathematics and Statistics**

**Jonathan G. Rudick**  
**Assistant Professor, Chemistry**

This dissertation is accepted by the Graduate School

Charles Taber  
Dean of the Graduate School

Abstract of the Dissertation

**Structure-Based Drug Design Targeting HIVgp41**

by

**Lingling Jiang**

**Doctor of Philosophy**

in

**Applied Mathematics and Statistics**

**(Computational Biology)**

Stony Brook University

**2015**

This dissertation presents method development and application of computational procedures for structure-based drug design, with a particular focus on the clinical relevant drug target HIVgp41.

Chapter 1 introduces the main computational techniques used in this study, including atomic-level molecular dynamics simulations, free energy calculations, and molecular docking. The importance of targeting HIVgp41, a viral envelope protein involved in viral entry and membrane fusion, is also discussed. In particular, the only FDA approved gp41 inhibitor, a peptide called T20, suffers from side effects, expense, and is subject to drug resistance. Thus, improved understanding of the binding mechanisms of T20 is of great interest, which provides the motivation for the computational work in this dissertation.

Chapter 2 describes a newly implemented DOCK scoring function termed pharmacophore matching similarity (FMS) score. By matching pharmacophore features in references, such as those in known peptide inhibitors of gp41, FMS score can help guide docking of small molecules to yield hits with desired properties. This new DOCK method, when used alone (FMS) and in combination with the standard single grid energy score (FMS+S<sub>GE</sub>), is validated via pose reproduction, crossdocking and enrichment studies with desirable outcomes using large molecular docking testsets.

Chapter 3 presents additional in-depth analyses of the FMS case studies for enrichment, as well as post-processing of virtual screening results targeting both the gp41 hydrophobic and inner pockets.

Chapter 4 presents preliminary applications of FMS score for *de novo* design using small (focused) fragment libraries generated for 50 small molecule test cases as well as peptide test cases targeting the two gp41 pockets.

Chapter 5 reports molecular dynamics simulation and free energy calculation results for T20 in complex with gp41 (T20-gp41) to help characterize the biological effects of a series of primary and secondary mutations. Per-residue energetic analyses and structural characterization of end-point simulations were employed to identify key residues at the T20-gp41 binding interface. Importantly, good agreement with experimental activity trends was observed for a series of T20 analogs with three gp41 variants, using a thermodynamic integration protocol, which further supports the atomistic model.

Chapter 6 summarizes the work presented in the dissertation in terms of scientific impact, challenges, and future studies to further aid structure-based drug design targeting HIVgp41.

## Table of Contents

List of Figures .....	viii
List of Tables .....	xiv
List of Abbreviations .....	xvi
Acknowledgments .....	xviii
Chapter 1. Introduction.....	1
1.1 Therapeutic Drug Target for HIV-1 Fusion: HIVgp41 .....	1
1.2 Computational Structure-Based Drug Design .....	7
1.3 Classical Molecular Mechanics .....	11
1.4 Molecular Dynamics and Free Energy Calculations.....	14
1.5 Molecular Docking .....	19
1.6 Research Projects.....	22
Chapter 2. Pharmacophore-Based Similarity Scoring Method for DOCK.....	23
Abstract.....	23
2.1 Introduction .....	24
2.2 Theoretical Methods .....	27
2.2.1 Pharmacophore Definitions. ....	27
2.2.2 Pharmacophore Matching Similarity (FMS) Scoring Function.....	34
2.3 Validation Metrics and Computational Details .....	39
2.3.1 Pose Reproduction Details. ....	39
2.3.2 Crossdocking Details. ....	41
2.3.3 Enrichment Details.....	42
2.4 Results and Discussions .....	45
2.4.1 Pose Reproduction Results. ....	45
2.4.2 Quadrant Partitioning using FMS Score.....	51
2.4.3 Crossdocking Results. ....	56
2.4.4 Enrichment Results. ....	64
2.4.5 Case Studies Targeting EGFR, IGF-1R, and HIVgp41. ....	69
2.5 Conclusion.....	74
Chapter 3. FMS-guided Virtual Screen to HIVg41 .....	78
3.1 Introduction .....	78
3.2 Methods and Computational Details.....	79

3.3 Results and Discussion.....	82
3.3.1 Enrichment Study Analyses on Four Representative Systems .....	82
3.3.2 Rescoring Virtual Screening to HIVgp41 .....	85
3.4 Conclusion.....	89
Chapter 4. FMS-guided <i>de novo</i> Design to HIVg41 .....	91
4.1 Introduction .....	91
4.2 Methods and Computational Details.....	93
4.3 Results and Discussions .....	98
4.3.1 Focused library <i>de novo</i> runs: small molecule reference reproduction. ....	98
4.3.2 Focused library <i>de novo</i> runs: DOCK outcomes and molecular properties .....	102
4.3.3 Focused <i>de novo</i> tests: HIVgp41 hydrophobic pocket.....	107
4.3.4 Focused <i>de novo</i> tests: HIVgp41 inner pocket .....	113
4.3.5 Future experiment: generic library <i>de novo</i> tests.....	115
4.4 Conclusion.....	116
Chapter 5. Quantitative Characterization of T20 Variants Affinity and Mutational effects.....	118
Abstract.....	118
5.1 Introduction .....	119
5.2 Theoretical Methods and Computational Details .....	123
5.2.1 Free Energy Calculations Using Thermodynamic Integration .....	123
5.2.2 Footprint Analysis: Energy Decomposition to Uncover Ligand Binding Profile.....	124
5.2.3 Model Construction and MD Simulation Protocol .....	124
5.3 Results and Discussions .....	127
5.3.1 Endpoint Simulation Behavior: RMSD.....	127
5.3.2 TI Simulation Behavior: RMSD.....	129
5.3.3 TI Simulation Behavior: $dV/d\lambda$ .....	131
5.3.4 TI Simulation Behavior: Null Transformation .....	133
5.3.5 TI Relative Binding Energies: Correlation with Experimental Results .....	136
5.3.6 T20 Binding: Footprint Analyses of the Wild-Type System.....	139
5.3.7 Why S138A? .....	144
5.4 Conclusion.....	149
Chapter 6. Dissertation Summary: Scientific Impact, Challenge and Future Direction. ....	150
6.1 Development and Application of FMS Docking Protocol .....	150
6.1.1 Scientific Impact.....	150

6.1.2 Challenge, Related Work and Future Direction.....	151
6.2 Computational Investigation of HIVgp41-T20 Binding .....	152
6.2.1 Scientific Impact.....	152
6.2.2 Challenge, Related Work and Future Direction.....	153
6.3 Summary .....	154
Bibliography .....	155
Appendix A. FMS-guided DOCKing Protocol.....	167
Appendix B. Visualization of Pharmacophore Models .....	176
Appendix C. Lipid-Bound HIVgp41-T20 Complex Simulations .....	179



# List of Figures

## Chapter 1

- Figure 1-1.** HIV fusion can be categorized as four stages: (1) Native pre-entry stage when no interactions between the virus and the host cell are observed, (2) pre-hairpin stage, (3) membrane fusion, and (4) post fusion. Figure adapted from Hughson *et al.*<sup>11</sup> ..... 3
- Figure 1-2.** Modeling the binding of T20 to gp41 NHR. (a) Schematic representation of positional relationship of T20, C34, FP, N-HR, C-HR and TM. (b) Linear sequences of T20, C34 and N-HR helices and visualization of T20 primary (red arrow) and secondary (orange and blue arrow) mutation sites. Charged residues are indicated by “+” and “-“ signs..... 5
- Figure 1-3.** Flow chart of computational structure-based drug design. Computational techniques used are highlighted in blue boxes. .... 8
- Figure 1-4.** Molecular recognition: identification of ligand with favorable binding affinity to the target receptor. .... 10
- Figure 1-5.** Molecular mechanics energy function terms: (1) bond length, (2) bond angle, (3) torsion angle, (4) VDW and (5) electrostatics. .... 13
- Figure 1-6.** Molecular dynamics: updating 3D coordinates of atoms in the molecular system ..... 15
- Figure 1-7.** Evaluating transformation energy using thermodynamic integration method. (a) Coupled systems in different  $\lambda$  windows. (b) Theoretic  $dV/d\lambda$  curve for transformation energy evaluation. .. 17
- Figure 1-8.** Thermodynamic cycle for relative binding energy calculation between two ligands Lig<sub>A</sub> and Lig<sub>B</sub> with a receptor (Rec) to form a complex (Com). The cycle depicted equates the experimental relative binding energy with the difference in transforming one of the two ligands to another in the bound and unbound state using TI..... 18
- Figure 1-9.** (a) DOCK anchor-and-grow algorithm. (b) Docking to grid. .... 20

## Chapter 2

- Figure 2-1.** 2D representations for three approved drugs (top) and corresponding DOCK pharmacophore (ph4) models (bottom). Features include: (i) hydrogen bond acceptor in red, (ii) hydrogen bond donor in blue (iii) hydrophobic atom/group in cyan, (iv) aromatic ring center and direction in orange, (v) non-aromatic ring center and direction in yellow, (vi) negatively charged group center in green, and (vii) positively charged group center in magenta. Structures of nevirapine, erlotinib, and zanamivir from PDB codes 1VRT, 1M17, and 1A4G, respectively..... 26
- Figure 2-2.** Pharmacophore feature assignment for rings: (a) aromatic (close to planar) and (b) non-aromatic (not planar). Ring center-to-vertex vectors shown as dashed blue lines, individual normal

vector shown as solid blue lines, averaged normal vectors shown in solid red lines. The angle between the blue and the red vectors are compared to a threshold to determine the planarity of the ring.....	30
<b>Figure 2-3.</b> Number of pharmacophore features computed by DOCK FMS scoring function using the SB2012 docking testset (N=1043 molecules). .....	34
<b>Figure 2-4.</b> Example pharmacophore matches for aromatic rings showing: (a) well matched case with same labels, small distance, and similar vector directions, (b) not matched case with same labels, large distance, and similar vector directions, and (c) not matched case with same labels, small distance, and different vector directions. ....	36
<b>Figure 2-5.</b> Flow chart schematic outlining pharmacophore-based virtual screening in DOCK. ....	38
<b>Figure 2-6.</b> Validation metrics used to evaluate DOCK scoring functions. (a) Pose reproduction cases with different outcomes: Success (top, PDB code 3CPA), Score Fail (middle, PDB code 1V2W), and Sample Fail (bottom, PDB code 1GKC). Crystal poses in orange, best scored poses in magenta, best RMSD pose in cyan. (b) Representative crossdocking heatmap showing docking outcome as a function of docking all ligands (Lig1, Lig2, ... LigN) to all receptors (Rec1, Rec2, ... RecN), for an aligned group of proteins with nearly identical sequence homology. (c) Hypothetical database enrichment results showing a partitioning of data based on FMS score ranking (0 to 6) for a group of ligands (left bottom, magenta curve) comprised of a known active ligand set (left middle, blue curve) and inactive decoy set (left top, red curve). The vertical dashed line represents a hypothetical FMS score cutoff dividing the total group into (X) predicted positive and (Y) predicted negative sets which can be partitioned in to four quadrants (I-IV) defined respectively as true positives (TP, I), false positives (FP, II), true negatives (TN, III), and false negatives (FN, IV). Also shown is an ROC curve, which for this example plots individual points which correspond to various FMS score cut-offs in the left panel. The coordinate of each point is determined by the false positive rate and true positive rate at that FMS score cut-off.....	40
<b>Figure 2-7.</b> Eleven scoring failures derived from FMS+SGE guided docking showing overlaid poses, PDB code identifier. RMSD in Å, and FMS scores in parentheses for the best FMS+SGE scored pose (first row, magenta) and the best FMS+SGE RMSD pose (second row, cyan) relative to the crystal pose in orange.....	47
<b>Figure 2-8.</b> SGE (top) and FMS (bottom) score histograms using ensembles derived from SGE (blue), FMS (red), or FMS+SGE (green) driven sampling methods. ....	50
<b>Figure 2-9.</b> 2D histograms of FMS score and RMSD for (a) all poses (N=239,486) and (b) best scored poses (N=1,041) generated using FMS guided sampling of 1043 systems. Poses without matches (FMS=20) not included in histograms. Color reflects density (population).....	52
<b>Figure 2-10.</b> Ten out of twenty-eight FP poses derived from FMS-guided docking with the largest RMSD values. Crystal poses in orange, best scored poses in magenta. RMSD in Å, and FMS scores in parentheses. ....	55

<b>Figure 2-11.</b> Ten out of twenty-six FN poses derived from FMS-guided docking with the largest FMS scores. Crystal poses in orange, best scored poses in magenta. RMSD in Å and FMS scores in parentheses. ....	56
<b>Figure 2-12.</b> Crossdocking outcomes averaged across the diagonal (left) or total matrix (right) for six protein families: carbonic anhydrase (CA), carboxy peptidase A (CPA), epidermal growth factor receptor (EGFR), thermolysin (THERM), HIV protease (HIVPR), and HIV reverse transcriptase (HIVRT) using SGE (top), FMS (middle), and FMS+SGE (bottom) protocols. Success in blue, scoring failure in green, sampling failure in red. ....	58
<b>Figure 2-13.</b> Crossdocking heatmaps using SGE, FMS and FMS+SGE protocols for carbonic anhydrase (29x29=841 combinations). ....	60
<b>Figure 2-14.</b> Crossdocking heatmaps using SGE, FMS and FMS+SGE protocols for thermolysin (26x26=676 combinations). ....	61
<b>Figure 2-15.</b> (a) FMS heatmap, using all crystallographic reference poses for thermolysin, with perfect overlap in dark blue (FMS=0) and poorest overlap (FMS>=8) in dark red. Group 1 sub-matrix defined by systems 1PE5, 1PE7, 1PE8, and 2TMN. Group 2 sub-matrix defined by systems 1KJO, 1KL6, 1KS7, and 1KKK. (b) Crystallographic reference overlays showing matched pharmacophore features for group 1 (left, orange), group 2 (right, magenta), and group 1 vs. group2 (middle). ....	62
<b>Figure 2-16.</b> ROC enrichment curves for 15 DUD-E systems using SGE, FMS and FMS+SGE protocol. ....	66
<b>Figure 2-17.</b> Pairwise Tanimoto heatmap for 15 DUD-E systems using FMS (top) and SGE (bottom) protocol. The color scheme in the heatmap represents the magnitude of Tanimoto similarity and the x/y axis represents the rank-ordered list (FMS or SGE) of unique active molecules for each system. ....	69
<b>Figure 2-18.</b> References (orange sticks, gray surface) used to rescore virtual screening results targeting (A) EGFR, (B) IGF-1R, (C) HIVgp41, and (D) HIVgp41 with Asp sidechain weighted 5 times. Matched pharmacophore features include: PHO in cyan; HBA (vertex and vector) in red; HBD vector in blue, hydrogen vertex in grey; ARO (vertex and vector) in orange; POS in magenta; NEG in green (see Theoretical Methods for definitions). ....	71
<b>Figure 2-19.</b> Histograms of rescoring results for the top 500 molecules selected from virtual screening targeting HIVgp41. ....	74

## Chapter 3

<b>Figure 3-1.</b> Flow chart represents the standard Rizzo lab virtual screening protocol. Colored boxes represents the approximate size of the compound set studied in each step of virtual screening. ....	80
<b>Figure 3-2.</b> HIVgp41 inner-pocket for FMS-guided virtual screening. (a) Mechanism of blocking N-helical trimer association via targeting inner pocket and blocking formation of the six helical bundle	

(6HB) via targeting the hydrophobic pocket. Figure adapted from work by Allen *et al.*<sup>120</sup> (b) Visualization of the hydrophobic pocket (gray ribbon) and pharmacophore reference for *de novo* design (cyan line representation) and virtual screening (orange stick representation). (c) Visualization of the inner pocket (gray ribbon) and the pharmacophore reference (orange stick representation). ..... 81

**Figure 3-3.** Early enrichment represented by the predicted binding poses of top molecules selected by SGE, FMS and FMS+SGE. Reference ligand shown in orange stick model; molecular volume of the reference ligand shown in gray surface model; top 50 molecules in DOCK predicted conformations shown in gray line model. PDB codes for the four systems are 3CCW, 2AA2, 1E66 and 1C8K. .... 84

**Figure 3-4.** Representative results from FMS-guided virtual screening targeting the HIVgp41 hydrophobic pocket (PDB ID: 1AIK). The Pharmacophore reference (orange) is the WWDI key residue side chains from C34. Molecules shown in the top row; Matched pharmacophore models in the bottom row. .... 86

**Figure 3-5.** Representative results from FMS-guided virtual screening targeting the HIVgp41 inner pocket. The pharmacophore reference (orange) is comprised of the IQLT key residues from one N helical peptide. Molecules shown in the top row; Matched pharmacophore models in the bottom row. .... 87

**Figure 3-6.** Histogram of (a) DCE score, (b) FMS score, (c) FMS+DCE score, (d) molecular weight, and (e) number of rotatable bonds for the top 500 scored molecules scored by DCE (blue), FMS (red), and FMS+DCE (green) from a virtual screen targeting the HIVgp41 inner pocket.<sup>14</sup> ..... 89

## Chapter 4

**Figure 4-1.** Illustration of *de novo* design protocols: (a) Horizontal pruning and guided growth in *de novo* design. Figure generated by William J. Allen.<sup>134</sup> (b) *de novo* growth of molecules by adding new segments from a user-defined fragment library instead of pre-defined fragments at all attachment points during the anchor-and-grow sampling processes. Anchor highlighted by the red box and the growing molecule at each layer shown in the grey boxes. .... 92

**Figure 4-2.** Histogram of (a) SGE, (b) FMS, (c) FMS+SGE, (d) Tanimoto and (e) Hungarian Sore for all six *de novo* experiments on all 50 systems. Tanimoto=1.0 stands for perfect Tanimoto overlap. Molecular properties and DOCK scores of ensembles generated with SGE (blue solid line), FMS (red solid line), FMS+SGE (green solid line), SGE+Tanimoto (blue dashed line), FMS+Tanimoto (red dashed line), and FMS+SGE+Tanimoto (green dashed line). .... 103

**Figure 4-3.** Histogram of (a) SGE, (b) FMS, (c) FMS+SGE, (d) Tanimoto and (e) Hungarian Sore for all six *de novo* experiments on the best scored molecule for each of the 50 systems. Molecular properties and DOCK scores of ensembles generated with SGE (blue solid line), FMS (red solid line), FMS+SGE (green solid line), SGE+Tanimoto (blue dashed line), FMS+Tanimoto (red dashed line), and FMS+SGE+Tanimoto (green dashed line). .... 105

- Figure 4-4.** Preliminary tests using *de novo* DOCK to 50 SB2012 targets with FMS-guided growth. Protein backbone is shown in tan ribbons; crystal reference molecule in cyan; *de novo* DOCK generated molecule in magenta. PDB IDs, Tanimoto and Hungarian RMSD values of the molecules (a)-(f) are provided. .... 107
- Figure 4-5.** Representative results from *de novo* growth targeting the HIVgp41 hydrophobic pocket. The pharmacophore reference is the extended peptide including WWDI key residue side chains (wire representation, C<sub>α</sub> shown in orange). .... 110
- Figure 4-6.** Histogram of (a) molecular weight, (b) number of rotatable bonds, (c) Tanimoto, (d) Hungarian Score, (e) SGE, (f) FMS and (g) FMS+SGE for *de novo* DOCK generated molecules targeting HIVgp41 hydrophobic pockets. Molecular properties and DOCK scores of ensembles generated with SGE (blue solid line), FMS (red solid line), FMS+SGE (green solid line), SGE+Tanimoto (blue dashed line), FMS+Tanimoto (red dashed line), and FMS+SGE+Tanimoto (green dashed line). .... 112
- Figure 4-7.** Representative results from *de novo* growth targeting the HIVgp41 inner pocket. Pharmacophore reference is the extended peptide including IQLT key residue side chains (wire representation, C<sub>α</sub> shown in orange). .... 114
- Figure 4-8.** Histogram of (a) molecular weight, (b) number of rotatable bonds, (c) Tanimoto, (d) Hungarian Score, (e) SGE, (f) FMS and (g) FMS+SGE for *de novo* DOCK generated molecules targeting HIVgp41 inner pockets. Molecular properties and DOCK scores of ensembles generated with SGE (blue solid line), FMS (red solid line), FMS+SGE (green solid line), SGE+Tanimoto (blue dashed line), FMS+Tanimoto (red dashed line), and FMS+SGE+Tanimoto (green dashed line). .... 115

## Chapter 5

- Figure 5-1.** (a) Preharipin stage in HIV fusion.<sup>12</sup> The three N-helices (shown in purple helices) formed a trimer inserted into target cell membrane while the three C-helices (shown in red) are yet to bind. (b) Illustration of positional alignment of T20 (orange tubes) to HIV CHR (red tubes). .... 120
- Figure 5-2.** (a) T20 interaction site with two primary mutations on the target peptide N43D (in cyan) and V38A (in red) and one secondary single-point mutation at residue 138 on CHR/T20 (letters a-e represent the residual position in the  $\alpha$ -helical secondary structure, symbol +/- highlights the charged residues); (b). Rotated (by 90°) view of the helical bundle formed by the gp41-T20 complex. (c) Corresponding wheel representation of CHR bound to NHR1, NHR2 and NHR3 (numbered 1-94 in NHRi, i=1,2,3 in gp41 sequencing).<sup>138</sup> .... 125
- Figure 5-3.** The RMSD plots of the six endpoint standard MD simulations of HIVgp41-T20 complex variants. RMSD values in Å. Time in picoseconds. Raw data in blue, running average shown in black. .... 129
- Figure 5-4.** Representative RMSD plots for TIMD simulations of mutating HIV<sub>WT</sub>-T20<sub>S138S</sub> to HIV<sub>WT</sub>-T20<sub>S138A</sub> in (a) bound state and (b) unbound state. RMSD values in Å. Time in picosecond. .... 130

<b>Figure 5-5.</b> The $dV/d\lambda$ value histogram in all $\lambda$ windows ( $\lambda = 0.05, \dots, 0.50, \dots, \text{and } 0.95$ , distinguished by color) simulations for mutating HIV <sub>WT</sub> -T20 <sub>S138S</sub> to HIV <sub>WT</sub> -T20 <sub>S138A</sub> in (a) bound state and (b) unbound state.....	132
<b>Figure 5-6.</b> The $dV/d\lambda$ plots with respect to time for all $\lambda$ windows ( $\lambda = 0.05, \dots, 0.50, \dots, \text{and } 0.95$ , distinguished by color with identical color scheme as shown in Figure 5-5) simulations for mutating HIV <sub>WT</sub> -T20 <sub>S138S</sub> to HIV <sub>WT</sub> -T20 <sub>S138A</sub> in (a) bound state and (b) unbound state. ....	133
<b>Figure 5-7.</b> The $dV/d\lambda$ curves with respect to $\lambda$ for run 1 (red), run1b (green), run2 (cyan) and run3 (magenta). Each point of the $dV/d\lambda$ plot is derived from ensemble average of a 2ns long TI run...	134
<b>Figure 5-8.</b> Correlation of calculated relative binding energy ( $\Delta\Delta G_{\text{calcd.}}$ , y axis) compared to experimental data ( $\Delta\Delta G_{\text{exptl.}}$ , x axis). Relative binding energy using TI method is calculated from mutating residue 138 on T20 when bound to the same receptor, averaging over the $2 \times 19$ 2ns production run for each complex. ....	137
<b>Figure 5-9.</b> (a) Key residues for VDW contact (heatmap for NHR1 WTS138S ( $>0.5\text{kcal/mol}$ )); (b) Key residues for VDW contact (heatmap for NHR3 WTS138S ( $>0.5\text{kcal/mol}$ )) .....	141
<b>Figure 5-10.</b> (a) Original and (b) modified model of T20-3NHR based on VDW footprint and heatmap analyses. ....	142
<b>Figure 5-11.</b> Electrostatic (ES) interaction heatmap for T20 binding to 3NHR.....	144
<b>Figure 5-12.</b> (a) VDW and (b) ES molecular footprint for V38A primary, and S138A secondary mutations. Error bar for each data point plotted as the dashed curves. HIV <sub>WT</sub> -T20 <sub>S138S</sub> in blue, HIV <sub>V38A</sub> -T20 <sub>S138S</sub> in red, HIV <sub>V38A</sub> -T20 <sub>S138A</sub> in green .....	146
<b>Figure 5-13.</b> (a) VDW and (b) ES molecular footprint for N43D primary, and S138A secondary mutations. Error bar for each data point plotted as the dashed curves. HIV <sub>WT</sub> -T20 <sub>S138S</sub> in blue, HIV <sub>N43D</sub> -T20 <sub>S138S</sub> in red, HIV <sub>N43D</sub> -T20 <sub>S138A</sub> in green. ....	148

## Appendix C

<b>Figure C-1.</b> Membrane-bounded HIVgp41-T20 complex prepared with CHARMM-GUI and AMBER14. ....	180
--	-----

## List of Tables

### Chapter 1

<b>Table 1-1.</b> List of HIV inhibitors targeting various steps in viral life cycle. <sup>#</sup> .....	2
<b>Table 1-2.</b> Experimental activities of HIVgp41-T20 complex analogs. ....	6
<b>Table 1-3.</b> Commonly used scoring functions in DOCK.....	21

### Chapter 2

<b>Table 2-1.</b> Pharmacophore type definitions in DOCK.....	28
<b>Table 2-2.</b> Examples of pharmacophore features derived from small molecules.....	32
<b>Table 2-3.</b> Systems used for enrichment tests. ....	44
<b>Table 2-4.</b> Pose reproduction results employing SGE, FMS, and FMS+SGE scoring functions.....	45
<b>Table 2-5.</b> Fold enrichment (FE) results at different percentages of the database (DB) screened. ....	67

### Chapter 4

<b>Table 4-1.</b> List of 50 systems for initial <i>de novo</i> validation from SB2012. <sup>100</sup> .....	96
<b>Table 4-2.</b> Reproduction rate of <i>de novo</i> design by Tanimoto cutoff in 50 systems tested in SB2012. ..	101
<b>Table 4-3.</b> Reproduction rate of <i>de novo</i> design by Hungarian RMSD cutoff in 50 systems tested in SB2012.....	101
<b>Table 4-4.</b> <i>De novo</i> design results for HIVgp41 hydrophobic and inner pocket. ....	109
<b>Table 4-5.</b> List of generic fragment libraries.....	116

### Chapter 5

<b>Table 5-1.</b> Binding energy of T20 targeting HIVgp41 calculated from experimental EC <sub>50</sub> values obtained from Izumi <i>et al.</i> <sup>18</sup> .....	122
<b>Table 5-2.</b> Calculated relative binding energy and transformation energy for null transformations .....	135
<b>Table 5-3.</b> Experimental vs. Calculated $\Delta\Delta G_{\text{bind}}$ for T20 analogs with HIVgp41 .....	138

## **Appendix B**

<b>Table B-1.</b> Pharmacophore feature represented in ph4.mol2 and ph4.bild. ....	178
--	-----



## List of Abbreviations

ACD, Available Chemical Directory

ACF, autocorrelation function

BASEM, block averaged standard errors of the mean

CHR, C-HR, C-terminal heptad repeat

DCE, DOCK Cartesian energy

DMS, distributed molecular surface

DUD-E, database of useful (docking) decoys - enhanced

ES, electrostatic energy

FEP, free energy perturbation

FLX, flexible ligand docking

FMS, pharmacophore matching similarity

FPS, footprint similarity

GB, Generalized Born

HIVgp41, human immunodeficiency virus glycoprotein 41

HTS, high-throughput screening

MD, Molecular Dynamics

MM-GBSA, Molecular Mechanics Generalized Born Surface Area

MOE, Molecular Operating Environment

MPER, membrane-proximal external region

MW, molecular weight

NHR, N-HR, N-terminal heptad repeat

NMR, nuclear magnetic resonance

PB, Poisson-Boltzmann

PDB, Protein Data Bank

QSAR, quantitative structure-activity relationship

RGD, rigid ligand docking

RMSD, root-mean-square deviation

ROC, receiver-operating characteristics

SB2010, SB2012, Stony Brook Pose Reproduction Database (2010, 2012)

SBDD, structure-based drug design

SGE, single grid energy

TI, thermodynamic integration

TIMD, thermodynamic integration molecular dynamics

VDW, van der Waals

VS, virtual screening

ZINC, Zinc Is Not Commercial (a small molecule database)

## Acknowledgments

I would like to thank my advisor Robert C. Rizzo for his support and guidance during my time in his research lab. I would like to also thank Dr. David F. Green, Dr. Yuefan Deng and Dr. Jonathan G. Rudick for agreeing to serve on my dissertation committee. I truly appreciate the feedback and assistance they have provided throughout my graduate study.

I am grateful to all past and present members of the Rizzo lab for the helpful discussions, collaborative works and their supports both inside and outside of work. I have enjoyed my experience in the Rizzo group thanks to all of them. I would especially like to thank Trent E. Balius and Sudipto Mukherjee for their work on DOCK code development and offering a lot of help with course works and research project developments in my early years of graduate school. Dr. William J. Allen has been working most closely with me towards my final years in the PhD program and has provided insightful advice and guidance with my DOCK development project, manuscript writing as well as my job search. In addition, I had the pleasure to interact with former and present members in the Green lab and Simmerling lab at Stony Brook University.

I express gratitude to all staff members in the Department of Applied Mathematics and Statistics at Stony Brook University and Computational Science Center at Brookhaven National Laboratories for their accommodations. And I would like to thank the funding agencies involved in sponsoring my graduate study and conferences/meetings traveling including NIH grants R01GM083669 (to RCR), the department of Applied Mathematics and Statistics, Graduate Student Organization at Stony Brook University, Society for Industrial and Applied Mathematic Computational Sciences and Engineering, CRA-W Grad Cohort and the Chinese Scholarship Council.

For my stay in Stony Brook during my graduate work, further acknowledgement and thanks is due to my friends. Linda Costanzo has given me her support and treated me like family. I have enjoyed spending time with members in the Suffolk Toastmasters Club and fellow students in the Island Budokan Dojo. A note of thanks is also due to my ex-roommates Song Feng and Yifan Peng for their encouragements and friendship.

Last and foremost, I would like to thank my parents Xiaomei Hu and Zhengjie Jiang who have always believed in me, given my love and supports throughout my academic career and life.

## **Chapter 1. Introduction**

This Chapter introduces a protein-peptide complex system of interest, HIVgp41-T20, which can provide valuable insights in blocking the membrane fusion of an important drug target HIV. The primary computational techniques used in this study include atomic-level molecular dynamics simulations and molecular docking.

### **1.1 Therapeutic Drug Target for HIV-1 Fusion: HIVgp41**

Human immunodeficiency virus (HIV), which causes the life-threatening disease called acquired immune deficiency syndrome (AIDS), has resulted in nearly 30 million deaths since the first HIV infection was diagnosed in 1981.<sup>1</sup> Great efforts have been made to prevent further spread of viral infection and thereby control viral load in patients with the condition. However, drug resistance arising from clinical treatment and the existing side effects of many current therapeutic strategies all call for continued development of next generation more potent drugs to fight the global epidemic.<sup>2-5</sup> Work presented in this dissertation applies atomic level computational modeling to help reach this important goal.

Like most viruses in the Retroviridae family, HIV has to invade into a host cell to complete viral replication. Targeting different steps of HIV host invasion, the current anti-HIV inhibitors can fall into five major categories: fusion and entry inhibitors, nucleotide reverse transcriptase inhibitors, non-nucleotide reverse transcriptase inhibitors, protease inhibitors and integrase inhibitors which target several stages in the viral life cycle at the same time. In addition, “cocktail” therapy employing combinations of antiretroviral from these categories are

designed to block HIV replications. Table 1-1 lists HIV medicines approved by the U.S Food and Drug Administration (FDA).<sup>6-8</sup> In this study, we are targeting the process of HIV fusion.

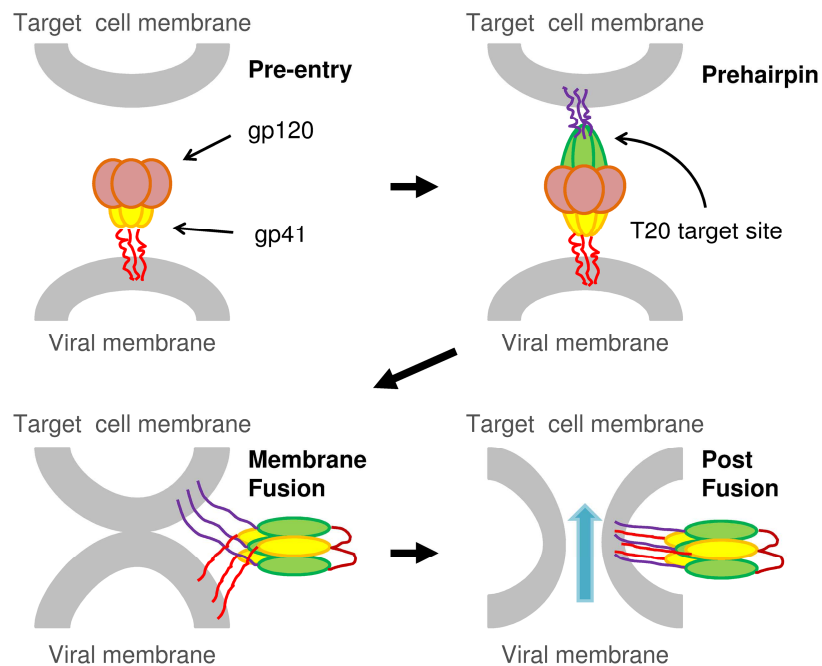
**Table 1-1.** List of HIV inhibitors targeting various steps in viral life cycle.<sup>#</sup>

<b>Category/Target</b>	<b>Drug (Generic Name)</b>
Fusion Inhibitor	enfuvirtide (T20)
Entry Inhibitor	maraviroc
Nucleoside Reverse Transcriptase Inhibitor	abacavir, didanosine, emtricitabine, lamivudine, stavudine, tenofovir disoproxil, zidovudine
Non-Nucleoside Reverse Transcriptase Inhibitor	delavirdine, efavirenz, etravirine, nevirapine, rilpivirine
Protease Inhibitor	atazanavir, darunavir, fosamprenavir, indinavir, nelfinavir, ritonavir, saquinavir, tipranavir
Integrase Inhibitor	dolutegravir, elvitegravir, raltegravir
Pharmacokinetic Enhancer	cobicistat
Combination Medicines	abacavir and lamivudine; abacavir, dolutegravir, and lamivudine; abacavir, lamivudine, and zidovudine; efavirenz, emtricitabine, and tenofovir disoproxil fumarate; elvitegravir, cobicistat, emtricitabine, and tenofovir disoproxil fumarate; emtricitabine, rilpivirine, and tenofovir disoproxil fumarate; emtricitabine and tenofovir disoproxil fumarate; lamivudine and zidovudine; lopinavir and ritonavir

<sup>#</sup>Data from AIDSinfo FDA-Approved HIV Medicines Fact Sheet, accessed May 15<sup>th</sup>, 2015

Prior work has resulted in a model of HIV fusion and entry that can be arranged as four distinct steps as shown in Figure 1-1.<sup>9-11</sup> Firstly, when HIV approaches a target cell, its envelope glycoprotein gp120 will recognize CD4 receptors together with chemokine co-receptors such as CCR5 and/or CXCR4, bringing the virus near the host cell (Figure 1-1, top left panel). Then gp120 will go through conformation changes that allows the N-terminal heptad repeat region (N-HR) of gp41 to expose itself and insert into the host cell while the gp41 C-terminal heptad repeat region (C-HR) is still attached to the viral membrane (the prehairpin stage, Figure 1-1, top right panel). In the next step, gp41 will undergo a series of conformational changes that lead to the binding of three C-HR helices to N-HR helices, forming a six-helical coiled-coil hairpin termed

the six helix bundle (Figure 1-1, bottom left panel). This process will bring the viral and native cell membranes in close proximity to each other and the two membranes will eventually merge and complete viral fusion (post fusion, Figure 1-1, bottom right panel). Importantly, in the prehairpin step, both N-HR and C-HR regions of gp41 are exposed to solvent and can interact with a variety of substrates and several inhibitors (both peptides and small molecules) has been shown to inhibit viral entry and reduce HIV infections.<sup>11,12</sup> The first FDA approved fusion inhibitor called enfuvirtide (T20) discussed more in Chapter 5 was originally designed by Jiang *et al*<sup>13</sup> to bind to the N-HR in the prehairpin stage in competition with the native viral C-HR. A new inner pocket in the interface within the N-HR helical coil has also been recently identified by Allen *et al*<sup>14</sup> in the Rizzo lab, which is also described in Chapter 5.



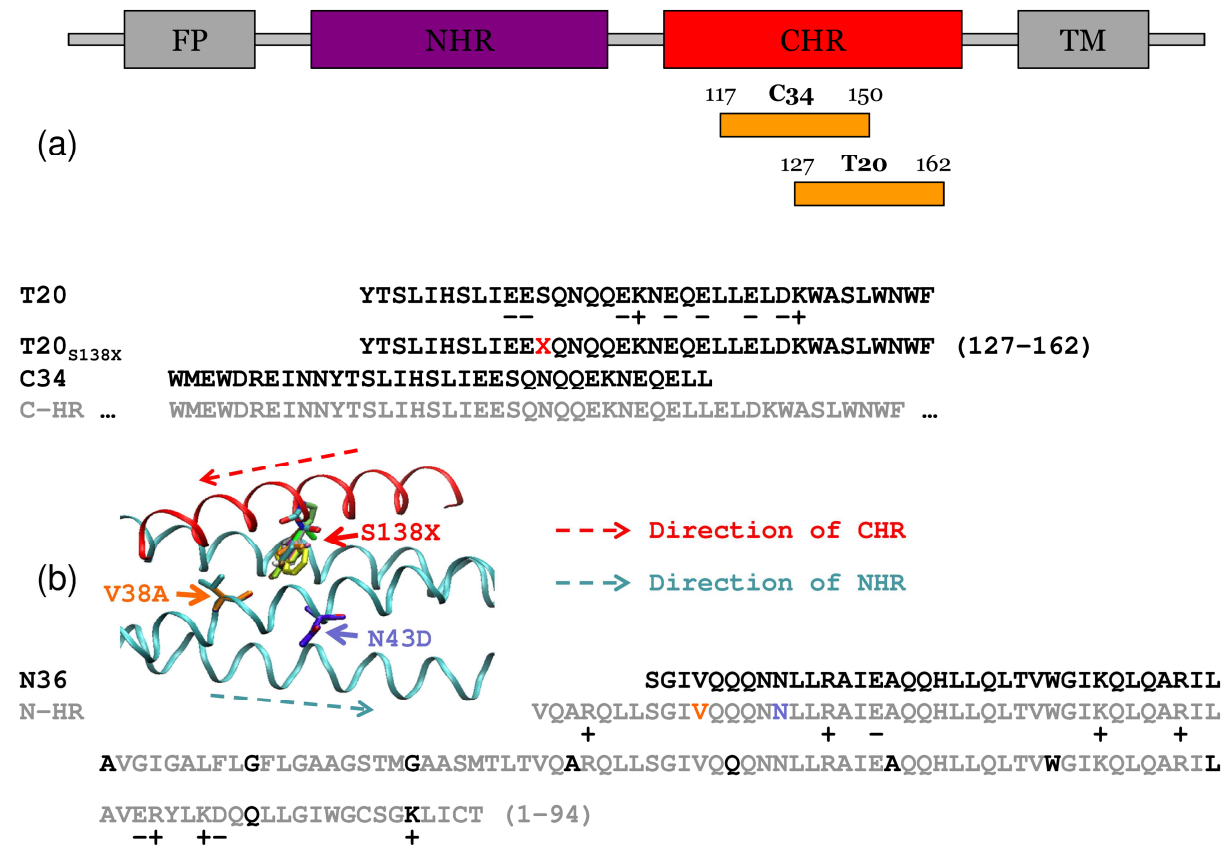
**Figure 1-1.** HIV fusion can be categorized as four stages: (1) Native pre-entry stage when no interactions between the virus and the host cell are observed, (2) pre-hairpin stage, (3) membrane fusion, and (4) post fusion. Figure adapted from Hughson *et al.*<sup>11</sup>

While clinical use of the peptide drug T20 is reported to successfully reduce viral loads in infected patients, the therapy also leads to resistant mutations in both the virus NHR and CHR sections.<sup>15</sup> As shown in Figure 1-2a, T20 is directly derived from the native C-HR of gp41. Thus, any primary resistant mutations that arise and affect T20 will also reduce the affinity between the native C-HR and the mutated N-HR helices. Interestingly, in order to restore fusion activity related to the coiled-coil hairpin formation, secondary mutations occurring in the C-HR of the “virus” have also been clinically observed. From a drug design standpoint, a greater understanding of the energetic effects of the primary and secondary mutation pair would be important. Figure 1-2a shows a schematic representation of T20 and C34 (another peptide inhibitor binding to a highly-reserved hydrophobic pocket in HIVgp41, used later to perform reference-based small molecule inhibitor design in Chapter 2 , 3 and 4) in alignment to the N-HR, C-HR, transmembrane domain (TM), and fusion peptide (FP) of HIVgp41.

It is important to note that the present HIVgp41-T20 studies employ a computational model of the complex as the crystallographic binding pose of T20 is difficult to obtain as the binding site of T20 is close to the membrane-associated regions of N-HR. And solving the conformation of the membrane-embedded regions in a protein and model their interaction with the lipid bilayers are known to be quite challenging for crystallographers. Based on sequence alignment to known crystal structures (PDB code 1IF3 and 1ENV) of gp41 N-HR bound to other peptide inhibitors, a computational model of T20-bound gp41 N-HR complex structure built and reported by McGillick *et al*<sup>16</sup> with the FP region modeled as  $\alpha$ -helices.<sup>17</sup> The amino acid sequence of T20 and three N-HR helices in the model are shown in Figure 1-2b. In the work presented in Chapter 5, a series of HIVgp41-T20 complex analogs have been constructed using the McGillick model by computationally mutating residue 138 (gp41 sequencing) on T20 and



amino acid groups corresponding to residue 38 and 43 (gp41 sequencing) on all three N-HR helices, as visualized in Figure 1-2b. These modified complex analogs structures are used for molecular dynamics calculations.



**Figure 1-2.** Modeling the binding of T20 to gp41 NHR. (a) Schematic representation of positional relationship of T20, C34, FP, N-HR, C-HR and TM. (b) Linear sequences of T20, C34 and N-HR helices and visualization of T20 primary (red arrow) and secondary (orange and blue arrow) mutation sites. Charged residues are indicated by “+” and “-“ signs.

In terms of available activities data for the HIVgp41-T20 systems, Izumi *et al*<sup>18</sup> has experimentally evaluated the activity (EC<sub>50</sub>) of a related series of 19 peptides based T20 mutants in which serine at residue 138 is replaced by all other natural amino acids except cysteine (Table 1-2), which represents the “secondary mutations”. The effects of two primary mutations V38A

and N43D relative to the wild-type receptor were also quantified. The changes in binding affinity due to structural variability can provide insight in the structure-activity relationship in the HIVgp41-T20 system to model HIV fusion. In this case, relative binding affinities rather than absolute binding affinities are of primary interest. In Chapter 5, computational binding energies for a subset of the complexes listed in Table 1-2 have been evaluated computationally for comparison to experimental activities and structural and energetic analyses to characterize the origins of T20 affinity and effects of mutations.

**Table 1-2.** Experimental activities of HIVgp41-T20 complex analogs.

Secondary Mutations	Primary Mutations		
	HIV <sub>WT</sub>	HIV <sub>V38A</sub>	HIV <sub>N43D</sub>
	EC50 (nM)	EC50 (nM)	EC50 (nM)
T20 <sub>S138S</sub>	2.4 ± 0.6	23 ± 8.2	49 ± 10
T20 <sub>S138A</sub>	0.6 ± 0.1	3.6 ± 1.7	3.5 ± 0.9
T20 <sub>S138D</sub>	210 ± 94	>1000	>1000
T20 <sub>S138E</sub>	283 ± 80	>1000	>1000
T20 <sub>S138F</sub>	9.4 ± 2.6	203 ± 89	393 ± 119
T20 <sub>S138G</sub>	1.3 ± 0.5	65 ± 8.8	141 ± 26
T20 <sub>S138H</sub>	210 ± 85	>1000	>1000
T20 <sub>S138I</sub>	0.5 ± 0.1	4.9 ± 2	2.9 ± 0.8
T20 <sub>S138K</sub>	708 ± 145	>1000	>1000
T20 <sub>S138L</sub>	0.7 ± 0.1	13 ± 6	2.9 ± 0.7
T20 <sub>S138M</sub>	0.7 ± 0.2	4.4 ± 0.1	1.7 ± 0.5
T20 <sub>S138N</sub>	19 ± 4	>1000	>1000
T20 <sub>S138P</sub>	446 ± 167	>1000	>1000
T20 <sub>S138Q</sub>	34 ± 11	>1000	>1000
T20 <sub>S138R</sub>	362 ± 114	>1000	>1000
T20 <sub>S138T</sub>	0.9 ± 0.2	39 ± 8.5	161 ± 35
T20 <sub>S138V</sub>	0.4 ± 0.2	31 ± 14	22 ± 3.5
T20 <sub>S138W</sub>	29 ± 14	>1000	>1000
T20 <sub>S138Y</sub>	25 ± 9	516 ± 223	>1000

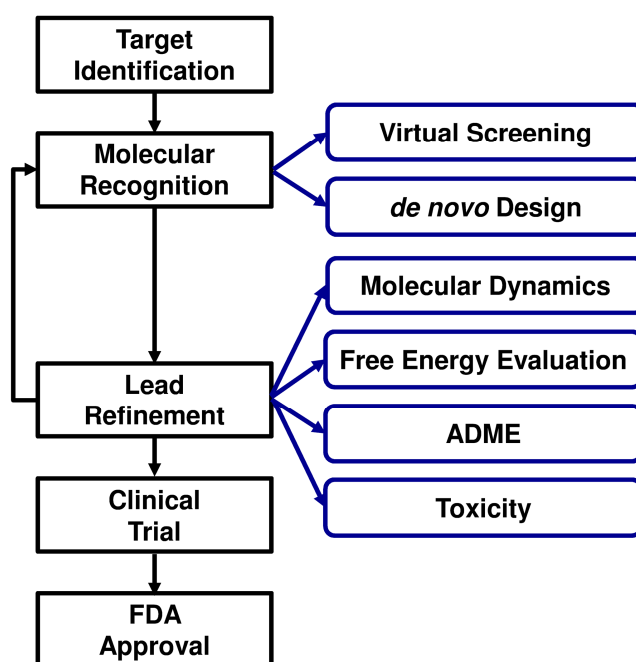
The ultimate goal of this dissertation is to design new generations of small molecule fusion inhibitors targeting HIVgp41 with development and applications of new computational structure-based drug design methods including molecular dynamics simulations, molecular docking (including *de novo* design) and pharmacophore-based similarity matching to known peptide inhibitors. The rest of this Chapter outlines the computational tools employed in the research with emphasis on molecular dynamics and molecular docking.

## 1.2 Computational Structure-Based Drug Design

Continuous advancements in modern drug design techniques are critical to combat drug resistance arising from already established therapeutic strategies as well as new emerging diseases. Figure 1-3 outlines a general framework for employing a computational drug design pipeline that includes target identification, molecular recognition, lead refinement, clinical trial and FDA approval. Importantly, computational molecular modeling methods can be used to determine the interactions between a ligand (often a small molecule) and the binding pocket in an identified target (often a protein).<sup>19-21</sup> For example, molecular docking with applications to virtual screening is a commonly used method to select initial hits, as an alternative to traditional experimental high throughput screening (HTS), to help speed up the initial steps in the drug design process which can reduce total cost. Alternatively, *de novo* design employing target-ligand interaction profiles can generate drug leads “from scratch”.

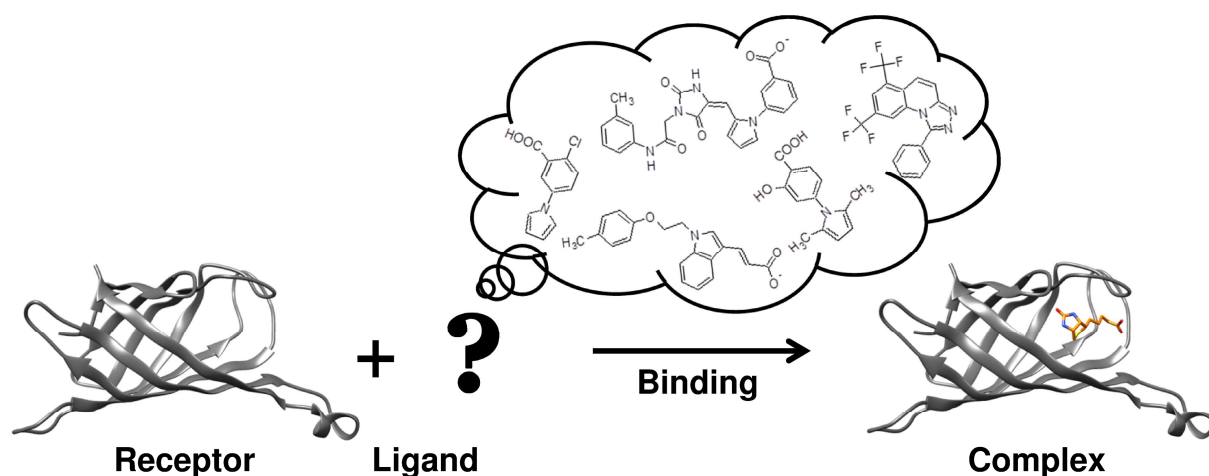
Leads selected from virtual screening or constructed from *de novo* design will ultimately be investigated and future refined to optimized properties. For example, atomic-level molecular dynamics simulations and free energy calculations can be performed to estimate the strength of noncovalent binding and assess the binding stability. Computational methods can also be used in

lead refinement by predicting the pharmacokinetic and pharmacodynamic properties of the hits including ADME profiles and toxicity.<sup>19,22,23</sup> In general, only a few molecules are selected for extensive lead refinement and even fewer molecules can pass through the several stages of clinical trials and eventually FDA approval. Reducing the cost is especially important as the entire drug design process can take up to 15 years and cost up to 1 billion dollars.<sup>24</sup> Successful examples in which computational methods were employed in the design of new drugs against various target systems include chronic myelogenous leukemia (Imatinib),<sup>25</sup> liver cancer (Thymitaq),<sup>26</sup> influenza (oseltamivir and zanamivir),<sup>27-29</sup> and HIV protease (Viracept and Aluviran).<sup>30-33</sup>



**Figure 1-3.** Flow chart of computational structure-based drug design. Computational techniques used are highlighted in blue boxes.

As illustrated in Figure 1-4, the goal of the molecular recognition step in drug design is to predict ligands with favorable interactions to a given target. Prior knowledge of both the receptor and available active ligands can be used to energetically and structurally predict the affinity of an arbitrary candidate ligand. For receptor-based molecular recognition, common methods like molecular mechanics-based energy calculations (discussed in more details later) allow physics-based predictions that are comparable to experimental results. The validated computational protocols can then provide more detailed insights into the binding profiles of the ligands, techniques such as energy component analyses based on electrostatic and van der Waals interactions or water-mediated hydrogen bonding.<sup>34</sup> For ligand-based molecular recognition, the similarities between known active molecules and candidate molecules can be quantified to guide reference-based drug design. It is hypothesized that drug activity is a function of molecular structure.<sup>35,36</sup> Thus similar molecules can potentially yield similar activities in binding. Pharmacophore modeling can potentially be used for ligand-based recognition. In this study, we will further discuss the use of a pharmacophore-based similarity metric in molecular docking in Chapter 2, Chapter 3 and Chapter 4.



**Figure 1-4.** Molecular recognition: identification of ligand with favorable binding affinity to the target receptor.

As shown in Figure 1-3, virtual screening and *de novo* design are two important tools in molecular recognition. Virtual screening is a process of rapidly testing large database of commercially available small molecules *in silico* for biological activities. Typically, virtual screenings are done using molecular docking programs with an efficient scoring function to sample ligand binding poses and rank-order the different poses of various molecules in the compound database.<sup>37</sup> The UCSF ZINC<sup>38,39</sup> database is one commonly used freely available compound resources for virtual screening. It provides about 22.7 million purchasable compounds as of November 2014 (<http://zinc.docking.org/browse/subsets>, accessed March 2015). In addition, Accelrys ACD<sup>40</sup> is a commercially available resource for over 7 million unique chemicals with 3D structure information. By rank-ordering sampled poses of molecules from these large compound databases, top-scored molecules can be selected for further inspection and filtered. Typically a selected subset of promising leads will be purchased and experimentally tested for initial activity. An alternative way of searching the chemical space for drug leads is to design drug-like molecule “from scratch” via *de novo* design, a technique that

uses fragment libraries and construction algorithms. Both “outside-and-in” and “inside-and-out” strategies can be used to either (1) probe the binding site to allocate favorable spots for different fragment binding and then link all fragments together to design a complete compound; or (2) seed the *de novo* growth with one component fragment and grow it by adding new fragments to fit the binding site energetically and geometrically.<sup>41</sup> The resulting compounds can then be synthesized and experimentally tested for biological activity. This study included drug design application employing both virtual screen and *de novo* design in Chapter 3.

