2017

Improving the accuracy and efficiency of docking methods

BOSTON UNIVERSITY

COLLEGE OF ENGINEERING

Dissertation

IMPROVING THE ACCURACY AND EFFICIENCY OF DOCKING METHODS

by

BING XIA

B.A., University of California, Berkeley, 2010
M.S., Boston University, 2015

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2017

Approved by

First Reader ____
Sandor Vajda, Ph.D.
Professor of Biomedical Engineering
Professor of Systems Engineering
Professor of Chemistry

Second Reader ____
Dmytro Kozakov, Ph.D.
Assistant Professor of Applied Mathematics and Statistics
Stony Brook University

Third Reader ____
Maxim D. Frank-Kamenetskii, Ph.D.
Professor of Biomedical Engineering
Professor of Materials Science and Engineering

Fourth Reader ____
Ioannis Ch. Paschalidis, Ph.D.
Professor of Electrical and Computer Engineering
Professor of Systems Engineering
Professor of Biomedical Engineering

Fifth Reader ____
Karen N. Allen, Ph.D.
Professor of Chemistry

DEDICATION

To my family

ACKNOWLEDGMENTS

I would like thank all the past and present members of the Structural Bioinformatics lab I have had the pleasure of working with during my time here at Boston University. I would especially like to thank David Hall for all the time has spent sharing his extensive knowledge of computational chemistry, when he was under no obligation to do so.

I am grateful for the chance that my advisors, Sandor Vajda and Dima Kozakov, gave me to work in this lab. You have given me so much help in guiding my research, and I have learned a lot during my time in this lab.

To my dear friends who have kept me sane, happy, and motivated over the years, thank you! I couldn’t have done this without you guys :)

IMPROVING THE ACCURACY AND EFFICIENCY OF DOCKING METHODS

BING XIA

Boston University College of Engineering, 2017

Major Professor: Sandor Vajda, Ph.D., Professor of Biomedical Engineering, Professor of Systems Engineering, Professor of Chemistry

ABSTRACT

Computational methods for predicting macromolecular complexes are useful tools for studying biological systems. They are used in areas such as drug design and for studying protein-protein interactions. While considerable progress has been made in this field over the decades, enhancing the speed and accuracy of these computational methods remains an important challenge. This work describes two different enhancements to the accuracy of ClusPro, a method for performing protein-protein docking, as well as an enhancement to the efficiency of global rigid body docking. SAXS is a high throughput technique collected for molecules in solution, and the data provides information about the shape and size of molecules. ClusPro was enhanced with the ability to SAXS data collected for protein complexes to guide docking by selecting conformations by how well they match the experimental data, which improved docking accuracy when such data is available. Various other experimental techniques, such as NMR, FRET, or chemical cross linking can provide information about protein-protein interfaces, and such information can be used to generate distance-based restraints between pairs of residues across the interface. A second enhancement to ClusPro enables the use of such distance restraints to improve docking accuracy. Finally, an enhancement to the efficiency of FFT based global docking programs was developed. This enhancement allows for the efficient search of multiple sidechain conformations, and this improved program was applied to the flexible computational solvent mapping program FTFlex.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CAPRI	Critical Assessment of Predicted Interfaces
FFT	Fast Fourier Transform
HADDOCK	High Ambiguity Driven biomolecular DOCKing
JSON	Javascript Object Notation
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
NOE	Nuclear Overhauser Effect
RMSD	Root-Square-Mean Deviation
SAXS	Small Angle X-ray Scattering

CHAPTER ONE: Background and Introduction

1.1 Motivation

Protein molecules are central to the functioning of a cell. They serve in roles from metabolism, to signal transduction, to DNA replication. In performing their functions in the cell, proteins often work together in complexes by interacting with other proteins or with nucleic acids. The three-dimensional structures of these complexes are frequently crucial for the mechanistic understanding of cell function. While techniques for determining the structures of these complexes have advanced significantly in recent years, it remains a challenging task, and the structure of complexes are still more challenging to obtain than the structures of unbound proteins. Thus, computational methods for predicting the structures of complexes are an important alternative method for studying proteins for many systems of interest.

Methods designed to predict the structures of protein complexes are not new. Since 2001, the CAPRI competition has been a way to evaluate the progress of protein docking methods. The results of CAPRI indicate substantial recent progress in methodology, including improved performance of automated docking methods (Lensink, Méndez, & Wodak, 2007; Lensink & Wodak, 2010, 2013). While such advances are very promising, it is still true that ab initio docking does not work well for some systems. In these cases, researchers often supplement the docking predictions generated by automated methods with knowledge about the system from other sources, such as chemical cross linking or NMR data. Automated methods that can use data from additional experimental sources can thus improve docking results. Chapters 2 and 3 describe extensions to an existing method to make use of additional experimental data to improve the accuracy of predictions.

In addition to the accuracy of docking methods, efficiency of these methods is another important consideration. While the constant growth of computational power has made many methods that used to be too expensive to run feasible, algorithmic advances that improve the scaling of running times are still important. Chapter 4 describes developments that improve the efficiency of rigid body sampling methods when multiple sidechain conformations are to be searched.

1.2 Background Methods

1.2.1 Global rigid-body docking using FFT

Methods for predicting the structure of protein complexes can be generally described as energy minimization methods. Given the structures of two unbound molecules, the goal is to find the structure of the complex, which is understood to be the structure with lowest free energy. There exist many different methods that pursue different strategies for predicting the structure of the complex. Sequence based methods use the sequence of the two proteins to search for similar complexes where the structure is known. By leveraging the information from homologous proteins, such methods are usually fast and accurate, assuming there are complexes where the homologs are similar enough. On the other hand, MD methods explicitly simulate the atoms in the proteins, which make it possible to study novel systems where there are no homologs with known structure. However, these simulations are costly to run, even on modern hardware.

This work is based on the rigid-body global sampling program, PIPER, which has been the object of much study (Brenke et al., 2012; Chuang, Kozakov, Brenke, Comeau, & Vajda, 2008; Kozakov, Brenke, Comeau, & Vajda, 2006). PIPER uses a FFT based approach to do a global rigid-body search of all possible relative orientations of one protein with respect to the other protein. It has been applied to both protein-protein systems and protein-small molecule systems. In FFT based methods, the scoring function is a sum of P different correlation functions. For each rotation of one of the molecules, termed the ligand, the score of a relative translation (α, β, γ) is given by the equation

$$E(\alpha, \beta, \gamma) = \sum_{p=1}^{P} \sum_{l,m,n} R_p(l, m, n) L_p(l + \alpha, m + \beta, n + \gamma)$$

where R_p and L_p are the components of the scoring function defined for the receptor and ligand, respectively. If we choose these component functions carefully, we end up with a scoring function that represents a pseudo-energy value of the configuration of the molecular complex.

This sum of correlation functions can be efficiently calculated using P forward Fourier transforms and one reverse transform, by rewriting the right-hand side of the previous equation as $E(\alpha, \beta, \gamma) = IFT\{\sum_{p=1}^{P} FT^*(R_p)FT(L_p)\}(\alpha, \beta, \gamma)$ (Katchalski-Katzir et al., 1992). The component functions R_p and L_p are computed on a grid on size (N_1, N_2, N_3), but if we make the simplifying assumption that these three values are roughly the same, then efficiency of the naïve approach is $O(N^6)$. The application of FFT reduces this to $O(N^3 \log(N^3))$, which was a great algorithm advance, making global rigid body methods feasible. Chapter 4 describes a method for decomposing the scoring function used in PIPER in a such a way as to allow for efficient search of multiple sidechain conformations.

PIPER takes as input a set of rotation matrices to apply to the ligand. For protein-protein systems, we use a set of 70,000 rotations that are quasi-uniformly distributed over the set of all Euler rotations with a grid size of 1 Å. For protein-small molecule systems, we use a smaller set of 500 rotations with a grid size of 0.8 Å. PIPER then produces as output one or more relative translations of the ligand with respect to the receptor which minimizes the scoring function for each rotation. Thus, we obtain a set of conformations of the ligand, which produce low values for the scoring function. Depending on the type of system we are trying to dock, we take between 500 to 2000 of the lowest energy conformations and apply a greedy RMSD based clustering algorithm to obtain a final set of 30 to 50 predictions, each of which corresponds to a low energy basin of conformations of the complex.

1.2.3 Computational Solvent Mapping

Computational solvent mapping is a computational method inspired by an experimental technique called Multiple Solvent Crystal Structures (MSCS) (Allen et al., 1996; Mattos & Ringe, 1996). In this experimental technique, crystals of a protein of interest are soaked in various solutions of small probe compounds, after which X-ray structures of the soaked crystals are obtained. The multiple crystal structures are superimposed, and it has been shown that regions where multiple different molecular probes bind tend to be hotspot regions. By docking with multiple small molecular probes, we can perform a computational analog of this experimental technique, which has proven effective for predicting the binding hotspots on a variety of different types of macromolecules

(Dennis, Kortvelyesi, & Vajda, 2002; Kozakov, Grove, et al., 2015; Landon, Lancia, Yu, Thiel, & Vajda, 2007). One application of the work in Chapter 4 is improving the efficiency of this method while considering multiple sidechain conformations in a binding pocket.

1.3 Contributions

The work in Chapter 2 was done in collaboration with Artem Mamonov, who helped test parameters for optimizing the protocol for docking proteins using SAXS data. The work on efficient docking with flexible sidechains was based on previous work by David Hall and Laurie Grove, and the method for generating alternate conformers of each residue was developed by Dmitri Beglov.

CHAPTER TWO: SAXS Guided Protein Docking

2.1 Background

As previously discussed in the background, computational methods still have uncertainties in structure determination. Although docking programs, including ClusPro, generate several near-native structures for a large fraction of interacting proteins, current scoring functions are not reliable enough for selecting the best models. It was shown that using ClusPro it may be necessary to retain up to 30 of the lowest energy models to assure that the set includes a near-native structure. Thus, additional information can be very useful for correct structure determination. Many users of ClusPro are aware of this limitation, and combine computational docking with information from a variety of experimental techniques, including site-directed mutagenesis, cross-linking, and radiolytic protein foot-printing with mass spectrometry.

Small Angle X-ray Scattering (SAXS) is emerging as an effective approach to obtaining low-resolution structural information that can increase the reliability of docking results (Graewert & Svergun, 2013). The basic idea of the method is observing the X-ray scattering of a macromolecule in solution as a function of the scattering angle. The results of the experiment are encoded in a one-dimensional scattering profile determined from the spherical averaging of random orientations that a biomolecule can adopt in aqueous solution, and contains information about the shape and size of the macromolecule (Yang, 2014). Without the need for obtaining protein crystals or for labeling the protein, obtaining data using SAXS is relatively easy, and thus very appealing. SAXS experiments can be performed under a wide variety of solution conditions, including near physiological conditions, and usually take only a few seconds per sample exposure time on a well-equipped synchrotron beam line. However, the information content from scattering is much lower than the one that can be obtained by X-ray crystallography, which makes docking a natural complement to SAXS for the determination of complex structures.

Recently, several groups reported combinations of SAXS with protein docking approaches. Pons et al. ranked docked structures by weighted docking energy and SAXS fit score as the combined scoring function. In the method developed by Sali and co-workers (Schneidman-Duhovny, Hammel, & Sali, 2011) rigid body solutions were filtered by a coarse SAXS fit score, clustered, and ranked by a combined scoring function. Thus, both methods used combinations of docking and SAXS fit to facilitate model selection. Here we take a slightly different approach, and combine the docking method implemented in the ClusPro server with SAXS experimental data without modifying the scoring function. This is achieved by generating a very large number (at least 70,000) of docked structures by global sampling of the conformational space on a dense grid, and retaining a smaller but still large number (at least 2000) configurations that best agree with the observed SAXS profile. These structures are then ranked by the scoring function that was shown to perform well in ClusPro, clustered, and the centers of several of the largest clusters are considered as models of the complex, as ordinarily done in ClusPro. The main motivation for this approach is that it is based on a well-established docking method that for many proteins provides good accuracy docked models without the use of any additional information (Comeau et al., 2007; Kozakov et al., 2010, 2013).

We account for the SAXS data by focusing on the regions of the configurational space containing the structures that are most compatible with the scattering results, but otherwise perform the docking as usual. This approach has the advantage that we avoid overfitting to the SAXS data, and hence the docking results will not get worse even in cases where the SAXS experiment provides very limited additional information. In fact, the information content of SAXS profiles substantially depends on the shape of the complex considered, and it is generally higher for elongated complexes than for ones with more spherical shapes. The parameters of the method, primarily the number of structures that should be retained after SAXS filtering, will be selected by considering a training set of protein-protein interactions with simulated SAXS data, and the resulting algorithm will be applied to a validation set of proteins with experimental SAXS information available.

Currently results of SAXS experiments can be found only for a few protein-protein complexes. Although the application of the method is simple, the main problem is that unless the binding is very strong, an experimental SAXS profile for a complex may be a mixture of values for the complex and the unbound component proteins, thus complicating the analysis. However, due to recent developments in the methodology, particularly the ability of obtaining more homogeneous samples using size exclusion chromatography, we expect that the popularity of SAXS for determining protein complex structures will substantially increase. Therefore, we believe that expanding the already well-tested docking server ClusPro by enabling it to account for SAXS data will be useful. The use of the server is free for academic and governmental research.

2.2 Methods

The method presented here addresses the docking problem restrained by a SAXS profile. Thus, given two structures of molecules (referred to as a receptor and a ligand) and the SAXS profile of their complex, we use ClusPro to find the complex structure. We assume at most moderate conformational changes, primarily in the side chains and backbones that can accounted for by using a smooth scoring function and by performing local energy minimization. The docking protocol involves three steps as described below.

Step 1: Generating docked structures.

PIPER, the docking program implemented in ClusPro, is based on the fast Fourier transform correlation approach, and uses a pairwise interaction potential as part of its scoring function $E = E_{attr} + w_1 E_{rep} + w_2 E_{elec} + w_3 E_{pair}$ (Kozakov et al., 2006). Here E_attr and E_{rep} denote the attractive and repulsive contributions to the van der Waals interaction energy E_{vdw}, E_{elec} is an electrostatic energy term, and the pairwise term E_{pair} represents the desolvation contributions. The repulsive term is designed to not penalize small conformational clashes, thus resulting in a “smooth” scoring function. The coefficients w_1, w_2, and w_3 specify the weights of the corresponding terms, and are optimally selected for different types of docking problems (Kozakov et al., 2013). Unless specified otherwise, ClusPro simultaneously generates four types of models using the scoring schemes called (1) balanced, (2) electrostatic-favored, (3) hydrophobic-favored, and (4) van der Waals + electrostatics. The balanced option works generally well for enzyme-inhibitor complexes, whereas options (2) and (3) are suggested for complexes where the association is primarily driven by electrostatic and hydrophobic interactions, respectively. The fourth option, van der Waals + electrostatics, means that $w_3 = 0$, that is, the pairwise potential E_{pair} is not used. For each parameter set, ClusPro explores 70,000 rotations of the ligand on a translational grid with 1 Å spacing, and retains the best (i.e., lowest energy) translation for each rotation, thus resulting in 70,000 structures. In addition to the above modes, the “others mode” can be selected as an advanced option for the so-called “other” type of complexes that primarily occur in signal transduction pathways (Chen, Tong, Mintseris, Li, & Weng, 2003), and generally have substantially less perfect shape and electrostatic complementarity than the enzyme-inhibitor complexes. Due to the diverse nature implied by the “other” classification, this mode uses three different sets of weighting coefficients, generating 70,000 structures for each.

Step 2: Calculation of the SAXS profile and SAXS based filtering of docked structures.

We calculate the theoretical SAXS profile using the Debye formula

$$I(q) = \sum_i \sum_j f_i(q) f_j(q) \frac{\sin(qd_{ij})}{qd_{ij}}$$

where the scattering intensity I is a function of the momentum transfer $q = \frac{4\pi \sin(\theta)}{\lambda}$ at the scattering angle θ, and I computed by summing over all pairs of atoms (Debye, 1915). The quantities $f_i(q)$ and d_{ij} are the scattering factor of atom i and the distance between atoms i and j, respectively. The scattering form factor is a function of the atom, as well as the displaced solvent and hydration layer, $f(q) = f^v(q) - c_1 f^s(q) + c_2 s f^w(q)$, where

f^v is the form factor in vacuo, f^s is the form factor of a dummy atom of solvent, s is the fraction of solvent accessible surface area, and f^w is the form factor of water. The two constants c_1 and c_2 adjust the volume of the dummy atom and the difference in density between the hydration layer and bulk water, respectively. The default values of these parameters are $c_1 = 1.0$ and $c_2 = 0$, and since the deviations from these values are small, they are fixed at the default values to reduce computational efforts as proposed by Sali and co-workers (Schneidman-Duhovny, Hammel, Tainer, & Sali, 2013). This simplification can be used here because we utilize the approximate SAXS profile only to select the region of conformational space, and do not directly incorporate SAXS values into the scoring function. The SAXS profile $I(q)$ is calculated for each structure generated in Step 1, and the difference between the this and the experimental profile $I_{exp}(q)$ is measured in terms of the χ score, defined by

$$\chi = \frac{\sqrt{\frac{1}{M}\sum_i^M \left(I_{exp}(q_i) - I(q_i)\right)^2}}{\sigma(q_i)}$$

where M is the number of points, and σ(q) is the error of the experimental profile. As described in Step 1, unless the “others mode” is used, 70,000 structures are saved for each of the four parameter sets used by ClusPro. The structures in each result file are ranked based on the χ score, and the 2000 structures that have the best fit to the experimental SAXS profile are retained. When the “others mode” is used, the structures are ranked based on the χ score in each of the three result files, resulting in 6,000 structures.

Step 3. Rescoring and clustering.

Unless the “others mode” is used, we have four result files from Step 2 for the different parameter sets, each containing 2000 structures. In each file the structures are re-ranked based on the PIPER energy, and the 1000 lowest energy structures are clustered as described previously. The standard ClusPro output shows the centers and populations of the 10 largest clusters for each of the four different parameter sets. In contrast, using the “others mode” we re-rank the 2,000 structures in each of the three result files, and select the 500 lowest energy structures from each file. The retained 1500 structures are merged and clustered, and the centers and populations of the 10 largest clusters are shown. Model selection based on filtering by χ values, followed by the selection and clustering of several low energy structures has two advantages relative to methods that seek structures with the lowest values of scoring functions combining an energy score and a SAXS fit score. First, retaining many structures that give a good fit to the SAXS profile eliminates overweighting dependence on this type of measurements that may carry very limited information for roughly spherical protein complexes. Second, retaining the largest clusters of low energy structures rather than the ones with the lowest scores makes our results less sensitive both to the inherent errors in the SAXS data and to the conformational variation in the structures generated by docking.

Training Data Set. The method was trained using simulated profiles generated from crystal structures of 49 "others type" complexes in the protein docking benchmark (Chen et al., 2003). The “others type” complexes, including cell surface receptors and signal transduction proteins, were chosen since they generally are the most challenging for docking. Simulated SAXS profiles were generated using $c_1 = 1.0$ and $c_2 = 0$, for a range of the q parameter between 0.0 and 0.3, with a step size of 0.05 using the method for computing theoretical SAXS profiles as described in Step 2. As will be described, the main goal of training is the selection of the number of structures with good fit to the experimental SAXS profile that should be retained to optimally account for the information provided by the SAXS data.

Experimental SAXS Data. The impact of accounting for SAXS information was demonstrated by applying the method to experimental data for a lysozyme-inhibitor complex, where the Protein Data Bank (PDB) code for the X-ray crystal structure of the complex is 4G9S, and for the inhibitor structure it is 4DY3. SAXS data for three homodimers suitable for use as tests cases were taken from the Bioisis database (*http://bioisis.net*) and from the SASBDB database (*http://www.sasbdb.org/*) (**Error! Reference source not found.**). The two dimers from Bioisis are a superoxide dismutase (Bioisis ID: APSODP) and the protein PYR1 (Bioisis ID: 1PYR1P). The dimer from SASBDB is a myomesin dimer (SASBDB ID: SASDAK5).

Homology modeling. Models were built using Modeller v9.0 (Sali & Blundell, 1993), using the templates shown in Table 1. Lys side chains that were not present in the template were not modeled since they have uncertain localization. Aromatic residues

(Tyr, Phe, and Trp) that were not present in the template were placed in the most probable non-clashing rotamer positions.

3.3 Results

Results for the training set. Figure 1 shows the histogram of docking performance, as compared to the *ab initio* docking approach, for the 49 test complexes with simulated SAXS data in the training set. These results show that accounting for SAXS profiles almost doubles the number of systems (from 12 to 21) that have a near-native structure in the first (largest) cluster. The top 10 clusters include near-native structures for 39 of the 49 systems if we use the SAXS-based filtering, but only for 30 if no SAXS data are considered. We have studied the performance of the method depending on the number of structures retained in the SAXS filtering step (Figure 4). As shown, the best performance occurs if 2000 structures with the best fit to the SAXS profile are selected. The detailed results show that in almost all cases, both the rank and the RMSD of the near-native structure is improved. In a few cases, we do not find any predictions within 10 Å RMSD from the native pose. However, in these cases the ab-initio prediction is also relatively far from the native pose, thus these predictions would have been filtered out during the SAXS filtering step. Retaining fewer structures, and thus putting more emphasis on SAXS data, results in worse performance for several complexes. The reason is that we use cluster size for model discrimination. Clustering requires many near-native structures that are close to each other in terms of the pairwise interface root mean square deviation

(RMSD). However, not all such structures have low SAXS scores, and thus we should retain enough structures within a SAXS score range for reliable clustering. On the other extreme, retaining too many structures in the SAXS filtering would yield results that are like those obtained by docking without considering the SAXS data. However, the results remain similar within the range of 500 to 5000 structures retained, demonstrating the robustness of the protocol.