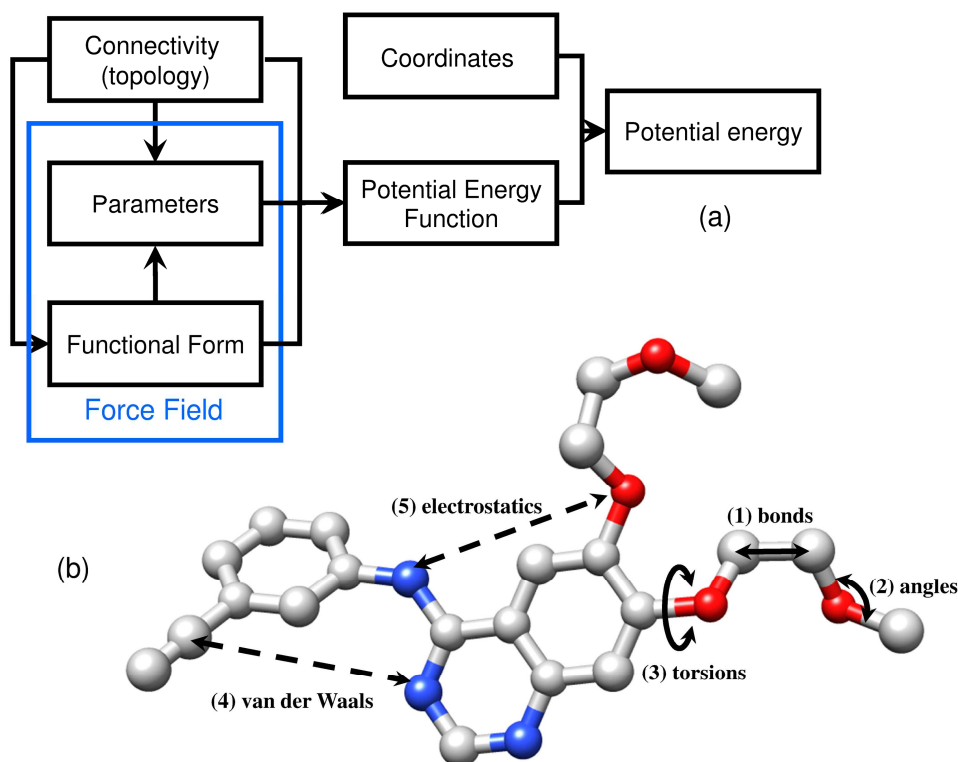
### 1.3 Classical Molecular Mechanics

All-atom molecular mechanics (MM) method is employed in this study to model molecular systems. A molecule is modeled as a set of atoms with specific topology defined by bonds. Each atom is also considered a particle with assigned radius, point charge and atom type. A set of parameters are defined in a certain force field for different atom types to evaluate the potential energy of a molecule as shown in Figure 1-5a.<sup>42</sup> The potential energy  $E$  is function of the atomic coordinates, and consists of both the bonded and non-bonded terms as illustrated in Figure 1-5b. The bonded terms compute the sum of bond length (Figure 1-5b (1)), bond angle (Figure 1-5b (2)) and torsion angle terms ((Figure 1-5b (3))). The nonbonded terms describe longer-range interactions between atoms that not directly connected by a specific bond and compute both the electrostatics (Figure 1-5b (4)) and van der Waals (Figure 1-5b (5)) interactions. The functional form of the total potential energy used with the Assisted Model Building and Energy Refinement (AMBER) force field described in Chapter 5 is as follows.<sup>43,44</sup>

$$\begin{aligned}
E_{total} = & \sum_{i \in bonds} k_r (r_i - r_0)^2 + \sum_{j \in angles} k_\theta (\theta_j - \theta_0)^2 + \sum_{k \in torsions} k_\phi [1 + \cos(n\phi - \gamma)] \\
& + \sum_{i < j} \left[ \frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} + \frac{q_i q_j}{\epsilon r_{ij}} \right] \tag{1-1}
\end{aligned}$$

Here, the bond length and bond angle terms are described by Harmonic functions where energetic penalties are assigned to values  $(r_i, \theta_j)$  that deviate from the corresponding equilibrium values  $(r_0, \theta_0)$ . And a sinusoidal function is used to describe the torsion term where  $n$  is the multiplicity parameter and  $\gamma$  is the phase factor. For the nonbonded terms,  $r_{ij}$  are the distance between atom  $i$  and  $j$ ;  $A_{ij}$  and  $B_{ij}$  are the van der Waals parameters defined by the well depth and radii of the two atoms;  $q_i$  is the charge of atom  $i$ . In AMBER, 1-4 interactions are usually used for both VDW and electrostatic terms to better match experimental measurements. For atoms in the same molecule, only those at least three bonds away are included in the through-space interactions calculation.



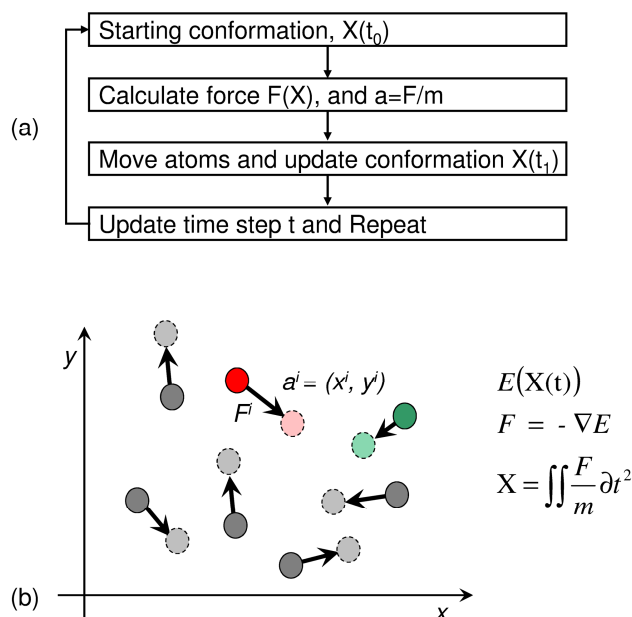


**Figure 1-5.** Molecular mechanics energy function terms: (1) bond length, (2) bond angle, (3) torsion angle, (4) VDW and (5) electrostatics.

To calculate the potential energy of a biological system with the potential energy function, the coordinates of the molecules are needed. In general, accurate initial starting structures are essential for molecular modeling studies, especially if the goal is to model protein-ligand interactions. Experimental methods including X-ray crystallography, nuclear magnetic resonance (NMR) and electron microscopy (EM) are commonly used to determine initial coordinates of the systems (usually proteins in bound or apo states). The Protein Data Bank (PDB),<sup>45</sup> which includes a total number of 95,375 X-ray crystal structures, 10862 NMR structures and 753 EM structures as of March 2015, is an expanding important resource for accessing experimental structures of proteins and other biomolecules.

## 1.4 Molecular Dynamics and Free Energy Calculations

**Molecular Dynamics.** Molecular dynamics (MD) is a method to simulate the motions of a biological system.<sup>46</sup> Derived from the potential energy of interaction within the molecular system, we can calculate the forces of particles in the molecules and mimic the dynamics of the system. As shown in Figure 1-6, the classical Newtonian equations of motion are employed for atomic-level MD simulations. Starting from an initial conformation ( $X(t_0)$ ), initial velocities are assigned for all atoms of the molecule at time step  $t_0$ . And the forces on the atoms  $F$  are calculated by taking the first derivatives of the potential energy  $E$  with respect to atomic coordinates  $X$ . Next, the acceleration  $a$  of each atom is computed using Newton's Law as  $a=F/m$  where  $m$  is the atomic mass. With a pre-determined time step  $\Delta t$ , the atomic coordinates for the next time step  $t_1 = t_0 + \Delta t$  is updated via  $X(t) = \iint \frac{F}{m} \partial t^2$ . The velocities of all atoms are also updated at each time step with intervals of dynamical relaxation to avoid "hotspot" in the system. Iterations of these calculations will update the three dimensional coordinates of a molecular system as a function of time. The performance and motions of a molecular dynamics (MD) simulation depends on the force field used in the molecular mechanics model. The physics-based MD simulations can potentially mimic the motions in the protein-ligand systems and provide energetic and structural insights of the system of interest.



**Figure 1-6.** Molecular dynamics: updating 3D coordinates of atoms in the molecular system

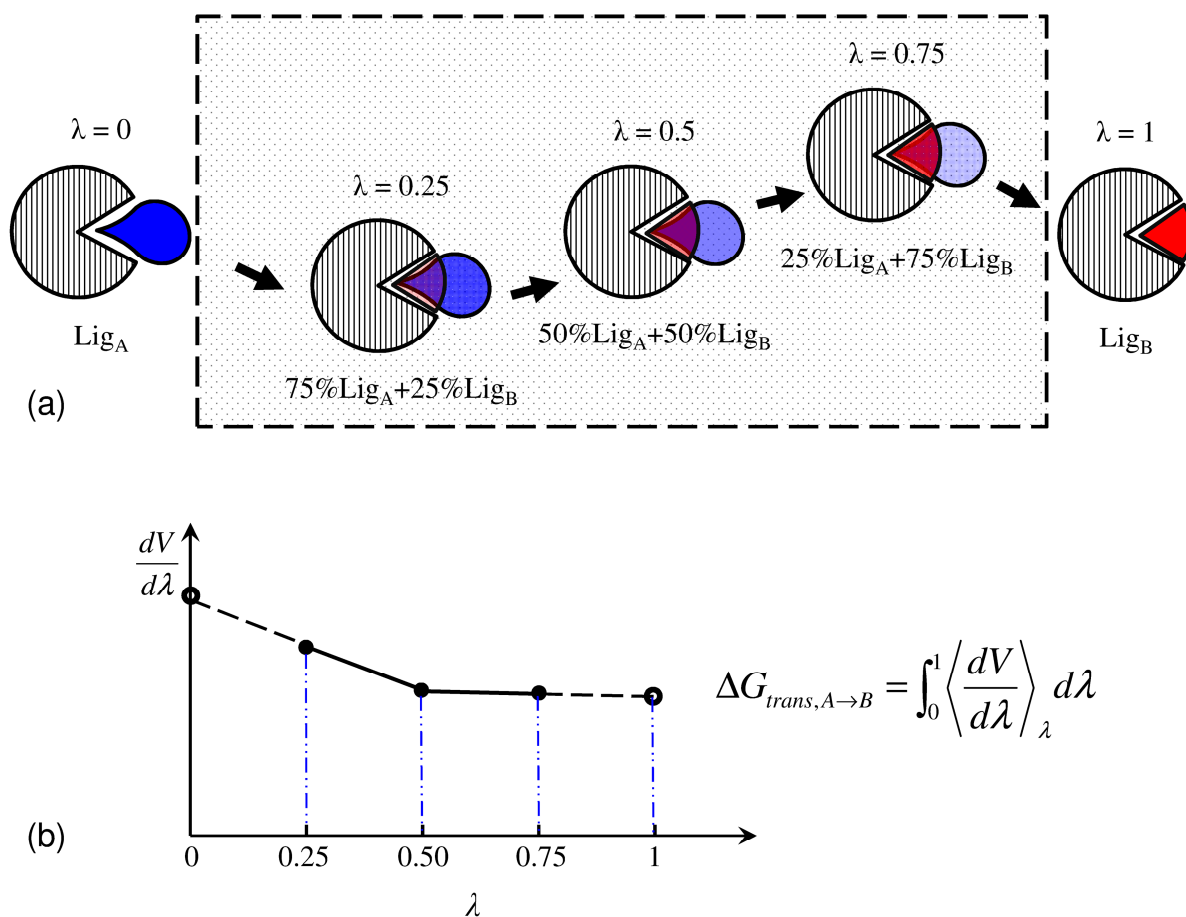
There are many software packages for performing molecular dynamics simulation such as AMBER,<sup>47-49</sup> CHARMM,<sup>50</sup> GROMOS,<sup>51</sup> GROMACS,<sup>52,53</sup> and NAMD.<sup>54</sup> Parallel computing tools including Message Passing Interface (MPI) and GPU CUDA are usually employed to run the simulations in parallel and yield higher simulation speed. In this study, molecular dynamics simulations are performed in AMBER with both MPI and CUDA.<sup>55,56</sup>

**Free energy calculations.** In order to estimate the relative binding free energies associated with conformational or compositional changes of a molecular system, such as the mutational effects in HIVgp41-T20 system introduced above, a variety of computational procedures based on a series of conformations sampled with methods such as MD simulations has been devised.<sup>57</sup> While in theory independent sampling of the two endstate-configurations associated with the mutation of interest can be performed to yield relative binding affinity. In

practice, MD simulations on the nanosecond timescale with the current computational power might not be sufficient for accurate evaluations. One key method, termed thermodynamics integration (TI),<sup>58</sup> is a procedure to simulate nonphysical transformations between endstate configurations and evaluate energy differences of two similar molecular systems such as shown in Figure 1-7. By gradually transforming with small intermediate steps, such ensemble sampling techniques will converge more quickly for reasonable energetic measurements.<sup>57</sup> Here, ligand  $\text{lig}_A$  is alchemically transformed to  $\text{lig}_B$  in a series of coupled simulations with a transformation parameter  $\lambda$  varying from 0 to 1. For each intermediate  $\lambda$  window (in the dashed box in Figure 1-7a), the dynamics of the coupled system is determined by potential energy and force calculated as weighted average of the two endstate systems (physical systems corresponding to  $\text{lig}_A$  and  $\text{lig}_B$ ). The closer  $\lambda$  is to 0, the more similar the mixed system is to the endstate system defined by  $\lambda=0$ , and vice versus. As virtually illustrated in Figure 1-7a, through the transparency of the ligand: the ligand in blue denotes  $\text{lig}_A$  when  $\lambda=0$  and the ligand in red denotes  $\text{lig}_B$  when  $\lambda=1$ . Typically, linear mixing energy functions can be used. In AMBER11,<sup>47</sup> a soft-core potential function is employed to address singularity problems for TI simulations close to the endstate when  $\lambda=0$  and  $\lambda=1$ . Detailed discussion will be provided in Chapter 5.

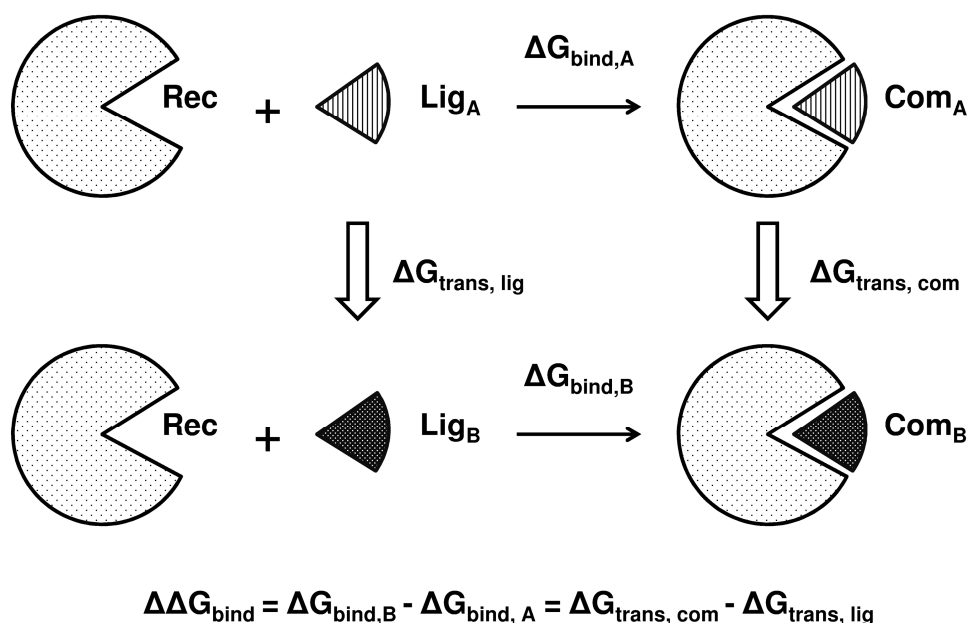
The value of the derivative of potential energy differences  $V$  with respect to  $\lambda$  ( $dV/d\lambda$ ) is evaluated for each snapshot in the TI simulation at each. The ensemble averaged  $dV/d\lambda$  value for each  $\lambda$  window between 0 and 1 is plot as a function of  $\lambda$  as shown in Figure 1-7b by connecting disjoint  $(\lambda, dV/d\lambda)$  points to approximate the continuous  $dV/d\lambda$  function. The value of  $dV/d\lambda$  curve at  $\lambda=0$  and  $\lambda=1$  are usually not directed calculated using the TI method but instead estimated by extending the curve based on function values near the edges (dashed section of the curve in Figure 1-7b). The signed area under the  $dV/d\lambda$  curve between  $\lambda=0$  and  $\lambda=1$  is

reported as the total transformation energy. In principle, alchemical transformation using TI, if done in a well-controlled and converged way with a reasonable partition for the intermediate states, can yield a relatively accurate energy. The coupled simulations can gradually sample in the energy landscape between the two physical states and also implicitly include the entropy term. In general, the two physical states being simulated are made to be similar to each other to obtain well-behaved molecular dynamics simulations and improve the accuracy and convergence of the energy measurements.



**Figure 1-7.** Evaluating transformation energy using thermodynamic integration method. (a) Coupled systems in different  $\lambda$  windows. (b) Theoretic  $dV/d\lambda$  curve for transformation energy evaluation.

In this study, we are particularly interested in computing the relative binding energy between two ligands to the same receptor. Thermodynamic integration molecular dynamics (TIMD) simulations are performed for calculating transformation energies from one ligand to another in both the bound and the unbound states. The relative binding energy can then be derived from the transformation energies as shown in Figure 1-8. The relative binding energy between ligand A and ligand B  $\Delta G_{\text{bind,B}} - \Delta G_{\text{bind,A}}$  is equivalent to the difference in transformation energy of ligand A to ligand B in the binding site and in the solvent alone  $\Delta G_{\text{trans,com}} - \Delta G_{\text{trans,lig}}$ .



**Figure 1-8.** Thermodynamic cycle for relative binding energy calculation between two ligands  $\text{Lig}_A$  and  $\text{Lig}_B$  with a receptor (Rec) to form a complex (Com). The cycle depicted equates the experimental relative binding energy with the difference in transforming one of the two ligands to another in the bound and unbound state using TI.

For each individual TI transformation, the combining energy function (mixing function)  $V$  to evaluate the potential energy of the coupled systems can be linear, polynomial, or any other reasonable form. It is found that linear mixing function for thermodynamic integration method

can cause the endpoint singularity problem<sup>58</sup> when the repulsive Lennard Jones term results in a large potential fluctuation as a result of clashing between solvent molecules and a disappearing atom in the solute. Thus we employed the soft-core potential mixing function implemented in AMBER since version AMBER11 to address this problem.<sup>58</sup> It allows the van der Waals term (Eq. 1-2, 1-3) and electrostatic term (Eq. 1-4, 1-5) being evaluated at the same time in one transformation step.

$$V_{V_0,disappearing,vdw} = 4\epsilon(1 - \lambda) \left[ \frac{1}{\left[\alpha\lambda + \left(\frac{r_{ij}}{\sigma}\right)^6\right]^2} - \frac{1}{\alpha\lambda + \left(\frac{r_{ij}}{\sigma}\right)^6} \right] \quad (1-2)$$

$$V_{V_1,appearing,vdw} = 4\epsilon\lambda \left[ \frac{1}{\left[\alpha(1-\lambda) + \left(\frac{r_{ij}}{\sigma}\right)^6\right]^2} - \frac{1}{\alpha(1-\lambda) + \left(\frac{r_{ij}}{\sigma}\right)^6} \right] \quad (1-3)$$

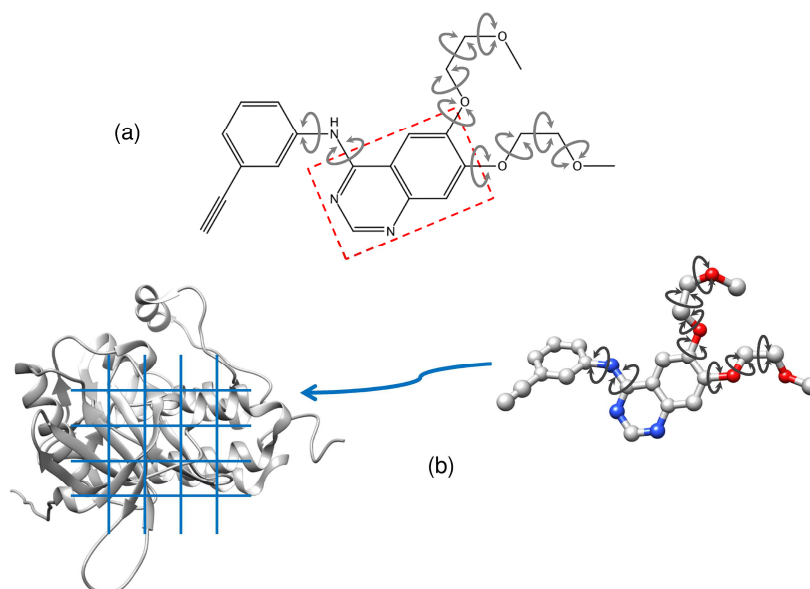
$$V_{V_0,disappearing,elec} = (1 - \lambda) \frac{q_i q_j}{4\pi\epsilon_0 \sqrt{\beta\lambda + r_{ij}^2}} \quad (1-4)$$

$$V_{V_1,appearing,elec} = \lambda \frac{q_i q_j}{4\pi\epsilon_0 \sqrt{\beta(1-\lambda) + r_{ij}^2}} \quad (1-5)$$

## 1.5 Molecular Docking

Fundamentally different from molecular dynamics, molecular docking is a conformational search method to predict individual snapshots of ligand binding poses. It has been historically described as a lock-and-key problem (Emil Fischer, 1894) in molecular recognition. Docking programs such as DOCK,<sup>59-61</sup> Surflex,<sup>62</sup> FlexX,<sup>63</sup> AutoDOCK,<sup>64,65</sup> Glide,<sup>66</sup>

FRED,<sup>67</sup> ICM,<sup>68</sup> and GOLD<sup>69</sup> are widely used to reproduce crystallographic binding poses and perform virtual screening to identify ligands with favorable binding poses to a specific target.



**Figure 1-9.** (a) DOCK anchor-and-grow algorithm. (b) Docking to grid.

Specifically for DOCK, it is designed to accomplish two major tasks: sampling and scoring. An on-the-fly algorithm termed anchor-and-grow was introduced in DOCK4.0 to sample ligand conformations in the binding site of a rigid receptor. As shown in Figure 1-9a, DOCK explores the ligand binding conformational space by first disassembles the small molecule ligand into rigid segments based on rotatable bonds. An anchor (as highlighted in the red box) is selected and then the rest of the fragments in the ligand will be added back at the connection points where the rotatable bonds were broken up previously. The growth processes are done layer by layer from the initial anchor. At each layer, the torsion angles of each added segment will be sampled. The set of partially grown molecules with varies conformations will be minimized, clustered and rank-ordered for pruning. To speed up the energy calculations



during sampling, a receptor grid (shown as blue grid lines) can be pre-generated with atom probes (Figure 1-9b), which is particularly useful in virtual screening when millions of ligands are docked into the same receptor.<sup>70</sup>

Scoring in DOCK is performed using different scoring functions to guide sampling and rank-ordering of ligand poses. Table 1-3 listed some of the more commonly used scoring functions currently available in DOCK. One category of scoring functions is force field-based scores such as grid energy score. The interaction energy of the ligand is evaluated on the pre-computed receptor grid and used to guide ligand growth by prioritize partially grown molecules in terms of geometric and chemical fitting to the binding pocket as well as rank-order complete ligand poses at the end of the docking event. With DOCK6.5, Balias *et al* introduced a per-residue energy decomposition method (footprint) to evaluate similarity of docked ligand pose to a reference active pose.<sup>71,72</sup> The footprint-based metric was shown to boost docking reproduction success rate and has been applied to select lead molecules in virtual screening.<sup>72</sup> In this study, we have implemented a new pharmacophore-based similarity metric as a DOCK scoring function. Further definition and application of the method are described in Chapter 2 and 3.

**Table 1-3.** Commonly used scoring functions in DOCK

Scoring Function	Definition	Version
contact score <sup>73</sup>	Summation of heavy atom contact	DOCK3.0
grid energy score <sup>70</sup>	Non-bonded MM FF terms calculated on the grid	DOCK3.0
continuous energy score	Non-bonded MM FF terms calculated in Cartesian space	DOCK3.0
Zou GB/SA score <sup>74-76</sup>	Fast algorithm for ligand binding affinity calculations	DOCK5
Hawkins GB/SA score <sup>77,78</sup>	MM-GBSA energy	DOCK6.0
AMBER score <sup>79</sup>	MM-GBSA energy calculated with AMBER force field	DOCK6.0
footprint similarity score <sup>71</sup>	Similarity of per-residue decomposition to a reference	DOCK6.5
multigrid FPS score <sup>80</sup>	Footprint similarity measured in multiple grids	DOCK6.6
SASA score	The percentage exposure of a ligand	DOCK6.6

## 1.6 Research Projects

This dissertation describes several research projects in both method development of structure-based drug design and application to potential clinical targets including HIVgp41 employing prior biochemical information of the target systems. Chapter 2 describes the implementation of a pharmacophore-based scoring function termed Pharmacophore Matching Similarity (FMS) in DOCK6 with validation using pose reproduction, crossdocking and enrichment study. Applications of FMS score in virtual screening and *de novo* design targeting the HIVgp41 hydrophobic and inner pocket are reported in Chapter 3 and Chapter 4. Chapter 5 reports the relative free energy calculation results in HIVgp41-T20 complex systems using thermodynamic integration method as well as per-residue energetic and structural analyses to characterize the binding profile of peptide fusion inhibitor T20. Chapter 6 discusses the scientific impact of this study, challenges encountered in each project, and future directions are presented. Appendix A documents the protocol, sample runs and parameter definitions associated with FMS code to be released in the next release of DOCK (DOCK6.8). Appendix B documents the procedure of generating pharmacophore models for visual inspection in Chimera. Appendix C documents an initial system preparation procedure using CHARMM-GUI lipid builder<sup>81</sup> and MD simulation protocols of membrane-bounded systems using *pmemd.cuda* in AMBER14.

## **Chapter 2. Pharmacophore-Based Similarity Scoring Method for DOCK**

This Chapter has been published as **Jiang, L.**; Rizzo, R. C. Pharmacophore-Based Similarity Scoring for DOCK. *J. Phys. Chem. B*, **2015**, *119*(3), 1083-1102. Copyright © 2014 American Chemical Society. [DOI: 10.1021/jp506555w](https://doi.org/10.1021/jp506555w) PMID: 25229837

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

### **Abstract**

Pharmacophore modeling incorporates geometric and chemical features of known inhibitors and/or targeted binding sites to rationally identify and design new drug leads. In this study, we have encoded a three-dimensional pharmacophore matching similarity (FMS) scoring function into the structure-based design program DOCK. Validation and characterization of the method are presented through pose reproduction, crossdocking, and enrichment studies. When used alone, FMS scoring dramatically improves pose reproduction success to 93.5% (~20% increase) and reduces sampling failures to 3.7% (~6% drop) compared to the standard energy score (SGE) across 1043 protein-ligand complexes. The combined FMS+SGE function further improves success to 98.3%. Crossdocking experiments using FMS and FMS+SGE scoring, for six diverse protein families, similarly showed improvements in success, provided proper pharmacophore references are employed. For enrichment, incorporating pharmacophores during sampling and scoring, in most cases, also yield improved outcomes when docking and rank-

ordering libraries of known actives and decoys to 15 systems. Retrospective analyses of virtual screenings to three clinical drug targets (EGFR, IGF-1R, and HIVgp41) using x-ray structures of known inhibitors as pharmacophore references are also reported, including a customized FMS scoring protocol to bias on selected regions in the reference. Overall, the results and fundamental insights gained from this study should benefit the docking community in general, particularly researchers using the new FMS method to guide computational drug discovery with DOCK.

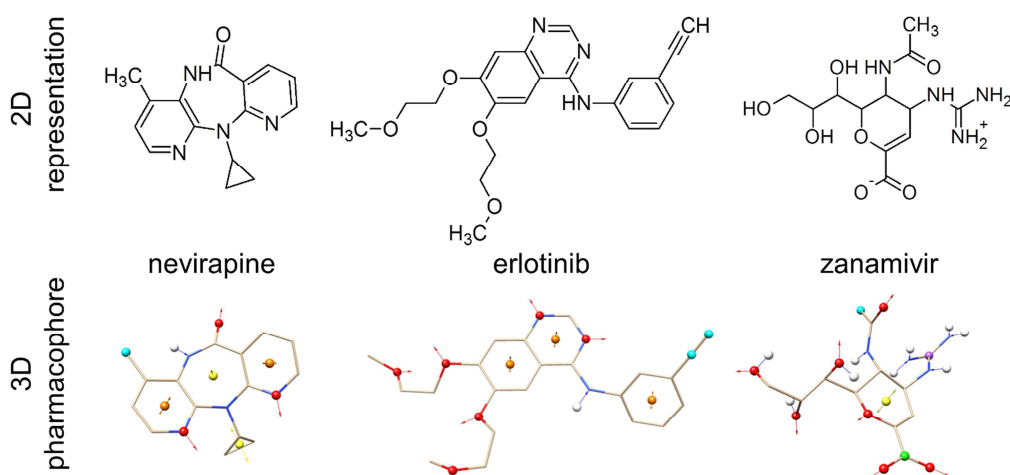
## 2.1 Introduction

Many docking and virtual screening programs, such as DOCK,<sup>60,82</sup> employ intermolecular interaction energy functions that contain non-bonded van der Waals and electrostatic terms to rank-order (i.e. score) small molecule binding geometries (poses) generated in the context of a defined protein binding site. Other physically reasonable scoring terms such as intermolecular hydrogen-bonding, ligand desolvation, numbers of ligand rotatable bonds, buried surface area, among others, have also been explored.<sup>83</sup> In all cases, the objective is to enrich for ligands with good geometric and chemical compatibility with the target so that promising drug-like leads can be identified.<sup>19-21</sup> Recently, Balias *et al*<sup>71,80</sup> reported a new DOCK scoring method termed *footprint similarity score* which can be used to identify compounds that match a specific molecular interaction energy pattern (i.e. footprint) based on a known reference ligand. Encouraged by the recent successes<sup>72,84</sup> from our laboratory, in which "footprints" were used to identify promising lead compounds, we have developed an analogous *similarity-based* scoring method for DOCK that employs "pharmacophores". Both methods yield enhanced docking outcomes but do so in an orthogonal sense (energy vs. geometry).

Historically, the concept of a pharmacophore is generally attributed to Ehrlich,<sup>85,86</sup> and has evolved to include the three dimensional spatial arrangements of key chemical features essential for compound affinity leading to a biological effect.<sup>87,88</sup> A thorough summary of the development of pharmacophores and early works on modeling can be found in a recent publication by Güner *et al.*<sup>87</sup> Reviews by Leach *et al.*,<sup>89</sup> Yang,<sup>90</sup> and Sanders *et al.*<sup>91</sup> also discuss technological advances and challenges of using different pharmacophore methods in modern drug discovery. In practice, pharmacophore features can be derived from known active ligand(s), a defined binding site geometry, or a combination of both. Importantly, the abundance of atomic-resolution structures publically available in the protein data bank (PDB)<sup>45</sup> can be used to derive pharmacophore models for compounds with verified experimental activity to help guide structure-based drug design. A partial list of programs that incorporate pharmacophore modeling includes CATALYST,<sup>92</sup> GASP,<sup>93</sup> LigandScout,<sup>94</sup> PHASE,<sup>95</sup> GALAHAD,<sup>96</sup> PhDOCK,<sup>97,98</sup> and MOE,<sup>99</sup> among others. While such prior efforts are important tools and represent different approaches for modeling, the goal of the present work is to provide a pharmacophore method that can leverage DOCK's powerful anchor-and-grow sampling strategy while taking advantage of different combinations of scoring functions.

The new DOCK pharmacophore scoring protocol termed Pharmacophore Matching Similarity (FMS) encodes useful chemical features including hydrogen bond acceptors/donors, hydrophobic groups, positively/negatively charged groups, and aromatic/non-aromatic rings. Initial pharmacophore types are generated based on atom type and chemical environment, defined by neighboring atoms in the same ligand molecule, and are processed to create a pharmacophore feature set (ph4 model) with coordinates and directionality as shown in Figure 2-1 for three representative drug-like compounds. Importantly, the amount of overlap (termed

FMS score) between a user-supplied reference ligand pharmacophore and candidate pharmacophores derived from docked compounds can be computed on-the-fly during docking (or rescoring) without the need for a separate pre-processing step. This enables large virtual screening libraries to be sorted (i.e. rank-ordered) with the function in an efficient manner.



**Figure 2-1.** 2D representations for three approved drugs (top) and corresponding DOCK pharmacophore (ph4) models (bottom). Features include: (i) hydrogen bond acceptor in red, (ii) hydrogen bond donor in blue (iii) hydrophobic atom/group in cyan, (iv) aromatic ring center and direction in orange, (v) non-aromatic ring center and direction in yellow, (vi) negatively charged group center in green, and (vii) positively charged group center in magenta. Structures of nevirapine, erlotinib, and zanamivir from PDB codes 1VRT, 1M17, and 1A4G, respectively.

Specific validation tests used in this work to evaluate the new scoring protocol include pose reproduction, crossdocking, and enrichment. All FMS results are compared relative to using the standard DOCK single grid energy (SGE) approach, as well as a combined scoring function (FMS+SGE) consisting of both terms. In pose reproduction, crystallographic ligand positions are used as a reference to test if a given method is capable of reproducing native-like poses (within 2 Å of the x-ray pose) using the large SB2012 validation database (update of SB2010)<sup>100</sup> developed in our laboratory. In crossdocking, select protein families from SB2012

(based on high sequence homology), are employed to evaluate docking accuracy across an NxN matrix when all ligands from a family are docked to each individual receptor. In enrichment, active ligands and accompanying decoy compounds taken from the DUD-E<sup>101</sup> database are docked to 15 different targets to assess the ability of the new scoring schemes to correctly rank-order active ligands earlier than decoys. Finally, retrospective analyses of three virtual screens to targets of pharmaceutical interest (EGFR, IGF-1R, and HIVgp41) are shown in which FMS-based scoring (FMS and FMS+S<sub>GE</sub>) was used as a data-mining tool to identify compounds with high pharmacophore overlap to small molecules or peptide ligand sidechains. Overall, the results of this comprehensive study suggest the new method will be a useful addition to the growing number of scoring and sampling methods available in DOCK.

## **2.2 Theoretical Methods**

### **2.2.1 Pharmacophore Definitions.**

Pharmacophore modeling in this study uses a two-step protocol involving: (1) assignment of a pharmacophore type definition to each ligand atom, followed by (2) construction of pharmacophore points with pharmacophore labels based on the type definitions. Inspired by chemical matching code previously developed for DOCK<sup>97,102</sup> we employ a type definition model based on SYBYL<sup>103</sup> atom types and environment (neighboring atoms). The finite list of pharmacophore type definitions is stored in the *ph4.defn* parameter file (Table 2-1) and can be customized to include other pharmacophore types. For clarity it is important to emphasize there is a distinction between pharmacophore type definitions (for the individually-typed atoms) and the pharmacophore label definitions (for the final constructed pharmacophore points) derived from the pharmacophore types. In the atom environment definition list in Table 2-1, parenthesis

( ) specify “atoms that *must be* bonded to the parent atom” while square brackets [ ] specify “atoms that *must not be* bonded to parent atom”.<sup>104</sup> The integer in the definition represents the number of atoms associated in the rule. For example the syntax “N.pl3 ( 2 \* ) [ H ]” specifies a trigonal planar nitrogen connected to at least two other atoms and not bound to any hydrogen atoms. For this work, eight pharmacophore types are assigned to individual atoms as outlined in Table 2-1: (1) null or no assignment, (2) hydrophobic, (3) hydrogen bond donor, (4) hydrogen bond acceptor, (5) aromatic ring member, (6) hydrogen bond acceptor in an aromatic ring, (7) negatively charged species, and (8) positively charged species. The resulting atom set is post-processed to generate pharmacophore points with coordinates that specify the position of the pharmacophore point center and vectors indicating the direction of potential interactions.

**Table 2-1.** Pharmacophore type definitions in DOCK.

<b>type name<sup>a</sup></b>	<b>environment definition<sup>b</sup></b>
(1) null	*
(2) hydrophobic	C. [ O. ] [ N. ] [ S. ] [ F ] [ P ] ( * ) C. ( N.pl3 ( 2 C. ) ) ( * ) N.pl3 ( 3 C. )
(3) donor	H ( O. ) H ( N. ) H ( S. ) H ( F )
(4) acceptor	O. ( * ) N.1 ( 1 * ) N.2 [ 3 * ] N.3 ( 3 * ) N.pl3 ( 2 * ) [ H ] S.2 [ O. ] [ N. ] S.3 ( 2 * ) F ( * ) Cl ( * )



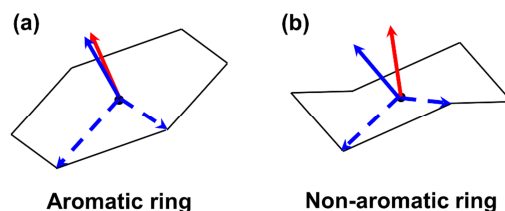
(5) aromatic	C.ar N.ar
(6) aroAcc	N.ar [ H ] [ 3 * ] ( * )
(7) negative	C. ( 2 O.co2 ) C.2 ( O.2 ) ( O.3 [ * ] ) P. ( 4 O. ) ( O.3 [ * ] ) S. ( 3 O. ) ( O.3 [ * ] ) S. ( 4 O. ) ( O.3 [ * ] ) F [ * ] Cl [ * ]
(8) positive	C.cat ( * ) N.4 ( * ) N.3 ( 4 * ) N.2 ( 3 * ) Zn [ * ] Mg [ * ] Ca [ * ] Mn [ * ] K [ * ] Fe [ * ]

---

<sup>a</sup>Types defined in DOCK *ph4.defn* parameter file. <sup>b</sup>Environments based on SYBYL atom types and atom connectivities.

Aromatic and non-aromatic rings are identified by checking for closed loops formed by connected atoms. The coordinate of the ring center, averaged over all ring member coordinates, is computed and saved as the pharmacophore point (Figure 2-2). The average normal vector of the plane defined by adjacent ring center-to-vertex vectors (Figure 2-2, dashed blue lines) is calculated and saved as the direction vector of the pharmacophore point. If the individual normal vectors (Figure 2-2, solid blue lines) of the ring are all within an angle cutoff  $\theta_c$  to the average normal vector (Figure 2-2a, solid red lines), then the pharmacophore point is marked as an aromatic ring (Figure 2-2a). Otherwise it is labeled as non-aromatic (Figure 2-2b). In practice,  $\theta_i$  is measured by directly computing the inner product of two vectors ( $x_i$ ) which is converted to degrees by the inverse function of cosine as  $\arccos(x_i) = \theta_i$ . Based on examining

crystallographic ligand coordinates containing aromatic and non-aromatics rings we use as a cutoff criteria  $\arccos(0.99) \approx 8.11$  degrees to determine if a ring is planar.

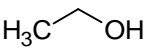
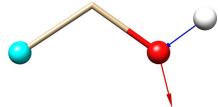
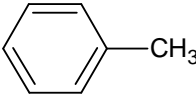
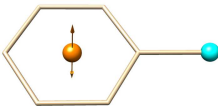
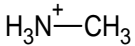
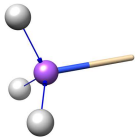
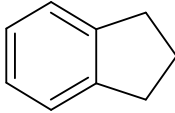
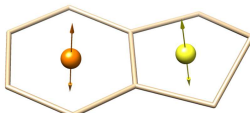
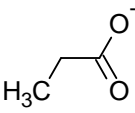
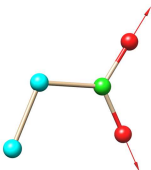
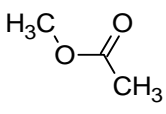
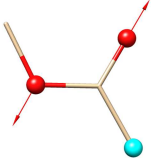


**Figure 2-2.** Pharmacophore feature assignment for rings: (a) aromatic (close to planar) and (b) non-aromatic (not planar). Ring center-to-vertex vectors shown as dashed blue lines, individual normal vector shown as solid blue lines, averaged normal vectors shown in solid red lines. The angle between the blue and the red vectors are compared to a threshold to determine the planarity of the ring.

Atoms with hydrophobic and positive/negative pharmacophore type definitions are saved individually as pharmacophore points inheriting the same type as their pharmacophore labels. For these cases, default direction vectors (which do not affect the score) are assigned to facilitate a common data structure. For the hydrogen bond acceptor, the coordinate of the polar atom is saved as the pharmacophore point. The average of vectors pointing from all neighbor atoms to the acceptor atom is saved as the direction vector, indicating the potential position of the coupling hydrogen bond donor, as indicated by red arrows in Table 2-2 which shows example pharmacophores derived for several small organic molecules. The hydrogen bond donor uses the coordinate of the hydrogen atom as that of the pharmacophore point. Similarly, the vector pointing from the donor hydrogen to the connecting polar atom is saved as a normalized direction vector, indicated as blue arrows in Table 2-2. The combined set of all pharmacophore points is called the molecular pharmacophore (ph4) model which may include hydrophobic (PHO), hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), aromatic ring (ARO),

non-aromatic ring (RNG), positively charged (POS) and negatively charged (NEG) features (see Table 2-2).

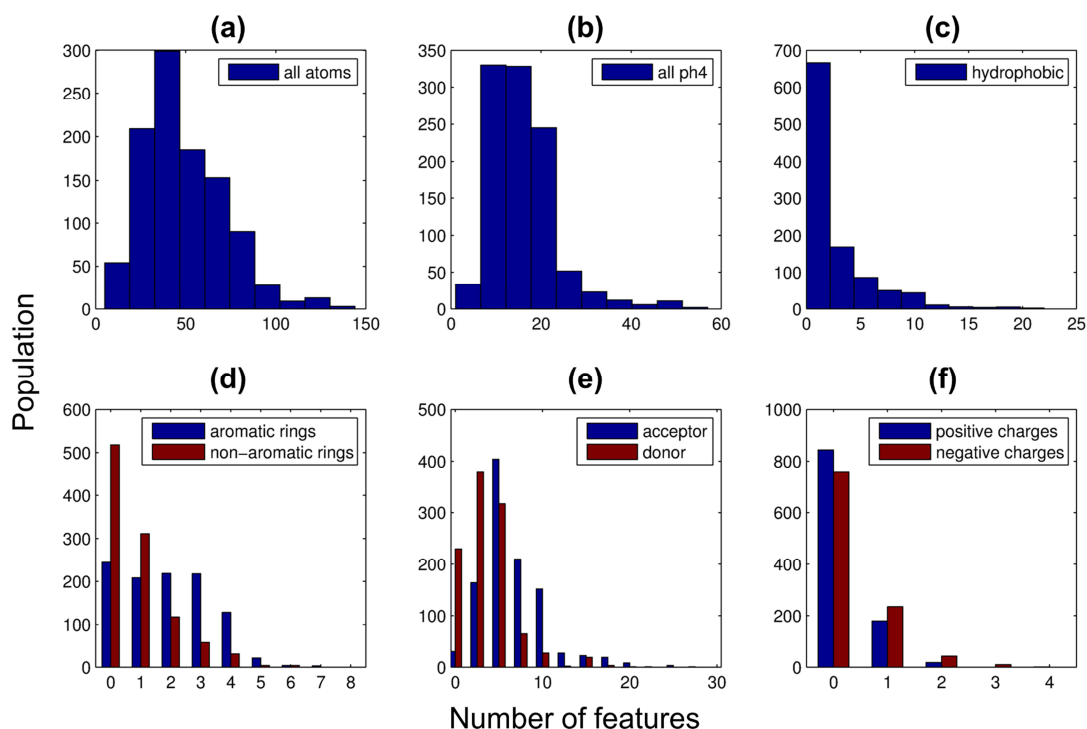
**Table 2-2.** Examples of pharmacophore features derived from small molecules.

Name	Label <sup>a</sup>	2D <sup>b</sup>	3D <sup>c</sup>
(a) ethanol	PHO HBD HBA		
(b) toluene	ARO PHO		
(c) methyl-amine	POS HBD		
(d) indane	ARO RNG		
(e) propanoic acid	PHO NEG HBA		
(f) methyl acetate	PHO HBA		

<sup>a</sup>PHO in cyan; HBA (vertex and vector) in red; HBD vector in blue, hydrogen vertex in gray; ARO (vertex and vector) in orange; RNG (vertex and vector) in yellow; POS in magenta; NEG in green. Direction vectors are shown in arrows generated using *Chimera*<sup>105</sup> *bild* files. <sup>b</sup>2D pictures generated with *ChemSketch*.<sup>106</sup> <sup>c</sup>3D molecules and pharmacophore visualization generated with *Chimera*.

To gauge how many pharmacophore features are present in typically sized compounds, Figure 2-3 plots histograms derived from 1043 molecules in their x-ray pose taken from the SB2012 testset used in this work to gauge pose reproduction and crossdocking accuracy. As a reduced representation, the pharmacophore model derived for each molecule contains (on average) a much smaller number of pharmacophore points (16.0) relative to the total number of atoms (49.2) as shown in Figure 2-3b vs. 3-3a. In terms of specific features, molecules in SB2012 contain on average of 1.9 aromatic rings, 0.9 non-aromatic rings, 2.6 hydrophobic groups, 3.6 hydrogen bond donors, and 4.6 hydrogen bond acceptors. Values for the two latter features are indicative of the drug-like characteristics of many of the compounds in SB2012 for which ~80% have less than 5 hydrogen bond donors and ~58% have less than 10 hydrogen bond acceptors in rough agreement with Lipinski-like<sup>107</sup> rules. About 1/4 of the testset contains molecules with positively (199) or negatively (287) charged functionality.

In principle, given the smaller feature space, use of pharmacophore models should yield faster run times than an all-atom based scoring function. In terms of rescoring poses without sampling, timing tests indicate that under the current conditions, computing the pharmacophore matching similarity (FMS) score between two molecules is faster than computing the standard energy score by about 3.5 fold. Comparing production times when using the FMS method to drive ligand sampling is less straightforward, due to the much larger numbers of poses generated when using FMS compared to SGE (discussed further below). However, when normalized by the size of the final pose ensemble retained using FMS or SGE methods, time per pose with FMS is faster by about 1.5 fold.



**Figure 2-3.** Number of pharmacophore features computed by DOCK FMS scoring function using the SB2012 docking testset (N=1043 molecules).

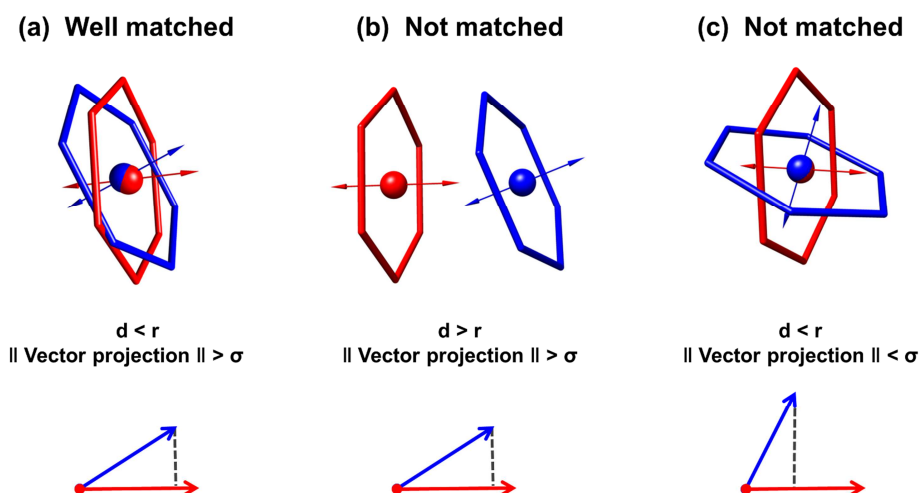
### 2.2.2 Pharmacophore Matching Similarity (FMS) Scoring Function.

After computing the pharmacophore (ph4) model using the protocol described above for both the reference and candidate poses, molecular similarity between the two poses is evaluated by the degree of pharmacophore overlap, termed here pharmacophore matching similarity score (FMS score). For each pharmacophore point  $A$  with pharmacophore label  $a$ , Cartesian coordinate  $\vec{x} = [x_1, x_2, x_3]$  and direction vector  $\vec{v} = [v_1, v_2, v_3]$  in the reference pharmacophore, is compared to every pharmacophore point  $B_i$  in the candidate pharmacophore in three steps: (i) label check, (ii) distance check and (iii) direction check. The pharmacophore label  $a$  is used to eliminate pharmacophore points in the candidate pharmacophore that have different labels. The distance between  $A$  and the candidate pharmacophore point  $B_i$ , computed

as  $d_i = \|\vec{x} - \vec{y}_i\| = \sqrt{\sum_{j=1}^3 (x_j - y_j^i)^2}$  where  $\vec{y}_i = [y_1^i, y_2^i, y_3^i]$  is the Cartesian coordinate of  $B_i$ , is compared to a distance cutoff  $r$ . Only when  $d_i \leq r$  will the corresponding pharmacophore point  $B_i$  be further investigated. A constant  $r$  value is assigned to all reference pharmacophore points as a default parameter, but for a ring (aromatic or non-aromatic) the radius of the ring is assigned as  $r$ . The scalar projection of the normalized direction vector  $\vec{v}$  onto that of  $B_i$ ,  $\vec{w}_i = [w_1^i, w_2^i, w_3^i]$  is calculated. The condition that the vector projection  $\vec{v} \cdot \vec{w}_i = \sum_{j=1}^3 v_j \times w_j^i \geq \sigma$  implies the angle between the two direction vectors  $\vec{v}$  and  $\vec{w}_i$  is within  $\arccos\sigma$ , which ensures that the two vectors are pointing in approximately the same direction. A perfect vector overlap (when  $\vec{v} = \vec{w}_i$ ) between two normalized direction vectors will be  $\vec{v} \cdot \vec{w}_i = \|\vec{v}\| = 1$ . By default, a scalar projection cutoff of  $\sigma = \cos(45^\circ) \approx 0.7071$  is used. Note that for hydrophobic and charged feature labeled points,  $\vec{v} \cdot \vec{w}_i \geq \sigma$  is always true as the same default value of (1,0,0) is assigned to both  $\vec{v}$  and  $\vec{w}_i$ . For a ring, the absolute value of the scalar projection  $|\vec{v} \cdot \vec{w}_i|$  is used to account for its orientation (i.e. vectors above and below the plane of the ring). If all of the above criteria are met, then the two pharmacophore points  $A$  and  $B_i$  are deemed a match.

In Figure 2-4, three ARO pharmacophore point pairs are shown to illustrate how the three criteria (label, distance, and direction) are used to identify matches in rings. The first criterion (same label) is met by all three pairs as the pharmacophore points shown are all labeled as aromatic rings (ARO). The first pair (Figure 2-4a) has both a small distance ( $d \leq r$ ) and good directional agreement ( $|\vec{v} \cdot \vec{w}_i| > \sigma$ ) and thus represents a well-matched case. The second pair (Figure 2-4b), although the ring vectors are well aligned, is not matched due to the large distance between the pharmacophore centers. For the third pair (Figure 2-4c), although the distance

between ring centers is small, this case is also not considered a match due to the large difference in ring vector orientation ( $|\vec{v} \cdot \vec{w}_i| < \sigma$ ).



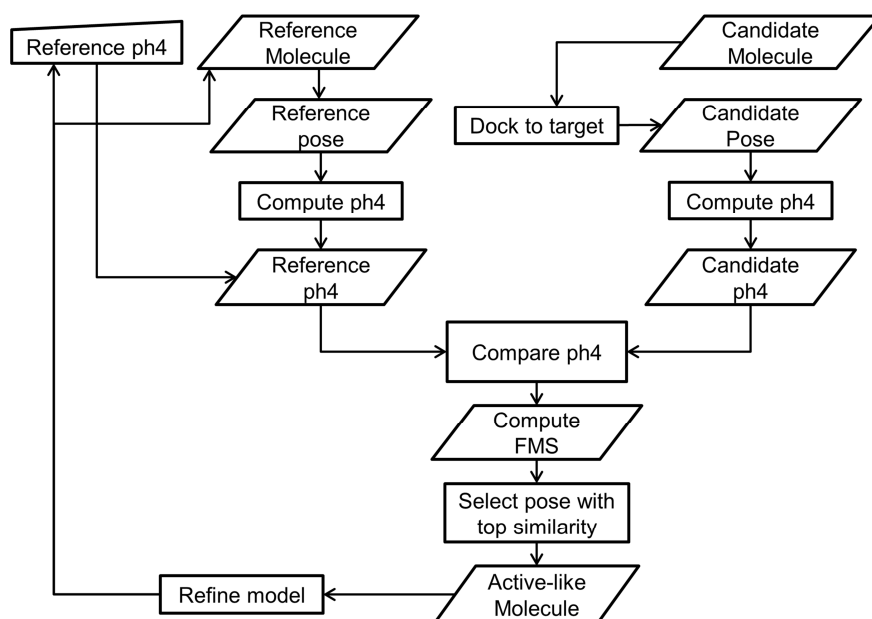
**Figure 2-4.** Example pharmacophore matches for aromatic rings showing: (a) well matched case with same labels, small distance, and similar vector directions, (b) not matched case with same labels, large distance, and similar vector directions, and (c) not matched case with same labels, small distance, and different vector directions.