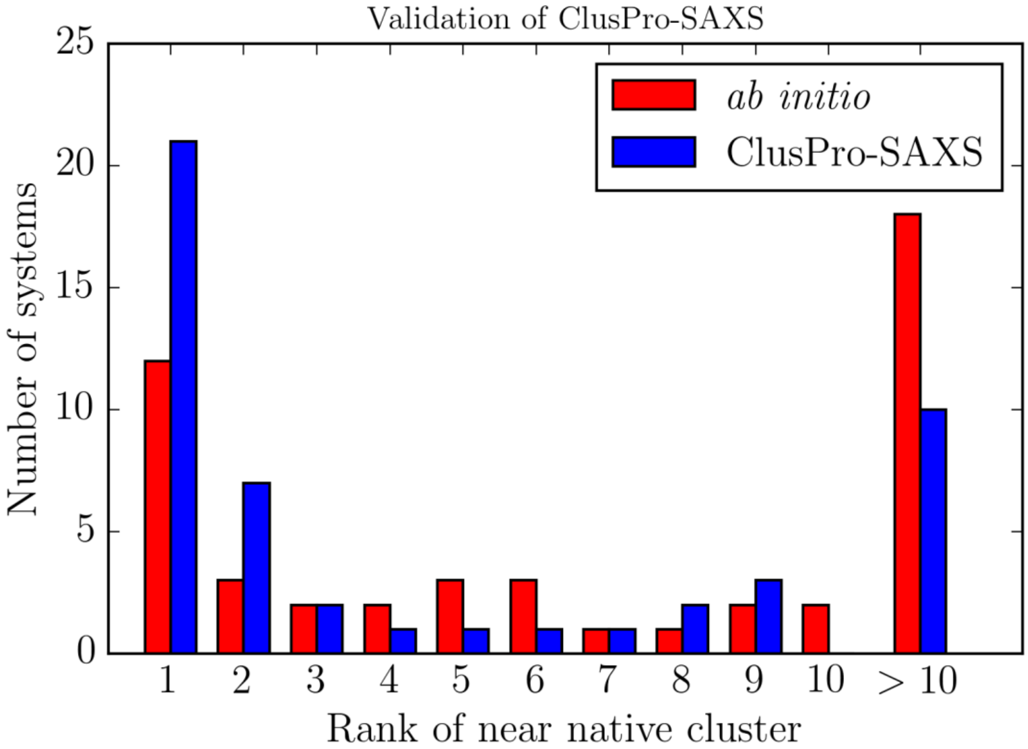

Figure 1. Validation using 49 complexes from the protein docking benchmark. Distribution of ranks of near-native models for ab initio docking shown in red, and SAXS docking in blue.

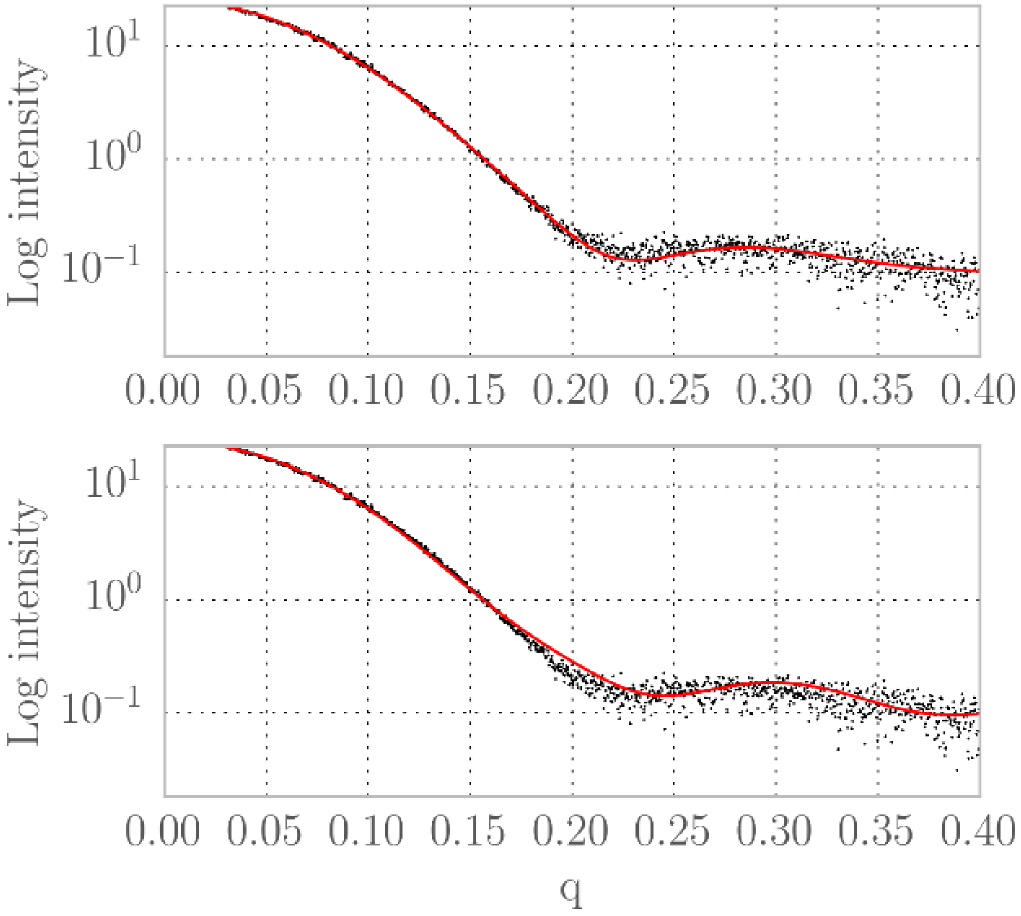

Figure 2. Top: SAXS profile of top ranked model, but which is far from native conformation, predicted by SAXS docking protocol (SAXS Fit score $\chi = 0.87$). Bottom: SAXS profile of near-native model (SAXS Fit score $\chi = 0.78$). This shows that the theoretical profile of an incorrect prediction can be fairly similar to the experimental profile. Thus, the SAXS fit score is not used to rank to the final outputs, but rather the cluster sizes are used.

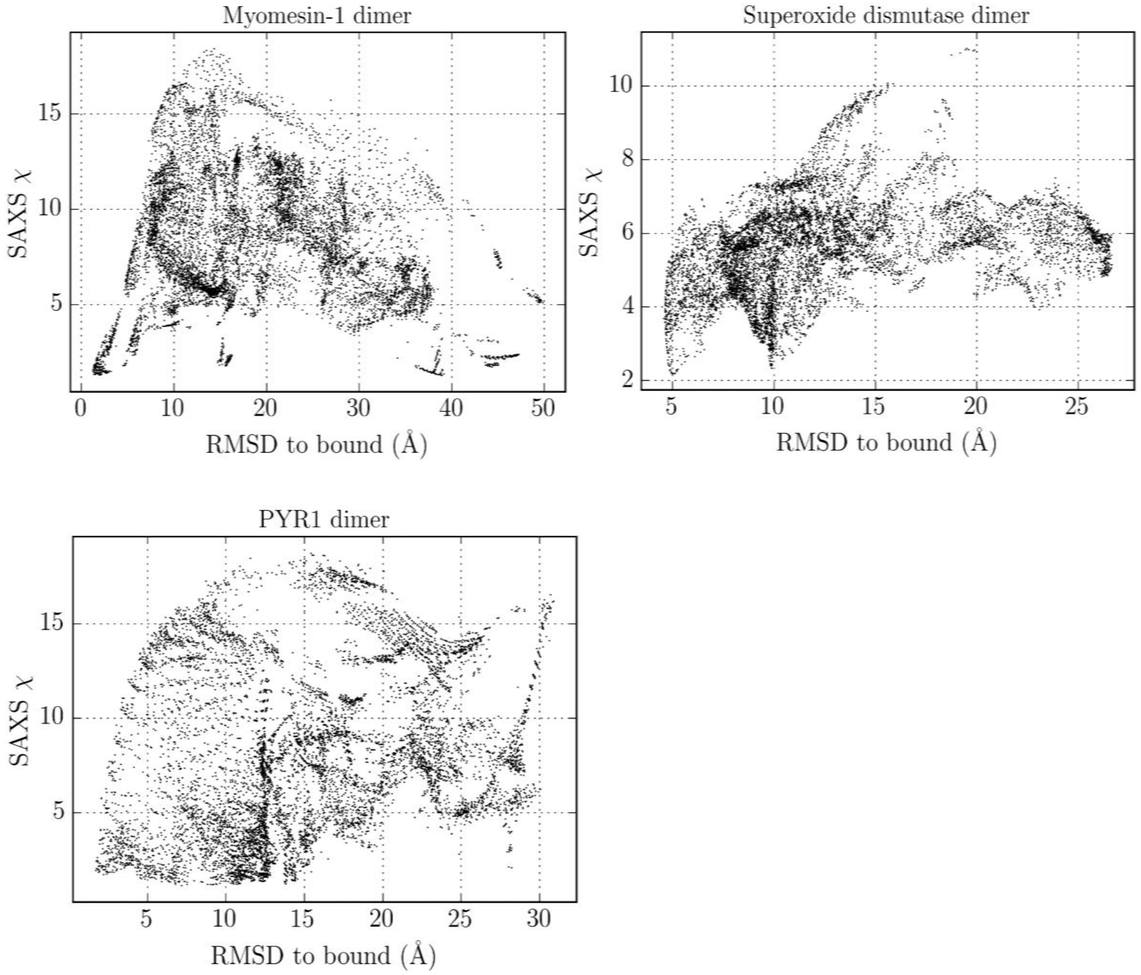

Figure 3. RMSD versus SAXS fit score for cases with experimental data. 2A: Myomesin-1 dimer (SASBDB ID: SASDAK5), 2B: superoxide dismutase dimer (Bioisis ID: APSODP), and 2C: PYR1 dimer (Bioisis ID: 1PYR1P). For each of the conformations predicted by PIPER, the SAXS χ score is plotted versus the RMSD to the native structure. While structures with low RMSD also tend to have low χ scores, there are also many structures with low χ score but high RMSD. This shows that the information content in SAXS is limited, as there are potentially many conformations with the same shape as the bound complex but with large RMSD to the native structure.

Results for Complexes with Experimental SAXS Data. Despite the potential of combining protein-protein docking with SAXS, experimental SAXS data on protein complexes remains scarce. However, as mentioned, recent methodology development such as size exclusion chromatography (SEC) SAXS, which allows for obtaining much more homogenous samples, should increase usage of SAXS for complex structure determination. Here we demonstrate the approach on one case of protein complex and 3 dimer test cases with experimental data (**Error! Reference source not found.**).

Table 1.

Experimental Case	Database ID	Template PDB ID	Sequence Identity	Original Rank	Final Rank
PliG-Lysozyme	N/A	1GBS	57.75%	6	3
Superoxide dismutase dimer	APSODP (Bioisis)	3F7K	62%	3	2
PYR1 dimer	1PYR1P (Bioisis)	3K3K	100%	3	3
Myomesin-1 dimer	SASDAK5 (SASBDB)	2RL5	99%	N/A	2

Table 1: The four validation cases using experimental data. The database ID can be used to find the SAXS data from the Bioisis or SASBDB databases. The template structures were used to build homology models of the ligand for the PliG-Lysozyme case, and of the monomer in the dimer cases. The ranks shown are the rank of the near native cluster as predicted by our method.

To get insight on how the approach works, we show SAXS fit score versus the RMSD values in Figure 2 for the systematic docking of *E. coli* PliG with the model of Atlantic salmon g-type lysozyme, where SAXS experimental data was available (Leysen, Vanderkelen, Weeks, Michiels, & Strelkov, 2013).