All matched point pairs between the reference and candidate pharmacophore models are investigated by their geometric relationships to obtain a quantitative measurement of matching. The residual between two matched points is defined as  $\delta_A^i = \sqrt{(d_i)^2 / |\vec{v} \cdot \vec{w}_i|}$  which takes into account both the distance and overlap in direction. After comparing pharmacophore point  $A$  with all candidate pharmacophore points  $B_i$ , the best matched point  $B_+$  with the lowest matching residual  $\delta_A^+$  will be retained for the pharmacophore matching similarity (FMS) score calculation. If no match was found for  $A$ , then it will not contribute to the residual term of FMS score. The residual term in combination with a match rate term defines the numerical value of the FMS score via eq 2-1.



$$FMS = \begin{cases} k \left(1 - \frac{n}{N}\right) + \sqrt{\frac{\sum_{j=1}^n (\delta_{A_j}^+)^2}{n}}, & n > 0 \\ X, & n = 0 \end{cases} \quad (2-1)$$

Here,  $k$  is a constant parameter;  $n$  stands for the total number of matches (note that for each reference pose pharmacophore point, one match is counted at most);  $N$  is the number of pharmacophore points in the reference pharmacophore;  $\delta_{A_j}^+$  represents the best matching residual of a matched reference pharmacophore point  $A_j$ . Based on similarity measurements in graph theory,<sup>108,109</sup> FMS score uses the match rate term  $k \left(1 - \frac{n}{N}\right)$  to prioritize poses with higher numbers of pharmacophore matches to the reference pose. Poses with similar numbers of matches will be differentiated by their root mean square matching residuals  $\sqrt{\frac{\sum_{j=1}^n (\delta_{A_j}^+)^2}{n}}$ . Note that the total number of matches  $n$  needs to be larger than zero for eq 2-1 to give a reasonable value. When no match is identified ( $n=0$ ), an arbitrary large score  $X$  is assigned ( $X$  is set to be larger than the upper bound of FMS score value when  $n>0$ ). For any reference and candidate pair of molecules, FMS score ranges between 0 (perfect match) and  $X$ , which depend on choices for  $k$ , distance cutoff  $r$ , and scalar projection cutoff  $\sigma$ . For pharmacophore-based docking, lower FMS scores are more desirable. Figure 2-5 outlines schematically the overall process using DOCK.



**Figure 2-5.** Flow chart schematic outlining pharmacophore-based virtual screening in DOCK.

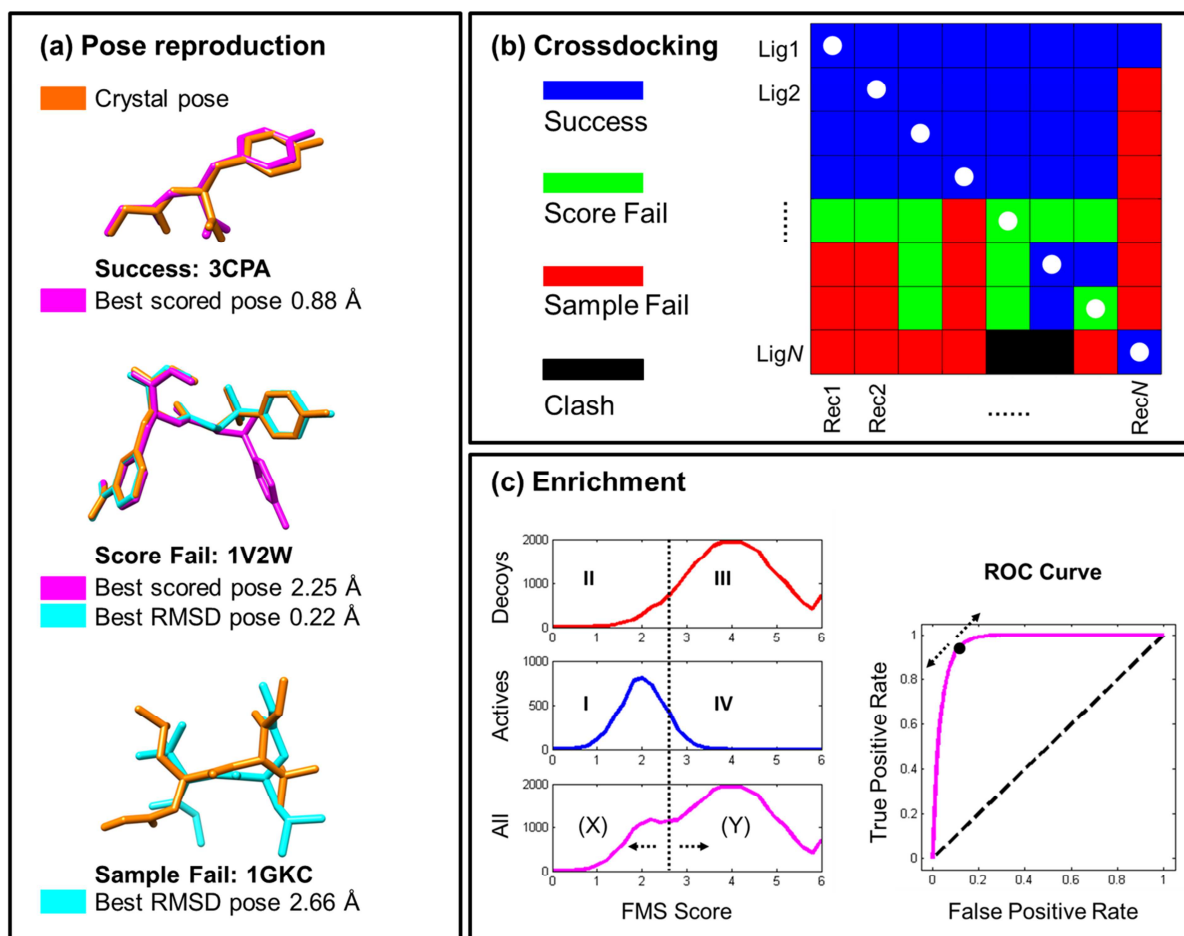
To determine a default set of values for  $k$ ,  $r$ , and  $\sigma$  in 2-1, we performed a series of rescoring tests using ligand geometries generated with the standard DOCK protocol, for comparison with crystallographic references, and pose reproduction success (defined in the next section) was determined. Four values for  $k$  (1, 2, 5, 10), three values for  $r$  (0.5Å, 1.0Å, 1.5Å), and three values for  $\sigma$  (30°, 45°, 60°) were examined. As a general rule, use of stricter matching criteria (shorter distance cutoff  $r$ , smaller angle cutoff  $\sigma$ ) led to lower docking success rates. In addition, the success rate increased as the matching rate term weight  $k$  was increased from 1 to 5, but remained relatively steady from  $k = 5$  to 10. Taking these results into consideration, the set comprising  $k = 5$ ,  $r = 1\text{Å}$ ,  $\sigma = 45^\circ$ , and  $X = 20$  yielded generally good pose reproduction success and had values which were roughly in-between the different ranges explored. Although other combinations might also have been suitable, this set was ultimately employed for all subsequent FMS sampling and scoring experiments used in this work.

## 2.3 Validation Metrics and Computational Details

### 2.3.1 Pose Reproduction Details.

In order to approximate the accuracy of ligand poses predicted by a given protocol for unknown systems, pose reproduction control experiments are performed over a large number of crystallographic complex structures. Ideally, the best-scored docked pose should agree with crystal pose. Following our previous work,<sup>100</sup> docking results are categorized as one of three outcomes: docking success (Success), scoring failures (Score Fail), and sampling failures (Sample Fail). Over a large dataset the percentage of Success + Score Fail + Sample Fail = 100%. Docking success is defined when the RMSD between the best scored pose and native (crystal) pose is  $\leq 2$  Å. A scoring failure is defined when a close-to-native pose is sampled but the best scored pose is  $> 2$  Å from the native pose. Finally, a sampling failure is defined if none of the sampled poses are within 2 Å of the native pose.

Representative visual examples of the three outcomes are shown in Figure 2-6a. For ligands of drug-like size, low RMSD values also typically correspond to good visual overlap between docked and reference ligand poses. All statistics reported in this work make use of "symmetry corrected" RMSDs to account for chemically identical functionality (i.e. symmetric ring flips, carboxylate flips, etc), or completely symmetric molecules, adopting visually indistinguishable conformations as described in detail previously.<sup>110</sup> The updated pose reproduction database termed SB2012 (an update of the SB2010 database),<sup>100</sup> was used for all pose reproduction and crossdocking (defined below) experiments. The set, derived from complexes in the protein databank (PDB), contains 1043 protein-ligand systems in ready to DOCK format and is freely available online at [www.rizzolab.org](http://www.rizzolab.org).



**Figure 2-6.** Validation metrics used to evaluate DOCK scoring functions. (a) Pose reproduction cases with different outcomes: Success (top, PDB code 3CPA), Score Fail (middle, PDB code 1V2W), and Sample Fail (bottom, PDB code 1GKC). Crystal poses in orange, best scored poses in magenta, best RMSD pose in cyan. (b) Representative crossdocking heatmap showing docking outcome as a function of docking all ligands (Lig1, Lig2, ... LigN) to all receptors (Rec1, Rec2, ... RecN), for an aligned group of proteins with nearly identical sequence homology. (c) Hypothetical database enrichment results showing a partitioning of data based on FMS score ranking (0 to 6) for a group of ligands (left bottom, magenta curve) comprised of a known active ligand set (left middle, blue curve) and inactive decoy set (left top, red curve). The vertical dashed line represents a hypothetical FMS score cutoff dividing the total group into (X) predicted positive and (Y) predicted negative sets which can be partitioned in to four quadrants (I-IV) defined respectively as true positives (TP, I), false positives (FP, II), true negatives (TN, III), and false negatives (FN, IV). Also shown is an ROC curve, which for this example plots individual points which correspond to various FMS score cut-offs in the left panel. The coordinate of each point is determined by the false positive rate and true positive rate at that FMS score cut-off.

All DOCK experiments in this work employed well-defined receptor and ligand setup protocols, in conjunction with the flexible ligand sampling protocol termed FLX, as previously described.<sup>100</sup> Briefly, in terms of receptor setup, several accessory programs are used to compute a molecular surface (DMS),<sup>111</sup> generate docking spheres to guide sampling (SPHGEN),<sup>112</sup> and pre-compute the potential energy on a grid which speeds up the docking calculations (GRID).<sup>70</sup> Key setup parameters include use of 6-9 Lennard-Jones and distance dependent dielectric ( $\epsilon=4r$ ), a 0.3Å resolution, and a grid box size extending 8Å in all directions based on the docking spheres (75 spheres max). Key docking parameters include use of the on-the-fly anchor-and-grow algorithm to orient and assemble ligands layer by layer, retaining a maximum of 5000 completely-grown conformers to be ranked by the primary scoring function, and saving a maximum of 100 conformers (after clustering to remove redundancy, RMSD  $\leq$  2 Å). Ligands were energy minimized at each stage of conformational search (500 iterations per cycle per anchor/step max) and those exceeding a total score cutoff of 100.0 were removed.

The different functions employed in this work include: (1) single grid energy (SGE) score, (2) DOCK Cartesian energy (DCE) score which is equivalent to SGE but in Cartesian space, (3) pharmacophore matching similarity (FMS) score, and (4) the combination of the two termed FMS+SGE (or FMS+DCE) score. For the combined function, the FMS score was weighted by 10-fold so when summed together the FMS and SGE (or DCE) terms would be more equally balanced.

### **2.3.2 Crossdocking Details.**

In addition to pose production experiments, crossdocking was employed in which highly homologous protein complexes, with nearly identical structure and sequence (termed here a

protein receptor family), are aligned into a common reference frame and each ligand is docked into each receptor as shown in Figure 2-6b. Such families inherently contain variability due to different crystallization conditions, co-crystallization with different ligands, as well as receptor point mutations, among others. Nevertheless, the hypothesis in crossdocking is that ligands should adopt similar binding geometries in highly homologous receptors, provided there are no large deformations in the binding site or incompatible mutations. The results are expressed as an  $N \times N$  heatmap ( $N$  = number of systems) with docking success plotted in blue, sampling failures plotted in red, and scoring failures are plotted in green (Figure 2-6b). As before, a 2 Å RMSD cut-off is used to evaluate success. The diagonal elements (Figure 2-6b, white dots) represent cognate protein-ligand pairs and thus represent experimental references. Off-diagonal elements are "theoretical" protein-ligand pairs and the reference, in some instances, may be incompatible. To identify incompatible elements, we employ a clash matrix check,<sup>100</sup> independent of the actual crossdocking experiment, in which all matrix complexes (representing cognate and theoretical references) are subject to a short restrained energy minimization. If the minimized ligand pose moves  $>2$  Å from the starting pose, or pose bears an unfavorable energy score ( $>0$  kcal/mol), the specific reference pair containing the clash is not included in crossdocking success evaluations (Figure 2-6b black squares). All crossdocking studies employed the FLX docking protocol, and results are reported for both the diagonal and the entire matrix.

### 2.3.3 Enrichment Details.

A third method used to evaluate docking methods is enrichment (Figure 2-6c). Databases such as the directory of useful decoys (DUD),<sup>113</sup> and the newer enhanced version called DUD-E,<sup>101</sup> contain large sets of known active compounds (and property-matched decoys) which are

docked to a specific target and the results are rank-ordered. Good enrichment is achieved when greater numbers of actives are ranked earlier in the list compared to the decoys. For more in-depth discussion on using DOCK to estimate enrichment interested readers should consult Brozell *et al.*<sup>114</sup> Briefly, for this work, ranked results were visualized as receiver operating characteristic (ROC) curves which plots how the true positive rate (true positive/total positive) changes relative to the false positive rate (false positive/total negative). Accompanying area under the curve (AUC) analysis was also performed and used to estimate fold enrichment values ( $FE = AUC/AUC_{\text{random}}$ ), relative to random, at 0.1%, 1%, 10%, and 100% of the database examined. For virtual screening, early enrichment is of particular importance, as typical applications will only focus on (i.e. purchase) small subsets of molecules ranked very early (i.e. 0.1-1%) in the database. In the theoretic example shown in Figure 2-6c, which employed FMS score to rank active and decoy ligands shown in the left panels, the ROC curve on the right represents a good enrichment case relative to random (Figure 2-6c magenta vs dashed line). By specifying a specific score cut-off (Figure 2-6c left bottom panel, dashed line) the data can also be partitioned into two groups for which molecules with smaller scores (better overlap) are defined as predicted positives (X), and molecules with higher scores (worse overlap) defined as predicted negatives (Y). If, as in the present example, the results are in fact known, this allows ligands in the active group to be classified as true positive (I) or false negative (IV), and ligands in the inactive (decoy) group classified as false positive (II) and true negative (III). By varying the cut-off, the number of molecules in the four subsets I–IV will change accordingly.

Enrichment studies employed the 15 DUD-E systems shown in Table 2-3.<sup>101</sup> The receptor PDB files were already available in SB2012 (same PDB code as DUD-E) and the active and decoy ligands were downloaded from the DUD-E website and used as is. It is important to

note that some ligands (active and decoys) for these systems contain multiple entries representing, for example, different tautomers or protonation states. For all enrichment analyses, in the case of duplicate id codes, only the best-scored molecule was retained. For each system, the native cognate ligand in the original PDB file is used as the pharmacophore reference for FMS scoring. As in the pose reproduction and crossdocking studies, the enrichment tests also employed the FLX docking protocol. With this protocol, predicted ligand poses with accompanying scores were obtained for approximately all but 2% of the actives and decoys listed in Table 2-3.

**Table 2-3.** Systems used for enrichment tests.

<b>PDB</b>	<b>System</b>	<b>#Actives<sup>a</sup></b>	<b>#Decoys<sup>a</sup></b>	<b>Description</b>
<b>2HZI</b>	abl1	295	10885	tyrosine-protein kinase ABL
<b>1E66</b>	aces	664	26373	acetylcholinesterase
<b>2VT4</b>	adrb1	458	15958	beta-1 adrenergic receptor
<b>1L2S</b>	ampc	62	2902	beta-lactamase
<b>1BCD</b>	cah2	835	31710	carbonic anhydrase II
<b>1R9O</b>	cp2c9	183	7574	cytochrome P450 2C9
<b>2RGP</b>	egfr	832	35442	epidermal growth factor receptor erbB1
<b>1SJ0</b>	esr1	627	20818	estrogen receptor alpha
<b>3CCW</b>	hmdh	299	8884	HMG-CoA reductase
<b>1UYG</b>	hs90a	125	4942	heat shock protein HSP 90-alpha
<b>2AA2</b>	mcr	193	5240	mineralocorticoid receptor
<b>1KVO</b>	pa2ga	127	5216	phospholipase A2 group IIA
<b>2GTK</b>	pparg	723	25867	peroxisome proliferator-activated receptor gamma
<b>1NJS</b>	pur2	201	2725	GAR transformylase
<b>1C8K</b>	pygm	114	4045	muscle glycogen phosphorylase

<sup>a</sup>Systems taken from DUD-E database. <sup>101</sup>



## 2.4 Results and Discussions

### 2.4.1 Pose Reproduction Results.

Table 2-4 shows pose reproduction outcomes computed for the three DOCK protocols tested (SGE, FMS, FMS+SGE) in which a given function was used for both sampling and scoring (diagonal blocks in gray box) or when rescored using the other two scoring functions (off-diagonal blocks). All experiments were performed under the same conditions except for the sampling and/or scoring method employed. It is important to emphasize that use of an alternative function to re-rank an ensemble of poses generated by any given method (Table 2-4, off-diagonal blocks) will, in most cases, lead to a different group of top-scored results, but the number of sampling failures remains unchanged.

**Table 2-4.** Pose reproduction results employing SGE, FMS, and FMS+SGE scoring functions.

scoring	Outcome	sampling <sup>a</sup>					
		SGE		FMS		FMS+SGE	
SGE	Success	756	72.5%	610	58.5%	854	81.9%
	Score Fail	185	17.7%	394	37.8%	182	17.4%
	Sample Fail	102	9.8%	39	3.7%	7	0.7%
FMS	Success	860	82.5%	975	93.5%	1035	99.2%
	Score Fail	81	7.8%	29	2.8%	1	0.1%
	Sample Fail	102	9.8%	39	3.7%	7	0.7%
FMS+SGE	Success	876	84.0%	719	68.9%	1025	98.3%
	Score Fail	65	6.2%	285	27.3%	11	1.1%
	Sample Fail	102	9.8%	39	3.7%	7	0.7%

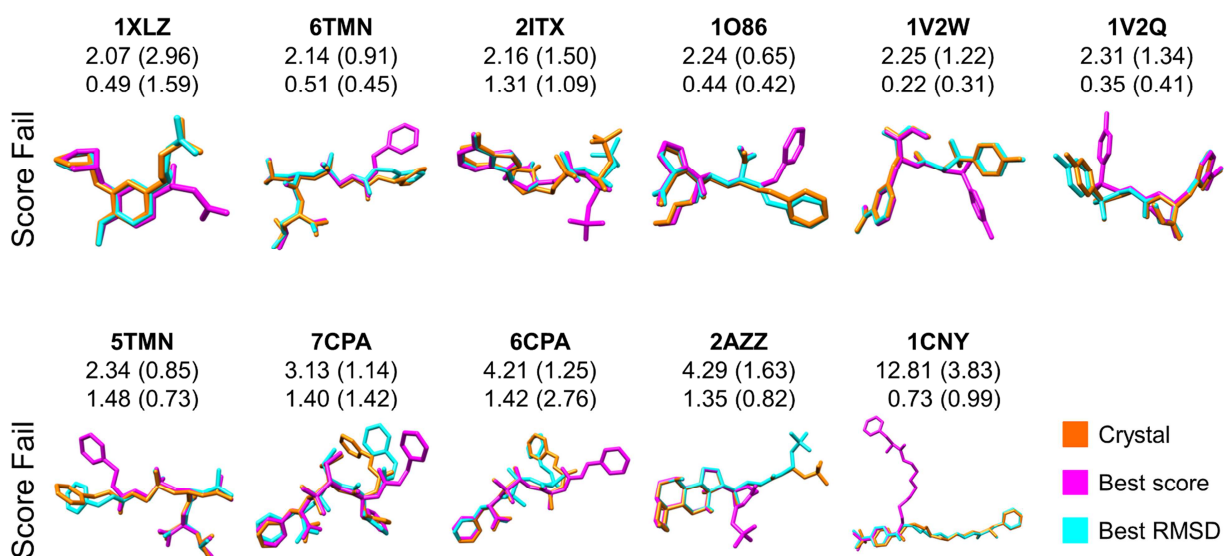
<sup>a</sup>SGE sampling size = 89,083 poses, FMS sampling size = 337,674 poses, FMS+SGE sampling size = 59,237 poses.

In general, the diagonal results (Table 2-4, gray boxes) using the three different methods yield high percentages of success across the 1043 systems in SB2012 with the FLX ligand

protocol. Importantly, the SGE success rate (72.5%) is consistent with earlier work from our laboratory,<sup>71</sup> using a smaller dataset (68.5%, N=780), indicating good reproducibility of DOCK. Overall, the diagonal results in Table 2-4 reveal a clear trend in terms of outcome with success following  $SGE < FMS < FMS+SGE$  and sampling and scoring failures following  $SGE > FMS > FMS+SGE$ . The very high success rates when using FMS (93.5%) or the combination FMS+SGE (98.3%) is significant and represents a 20-25% improvement over the standard DOCK method employing SGE (72.5%). On one hand, such high success rates are expected given that for any system the x-ray reference ligand and docked ligand are the same molecule in terms of topology and thus have the exact same number of pharmacophore features. In actual practice, for virtual screening, the number of features between a reference and candidate would change as each new ligand was docked. Nevertheless, the good correspondence in these validation tests provides strong evidence the newly implemented DOCK pharmacophore labeling, modeling, and overlap routines are behaving as expected and yield robust results over a large pose reproduction testset. Importantly, the FMS method is straightforward to use and only requires that the user input a reference molecule consisting of a single 3D conformation. The processing of the candidate pose(s) to determine FMS scores is done automatically and on-the-fly. Ongoing work to allow a text-based pharmacophore reference to be used as a query will further simplify the procedure of customizing inputs for FMS score calculation.

*Systems with failures.* Of the three methods tested, the FMS+SGE protocol yields the lowest sampling (0.7%) and scoring (1.1%) failure rates on the diagonal. In an attempt to understand what led to the small subset of failures (N=18), docked poses for the group were examined. Out of the 7 sampling failures, one system did not complete growth, which, although infrequent, can happen using DOCK under some circumstances. And for the remaining 6

sampling failures, 4 are relatively large molecules with up to 35 rotatable bonds, and thus extremely challenging for any docking protocol. In terms of the 11 scoring failures, a noteworthy result (Figure 2-7) is that 7 out of the 11 systems (PDB code: 1XLZ, 6TMN, 2ITX, 1O86, 1V2W, 1V2Q, 5TMN) actually show good correspondence both in terms of visual overlap as well as RMSD (2.07 - 2.34 Å). Thus, these 7 can be classified as "near misses" for which only a part of the ligand geometry adopts a conformation different than the x-ray pose. Consistent with expectation, in all but two cases (7CPA, 6CPA), geometries corresponding to the best RMSD also have a lower FMS score. The fact that the FMS+SGE protocol correctly identifies a native-like pose in nearly all 1043 cases is notable.



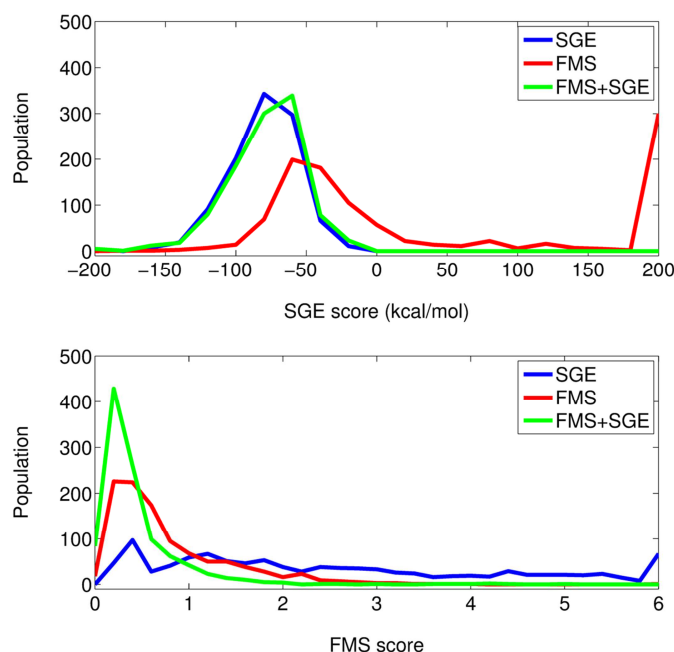
**Figure 2-7.** Eleven scoring failures derived from FMS+SGE guided docking showing overlaid poses, PDB code identifier. RMSD in Å, and FMS scores in parentheses for the best FMS+SGE scored pose (first row, magenta) and the best FMS+SGE RMSD pose (second row, cyan) relative to the crystal pose in orange.

**Rescoring.** In terms of the off-diagonal blocks (Table 2-4), rescoring the standard SGE results (72.5%) with FMS (82.5%) or FMS+SGE (84.0%) reveals a similar trend with SGE < FMS < FMS+SGE as in the diagonal experiments. Here, as rescoring cannot "rescue" incorrectly sampled geometries, the maximum success rate attainable is a function of the poses originally sampled which for SGE is 90.2% (e.g. 100% - 9.8% sampling failures). This specific experiment is important as the improvement in success when rescoring SGE-derived results with FMS or FMS+SGE (10-11%) suggests the current implementation is a viable way to post-process docked poses and identify those compounds with good pharmacophore overlap to a reference. This procedure would be a particularly useful tool to aid virtual screening as discussed further below. Rescoring results for the group derived from FMS+SGE sampling shows similar results, with FMS (99.2%) yielding a significantly higher success rate than SGE (81.9%).

The most dramatic changes in terms of pose reproduction involve using SGE (58.5%) or FMS+SGE (68.9%) to rescore the pose ensembles derived from FMS-only sampling (93.5%). These reduced success rates likely stem from the fact that the FMS score accounts only for overlap between pharmacophore features derived from the reference ligand structure and the receptor is "invisible" during sampling. The end result is that poses generated using FMS alone may clash with the target protein when rescored in "energy space" despite high pharmacophore overlap. However, as the pairing of energy and pharmacophore overlap (FMS+SGE) leads to relatively high success rates when rescored in SGE-space, as noted above, the combined function is likely to be preferred when a receptor structure is available. Nonetheless, the 58% success rate obtained with SGE rescoring can be considered encouraging considering that ligand sampling with the anchor-and-grow algorithm was done in the absence of a receptor. Thus, for ligand-

only based design, the FMS protocol appears to be capable of enriching for energetically favorable poses by matching only to a reference pharmacophore. The caveat of course is identifying suitable pharmacophores in the absence of crystallographic information.

***Ensemble properties.*** A protocol designed to enrich for ligands with poses close to a native structure should, in theory, yield favorable scores using any reasonable scoring function. To examine in more detail how properties of molecules generated with one protocol may differ when rescored with another, histograms of the resultant SGE and FMS scores were plotted using each of the three different pose ensembles obtained with SGE, FMS, or FMS+SGE methods. As expected, and consistent with the rescoring results in Table 2-4, use of the FMS function alone to derive poses does lead to overall less favorable DOCK energies (Figure 2-8 top, red) when rescored in SGE-space compared to FMS+SGE (Figure 2-8 top, green) or SGE (Figure 2-8 top, blue). The large positive peak at 200 kcal/mol (Figure 2-8 top, red) represent those systems for which large positive energies were obtained due to geometric clashes occurring between ligand and protein. However, an encouraging number of the poses derived from FMS sampling do yield favorable energies. At first glance, the fact that the SGE and FMS+SGE energy histograms (Figure 2-8 top, blue and green) are nearly superimposable is somewhat surprising, especially considering the two ensembles yield substantially different success rates (SGE = 72.5% vs FMS+SGE = 98.3%). However, given the underlying complexity of binding energy landscapes, ligand poses with distinctly different binding geometries may in fact yield similar energy scores (and vice versa), thus the observed SGE overlap in Figure 2-8 (top panel) is not unreasonable.



**Figure 2-8.** SGE (top) and FMS (bottom) score histograms using ensembles derived from SGE (blue), FMS (red), or FMS+SGE (green) driven sampling methods.

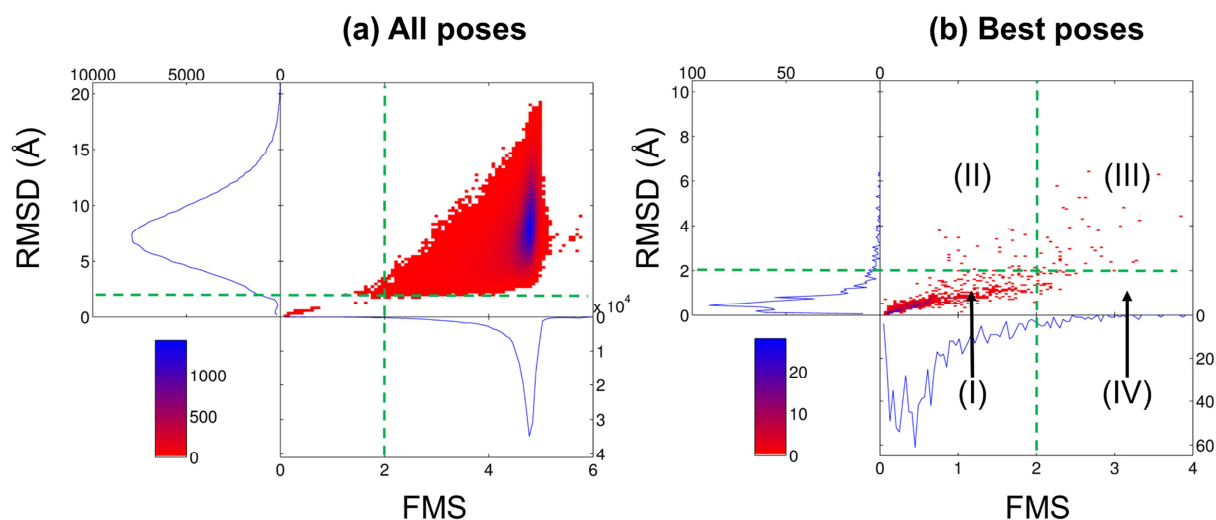
As shown in Figure 2-8 (bottom), FMS score distributions show much greater separation, indicating greater sensitivity, in contrast to the SGE score distributions shown in Figure 2-8 (top). Here, SGE sampled poses yield a much wider almost uniformly-distributed range of FMS scores (Figure 8 bottom, blue) compared to FMS (Figure 2-8 bottom, red) or FMS+SGE (Figure 2-8 bottom, green) sampled poses which have large peaks around 0.5, indicative of high pharmacophore overlap. Importantly, the FMS+SGE combination containing both geometric and energetic components to guide growth yields energy scores on par with standard SGE-guided docking poses (Figure 2-8 top, green vs. blue) and matches the pharmacophore models even better than FMS-only docking (Figure 2-8 bottom, green vs. red).

**Ensemble sizes.** An additional interesting observation from the results in Table 2-4 is the larger number of final docked poses obtained using FMS (337,674) compared to SGE (89,083)

or FMS+SGE (59,237). The much larger ensemble generated with FMS corresponds to an increase in total docking time, which could be of concern, although when normalized by the number of poses kept, the FMS function is actually faster than SGE by about 1.5 fold. The most likely explanation for the increased size involves reduced pruning. Current experiments employed a standard DOCK input file specifying a maximum score cutoff of 100.0, larger than the upper bound of the FMS function [0, 20]. Thus, poses are not as vigorously pruned during growth compared to protocols that employ energy-based functions (themselves not bounded). The significantly larger ensemble from FMS-sampling also likely contributes to the reduction in docking success rate associated with SGE rescoring because of the greater number of alternative (decoy) poses associated with system. Future studies to optimize the maximum score cutoff parameter would be worthwhile.

#### **2.4.2 Quadrant Partitioning using FMS Score.**

Although no score cutoffs were used to define success in the pose reproduction tests in Table 2-4, if both a RMSD cutoff and score cutoff are defined then the results can be classified in one of four different quadrants (see Figure 2-9b) defined as: (I) True Positive (TP), good score and low RMSD; (II) False Positive (FP), good score and high RMSD; (III) True Negative (TN), bad score and high RMSD; (IV) False Negative (FN), bad score and low RMSD. To highlight properties of the new DOCK pharmacophore function, Figure 2-9 focuses on the results derived using only the FMS-guided sampling protocol discussed above (success = 93.5%, sampling failure = 3.7%, scoring failure = 2.8%). Dashed green lines at RMSD=2 Å and FMS=2 delineate the four quadrants.



**Figure 2-9.** 2D histograms of FMS score and RMSD for (a) all poses (N=239,486) and (b) best scored poses (N=1,041) generated using FMS guided sampling of 1043 systems. Poses without matches (FMS=20) not included in histograms. Color reflects density (population).

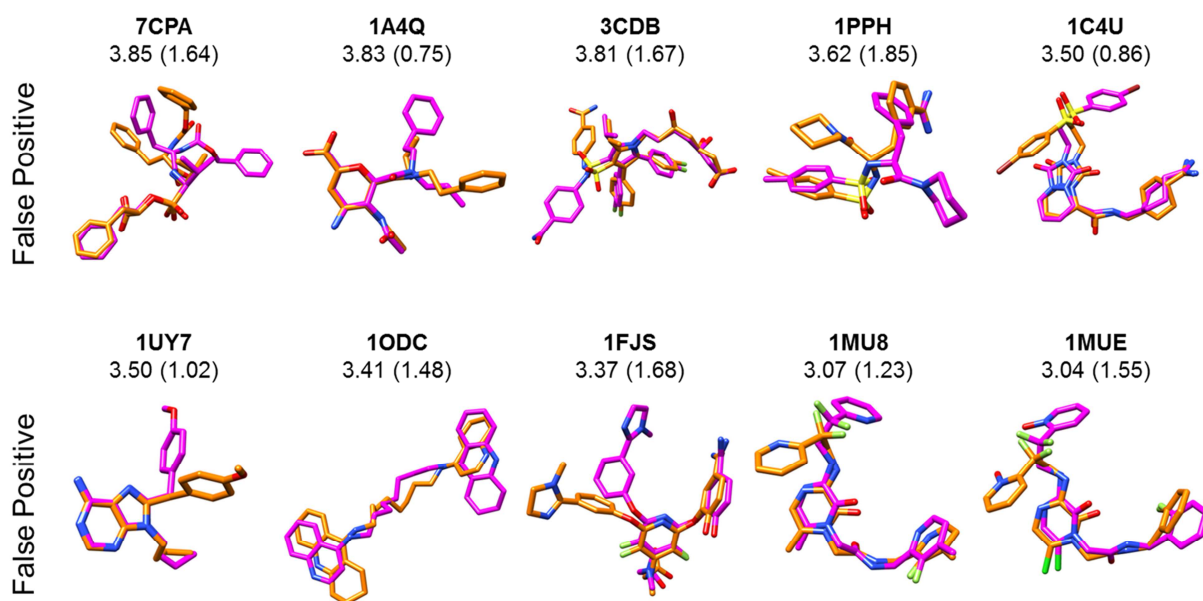
Figure 2-9a plots the large "all poses" set consisting of 239,486 ligand conformations with  $FMS < 20$  out of the total sampled space obtained with FMS sampling (337,674 poses). Here, the small separate cluster of points located in the TP region (lower left quadrant), which shows roughly linear correlation with RMSD, corresponds to mostly docking successes compared to the highly populated TN region (upper right quadrant) containing many thousands of points for which the correlation between RMSD and FMS begins to diverge as FMS values increase. Unlike the standard SGE function, which typically shows little correlation with RMSD, the FMS method behaves more like RMSD given the geometric nature of the function. Importantly, the results in Figure 2-9a indicate that the FMS protocol is not only able to identify close-to-native ligand conformation with favorable scores (region I), but also correctly characterize poses that are geometrically different from the reference by assigning unfavorable scores (region III).



Figure 2-9b plots the "best poses" set consisting of 1041 ligand conformations (1 system failed to dock, 1 system with FMS=20 for the best score pose). As in Figure 2-9a, poses in the TP region again show roughly linear correlation with RMSD. In this case however, as only a single pose for each systems is retained, unlike the "all poses" case, the TN region is sparse. Ideally, a good function should maximize TP and minimize FP. With the present RMSD (2.0 Å) and FMS (2) cutoffs, 949 points are classified as TP and 26 are classified as FN. The remaining points (1042-975 = 67) are divided into 39 TN cases and 28 FP cases. Overall, the 97.3% TP rate (949/975) and 41.8% FP rate (28/67) indicates good quadrant partitioning. And, as expected, use of a smaller score cutoff will yield a reduction in TP but an improvement in FP. For example, use of an FMS cutoff = 1.5 yields TP rate = 91.4% and FP rate = 20.9%, and use of an FMS cutoff =1.0 yields TP rate = 78.1% and FP rate = 9.8%. As a point of comparison, comparable analysis by Balius *et al*<sup>71</sup> for a similar TP rate = 79.8% yielded a higher FP rate = 46.2% using DOCK's footprint similarity method with a 0.6 score cutoff (based on normalized Euclidean distance) and 2 Å RMSD cutoff across 780 protein-ligand systems. In practice, the optimal choice of a numerical value for score cutoff to employ in a study to yield compounds with the desired properties is system dependent. For example, in typical virtual screening applications, FMS score between candidate compounds and a reference would be expected to be higher (i.e. less overlap) than under the present pose reduction tests which compare compounds with identical topologies but different conformations.

***False Positive (FP) cases with FMS.*** While FMS in general yields excellent quadrant partitioning, an examination of the results was undertaken to determine the underlying cause of FP and FN classifications. Focusing on results from the "best poses" set (Figure 2-9b), Figure 2-10 presents the ten out of twenty-eight FP results (RMSD > 2 Å, FMS <= 2) with the highest

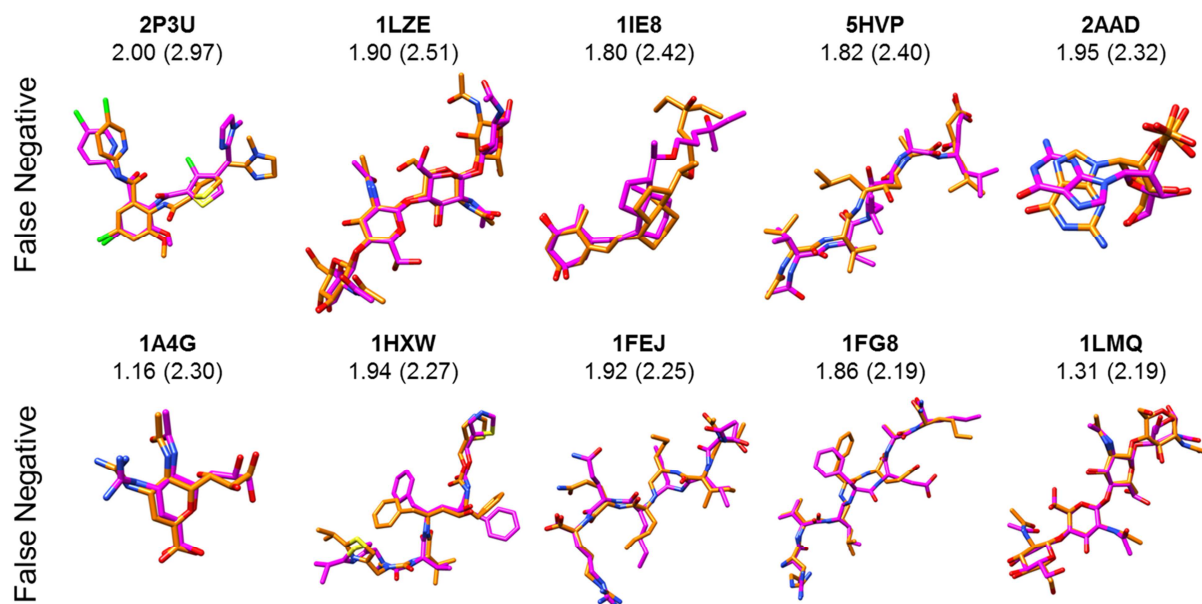
RMSD. Analogous to that observed with the FMS+SGE scoring failures (Figure 2-7), FP poses derived with FMS-sampling show, for the most part, remarkably high overlap except for one end of the molecule. And in all ten cases, the poorly-overlapped groups contain rings, which are weighted heavier by the RMSD function than FMS. System 1ODC is a particularly interesting case. Here, the ligand pose is semi-symmetric and flipped by ca. 180° relative to the reference (magenta vs. orange) resulting in overlap between two rings on one end with three rings from the other. Although the Hungarian algorithm used here in DOCK<sup>110</sup> to compute symmetry-corrected RMSD effectively accounts for the swap of functionality having identical chemical properties, the resultant value of 3.41 Å is still classified as a failure, largely as a result of one ring on either end (8 atoms total) not being matched. In contrast, the FMS score not only accounts for the symmetry but the good overlap between four out of six ring centers (and associated vector directions), which leads to a relatively low FMS score of 1.48. Overall, visual examination of these ten worst FP cases reveals a significant amount of physically-reasonable matches and minimal mismatch and the classification of poses to this quadrant is, in most cases, understandable.



**Figure 2-10.** Ten out of twenty-eight FP poses derived from FMS-guided docking with the largest RMSD values. Crystal poses in orange, best scored poses in magenta. RMSD in Å, and FMS scores in parentheses.

**False Negative (FN) cases with FMS.** In terms of the FN examples (RMSD < 2 Å, FMS > 2), Figure 2-11 presents the ten out of twenty-six poses with the highest FMS scores. Immediately obvious compared to the FP examples, is that the molecules here contain fewer aromatic rings, for the most part are larger and more extended, and have a higher number of more loosely matched hydrogen-bonding functional groups (most polar atoms in the FP cases are either tightly matched or not matched at all). This latter point is particularly important as relatively small changes in position of a hydrogen bonding functional group can lead to relatively large changes in FMS overlap but minor effects on RMSD which is computed using only heavy (non-hydrogen) atoms. Although our standard preparation protocol for FMS scoring employs an energy minimization step to relax any hydrogen atoms added to the system, the positions adopted

as a result of ligand sampling during growth may result in the candidate and reference poses having different hydrogen directions. This result highlights the need for care when preparing a molecule to be used as a "reference" for scoring candidate compounds. Despite being a distinctly different type of function, a similar conclusion was reached by employing the DOCK footprint function.<sup>71</sup> Despite this sensitivity, however, most of the FN cases have scores close to 2 that could easily be rescued by a minor increase in FMS cutoff to 2.5.

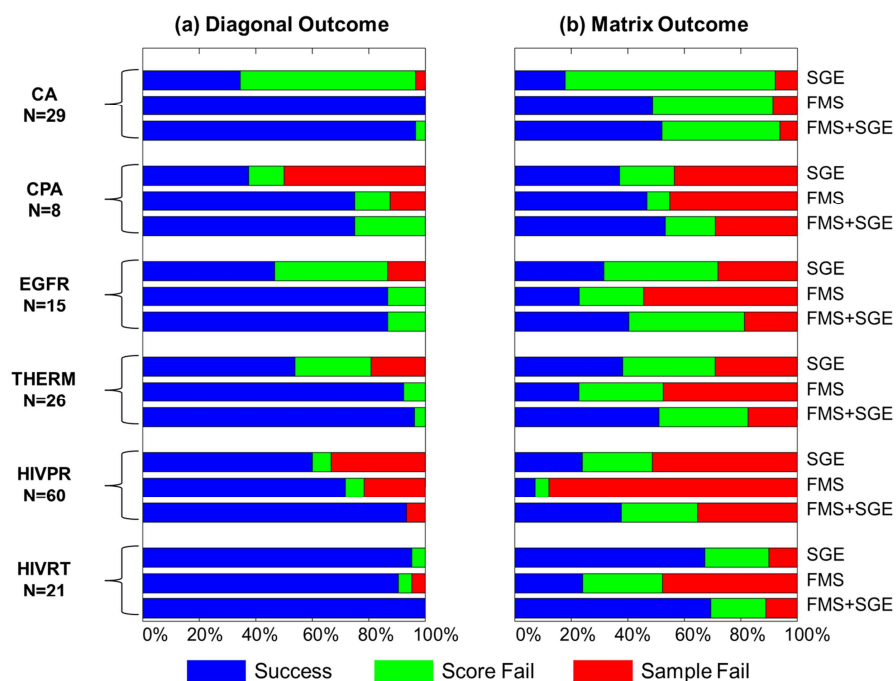


**Figure 2-11.** Ten out of twenty-six FN poses derived from FMS-guided docking with the largest FMS scores. Crystal poses in orange, best scored poses in magenta. RMSD in Å and FMS scores in parentheses.

### 2.4.3 Crossdocking Results.

In addition to pose reproduction, crossdocking experiments are a useful way to determine if different protocols can reproduce native-like poses when ligands are docked to highly homologous protein binding sites from different crystallographic structures (see Figure 2-6b).

Figure 2-12 displays outcomes across six protein families: carbonic anhydrase (CA,  $N=29$ ), carboxypeptidase A (CPA,  $N=8$ ), epidermal growth factor receptor (EGFR,  $N=15$ ), thermolysin (THERM,  $N=26$ ), HIV protease (HIVPR,  $N=60$ ), and HIV reverse transcriptase (HIVRT,  $N=21$ ). For comparison, both the diagonal (cognate protein-ligand pairs) and the entire matrix (all combinations) are shown. As before, three docking protocols were tested (SGE, FMS, and FMS+SGE). As shown in Figure 2-12a, this is a particularly challenging group of proteins with the standard SGE protocol yielding low diagonal successes (34.5-60.0%) for 5 out of 6 families. The exception is HIVRT for which the SGE success rate = 95.2%. In contrast, use of FMS (71.7-100.0%) or FMS+SGE (75.0-100.0%) yields significant improvement for cognate receptor-ligand pairs. Carbonic anhydrase is a particularly noteworthy example as the SGE diagonal success increases from only 34.5% to near 100.0% using the FMS or FMS+SGE functions. Comparable enhancements in success for carbonic anhydrase were also reported by Balius *et al*<sup>71</sup> when using the DOCK footprint similarity scoring function (82.8%) compared to SGE (31.0%).

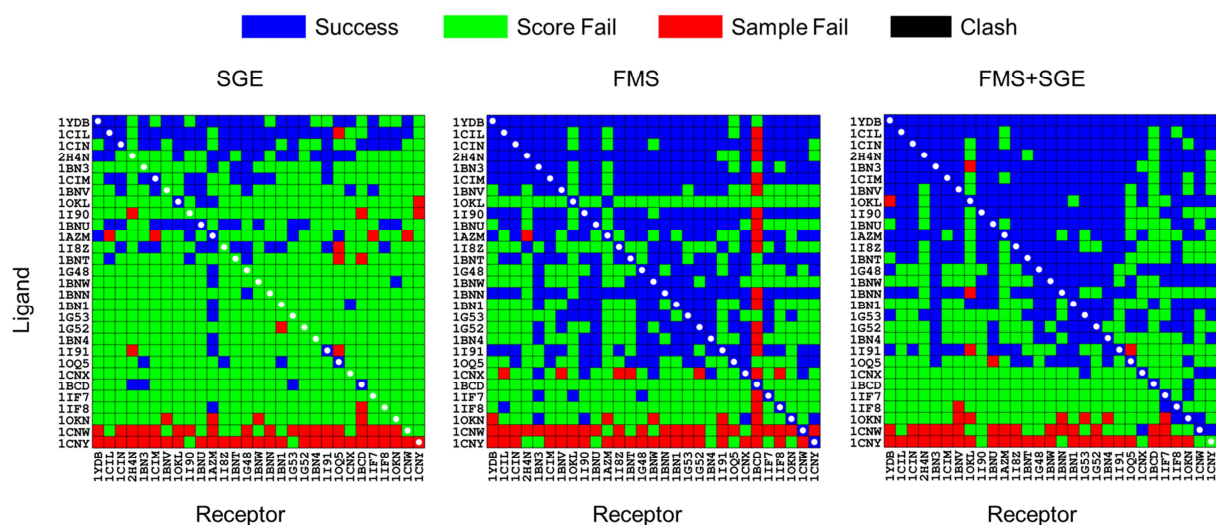


**Figure 2-12.** Crossdocking outcomes averaged across the diagonal (left) or total matrix (right) for six protein families: carbonic anhydrase (CA), carboxy peptidase A (CPA), epidermal growth factor receptor (EGFR), thermolysin (THERM), HIV protease (HIVPR), and HIV reverse transcriptase (HIVRT) using SGE (top), FMS (middle), and FMS+SGE (bottom) protocols. Success in blue, scoring failure in green, sampling failure in red.

As expected, for more challenging crossdocking experiments, matrix success (Figure 2-12b) using any of the scoring functions are in general significantly lower than their diagonal counterparts (Figure 2-12a). As a baseline, use of SGE yields an averaged matrix success of 36.0% compared to the diagonal at 54.6%. In contrast to the diagonal results, interestingly, use of FMS alone for crossdocking shows improvement over SGE in only two cases (CA and CPA). However, in all cases, the combined FMS+SGE function always yields a better matrix success than does SGE. Analogous to the diagonal results, the matrix outcomes (Figure 2-12b) similarly reveal carbonic anhydrase has the lowest overall matrix SGE success rate (17.8%) which increased the most among all systems tested when using FMS (48.8%) or FMS+SGE (52.1%).

Figure 2-13 compares the heatmaps for carbonic anhydrase, derived from three independent docking sets of size  $29 \times 29 = 841$  combinations, using SGE, FMS, and FMS+SGE methods. The maps visually highlight that SGE failures are primarily due to scoring (green squares), pinpoint which specific systems are involved, and indicate that FMS and FMS+SGE protocols significantly improve docking outcomes (more blue squares).

Additionally visible in the FMS heatmap for carbonic anhydrase (Figure 2-13, middle), is the appearance of previously unseen sampling failures specifically localized to column 1BCD. It is important to note that the RMSD calculations in both diagonal and off-diagonal experiments always involve compounds of the same topology. However, for pharmacophore overlap calculations involving off-diagonal elements the pharmacophore reference and the candidate molecule being docked are usually of different topology. In such cases, FMS-guided docking may drive sampling in a direction that will not necessarily agree with the RMSD reference. Calculation of the pharmacophore overlap between all aligned crystallographic references for carbonic anhydrase indeed shows 1BCD has the poorest reference FMS scores (between the pharmacophore reference and the RMSD reference) when averaged across all columns (FMS=5.15) or all rows (FMS=5.51) which is appreciably above the overall average (FMS=3.38) across all reference pairs. Inspection further revealed that the ligand from 1BCD has only one rotatable bond and a molecular weight of 148.1 g/mol, which is markedly smaller than the average ligand in this family with 5.1 rotatable bonds and molecular weight of 339.7 g/mol. Thus, crossdocking of ligands to receptor 1BCD, using FMS alone, is not expected to be consistent with the 1BCD reference sampling space, which leads to the observed sampling failures.

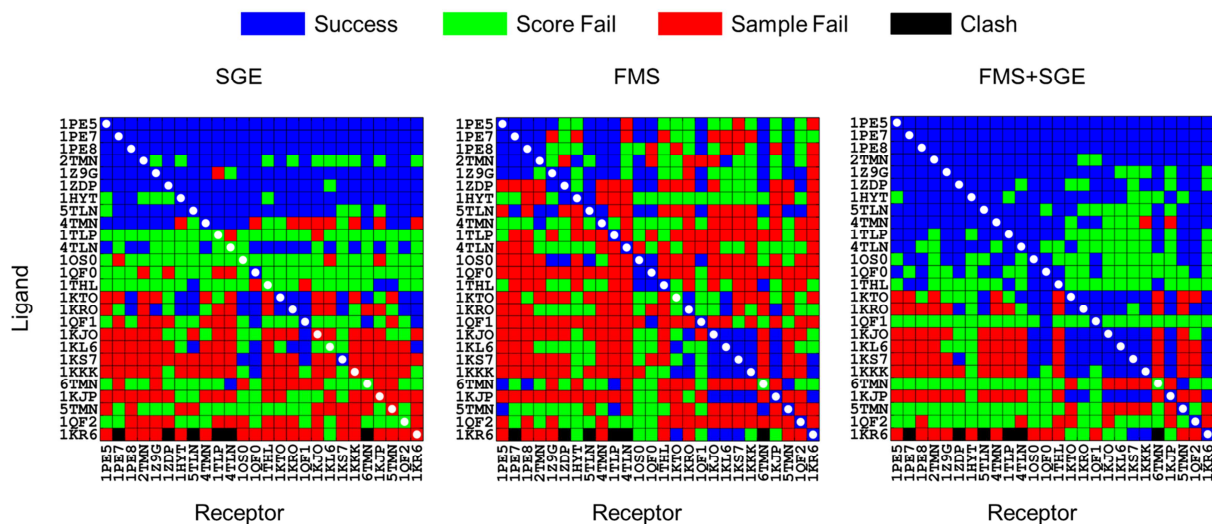


**Figure 2-13.** Crossdocking heatmaps using SGE, FMS and FMS+SGE protocols for carbonic anhydrase (29x29=841 combinations).

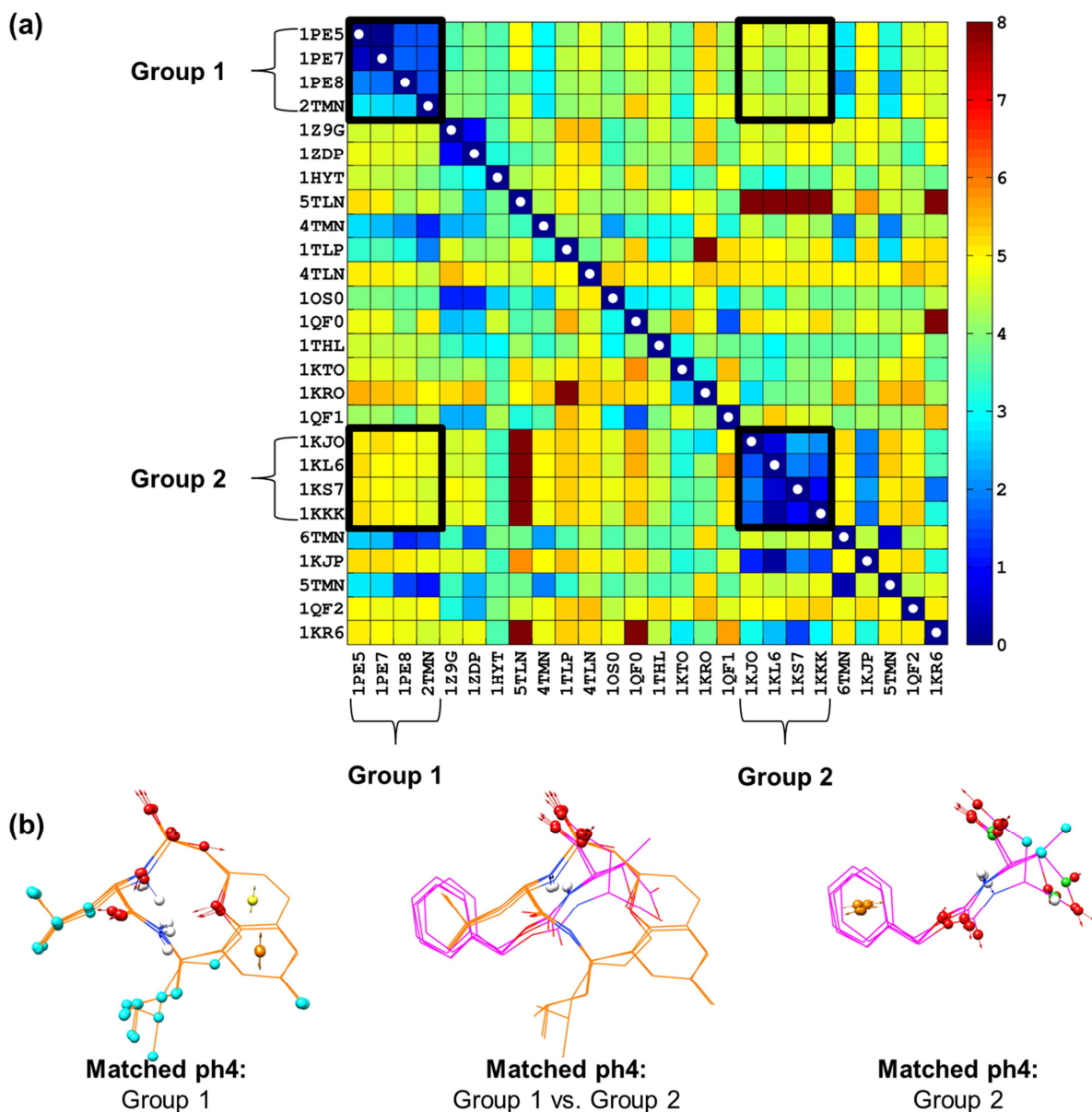
Additionally, more dramatic examples of this phenomenon manifest themselves in the heatmaps for thermolysin as shown in Figure 2-14. Here, in contrast to carbonic anhydrase, crossdocking with SGE yields a higher overall success rate of 38.2% (Figure 2-14 left, blue) but with a higher percentage of sampling failures (29.1%, red). And, while the combined function FMS+SGE yields the overall best docking success rate (51.0%) for this family, use of FMS alone actually increases sampling failures (47.5%) relative to SGE (Figure 2-14 left vs middle, red) which, as described below, likely involves poor reference pharmacophore overlap. Close inspection of the crossdocking heatmaps reveals sub-matrices of size 4x4, defined here as group 1 (1PE5, 1PE7, 1PE8, 2TMN) and group 2 (1KJO, 1KL6, 1KS7, 1KKK), for which FMS sampling relative to SGE: (i) maintains docking success and/or (ii) rescues previously unsuccessful docking outcomes involving systems within the same group, and (iii) introduces docking failures for systems from different groups. To aid the discussion, Figure 2-15 shows a



heatmap of FMS scores (as opposed to docking outcomes), derived from the x-ray references, with diagonal and off-diagonal sub-matrix blocks for groups 1 and 2 outlined as black boxes.



**Figure 2-14.** Crossdocking heatmaps using SGE, FMS and FMS+SGE protocols for thermolysin (26x26=676 combinations).



**Figure 2-15.** (a) FMS heatmap, using all crystallographic reference poses for thermolysin, with perfect overlap in dark blue (FMS=0) and poorest overlap (FMS $\geq$ 8) in dark red. Group 1 sub-matrix defined by systems 1PE5, 1PE7, 1PE8, and 2TMN. Group 2 sub-matrix defined by systems 1KJO, 1KL6, 1KS7, and 1KKK. (b) Crystallographic reference overlays showing matched pharmacophore features for group 1 (left, orange), group 2 (right, magenta), and group 1 vs. group2 (middle).

The FMS scores computed between all reference pairs indicate perfect overlap on the diagonal (FMS=0, dark blue) but for the most part the majority of pairs have poor overlap

(FMS=3-8, green to dark red). A striking exception are the cases defined by group 1 and 2 (Figure 2-15, black boxes) which all have relatively good reference FMS scores within the same group (two blue sub-matrices near the diagonal) but poor FMS scores between different groups (two green to yellow sub-matrices on the off-diagonal). This observation help explains why FMS-guided docking yields 100% success across the sub-matrices formed within the same group (Figure 2-14 middle), but when using group 1 systems as a reference to guide docking of ligands in group 2, no matrix success is reported and only 1 success is obtained for the opposite case (other symmetric block). Structurally, the molecular cluster formed by ligands in group 1, occupies an extended space in the thermolysin binding pocket (Figure 2-15b left) and contain additional hydrophobic groups compared to group 2 (Figure 2-15b right). Group 2 ligands cluster into a more slender volume anchored by an aromatic ring at one end and hydrogen bond acceptor on the other. As a consequence, groups 1 and 2 share only a few (1-3) matched pharmacophore points (Figure 2-15b middle) which explains the poor FMS scores between off-diagonal reference ligands in addition to the poor docking outcomes. Interestingly, the addition of the energy term to the pharmacophore overlaps score (FMS+SGE score), using group 2 as a reference to dock group 1, yields 100% docking success. In contrast, using group 1 as reference to dock group 2 yields 100% sampling failure (Figure 2-14, right panel).

Finally, the overall poorest matrix success results using FMS (7.2%) or FMS+SGE (37.7%) docking is seen with HIVPR. Although high ligand flexibility is expected to play a role in the large number of sampling failures seen in the FMS matrix (Figure 2-12b red) relative to other systems (31/60 of ligands have  $\geq 15$  rotatable bonds), the most likely cause is poor pharmacophore overlap between all pairwise combinations. Consistent with the discussions above, out of the 3600 pairwise combinations in the HIVPR crossdocking reference FMS matrix

derived from crystallographic poses, only 220 pairs yielded reasonable pharmacophore overlap ( $FMS \leq 3$ ). In contrast, 2493 pairs have poor pharmacophore overlap ( $FMS \geq 4.5$ ) which, interestingly in this case, is about the same as the number of sampling failures (2816).

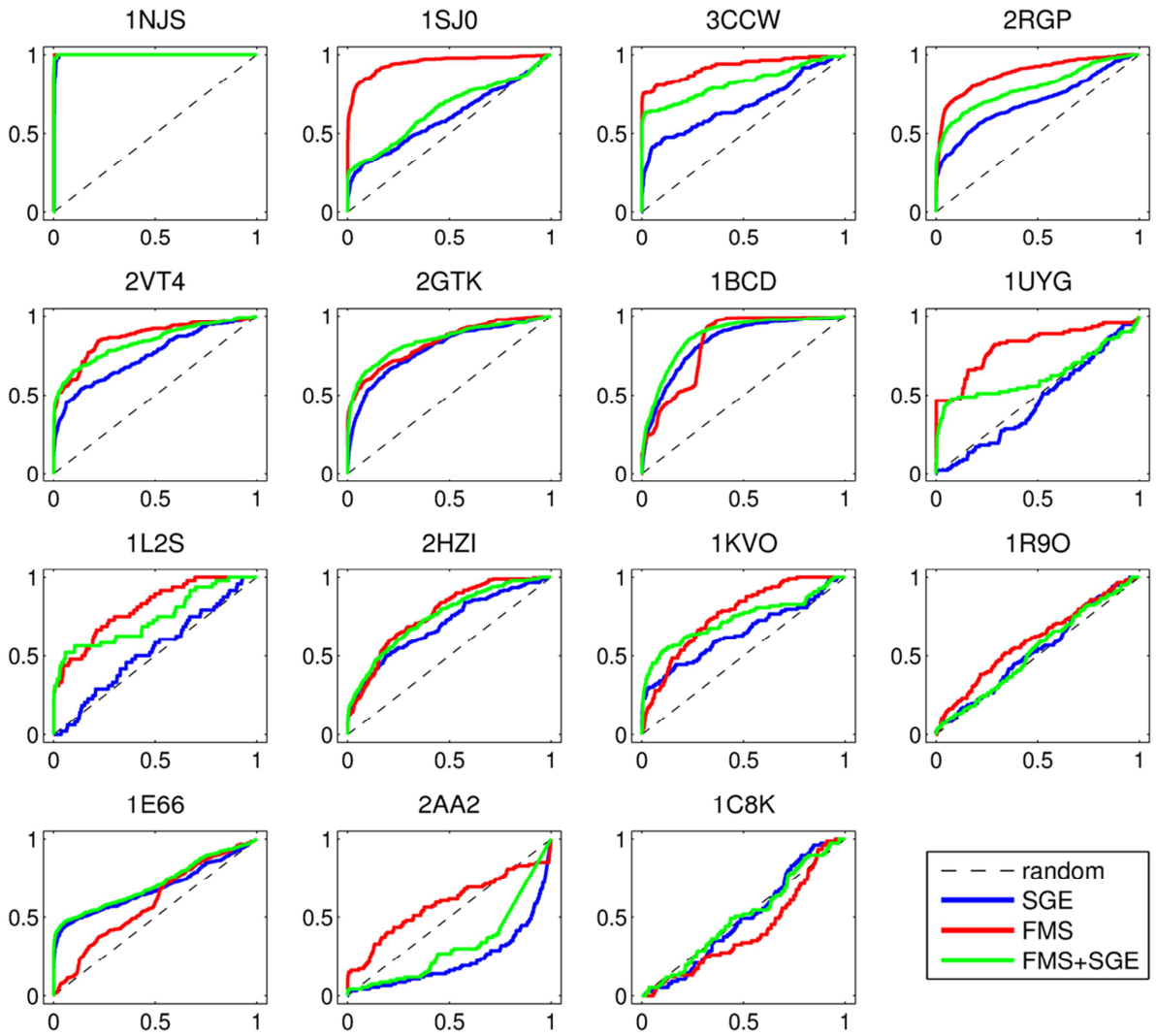
Overall, two key points have emerged from the current crossdocking studies: (1) FMS-guided success rates, particular for off-diagonal elements, are dependent on the similarity between the pharmacophore reference and the RMSD reference. (2) The FMS+SGE protocol generally improves crossdocking performance, relative to SGE or FMS, by integrating known binding profiles into the standard DOCK energy score.

#### **2.4.4 Enrichment Results.**

Results for enrichment experiments, used to gauge how DOCK would perform in a virtual screening using SGE, FMS, or FMS+SGE protocols are shown in Figure 2-16 and Table 2-5. Receiver operating characteristic (ROC) curves and area under the curve (AUC) analyses were used to compute fold enrichment ( $FE = AUC_{curve} / AUC_{random}$ ) values for docking active and decoy ligands taken from the DUD-E database.<sup>101</sup> For virtual screening applications, good early enrichment is considered to be critically important thus FE was also computed at 0.1%, 1%, and 10% of the ranked database. For the current tests, the overall shape of the ROC curves vary from essentially perfect enrichment (1NJS) to random enrichment (1C8K) with most systems exhibiting good overall enrichment but with a visible dependence on which of the three docking functions was used. For the majority of systems, depending on which ROC region is examined, FMS (red curves) shows higher enrichment than SGE (blue curves), with FMS+SGE (green curves) being roughly in between (Figure 2-16). Across different ranges of the database, based on numerical AUC values, use of FMS or FMS+SGE consistently yield higher FE rates relative

to SGE (Table 2-5). For example, at 0.1% of the database, 11/15 FE values using FMS and 11/15 FE values using FMS+SGE are enhanced relative to SGE (Table 2-5 column A). Similarly, at 1% of the database, 10/15 FE values using FMS and 13/15 FE values using FMS+SGE are enhanced relative to SGE (Table 2-5 column B). Comparable results are obtained at 10% and 100% of the database.

The fact that use of FMS+SGE yields generally lower enrichment outcomes than FMS is somewhat surprising given that FMS+SGE yielded higher success rates than FMS in pose reproduction experiments. However, it is important to note that the role of the SGE term in FMS+SGE is fundamentally different for pose reproduction given that different molecular conformers, as opposed to the different chemical species for enrichment, are what is rank-ordered. The most likely contributing factor as to why FMS scoring yields enhanced enrichment involves the fact that use of a crystallographic reference captures elements of what is important for activity for at least one active ligand. Because rank-ordering of "actives" using FMS scoring are biased towards the known binder, higher enrichments can be obtained. With the addition of the SGE term, sampling and rank-ordering using FMS+SGE will change as a result of, for example MW bias, which leads to different enrichment results (less-favorable in most cases for the present tests). Overall, the enrichment tests validate the ability of FMS and FMS+SGE protocols to enrich for true actives relative to SGE alone by prioritizing molecules with similar binding profiles as a known ligand. This strongly suggests use of a pharmacophore reference to help guide virtual screening is a viable alternative to the standard DOCK protocol.



**Figure 2-16.** ROC enrichment curves for 15 DUD-E systems using SGE, FMS and FMS+SGE protocol.

**Table 2-5.** Fold enrichment (FE) results at different percentages of the database (DB) screened.

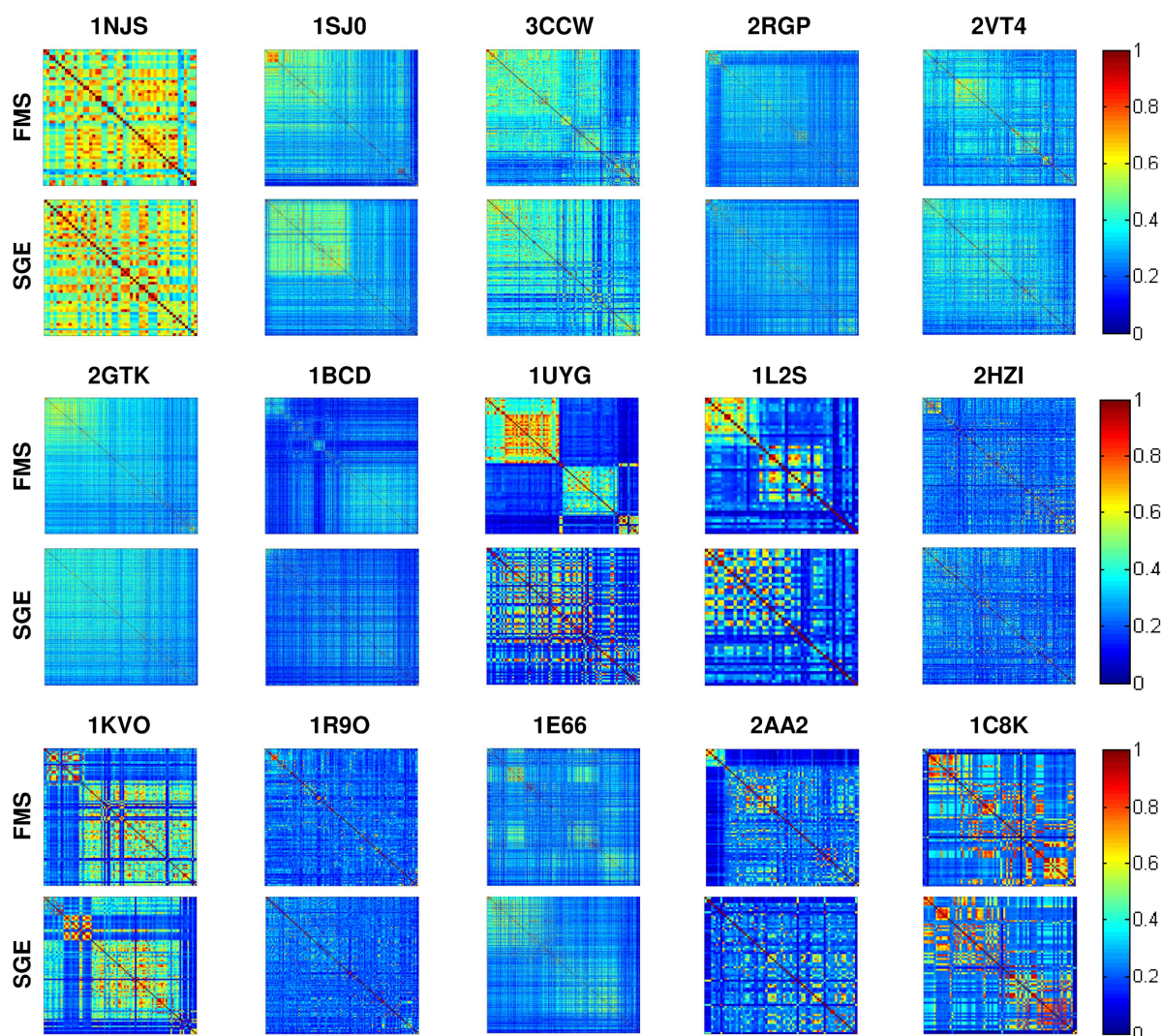
		(A) FE @ 0.1% of DB <sup>b</sup>	(B) FE @ 1% of DB <sup>b</sup>	(C) FE @ 10% of DB <sup>b</sup>	(D) FE @ 100% of DB <sup>b</sup>
<b>System<sup>a</sup></b>	<b>Random</b>	1.00	1.00	1.00	1.00
	<b>Maximum</b>	2000.00	200.00	20.00	2.00
<b>1NJS</b>	<b>SGE</b>	0.00	111.36	19.05	1.99
	<b>FMS</b>	1009.65	184.26	19.88	2.00
	<b>FMS+SGE</b>	0.00	150.85	19.52	2.00
<b>1SJ0</b>	<b>SGE</b>	88.96	22.09	4.98	1.21
	<b>FMS</b>	804.91	114.91	15.39	1.90
	<b>FMS+SGE</b>	382.87	48.69	5.90	1.30
<b>3CCW</b>	<b>SGE</b>	80.74	31.49	7.20	1.43
	<b>FMS</b>	1167.99	144.49	15.46	1.77
	<b>FMS+SGE</b>	932.51	116.07	12.68	1.61
<b>2RGP</b>	<b>SGE</b>	218.33	41.28	6.76	1.40
	<b>FMS</b>	225.70	46.36	11.77	1.77
	<b>FMS+SGE</b>	517.05	67.51	9.90	1.59
<b>2VT4</b>	<b>SGE</b>	166.11	36.47	7.51	1.50
	<b>FMS</b>	223.36	65.93	10.45	1.73
	<b>FMS+SGE</b>	376.70	78.04	11.07	1.67
<b>2GTK</b>	<b>SGE</b>	53.17	22.27	7.21	1.59
	<b>FMS</b>	613.72	75.99	10.30	1.67
	<b>FMS+SGE</b>	319.87	60.89	10.75	1.69
<b>1BCD</b>	<b>SGE</b>	6.40	14.38	6.35	1.67
	<b>FMS</b>	147.94	25.25	5.43	1.65
	<b>FMS+SGE</b>	43.48	25.22	7.74	1.74
<b>1UYG</b>	<b>SGE</b>	0.00	3.56	0.65	0.92
	<b>FMS</b>	590.68	90.43	9.32	1.62
	<b>FMS+SGE</b>	75.01	35.68	7.90	1.25
<b>1L2S</b>	<b>SGE</b>	0.00	0.00	0.54	1.07
	<b>FMS</b>	264.83	55.17	7.82	1.61
	<b>FMS+SGE</b>	235.40	56.64	8.84	1.48
<b>2HZI</b>	<b>SGE</b>	86.92	23.91	5.03	1.40
	<b>FMS</b>	149.30	23.72	4.43	1.53
	<b>FMS+SGE</b>	103.28	28.80	5.51	1.50
<b>1KVO</b>	<b>SGE</b>	62.81	35.80	5.84	1.29
	<b>FMS</b>	39.26	6.16	4.04	1.53
	<b>FMS+SGE</b>	51.04	32.23	8.04	1.47
<b>1R9O</b>	<b>SGE</b>	31.34	4.59	1.53	1.07
	<b>FMS</b>	0.00	2.17	2.45	1.20
	<b>FMS+SGE</b>	15.67	5.33	1.63	1.07
<b>1E66</b>	<b>SGE</b>	247.06	52.18	8.20	1.37
	<b>FMS</b>	11.28	3.43	1.56	1.19
	<b>FMS+SGE</b>	508.10	70.25	8.99	1.43
<b>2AA2</b>	<b>SGE</b>	4.13	5.00	0.74	0.45
	<b>FMS</b>	190.19	26.71	3.32	1.17
	<b>FMS+SGE</b>	62.02	7.07	0.90	0.64
<b>1C8K</b>	<b>SGE</b>	0.00	0.00	0.63	0.98
	<b>FMS</b>	0.00	0.00	0.47	0.83
	<b>FMS+SGE</b>	0.00	0.00	0.82	1.00

<sup>a</sup>PDB codes used with accompanying DUD-E libraries (actives + decoys). <sup>b</sup>FE=  $AUC_{curve}/AUC_{random}$  thus baseline random selection always yields a FE = 1.00.

As an additional point, in general, good enrichment should depend only on actives being ranked earlier than decoys without regards to there being "similarity" among groups of compounds. However, use of the FMS function might be expected to yield higher early similarity, compared to the entire set of actives as a whole, provided the composition of active molecules in a given database does contain subsets with 2D similarity and a larger than average number of docked compounds yield good 3D overlap with the reference pharmacophore. To explore this issue, among rank-ordered active compounds, we computed all possible pairwise Tanimoto coefficients<sup>115</sup> using the DOCK fingerprinting method motivated by the MOLPRINT algorithm<sup>116,117</sup> and plotted the data as heatmaps (Figure 2-17).

While additional studies should be pursued, especially those employing more than one reference per system as was done in the current study, Figure 2-17 reveals that in a number of cases, active molecules do in fact appear to have higher similarity earlier in rank-ordered list when using FMS vs. SGE scoring (Figure 2-17 red/yellow vs. blue, top vs. bottom rows). Rank-ordering with FMS also shows a tendency to cluster similar molecules together. Particularly interesting examples include 1SJO, 1UYG, and 1L2S for which SGE shows poor (random in 2 cases) enrichment compared to FMS as gauged by the shape of the ROC curves in Figure 2-16.



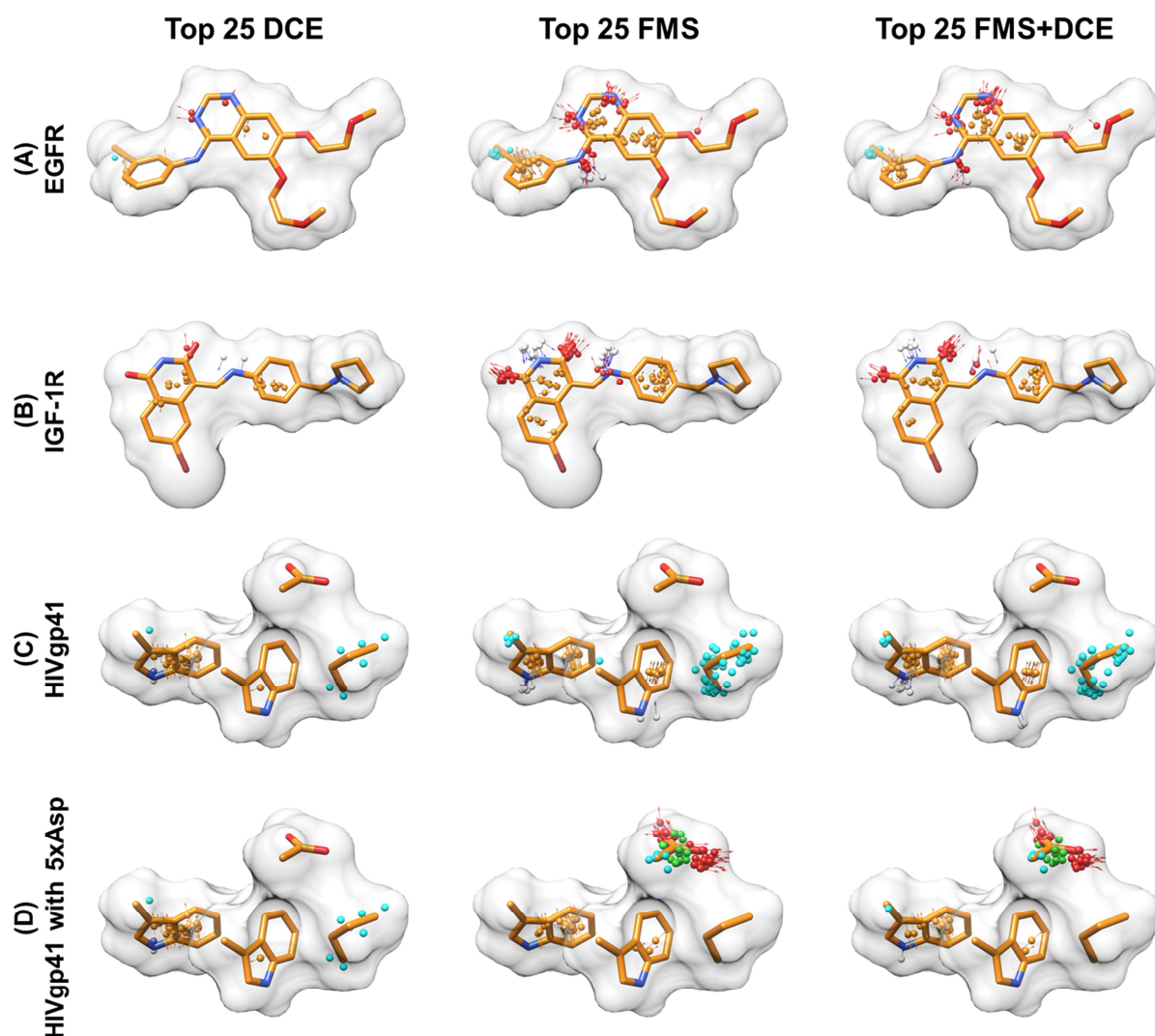


**Figure 2-17.** Pairwise Tanimoto heatmap for 15 DUD-E systems using FMS (top) and SGE (bottom) protocol. The color scheme in the heatmap represents the magnitude of Tanimoto similarity and the x/y axis represents the rank-ordered list (FMS or SGE) of unique active molecules for each system.

#### 2.4.5 Case Studies Targeting EGFR, IGF-1R, and HIVgp41.

To further gauge the utility of using FMS methods, we rescored virtual screening results for three systems being targeted in our laboratory: epidermal growth factor (EGFR),<sup>34,118</sup> insulin-like growth factor 1 receptor (IGF-1R), and human immunodeficiency virus glycoprotein 41 (HIVgp41)<sup>72,119</sup> and visually examined the number of pharmacophore matches for top-ranked

molecules under different conditions (Figure 2-18). The FMS references employed for EGFR (erlotinib) and IGF-1R (isoquinolinedione analog) were based on known small molecule inhibitors, while the HIVgp41 reference was based on four key amino acid sidechains (WWDI) from a known peptide inhibitor. The receptors and accompanying references were derived from crystallographic structures (PDB codes 1M17, 2ZM3, and 1AIK, respectively), and the molecules docked to each target were taken from the publically available ZINC<sup>39</sup> collection of purchasable organic compounds. For each screen, the top 100,000 ranked compounds obtained with the standard docking protocol (grid score with FLX protocol) were retained and then rescored and re-ranked using DOCK Cartesian energy (DCE, which is comparable to SGE but in Cartesian space), FMS, and FMS+DCE scoring protocols.



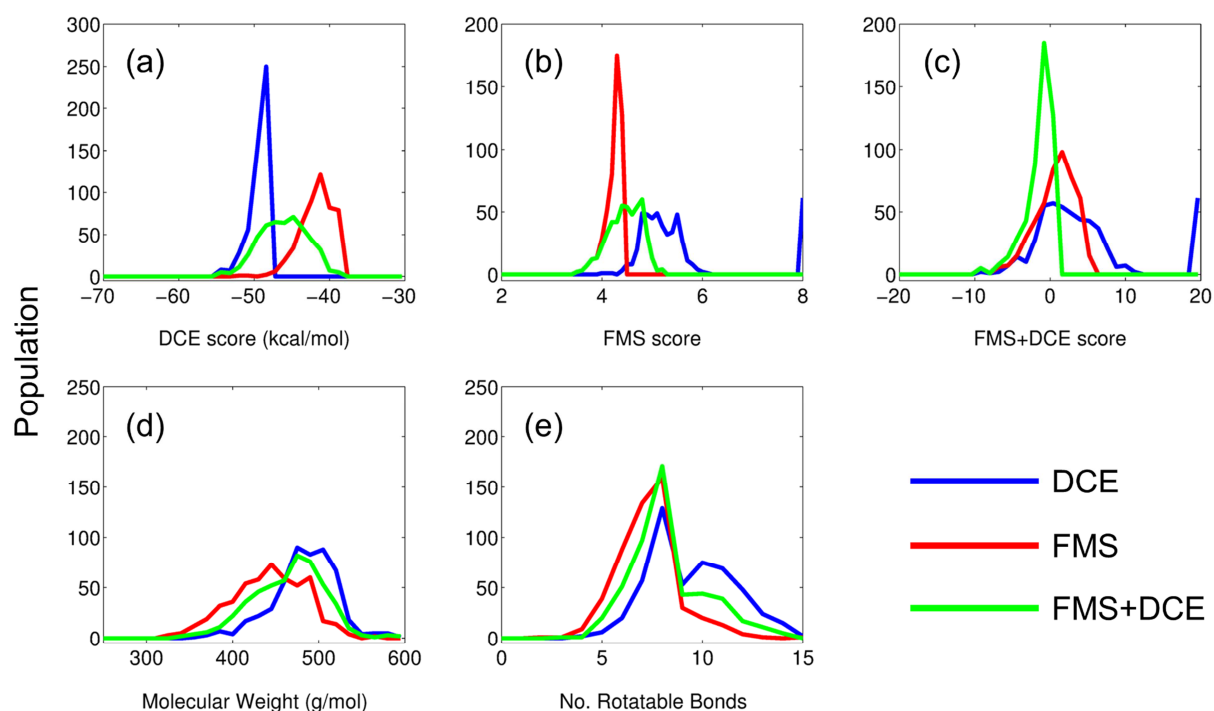
**Figure 2-18.** References (orange sticks, gray surface) used to rescore virtual screening results targeting (A) EGFR, (B) IGF-1R, (C) HIVgp41, and (D) HIVgp41 with Asp sidechain weighted 5 times. Matched pharmacophore features include: PHO in cyan; HBA (vertex and vector) in red; HBD vector in blue, hydrogen vertex in grey; ARO (vertex and vector) in orange; POS in magenta; NEG in green (see Theoretical Methods for definitions).

As shown in Figure 2-18, the number of pharmacophores matched for the top 25 ranked compounds is relatively small using DCE. In sharp contrast, use of FMS or FMS+DCE show, for example, many more matched HBD (blue arrows), HBA (red arrows), ARO (orange arrows), and PHO (cyan spheres) features. It is important to note that the plots in Figure 2-18 show how

many "matched" pharmacophores were obtained, *relative to the reference*, but candidate compounds can contain "unmatched features" that extend beyond the volume defined by the reference compound; the functional form of eq 2-1 does not necessarily penalize unmatched features *relatively to the candidate*. This behavior could be changed, for example, by including a simply penalty term based on the number of unmatched groups in the candidate however this was not explored in great detail. Other functional forms besides eq 2-1 could also be investigated. In any event, the number of matched and unmatched features, including types, for each docked pose, is printed to the DOCK output, which can be useful to determine whether particular characteristics have been satisfied.

As a specific example, an interesting result from the present analysis is a lack of matched pharmacophore features to the Asp carboxylate group in the HIVgp41 reference (Figure 2-18 row C). An examination of ranked poses higher up the FMS and FMS+DCE lists did indeed reveal compounds with overlap to the reference carboxylate but they were not ranked as well as compounds with multiple matches involving two Trp indole rings and a hydrophobic Ile (Figure 2-18 row C). Given the biological importance of the Asp group in this system, an effective small molecule mimic would reasonably be expected to contain a negatively charged or hydrogen bonding group at this position.<sup>120,121</sup> A straightforward way to enforce this requirement was devised by using a modified HIVgp41 reference that simply included 5 copies of the Asp carboxylate which had the effect of weighting this feature more heavily as shown in Figure 2-18 row D. For this particular test, weighting the Asp more highly had the desired effect but at the expense of losing hydrophobic matches to Ile (Figure 2-18 FMS and FMS+DCE, row C vs D). As a general point, this example demonstrates the ease with which specific pharmacophore features can be emphasized over others using the current DOCK infrastructure.

Finally, in terms of additional ligand properties, Figure 2-19 plots results from the HIVgp41 screen for different groups of top-ranked molecules (N=500) each obtained by one of the ranking protocols. Consistent with previous studies from our laboratory, use of DCE (or SGE) shows a bias towards larger molecules. In contrast, compounds ranked by FMS score are smaller in size as demonstrated by ligands with lower molecular weights (Figure 2-19d) and fewer numbers of rotatable bonds (Figure 2-19e). As anticipated, use of FMS+DCE yields molecular weights and numbers of rotatable bonds roughly in-between DCE and FMS. For scoring, use of DCE results in more favorable DCE energies (Figure 2-19a blue vs. red or green), FMS results in more favorable FMS scores (Figure 2-19b red vs. blue or green), and FMS+DCE results in more favorable FMS+DCE scores (Figure 2-19c green vs. blue or red). And, rescore molecules obtained with one function with another function leads to the expected results. For example, DCE score distributions for top ranked FMS+DCE molecules are in between that of DCE and FMS (Figure 2-19a green), FMS score distributions for top ranked FMS+DCE molecules are in between that of FMS and DCE (Figure 2-19b green), and FMS+DCE score distributions for top ranked FMS molecules are in between that of FMS+DCE and DCE (Figure 2-19c red). Importantly, use of the combined FMS+DCE function to rescore virtual screening results yield both favorable FMS scores and DOCK energies. This suggests use of a reference to rescore screening results could also be a viable way to identify compounds that make known interaction patterns, with favorable interaction energies, while reducing molecular weight bias.



**Figure 2-19.** Histograms of rescored results for the top 500 molecules selected from virtual screening targeting HIVgp41.

## 2.5 Conclusion

In conclusion, the primary goal of this study was to develop, implement, and thoroughly test a pharmacophore-based scoring function for the docking program DOCK. The resulting method, termed pharmacophore matching similarity (FMS) score, was validated using experiments that help gauge accuracy relative to the standard DOCK single energy grid (SGE) protocol, and the combination score FMS+SGE. Three groups of validation experiments were performed: (i) pose reproduction (Figures 3.7-11, Table 2-4), (ii) crossdocking (Figures 3.12-15), and (iii) enrichment (Figures 3.16, Table 2-5). Importantly, in terms of pose reproduction, use of FMS (93.5%) or FMS+SGE (98.3%) functions yielded significantly higher success rates than the standard SGE (72.5%) method when evaluated using 1043 systems in the SB2012 testset. The

nearly perfect success rate obtained with the combined FMS+SGE function, which biases sampling to match a reference while simultaneously including energetic constraints imposed by a binding site, is notable and strongly suggests the method will have applicability for structure-based drug design provided a "suitable" reference can be identified. Tests using FMS alone for pose reproduction showed relatively few ligand poses falling into false positive (FP) and false negative (FN) regions defined by quadrant partition using specific RMSD and FMS score cutoff criteria (Figure 2-9). Interestingly, visual examination of the worst FP cases (Figure 2-10) revealed, in most instances, that the candidate and reference poses were in fact well overlaid and that only one part of molecule was not well matched. Unlike the standard DOCK energy function, the geometry-based FMS scores show reasonable correlation with RMSD.

For crossdocking, while use of FMS scoring alone showed significant improvement with regards to systems on the diagonal (cognate protein-ligand pairs), the overall matrix success rate in 4 out of 6 cases was significantly lower than SGE. Examination of the underlying reference structures showed that FMS docking success is highly dependent on how well the pharmacophore reference overlays with the RMSD reference (Figure 2-12~2-15). Thus, while use of FMS scoring alone to drive sampling of a ligand using a reference without possibilities for good overlap yields poor results, such behavior makes physical sense. More importantly, the results dramatically emphasize that the FMS function works best when the goal is identification of molecules that resemble the reference, as was the original intent. As expected, use of the combined FMS+SGE function provides more of a balance and yields the highest crossdocking matrix success rates (Figure 2-12).

In terms of enrichment, receiver operator characteristic (ROC), area under the curve (AUC), and fold enrichment (FE) analyses, in general, showed that FMS and FMS+SGE

functions yield better performance than SGE alone (and random selection) for both early and total enrichment (Figure 2-16 and Table 2-5) when evaluated over 15 systems taken from the DUD-E database. For several systems FMS+SGE enrichment appears roughly in between that obtained using FMS or SGE alone (Figure 2-16). Importantly, FE values computed very early in rank-ordered lists (0.1% and 1%) showed using FMS and FMS+SGE yielded 10-13 out of 15 FE values enhanced relative to the standard protocol SGE (Table 2-5 column A, B) despite the fact that only a "single" reference (cognate ligand) was used to guide sampling of compounds. Future studies should evaluate enrichment outcomes using multiple FMS references.

In terms of virtual screening, rescoring results obtained from standard docking to three target of pharmaceutical interest (EGFR, IGF-1R, and HIVgp41) showed that the FMS and FMS+DCE (equivalent to FMS+SGE) methods yielded more compounds with greater numbers of pharmacophore matches when the top 25 compounds from each method were examined (Figure 2-18). The example also demonstrated how FMS scoring could utilize small organic molecules or non-contiguous protein sidechains as references. For gp41 in particular, examination of top poses revealed that none of the compounds matched an important Asp sidechain in the initial pharmacophore model. A simple modification of the reference to include multiple copies of the Asp weighted this functionality more highly, and when rescored, yielded top-ranked compounds with the desired interaction. Importantly, this result further establishes the importance of the FMS "reference" in addition to demonstrating how pharmacophores could be customized.

Finally, the current results suggest several directions for future research including exploring other functional forms of the main FMS equation (eq 2-1), testing FMS score in combination with other scoring functions (i.e. footprint similar scoring), development of a



receptor-based<sup>122</sup> as opposed to the current ligand-based method, and implementation of routines to address multiple pharmacophore references simultaneously.<sup>123</sup> Ongoing work involves incorporation of FMS scoring into a *de novo* design version of DOCK, currently under development in our laboratory, to allow pharmacophore-guided *de novo* growth of new ligands from scratch having similar binding profiles as a known reference.

## Chapter 3. FMS-guided Virtual Screen to HIVgp41

This Chapter provides additional analyses on application case studies employing pharmacophore-based scoring in database enrichment and virtual screening targeting HIVgp41.

### 3.1 Introduction

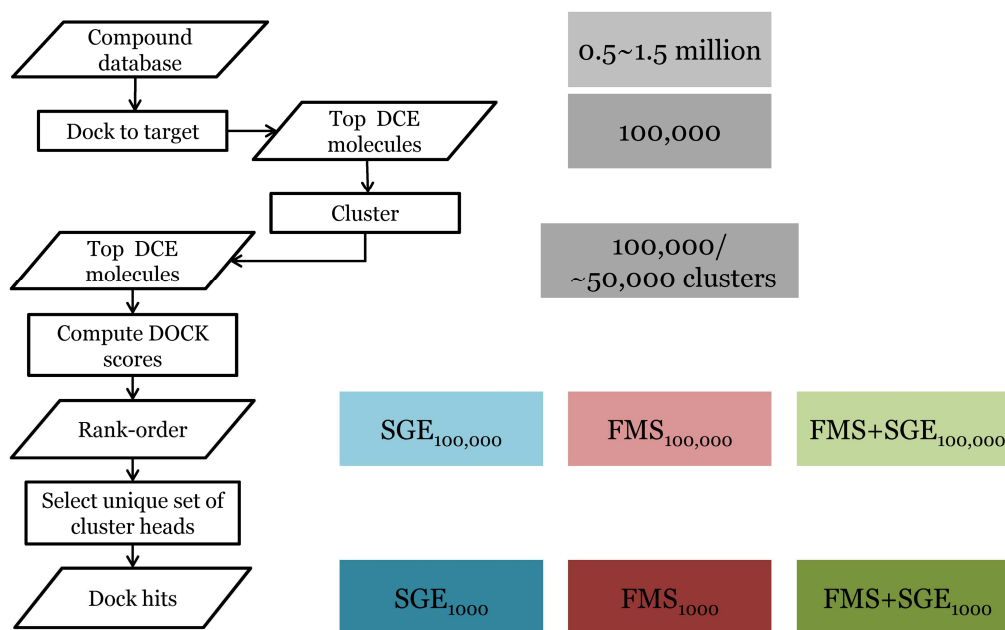
The ultimate goal of structure-based drug design is to obtain potent ligands with desirable interactions against a given target. In this Chapter, we present detailed visual inspection of enrichment studies, which is an important indicator of the expected outcomes for the computer-aided drug design approach virtual screening, and application studies of FMS-guided docking. Rescoring virtual screenings results with different DOCK scoring functions were performed targeting HIVgp41, to both the well-known hydrophobic binding pocket as well as a new NHR inner pocket recently identified by Allen *et al.*<sup>14</sup>

As introduced in Chapter 1, the viral protein HIVgp41 is an attractive anti-HIV drug target. To date, the only FDA approved HIV fusion inhibitor, T20, is a peptide-based drug. However, small molecule drugs targeting HIVgp41 are of great interest. In addition to the known conserved hydrophobic pocket on the surface of the NHR trimer where the known peptide inhibitor C34 binds (Figure 2-18C,D, Figure 2-1a), our lab has recently identified an inner pocket found at the internal interface of the three NHR helices.<sup>14</sup> Virtual screening to the inner pocket and experimental efforts to validate the mechanism of NHR trimer formation and confirm the target eligibility of this pocket are ongoing in collaboration with researchers Dr. Amy Jacobs (SUNY Buffalo) and Dr. Miriam Gochin (University of California, San Francisco).

Validation tests and application tests for small molecule inhibitor design can help to provide additional insight into the FMS scoring protocol and guide protocol refinement to aid future works. By comparing the sampling sizes and hit properties in virtual screening, we can evaluate the effects of using different scoring functions such as single grid energy (SGE), pharmacophore matching similarity (FMS) and the combination FMS+SGE. Our hypothesis is that FMS and FMS+SGE procedure can serve as robust scoring and sampling protocols for alternative virtual screening compound selection outcomes.

### **3.2 Methods and Computational Details**

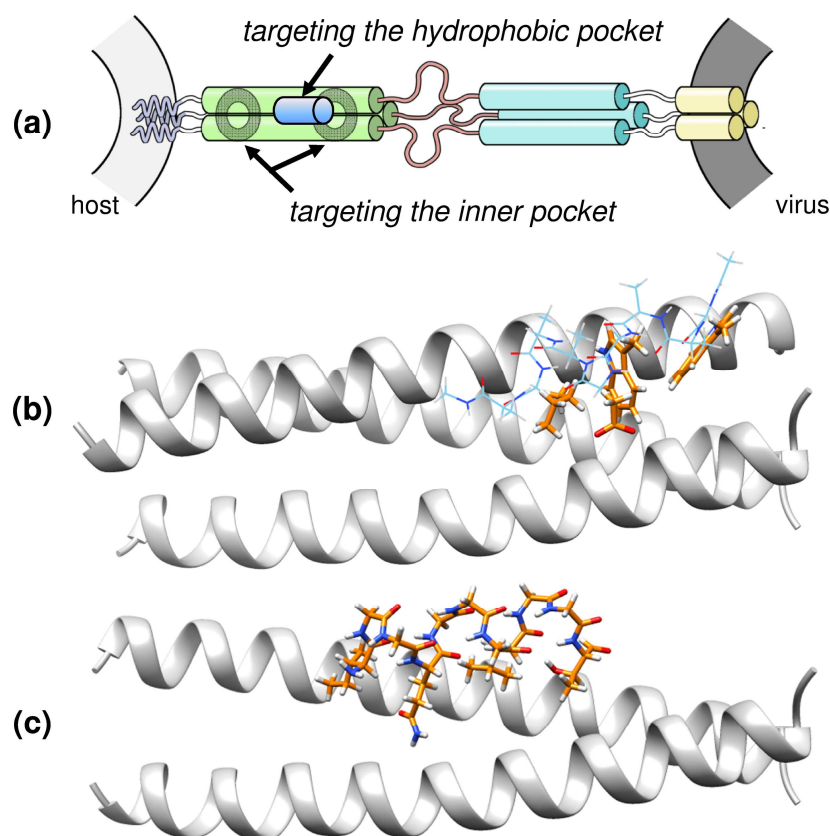
*Structured-based drug design techniques and protocols.* Previously in Chapter 2, enrichment studies were performed on 15 DUD-E systems using three different docking protocols (SGE, FMS and FMS+SGE). And as was discussed, FMS and FMS+SGE can significantly enhance enrichment rates. In this chapter, we provide more in-depth visual inspection of docked poses for top scored molecules in a subset of the enrichment systems.



**Figure 3-1.** Flow chart represents the standard Rizzo lab virtual screening protocol. Colored boxes represents the approximate size of the compound set studied in each step of virtual screening.

Figure 3-1 illustrates the general procedure for virtual screening in this study. In the Rizzo lab, virtual screening is performed via the following five steps: 1) prepare the target protein for docking and download a compound database for screening; 2) perform individual on-the-fly docking for all molecules in the database using the DOCK single grid energy (SGE) score followed by restrained minimization using DOCK Cartesian energy (DCE) score to eliminate molecules with unfavorable affinity to the target; 3) cluster top DCE scored molecules according to molecular properties such as molecular weight, 2D fingerprint etc.; 4) rank-order clustered heads by different DOCK scoring functions; 5) select top hits from each rank-ordered list by visual inspection, and purchase samples for experimental testing. Typically, the initial set for screening contains 0.5 to 1.5 million compounds (step 1). The top 100,000 DCE scored molecules will be evaluated in step 2 through step 4. Finally, about 1000 top scored molecules using different metric will be inspected in step 5 (Figure 3-1, right).

Using the virtual screening protocol in Figure 3-1 that has been previously shown by our group to yield promising hits for the known HIVgp41 hydrophobic pocket<sup>72</sup> (Chapter 2), Allen *et al* recently performed a screen to a newly identified inner pocket on the gp41 NHR region using an IQLT peptide derived from one N helix as the pharmacophore reference.<sup>14</sup> In this Chapter, we rescore these virtual screening results to both gp41 pocket using FMS-based methods. Figure 3-2b-c shows the two targets structurally.



**Figure 3-2.** HIVgp41 inner-pocket for FMS-guided virtual screening. (a) Mechanism of blocking N-helical trimer association via targeting inner pocket and blocking formation of the six helical bundle (6HB) via targeting the hydrophobic pocket. Figure adapted from work by Allen *et al.*<sup>120</sup> (b) Visualization of the hydrophobic pocket (gray ribbon) and pharmacophore reference for *de novo* design (cyan line representation) and virtual screening (orange stick representation). (c) Visualization of the inner pocket (gray ribbon) and the pharmacophore reference (orange stick representation).

### 3.3 Results and Discussion

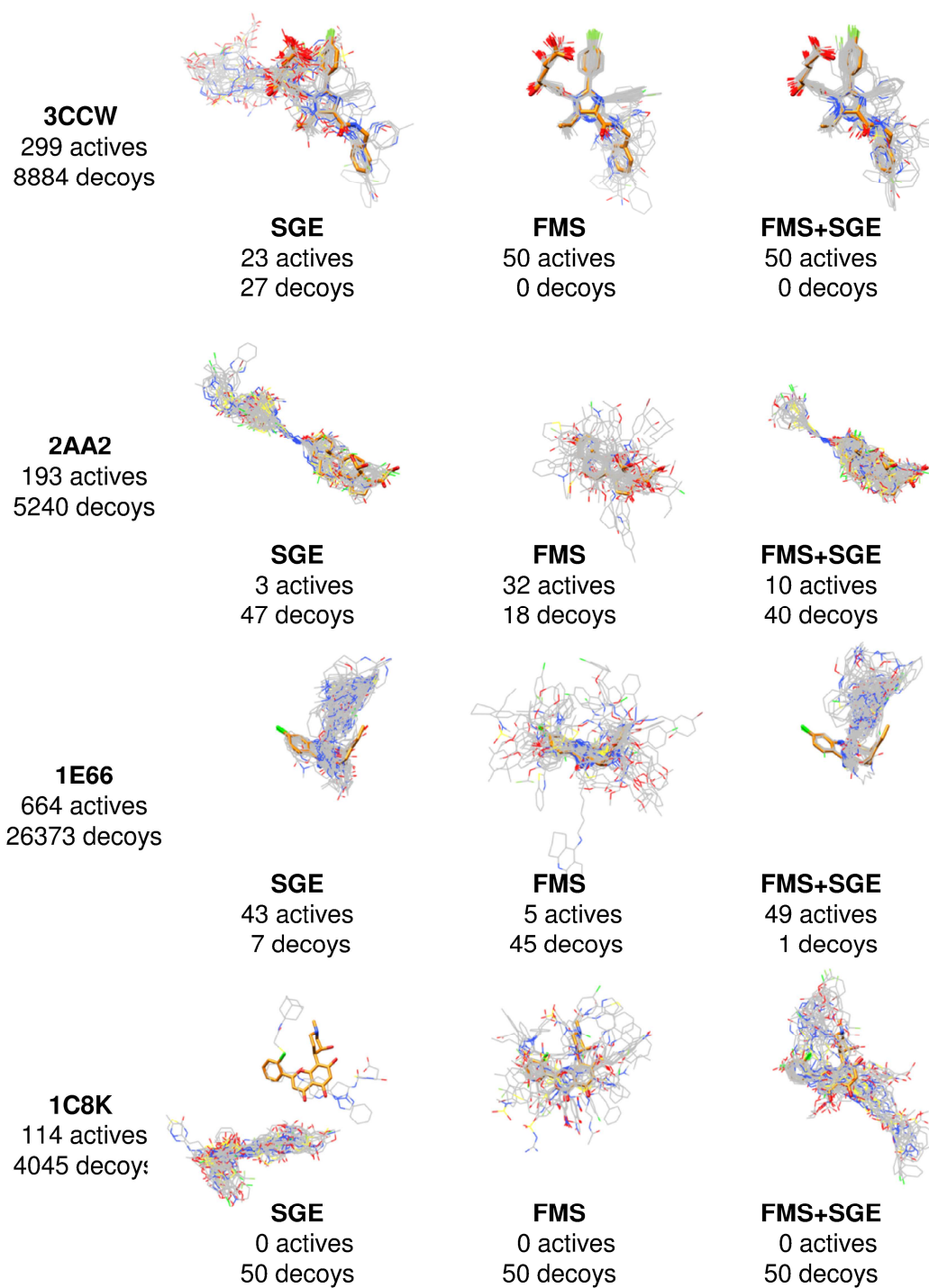
#### 3.3.1 Enrichment Study Analyses on Four Representative Systems

The top 50 molecules picked by the three DOCK scoring functions SGE, FMS and FMS+SGE from four DUD-E systems 3CCW, 2AA2, 1E66, 1C8K contain different numbers of actives and decoys, representing different stages of early enrichment rate. The enrichment data and molecular visualization of the top 50 hits are shown in Figure 3-3. Note that molecules in DUD-E may contain multiple compounds with the same molecular ID, based on different protonation states or tautomer, but only the best scored candidate compound is retained for rank ordering and “hit” selection. Overall, the binding poses of the top 50 molecules (carbon atoms in grey in Figure 3-3) picked by FMS score overlay more tightly to the reference molecules shown in orange. In contrast, molecular clusters formed by the top 50 molecules picked by SGE protocol tend to occupy additional volumes that are not filled with the crystal pose.

For systems 3CCW and 2AA2, FMS score has the best overall and early enrichment performance in all stages as reported previously in Chapter 2 (Figure 2-16 and Table 2-5). For system 3CCW, both FMS and FMS+SGE yields very promising results with 100% true actives for the top 50 molecules when only ~50% of the top SGE molecules are true actives. And, the top FMS and FMS+SGE hits all have almost perfect overlap in terms of binding poses to the reference molecule (row 3CCW middle and right panel in Figure 3-3). In the system 2AA2, top molecules selected by SGE and FMS+SGE both favors a extra pocket adjacent to where the reference molecule binds, as shown by the some molecular cluster to the top left of the crystal ligand in Figure 3-3 row 2AA2. However, the top FMS molecules generate poses with ring structures tightly clustered around the center of the reference pose (the gray clouds in the center for row 2AA2 middle panel in Figure 3-3) with few occupancy to that top left region filled by

top SGE and FMS+SGE molecules. The total molecular volume with high occupancy by top molecules for FMS+SGE protocol is in between that for SGE and FMS protocol. For 1E66, although FMS protocol has the worst early enrichment (for 0.1%, 1% and 10% of database, Table 2-5) and has the least number of actives in the top 50 molecules, the top scored poses still yields the best overlap visually with the reference pose at the matched region (Figure 3-3, row 1E66, middle panel) with unmatched segments scattered in all directions from the reference. Top SGE and FMS+SGE molecules cluster into a molecular volume that only intersects with the reference pose in a small percentage (Figure 3-3, row 1E66, left and right panel) when the unmatched segments tightly overlapped. The only system where all three protocols fail to do better than random selection in terms of enrichment rate is 1C8K. The top 50 molecules picked by SGE, FMS and FMS+SGE protocols are all decoys instead of actives. Although the early enrichment rates are all poor, poses generated with FMS protocol and FMS+SGE protocol are still reasonably well overlapped with the reference pose, while only 2 of the top 50 SGE molecules are docked to the same pocket as the reference molecule.