	CLUSPRO-SAXS		CLUSPRO ONLY	
PDB	Rank	RMSD	Rank	RMSD
1A2K	1	3.80	5	4.24
1AKJ	9	4.02	6	6.05
1ATN	2	6.66	N/A	N/A
1AZS	9	5.07	21	2.73
1B6C	1	3.57	1	4.01
1BUH	1	6.51	28	3.84
1E96	1	9.86	5	4.70
1EER	1	8.64	16	6.60
1F51	N/A	N/A	N/A	N/A
1FFW	2	9.17	9	8.80
1GLA	1	4.03	1	9.26
1GPW	1	1.88	1	3.28
1GRN	2	4.19	7	5.22
1H9D	N/A	N/A	N/A	N/A
1HE1	3	4.68	10	6.42
1I2M	6	7.01	44	4.22
1J2J	2	8.86	1	8.42
1JK9	1	9.41	2	9.75
1JWH	1	4.10	5	4.83
1JZD	15	6.08	31	4.32
1K5D	1	7.39	N/A	N/A
1K74	1	3.89	1	3.36
1KXP	1	4.76	1	3.77
1LFD	32	8.36	N/A	N/A
1ML0	1	6.89	1	4.97
1OFU	1	3.66	1	4.04
1R6Q	1	7.41	N/A	N/A
1RLB	19	4.70	8	6.45
1RV6	8	9.19	N/A	N/A
1SYX	2	4.78	1	6.58
1WQ1	8	9.16	10	8.03
1XD3	N/A	N/A	1	2.98
1XQS	1	5.56	6	7.53
1Z0K	15	2.51	26	3.40
1Z5Y	N/A	N/A	2	4.08
1ZHI	3	4.04	9	7.28
2A5T	9	9.43	N/A	N/A
2AYO	1	5.76	4	5.74
2BTF	1	5.29	14	8.22
2CFH	7	6.45	3	4.72
2G77	2	6.45	14	7.51
2HLE	N/A	N/A	3	4.58
2HRK	1	4.50	4	6.91
2I9B	2	7.45	11	6.05
2NZ8	4	7.70	1	9.78
2OT3	N/A	N/A	22	9.80
3BP8	5	6.52	6	8.73
3CPH	1	9.22	2	8.06
3D5S	1	3.48	1	3.94

Table 2. Ranks and RMSD of the near-native docked pose for all 49 training set cases. In the cases where there were no poses within 10.0 angstrom RMSD from the

bound pose the rank and RMSD are reported as N/A.

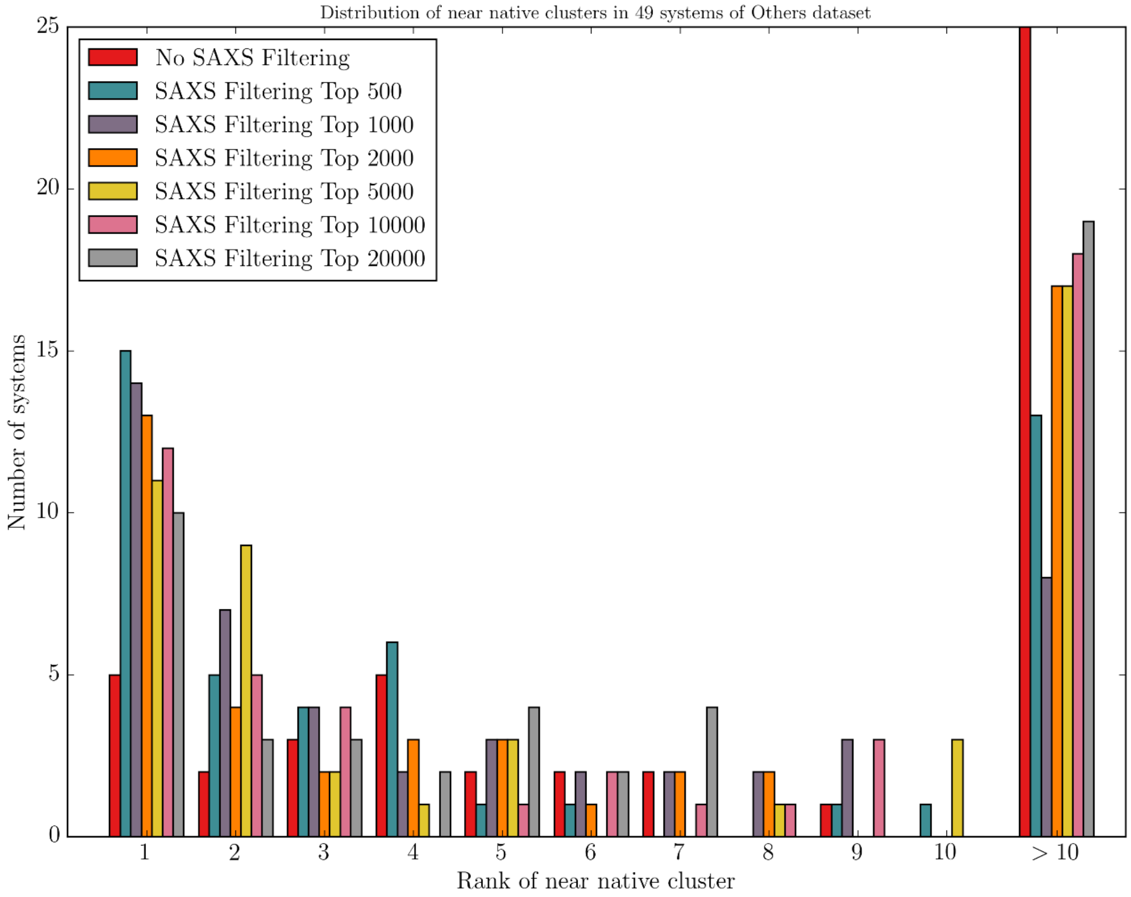

Figure 4: Distribution of ranks of near native models for docking with SAXS using different cutoff points for the SAXS filtering step. The best performance is found when we retain the top 2000 conformations by chi score.

2.4 Discussion

Due to spherical averaging, the SAXS data frequently provide limited information for protein docking. In fact, two conformations can have equally low SAXS fit scores but very different RMSDs from the native structure. Plots for the other experimental cases are shown on the Figure S2. Like the lysozyme case, discrimination of the near-native conformations by SAXS chi-score is limited for the globular system PYR1. However, when the geometry of the complex is more elongated (myomesin-1 and superoxide dismutase), the SAXS chi-score becomes more discriminative and we can see sharper funnels in a neighborhood of the native structure (with 10Å RMSD for the myomesin-1 dimer and 7Å RMSD for the superoxide dismutase dimer). In Figure 2 we show the SAXS profile of an incorrect model with a relatively low SAXS fit score, compared to near-native model to demonstrate that they both satisfy the SAXS constraints. Nevertheless, if we dock the PliG protein to lysozyme without the SAXS filtering step, the near native model is ranked 6th, whereas it is ranked 3rd if the SAXS data are considered. Improvement was also observed for two of the three dimers in Table 1. Although the improvement may be moderate, the docking did not yield any near-native structure for Myomesin-1 dimer without the SAXS constraints, and thus accounting for the additional information was crucial.

CHAPTER THREE: Spatial Restraint Guided Protein Docking

3.1 Background

Despite the significant progress, docking methods generally cannot be fully trusted when used without any experimental validation. The main reason is that the current scoring functions are not accurate enough for finding the best models among the ones generated by the sampling. Thus, additional information can be very useful for improving the reliability of structure determination. Accordingly, in the scoring function used by ClusPro we have the option to apply extra attraction terms to residues that are a priori known to be the inter-face. Conversely, repulsion terms are applied to residues that are not expected to be in the interface. However, what ClusPro was lacking so far was the ability to define distance restraints between pairs of atoms or residues. Such restraints can be derived, e.g., from NMR Nuclear Overhauser effect (NOE) experiments, by FRET, or by chemical cross-linking, and are very useful as they provide information on the relative orientation of the two proteins. In fact, the use of restraints is central to the popular HADDOCK server. HADDOCK incorporates the interaction restraints into the scoring function to guide the search toward regions of the conformational space in which the restraints are satisfied. HADDOCK applications generally involve interaction restraints based on 10 to 25 residues on the two sides of the interface.

While the extra terms in the scoring function due to the restraints do not significantly increase the computational burden if the sampling is based on Monte Carlo or molecular dynamics algorithms, a similar approach is very costly when used with FFT based sampling. The problem is that each pairwise restraint in the scoring function requires a new correlation function term, and thus an additional Fourier transform. Since the expression used for scoring generally includes only four or five correlation functions, representing the various energy contributions, adding just five distance constraints would double the computational burden. Thus, it is not surprising that none of the successful FFT based docking programs has the option of accounting for pairwise restraints. However, since FFT performs global sampling, there is no need for guiding the search toward feasible regions. Based on this observation we solve the problem by directly selecting low energy solutions that also satisfy the restraints. As will be shown, this implies that frequently only portions of configurational space need to be examined, and hence in some cases the computational efforts are reduced. A further advantage is that the scoring function is not affected, and thus we retain the favorable properties of the ClusPro server, validated in many rounds of the CAPRI docking experiment. In this note we consider pairs of proteins from the protein docking benchmark, and show that accounting for a varying number of simulated distance restraints significantly improves the results. Additional validation is presented for two systems with experimentally determined distance restraints.

3.2 Methods

A pairwise distance restraint can be defined by two sets of atoms, S_1 and S_2 and a distance range, d_{min} to d_{max}. The restraint is considered satisfied if there is at least one atom in S_1 and at least one atom in S_2 such that the distance between them falls in this range. While the implementation allows for arbitrary sets of atoms to be used to define a restraint, most frequently these involve a single atom or residue on each side of the interface. Given multiple restraints, users may wish to require a certain number of restraints out of a group to be satisfied. In addition, restraints may be based on sources with varying reliability, requiring different cutoff values. Our implementation allows for grouping restraints into restraint groups, and restraint groups into restraint sets. Restraint groups are considered satisfied when more than a user specified number of restraints in the group are satisfied, and a restraint set is satisfied when more than a user specified number of its groups are satisfied. This hierarchical definition is flexible enough to provide options that are like the ones used by HADDOCK. We have developed a JSON based file format for specifying groups of restraints used by our restraint library, as well a script for converting data in the NOE format into our JSON format. A full description of the file format is provided in Appendix 1.

Docking is performed using PIPER, which samples all translations and rotations of a ligand protein with respect to a receptor protein. When a restraint set is provided, PIPER will only report solutions that satisfy the restraints. To do this efficiently, we first generate the set of translations that satisfy each individual restraint, called the feasible translation set for that restraint. We then consider the intersection of feasible translation sets for the restraints in each restraint group, and select the translation that appears more often than the cutoff for the restraint group. The selected feasible translation sets for each restraint group are merged in a similar way to generate the feasible translation set for an entire restraint set.

We note that providing restraints can decrease the running times by using the restraint set to generate a feasible translation set for each rotation. For each feasible translation, the van der Waals interaction energy is computed and is used to filter out translations that result in unacceptable clashes. If there are no feasible translations leading to an acceptable van der Waals energy, the rotation is skipped and no other energy terms are evaluated. In practice, this often results in skipping many rotations. When the cost of generating the feasible translations is less than the cost of evaluating the additional energy terms, fast rotation skipping results in an overall speedup. After selecting the solutions that satisfy the restraints, 1000 structures with the lowest PIPER energies are clustered and minimized as customary in ClusPro.