Overall, the in-depth conformational inspection of the enrichment study results again validates the reliability of FMS and FMS+SGE scoring protocol to enrich for true binders while competing with decoy compounds with similar physico-chemical properties. These results further demonstrate the robustness of our pharmacophore-based scoring protocol and the utility to aid virtual screening.

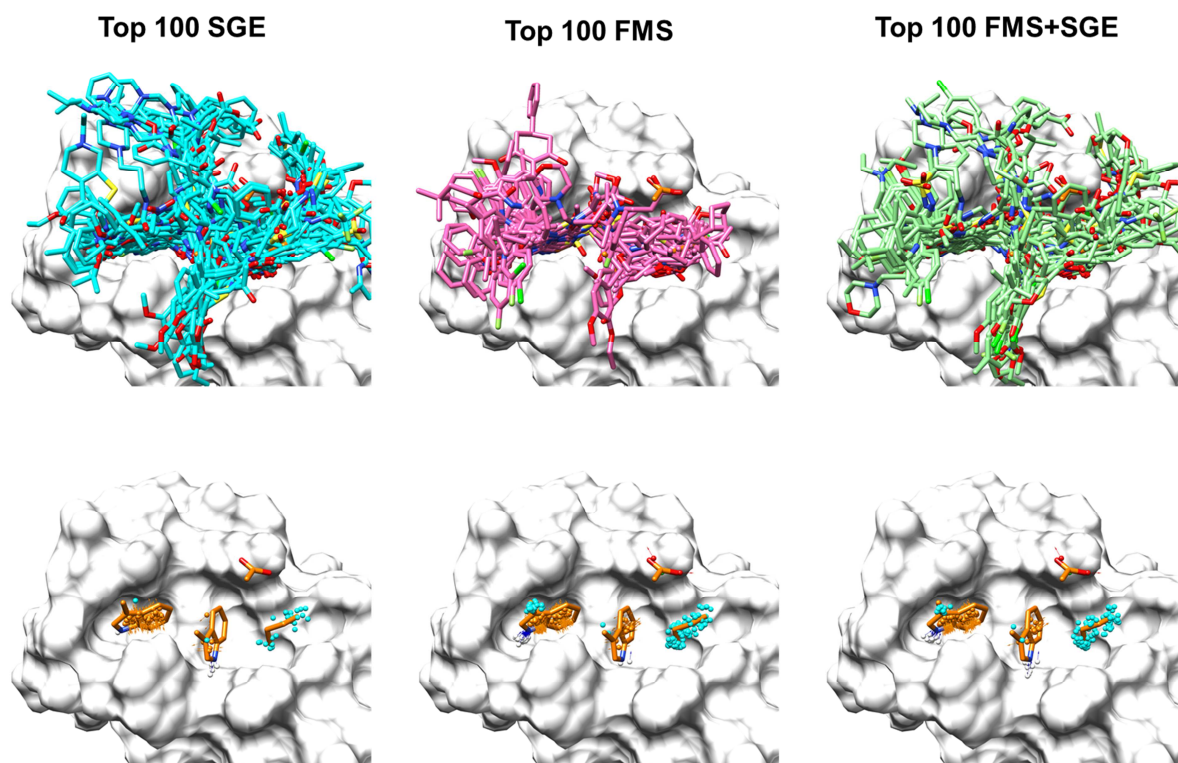


**Figure 3-3.** Early enrichment represented by the predicted binding poses of top molecules selected by SGE, FMS and FMS+SGE. Reference ligand shown in orange stick model; molecular volume of the reference ligand shown in gray surface model; top 50 molecules in DOCK predicted conformations shown in gray line model. PDB codes for the four systems are 3CCW, 2AA2, 1E66 and 1C8K.



### 3.3.2 Rescoring Virtual Screening to HIVgp41

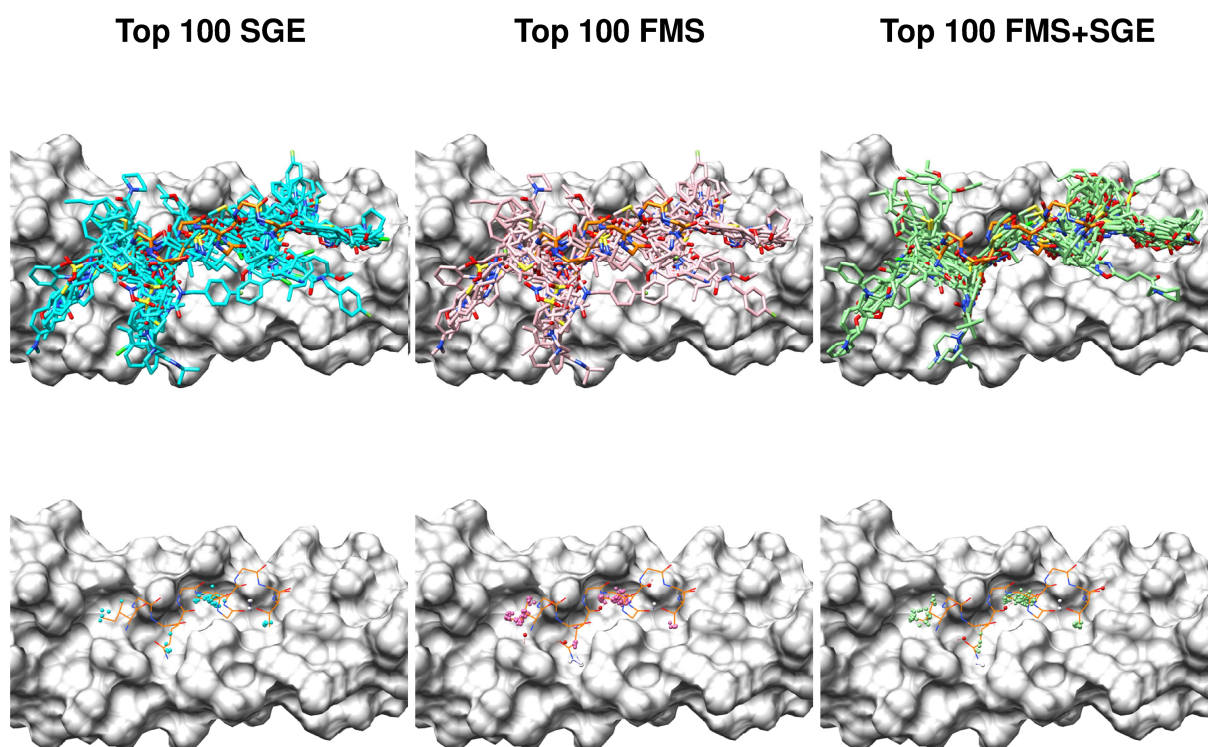
***HIVgp41 hydrophobic pocket.*** As described in Chapter 2, ~0.5 million compounds previously screened to the HIVgp41 hydrophobic pocket by Holden *et al*<sup>72,119</sup> were rescored using SGE, FMS and FMS+SGE scoring functions. Here, visualizations of the top 100 molecules from these rescoring tests along with their matched pharmacophore models to the reference ligand pose in the pocket are shown in Figure 3-4. In general, the top molecule poses generated with FMS (carbon atoms colored in pink) are consistently more tightly clustered around the reference, which in this case is four (WWDI) isolated amino acid side chains (Figure 3-4, top row). Visualization of the actual matched pharmacophore points yield similar results with FMS yielding more matches compared to SGE (Figure 3-4, bottom row, middle panel vs. left panel). The top SGE molecules (carbon atoms colored in cyan) fill not only the part of the binding site that has been occupied by the reference (central pocket) but also several adjacent pockets (Figure 3-4, top left panel). Visually, the top SGE molecules are also larger than the other two groups of molecules. The top 100 molecules picked by FMS+SGE are more medium-sized compared to the top SGE or top FMS scored molecules with fewer occupancy outside of the central pocket and reasonably well matched pharmacophore models to the reference. (Figure 3-4, bottom row)



**Figure 3-4.** Representative results from FMS-guided virtual screening targeting the HIVgp41 hydrophobic pocket (PDB ID: 1AIK). The Pharmacophore reference (orange) is the WWDI key residue side chains from C34. Molecules shown in the top row; Matched pharmacophore models in the bottom row.

*HIVgp41 inner pocket.* To gauge virtual screening rescoring outcomes for a different target system, SGE, FMS and FMS+SGE methods were used to rescore results generated by Allen *et al*<sup>14</sup> from docking 1 million ligands to the HIVgp41 inner pocket introduced in Figure 3-3. Visualizations of the top 100 molecules with their matched pharmacophore models to the reference poses for the HIVgp41 inner pocket are shown in Figure 3-5. Here, this binding pocket occupied by the reference peptide (IQLT) contains two hydrophobic residues and two polar residues with hydrogen bond acceptors and hydrogen bond donors. In contrast to the previous example, no ring-containing residues were used for this reference. Visually, top hits from all three methods occupy the full span of the binding pocket. In addition to overlaps with reference

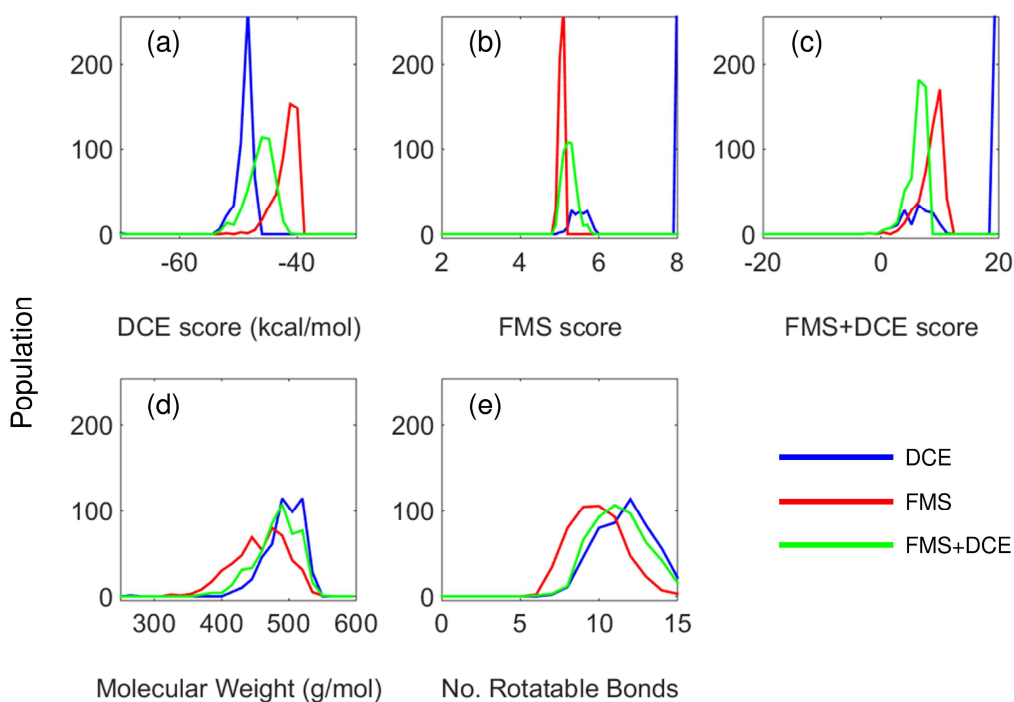
residues, top molecules picked by FMS and FMS+SGE have more occupancy on the left corner of the binding site as shown by the hydrophobic pharmacophore point clusters close to the Ile group (Figure 3-5, bottom middle and bottom right). Compared to re-screening to the hydrophobic pocket (Figure 3-4), the results here are less dramatic.



**Figure 3-5.** Representative results from FMS-guided virtual screening targeting the HIVgp41 inner pocket. The pharmacophore reference (orange) is comprised of the IQLT key residues from one N helical peptide. Molecules shown in the top row; Matched pharmacophore models in the bottom row.

To further gauge the differences in top-scored hits selected by FMS and energy scores, molecular properties of the top 500 molecules picked by DCE (equivalent to SGE), FMS, and FMS+DCE targeting the gp41 inner pocket are shown in Figure 3-6. Consistent with the analyses in Chapter 2 (Figure 2-19), use of DCE score yielded most favorable DCE results (Figure 3-6a, blue vs. red and green curves), use of FMS score yielded best pharmacophore

overlap (Figure 3-6b, red vs. blue and green curves) and use of FMS+DCE yielded overall lowest FMS+DCE scores which is as expected (Figure 3-6c, green vs. blue and red curves). The best FMS scores for this inner pocket screen are around 5.0 (Figure 3-6b, red curve), compared to 4.2 (Figure 2-19, red curve) for the hydrophobic pocket. This indicates the less overlap in this pocket may be related to the fact that the reference compound does not contain rings. In addition, FMS again showed the least amount of bias on the size of the molecules (Figure 3-6d-e, red vs. blue and green curves). Compared to the HIVgp41 hydrophobic pocket, top hits in the inner pocket are generally more flexible as the number of rotatable bonds peaks at around 10 in Figure 3-6e while hits in the hydrophobic pocket on average have around 8 rotatable bonds (Figure 2-19e). The molecular weights of hits for the two pockets, however, both are centered around 500 g/mol, which agrees well with known “drug-like” properties.<sup>22</sup> This observation is likely to correspond to the binding profiles of the two binding pockets. For instance, the native peptide inhibitor for the inner pocket consists of amino acids that are relatively smaller (IQLT), and the native substrate for the hydrophobic pocket (WWDI) consists of indole-containing residues that are much larger in size.



**Figure 3-6.** Histogram of (a) DCE score, (b) FMS score, (c) FMS+DCE score, (d) molecular weight, and (e) number of rotatable bonds for the top 500 scored molecules scored by DCE (blue), FMS (red), and FMS+DCE (green) from a virtual screen targeting the HIVgp41 inner pocket.<sup>14</sup>

### 3.4 Conclusion

In summary, in this Chapter we performed enrichment structural results analyses as well as virtual screening rescoring tests targeting the HIVgp41 hydrophobic and inner pockets. Judged by the ability of FMS to enrich for known actives, this method is likely to be an important tool to aid virtual screening.

It is important to emphasize a new DOCK descriptor score that allows scoring of molecules with different DOCK scoring functions at the same time was employed in this Chapter. Customized weight can be assigned to different components in the descriptor score. This is particularly useful in driving the sampling with a hybrid score in future screening. However, complexities can also arise when optimal weights on the individual score component

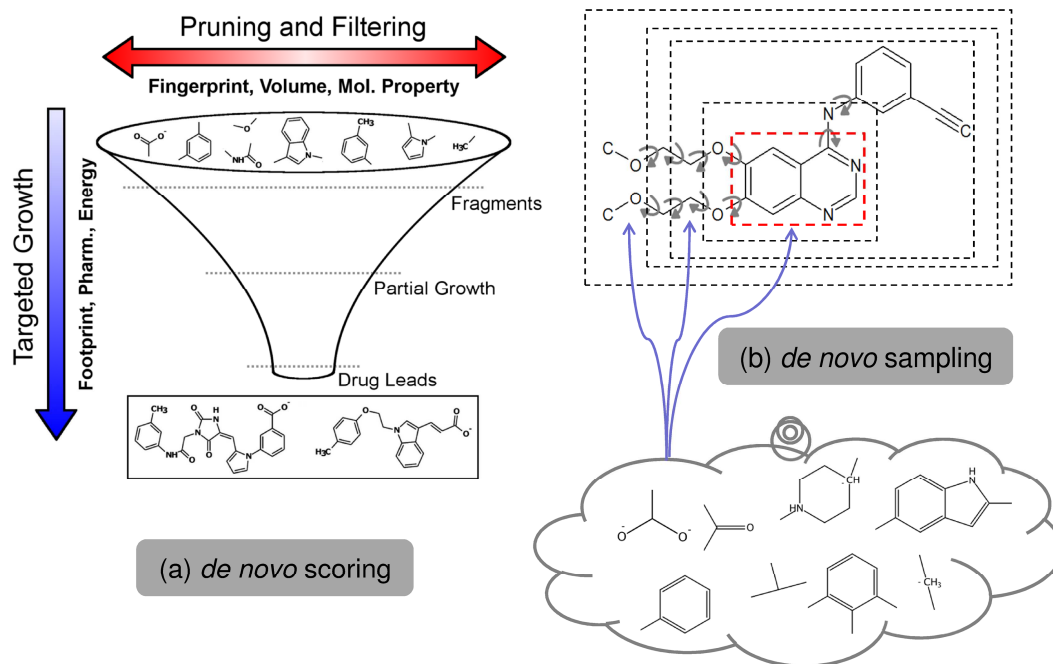
for the combinations are not employed. Thus, it is recommended that the users also perform their own validation tests when deviating from the values used here.

## Chapter 4. FMS-guided *de novo* Design to HIVg41

This Chapter provides analyses on preliminary application case studies employing pharmacophore-based scoring for *de novo* design.

### 4.1 Introduction

In addition to standard ligand docking, which relies primarily on robust sampling and scoring routines to search conformational space, an alternative technique used computer-aided approaches for ligand discovery is termed *de novo* design as discussed below. The *de novo* design of a novel molecule from scratch requires two principle tasks: (1) obtaining molecular “building blocks”, and (2) assembling the “building blocks” into physically reasonable molecules with desirable pharmacological properties in a timely manner.<sup>124</sup> Compared to virtual screening, *de novo* design does not *a priori* limit the search space to a chosen compound database, which makes it a useful alternative tool to search over very large chemical sub-spaces which are estimated to contain approximately  $10^{60}$  (<500 Da) molecules.<sup>125</sup> In recent years, many *de novo* design drug design programs, such as LEGEND,<sup>126</sup> *LeapFrog*,<sup>127</sup> LEA3D,<sup>128</sup> BOMB,<sup>129,130</sup> and LigBuilder,<sup>131,132</sup> have been developed.<sup>133</sup> Ongoing development in the Rizzo lab lead by postdoctoral fellow Dr. William J. Allen<sup>134</sup> aims to implement a robust *de novo* design algorithms into the docking program DOCK to leverage recent advancements in on-the-fly sampling and new scoring functions. *De novo* DOCK version 2015-01-15<sup>134</sup> (which included the FMS scoring function) has been used in this Chapter to perform FMS-guided *de novo* design for both method validation and case studies targeting HIVgp41.



**Figure 4-1.** Illustration of *de novo* design protocols: (a) Horizontal pruning and guided growth in *de novo* design. Figure generated by William J. Allen.<sup>134</sup> (b) *de novo* growth of molecules by adding new segments from a user-defined fragment library instead of pre-defined fragments at all attachment points during the anchor-and-grow sampling processes. Anchor highlighted by the red box and the growing molecule at each layer shown in the grey boxes.

The *de novo* design strategy implemented into DOCK constructs novel molecules from scratch using a modified version of the anchor-and-grow algorithm. The procedure requires “building blocks” (*i.e.* fragments), which are obtained from common molecular segments generated from a set of existing and purchasable drug-like molecules. The frequency of each fragment is saved for later reference. Also, the atom environments on both sides of the “breaking point”, which corresponds to a rotatable bond identified by DOCK, are documented and saved as an “allowable bond library”. In the assembly step of *de novo* growth, only bonds contained in the “allowable bond library” can be formed. Importantly, this restriction helps to enforce physically reasonable molecules and increases the chemical feasibility of final hits generated by the algorithm. Similar to the standard DOCK anchor-and-grow algorithm, *de novo*



growth employs an “inside-to-outside” strategy. Starting from the placement of an anchor (highlighted by the red box in Figure 4-1b), for each attachment point on an anchor (layer 0), a series of fragments from the fragment library will be evaluated. The compatible fragments satisfying the conditions defined in the “allowable bond library” after attachment will be added to the molecule and the new structure will be saved until the maximum number of constructs is achieved at any attachment point. After adding the first layer of fragments, all new constructs are evaluated together. If there are no attachment points (tagged by dummy atoms) in the construct, then the complete molecule is saved as final molecule for output. Otherwise, it will be clustered and filtered (horizontal pruning, Figure 4-1a) by scoring functions and other metrics that define the molecular properties of the constructs. The top scored (guided growth, Figure 4-1a) clusterheads will then be used for growing the next layer. By design, different DOCK scoring functions can be used for pruning at each step of *de novo* growth, as well as guide growth to optimize the final molecules for certain molecular properties such as similarity to a known reference molecule. Importantly, *de novo* design is often considered to rely even more on accuracy of the scoring method used compared to virtual screening.<sup>131</sup> In this study, we evaluated the robustness of FMS and FMS+SGE as the scoring and sampling protocols for *de novo* design. Overall, the preliminary tests yielded promising results, indicating that FMS-guided *de novo* design has the potential to construct small molecule ligands with similar binding profiles to the known references in an extended chemical space.

## 4.2 Methods and Computational Details

***Focused library de novo tests.*** As a first battery of tests to validate the performance of new FMS-guided *de novo* DOCK protocols, we have run *de novo* reproduction tests for 50

SB2012 systems as listed in Table 4-1. For each reproduction test case, we first generated a highly restricted “focused” fragment library by decomposing each crystallographic ligand into scaffolds (>2 attachment points), linkers (2 attachment points), and sidechains (1 attachment point) by breaking the rotatable bonds in each molecule and storing each non-redundant fragment. The combined set of scaffolds, linkers, and sidechains are saved as “anchors”. The total number of fragments in the anchor file is shown in Table 4-1 (Column 3 and Column 6). Each segment in the anchor file is oriented in the binding site to obtain the initial set of roots (molecular construct with attachment points) for growth. The maximum number of orientations for the current tests was set to 10000. For *de novo* sampling, a “graph” method was used to select new fragments. Briefly, a fragment graph that includes the pair-wise Tanimoto and rank-ordered lists of similarity among all the input fragments are prepared in the beginning of the *de novo* procedure to optimize how fragments are chosen for each attaching event. The maximum number of starting points to try in the fragment graph is set to 10 while the breath and depth of the graph are set to 5 and 2, respectively. If adding one fragment at a given attachment point favorably enhances the overall score, then the attaching event is accepted and fragments that similar to the current fragment will be chosen to generate more constructs at the same attachment point. Otherwise, this fragment is kept based on a acceptance probability calculated using equation  $P = e^{(-\Delta E/KT)}$ . Here,  $\Delta E$  represents the difference in energy before and after adding the fragment; T is the annealing temperature initially set to 100 and gradually decreases in the growth procedure. As a result, an unfavorable fragment is more likely to be kept earlier on in the growth step and less and less likely to be kept later on. After adding fragments to generate a maximum number of 50 next layer partially grown molecules for each root construct, a molecular weight restraint of 1000 and a maximum number of 15 rotatable bonds will be used to

filter newly built constructs. Then the subset of new layer molecules is pruned by Tanimoto (cutoff set to 1.0) and the Hungarian RMSD heuristic<sup>110</sup> (unmatched number cutoff of 0 and matched region RMSD cutoff of 2.0 Å). Next layer constructs that have no attachment points at this step will be written to file, while a maximum number of 50 (maximum root size) of the remaining partially-built molecules are returned to root for the next iteration of growth. In this study, the maximum number of growth layers is set to 7. As with standard flexible ligand docking, internal energy is used during ligand growth to avoid internal clashes with a repulsive-only VDW potential with exponent 12.

**Table 4-1.** List of 50 systems for initial *de novo* validation from SB2012.<sup>100</sup>

<b>PDB</b>	<b># Fragment</b>	<b>PDB</b>	<b># Fragment</b>	<b>PDB</b>	<b># Fragment</b>
<b>1ACM</b>	6	<b>1FH7</b>	4	<b>1O2Q</b>	7
<b>1AID</b>	6	<b>1FHD</b>	4	<b>1O37</b>	7
<b>1BIR</b>	7	<b>1G9V</b>	8	<b>1O5G</b>	7
<b>1BJU</b>	6	<b>1H46</b>	7	<b>1PMN</b>	8
<b>1BN4</b>	8	<b>1HAK</b>	6	<b>1Q95</b>	6
<b>1BR5</b>	5	<b>1HDQ</b>	7	<b>1RDN</b>	5
<b>1C8V</b>	6	<b>1IEP</b>	8	<b>1RGK</b>	7
<b>1C9D</b>	5	<b>1JJE</b>	5	<b>1RGL</b>	7
<b>1CPS</b>	6	<b>1JJT</b>	4	<b>1RKG</b>	6
<b>1CW2</b>	5	<b>1JLA</b>	5	<b>1RNT</b>	7
<b>1CX9</b>	5	<b>1JLG</b>	8	<b>1ROB</b>	7
<b>1D09</b>	6	<b>1K3U</b>	6	<b>1RT2</b>	5
<b>1D4P</b>	7	<b>1NFU</b>	7	<b>1S1T</b>	8
<b>1DY4</b>	7	<b>1NFX</b>	6	<b>1T40</b>	7
<b>1E6S</b>	4	<b>1NFY</b>	7	<b>1T46</b>	8
<b>1E72</b>	4	<b>1O2H</b>	7	<b>1TZ8</b>	4
<b>1EJN</b>	7	<b>1O2I</b>	7		

Scoring functions used in the *de novo* reproduction tests presented in this Chapter include the following: (1) single grid energy (SGE) score, where 6-9 Lennard-Jones, distance dependent dielectric ( $\epsilon=4r$ ) and a grid box with 8.0 Å extension from all sphere points and 0.3 Å resolution are used; (2) Pharmacophore matching similarity (FMS) score; (3) FMS+SGE score, where the weight parameters for FMS and SGE are 10 and 1, respectively. Additional combinations methods include (4) SGE+Tanimoto, where the weight parameters for SGE and Tanimoto are 1 and -50; (5) FMS+Tanimoto, where the weight parameters for FMS and Tanimoto are 1 and -5; (6) FMS+SGE+Tanimoto, where the weight parameters for FMS, SGE, and Tanimoto are 10, 1 and -100.

Specifically for the FMS-guided *de novo* tests, we merged the FMS scoring protocol with the development version of *de novo* DOCK (version 2015-01-15).<sup>134</sup> The intention is to release the FMS and *de novo* functionality as DOCK6.8. It should also be noted that in this study, scoring functions tested are used to guide vertical *de novo* growth only (See Figure 4-1a). Currently, horizontal pruning is performed with the repulsive internal energy, pairwise Hungarian RMSD<sup>110</sup> and Tanimoto function among the partially grown molecules at each layer.

**Targeting HIVgp41.** Similarly to the virtual screen that used peptide-based references to guide compound selection targeting the HIVgp41 hydrophobic pocket and the inner pocket in Chapter 3, here we performed FMS-guided *de novo* design to bias “from-scratch” ligand growth also guided by peptides. The focused fragment libraries derived from the continuous peptides with key residues IQLT (inner pocket, Figure 3-2b) and WWDI (hydrophobic pocket, Figure 3-2c) were retained at the interface. For the hydrophobic pocket, the peptide used as pharmacophore reference and for fragment library generation is obtained by keeping the residues from residue Trp117 (W, gp41 sequencing) to residue Ile124 (I, gp41 sequencing) while mutating intermediate residues to Alanine other than the four key amino acids (WWDI). Note that this reference molecule is slightly different from what was used in the prior the virtual screening, which contains only the side chains of the four disjoint residues (WWDI). Construction of the inner pocket reference (IQLT) followed the same protocol. The focused fragment library generated for the hydrophobic pocket (1AIK-WWDI) includes 1 scaffold, 4 linkers and 5 sidechains. The focused fragment library generated for the inner pocket (1AIK-IQLT) includes 1 scaffold, 4 linkers, and 6 sidechains.

## 4.3 Results and Discussions

### 4.3.1 Focused library *de novo* runs: small molecule reference reproduction.

The focused library *de novo* design tests on the 50 SB2012 systems can serve as validation for working protocol employing different scoring metrics including FMS. For each system, a subset of novel molecules with various molecular fingerprints were generated. These *de novo* output molecules were then evaluated using different DOCK scores as well as their molecular properties including Tanimoto and Hungarian score to the corresponding crystal ligand in SB2012. Note that standard Hungarian RMSD can be directly derived from the Hungarian score for molecules with Tanimoto of 1.0 to the reference molecule because for two molecules with a Tanimoto of 1.0, the Hungarian score is equal to  $-5 + \text{Hungarian RMSD value}$ . Thus, if at least one molecule generated by given *de novo* DOCK has a Tanimoto of 1.0 to the crystal ligand, we identify this system as being successfully reproduced in terms of Tanimoto (Tanimoto reproduction) by the given *de novo* protocol. In addition, if for at least one of the reproduced molecules, its binding pose predicted by *de novo* DOCK is within 2Å (Hungarian RMSD) to the crystal pose, then we identify this system as being successfully reproduced in terms of RMSD (RMSD reproduction) by the *de novo* protocol. The objective in optimizing the protocols is to tune input parameters and setups to maximize the number of systems that reproduce both Tanimoto and RMSD values.

Table 4-2 shows the total sampling sizes (column b) resulting from *de novo* design for all 50 systems as well as the number of systems that are Tanimoto reproduced (column c1) by *de novo* DOCK using six scoring protocols: SGE, FMS, FMS+SGE, SGE+Tanimoto, FMS+Tanimoto, and FMS+SGE+Tanimoto. By relaxing the Tanimoto cutoff, we find increased Tanimoto reproduction rate, depending on what function is used, as shown by the increased

number of compounds going from column c2 (Tanimoto cutoff of 0.95), column c3 (Tanimoto cutoff of 0.8), column c4 (Tanimoto cutoff of 0.7) to column c5 (Tanimoto cutoff of 0.6). In most cases, the current protocol can reproduce the original ligand to a Tanimoto of 1.0, especially if growth is driven using Tanimoto as a component of the scoring function (Table 4-2 column c1, row SGE+Tanimoto: 45, FMS+Tanimoto: 47 and FMS+SGE+Tanimoto: 45). And, if the Tanimoto cutoff is loosened to 0.6, most systems can be rebuilt for all six protocols (reproduction between 47 and 50 systems, Table 4-2, column c5). In all cases, FMS-alone growth yielded the highest reproduction rate compared to SGE and FMS+SGE. For example, in column c1 in Table 4-2, the FMS protocol reproduced 45 systems while SGE reproduced 36 and FMS+SGE reproduced 37. When boosted by Tanimoto, FMS+Tanimoto protocol reproduced 47 systems while SGE+Tanimoto and FMS+SGE+Tanimoto both reproduced 45. Future studies however should examine if sample size is leading to differences in the Tanimoto being reproduced, through using consistently sized ensembles (i.e. top scoring compounds only). Additionally, the focused (small) libraries are “artificial” in the sense that to be of practical use, de novo design protocols must also behave well when using much larger generic libraries. These initial tests are meant only to be a first step in the overall validation procedure.

Note that for one particular system 1AID, *de novo* DOCK using SGE, FMS+SGE, SGE+Tanimoto and FMS+SGE+Tanimoto protocols all terminated growth before layer 7 with 0 output molecules. Interestingly, FMS and FMS+Tanimoto were able to generate 59 molecules (up to layer 7) and 3 molecules (up to layer 4) for 1AID. Visual inspection showed that 1AID is a relatively large molecule with few rotatable bonds (rotN=6, molecular weight = 453.1 g/mol). The focused fragment library for 1AID includes 4 large ring-containing fragments (molecular weight 86.2-112.6 g/mol) and two small fragment containing single heavy atoms (molecular

weight 15.0-17.0 g/mol). It is likely that partially grown molecules for system 1AID already exceeded the maximum molecular weight limit during the *de novo* growth before layer 7, especially when driven by energetic scoring functions. Thus only FMS and FMS+Tanimoto were able to complete *de novo* growth and reproduce the crystal ligand in terms of Tanimoto. This suggests that for some applications use of the FMS protocol without an accompany energy term (i.e. SGE) can be useful.

Table 4-3 shows the number of RMSD reproduction cases out of the Tanimoto reproduced systems with Tanimoto cutoff of 1.0 and RMSD cutoff of 2Å (column d1), 2.5Å (column d2) and 3Å (column d3). One obvious observation is the significant decrease in the number of RMSD reproduced cases for all six protocols. This indicates that sufficient sampling in the *de novo* protocol to re-generate the 3D binding geometry (measured by RMSD) is a more challenging problem than reproducing the 2D fingerprint (measured by Tanimoto) alone. However, by relaxing the RMSD cutoff to 3Å (column d3), over 50% of the Tanimoto reproduced cases (Tanimoto =1.0) and nearly 50% of the total test systems (N=50) can be RMSD reproduced.



**Table 4-2.** Reproduction rate of *de novo* design by Tanimoto cutoff in 50 systems tested in SB2012.

<b>a. DOCK scoring protocol</b>	<b>b. sample size</b>	<b>c1. # Tanimoto reproduction (1.0)</b>	<b>c2. # Tanimoto reproduction (0.95)</b>	<b>c3. # Tanimoto reproduction (0.8)</b>	<b>c4. # Tanimoto reproduction (0.7)</b>	<b>c5. # Tanimoto reproduction (0.6)</b>
<b>SGE</b>	2500	36	36	45	45	47
<b>FMS</b>	4595	45	45	48	49	50
<b>FMS+SGE</b>	2752	37	38	43	44	48
<b>SGE+Tanimoto</b>	3119	45	45	47	47	49
<b>FMS+Tanimoto</b>	5430	47	48	48	48	49
<b>FMS+SGE+Tanimoto</b>	3642	45	46	46	47	48

*de novo* Tanimoto reproduction defined as creating molecules with Tanimoto=1.0 to reference molecule.

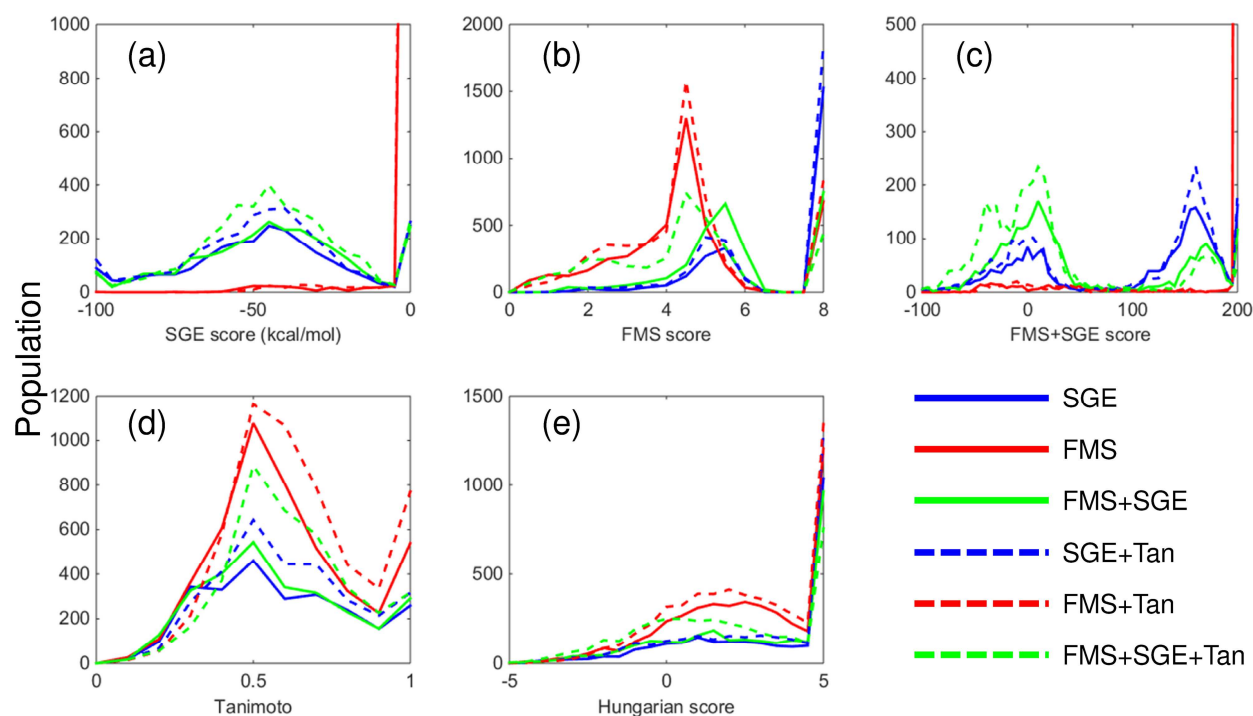
**Table 4-3.** Reproduction rate of *de novo* design by Hungarian RMSD cutoff in 50 systems tested in SB2012.

<b>a. DOCK scoring protocol</b>	<b>b. sample size</b>	<b>c1. # Tanimoto reproduction (1.0)</b>	<b>d1. # RMSD reproduction (2 Å)</b>	<b>d2. # RMSD reproduction (2.5 Å)</b>	<b>d3. # RMSD reproduction (3 Å)</b>
<b>SGE</b>	2500	36	14	19	23
<b>FMS</b>	4595	45	10	15	23
<b>FMS+SGE</b>	2752	37	17	22	23
<b>SGE+Tanimoto</b>	3119	45	15	19	22
<b>FMS+Tanimoto</b>	5430	47	13	21	24
<b>FMS+SGE+Tanimoto</b>	3642	45	26	29	33

All experiments in this table employ a Tanimoto cutoff of 1.0 and the listed Hungarian RMSD cutoff.

### 4.3.2 Focused library *de novo* runs: DOCK outcomes and molecular properties

*Total ensemble properties.* Histograms of SGE, FMS, and FMS+SGE, Tanimoto and Hungarian (RMSD) scores for all *de novo* output molecules for the 50 SB2012 systems are shown Figure 4-2 to examine overall global trends. As the current FMS protocol does not penalize clashes between a candidate ligand and the receptor, FMS and FMS+Tanimoto generated molecules yield mostly energetically unfavorable molecules (Figure 4-2a, red solid and red dashed lines) when evaluated in the contact of the receptor (*i.e.* SGE score). The FMS+SGE guided *de novo* protocol yielded similar energy profiles to the SGE protocol (Figure 4-2a, green solid and green dashed lines vs. blue solid and blue dashed lines). The FMS and FMS+Tanimoto protocol yielded the best pharmacophore overlap with the reference, with the FMS score peak around FMS=4.2 (Figure 4-2b, red solid and dashed lines) while SGE and SGE+Tanimoto yielded mostly molecules with little pharmacophore overlaps (Figure 4-2b, peak at FMS=8 blue solid and dashed lines). The FMS+SGE and FMS+SGE+Tanimoto FMS histogram fell in between the SGE guided protocol and the FMS guided protocol.



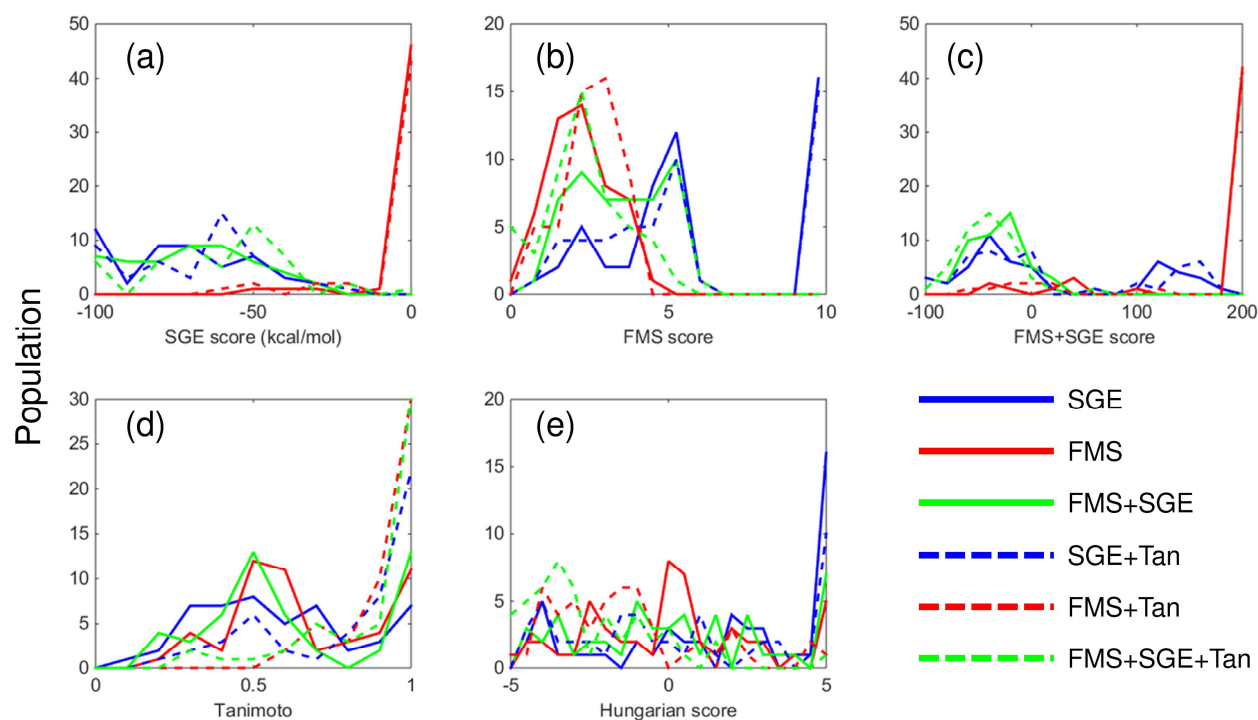
**Figure 4-2.** Histogram of (a) SGE, (b) FMS, (c) FMS+SGE, (d) Tanimoto and (e) Hungarian Score for all six *de novo* experiments on all 50 systems. Tanimoto=1.0 stands for perfect Tanimoto overlap. Molecular properties and DOCK scores of ensembles generated with SGE (blue solid line), FMS (red solid line), FMS+SGE (green solid line), SGE+Tanimoto (blue dashed line), FMS+Tanimoto (red dashed line), and FMS+SGE+Tanimoto (green dashed line)

Interestingly, by adding the Tanimoto term, FMS+SGE+Tanimoto yielded a notably enhanced FMS distribution (Figure 4-2b green dashed line) compared to FMS+SGE protocol (Figure 4-2b green solid line). This result indicates that growth towards favorable Tanimoto space can in principle promote better pharmacophore overlap. In Figure 4-2c, all protocols showed two separate peaks, one shifted to the left indicating good overall scores; another shifted to the right indicating bad overall scores. The FMS+SGE guided protocols (green solid and dashed lines) contains more molecules in the left peak while the FMS guided protocols (red solid and dashed lines) contains mostly molecules in the right peak. This is consistent with the energetic clashes of FMS-only ensembles observed in the SGE histogram in Figure 4-2a. For FMS+SGE and SGE

guided protocols, the right peak (FMS+SGE>100) likely represents molecules with dominating bad FMS scores of 20 (with a score penalty contribution of  $5 \times 20 = 100$  to FMS+SGE score), indicating no pharmacophore overlaps between the generated molecule and the crystal ligand pose. The left shifted peaks contain molecules with both good pharmacophore overlap to the reference ligands and favorable energetic affinity to the target proteins. Tanimoto histograms (Figure 4-2d) for all six protocols all peak at around 0.5, while those driven with Tanimoto (dashed lines) are slightly shifted towards an improved Tanimoto score. Overall, the FMS and FMS+Tanimoto protocols generated the most ensembles with Tanimoto of 1.0 to the reference ligand (Figure 4-2d, peaks of red lines at Tanimoto=1.0). This explains the high Tanimoto reproduction rate of the FMS guided protocol shown in Table 4-2. Finally, Figure 4-2e showed the Hungarian score where a perfect overlap of two poses of the same molecule would yield a score of -5. Very few molecules have perfect Hungarian scores in the total ensembles.

***Best-scored molecule properties.*** Figure 4-3 shows the same results from Table 4-3 but when only the best-scored *de novo* grown molecule is retained instead of the entire ensemble across the 50 SB2012 test systems. Notably, top-scored molecules (Figure 4-3) for all six protocols bear improved scores compared to the total ensembles (Figure 4-2). For example, a new peak ranging from 1 to 3 in the FMS histogram (Figure 4-3, blue solid line) of the top scored molecules obtained using the SGE protocol indicate an increased population of compounds with good pharmacophore overlaps. In fact, FMS score histograms for all six protocols showed significant improvement for the best-scored molecule sets (Figure 4-3b vs. 4-2b). Best-scored poses also showed great improvement in terms of the Hungarian score with more diverse distributions shifted towards lower scores (Figure 4-3e vs. 4-2e). Overall the FMS+SGE+Tanimoto ensemble (green dashed line)

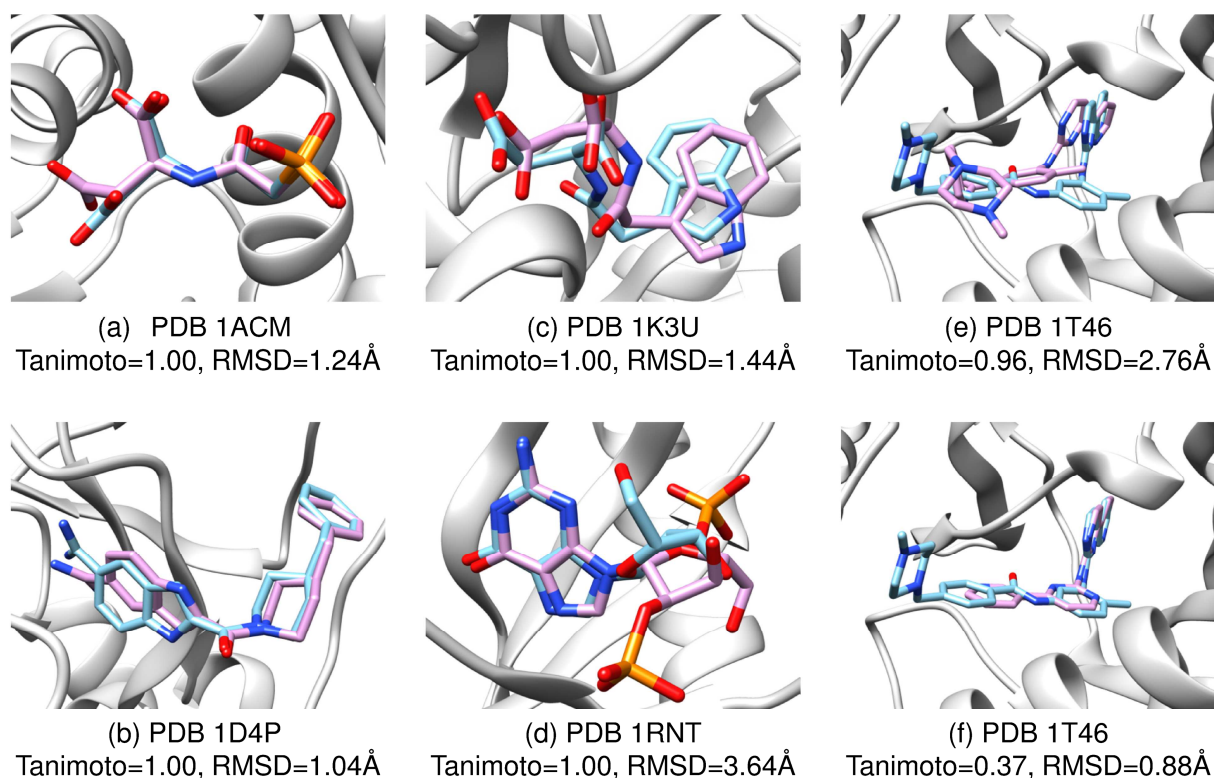
contain the most closely overlapped molecules indicated by the largest peak near the perfect Hungarian score of -5 in Figure 4-3e. This supports the RMSD reproduction results shown in Table 4-3 that FMS+SGE+Tanimoto consistently yields the highest RMSD reproduction rates.



**Figure 4-3.** Histogram of (a) SGE, (b) FMS, (c) FMS+SGE, (d) Tanimoto and (e) Hungarian Score for all six *de novo* experiments on the best scored molecule for each of the 50 systems. Molecular properties and DOCK scores of ensembles generated with SGE (blue solid line), FMS (red solid line), FMS+SGE (green solid line), SGE+Tanimoto (blue dashed line), FMS+Tanimoto (red dashed line), and FMS+SGE+Tanimoto (green dashed line).

**Representative structural analysis.** In Figure 4-4, six representative *de novo* generated molecules across five systems (PDB codes 1ACM, 1D4P, 1K3U, 1RNT and 1T46) from the 50 SB2012 test cases using FMS+SGE+Tanimoto protocol are illustrated. Systems 1ACM, 1D4P and 1K3U are RMSD reproduced cases where the best (in terms of Tanimoto and RMSD) grown molecules (Figure 4-4 a-c) have a Tanimoto of 1.0 and Hungarian RMSD value within 2Å relative

to the crystal ligand. The poses (carbon atoms shown in magenta) all visually overlap well with the crystal ligand (carbon atoms shown in cyan). System 1RNT is a Tanimoto reproduced case but not RMSD reproduced. Here, the molecule shown in Figure 4-4d is identical to the crystal ligand in term of 2D fingerprint (Tanimoto=1.0). However, the best Hungarian RMSD for the Tanimoto reproduced molecules is larger than 2 Å (RMSD=3.64 Å for the molecule shown). Visually, this pose has reasonable FMS overlap (FMS=2.90) and good geometrical overlap with the reference except for the sulfate group to the right side of the pocket. In contrast, for system 1T46, although the obtained Tanimoto value of 0.96 for the pose in Figure 4-4e is nearly identical to 1.0, it yielded a somewhat poor RMSD of 2.76 Å relative to the crystal ligand. Figure 4-4f also shows an example in which the best Hungarian RMSD (0.88 Å) for 1T46 (for only the matched part of the molecule with Tanimoto<1.0) yielded a poor Tanimoto of 0.37. Encouragingly, many molecules generated with a function combining the standard DOCK scores with FMS achieved energetic fits in the target pocket and had enhanced overlap to the reference. Particularly, use of FMS+SGE+Tanimoto appeared to yield the most well behaved molecule sets in these very focused *de novo* validation tests. It thus shows potential as a protocol to be used in *de novo* design although additional tests are needed.



**Figure 4-4.** Preliminary tests using *de novo* DOCK to 50 SB2012 targets with FMS-guided growth. Protein backbone is shown in tan ribbons; crystal reference molecule in cyan; *de novo* DOCK generated molecule in magenta. PDB IDs, Tanimoto and Hungarian RMSD values of the molecules (a)-(f) are provided.

#### 4.3.3 Focused *de novo* tests: HIVgp41 hydrophobic pocket

The focused *de novo* protocol using SGE, FMS, FMS+SGE, SGE+Tanimoto, FMS+Tanimoto and FMS+SGE+Tanimoto score functions were also applied to the HIVgp41 hydrophobic pocket and inner pocket target systems as a preliminary application test. Here, the focused fragment libraries were generated from the native peptide inhibitors shown in Figure 3-2b (orange stick representation) and Figure 3-2c (cyan wire representation). These peptide inhibitors were also used as pharmacophore reference for FMS guided growth. The focused fragment library generated for the inner pocket (1AIK-IQLT) includes 11 fragments in total; the focused fragment library generated for the hydrophobic pocket (1AIK-WWDI) includes 10 fragments in total.

Compared to the virtual screening protocol used in the Rizzo lab as described in the study by Holden *et al*<sup>72</sup> targeting HIVgp41, the sampling procedure and computational complexity in *de novo* design are quite different. Here, it generally takes about 2~7 seconds to sample one new molecule (i.e. generate a fully/partially grown molecule in one growth layer). Total sampling size, final hit (output fully grown molecules) ensemble size and run time (total time for each complete *de novo* experiment and average time for each sampling event) are reported in Table 4-5. Note that for focused *de novo* design targeting the HIVgp41 hydrophobic pocket (1AIK-WWDI), the total time exceeded the wall clock time, which was set to 120 hours. The experiment was terminated at growth layer 6. Thus the sampling size and run time are not shown. Overall, the average run time for targeting the inner pocket (2~5 seconds) is much less than that for targeting the hydrophobic pocket (4~7 seconds). And the sampling size for targeting the inner pocket is much larger except for the case using the FMS+Tanimoto scoring function. On average, about 200~500 fully grown molecules were output after approximately 7~75 hours (except for FMS-guided *de novo* design to the hydrophobic pocket). It is important to note that the sampling sizes and run time variability will depend on the properties of the fragments such as molecular size, number of attachment points, number of allowable bonds at each attachment point, and environment. To address these challenges, evolutionary sampling procedures (*i.e.* genetic algorithm) and refined generic fragment libraries are also being tested in the Rizzo lab to help improve sampling and increase efficiency of the *de novo* protocol in DOCK.