Data Preparation

E2A-Hpr complex. For this test case we considered the restraints based on NMR experimental data (Garrett, et al., 1997). An AIR restraints file was generated for the application of HADDOCK to this problem (Dominguez, et al., 2003). The AIR restraints file was converted to a JSON file using a Python script, which added 1.5 Å to the top end of the distance range, for a range of 0 to 4.5 Å for each restraint. In addition, since the domains to be docked are defined as chain A in both E2A and Hpr, we set the ligand chain in the restraint file to A for every restraint.

Nucleosome complex. The docking of the UbcH5c subunit of the PRC1 complex to histone H2A of nucleosome was target 95 of the CAPRI docking experiment. Restraints were generated after examining the evidence available from the literature. Based on Bently et al. (2011) we knew that there is an interaction between Lys119 of histone H2A and Cys85 of the UbcH5c subunit, and hence we created one restraint that had to be satisfied between these two residues. The required range, 0 to 8 Å, was fairly large, because these residues were located in flexible tail regions of the proteins. To assure that Lys97 and Arg98 interact with the histone in the nucleosome, we created a second restraint group with multiple restraints, from Lys97 to the set of surface residues on the histone. In this second group, we only require one of the restraints to be satisfied, since it is not known which of the residues on the surface of the histone interact with Lys97.

While these restraints were generated manually within an interactive Python session, we have created an interactive web application that can aid in the creation of similar JSON restraint files for other users. This tool can be found at https://cluspro.bu.edu/generate_restraints.html. Using the web form there, users can easily create complex restraint sets.

For example, the restraints used for the nucleosome test case could have been created using the web application as follows:

- Set “Required percent of groups” to 100. We want both the specific restraint and the Lys97 to surface restraint group to be satisfied.
- Create the Lys119 to Cys85 restraint. Set “Required percentage of restraints” to 100 for the first restraint group, and add a restraint from “G 118” to “A 85” (these are the residue identifiers from the PDB files).
- Create the Lys97 to surface restraints. Click “Add Group” to add a new restraint group, then add 46 restraints from ligand residue “C 97” to the following list of receptor residues: "E 73", "E 76", "E 77", "E 80", "E 134", "F 25", "F 27", "F 52", "F 56", "F 59", "F 67", "F 74", "G 14", "G 19", "G 22", "G 61", "G 64", "G 65",

"G 68", "G 71", "H 44", "H 47", "H 48", "H 89", "H 96", "H 102", "H 105", "H 106", "H 109", "H 113", "H 116", "H 117", "H 120", "I 5", "I 6", "I 15", "I 46", "I 47", "I 56", "I 57", "J -52", "J -51", "J -42", "J -41", "J -11", and "J -10". Set the required percentage to 2, which should result in only 1 restraint being required in this group.

Click “Create Restraints”. A JSON formatted restraint set should appear below the form. The user can then copy and paste this into a text file, or click the “Save As…” button to save the file to disk.

Benchmark for docking with simulated restraints

Receptor and ligand PDB files were acquired from the ZLAB Benchmark 4, a curated set of protein-protein complexes with known structure (Hwang, Vreven, Janin, & Weng, 2010). Using the structures super-imposed into the bound pose, we ordered pairs of residues across the interface by their C-alpha distance, and chose the top 20 residue pairs with minimum distance to use in restraint set.

3.3 Results

We tested the impact of restraints on the ClusPro results for 101 rigid enzyme-inhibitor and “other” type complexes from version 4 of the protein docking benchmark. For 55 out of these 101 cases, ClusPro without restraints did not produce a near native structure in the top 5 predictions. Near-native structures were defined as the ones with less than 10.0 Å interface root mean square deviation (IRMSD) between X-ray and predicted ligand positions after superimposing the receptor structures. For each of these 55 cases a set of restraints was generated by selecting the 20 closest residue pairs across the known interface for creating a restraint group. For each restraint, d_{max} was set to the actual C_α to C_α distance plus 2.0 Å, and d_{min} to half the distance. To test the effectiveness of adding restraints, we ran docking calculations using different requirements for the number of restraints to be satisfied. With the addition of restraint sets, even at a relatively non-stringent requirement of 50% of the restraints satisfied, we start to see near native poses ranked within the top 5 clusters (Figure 5). For the chosen test set, using 20 restraints across the interface and requiring all restraints to be satisfied was sufficient to produce near- native conformations within the top 5 predictions for all cases, and as the top prediction for 53 of the 55 complexes.

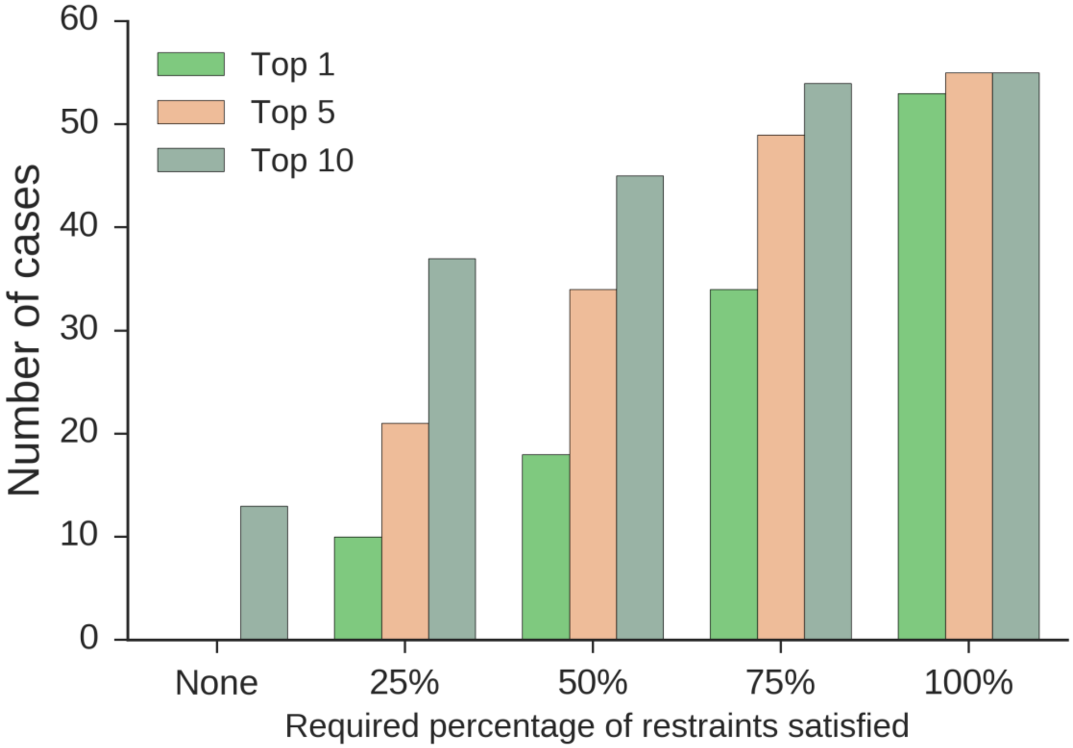

Figure 5. Ranking of clusters that include the first near-native structure when satisfying varying fraction of the restraints in docking the test set with simulated restraints.

E2A-Hpr complex. Our first test case with restraints based on experimental data is the E2A-Hpr complex, also studied using HADDOCK. The Ambiguous Interaction Restraints (AIRs) were converted into a restraint set (see Data Preparation Section above). We compared docking without any restraint to docking using the restraint set (Figure 6). ClusPro works very well for this complex even without restraints, as the second ranked cluster includes a near-native structure. Accounting for the restraints further improves the result, and the first near-native solution is contained in the top ranked cluster.

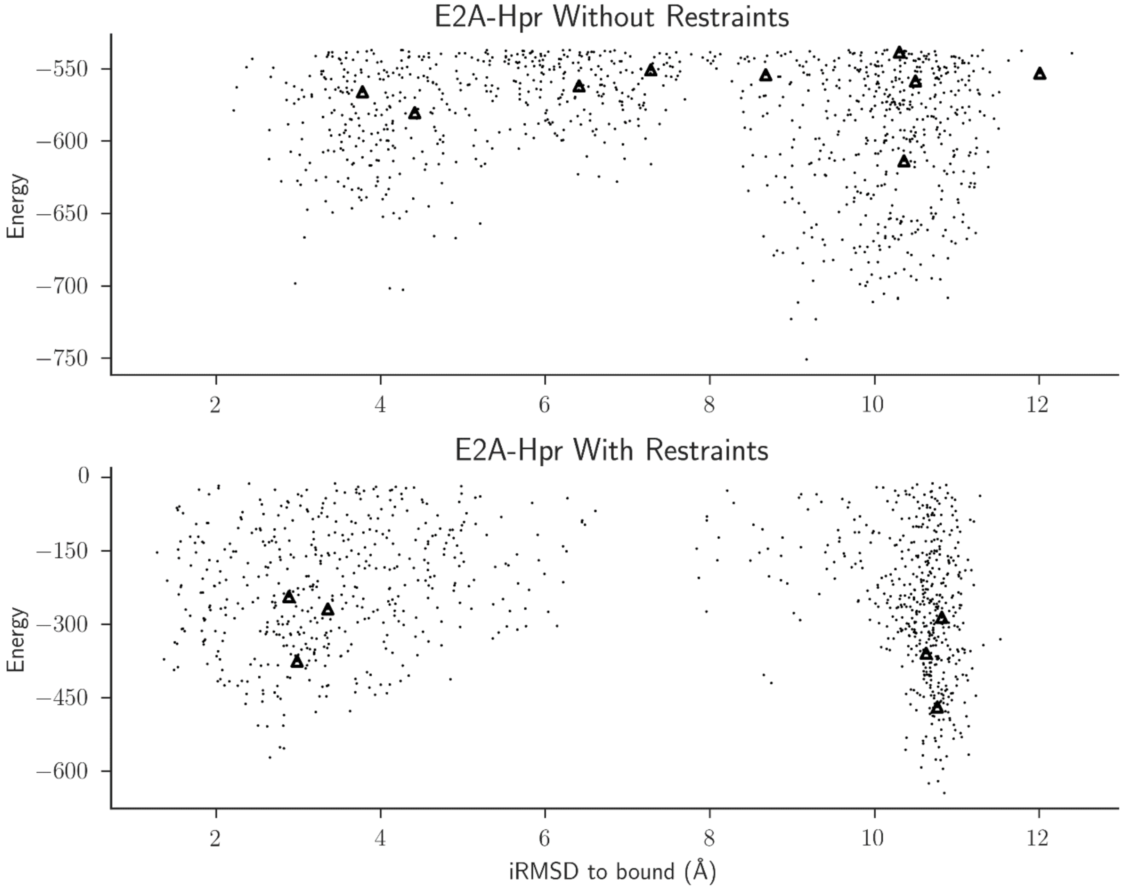

Figure 6. iRMSD vs. energy plots for pre-clustering results for the E2A-Hpr test case. iRMSD is calculated using alpha carbon in the interface of the complex. Each docking result is shown as a dot, and the cluster centers are shown as open triangles. Docking with restraints shifts the distribution to the left, but also increases the range of the energies observed for the lowest 1000 results. The cluster center of the near-native funnel is shifted from 3.78 Å iRMSD down to 2.88 Å iRMSD.