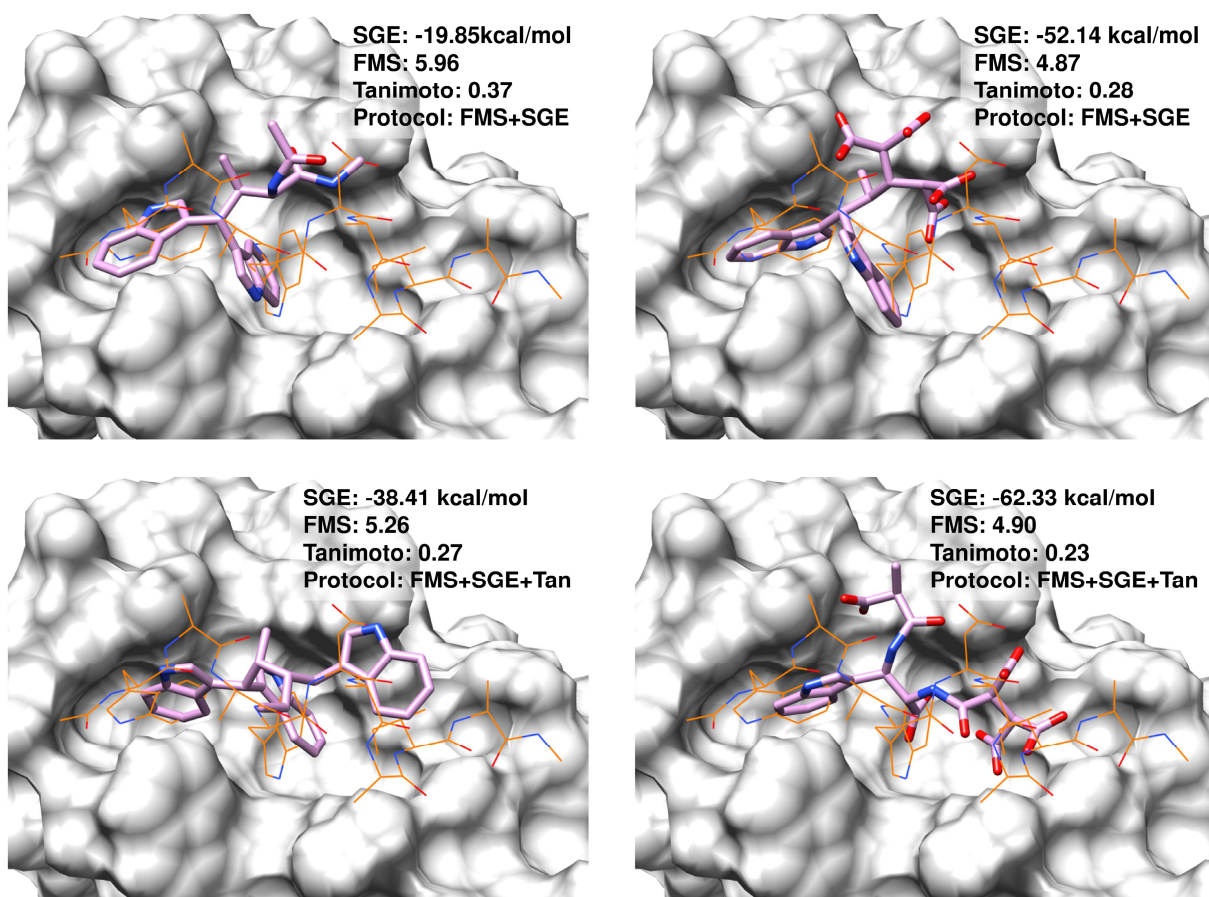
**Table 4-4.** *De novo* design results for HIVgp41 hydrophobic and inner pocket.

		SGE	FMS	FMS +SGE	SGE +Tan	FMS +Tan	FMS+ SGE+Tan
<b>Sampling size</b>	<b>1AIK-WWDI</b>	14878	--	14783	10107	17862	9159
	<b>1AIK-IQLT</b>	34204	89520	31561	39794	11862	13196
<b>Final hit size</b>	<b>1AIK-WWDI</b>	237	119	201	238	282	230
	<b>1AIK-IQLT</b>	229	231	214	323	484	501
<b>Run time (seconds)</b>	<b>1AIK-WWDI</b>	90696	--	100572	51171	118691	60229
	<b>1AIK-IQLT</b>	107154	270788	145859	171379	25513	37302
<b>Run time (hours)</b>	<b>1AIK-WWDI</b>	25.19	--	27.94	14.21	32.97	16.73
	<b>1AIK-IQLT</b>	29.77	75.22	40.52	47.61	7.09	10.36
<b>Avg. time (seconds)</b>	<b>1AIK-WWDI</b>	6.10	--	6.80	5.06	6.64	6.58
	<b>1AIK-IQLT</b>	3.13	3.02	4.62	4.31	2.15	2.83

1AIK-WWDI: hydrophobic pocket; 1AIK-IQLT: inner pocket. Average run time is calculated based on sampling size.

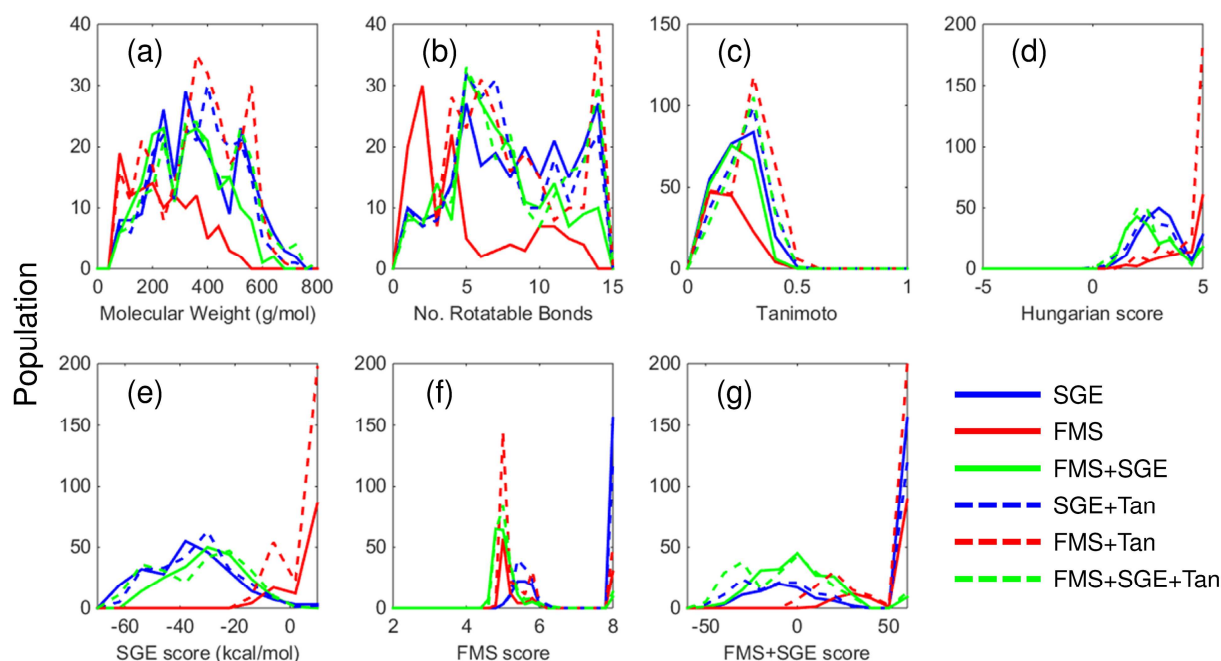
Figure 4-5 shows four representative compounds targeting the HIVgp41 hydrophobic pocket designed by the *de novo* DOCK protocol guided with FMS+SGE (top row) and FMS+SGE+Tanimoto (bottom row) scoring functions. All four molecules yield favorable binding energies ranging from -62 kcal/mol to -20 kcal/mol, which are comparable to top SGE scores obtained in a virtual screen to this pocket (peak at around -50 kcal/mol in Figure 2-19a, blue curve). Encouragingly, several of these representative molecules contain indole rings that overlap with the Trp117 position on the left side of the binding pocket (Figure 4-5) as well as a charged acid group positioned roughly near the Asp121 residue of the reference. Particularly, the molecule on the top right also contains a secondary indole ring that overlaps with the Trp120. While the synthetic feasibility of these molecules is yet to be determined, their binding poses are quite unique and not yet observed in the virtual screening results. An unusual occurrence that needs more investigation

is the presences of compounds with multiple acid groups and large magnitudes of net charges (compound with net formal charge of -4 in Figure 4-5 top right and -5 in Figure 4-5 bottom right). As expected, these two negatively charged molecules yielded much more favorable energy scores (SGE=-62.33 ~-52.14 kcal/mol) compared to the two neutral molecules (Figure 4-5 top left and bottom left, SGE=-38.41~-19.85 kcal/mol). This is consistent with the fact that the crystal receptor carries a positive net formal charge of +6. In practice, however, restraints in the total charges should be assigned in the *de novo* protocol to yield more drug-like molecules.



**Figure 4-5.** Representative results from *de novo* growth targeting the HIVgp41 hydrophobic pocket. The pharmacophore reference is the extended peptide including WWDI key residue side chains (wire representation, C<sub>α</sub> shown in orange).

Molecular properties and DOCK scores of all the *de novo* generated molecules targeting the HIVgp41 hydrophobic pocket using all six scoring functions are shown in Figure 4-6. Molecules generated using the FMS-alone are in general smaller in size (Figure 4-6a-b, red solid lines). Interestingly, use of FMS+Tanimoto score instead of FMS alone seemed to yield the best FMS score results as shown in Figure 4-6f red solid and dashed lines. However, as the ensemble size for each *de novo* tests could vary (Table 4-2 and 4-3, column b), it is likely that the enhanced peak for FMS+Tanimoto is partially due to its larger sampling size (5430) compared to FMS alone (4595). Future studies should enforce a “common” ensemble size to facilitate comparisons. In contrast to virtual screening when sampling of a single molecule is performed with energy score and the six scoring functions are used for rescoring only; in *de novo* design, the sampling is heavily influenced by the given scoring function. Thus the enhancement in FMS score for FMS+Tanimoto protocol demonstrated the synergy of FMS and Tanimoto scores in *de novo* sampling.



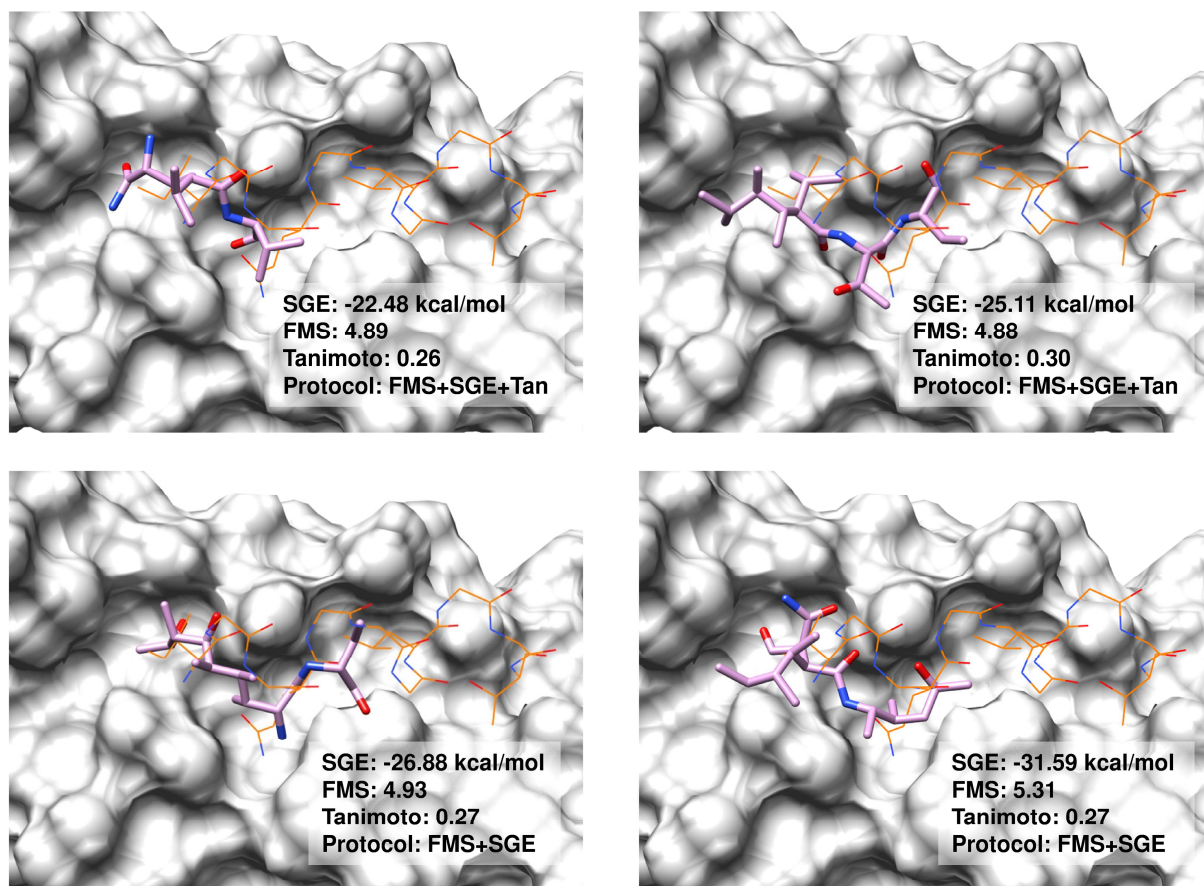
**Figure 4-6.** Histogram of (a) molecular weight, (b) number of rotatable bonds, (c) Tanimoto, (d) Hungarian Score, (e) SGE, (f) FMS and (g) FMS+SGE for *de novo* DOCK generated molecules targeting HIVgp41 hydrophobic pockets. Molecular properties and DOCK scores of ensembles generated with SGE (blue solid line), FMS (red solid line), FMS+SGE (green solid line), SGE+Tanimoto (blue dashed line), FMS+Tanimoto (red dashed line), and FMS+SGE+Tanimoto (green dashed line).

Although some aspect of these *de novo* test results are encouraging (i.e. overlap to indole and acid groups in the reference), it is important to emphasize that the focused *de novo* test for reproducing the peptide reference in these computational experiments were not nearly as successfully as that for the small molecule test cases discussed earlier in the section (Table 4-2, 4-3 and Figure 4-3). Focusing on the Tanimoto histogram (Figure 4-6c) and the Hungarian score histogram (Figure 4-6d), none of the six protocols could regenerate the peptide in terms of Tanimoto (perfect Tanimoto score:1.0) or Hungarian score (perfect Hungarian score: -5, corresponding to Tanimoto =1.0 and RMSD = 0 Å). One of the reasons for this result is the fact that rebuilding a peptide reference from fragments is a far more complicated problem. Not only are the peptides much larger than the small molecules (< 500 g/mol in molecular weight, <7 rotatable

bonds) used in the initial test case, but the helical conformation of the peptide taken from the crystal structure is part of a much larger protein helix. It may also be challenging to reproduce the exact 2D fingerprint and 3D conformation of the peptide with the limited number of *de novo* growth steps (up to layer 7). One potential modification to the current protocol would be to use the crystal pose of the fragments as the initial anchor placement instead of orienting the fragment from scratch before growth in layer 1. Additional future refinement to both the fragment library and sampling protocols can be explored to improve the reproduction rate for peptide reference guided *de novo* design.

#### **4.3.4 Focused *de novo* tests: HIVgp41 inner pocket**

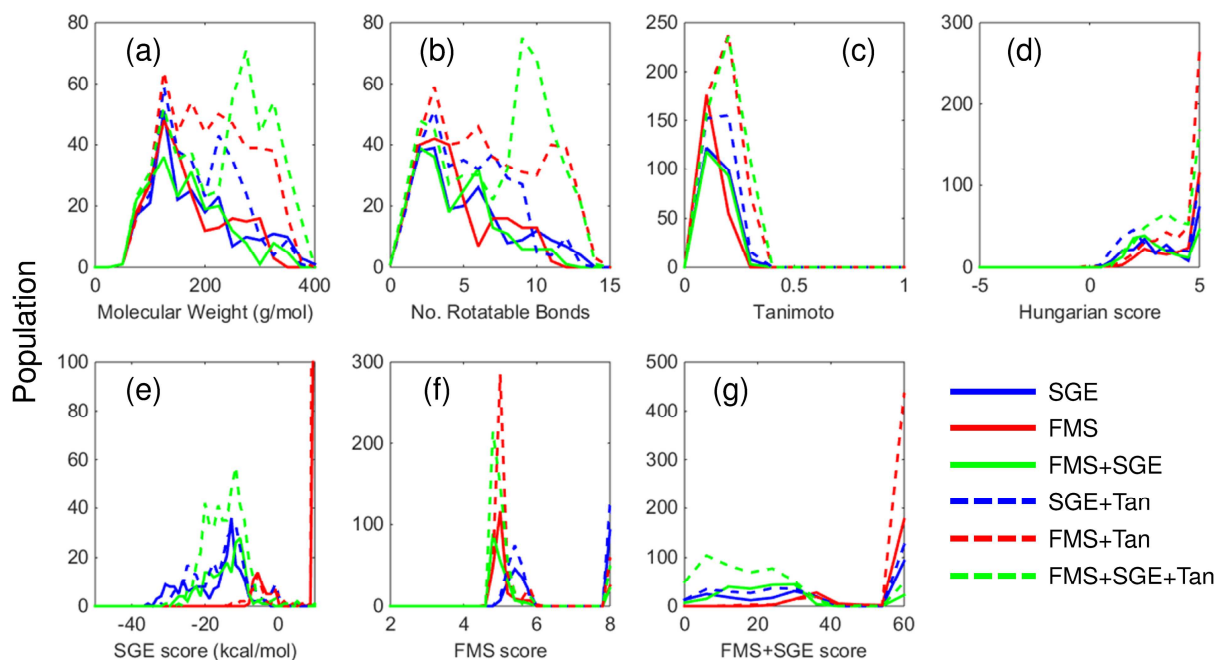
Figure 4-7 shows four representative hits targeting the alternative HIVgp41 inner pocket designed by the *de novo* protocol guided by FMS+SGE (bottom row) and FMS+SGE+Tanimoto (top row) scoring functions. Interestingly, hits from these preliminary focused *de novo* design experiments occupied only about half of the binding pocket. And, none of the hits contain ring structures as the crystal ligand used to generate the fragment library itself does not contain rings. Thus, the small sizes of the hits lead to less strong binding interaction energies between the ligands and binding site, which range from approximately -32 to -22 kcal/mol.



**Figure 4-7.** Representative results from *de novo* growth targeting the HIVgp41 inner pocket. Pharmacophore reference is the extended peptide including IQLT key residue side chains (wire representation, C<sub>α</sub> shown in orange).

As before, figure 4-7 shows the histograms of molecular properties and DOCK scores of all the *de novo* generated molecules targeting the HIVgp41 inner pocket with the six different scoring functions. The molecular sizes of the molecules designed in the inner pocket are significantly smaller compared to that of the hydrophobic pocket. While the molecular weights are as large as 400 g/mol and they peak at approximately 120 g/mol except for the FMS+Tanimoto and FMS+SGE+Tanimoto protocols (Figure 4-7a). Note that the reference peptide for the hydrophobic pocket contains 36 rotatable bonds (as determined by molecular modeling program MOE) and has a molecular weight of 1028.2 g/mol; the reference peptide for the inner pocket, contains 39 rotatable

bond (as determined by MOE) and has a molecular weight of 856.9 g/mol. Thus, the average molecular weight of each fragment in the focused fragment library for the inner pocket is much smaller. Most hits contain less than 5 rotatable bonds. This along with the visual inspection in Figure 4-7 suggests that more layers and more sampling needs to be added for the inner pocket with the focused *de novo* protocol to successfully reproduce molecules similar in size to the two reference peptides employed in these preliminary tests.



**Figure 4-8.** Histogram of (a) molecular weight, (b) number of rotatable bonds, (c) Tanimoto, (d) Hungarian Score, (e) SGE, (f) FMS and (g) FMS+SGE for *de novo* DOCK generated molecules targeting HIVgp41 inner pockets. Molecular properties and DOCK scores of ensembles generated with SGE (blue solid line), FMS (red solid line), FMS+SGE (green solid line), SGE+Tanimoto (blue dashed line), FMS+Tanimoto (red dashed line), and FMS+SGE+Tanimoto (green dashed line).

#### 4.3.5 Future experiment: generic library *de novo* tests.

While *de novo* DOCK should be capable of re-generating the crystal ligand given a very focused fragment library derived from the original molecule itself, the ultimate goal of *de novo*

design is to yield a diverse ensemble of novel molecules with desirable molecular properties and binding affinities to the target protein using a “generic” fragment library. Table 4-5 shows a series of generic libraries, pre-computed by Dr. William J. Allen using a ZINC drug-like compound set (*3\_t90.both.mol2*, with 205,792 molecules in total) currently being evaluated for different growth protocols. Using a frequency (times the fragment is seen in the original database) cutoff of 1, 10, 50, 100, 250, 500 and 1000, the number of fragments in these generic libraries can be, with additional studies, “tuned” so that a reasonable sized library (*i.e.* 250~500 fragments) can be used to re-generate a known compound (*i.e.* Tanimoto >0.90) in a timely manner (*i.e.* 1~2 days).

**Table 4-5.** List of generic fragment libraries.

<b>Library Name</b>	<b>Frequency cutoff</b>	<b># Scaffold</b>	<b># Linker</b>	<b># Sidechain</b>
<b>Fraglib_1</b>	1	1326	7295	21165
<b>Fraglib_2</b>	10	101	673	1746
<b>Fraglib_3</b>	50	26	213	520
<b>Fraglib_4</b>	100	15	145	300
<b>Fraglib_5</b>	250	11	76	143
<b>Fraglib_6</b>	500	8	40	88
<b>Fraglib_7</b>	1000	4	27	49

Libraries generated by Dr. William J. Allen.

## 4.4 Conclusion

In summary, in this Chapter we performed preliminary *de novo* design tests targeting the HIVgp41 hydrophobic and inner pockets. The focused *de novo* protocol has been validated by reproduction tests with 50 SB2012 systems. In most cases, the small molecule ligands, with a very small fragment library, can be reproduced in terms of Tanimoto overlap. And, reasonable structure overlap to the crystal poses of the ligands can be obtained. FMS is shown to be an effective scoring function not only works well in virtual screening but also now compatible with *de novo* design.



The ability to reproduce known poses is essential for the validation of *de novo* design protocols and can significantly improve the potency and feasibility of *de novo* designed molecules. Finally, when used in combination, synergies among FMS, SGE and Tanimoto scores can significantly improve the properties of the hits generated in *de novo* design.

Future studies with FMS-guided *de novo* design should include: (1) an evolutionary strategy to speed up the *de novo* sampling procedure; (2) using optimized generic library generation to improve hit properties; (3) gradually increase the weight on the matching residual term for FMS for each *de novo* growth layer so that partially grown molecules with more numbers of matches are favored compared to molecules with fewer exact matches.

## **Chapter 5. Quantitative Characterization of T20 Variants Affinity and Mutational effects**

Chapter 5 reports molecular dynamics simulation and free energy calculation results using thermodynamic integration and molecular footprinting to study the interactions between HIV fusion protein gp41 and peptide inhibitor T20.

### **Abstract**

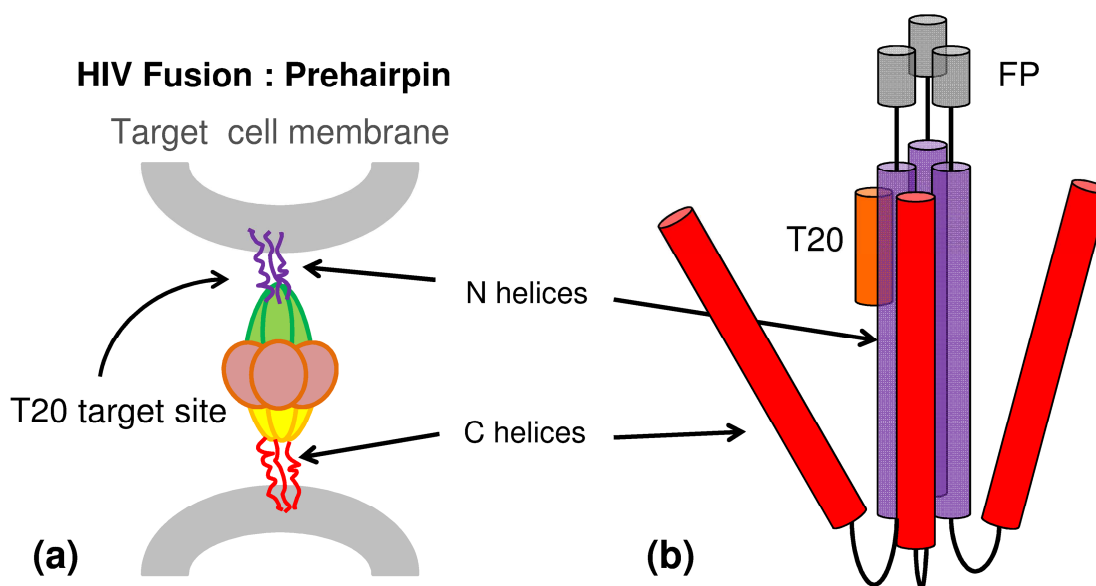
A critical step in the HIV life cycle is viral fusion, involving the binding of three C-helices in HIVgp41 to three N-helices to form a six-helical bundle.<sup>12</sup> Derived from the outer gp41 C-helix, the first FDA approved HIV fusion inhibitor T20 (enfuvirtide) competes with the native C-helices for binding to the viral protein N-helices<sup>12</sup> and effectively blocks viral replication in clinical use.<sup>18</sup> However, resistance to T20 can arise from primary mutations in the gp41 N-helices and secondary mutations in the gp41 C-helices which reduce affinity for T20 while retaining viral infectivity. This interesting observation underscores the importance of understanding the origins of the binding affinities and mutational effects in the HIVgp41-T20 complex system in order to design new fusion inhibitors that can overcome such drug resistance. In this study, we have successfully predicted the effect of both favorable and unfavorable mutations compared to experiments using all-atom molecular simulations and relative free energy calculations employing the thermodynamic integration (TI) method. In addition, several key interactions between specific residue pairs in T20 with gp41 N-helices has been identified which can help explain the underlying energetic and structural effects of primary mutations and secondary mutations for 19 variants of T20 with 3

different gp41 N-helices analogs. These computational results can be used to help guide design of potential peptide and small molecule HIV fusion inhibitors with improved binding profiles.

## 5.1 Introduction

As the world is entering the fourth decade of fighting the HIV/AIDS pandemic, the virus has been estimated to result in over 36 million deaths worldwide throughout the years.<sup>1,4,5</sup> Due in large part to the increased clinical use of antiretroviral therapy developed starting in the 1990s and introduced worldwide since the early 2000s, provided they can afford it, HIV-infected patients can expect a reasonable life-span with proper treatment.<sup>2,5</sup> A key challenge with treatment however is the development of drug resistant mutations that reduce drug potency. In addition, with 2.5 million new infections observed worldwide in 2011,<sup>1</sup> it is urgent for researchers to continue working towards more effective HIV treatment strategy with improved resistance profiles to prevent further spread of the epidemic. Among the various clinical strategies to block the HIV life cycle, targeting membrane fusion via compounds that bind to the glycoprotein gp41, is of great interest.<sup>9,12,13,72,120,135,136</sup> Gp41 plays an important role in facilitating the HIV fusion process through bringing the viral and host cell membranes close to each other. The end result is that the two membranes fuse together, resulting in a pore pathway that allows the viral core to enter the host cell. It is believed the early stages of this process include a pre-hairpin intermediate in which the three inner gp41 N-terminal helices (shown in purple helices/tubes in Figure 5-1) become inserted into the target cell membrane while the C-terminal helices (shown in red helices/tubes in Figure 5-1) remain attached to the viral membrane. Later the C helices will bend over and bind to the N helices and form a coiled-coil hairpin known as the six-helical bundle (6HB).<sup>137,138</sup> Notably, at this pre-hairpin stage, the N helical surfaces are exposed and susceptible to fusion inhibitors that block

formation of the 6HB and prevent fusion.<sup>11,12</sup> The first and for now the only FDA-approved HIV fusion inhibitor T20 has been designed with this strategy.<sup>18,139,140</sup> Derived from the outer C-helical sequence of the native virus, this 36 amino acid (residue 127 to residue 162, shown as orange tube in Figure 5-1b) peptide has shown clinical efficacy in controlling viral load although its use also results in drug resistance. While it is believed that T20 binds to N-HR during the prehairpin stage of HIV fusion, currently, there is no complete T20-gp41 crystal structure available. However, a well-validated computational T20-gp41 complex model developed in the Rizzo lab as reported by McGillick *et al.*<sup>16</sup> is available to help characterize T20 binding profiles. Importantly the model is in good agreement with an experimental crystal structure subsequently reported by Buzon *et al.*<sup>141</sup>



**Figure 5-1.** (a) Prehairpin stage in HIV fusion.<sup>12</sup> The three N-helices (show in purple helices) formed a trimer inserted into target cell membrane while the three C-helices (show in red) are yet to bind. (b) Illustration of positional alignment of T20 (orange tubes) to HIV CHR (red tubes).

As a RNA virus, HIV has a tendency to mutate and can easily develop drug resistance.<sup>142</sup> For gp41, one important example is the development of T20-resistant mutations at positions V38A or N43D (gp41 sequencing) on the gp41 N-terminal region. Considering the fact that the T20

sequence is identical to the native virus, for the mutated virus to confine function, a secondary mutation S138A is also observed on the C-terminal of gp41 that restores binding to the mutated N-terminal. Thus, T20 can be modified correspondingly to inhibit the mutated virus. Izumi *et al*<sup>18</sup> has designed an experiment to approximate the change of binding affinity across a matrix of gp41 recombinant systems formed by 19 T20-derived peptides (S138X in Table 5-1) and several gp41 N-helical mutants. Table 5-1 shows binding affinities estimated from the original experimental EC<sub>50</sub> data. Importantly, different single-point mutation pairs show dramatic differences in binding affinity, which can be used to help provide biological clues as to the origins in T20-gp41 binding and fusion inhibition. For example, the clinical observed primary mutation N43D reduces the binding affinity of T20 by almost 2 kcal/mol (-11.75 to -9.97 kcal/mol) relative to the wild-type receptor. And the secondary compensatory mutation S138A on T20 restores affinity to almost the wild-type level (-11.53 kcal/mol).

**Table 5-1.** Binding energy of T20 targeting HIVgp41 calculated from experimental EC<sub>50</sub> values obtained from Izumi *et al.*<sup>18</sup>

Receptors	HIV-1 <sub>WT</sub>	HIV-1 <sub>V38A</sub>	HIV-1 <sub>N43D</sub>
Ligands	$\Delta G_{\text{bind}}$ (kcal/mol)	$\Delta G_{\text{bind}}$ (kcal/mol)	$\Delta G_{\text{bind}}$ (kcal/mol)
T20 <sub>S138S</sub>	-11.75	-10.41	-9.97
T20 <sub>S138A</sub>	-12.57	-11.51	-11.53
T20 <sub>S138D</sub>	-9.10	>-8.18	>-8.18
T20 <sub>S138E</sub>	-8.93	>-8.18	>-8.18
T20 <sub>S138F</sub>	-10.94	-9.12	-8.73
T20 <sub>S138G</sub>	-12.11	-9.80	-9.34
T20 <sub>S138H</sub>	-9.10	>-8.18	>-8.18
T20 <sub>S138I</sub>	-12.68	-11.33	-11.64
T20 <sub>S138K</sub>	-8.38	>-8.18	>-8.18
T20 <sub>S138L</sub>	-12.48	-10.75	-11.64
T20 <sub>S138M</sub>	-12.48	-11.39	-11.96
T20 <sub>S138N</sub>	-10.53	>-8.18	>-8.18
T20 <sub>S138P</sub>	-8.66	>-8.18	>-8.18
T20 <sub>S138Q</sub>	-10.18	>-8.18	>-8.18
T20 <sub>S138R</sub>	-8.78	>-8.18	>-8.18
T20 <sub>S138T</sub>	-12.33	-10.10	-9.26
T20 <sub>S138V</sub>	-12.81	-10.24	-10.44
T20 <sub>S138W</sub>	-10.28	>-8.18	>-8.18
T20 <sub>S138Y</sub>	-10.36	-8.57	>-8.18

Binding energies  $\Delta G_{\text{bind}}$  are estimated using the equation  $\Delta G_{\text{bind}} = RT \cdot \ln(\text{EC}_{50})$  in kcal/mol at 298.15K using experimentally evaluated EC<sub>50</sub> values from Table 1-2 in Chapter 1.<sup>13</sup>

In this chapter, we present a computational study to evaluate T20-gp41 binding affinities and determine the biological effects of T20-resistant primary mutations V38A and N43D as well as compensatory mutations on T20 analogs. Our hypothesis is that, since the change of affinities among the mutants are derived from single-point mutations, there should be a subset of key residues in the binding pocket or close to the mutated regions that dominate changes in antiviral activity. An examination of residues was performed using per-residue heatmaps based on molecular dynamics simulations of different T20-gp41 complex systems at each endstate. Concurrently, atomic-level molecular dynamics simulations were performed using the thermodynamic integration (TI) method

in an effort to reproduce the clinical observations through direct comparison to the experimental binding data shown in Table 5-1. The goal of this study is to use such information to help guide the development of new HIV fusion inhibitors, either peptides or small molecules, using complementary computational approaches such as pharmacophore modeling (discussed in Chapter 2-4)<sup>137</sup> and *de novo* design.

## 5.2 Theoretical Methods and Computational Details

### 5.2.1 Free Energy Calculations Using Thermodynamic Integration

Thermodynamic integration (TI) methods calculate energy changes for nonphysical transformation processes (See Chapter 1, section 1.4 Molecular Dynamics and Free Energy Calculation). By artificially simulating the transition from one ligand to the other in both the bound and unbound states, one can obtain the potential energy differences of the processes, which are defined as the relative free energy of binding. Importantly, using thermodynamic cycles shown in Figure 1-8, the computed relative binding energy  $\Delta\Delta G_{\text{bind}}$  obtained from differences in transformation energies can be directly compared to the differences in two absolute binding energies measured experimentally. Sufficient sampling, with carefully chosen intermediate windows can, in principle, yield very accurate TI predictions provided that the two states are similar. The transformation free energy between any two similar systems is computed as a coupled function of the endstate potential energies  $V_0$  and  $V_1$  with respect to a mixing parameter  $\lambda$ , which varies from 0 to 1 as introduced in Chapter 1, section 1.4. The potential energy in each window is a weighted combination between that of the two endstate-systems. A series of transition simulations for all  $\lambda$  windows will be performed. The soft-core potential mixing function, as implemented in

AMBER11 (See Chapter 1) is used with the parameters  $\alpha$  and  $\beta$  set to the default values in ( $\alpha=0.5$ ,  $\beta=12$ ).

### 5.2.2 Footprint Analysis: Energy Decomposition to Uncover Ligand Binding Profile

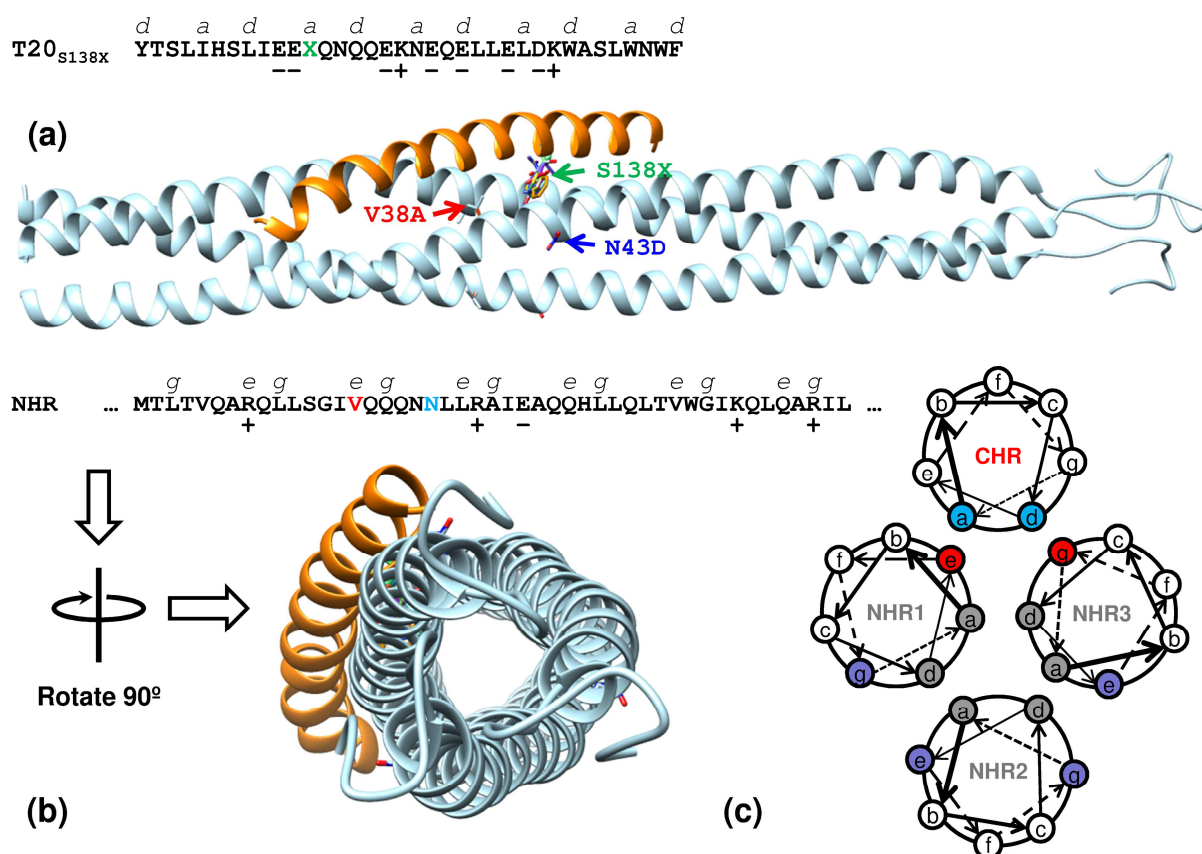
In addition to the TI simulations, standard molecular dynamics simulations for each T20-gp41 complex were also performed at the endstates. The resulting dynamics as well as molecular footprint signatures were analyzed to provide insight into specific protein-ligand interactions.<sup>16,71,72</sup> For each complex system, a per-residue energy decomposition based on either the protein receptor or the peptide ligand yielded what is called a footprint, essentially an energy density string of residue numbers. In this study, both van der Waals (VDW) and electrostatic (ES) footprints were computed which can be plotted as an interaction energy matrix (heatmap) in an attempt to identify key residues and energy components most important for T20 binding. Besides MD, use of such energy signals can be an alternative way to identify hits in docking studies.<sup>71,80</sup>

### 5.2.3 Model Construction and MD Simulation Protocol

All-atom gp41-T20 structures were based on those developed by McGillick *et al.*,<sup>16</sup> which in turn were originally constructed based on a model reported by Caffery *et al.*<sup>143</sup> Briefly, the molecule was constructed from PDB entry 1IF3 and PDB structure 1ENV by superimposing the common regions and matching 1ENV to the correct gp41 sequence. The derived complex structure included the membrane-proximal and fusion peptide regions, which were modeled as  $\alpha$ -helical.<sup>16</sup> In the presented study, eighteen analogs of T20 (19 peptides in total) and three variants of the gp41 receptor were constructed using the McGillick model as a starting point by manually mutating the T20 residue 138 to 18 different natural amino acids and the gp41 residue 38 from VAL to ALA, or



the gp41 residue 43 from ASN to ASP using the program MOE.<sup>99</sup> The initial conformations were optimized and tested via energy minimization and equilibration using *sander* in AMBER11. The final model to be simulated using both the standard MD protocol and the TI protocol are shown in Figure 5-2.



**Figure 5-2.** (a) T20 interaction site with two primary mutations on the target peptide N43D (in cyan) and V38A (in red) and one secondary single-point mutations at residue 138 on CHR/T20 (letters a-e represent the residual position in the  $\alpha$ -helical secondary structure, symbol +/- highlights the charged residues); (b). Rotated (by 90°) view of the helical bundle formed by the gp41-T20 complex. (c) Corresponding wheel representation of CHR bound to NHR1, NHR2 and NHR3 (numbered 1-94 in NHR<sub>i</sub>, i=1,2,3 in gp41 sequencing).<sup>138</sup>

The coupled simulation for the thermodynamic integration MD simulation in AMBER11 required a defined “soft-core mask” which includes the regions to be transformed. The other atoms,

which are outside the mask, have to be identical in the two coupled systems through the whole simulation. In order to meet this requirement, the initial structures of the solute were fit to the same scaffold (as shown in Figure 5-2a), and an identical solvent box was assigned. For both the unbound and the bound states, the wild-type complex/ligand was solvated with TIP3P<sup>144</sup> waters via *tLeap*. Then the solvent box was saved as a separate pdb file and later shifted to match the original gas phase complex/ligand variant models. This way of assigning the water box enables all of the mutant systems to differ only in the mutated region. The size of the water box for the complex systems (320 residues) was  $59 \times 58 \times 173 \text{ \AA}^3$  (13879 TIP3P waters) and for the ligand systems (38 residues including capping groups on both ends) was  $42 \times 58 \times 82 \text{ \AA}^3$  (4774 TIP3P waters). The resulting starting conformations were used for all the molecular dynamic simulations in this study.

The systems were equilibrated with a periodic boundary in 9 steps including several minimization steps of 1000 cycles and short MD simulations (50ps each) before the production runs. The Partial Mesh Ewald (PME)<sup>145</sup> method is used for calculating the electrostatic energy in the periodic box. Equilibration started with a minimization and a MD simulation step with positional restraints on all the heavy atoms in the solute (restraint weight  $5.0 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$ ), followed by three minimization runs with the same restraint mask, but the restraint weight decreased to 2.0, 0.1 and, finally,  $0.05 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$ . Then, two additional MD simulations were performed with reduced restraint weight  $1.0$  and  $0.5 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$ , respectively. The final two equilibration steps had relaxed restraint masks where only the back bone heavy atoms were included. The restraint weights were set to  $0.1 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{\AA}^{-2}$ . Production runs were done after the equilibration with identical simulation protocol to the last equilibration runs for data collection. The reference conformation used for the restraints in each step was the last snapshot from the

previous step up to step 6. For steps subsequent, the last frame of step 5 was always used as the restraint reference.

All TI simulations were accomplished using *sander* module in AMBER11 with a 1fs time step. Throughout the NPT (constant number of atoms, constant pressure, and constant temperature) MD simulation, the temperature of the systems was kept to 298.15K regulated by the Langevin dynamics method with collision frequency of  $1\text{ps}^{-1}$ . System pressure was relaxed every 0.5ps. Default Lennard Jones and Coulombic parameters were used. The MD trajectory snapshots were saved every 1ps for further data analyses. Specifically for TIMD simulations, a soft-core mask was defined for the single mutation site at residue 138 on T20. A total number of 19  $\lambda$  windows ( $\lambda = 0.05, 0.10, \dots, 0.95$ ) were simulated for each transformation. Each of the data point on the  $dV/d\lambda$  plot represents results from nineteen 2ns-long TI runs with varying mixing parameter  $\lambda$ .

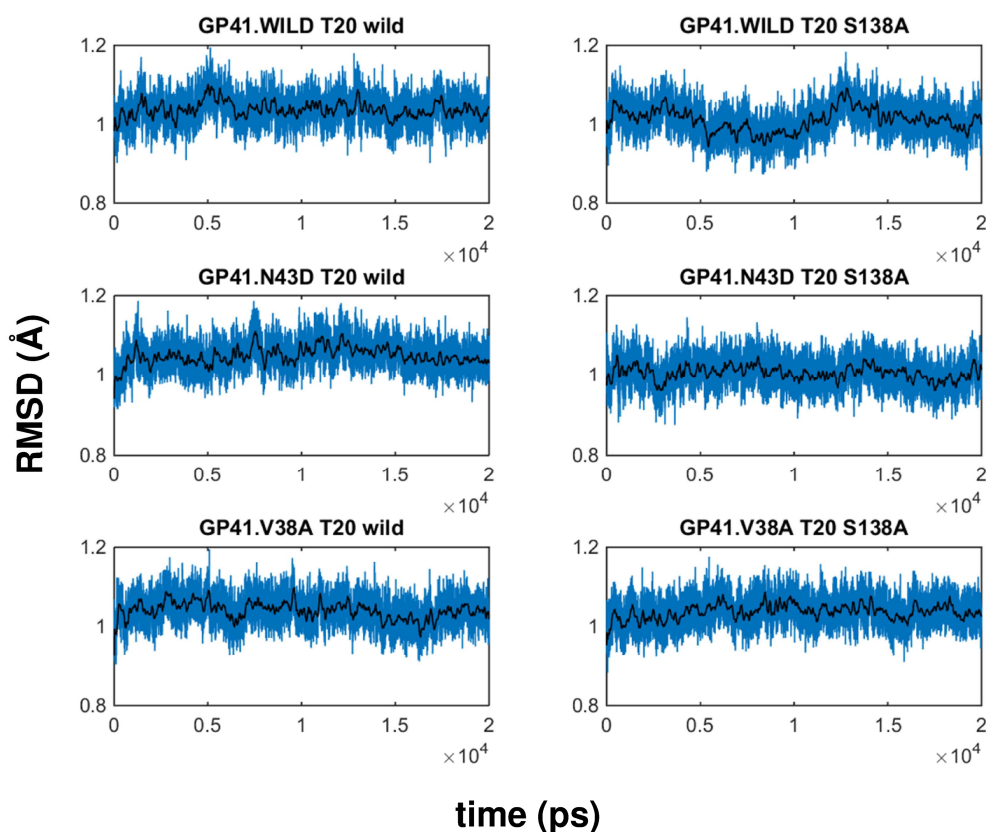
For extended endpoint MD simulations, *pmemd.cuda* in AMBER14 was used with GPU acceleration.<sup>49,56</sup> The longer 20ns endpoint simulations were done for six ligand-protein systems involved in clinical observed primary and secondary mutations (HIV<sub>WT</sub>-T20<sub>S138S</sub>, HIV<sub>WT</sub>-T20<sub>S138A</sub>, HIV<sub>V38A</sub>-T20<sub>S138S</sub>, HIV<sub>V38A</sub>-T20<sub>S138A</sub>, HIV<sub>N43D</sub>-T20<sub>S138S</sub>, HIV<sub>N43D</sub>-T20<sub>S138A</sub>). Longer endstate simulations help ensure improved convergence and accuracy in per-residue energetics. Snapshots of the endpoint simulations for every 1ps are saved for the structural and energetic analyses.

## 5.3 Results and Discussions

### 5.3.1 Endpoint Simulation Behavior: RMSD

To evaluate the behavior of the molecular models, stabilities of the endpoint systems for the bound states were gauged through monitoring RMSD vs. time. While stability of the wild type complex HIV<sub>WT</sub>-T20<sub>S138S</sub> was investigated in the original study by McGillick *et al.*,<sup>16</sup> the different

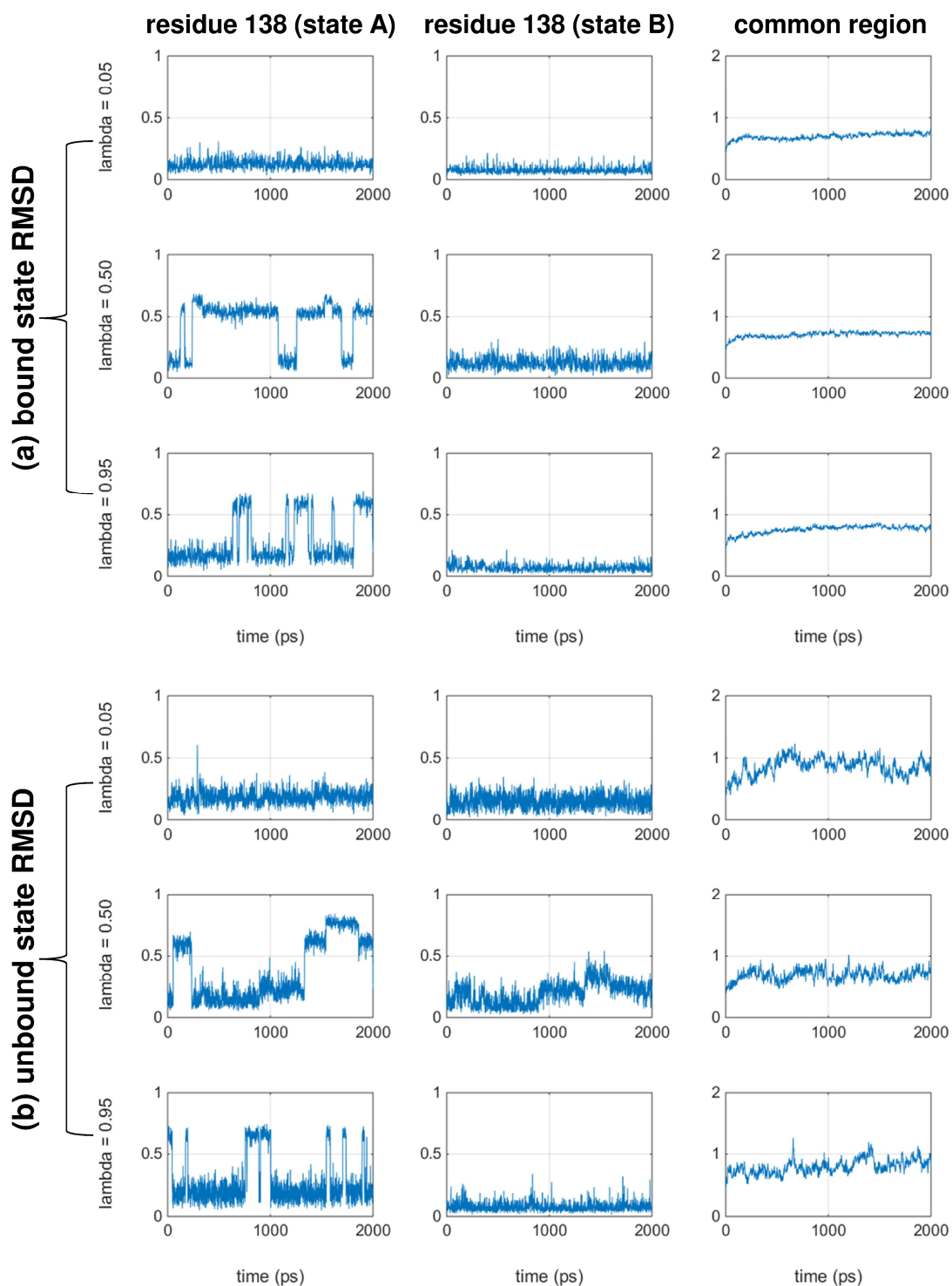
mutants studied here have not yet been examined. Figure 5-3 shows representative trajectories for six different systems (HIV<sub>WT</sub>-T20<sub>S138S</sub>, HIV<sub>WT</sub>-T20<sub>S138A</sub>, HIV<sub>V38A</sub>-T20<sub>S138S</sub>, HIV<sub>V38A</sub>-T20<sub>S138A</sub>, HIV<sub>N43D</sub>-T20<sub>S138S</sub>, HIV<sub>N43D</sub>-T20<sub>S138A</sub>, 20ns long each) with instantaneous results in blue and running averages (block size = 100ps) in black. Here, the RMSD values are computed for C<sub>α</sub> atoms in the complex structures with “fitting” to the initial constructed structure (the first frame of step 1 in the standard MD simulations). As expected, given the weak restraints employed, none of the restrained MD simulations showed significant backbone conformational deviations from the initial conformation with relatively low RMSD values (<1.2 Å). Thus we hypothesize that the ensemble of conformations of the stable complexes can be used to help decipher the binding profiles of T20 in terms of per-residue VDW and ES interaction patterns. Detailed analyses on the average molecular footprint and heatmap of binding and the associated error estimation are discussed below.



**Figure 5-3.** The RMSD plots of the six endpoint standard MD simulations of HIVgp41-T20 complex variants. RMSD values in Å. Time in picoseconds. Raw data in blue, running average shown in black.

### 5.3.2 TI Simulation Behavior: RMSD

We also performed RMSD calculations using the TIMD trajectories to evaluate simulation stability. Figure 5-4 illustrates the RMSD values of the mutating region (residue 138 on T20) in both the initial state A (T20<sub>S138S</sub>, first column) and the final state B (T20<sub>S138A</sub>, second column) as well as those for the common region (complex structures bound to HIV<sub>WT</sub> that are identical for both state A and B, third column) as a function of time step (in picosend). For the mutating region, all heavy atoms are used for RMSD calculation; for the common region, C<sub>α</sub> atoms are used. All results are computed for complex coordinates “fitted” to the last frame of MD equilibration runs (step 9 in the TIMD protocol). Representative windows of  $\lambda = 0.05, 0.50$  and  $0.95$  are plotted.

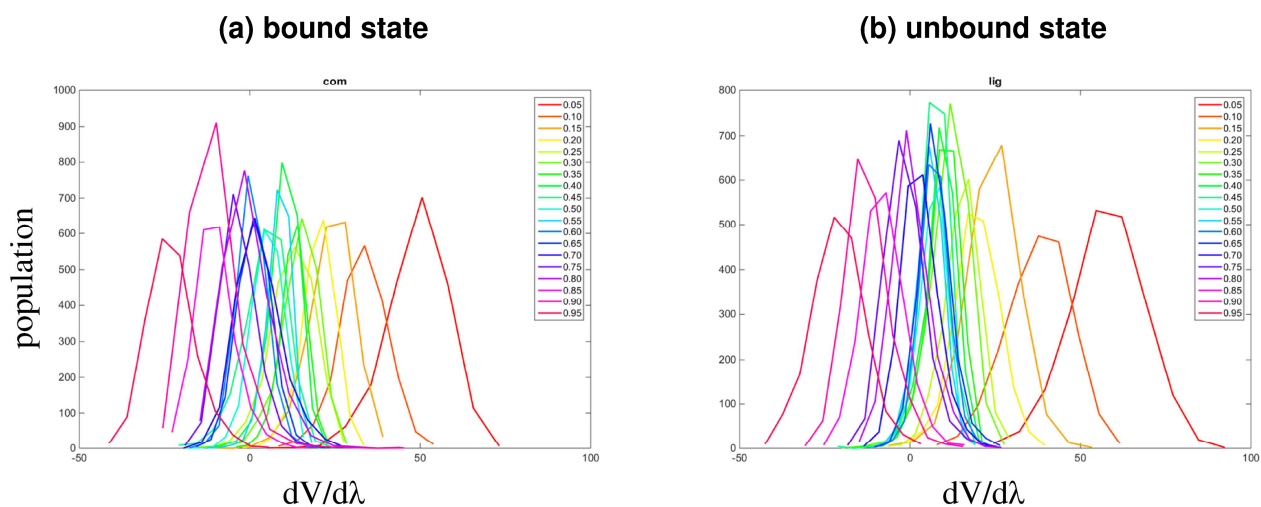


**Figure 5-4.** Representative RMSD plots for TIMD simulations of mutating HIV<sub>WT</sub>-T20<sub>S138S</sub> to HIV<sub>WT</sub>-T20<sub>S138A</sub> in (a) bound state and (b) unbound state. RMSD values in Å. Time in picosecond.

Again, in all cases, both the mutating region RMSD ( $<1 \text{ \AA}$ ) and common region RMSD ( $< 2 \text{ \AA}$ ) values are very low, likely a function of the weak energetic restraints on the protein and peptide backbones. This suggests the conformational differences between the two endstate T20 analogs in both the bound and unbound states are quite small, which is desirable for well-behaved TI simulations and transformation energy calculations. Interestingly for the RMSD of the mutating regions (first and second columns), although most of the TI runs yield extremely low RMSD value for the single residue (close to  $0 \text{ \AA}$ ), simulation at  $\lambda = 0.50$  and  $0.95$  for both the bound and unbound states yielded a subset of trajectories with RMSD values of about  $0.5 \text{ \AA}$ . This suggests that across different  $\lambda$  windows, the TI simulations likely sampled slightly different orientations of residue 138, which could contribute to the energetic calculations in the non-physical paths of transformation.

### 5.3.3 TI Simulation Behavior: $dV/d\lambda$

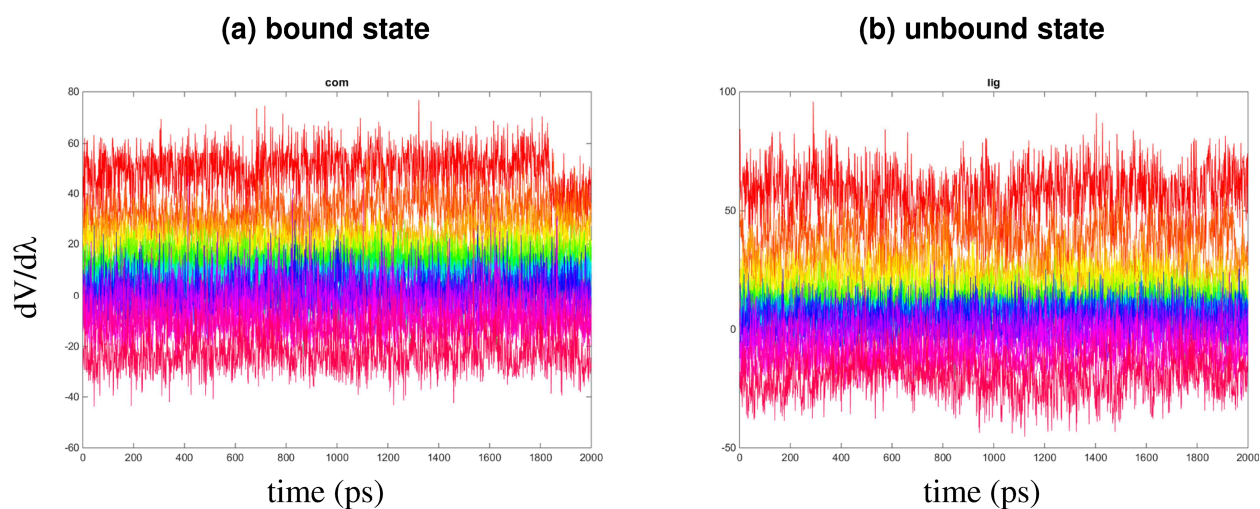
Another procedure to evaluate TIMD protocols are to inspect distributions of  $dV/d\lambda$  across different  $\lambda$  windows. Sufficient sampling across intermediate states, with a reasonable number of  $\lambda$  windows along the complete transformation path, should yield good overlap in terms of the  $dV/d\lambda$  values sampled for each pair of adjacent  $\lambda$  windows. Figure 5-5 shows the histograms of  $dV/d\lambda$  values for each of the 19  $\lambda$  windows (different colors represents different  $\lambda$  windows) of TIMD simulations of mutating T20<sub>S138S</sub> to T20<sub>S138A</sub> when bound to (Figure 5-5b) to HIV<sub>WT</sub> and in the unbound states. Overall, histograms for adjacent  $\lambda$  windows are well overlapped, especially for windows closer to  $\lambda=0.50$ . The total range of  $dV/d\lambda$  values of all 19 windows ( $-40\sim 70 \text{ kcal/mol}$  for bound state and  $-40\sim 90 \text{ kcal/mol}$  for unbound state) are continuously covered by all  $\lambda$  windows, suggesting that our choice of  $\Delta\lambda$  of  $0.05$  is reasonable.



**Figure 5-5.** The  $dV/d\lambda$  value histogram in all  $\lambda$  windows ( $\lambda = 0.05, \dots, 0.50, \dots, \text{and } 0.95$ , distinguished by color) simulations for mutating  $\text{HIV}_{\text{WT}}\text{-T20}_{\text{S138S}}$  to  $\text{HIV}_{\text{WT}}\text{-T20}_{\text{S138A}}$  in (a) bound state and (b) unbound state.

In addition, the stability of TIMD runs are illustrated by the  $dV/d\lambda$  plots as a function of time steps (in picosecond) for all  $\lambda$  windows in the bound and unbound states of  $\text{T20}_{\text{S138S}}$  to  $\text{T20}_{\text{S138A}}$  mutations (Figure 5-6). Corresponding to the good behavior in terms of RMSD fluctuation in Figure 5-4, all  $dV/d\lambda$  plots are consistently stable across all frames in the TIMD simulations. The gap in terms of  $dV/d\lambda$  values for adjacent  $\lambda$  windows are minimal for windows closer to  $\lambda=0.50$  and grow gradually as  $\lambda$  approaches the two physical endstate system when  $\lambda=0$  and 1.





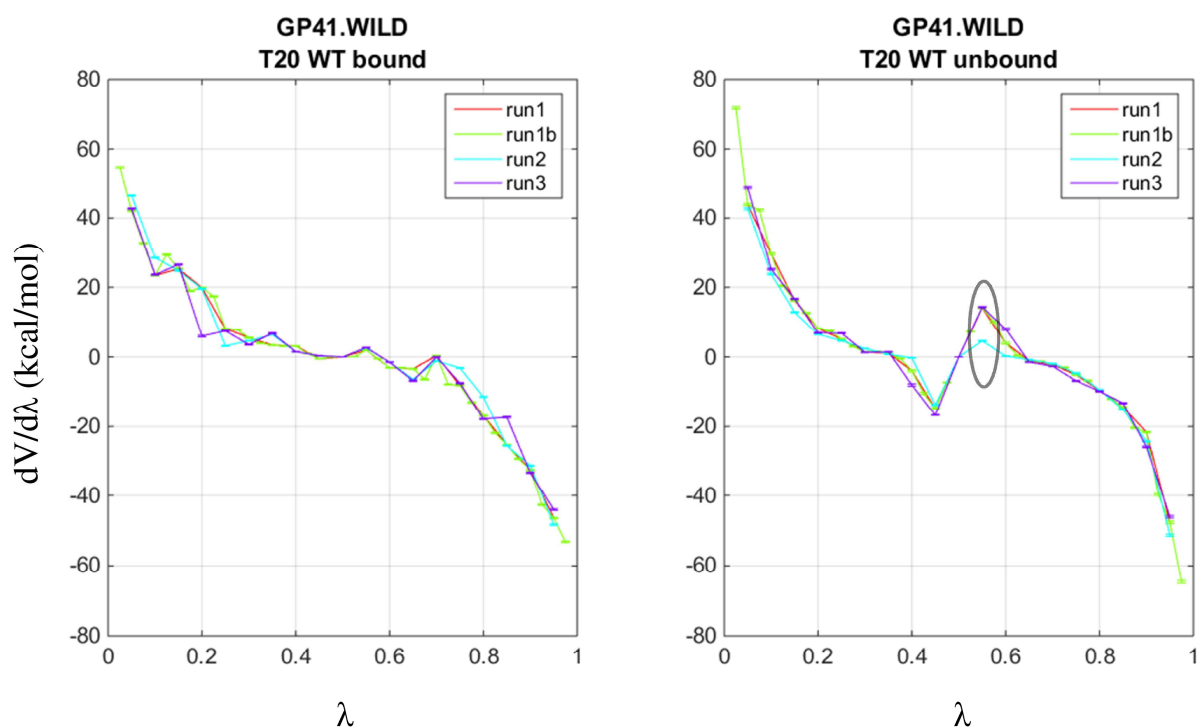
**Figure 5-6.** The  $dV/d\lambda$  plots with respect to time for all  $\lambda$  windows ( $\lambda = 0.05, \dots, 0.50, \dots, \text{and } 0.95$ , distinguished by color with identical color scheme as shown in Figure 5-5) simulations for mutating HIV<sub>WT</sub>-T20<sub>S138S</sub> to HIV<sub>WT</sub>-T20<sub>S138A</sub> in (a) bound state and (b) unbound state.

### 5.3.4 TI Simulation Behavior: Null Transformation

Comparison to experimental measurements is a common method to evaluate TI energy calculation accuracy. However, experimental values may also have associated noises related to the theoretic true value. A straightforward test to help validate a given TI protocol, which does not rely on the quality of experimental measurements, is the null transformation test where by transformation simulations are run to mutate a molecule to itself. The transformation energy of such null-transformation runs should be equivalent to exactly zero. And, relative binding energy derived from the difference between null transformation simulations in the bound and unbound states should also be zero.

Ensemble averaged  $dV/d\lambda$  plots with respect to  $\lambda$  from four different TI null transformation simulations when T20<sub>S138S</sub> is mutated to itself when either bound to wild type receptor HIV<sub>WT</sub> (Figure 5-7, left panel) and in the unbound state (Figure 5-7, right panel) are shown in Figure 5-7 and Table 5-2. Here, run 1, run 2 and run 3 each consisted of 19  $\lambda$  windows ( $\Delta\lambda=0.05$ ) with

different random seeds. To further evaluate convergence, run1 was expanded to include 39 windows with a window size of 0.025, and denoted as run1b. Similarly, results for one TI null transformation run (run1) for T20<sub>S138S</sub> binding to HIV<sub>V38A</sub> and four TI null-transformation runs (run1, run1b, run2 and run3) for T20<sub>S138S</sub> binding to HIV<sub>N43D</sub> are also reported in Table 5-2.



**Figure 5-7.** The  $dV/d\lambda$  curves with respect to  $\lambda$  for run 1 (red), run1b (green), run2 (cyan) and run3 (magenta). Each point of the  $dV/d\lambda$  plot is derived from ensemble average of a 2ns long TI run.

The  $dV/d\lambda$  curves from all four null transformation runs for T20<sub>S138S</sub> binding to HIV<sub>WT</sub> overlap well with each other in most regions for both bound state simulations and unbound state simulations as shown in Figure 5-7. The calculated transformation energies as well as relative binding energies for almost all null transformation runs in Table 5-2 are close to zero with the exception of run2 for HIV<sub>WT</sub> and run1 for HIV<sub>N43D</sub>. Thus our TI protocol is generally robust and

can correct predict the theoretic transformation energy and relative binding energy of a null transformation in a large protein-peptide complex system with a relatively large transformation mask (a entire amino acid group). Interestingly, for the two cases not close to zero, they are due to either the unbound transformation only (HIV<sub>WT</sub>) or the bound transformation only (HIV<sub>N43D</sub>) as shown in Table 5-2. Referring to the dV/dλ curve in Figure 5-7 (right panel) for HIV<sub>WT</sub>, the error in unbound transformation energy for HIV<sub>WT</sub> is primary due to the TI simulation at window λ = 0.55 (gray circle in Figure 5-7, right panel) where the dV/dλ curve is most steep. Potential improvements to the current TI protocol include (1) performing multiple independent transformation runs and using the average value; (2) performing additional TI simulations with intermediate λ values near the λ windows corresponding to sharply angled regions in the initial dV/dλ curve with uniformly distributed λ windows. For example, for HIV<sub>WT</sub> run2, TI simulations can be performed with λ = 0.525 and 0.575 which presumably could reduce the offset of calculated unbound state transformation energy as well as the final relative binding energy.

**Table 5-2.** Calculated relative binding energy and transformation energy for null transformations

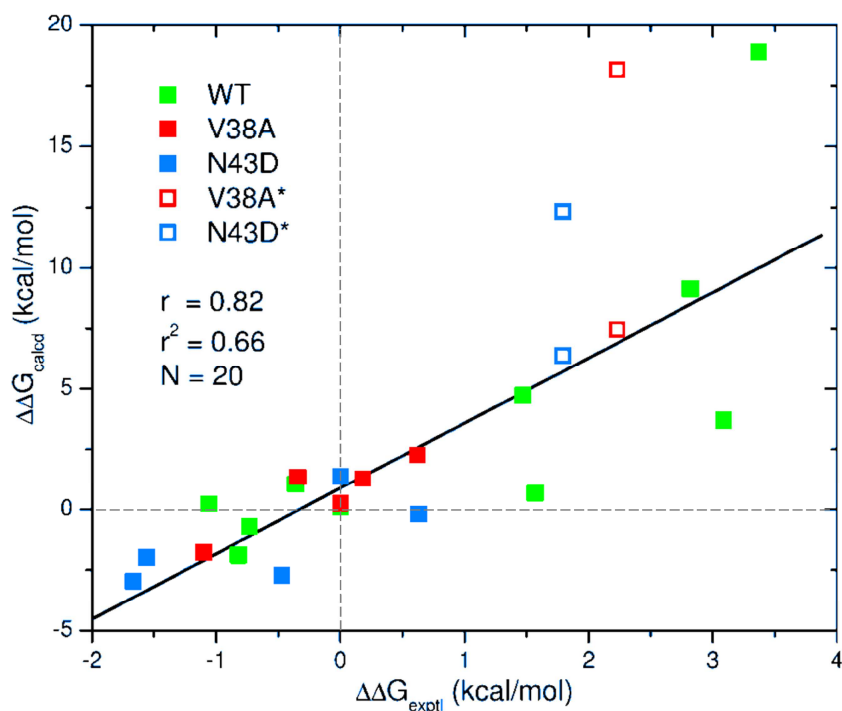
	<b>Runtime</b>	<b>Run1</b>	<b>Run1b</b>	<b>Run2</b>	<b>Run3</b>
HIV <sub>WT</sub>	Bound transformation energy	0.16±0.15	0.22±0.15	0.43±0.14	0.18±0.14
	Unbound transformation energy	0.23±0.18	0.58±0.18	1.56±0.16	0.10±0.19
	Relative binding energy	0.08±0.24	0.80±0.23	1.99±0.21	0.29±0.23
HIV <sub>V38A</sub>	Bound transformation energy	0.03±0.15	--	--	--
	Unbound transformation energy	0.23±0.18	--	--	--
	Relative binding energy	0.26±0.24	--	--	--
HIV <sub>N43D</sub>	Bound transformation energy	1.13±0.14	0.01±0.14	0.47±0.14	0.01±0.13
	Unbound transformation energy	0.23±0.18	0.35±0.18	0.63±0.19	0.18±0.17
	Relative binding energy	1.36±0.23	0.36±0.23	0.15±0.24	0.18±0.22

Run 1b has 39 λ windows instead of 19 for run 1, run2 and run3. Run 2 and 3 have different random seeds from Run 1. Average energy and associated error (standard error of the mean) were reported in kcal/mol.

### 5.3.5 TI Relative Binding Energies: Correlation with Experimental Results

Figure 5-8 and Table 5-3 show relative binding energies computed from the TIMD calculations for 24 systems and comparisons with the experimental measurements.<sup>18</sup> The data in Table 5-3 reports the relative binding energies of the T20 analogs compared to that of wild type T20 (T20<sub>S138S</sub>) for binding to a given form of the receptor (HIV<sub>WT</sub>, HIV<sub>N43D</sub> and HIV<sub>V38A</sub> N-HR). Here, each of the computed relative binding energy value is derived from two sets of TI transformations (bound and unbound), each consisting of 19  $\lambda$  windows that are 2ns long per window. The experimental relative binding energies are computed via subtracting the T20<sub>S138S</sub> absolute binding energy from that of each T20<sub>S138X</sub> analog shown in Table 5-1. For consistency, data for null-transformations reported in Table 5-3 were all obtained using only a single run (run1 in Table 5-2). Overall, 10 system with wild type receptor (green squares) and 7 each with the two mutated receptors (blue squared for HIV<sub>N43D</sub> and red squares for HIV<sub>V38A</sub>) were simulated.

Despite potential issues with null transformation not being zero for N43D. Remarkably, the computational results correctly reproduce the trend of affinity among the complex analogs with a correlation coefficient  $r^2=0.66$  (N=20). Note that four of the data points labeled with stars in Figure 5-8 (hollow squares) were not considered for the  $r^2$  calculation as the corresponding experimental data (see Table 5-2) were reported as approximate ranges (HIV<sub>N43D</sub>-T20<sub>S138K</sub>, HIV<sub>N43D</sub>-T20<sub>S138E</sub>, HIV<sub>V38A</sub>-T20<sub>S138K</sub>, and HIV<sub>V38A</sub>-T20<sub>S138E</sub>).



**Figure 5-8.** Correlation of calculated relative binding energy ( $\Delta\Delta G_{\text{calcd}}$ , y axis) compared to experimental data ( $\Delta\Delta G_{\text{exptl}}$ , x axis). Relative binding energy using TI method is calculated from mutating residue 138 on T20 when bound to the same receptor, averaging over the  $2 \times 19$  ns production run for each complex.

It is also encouraging to find out with only a few exceptions, our calculations were able to correctly distinguish the favorable T20 mutations from the unfavorable ones. As shown in Figure 5-8, the data points to the top right of point (0,0) are unfavorable both experimentally and computationally while data points to the bottom left are favorable. Only four data points fall into the region where the sign of the relative binding energy are not consistent (top left and bottom right from point (0,0)). And, their experimental or calculated energies were very close to zero. Although the correlation is reasonable, it is worth noting that the calculated results tend to over-predict the magnitude of the experimental results. In particular, there is significant overestimation of the effects of the T20<sub>S138K</sub> mutation, although the experimental trend is in fact correct (see Table 5-3 for row T20<sub>S138K</sub>,  $\Delta\Delta G_{\text{calcd}}$  ranging from 12.32 to 18.90 kcal/mol). Possible causes of this

overestimation are that the size of this mutation (Lys) is relatively large, which could lead to a decrease in favorable VDW energy as a result of unfavorable contacts compared to Ser, arising from our use of backbone restraints. Another possibility is the change in net charge, which is likely to affect long-range electrostatic interactions, as our current simulation protocol does not mutate another residue to enforce a consistent total charge for the system. An examination of the results shows that the other charged mutations S138E lead to a similar, although not as large, overestimation ( $\Delta\Delta G_{\text{calcd}}$  ranging from 6.39 to 9.13 kcal/mol). Despite these discrepancies, overall the TI results yield reasonable agreement with experimental trends. Thus, the model and simulation protocols used in this Chapter can be used to perform structural and per-residue energetic analyses to help understand why specific mutations lead to loss of affinity and to aid the design of new HIV fusion inhibitors using analogs binding energy calculations.

**Table 5-3.** Experimental vs. Calculated  $\Delta\Delta G_{\text{bind}}$  for T20 analogs with HIVgp41

Receptors Ligands	HIV <sub>WT</sub>		HIV <sub>V38A</sub>		HIV <sub>N43D</sub>	
	Exptl.	Calcd.	Exptl.	Calcd.	Exptl.	Calcd.
T20 <sub>S138S</sub>	0.00	0.08±0.24	0.00	0.26±0.24	0.00	1.36±0.23
T20 <sub>S138A</sub>	-0.82	-1.88±0.19	-1.10	-1.75±0.19	-1.56	-1.96±0.19
T20 <sub>S138E</sub>	2.82	9.13±0.40	2.23	7.47±0.39	1.79	6.39±0.41
T20 <sub>S138G</sub>	-0.36	1.05±0.22	0.62	2.26±0.21	0.63	-0.19±0.21
T20 <sub>S138K</sub>	3.37	18.90±0.38	2.23	18.17±0.38	1.79	12.32±0.38
T20 <sub>S138L</sub>	-0.73	-0.70±0.27	-0.34	1.34±0.27	-1.67	-2.96±0.26
T20 <sub>S138P</sub>	3.09	3.71±0.25	2.23	--	1.79	--
T20 <sub>S138Q</sub>	1.57	0.68±0.30	2.23	--	1.79	--
T20 <sub>S138V</sub>	-1.06	0.22±0.24	0.18	1.28±0.25	-0.47	-2.71±0.24
T20 <sub>S138W</sub>	1.47	4.73±0.35	2.23	--	1.79	--

Calcd.  $\Delta\Delta G_{\text{bind}}$  computed using TIMD results from mutating residue 138 on T20 when bound to the same receptor. Exptl.  $\Delta\Delta G_{\text{bind}}$  computed via subtracting the binding energy of wild type T20 from that of the mutated T20 when bound to the same receptor. Energy unit in kcal/mol.

### 5.3.6 T20 Binding: Footprint Analyses of the Wild-Type System

The 3-fold symmetry of gp41 presents challenges with respect to determine which amino acids interactions are most important for ligand binding in this system. In an attempt to identify which residues are most important, per-residue energy decomposition of the VDW and ES interactions between the three different gp41 NHR helices (the receptor) and T20 (the ligand) were computed and plotted as both one-dimensional and two dimensional (termed molecular heatmap) molecular footprints. The gp41 receptor residue numbers range from 1 to 94 ( $\times 3$  chains) for each of the individual NHR monomers while the T20 ligand residue numbers range from 127 to 162.

**Key residues: VDW interactions.** In Figure 5-9, VDW energies are averaged across all frames (20,000 snapshots) of the 20ns long endstate MD simulations (frames saved every 1ps) of the wild type complex (HIV<sub>WT</sub> -T20<sub>S138S</sub>), and the color scheme represents the magnitude of interactions (favorable ones in red squares, unfavorable ones in blue squares) provided they exceed a threshold of 2.0 kcal/mol (interactions  $< 2.0$  kcal/mol in magnitude are shown in white squares).

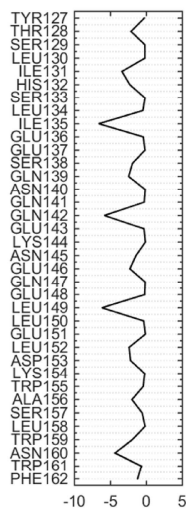
Of the three N-HR helices, NHR2 is not directly in contact with T20, thus the threshold VDW molecular footprint for NHR2-T20 contained all white squares ( $< 2.0$  kcal/mol) and the data is not shown. In contrast, threshold VDW molecular heatmaps for NHR1 (Figure 5-9a) and NHR3 (Figure 5-9b) show numerous favorable interactions (red squares), and the differences in relative position of T20 interacting with each NHR yielded different interaction patterns. For the NHR1-T20 interactions, 3 T20 residues Ile135, Gln142 and Leu149 (indicated by the horizontal arrows in Figure 5-9a) are the most significant in terms of the ligand footprint (VDW interactions decomposed by each ligand residue, top left panel, Figure 5-9a), and correspond to 3 out of 5 key residues Arg31, Val38 and Leu45 (indicated by vertical arrows in the heatmap, Figure 5-9a) near

the site of mutations (V38A, N43D) on the corresponding receptor footprint (VDW interactions decomposed on each receptor residue, bottom right panel, Figure 5-9b).

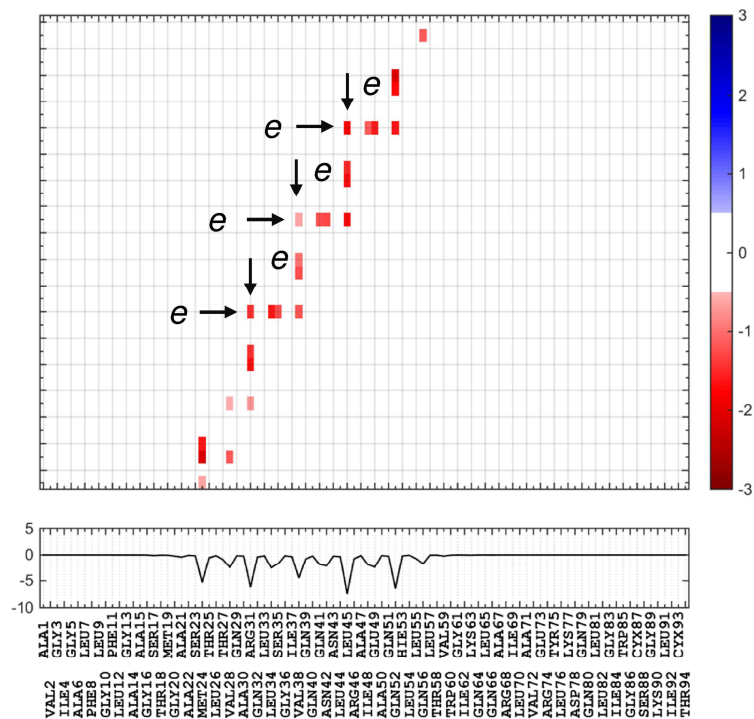
For the 3 residues on T20, they are all 7 residues apart in terms of linear sequence, and thus adopt the same position corresponding to letter *e* denoted in Figure 5-2 in the  $\alpha$ -helical T20 secondary structure. For the 3 residues on NHR1 (Figure 5-9a), they are exactly 7 residues apart, and also all adopt positions corresponding to letter *e* in the  $\alpha$ -helix NHR. These primary interactions involve the following residue pairs for the 3 key “*e*” residues on T20: (1) Ile135 consist of: Ile135-Leu45 (*e*), Ile135-Ile48 (*a*), Ile135-Glu49 (*b*), Ile135-Gln52 (*e*); (2) Gln142 consist of: Gln142-Val38 (*e*), Gln142-Gln41 (*a*), Gln142-Asn42 (*b*), Gln142- Leu45 (*e*); (3) Leu149 consist of: Leu149-Arg31 (*e*), Leu149-Leu34 (*a*), Leu149-Ser35 (*b*), Leu149-Val38 (*e*). From the receptor point of view, for the 3 key “*e*” residues on NHR1, key residue pairs include: (1) for Arg31: Arg31-Leu152 (*a*), Arg31-Asp153 (*b*), Arg31-Ala156 (*e*); (2) for Val38: Val38-Asn145 (*a*), Val38-Glu146 (*b*), Val38-Leu149 (*e*); (3) for Leu45: Leu45-Ser138 (*a*), Leu45-Gln139 (*b*), Leu45-Gln142 (*e*).

For the NH3-T20 interactions (Figure 5-9b), key residue pairs are emphasized using the square boxes instead of arrows in the molecular heatmap. Here, the top edge of the box corresponds to T20 residues at position *d*; the bottom edge corresponds to T20 residues at position *a*; the left edge corresponds to NHR3 residues at position *c* and the right edge corresponds to NHR3 residues at position *g*. Most of interactions in this heatmap involve residues at these four positions. Specifically, T20 residues at position *d* (top) interact most strongly with NHR3 residues at position *g* (right), as shown by the red squares overlapping with the top right corner of the black boxes. And T20 residues at position *a* (bottom) interact mostly with NHR3 residues at position *c* (left), showing by the red squares overlapping with the bottom left corner of the black boxes.

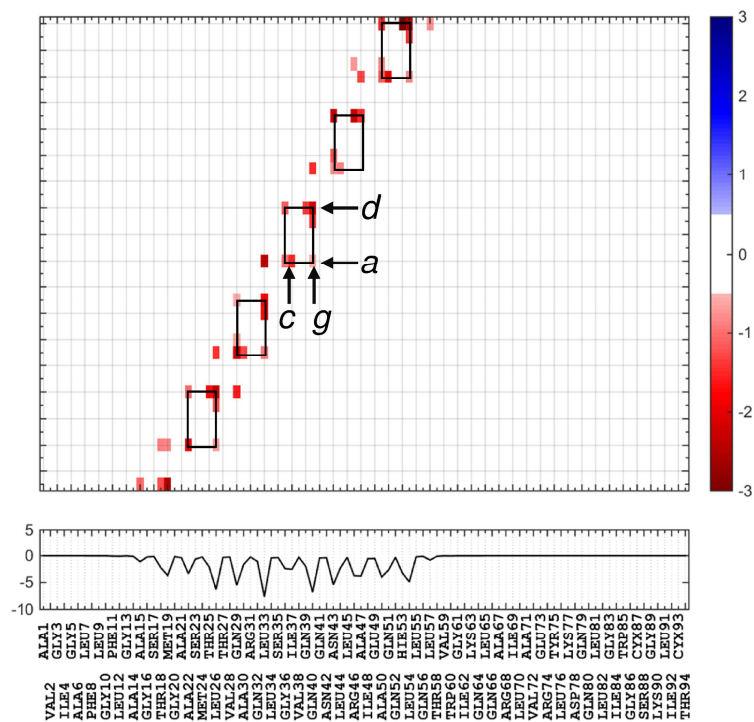
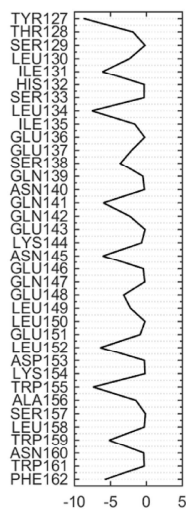




(a) NHR1

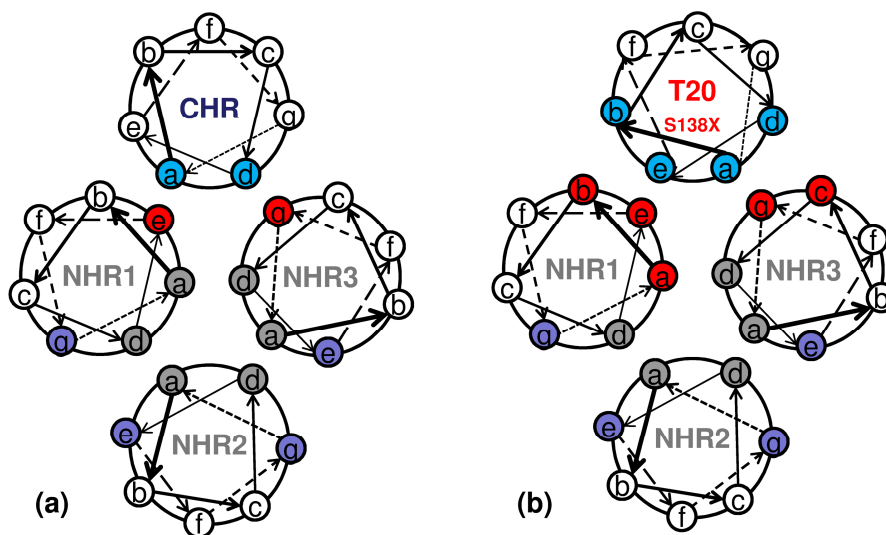


(b) NHR3



**Figure 5-9.** (a) Key residues for VDW contact (heatmap for NHR1 WTS138S (>0.5kcal/mol)); (b) Key residues for VDW contact (heatmap for NHR3 WTS138S (>0.5kcal/mol))

Interestingly, the analyses on binding interaction patterns shown in Figure 5-9 suggest a modified wheel representation for the T20-3NHR bundle is more applicable, as shown in Figure 5-10b, with T20 rotated by 1 position counter-clockwise resulting in an alternative interaction pattern. Using the traditional model (Figure 5-10a), CHR residues at position *a* should be directly interacting with NHR1 residues at position *e* and those at position *d* should be making close contact with NHR3 residues at position *g*. The CHR residues at positions *e* and *b* are more distal to either NHR1 or NHR3. However, based on the energy footprints shown in Figure 5-9, T20 residues at position *e* and *b* also appear to be essential, especially for interactions with NHR1, suggesting the modified interaction wheel in Figure 5-10b. Overall, use of molecular heatmaps, such as the ones shown in Figure 5-9, may be more effective than use of wheel representations (Figure 5-10a, b) if the goal is to more precisely identify residues pairs involving alpha-helical bundles.

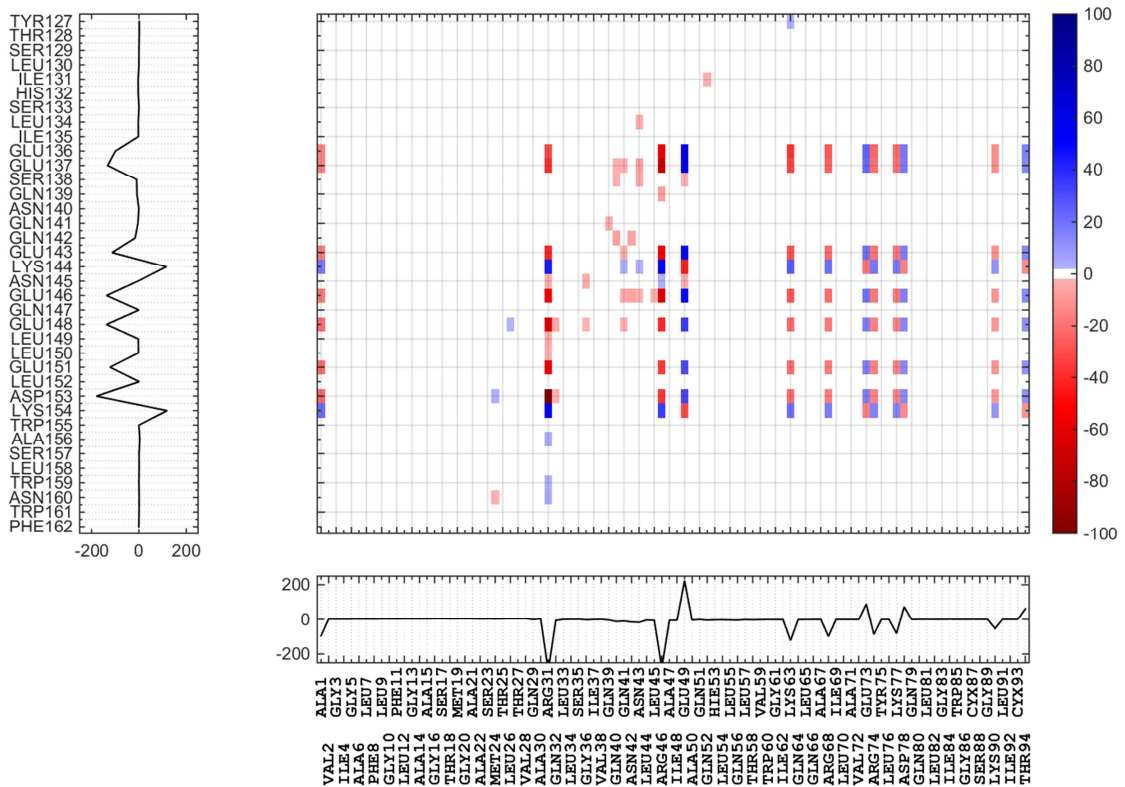


**Figure 5-10.** (a) Original and (b) modified model of T20-3NHR based on VDW footprint and heatmap analyses.

***Charged residues: electrostatics interactions.*** Compared to VDW interactions, ES interactions are more long-range and the three NHR ES heatmaps were in general very similar. Thus, to simplify the analysis, the ES energies across all three helices ( $3\text{NHR}=\text{NHR1}+\text{NHR2}+\text{NHR3}$ ) were computed and shown in Figure 5-11. As expected, inspection of the electrostatic T20 binding profile using the ES heatmap shows that charged residues (indicated in Figure 5-2a) contribute most to the electrostatics interactions profile. Large peaks in both the ligand and receptor footprints all correspond to these charged residues. As simulated here, T20 has a net formal charge of -5 while the wild type gp41 NHR has a total net charge of +12 (+4 for each monomer). Thus, all the negatively charged T20 residues, including Glu136 (*f*), Glu137 (*g*), Glu143 (*f*), Glu146 (*b*), Glu148 (*d*), Glu151 (*g*), Asp153 (*b*), where the letter in the bracket indicates the residue position in the wheel representation of the T20-3NHR bundle, will yield overall favorable per-residue electrostatic energy contributions. Correspondingly, the positively charged T20 residues Lys144 (*g*) and Lys154(*c*) yielded unfavorable per-residue electrostatic energy. Analogous observations are made from the receptor point of view in terms of negatively charged residues (Glu49 (*b*), Glu73 (*e*), Asp78 (*c*)) and positively charged residues (Arg31 (*e*), Arg46 (*f*), Lys63 (*b*), Arg68 (*g*), Arg74 (*f*), Lys77 (*b*)). Note that the interactions involving residues Ala1 and Thr94 are due to the fact that the terminal residues on the receptor are left uncapped.

Notably, none of the positive charges on T20 are located directly in the binding interface (only at position *c* and *g*) and thus avoid strong repulsive electrostatic interactions with the overall positively charged receptor. Focusing on the Ala15-Leu57 region (on the receptor), corresponding to the interface with direct contact to T20 in terms of the VDW interaction heatmap in Figure 5-9, residues Arg31 (*e*), Arg46 (*f*) and Glu49 (*b*) yield the most significant per-residue ES interactions as shown in Figure 5-11. In addition to the charged residues, polar residues at this interface also make

important contributions. For example, favorable interactions (red squares) are observed for receptor residues Gly36 (c), Gln39 (f), Gln40 (g), Gln41 (a), Asn42 (b), Asn43 (e) with T20 residues Ser138 (a), Gln139 (b), Gln141 (d), Gln142 (e). And, unfavorable interactions are reported between Lys144 (g) on T20 and Gln41 (a) /Asn43 (c) on gp41.



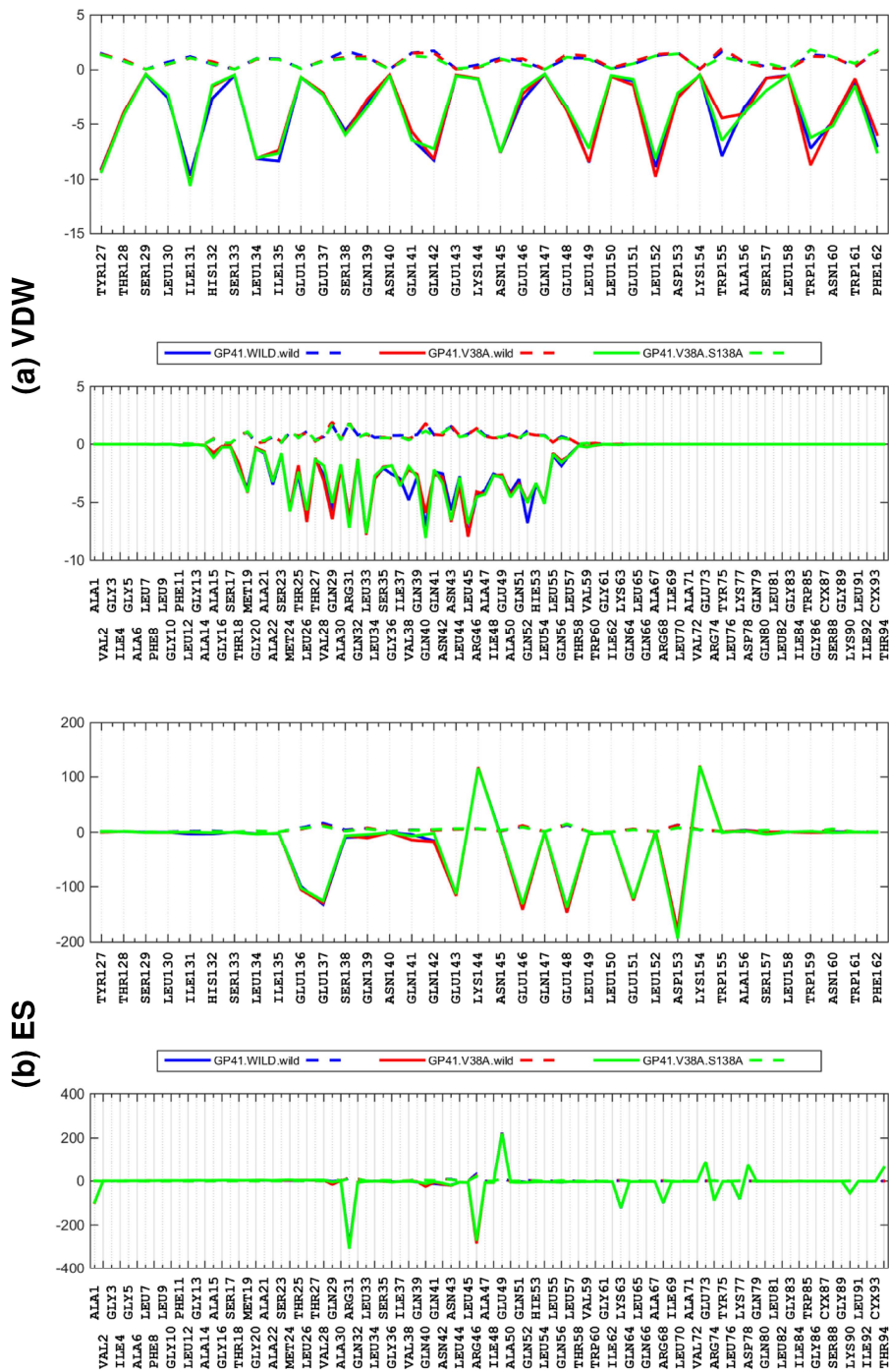
**Figure 5-11.** Electrostatic (ES) interaction heatmap for T20 binding to 3NHR.

### 5.3.7 Why S138A?

As predicted both experimentally by Izumi et al<sup>18</sup> and computationally in this study (Table 5-3 and Figure 5-8), S138A mutation can restore binding caused by primary mutations in gp41 NHR to approximately the wild type complex level. Ala is more hydrophobic and examination of the data in Table 5-3 shows other hydrophobic mutations (e.g. S138V) also usually enhance

binding. According to the experimental data, the loss in binding affinity of T20 is ~1.34 kcal/mol for the gp41 V38A mutation, and ~1.78 kcal/mol for the gp41 N43D mutation (Table 5-1). The estimated experimental gains in binding energy of T20 with the gp41 S138A secondary mutation for these two primary mutants are 1.75 kcal/mol and 1.96 kcal/mol respectively. It is thus of particular interest to identify the origins of affinity and the change in affinity between the mutating systems. Here, we decomposed the total computational binding energy into individual residues to generate both the receptor gp41 (sum of all three helices, 3NHR) and T20 ligand footprints (Figure 5-12 and 5-13). The values of the associated standard error of the mean for each per-residue decomposition data point are plotted as the dashed curves.

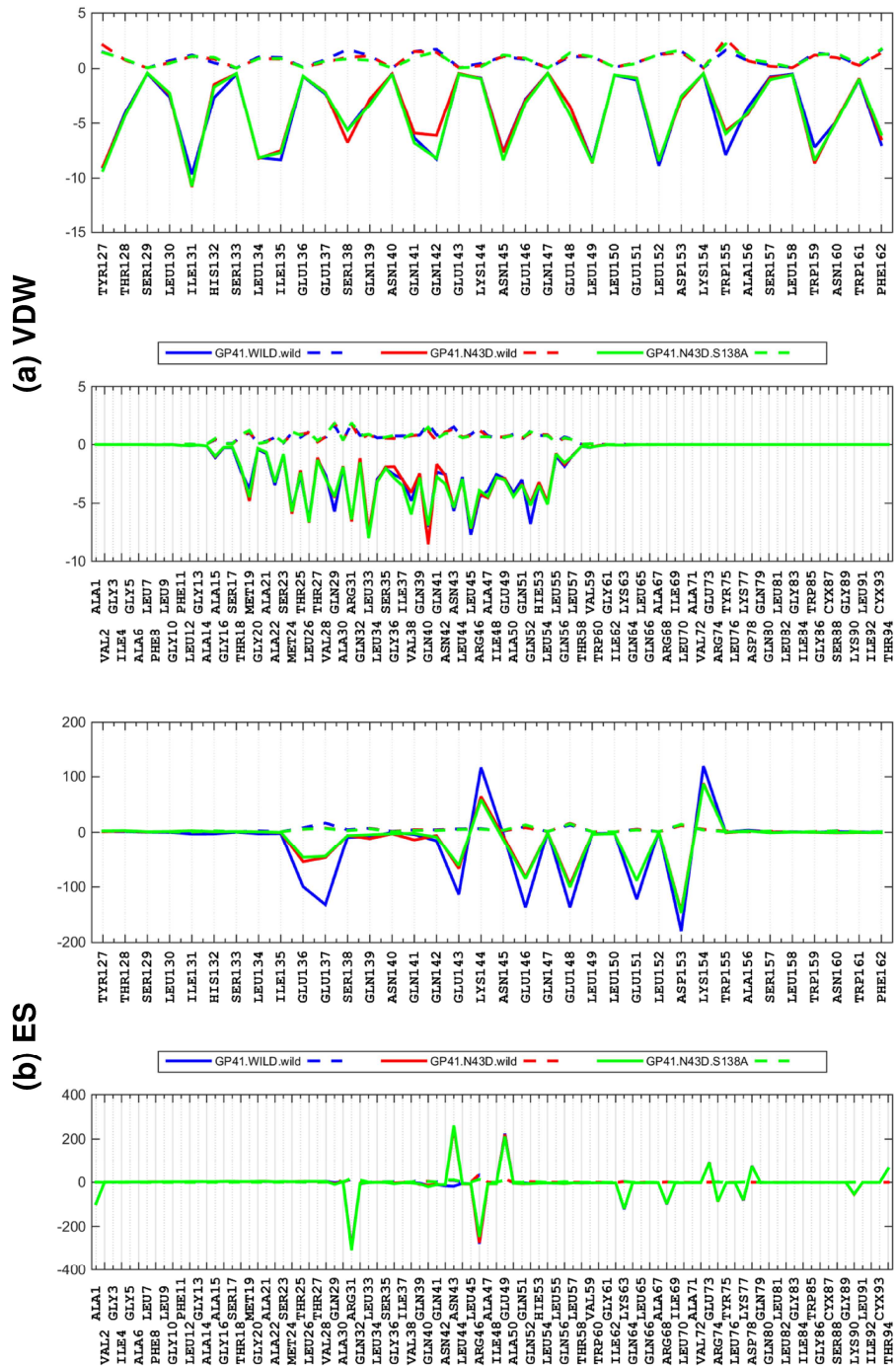
**Primary (V38A) and Secondary Mutations.** The footprints for the three systems involved in the V38A primary and S138A secondary mutations are shown in Figure 5-12. An obvious loss of VDW affinity was observed at gp41 residues 38 and 52, both at position *e* on 3NHR, for the V38A primary mutation (blue to red curve, Figure 5-12a bottom). Interestingly, the VDW footprint for the T20 ligand (Figure 5-12a top) showed the most significant changes were at residue 155 (*d*), which is directly interacting with gp41 residues Ala22 (*c*), Thr25 (*f*), Leu26 (*g*) and Gln29 (*e*), the latter being one turn apart from the primary mutation site at Val38 (*e*). Here, the VDW interactions at Trp155 get less favorable due to primary mutation V38A (blue to red curve) and then improve as a result of the secondary mutation S138A (red to green curve). In contrast, only minor changes are observed in the corresponding ES footprints (blue, red and green curves in Figure 5-12b). This indicates that the hydrophobic V38A and S138A mutations have little direct effects on the electrostatic interactions in the complex system. While more structural analysis is needed, loss and gain of physical contact at residue Trp155 could be a contributing factor in the V38A/S138A mutations.



**Figure 5-12.** (a) VDW and (b) ES molecular footprint for V38A primary, and S138A secondary mutations. Error bar for each data point plotted as the dashed curves. HIV<sub>WT-T20<sub>S138S</sub></sub> in blue, HIV<sub>V38A-T20<sub>S138S</sub></sub> in red, HIV<sub>V38A-T20<sub>S138A</sub></sub> in green

**Primary (N43D) and Secondary Mutations.** The footprints for the three systems involved in the N43D primary and S138A secondary mutations are shown in Figure 5-13. Here, a loss (blue to red curve) of VDW affinity is observed again at residues 38 and 52 as well as 41. In terms of the secondary mutation, only position 38 and 41 on the receptor show a restoration of affinity (red to green curve). For the T20 ligand, at residue Gln142, the per-residue VDW interactions (Figure 5-13a, top) become significantly less favorable due to the primary mutation N43D (blue to red curve) and then are restored as a result of secondary mutation S138A (red to green curve). Interestingly, Gln142 is predicted to directly interact with residues Val38, Gln41 (with affinity changes in the receptor footprints) as well as Asn42 and Leu45, but not N43D, as shown in the VDW heatmap in Figure 5-9a. Thus, loss and gain of physical contacts at residue Gln142, likely paired with residue V38A and Gln41, could be contributing factors in the N43D/S138A mutations.

In sharp contrast to V38A, as shown in Figure 5-13b, the charged primary mutation N43D has significantly changed the magnitude of ES interactions for all charged residues, especially residue Glu136 and Glu137, on T20 (Figure 5-13b top). Correspondingly, the mutation introduced a large unfavorable ES interaction at residue 43 (Figure 5-13b bottom, blue to red/green). And, the small neutral secondary S138A mutation has only limited effects on the ES interactions, similar to what was observed in the V38A/S138A mutations case. The fact that the large loss in ES energy is not restored by S138A, in terms of only the interaction energy examined here (despite the TI calculations yielding the correct experimental trends), emphasizes challenges with computationally pinpointing the biological effects of charged vs. neutral mutations. Additional computational studies are warranted, more specifically, use of footprints that includes desolvation energy penalties should be examined.



**Figure 5-13.** (a) VDW and (b) ES molecular footprint for N43D primary, and S138A secondary mutations. Error bar for each data point plotted as the dashed curves. HIV<sub>WT</sub>-T20<sub>S138S</sub> in blue, HIV<sub>N43D</sub>-T20<sub>S138S</sub> in red, HIV<sub>N43D</sub>-T20<sub>S138A</sub> in green.



## 5.4 Conclusion

In summary, in Chapter 5, we reported a thermodynamic integration molecular dynamics protocol of a large complex system HIVgp41-T20 with a transformation mask set to an entire amino acid residue on the peptide ligand T20 to study both primary (V38A, N43D) and compensatory (S138X) mutations. Our protocol is shown to have reasonable behavior in terms of simulation stability. Relative binding energy calculations yielded good correlation with experimental measurements for a series of complex analogs with ligand variants (secondary mutations) binding to receptor variants (primary mutations).

In addition, one-dimensional molecular footprints as well as two-dimensional footprints (heatmap) for the protein-peptide (gp41-T20) system are calculated for both the wild type and mutant complexes. Preliminary discussions on the VDW and ES interactions between T20 and the gp41 N-HR leading to the computational evaluation of peptide binding and changes in affinity due to primary and secondary mutations are provided. With the present results, the neutral mutation pair V38A/S138A was more easily interpretable than the charged mutation pair N43D/S138A in terms of the energetic effects of loss and gain of affinity.

Future studies to yield more accurate relative binding energy estimates and characterization of the origins of affinity in the gp41-T20 system include: (1) multiple independent TI simulations to obtain more converged “average” energies; (2) additional  $\lambda$  windows in TIMD near the steepest regions of the initial  $dV/d\lambda$  curve as well as for  $\lambda$  close to 0 or 1; (3) construction of lipid-bound complex systems to enhance the robustness of the structure, better mimic the dynamics of the system (potentially with non-restrained simulation protocol for the new construct), and evaluate the role of lipids in T20 binding; (4) additional structural analyses at the per-residue level.

## **Chapter 6. Dissertation Summary: Scientific Impact, Challenge and Future Direction.**

Computer-aided structure-based drug design is a rising force in modern pharmaceutical industry. Advanced technology to resolve complicated drug target structures, as well as new and improved computational modeling methods can all contribute to the enhancement of the overall drug discovery pipeline. Studies discussed in Chapters 2, 3, 4, and 5 provided a new perspective of molecular recognition; investigated the HIVgp41 NHR-CHR binding interface; and introduced novel protocols for molecular docking. This Chapter summarizes the scientific impact of works in this dissertation and points out future directions of this study.

### **6.1 Development and Application of FMS Docking Protocol**

#### **6.1.1 Scientific Impact**

In Chapter 2, we introduced a new pharmacophore-based scoring function in the docking program DOCK. This FMS scoring method encodes the geometric arrangement of key chemical features in a reference molecule (known ligand) to enrich for molecules with desirable binding profiles. Importantly, as shown by validation test results, FMS score when used alone, and in combination with other scoring functions such as single grid energy (SGE) score, can in most cases improve docking performance compared to the standard force field based approach (*e.g.* SGE alone). For pose reproduction, FMS and FMS+SGE scores yield close to 100% success rate (93.5% and 98.3%, respectively) compared to 72.5% success for SGE using the docking testset SB2012 (N=1043), where success is defined as generating a close-to-native ( $\text{RMSD} \leq 2\text{\AA}$  to crystal) pose in docking. For crossdocking, FMS/FMS+SGE also yields promising results, given that the

pharmacophore and RMSD references overlap reasonably. In general, for enrichment study, FMS shows the most favorable early and total enrichment followed by FMS+SGE. Finally, in virtual screening, FMS score tends to select a unique set of top molecules different from SGE score. Additional computational experiments also show that with an engineered pharmacophore reference, which includes multiple copies of a certain functional group, FMS-guided docking can be customized to prioritize a hotspot for binding. Overall, we have developed an alternative scoring protocol for the docking community, with the added possibility for use as a ligand-only (i.e. without use of a receptor) scoring tool. The final FMS code is slated to be released in DOCK6.8.