Nucleosome complex. The Polycomb repressive complex (PRC1) binds and ubiquitinates the nucleosome in histone H2A on Lys119. Cys85 in the UbcH5c subunit of the PRC1 complex was known to interact with Lys119 of H2A during the ubiquitination reaction. In addition, Lys97 and Arg98 of the PRC1 complex were shown to be required for activity, although it was not known where on the nucleosome these residues interacted. Based on this experimental data, we constructed a restraint, which required Cys85 to be within 5 Å of Lys119, and supplemented it with restraints that required Lys97 to be close to the surface of the nucleosome complex (see Data Preparation Section above). Docking using this restraint set produced the native pose ranked 2 with a C_α RMSD of 6.8 Å (Figure 7).

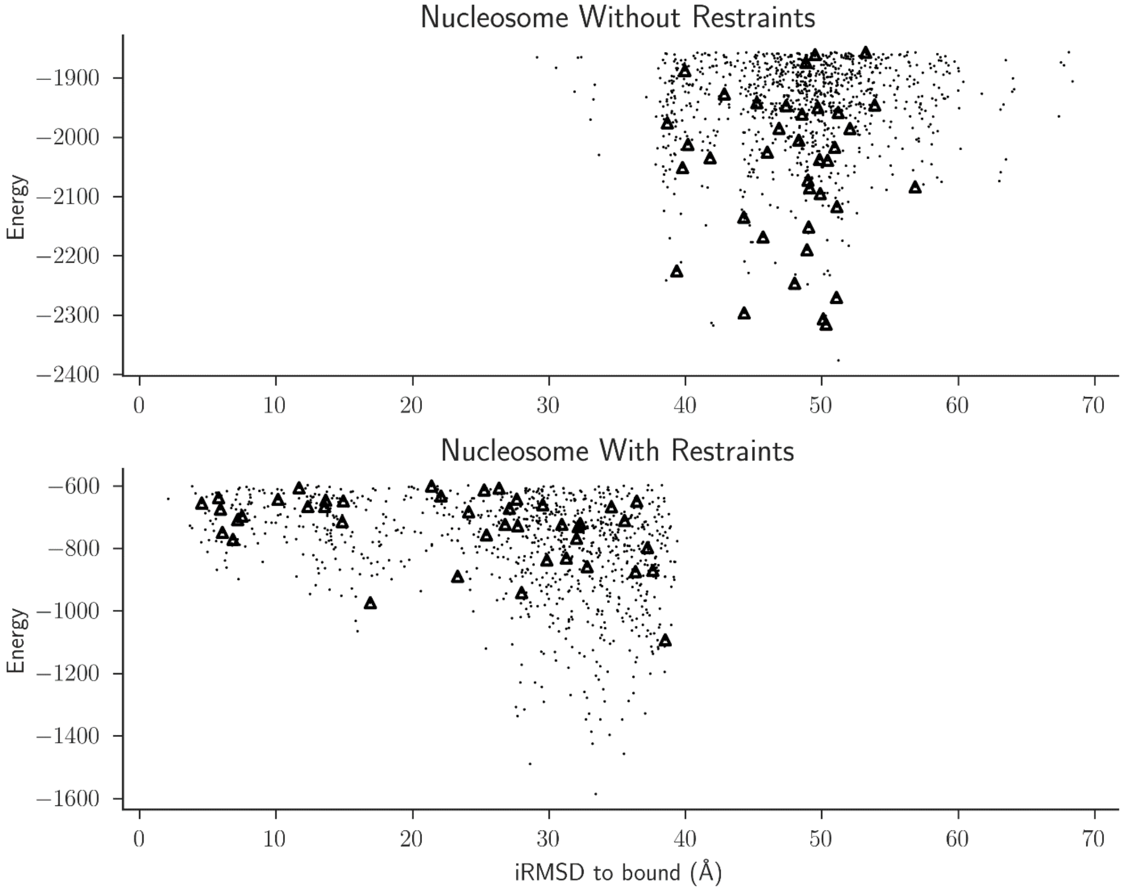

Figure 7. iRMSD vs energy plots for pre-clustering results for the nucleosome test case. iRMSD is calculated using alpha carbon in the interface of the complex. Each docking

result is shown as a dot, and the cluster centers are shown as open triangles. In this test case docking with restraints drastically shifts the distribution to the left, with the iRMSD of the best cluster center moving from 38.68 Å to 4.53 Å.

	NO RESTRAINTS		**25% SATISFIED**		**50% SATISFIED**		**75% SATISFIED**		**100% SATISFIED**	
ID	Rank	RMSD (Å)	Rank	RMSD (Å)	Rank	RMSD (Å)	Rank	RMSD (Å)	Rank	RMSD (Å)
1A2K	5	4.13								
1AK4			21	9.83	11	6.3	1	4.34	1	4.65
1AKJ	6	6.06	5	3.73	4	4.19	3	5.24	1	3.23
1AZS	21	2.56	5	2.44	5	2.44	1	2.89	1	1.46
1B6C	1	2.96								
1BUH	27	9.36	1	5.51	1	5.19	1	4.29	1	1.66
1E96	7	4.98	1	5.08	1	6.3	1	3.99	1	2.36
1EFN					12	6.54	3	5.97	1	2.49
1F51			7	2.35	6	2.5	1	5.35	1	2.99
1FC2			4	8.88	1	9.18	2	9	1	6.26
1FCC			28	9.2	4	6.05	1	4.7	1	3.36
1FFW	1	9.68								
1FQJ			15	4.74	3	5.49	1	2.96	1	2.54
1GCQ					11	7.43	1	3.86	1	3.14
1GHQ			19	8.6	15	7.1	12	7.42	1	1.61
1GLA	1	8.9								
1GPW	1	5.48								
1H9D	19	9.92	14	9.37	6	5.57	3	5.02	1	3.54
1HCF			9	8.33	12	4.9	3	3.14	1	1.15
1HE1	10	6.24	5	3.57	1	4.43	1	3.75	1	2.32
1I4D			19	9.9	6	7.98	3	6.39	1	4.7
1J2J	1	8.17								
1JWH	5	4.02								
1K74	1	3.26								
1KAC			8	4.79	1	2.15	1	1.86	1	1.17
1KLU					15	9.95	5	3.77	1	3.28
1KTZ			8	5.49	4	5.87	1	5.56	1	1.73
1KXP	1	2.97								
1ML0	1	4.49								
1OFU	1	3.98								
1PVH			7	9.2	7	8.74	7	8.07	1	2.62
1QA9			8	4.22	2	5.02	1	1.96	1	4.55
1RLB	7	6.14	17	3.87	6	6.28	6	4.2	1	2.75
1RV6			7	6.52	4	4.71	1	3.35	1	1.07
1S1Q			7	5.07	5	5.89	1	2.75	1	2.97
1SBB					15	2.17	5	5.45	1	2.66
1T6B			6	5.3	1	4.22	1	2.27	1	2.96

1US7			25	6.69	6	6.56	10	5.01	2	5.45
1WDW	1	6.23								
1XD3	1	2.96								
1XU1			6	3.04	1	2.34	1	2.23	1	1.63
1Z0K	2	9.26								
1Z5Y	2	3.77								
1ZHH			7	6.85	5	6.22	1	4.74	1	2.46
1ZHI	9	7.34	1	2.81	3	5.19	2	4.66	1	1.41
2A5T			22	5.96	13	6.29	3	6.04	1	3.54
2A9K			1	4.53	8	4.08	1	3.47	1	2.83
2AJF			25	5.16	14	6.47	2	2.05	1	4.18
2AYO	4	5.45								
2B4J					1	8.05	6	7.3	1	4.26
2BTF	18	8.15	1	4.64	1	2.54	1	2.54	1	4.16
2FJU			24	5.96	9	3.48	1	5.39	1	3.82
2G77	13	6.89	2	3.59	1	3.7	1	4.4	1	3.18
2HLE	3	8.36								
2HQS			2	4.79	1	3.22	1	2.66	1	3.85
2OOB			5	7.28	6	3.49	2	5.94	2	5.46
2OOR			3	5.3	2	5.3	1	3	1	3.71
2VDB			3	7.02	1	5.01	1	4.89	1	1.44
3BP8	6	8.46	1	6.92	3	6.85	1	5.43	1	4.89
3D5S	1	3.3								
1AVX	1	3.3								
1AY7	3	6.14								
1BVN	1	6.01								
1CGI	2	9.13								
1CLV	1	5.05								
1D6R	18	7.62	8	6.06	8	5.07	4	2.66	1	2.77
1DFJ	2	3.79								
1E6E	1	3.42								
1EAW	2	5.71								
1EWY	4	7.26								
1EZU	8	3.2	5	3.18	1	3.63	1	3.75	1	3.34
1F34	13	5.99	2	6.01	2	5.11	1	5.01	1	2.22
1FLE	2	5.04								
1GL1	1	9.22								
1GXD			1	8.13	1	7.23	2	5.62	1	3.73
1HIA	7	8.01	8	7.99	3	5.22	1	5.31	1	3.07
1JTG	1	3.75								

1MAH	2	6.68								
1N8O	1	6.11								
1OC0	3	7.99								
1OPH	8	8.67	6	7.72	3	7.13	7	4.36	1	3.88
BOYV			22	4.06	7	7.08	2	4.03	1	1.79
1OYV	1	3.3								
1PPE	1	2.33								
1R0R	4	1.65								
1TMQ	7	2.87	1	3.61	1	2.95	1	3.49	1	3.27
1UDI	1	2.54								
1YVB	3	3.81								
2ABZ	19	5.51	6	5.52	1	5.67	1	2.69	1	2.74
2B42	2	4.97								
2J0T	1	9.6								
2MTA	3	6.05								
2O8V	13	5.55	6	6.46	4	5.16	1	5.38	1	5.48
2OUL	1	2.93								
2PCC	6	5.76	12	6.01	16	9.61	1	5.07	1	2.79
2SIC	1	4.26								
2SNI	1	3.81								
2UUY	7	7.14	1	6.5	1	5.15	1	6.17	1	3.61
3SGQ	4	9.96								
4CPA	6	2.95	1	4.35	1	3.43	1	2.64	1	1.9
7CEI	3	4.89								

Table 3. Rank and RMSD values for all benchmark cases. For cases where ClusPro without restraints already produced a near native prediction with rank less than or equal to 5, we did not test the docking with restraints. For the other 55 cases where ClusPro did not produce a highly ranked near native prediction, we tested our restraints method with different restraint sets. Using the most stringent restraints produces the best ranking results overall. The ID is the taken from the protein benchmark in a few cases different chains from the same PDB entity were used for docking, so these have a different ID than found in the PDB (for example, BOYV which corresponds to 1OYV). Using the most stringent synthetic restraints results in the near native pose in either rank 1 or 2. While the restraints

in all these cases were the same, we can clearly see that requiring fewer of the restraints to be satisfied does not sufficiently constrain the ligand to the native pose.

3.4 Discussion

We describe implementation of pairwise restraints to the FFT sampling approach. Unlike other approaches which bias the energy function to steer the docking results towards satisfying the restraints, we leave the energy function intact and restrain the search space. Our implementation allows the user to vary the confidence in the restraints by varying the number of restraints to be satisfied, as well as specifying restraints in multiple groups to account for multiple possible interfaces. Accounting for restraints we demonstrate that this approach improves results even with spurious restraints. This was shown by using simulated restraints on a well-known docking benchmark, as well as in applications with restraints based on experimental data. The method is freely available as part of ClusPro protein docking server.

CHAPTER FOUR: Efficient Global Sampling of Flexible Sidechains

4.1 Background

While FFT based methods have proven very effective for protein docking, they are still limited by the rigid body nature of the global sampling. While this limitation does not prevent the method from working for many systems, we can still obtain better results when taking flexibility into account. This is seen in the way we implement the shape complementarity term in the scoring function, where we “soften” the surface layer of the proteins so that clashes at the surface are not as heavily penalized. This results in the surfaces of the initial predicted structures overlapping slightly, which is afterwards corrected by minimizing using a more precise energy function.

Previous work in the lab showed that considering multiple conformations of the sidechains of key residues in the binding pocket can improve the quality of mapping results (Grove, Hall, Beglov, Vajda, & Kozakov, 2013). However, the method was implemented by repeating the global rigid body docking stage for all conformations of the key sidechains, which proved to be computationally expensive. In addition to this use case, there are other types of systems where the ability to efficiently sample multiple sidechain conformations while still performing global systematic search would be useful. For example, there exist protein-protein interactions where an anchor residue must exist in a specific conformation for binding (Rajamani, Thiel, Vajda, & Camacho, 2004). For these types of interactions, a systematic search of the conformations of the anchor residue may lead to improved docking accuracy.

4.2 Methods

Previous work has shown how to efficiently calculate energy-like scoring functions using the FFT if the functions are in the form of a correlation functions. This insight significantly improved the efficiency of exhaustive global sampling from $O(N^6)$ to $O(N^3 \ln N)$. Using this technique, it becomes possible to perform an exhaustive global search for energy minima in the docking of two macromolecules. This method has been effectively used in computational solvent mapping (Kozakov, Grove, et al., 2015). By performing multiple mappings with different sidechain conformations, we can simulate sidechain flexibility. However, there is significant redundant work being done using this approach. When altering sidechain conformations, most of the atoms in a protein stay fixed. We can use this insight to greatly improve the efficiency of sidechain sampling when using the FFT method.

We term the fixed portion of each macromolecule being docked the template, and the moving sidechains the key sidechains. Thus, we have the receptor template, key receptor sidechains, ligand template, and key ligand sidechains. We can then decompose the scoring function used in rigid body global docking:

$$E(\alpha,\beta,\gamma) = \sum_{i,j,k} R(i,j,k)L(i+\alpha,j+\beta,k+\gamma)$$

The above equation can be rewritten after breaking up $R(i,j,k) = R_T(i,j,k) + \sum_u R_u(i,j,k)$ and $L(i,j,k) = L_T(i,j,k) + \sum_v L_v(i,j,k)$. R_T and L_T are contributions to the scoring function from just the template portions of the receptor and ligand respectively, and R_u and L_v are the contributions of the movable sidechains of the receptor and ligand. We can then rewrite the first equation as below:

$$\begin{aligned} E(\alpha,\beta,\gamma) &= \sum_{i,j,k} R(i,j,k)L(i+\alpha,j+\beta,k+\gamma) \\ &= \sum_{i,j,k}\left(R_T(i,j,k)+\sum_u R_u(i,j,k)\right)\left(L_T(i+\alpha,j+\beta,k+\gamma)+\sum_v L_v(i+\alpha,j+\beta,k+\gamma)\right) \\ &= E_{R_tL_t}(\alpha,\beta,\gamma)+\sum_i E_{R_iL_t}(\alpha,\beta,\gamma)+\sum_j E_{R_tL_j}(\alpha,\beta,\gamma)+\sum_i\sum_j E_{R_iL_j}(\alpha,\beta,\gamma) \end{aligned}$$

In the case of mapping small molecular probes, there are no movable ligand sidechains so the second and third summation terms in Equation 2 drop out, and we are left with just the following:

$$E(\alpha,\beta,\gamma) = E_{R_tL_t}(\alpha,\beta,\gamma)+\sum_i E_{R_iL_t}(\alpha,\beta,\gamma)$$

This final expression contains one term for correlation between the non-moving template portions of the receptor and ligand ($E_{R_tL_t}(\alpha,\beta,\gamma)$), and one term for each of the movable sidechains ($E_{R_iL_t}(\alpha,\beta,\gamma)$). This formulation allows us to greatly reduce the amount of redundant calculations made, and is illustrated by the conceptual Figure 8. Without the decomposition described, to evaluate the energy of M variants of a protein we would need to compute M correlation functions of a N^3 sized grid. Using the decomposition, we need only compute one N^3 grid and M n^3 grids, where n will typically be much smaller than N as each subgrid only needs to be big enough to cover one sidechain of a protein. The global rigid body docking program, PIPER, was modified to incorporate this faster grid calculation technique when run with a template structure and multiple rotamers of movable sidechains.

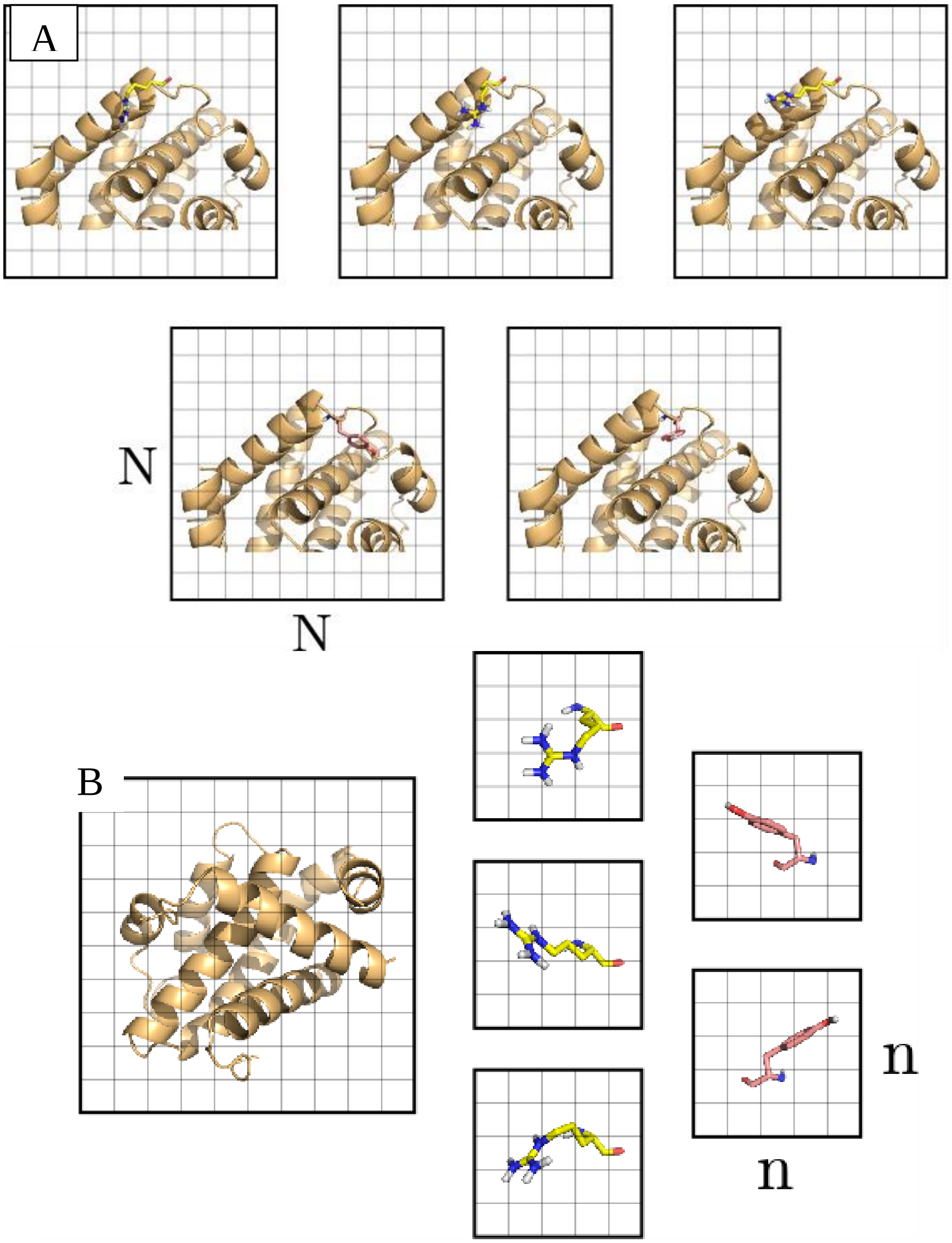

Figure 8. Conceptual figure of the decomposition of FFT grids. As depicted in (A), without decomposition of the grid we are forced to recompute the scoring function on the full grid of size $N \times N$. When we decompose the grid as in (B), we can compute the scoring function for the template on the full $N \times N$ grid, and compute the scoring function for each sidechain conformation on a smaller $n \times n$ grid.

Sidechain Selection

Selection of movable chains is performed using the method described in (Grove et al., 2013). First, the unliganded structure of each case is initially mapped using FTMap to determine the binding pocket. In the general case, a user would select one or more consensus clusters in a region of interest. These selected clusters define a pocket of interest, which are all residues within 6 angstroms of the selected clusters. From this pocket, residues of the amino acids Lys, Arg, Tyr, Phe, Trp, His, Met, Gln, Asn, and Asp which satisfies cavity and hydrophobicity cutoffs are selected for rotamer generation. Rotamer generation is performed using an end group library minimization method (EGLM), which starts with a library of pre-generated sidechain conformations and uses minimization to generate an ensemble of sidechain conformations, or rotamers (Beglov et al., 2012). Clusters with both high population and low energy are selected as the rotamer set for each residue.

Final Rotamer Selection

In the original FTFlex protocol, a separate mapping was done for each rotamer of each of the movable sidechains. That is, for each movable sidechain, each of the rotamers for that sidechain is placed into the original unbound structure to obtain a structure that is different in structure for only that residue. Each of these modified structures is submitted to FTMap. The final conformation for each of the movable sidechains is selected by counting the number of consensus clusters within 6 angstroms of the selected pocket, and choosing the rotamer that has the highest count. This could be the original unbound rotamer, in which case no change is made for that residue. The final structure is obtained by using the best rotamer for each of the movable sidechains.

Using the fast PIPER program, we can generate the same results for all rotamers at once. The initial mapping and selection of movable sidechains is performed the same way as in FTFlex. Instead of running an additional FTMap step for each rotamer, we use the enhanced PIPER program to generate rigid body docking results for all rotamers of all the movable sidechains in one step. Using these docking results, we count the number of probe atoms within 6 angstroms of the pocket residues. The rotamer with the highest number of probe atoms contacts is chosen, and the final structure is obtained by using the chosen rotamer for all movable sidechains. In both versions of the FTFlex protocol, the final structure is mapped one last time using FTMap to get the final consensus clusters.