In Chapter 3 and 4, more applications of the FMS docking protocol have been implemented and closely investigated. Detailed structure visualization of virtual screening further supported the robustness and utility of the FMS docking protocol. Top molecules scored by various scoring functions including FMS score have been prioritized and, pending additional study, may be purchased for experimental testing by collaborating labs to identify new small molecule leads that inhibit HIV fusion. Another important component of this study was merging the FMS scoring function into the *de novo* DOCK code under active development by Dr. William Joseph Allen and colleagues in the Rizzo lab. As proof of principle, the merged code was used to guide novel ligand growth to match pharmacophores of reference in rebuilding tests using 50 diverse small molecule inhibitors and 2 peptides targeting either the hydrophobic or the inner pockets in HIVgp41.

### **6.1.2 Challenge, Related Work and Future Direction**

To expand functionality, future implementation of new features of FMS scoring such as including matching to multiple references or use of a receptor-based reference should be explored. By matching pharmacophore features of multiple ligand references with known binding poses in the

same target site, FMS score can capture the hotspots that are statistically more likely to contribute to ligand binding and yield more potent leads. And, extension to a receptor-based pharmacophore would enable identification of ligands that make mirror-image, favorable interactions with the target. A receptor-based FMS protocol could also be used in combination with a ligand-based FMS protocol to eliminate spatial clashes between docked ligands and the receptor (i.e. similar to excluded volume) while ensuring good ligand overlap. Implementation of a text-based input format would also be worthwhile. Also, the *de novo* FMS docking protocol should also be continuously updated to match ongoing developments of the *de novo* DOCK code.

## **6.2 Computational Investigation of HIVgp41-T20 Binding**

### **6.2.1 Scientific Impact**

In Chapter 5, we employed overall and per-residue energetic analyses on a series of HIVgp41-T20 complex analogs with point mutations on the target protein as well as on the peptide ligand. With thermodynamic integration simulations, we obtained good correlation with the experimentally determined relative free binding energies due to primary and secondary mutations. With molecular footprint and heatmap calculations, we identified per residue interaction patterns of T20 binding to HIVgp41 that suggest a modified wheel depiction may be a more appropriate description of alpha-helical bundles in the case of gp41. Although our detailed structural analyses help provide a deeper understanding of the origins of T20 binding affinity, and the effects of a neutral primary and secondary mutation pair in the system, challenges associated with a charged mutation were also encountered. Importantly, as one of the few TI case studies where reasonable correlations of free energy measurements are obtained when mutating an entire amino acid residue at once in terms of the mutation mask and VDW and ES terms simultaneously, this study can aid

the community in developing more straightforward and efficient alchemical simulation protocols. It is also worth noting that, the simulation behavior as well as free energy results of the constructed protein-peptide systems indicate the robustness of our computational model of the HIVgp41-T20 complex.

### **6.2.2 Challenge, Related Work and Future Direction**

While prior studies from the Rizzo lab have explored membrane-bound HIVgp41-T20 simulations and the roles of membrane at the fusion peptide insertion interface of gp41, more extended simulations using a more rigorous model for relative binding energy calculations (*i.e.* TI vs. MM-GBSA) would be worthwhile. Preliminary efforts have been made in constructing a membrane-bound gp41-T20 system using the online CHARM-GUI and simulating the structure with GPU-accelerated *pmemd* using AMBER14.<sup>49</sup> The model makes use of a DOPC lipid bilayer prepared with the new Lipid14<sup>146</sup> force field. Works to adjust the insertion depth of the complex into the membrane bilayer, identify a reasonable solvent-to-solute ratio, and obtain a good simulation protocol including equilibration setups is ongoing. Simulations with explicit lipid should further aid the study of dynamics of the system and in characterizing the nature of T20 binding and drug resistance.

In addition, the identification of which key residues contribute most to binding of T20 will be useful as references to guide small molecule lead discovery. The Rizzo lab has already successfully screened for small molecule leads to match molecular footprints of key residues using peptide substrates to yield hits with experimental activities. Similar strategies using the newly developed FMS scoring function as described in Chapter 2, 3 and 4 are envisioned to target the T20 binding interface.

### 6.3 Summary

As presented in this dissertation, although considerably challenges remain, atomic-level molecular docking, molecular dynamics simulations, free energy calculations, and molecular footprint studies have been employed to provide insight into protein-ligand binding for the drug target HIVgp41. Of particular note is development of a new powerful yet easy-to-use pharmacophore-based docking method for the program DOCK which we believe will become an important tool of benefit to the community performing both virtual screening and *de novo* design projects.

## Bibliography

- (1) PubMed AIDS-PubMed Health. **2011**.
- (2) AVERT HIV and AIDS Charity AVERT. **2011**.
- (3) UNAIDS AIDS fact sheet. **2011**.
- (4) UNAIDS Global AIDS Epidemic Facts and Figures. **2012**.
- (5) UNAIDS Global Report: UNAIDS Report on the Global AIDS Epidemic **2013**.
- (6) AIDSinfo FDA-Approved HIV Medicines Fact Sheet. **2014**.
- (7) FDA Antiretroviral Drugs Used in the Treatment of HIV Infection. **2014**.
- (8) NIAID Drugs That Fight HIV-1. **2015**.
- (9) Liu, S.; Lu, H.; Niu, J.; Xu, Y.; Wu, S.; Jiang, S. Different from the HIV fusion inhibitor C34, the anti-HIV drug Fuzeon (T-20) inhibits HIV-1 entry by targeting multiple sites in gp41 and gp120. *J. Biol. Chem.* **2005**, *280*, 11259-11273.
- (10) Chan, D. C.; Fass, D.; Berger, J. M.; Kim, P. S. Core Structure of gp41 from the HIV Envelope Glycoprotein. *Cell* **1997**, *89*, 263-273.
- (11) Hughson, F. M. Enveloped viruses: A common mode of membrane fusion? *Current Biology* **1997**, *7*, R565-R569.
- (12) Eckert, D. M.; Kim, P. S. Design of potent inhibitors of HIV-1 entry from the gp41 N-peptide region. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 11187-11192.
- (13) Jiang, S.; Lin, K.; Zhang, L.; Debnath, A. K. A screening assay for antiviral compounds targeted to the HIV-1 gp41 core structure using a conformation-specific monoclonal antibody. *Journal of Virological Methods* **1999**, *80*, 85-96.
- (14) Allen, W. J.; Yi, H. A.; Gochin, M.; Jacobs, A.; Rizzo, R. C. Small Molecule Inhibitors of HIVgp41 N-heptad Repeat Trimer Formation. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 2853-2859.

- (15) Greenberg, M. L. Resistance to enfuvirtide, the first HIV fusion inhibitor. *Journal of Antimicrobial Chemotherapy* **2004**, *54*, 333-340.
- (16) McGillick, B. E.; Balias, T. E.; Mukherjee, S.; Rizzo, R. C. Origins of Resistance to the HIVgp41 Viral Entry Inhibitor T20. *Biochemistry* **2010**, *49*, 3575-3592.
- (17) Shi, W.; Bohon, J.; Han, D. P.; Habte, H.; Qin, Y.; Cho, M. W.; Chance, M. R. Structural Characterization of HIV gp41 with the Membrane-proximal External Region. *J. Biol. Chem.* **2010**, *285*, 24290-24298.
- (18) Izumi, K.; Kodama, E.; Shimura, K.; Sakagami, Y.; Watanabe, K.; Ito, S.; Watabe, T.; Terakawa, Y.; Nishikawa, H.; Sarafianos, S. G.; Kitaura, K.; Oishi, S.; Fujii, N.; Matsuoka, M. Design of peptide-based inhibitors for human immunodeficiency virus type 1 strains resistant to T-20. *J. Biol. Chem.* **2009**, *284*, 4914-4920.
- (19) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813-1818.
- (20) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature* **2004**, *432*, 862-865.
- (21) Kuntz, I. D. Structure-based Strategies for Drug Design and Discovery. *Science* **1992**, *257*, 1078-1082.
- (22) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3-26.
- (23) Hou, T. J.; Xu, X. J. ADME Evaluation in Drug Discovery. 3. Modeling Blood-Brain Barrier Partitioning Using Simple Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2137-2152.
- (24) Dickson, M.; Gagnon, J. P. Key factors in the rising cost of new drug discovery and development. *Nat. Rev. Drug Discov.* **2004**, *3*, 417-429.
- (25) van Oosterom, A. T.; Judson, I.; Verweij, J.; Stroobants, S.; di Paola, E. D.; Dimitrijevic, S.; Martens, M.; Webb, A.; Scot, R.; Van Glabbeke, M.; Silberman, S.; Nielsen, O. S. Safety and efficacy of imatinib (STI571) in metastatic gastrointestinal stromal tumours: a phase I study. *The Lancet* **2001**, *358*, 1421-1423.
- (26) NIH; Structure-Based Drug Design Fact Sheet: 2011.



- (27) Ward, P. Oseltamivir (Tamiflu(R)) and its potential for use in the event of an influenza pandemic. *Journal of Antimicrobial Chemotherapy* **2005**, *55*, i5-i21.
- (28) Zanamivir: from drug design to the clinic. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **2001**, *356*, 1885-1893.
- (29) von Itzstein, M. The war against influenza: discovery and development of sialidase inhibitors. *Nat. Rev. Drug Discov.* **2007**, *6*, 967-974.
- (30) Zhang, K. E.; Wu, E.; Patick, A. K.; Kerr, B.; Zorbas, M.; Lankford, A.; Kobayashi, T.; Maeda, Y.; Shetty, B.; Webber, S. Circulating Metabolites of the Human Immunodeficiency Virus Protease Inhibitor Nelfinavir in Humans: Structural Identification, Levels in Plasma, and Antiviral Activities. *Antimicrobial Agents and Chemotherapy* **2001**, *45*, 1086-1093.
- (31) Hardy, L. W.; Malikayil, A. The impact of structure-guided drug design on clinical agents. *Curr. Drug. Discov.* **2003**, *15*, 15-20.
- (32) Sham, H. L. ABT-378, a Highly Potent Inhibitor of the Human Immunodeficiency Virus Protease. *Antimicrobial Agents and Chemotherapy* **1998**, *42*, 3218-3224.
- (33) Tomasselli, A. G.; Heinrikson, R. L. Targeting the HIV-protease in AIDS therapy: a current clinical perspective<sup>1</sup>. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **2000**, *1477*, 189-214.
- (34) Huang, Y.; Rizzo, R. C. A Water-Based Mechanism of Specificity and Resistance for Lapatinib with ErbB Family Kinases. *Biochemistry* **2012**, *51*, 2390-2406.
- (35) Kubinyi, H. Similarity and Dissimilarity: A Medicinal Chemist's View. *Perspectives in Drug Discovery and Design* **1998**, *9-11*, 225-252.
- (36) Kubinyi, H. Chemical similarity and biological activities. *Journal of the Brazilian Chemical Society* **2002**, *13*, 717-726.
- (37) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912-5931.
- (38) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757-1768.

- (39) Irwin, J. J.; Shoichet, B. K. ZINC – A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2004**, *45*, 177-182.
- (40) Accelrys Available Chemical Directory, website accessed March 2015; <http://accelrys.com/products/databases/sourcing/available-chemicals-directory.html>, 2015.
- (41) Pegg, S. H.; Haresco, J.; Kuntz, I. A genetic algorithm for structure-based de novo design. *J. Comput. Aided Mol. Des.* **2001**, *15*, 911-933.
- (42) Mackerell, A. D., Jr. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* **2004**, *25*, 1584-1604.
- (43) 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. *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.
- (44) Pang, Y. P. Use of 1-4 interaction scaling factors to control the conformational equilibrium between alpha-helix and beta-strand. *Biochemical and biophysical research communications* **2015**, *457*, 183-186.
- (45) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235-242.
- (46) Karplus, M.; Petsko, G. A. Molecular dynamics simulations in biology. *Nature* **1990**, *347*, 631-639.
- (47) Case, D. A.; Darden, T. A.; Cheatham, I., T. E.; Simmerling, C. L.; Wang, J.; Duke, R. E.; Luo, R.; Walker, R. C.; Zhang, W.; Merz, K. M.; Roberts, B.; Wang, B.; Hayik, S.; Roitberg, A.; Seabra, G.; Kolossváry, I.; Wong, K. F.; Paesani, F.; Vanicek, J.; Liu, J.; Wu, X.; Brozell, S. R.; Steinbrecher, T.; Gohlke, H.; Cai, Q.; Ye, X.; Wang, J.; Hsieh, M.-J.; Cui, G.; Roe, D. R.; Mathews, D. H.; Seetin, M. G.; Sagui, C.; Babin, V.; Luchko, T.; Gusarov, S.; Kovalenko, A.; Kollman, P. A.; AMBER 11, University of California: San Francisco, 2010.
- (48) 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.
- (49) D.A. Case, V. B., J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T.E. Cheatham, III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I.

Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wu and P.A. Kollman (2014) AMBER 14, University of California, San Francisco. **2014**.

(50) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187-217.

(51) Christen, M.; Hünenberger, P. H.; Bakowies, D.; Baron, R.; Bürgi, R.; Geerke, D. P.; Heinz, T. N.; Kastenholz, M. A.; Kräutler, V.; Oostenbrink, C.; Peter, C.; Trzesniak, D.; van Gunsteren, W. F. The GROMOS software for biomolecular simulation: GROMOS05. *J. Comput. Chem.* **2005**, *26*, 1719-1751.

(52) Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J. C. GROMACS: Fast, flexible, and free. *J. Comput. Chem.* **2005**, *26*, 1701-1718.

(53) Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* **2008**, *4*, 435-447.

(54) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781-1802.

(55) Gotz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *J. Chem. Theory Comput.* **2012**, *8*, 1542-1555.

(56) Salomon-Ferrer, R.; Götz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878-3888.

(57) Reddy, R.; Erion, M. D. *Free Energy Calculations in Rational Drug Design*; Springer, 2001.

(58) Steinbrecher, T.; Mobley, D. L.; Case, D. A. Nonlinear scaling schemes for Lennard-Jones interactions in free energy calculations. *J. Chem. Phys.* **2007**, *127*, 214108.

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

- (60) 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*, 1219-1230.
- (61) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269-288.
- (62) Jain, A. N. Surflex-Dock 2.1: robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput. Aided Mol. Des.* **2007**, *21*, 281-306.
- (63) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins: Structure, Function, and Bioinformatics* **1999**, *37*, 228-241.
- (64) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: Applications of autodock. *Journal of Molecular Recognition* **1996**, *9*, 1-5.
- (65) Trott, O.; Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455-461.
- (66) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739-1749.
- (67) McGann, M. FRED pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2011**, *51*, 578-596.
- (68) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488-506.
- (69) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics* **2003**, *52*, 609-623.
- (70) Meng, E. C.; Shoichet, B. K.; Kuntz, I. D. Automated Docking with Grid-based Energy Evaluation. *J. Comput. Chem.* **1992**, *13*, 505-524.
- (71) Balias, T. E.; Mukherjee, S.; Rizzo, R. C. Implementation and Evaluation of a Docking-rescoring Method Using Molecular Footprint Comparisons. *J. Comput. Chem.* **2011**, *32*, 2273-2289.

- (72) Holden, P. M.; Kaur, H.; Goyal, R.; Gochin, M.; Rizzo, R. C. Footprint-based Identification of Viral Entry Inhibitors Targeting HIVgp41. *Bioorg. Med. Chem. Lett.* **2012**, *22*, 3011-3016.
- (73) Shoichet, B. K.; Kuntz, I. D.; Bodian, D. L. Molecular docking using shape descriptors. *J. Comput. Chem.* **1992**, *13*, 380-397.
- (74) Liu, H.-Y.; Kuntz, I. D.; Zou, X. Pairwise GB/SA Scoring Function for Structure-based Drug Design. *J. Phys. Chem. B* **2004**, *108*, 5453-5462.
- (75) Zou, X.; Yaxiong; Kuntz, I. D. Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model. *J. Am. Chem. Soc.* **1999**, *121*, 8033-8043.
- (76) Hawkins, G. D.; Cramer, C. J.; Truhlar, D. G. Pairwise solute descreening of solute charges from a dielectric medium. *Chemical Physics Letters* **1995**, *246*, 122-129.
- (77) 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. Calculating Structures and Free Energies of Complex Molecules: Combining Molecular Mechanics and Continuum Models. *Accounts of Chemical Research* **2000**, *33*, 889-897.
- (78) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate–DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401-9409.
- (79) Graves, A. P.; Shivakumar, D. M.; Boyce, S. E.; Jacobson, M. P.; Case, D. A.; Shoichet, B. K. Rescoring docking hit lists for model cavity sites: predictions and experimental testing. *J. Mol. Biol.* **2008**, *377*, 914-934.
- (80) Balius, T. E.; Allen, W. J.; Mukherjee, S.; Rizzo, R. C. Grid-based Molecular Footprint Comparison Method for Docking and De Novo Design: Application to HIVgp41. *J. Comput. Chem.* **2013**, *34*, 1226-1240.
- (81) Wu, E. L.; Cheng, X.; Jo, S.; Rui, H.; Song, K. C.; Dávila-Contreras, E. M.; Qi, Y.; Lee, J.; Monje-Galvan, V.; Venable, R. M.; Klauda, J. B.; Im, W. CHARMM-GUI Membrane Builder toward realistic biological membrane simulations. *J. Comput. Chem.* **2014**, *35*, 1997-2004.
- (82) 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. *J. Comput. Aided Mol. Des.* **2006**, *20*, 601-619.
- (83) Klebe, G. Virtual Ligand Screening: Strategies, Perspectives and Limitations. *Drug Discov. Today* **2006**, *11*, 580-594.

- (84) Berger, W. T.; Ralph, B. P.; Kaczocha, M.; Sun, J.; Balius, T. E.; Rizzo, R. C.; Haj-Dahmane, S.; Ojima, I.; Deutsch, D. G. Targeting Fatty Acid Binding Protein (FABP) Anandamide Transporters - A Novel Strategy for Development of Anti-inflammatory and Anti-nociceptive Drugs. *PLoS one* **2012**, *7*, e50968.
- (85) Ehrlich, P. Über die Constitution des Diphtheriegiftes. *Deut. Med. Wochschr.* **1898**, *24*, 597-600.
- (86) Ehrlich, P. Über den jetzigen Stand der Chemotherapie. *Ber. Dtsch. Chem. Ges.* **1909**, *42*, 17-47.
- (87) Güner, O. F.; Bowen, J. P. Setting the Record Straight: The Origin of the Pharmacophore Concept. *J. Chem. Inf. Model.* **2014**, *54*, 1269-1283.
- (88) Wermuth, C. G.; Ganellin, C. R.; Lindberg, P.; Mitscher, L. A. Glossary of Terms Used in Medicinal Chemistry (IUPAC Recommendations, 1998). *Pure Appl. Chem.* **1998**, *70*, 1129-1143.
- (89) Leach, A. R.; Gillet, V. J.; Lewis, R. A.; Taylor, R. Three-dimensional Pharmacophore Methods in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 539-558.
- (90) Yang, S. Y. Pharmacophore Modeling and Applications in Drug Discovery: Challenges and Recent Advances. *Drug Discov. Today* **2010**, *15*, 444-450.
- (91) Sanders, M. P. A.; Barbosa, A. J. M.; Zarzycka, B.; Nicolaes, G. A. F.; Klomp, J. P. G.; de Vlieg, J.; Del Rio, A. Comparative Analysis of Pharmacophore Screening Tools. *J. Chem. Inf. Model.* **2012**, *52*, 1607-1620.
- (92) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations among Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563-571.
- (93) Jones, G.; Willett, P.; Glen, R. C. A Genetic Algorithm for Flexible Molecular Overlay and Pharmacophore Elucidation. *J. Comput. Aided Mol. Des.* **1995**, *9*, 532-549.
- (94) Wolber, G.; Langer, T. LigandScout: 3-D Pharmacophores Derived from Protein-bound Ligands and Their Use as Virtual Screening Filters. *J. Chem. Inf. Model.* **2005**, *45*, 160-169.
- (95) Dixon, S. L.; Smondyrev, A. M.; Knoll, E. H.; Rao, S. N.; Shaw, D. E.; Friesner, R. A. PHASE: A New Engine for Pharmacophore Perception, 3D QSAR Model Development, and 3D Database Screening: 1. Methodology and Preliminary Results. *J. Comput. Aided Mol. Des.* **2006**, *20*, 647-671.

- (96) Richmond, N. J.; Abrams, C. A.; Wolohan, P. R.; Abrahamian, E.; Willett, P.; Clark, R. D. GALAHAD: 1. Pharmacophore Identification by Hypermolecular Alignment of Ligands in 3D. *J. Comput. Aided Mol. Des.* **2006**, *20*, 567-587.
- (97) Joseph-McCarthy, D.; Alvarez, J. C. Automated Generation of MCSS-derived Pharmacophoric DOCK Site Points for Searching Multiconformation Databases. *Proteins: Structure, Function, and Bioinformatics* **2003**, *51*, 189-202.
- (98) Joseph-McCarthy, D.; Thomas, B. E.; Belmarsh, M.; Moustakas, D.; Alvarez, J. C. Pharmacophore-based Molecular Docking to Account for Ligand Flexibility. *Proteins: Structure, Function, and Bioinformatics* **2003**, *51*, 172-188.
- (99) MOE; Version 2012.10, Chemical Computing Group Inc.: Montreal, Canada.
- (100) Mukherjee, S.; Balias, T. E.; Rizzo, R. C. Docking Validation Resources: Protein Family and Ligand Flexibility Experiments. *J. Chem. Inf. Model.* **2010**, *50*, 1986-2000.
- (101) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582-6594.
- (102) Shoichet, B. K.; Kuntz, I. D. Matching Chemistry and Shape in Molecular Docking. *Protein Engineering* **1993**, *6*, 723-732.
- (103) Tripos; Mol2 File Format: St. Louis, MO, 2009.
- (104) DOCK6.6, user manual, <http://dock.compbio.ucsf.edu>.
- (105) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605-1612.
- (106) ACD/ChemSketch, Version 14.01, Advanced Chemistry Development, Inc., Toronto, ON, Canada, <http://www.acdlabs.com>, **2014**.
- (107) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3-26.
- (108) Bunke, H.; Shearer, K. A Graph Distance Metric Based on the Maximal Common Subgraph. *Pattern Recogn. Lett.* **1998**, *19*, 255-259.

- (109) Bunke, H. Graph Matching: Theoretical Foundations, Algorithms, and Applications. *International Conference on Vision Interface* **2000**, 82-84.
- (110) Allen, W. J.; Rizzo, R. C. Implementation of the Hungarian Algorithm to Account for Ligand Symmetry and Similarity in Structure-based Design. *J. Chem. Inf. Model.* **2014**, *54*, 518-529.
- (111) DMS; UCSF Computer Graphics Laboratory: San Francisco, CA.
- (112) DesJarlais, R. L.; Sheridan, R. P.; Seibel, G. L.; Dixon, J. S.; Kuntz, I. D.; Venkataraghavan, R. Using Shape Complementarity as an Initial Screen in Designing Ligands for a Receptor Binding Site of Known Three-dimensional Structure. *J. Med. Chem.* **1988**, *31*, 722-729.
- (113) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789-6801.
- (114) 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**, *26*, 749-773.
- (115) Willett, P. Similarity-based approaches to virtual screening. *Biochemical Society transactions* **2003**, *31*, 603-606.
- (116) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708-1718.
- (117) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2003**, *44*, 170-178.
- (118) Balius, T. E.; Rizzo, R. C. Quantitative Prediction of Fold Resistance for Inhibitors of EGFR. *Biochemistry* **2009**, *48*, 8435-8448.
- (119) Holden, P. M.; Allen, W. J.; Gochin, M.; Rizzo, R. C. Strategies for Lead Discovery: Application of Footprint Similarity Targeting HIVgp41. *Bioorg. Med. Chem.* **2014**, *22*, 651-661.
- (120) Allen, W. J.; Rizzo, R. C. Computer-Aided Approaches for Targeting HIVgp41. *Biology* **2012**, *1*, 311-338.
- (121) He, Y.; Liu, S.; Li, J.; Lu, H.; Qi, Z.; Liu, Z.; Debnath, A. K.; Jiang, S. Conserved Salt Bridge Between the N- and C-terminal Heptad Repeat Regions of the Human Immunodeficiency



Virus Type 1 gp41 Core Structure Is Critical for Virus Entry and Inhibition. *Journal of virology* **2008**, *82*, 11129-11139.

(122) Hu, B.; Lill, M. A. Exploring the Potential of Protein-based Pharmacophore Models in Ligand Pose Prediction and Ranking. *J. Chem. Inf. Model.* **2013**, *53*, 1179-1190.

(123) Gardiner, E. J.; Cosgrove, D. A.; Taylor, R.; Gillet, V. J. Multiobjective Optimization of Pharmacophore Hypotheses: Bias Toward Low-energy Conformations. *J. Chem. Inf. Model.* **2009**, *49*, 2761-2773.

(124) Gisbert Schneider, M.-L. L., Martin Stahl, Petra Scheider De novo Design of Molecular Architectures by Evolutionary Assembly of Drug-Derived Building Blocks. *J. Comput. Aided Mol. Des.* **2000**, *14*, 487-494.

(125) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823-823.

(126) Yoshihiko Nishibata, A. I. Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation. *Tetrahedron* **1991**, *47*, 8885-8990.

(127) LeapFrog; Tripos: St. Louis, MO.

(128) Douguet, D.; Munier-Lehmann, H.; Labesse, G.; Pochet, S. LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design. *J. Med. Chem.* **2005**, *48*, 2457-2468.

(129) Jorgensen, W. L.; Ruiz-Caro, J.; Tirado-Rives, J.; Basavapathruni, A.; Anderson, K. S.; Hamilton, A. D. Computer-aided design of non-nucleoside inhibitors of HIV-1 reverse transcriptase. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 663-667.

(130) Jorgensen, W. L. Yale University, New Haven, CT (2004), **2004**.

(131) Yuan, Y.; Pei, J.; Lai, L. LigBuilder 2: a practical de novo drug design approach. *J. Chem. Inf. Model.* **2011**, *51*, 1083-1091.

(132) Renxiao Wang, Y. G., Luhua Lai LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. *Journal of Molecular Modeling* **2000**, *6*, 498-516.

(133) Schneider, G.; Fechner, U. Computer-based de novo design of drug-like molecules. *Nat. Rev. Drug Discov.* **2005**, *4*, 649-663.

(134) Allen, W. J.; Rizzo, R. C. Implementation of de novo DOCK. **2015**, *In preparation*.

- (135) Eckert, D. M.; Malashkevich, V. N.; Hong, L. H.; Carr, P. A.; Kim, P. S. Inhibiting HIV-1 Entry: Discovery of D-Peptide Inhibitors that Target the gp41 Coiled-Coil Pocket. *Cell* **1999**, *99*, 103-115.
- (136) Cai, L.; Jiang, S. Development of Peptide and Small-Molecule HIV-1 Fusion Inhibitors that Target gp41. *Chem. Med. Chem.* **2010**, *5*, 1813-1824.
- (137) Zheng, B.; Wang, K.; Lu, L.; Yu, F.; Cheng, M.; Jiang, S.; Liu, K.; Cai, L. Hydrophobic mutations in buried polar residues enhance HIV-1 gp41 N-terminal heptad repeat-C-terminal heptad repeat interactions and C-peptides' anti-HIV activity. *Aids* **2014**, *28*, 1251-1260.
- (138) Lu, L.; Tong, P.; Yu, X.; Pan, C.; Zou, P.; Chen, Y. H.; Jiang, S. HIV-1 variants with a single-point mutation in the gp41 pocket region exhibiting different susceptibility to HIV fusion inhibitors with pocket- or membrane-binding domain. *Biochimica et biophysica acta* **2012**, *1818*, 2950-2957.
- (139) Cooper, D. A.; Lange, J. M. A. Peptide inhibitors of virus?cell fusion: enfuvirtide as a case study in clinical discovery and development. *The Lancet Infectious Diseases* **2004**, *4*, 426-436.
- (140) Kilby, J. M.; Eron, J. J. Novel Therapies Based on Mechanisms of HIV-1 Cell Entry. *New England Journal of Medicine* **2003**, *348*, 2228-2238.
- (141) Buzon, V.; Natrajan, G.; Schibli, D.; Campelo, F.; Kozlov, M. M.; Weissenhorn, W. Crystal structure of HIV-1 gp41 including both fusion peptide and membrane proximal external regions. *PLoS Pathog.* **2010**, *6*, e1000880.
- (142) Wensing, A. M.; Calvez, V.; Günthard, H. F.; Johnson, V. A.; Paredes, R.; Pillay, D.; Shafer, R. W.; Richman, D. D. Special Contribution 2014 Update of the Drug Resistance Mutations in HIV-1. *Topics in Antiviral Medicine* **2014**, *22*, 642-650.
- (143) Caffrey, M. Model for the structure of the HIV gp41 ectodomain: insight into the intermolecular interactions of the gp41 loop. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **2001**, *1536*, 116-122.
- (144) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926-935.
- (145) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An  $N \log(N)$  method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089.
- (146) Dickson, C. J.; Madej, B. D.; Skjevik, A. A.; Betz, R. M.; Teigen, K.; Gould, I. R.; Walker, R. C. Lipid14: The Amber Lipid Force Field. *J. Chem. Theory Comput.* **2014**, *10*, 865-879.

## Appendix A. FMS-guided DOCKing Protocol

Dock input files for FMS and FMS+SGE guided standard molecular docking and *de novo* DOCK are provided below. This version of DOCK 6 that incorporates a new descriptor score to combine individual DOCK scores and FMS score will be released in DOCK6.8.

For FMS-guided rescoring, the sample DOCK input file used is shown below.

```
conformer_search_type          rigid
use_internal_energy            no
ligand_atom_file               ligand.mol2
limit_max_ligands              no
skip_molecule                 no
read_mol_solvation             no
calculate_rmsd                 no
use_database_filter            no
orient_ligand                  no
bump_filter                    no
score_molecules                 yes
contact_score_primary          no
contact_score_secondary        no
grid_score_primary             no
grid_score_secondary           no
multigrid_score_primary        no
multigrid_score_secondary      no
dock3.5_score_primary          no
dock3.5_score_secondary        no
continuous_score_primary       no
continuous_score_secondary     no
footprint_similarity_score_primary no
footprint_similarity_score_secondary no
Ph4_primary                    yes
Ph4_secondary                  no
use_ph4_ref_mol2               yes
Ph4_ref_mol2_filename          ph4_ref.mol2
write_out_reference_ph4        no
write_out_candidate_ph4        no
write_out_matched_ph4          no
ph4_compare_type               o
ph4_full_match                  yes
descriptor_score_secondary     no
gbsa_zou_score_secondary       no
gbsa_hawkins_score_secondary   no
SASA_descriptor_score_secondary no
amber_score_secondary           no
minimize_ligand                no
atom_model                     all
```

vdw_defn_file	vdw_AMBER_parm99.defn
flex_defn_file	flex.defn
flex_drive_file	flex_drive.tbl
ph4_defn_file	ph4.defn
ligand_outfile_prefix	FMS_output_re
write_orientations	no
num_scored_conformers	1
rank_ligands	no

For FMS+SGE-guided rescoring using descriptor score, the sample DOCK input file used is shown

below.

conformer_search_type	rigid
use_internal_energy	no
ligand_atom_file	ligand.mol2
limit_max_ligands	no
skip_molecule	no
read_mol_solvation	no
calculate_rmsd	no
use_database_filter	no
orient_ligand	no
bump_filter	no
score_molecules	yes
contact_score_primary	no
contact_score_secondary	no
grid_score_primary	no
grid_score_secondary	no
multigrid_score_primary	no
multigrid_score_secondary	no
dock3.5_score_primary	no
dock3.5_score_secondary	no
continuous_score_primary	no
continuous_score_secondary	no
footprint_similarity_score_primary	no
footprint_similarity_score_secondary	no
Ph4_secondary	no
descriptor_score_primary	yes
descriptor_score_secondary	no
descriptor_use_grid_score	yes
descriptor_use_tanimoto	yes
descriptor_use_pharmacophore_score	yes
descriptor_grid_score_rep_rad_scale	1
descriptor_grid_score_vdw_scale	1
descriptor_grid_score_es_scale	1
descriptor_grid_score_grid_prefix	receptor
descriptor_fingerprint_ref_filename	tan_ref.mol2
use_ph4_ref_mol2	yes
Ph4_ref_mol2_filename	ph4_ref.mol2
ph4_compare_type	o
ph4_full_match	yes
descriptor_weight_grid_score	1
descriptor_weight_fingerprint_tanimoto	0

descriptor_weight_pharmacophore_tanimoto	10
gbsa_zou_score_secondary	no
gbsa_hawkins_score_secondary	no
SASA_descriptor_score_secondary	no
amber_score_secondary	no
minimize_ligand	no
atom_model	all
vdw_defn_file	vdw_AMBER_parm99.defn
flex_defn_file	flex.defn
flex_drive_file	flex_drive.tbl
ph4_defn_file	ph4.defn
ligand_outfile_prefix	FMS+SGE_output_re
write_orientations	no
num_scored_conformers	1
rank_ligands	no

For FMS-guided standard flexible ligand docking (FLX), the sample DOCK input file used is shown below.

conformer_search_type	flex
user_specified_anchor	no
limit_max_anchors	no
min_anchor_size	5
pruning_use_clustering	yes
pruning_max_orients	1000
pruning_clustering_cutoff	100
pruning_conformer_score_cutoff	100.0
use_clash_overlap	no
write_growth_tree	no
write_fragment_libraries	no
use_internal_energy	yes
internal_energy_rep_exp	12
ligand_atom_file	ligand.mol2
limit_max_ligands	no
skip_molecule	no
read_mol_solvation	no
calculate_rmsd	yes
use_rmsd_reference_mol	no
use_database_filter	no
orient_ligand	yes
automated_matching	yes
receptor_site_file	spheres.sph
max_orientations	1000
critical_points	no
chemical_matching	no
use_ligand_spheres	no
bump_filter	no
score_molecules	yes
contact_score_primary	no
contact_score_secondary	no
grid_score_primary	no

grid_score_secondary	no
dock3.5_score_primary	no
dock3.5_score_secondary	no
continuous_score_primary	no
continuous_score_secondary	no
footprint_similarity_score_primary	no
footprint_similarity_score_secondary	no
Ph4_primary	yes
Ph4_secondary	no
use_ph4_ref_mol2	yes
Ph4_ref_mol2_filename	ph4_ref.mol2
ph4_compare_type	o
descriptor_score_secondary	no
gbsa_zou_score_secondary	no
gbsa_hawkins_score_secondary	no
amber_score_secondary	no
minimize_ligand	yes
minimize_anchor	yes
minimize_flexible_growth	yes
use_advanced_simplex_parameters	no
simplex_max_cycles	1
simplex_score_converge	0.1
simplex_cycle_converge	1.0
simplex_trans_step	1.0
simplex_rot_step	0.1
simplex_tors_step	10.0
simplex_anchor_max_iterations	500
simplex_grow_max_iterations	500
simplex_grow_tors_premin_iterations	0
simplex_random_seed	0
simplex_restraint_min	no
atom_model	all
vdw_defn_file	vdw_AMBER_parm99.defn
flex_defn_file	flex.defn
flex_drive_file	flex_drive.tbl
ph4_defn_file	ph4.defn
ligand_outfile_prefix	
flex.dock2grid.orient.FMS	
write_orientations	no
num_scored_conformers	5000
write_conformations	no
cluster_conformations	yes
cluster_rmsd_threshold	2.0
rank_ligands	no

For FMS+SGE-guided standard flexible ligand docking (FLX), the sample DOCK input file used is shown below.

conformer_search_type	flex
user_specified_anchor	no
limit_max_anchors	no
min_anchor_size	5

pruning_use_clustering	yes
pruning_max_orients	1000
pruning_clustering_cutoff	100
pruning_conformer_score_cutoff	100.0
use_clash_overlap	no
write_growth_tree	no
write_fragment_libraries	no
use_internal_energy	yes
internal_energy_rep_exp	12
ligand_atom_file	ligand.mol2
limit_max_ligands	no
skip_molecule	no
read_mol_solvation	no
calculate_rmsd	yes
use_rmsd_reference_mol	no
use_database_filter	no
orient_ligand	yes
automated_matching	yes
receptor_site_file	spheres.sph
max_orientations	1000
critical_points	no
chemical_matching	no
use_ligand_spheres	no
bump_filter	no
score_molecules	yes
contact_score_primary	no
contact_score_secondary	no
grid_score_primary	no
grid_score_secondary	no
dock3.5_score_primary	no
dock3.5_score_secondary	no
continuous_score_primary	no
continuous_score_secondary	no
footprint_similarity_score_primary	no
footprint_similarity_score_secondary	no
Ph4_primary	no
Ph4_secondary	no
descriptor_score_primary	yes
descriptor_score_secondary	no
descriptor_use_grid_score	yes
descriptor_use_tanimoto	no
descriptor_use_pharmacophore_score	yes
descriptor_grid_score_rep_rad_scale	1
descriptor_grid_score_vdw_scale	1
descriptor_grid_score_es_scale	1
descriptor_grid_score_grid_prefix	receptor
use_ph4_ref_mol2	yes
Ph4_ref_mol2_filename	ph4_ref.mol2
ph4_compare_type	o
ph4_full_match	yes
descriptor_weight_grid_score	1
descriptor_weight_pharmacophore_tanimoto	10
gbsa_zou_score_secondary	no
gbsa_hawkins_score_secondary	no
amber_score_secondary	no

minimize_ligand	yes
minimize_anchor	yes
minimize_flexible_growth	yes
use_advanced_simplex_parameters	no
simplex_max_cycles	1
simplex_score_converge	0.1
simplex_cycle_converge	1.0
simplex_trans_step	1.0
simplex_rot_step	0.1
simplex_tors_step	10.0
simplex_anchor_max_iterations	500
simplex_grow_max_iterations	500
simplex_grow_tors_premin_iterations	0
simplex_random_seed	0
simplex_restraint_min	no
atom_model	all
vdw_defn_file	vdw_AMBER_parm99.defn
flex_defn_file	flex.defn
flex_drive_file	flex_drive.tbl
ph4_defn_file	ph4.defn
ligand_outfile_prefix	
flex.dock2grid.orient.FMS+SGE	
write_orientations	no
num_scored_conformers	5000
write_conformations	no
cluster_conformations	yes
cluster_rmsd_threshold	2.0
rank_ligands	no

For FMS-guided *de novo* growth, the sample DOCK input file used is shown below.

conformer_search_type	denovo
dn_fraglib_scaffold_file	fraglib_scaffold.mol2
dn_fraglib_linker_file	fraglib_linker.mol2
dn_fraglib_sidechain_file	fraglib_sidechain.mol2
dn_fraglib_rigid_file	fraglib_rigid.mol2
dn_user_specified_anchor	yes
dn_fraglib_anchor_file	fraglib_anchor.mol2
dn_use_torenv_table	yes
dn_torenv_table	fraglib_torenv.dat
dn_sampling_method	graph
dn_graph_starting_points	10
dn_graph_breadth	5
dn_graph_depth	2
dn_graph_temperature	100
dn_constraint_mol_wt	1000
dn_constraint_rot_bon	15
dn_tanimoto_cutoff	1
dn_heur_unmatched_num	0
dn_heur_matched_rmsd	2.0
dn_max_grow_layers	7
dn_max_current_aps	7
dn_max_root_size	50



dn_max_layer_size	50
dn_write_checkpoints	yes
dn_output_prefix	FMS_denovo.final
use_internal_energy	yes
internal_energy_rep_exp	12
use_database_filter	no
orient_ligand	yes
automated_matching	yes
receptor_site_file	spheres.sph
max_orientations	10000
critical_points	no
chemical_matching	no
use_ligand_spheres	no
bump_filter	no
score_molecules	yes
contact_score_primary	no
contact_score_secondary	no
grid_score_primary	no
grid_score_secondary	no
multigrid_score_primary	no
multigrid_score_secondary	no
dock3.5_score_primary	no
dock3.5_score_secondary	no
continuous_score_primary	no
continuous_score_secondary	no
footprint_similarity_score_primary	no
footprint_similarity_score_secondary	no
Ph4_primary	yes
Ph4_secondary	no
use_ph4_ref_mol2	yes
Ph4_ref_mol2_filename	ph4_ref.mol2
ph4_compare_type	o
ph4_full_match	yes
descriptor_score_secondary	no
gbsa_zou_score_secondary	no
gbsa_hawkins_score_secondary	no
SASA_descriptor_score_secondary	no
amber_score_secondary	no
minimize_ligand	yes
minimize_anchor	yes
minimize_flexible_growth	yes
use_advanced_simplex_parameters	no
simplex_max_cycles	1
simplex_score_converge	0.1
simplex_cycle_converge	1.0
simplex_trans_step	1.0
simplex_rot_step	0.1
simplex_tors_step	10.0
simplex_anchor_max_iterations	500
simplex_grow_max_iterations	500
simplex_grow_tors_premin_iterations	0
simplex_random_seed	0
simplex_restraint_min	no
atom_model	all
vdw_defn_file	vdw.defn

flex_defn_file	flex.defn
flex_drive_file	flex_drive.tbl
ph4_defn_file	ph4.defn

For FMS+SGE-guided *de novo* growth, the sample DOCK input file used is shown below.

conformer_search_type	denovo
dn_fraglib_scaffold_file	fraglib_scaffold.mol2
dn_fraglib_linker_file	fraglib_linker.mol2
dn_fraglib_sidechain_file	fraglib_sidechain.mol2
dn_fraglib_rigid_file	fraglib_rigid.mol2
dn_user_specified_anchor	yes
dn_fraglib_anchor_file	fraglib_anchor.mol2
dn_use_torenv_table	yes
dn_torenv_table	fraglib_torenv.dat
dn_sampling_method	graph
dn_graph_starting_points	10
dn_graph_breadth	5
dn_graph_depth	2
dn_graph_temperature	100
dn_constraint_mol_wt	1000
dn_constraint_rot_bon	15
dn_tanimoto_cutoff	1
dn_heur_unmatched_num	0
dn_heur_matched_rmsd	2.0
dn_max_grow_layers	7
dn_max_current_aps	7
dn_max_root_size	50
dn_max_layer_size	50
dn_write_checkpoints	yes
dn_output_prefix	FMS+SGE_denovo.final
use_internal_energy	yes
internal_energy_rep_exp	12
use_database_filter	no
orient_ligand	yes
automated_matching	yes
receptor_site_file	spheres.sph
max_orientations	10000
critical_points	no
chemical_matching	no
use_ligand_spheres	no
bump_filter	no
score_molecules	yes
contact_score_primary	no
contact_score_secondary	no
grid_score_primary	no
grid_score_secondary	no
multigrid_score_primary	no
multigrid_score_secondary	no
dock3.5_score_primary	no
dock3.5_score_secondary	no
continuous_score_primary	no
continuous_score_secondary	no

footprint_similarity_score_primary	no
footprint_similarity_score_secondary	no
Ph4_primary	no
Ph4_secondary	no
descriptor_score_primary	yes
descriptor_score_secondary	no
descriptor_use_grid_score	yes
descriptor_use_pharmacophore_score	yes
descriptor_use_tanimoto	no
descriptor_use_hungarian	no
descriptor_grid_score_rep_rad_scale	1
descriptor_grid_score_vdw_scale	1
descriptor_grid_score_es_scale	1
descriptor_grid_score_grid_prefix	receptor
use_ph4_ref_mol2	yes
Ph4_ref_mol2_filename	ph4_ref.mol2
ph4_compare_type	o
ph4_full_match	yes
descriptor_weight_grid_score	1
descriptor_weight_pharmacophore_tanimoto	1
gbsa_zou_score_secondary	no
gbsa_hawkins_score_secondary	no
SASA_descriptor_score_secondary	no
amber_score_secondary	no
minimize_ligand	yes
minimize_anchor	yes
minimize_flexible_growth	yes
use_advanced_simplex_parameters	no
simplex_max_cycles	1
simplex_score_converge	0.1
simplex_cycle_converge	1.0
simplex_trans_step	1.0
simplex_rot_step	0.1
simplex_tors_step	10.0
simplex_anchor_max_iterations	500
simplex_grow_max_iterations	500
simplex_grow_tors_premin_iterations	0
simplex_random_seed	0
simplex_restraint_min	no
atom_model	all
vdw_defn_file	vdw.defn
flex_defn_file	flex.defn
flex_drive_file	flex_drive.tbl
ph4_defn_file	ph4.defn

## Appendix B. Visualization of Pharmacophore Models

This section describes the general procedure to visualize pharmacophore models generated by the FMS protocol using Chimera *bild* files<sup>105</sup> as introduced in Chapter 2.

First, the following DOCK input parameters need to be set. By default, the parameters “write\_out\_reference\_ph4”, “write\_out\_candidate\_ph4” and “write\_out\_matched\_ph4” are set to “no” and no output pharmacophore *mol2* file will be generated.

write_out_reference_ph4	yes
reference_ph4_out_filename	ref_ph4.mol2
write_out_candidate_ph4	yes
candidate_ph4_out_filename	cad_ph4.mol2
write_out_matched_ph4	yes
matched_ph4_out_filename	mat_ph4.mol2

In the pharmacophore output *mol2* files “ref\_ph4.mol2”, “cad\_ph4.mol2” and “mat\_ph4.mol2”, each pharmacophore point is represented by a set of atoms (Table B-1, column c). Note for HBD, HBA, ARO and RNG labeled pharmacophore points, more than one atom is used because both the center of the point (denoted by the first atom in the list) and the directionality (derived from all the atoms in the list) need to be recorded. The output pharmacophore *mol2* file will then be converted to a *bild* file using a python script *mol2bild.py* (to be released in DOCK6.8). The *bild* file contains the directional vectors derived for all the pharmacophore points with directionality (HBD, HBA, ARO, RNG) in the original *mol2* file. Different colors are used to represent different pharmacophore labels as shown in Table B-1 column d. The directional vectors are modeled as 3D arrow geometric objects in the molecular modeling software Chimera, each consists of a cylinder (from the start point to an intermediary point) and a cone (from the

intermediary point to the end point) representing the arrowhead. As an example, a 3D arrow for a ring pharmacophore feature is written as follows in the *bild* file.

```
.color orange
.arrow 0.64 -17.12 -10.61 0.99 -17.69 -9.86 0.01 0.04 0.75
.color orange
.arrow 0.64 -17.12 -10.61 0.30 -16.55 -11.36 0.01 0.04 0.75
.color orange
.arrow -0.04 -15.65 -9.18 0.30 -16.22 -8.43 0.01 0.04 0.75
.color orange
.arrow -0.04 -15.65 -9.18 -0.39 -15.09 -9.93 0.01 0.04 0.75
.color orange
.arrow -0.22 -16.30 -4.16 -0.73 -16.46 -3.32 0.01 0.04 0.75
.color orange
.arrow -0.22 -16.30 -4.16 0.29 -16.15 -5.01 0.01 0.04 0.75
.color orange
.arrow -1.30 -14.47 -4.50 -1.81 -14.62 -3.65 0.01 0.04 0.75
.color orange
.arrow -1.30 -14.47 -4.50 -0.79 -14.33 -5.35 0.01 0.04 0.75
.color red
.arrow -7.12 -15.42 -2.60 -7.25 -14.95 -1.73 0.01 0.04 0.75
.color red
.arrow -7.11 -15.54 -4.81 -7.24 -15.16 -5.73 0.01 0.04 0.75
.color blue
.arrow 2.65 -16.61 -11.14 1.71 -16.86 -10.90 0.01 0.04 0.75
.color blue
.arrow 1.53 -15.84 -3.04 0.72 -16.07 -3.57 0.01 0.04 0.75
```

Here, each Chimera object is defined by two lines. The first line defines the color of the Chimera object (e.g. “.color orange”). The second line defines the type of the geometric object (“.arrow”), the start and end point of the arrow ( $(x1, y1, z1) = (0.64, -17.12, -10.61)$  and  $(x2, y2, z2) = (0.99, -17.69, -9.86)$  in the first object in the above example, arrow pointing from  $(x1, y1, z1)$  to  $(x2, y2, z2)$ ), as well as the size of the arrow (radius of the cylinder  $r1 = 0.01$ , the radius of the base of the cone  $r2 = 0.04$ , and the ratio of the length of cylinder to that of the complete arrow  $rho = 0.75$ ). For the most recent description, please consult the Chimera website (<http://www.cgl.ucsf.edu/chimera/>).

**Table B-1.** Pharmacophore feature represented in ph4.mol2 and ph4.bild.

<b>a. Label</b>	<b>b. definition</b>	<b>c. mol2</b>	<b>d. bild</b>
PHO	Hydrophobic	C	-
HBD	Hydrogen bond donor	HD, N	blue
HBA	Hydrogen bond acceptor	O, HA	red
ARO	Aromatic ring	S, H1, H2	orange
RNG	Non-aromatic ring	P, H3, H4	yellow
POS	Positively charged	Na	-
NEG	Negatively charged	Cl	-

The complete *bild* files include all the directional vectors derived from the pharmacophore mol2 files. Both the pharmacophore *mol2* and *bild* files will be opened in Chimera for visualization of the pharmacophore model, as shown in Figure 2-1, Table 2-1 and Figure 2-18 in Chapter 2.

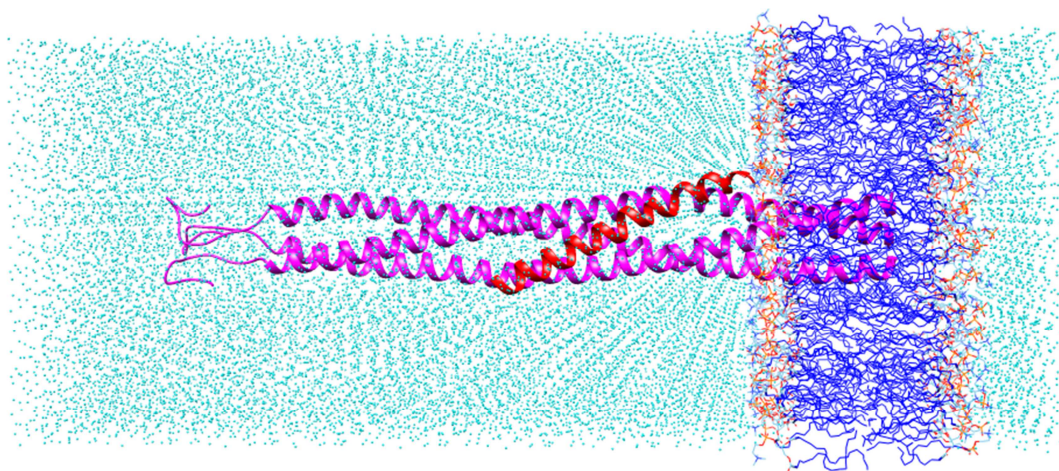
## Appendix C. Lipid-Bound HIVgp41-T20 Complex Simulations

This section outlines the DOPC lipid-bound HIVgp41-T20 complex (as visualized in Figure C-1) construction steps using the CHARMM-GUI lipid builder<sup>81</sup> and simulation protocols using *pmemd.cuda* in AMBER14 on GPU-accelerated machines. The peptide ligand (red helix in Figure C-1) and protein (magenta helices in Figure C-1) structures are obtained as described in Chapter 5.

First, the gas phase protein-ligand complex structure is uploaded to the CHARMM-GUI lipid builder to generate the lipid bilayer. “Heterogeneous Lipid” system with “Rectangular” box type is chosen. The length of Z axis for the solvent box is determined by assigning the “water thickness”, which is defined as the minimum water height on top and bottom of the complex system, to “17.5”. In order to obtain a reasonable conformation with the fusion peptide region correctly embedded in between the lipid bilayer, several values of the NHR insertion depth were tested and a final value of  $-75.5 \text{ \AA}$  was selected. To achieve similar lipid density in the upper and lower lipid layers, the “number of lipid components” is set to 80 for the upper leaflet and 74 for the lower leaflet. No ions are added for the initial test. The solvated lipid-bound system is pre-equilibrated briefly in the CHARMM-GUI platform and final structure is downloaded.

The solvent box along with the lipid bilayer from the resulting structure was saved as a separate pdb file and later shifted to match the original gas phase complex variant models using molecular modeling software *Chimera*. Also, the CHARMM-GUI PDB format structure is converted into AMBER compatible format using a python script *charmm lipid2amber.py* provided in AMBERTools14. Another bash script *vmd\_box\_dims.sh* is used to estimate the periodic box dimensions by measuring the range of water molecules’ coordinates. The size of the water box for

the complex system (320 residues) is set to  $80 \times 80 \times 1190 \text{ \AA}^3$  (24399 TIP3P waters, 154 DOPC lipids) for MD simulations in AMBER14. Force field ff99SB is employed for proteins, Lipid14 for lipids and TIP3P for explicit waters. The resulting starting conformations were used for 200ns-long molecular dynamic simulations. The 9-step equilibration and minimization protocols are identical to that described in Chapter 5. 20 production runs (10ns each) of MD simulations were performed with restraint weight of  $0.5 \text{ kcal} \cdot \text{mol}^{-1} \cdot \text{ \AA}^{-2}$  on TM region (residues 82-94 on NHR1, residues 176-188 on NHR2, and residues 270-282 on NHR3, here the residue numbers are from the actual complex file) heavy atoms. The average simulation time is about 19ns/day per GPU card. Future tests and studies can be done to further refine the lipid-bound structure and optimize the molecular dynamics simulation protocols for better initial setup and equilibration of the solvated complex system. Energetic and structural analyses specifically on the lipids can uncover the role of lipids in the mechanism of T20 binding.



**Figure C-1.** Membrane-bounded HIVgp41-T20 complex prepared with CHARMM-GUI and AMBER14.