Calculation of Profile Correlation

To quantitatively measure the quality of a mapping, we turn to mapping fingerprints, a metric previously developed to assess the similarity of mapping results (Bohnuud, Kozakov, & Vajda, 2014). We first count the number of non-bonded contacts each residue in a mapped structure makes with the consensus clusters. This number is normalized by the total number of contacts made to get a vector of the fraction of non-bonded contacts for each residue in the protein being mapped. For a protein with n residues, this vector of length n is termed the mapping fingerprint. From a fingerprint, we select the terms for each of the residues in the pocket of interest to obtain a pocket fingerprint, which is illustrated in Figure 9. Finally, we compare how similar two pocket fingerprints are by computing the Pearson correlation between them, which is a quantitative measure of mapping similarity. The correlation measured between mapping results for the bound structure and a variant is defined as the bound-state similarity coefficient (BSSC). The BSSC ranges from 0 to 1, with values closer to 1 representing results more similar to the bound mapping.

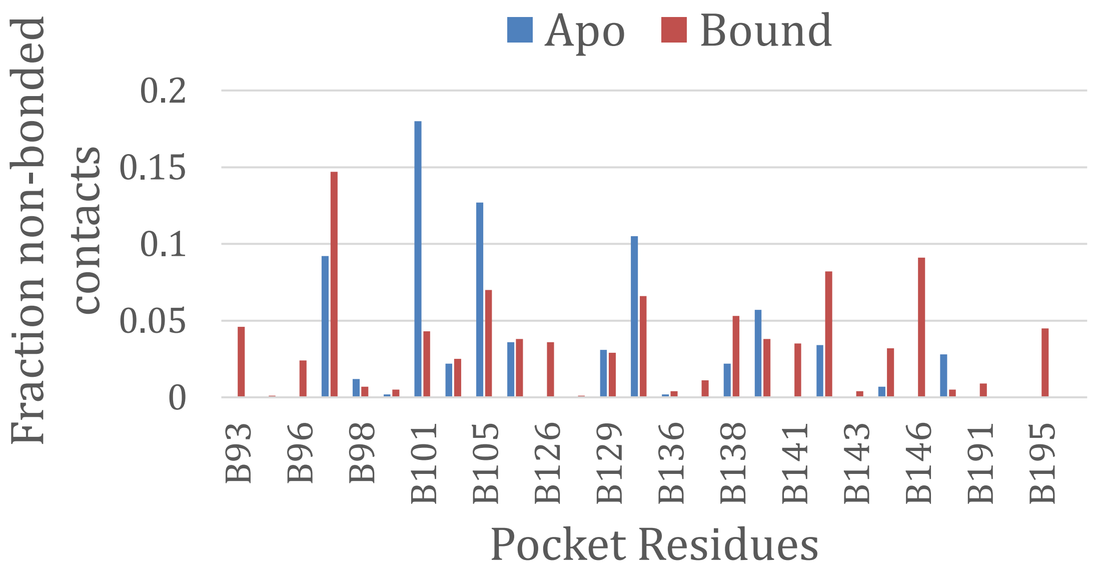

Figure 9: An example of pocket fingerprints for mapping of the bound and unliganded structures. These pocket fingerprints are for the Bcl-xl case, and the BSSC for the unliganded mapping is 0.498.

4.3 Results

To test and validate our method, 17 systems were selected from (Grove et al., 2013) and (Kozakov, Hall, et al., 2015), listed in Table 4. To highlight the efficiency increases for multiple rotamers, these 17 systems were chosen based on significant number of alternate sidechain conformations for movable sidechains within the binding pocket. Each of these 17 systems have both a bound and unliganded structure available. We selected pockets of interest on each system using mapping results of the unliganded structure. After selection of movable sidechains as described above, alternate conformers for each movable sidechain were generated using the EGLM method. The optimal rotamer for each movable sidechain were then selected using the enhanced PIPER program. The BSSC results for the unliganded and optimized structures are presented in Table 4 and Figure 11. Our results show moderate improvement on many systems, and large improvements on a few systems. For example, the Bcl-xl system goes from a BSSC of 0.5 to 0.67. We can see in Figure 12 that by optimizing the rotamers, the pocket becomes more open on the right side and allows for FTMap to find a consensus cluster that could not be found on the unliganded structure.

Speed

Our algorithm achieves a speedup of between 5 to 20 times faster for the global rigid body sampling step. The amount of time spent on mapping is displayed in Table 4. The mapping program can take advantage of multiple cores by using one thread per core. Timings are total time spent across all threads. The runs were performed on machines with 16 core Xeon chips. The decrease in computation time achieved is dependent on the number of rotamers as well as the size of the system, and we see a greater speedup for systems with more rotamers to search, as expected. This is shown graphically below in Figure 10.

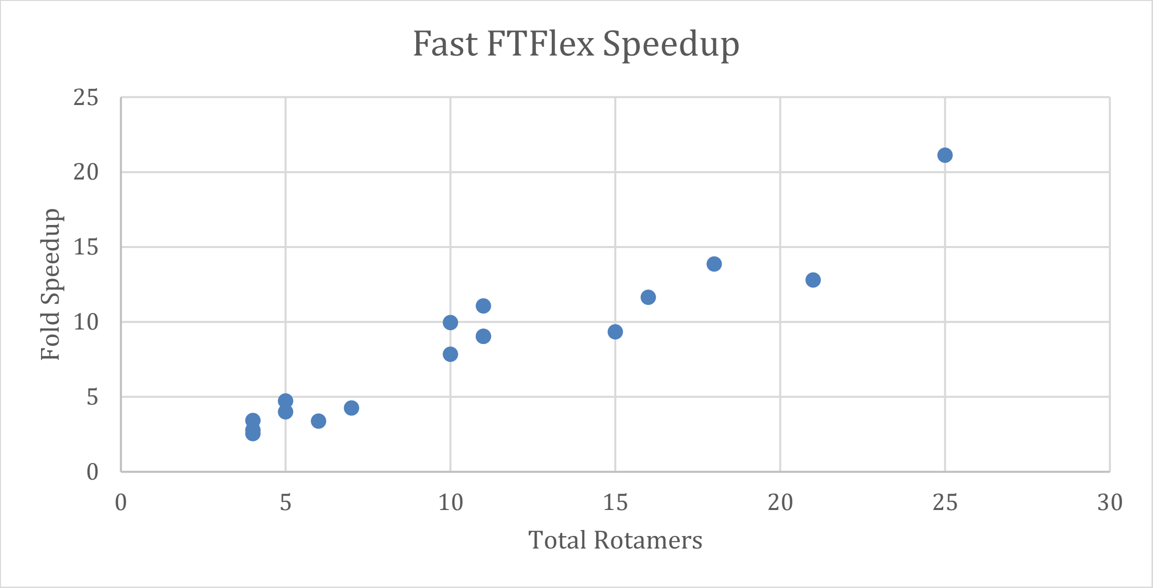

Figure 10: Graph of the relative speed up using the fast PIPER program for rigid docking versus repeated application of classic PIPER. As expected, we see a near linear speed up depending on the number of rotamers searched.

		Correlations		**Timings in seconds**			
PDB	Total rotamers	Unliganded	Fast FTflex	Fast FTFlex Time	Single PIPER TIme	Total PIPER Time	Fold Speedup
1ai9	7	0.81	0.93	18,906	11,519	80,632	4.26
1e15	18	0.91	0.93	11,615	8,954	161,163	13.87
1ea5	16	0.43	0.84	14,896	10,840	173,447	11.64
1jcz	4	0.91	0.82	44,007	28,142	112,569	2.56
1nsb	4	0.95	0.98	27,972	19,424	77,697	2.78
1ob3	25	0.98	0.97	8,781	7,418	185,440	21.12
1pdb	10	0.93	0.84	19,835	19,761	197,609	9.96
1pfq	15	0.98	0.88	30,234	18,834	282,510	9.34
1pud	11	0.99	0.99	11,147	9,151	100,662	9.03
1pw2	11	0.89	0.93	16,250	16,344	179,785	11.06
2bls	6	0.80	0.77	21,350	12,026	72,154	3.38
2nxr	5	0.98	0.98	9,665	7,750	38,749	4.01
1phc	5	0.92	0.92	16,965	16,040	80,199	4.73
1zvi	4	0.96	0.95	15,169	13,004	52,016	3.43
1r2d	11	0.50	0.67	6,722	5,515	60,670	9.03
1r6k	21	0.81	0.80	12,851	7,835	164,531	12.8
1cqr	10	0.67	0.57	8,696	6,819	68,191	7.84

Table 4. Summary of test cases and their results. The timings shown are the best of three runs, and the fold speedup is the ratio of total PIPER time versus fast FTFlex time.

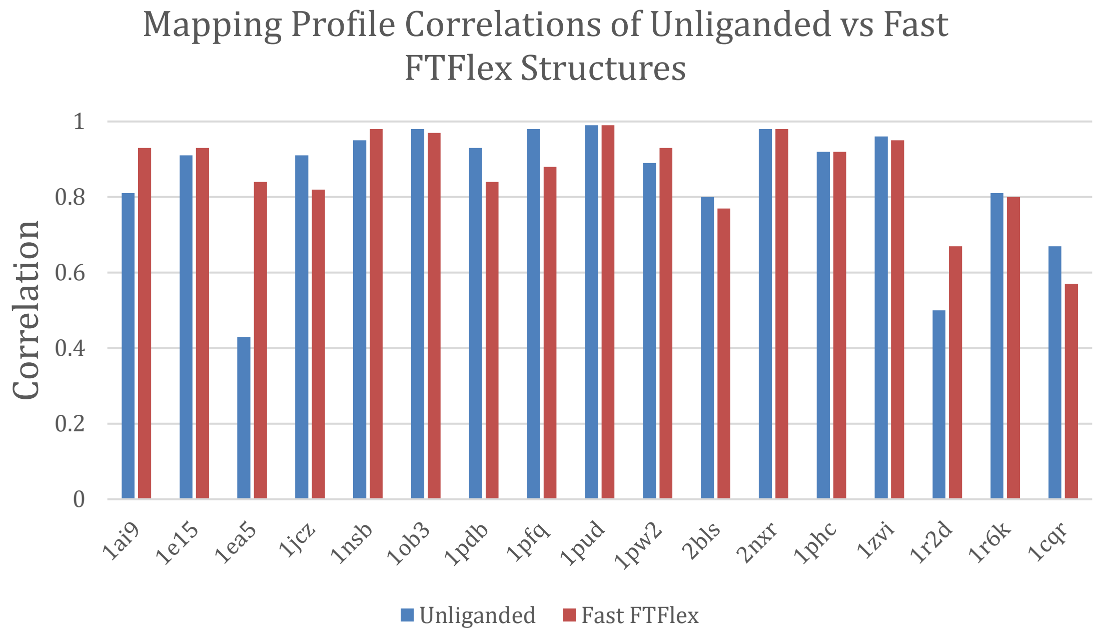

Figure 11: Mapping profile correlations for test cases in Table 4. We can see that the application of the FTFlex algorithm with fast sidechain search generally does not decrease the correlation when it was already high. In a few cases, such as 1ea5 and 1r2d, the correlation increases quite significantly, signifying that the mapping results are much more similar to the mapping results of the bound structure.

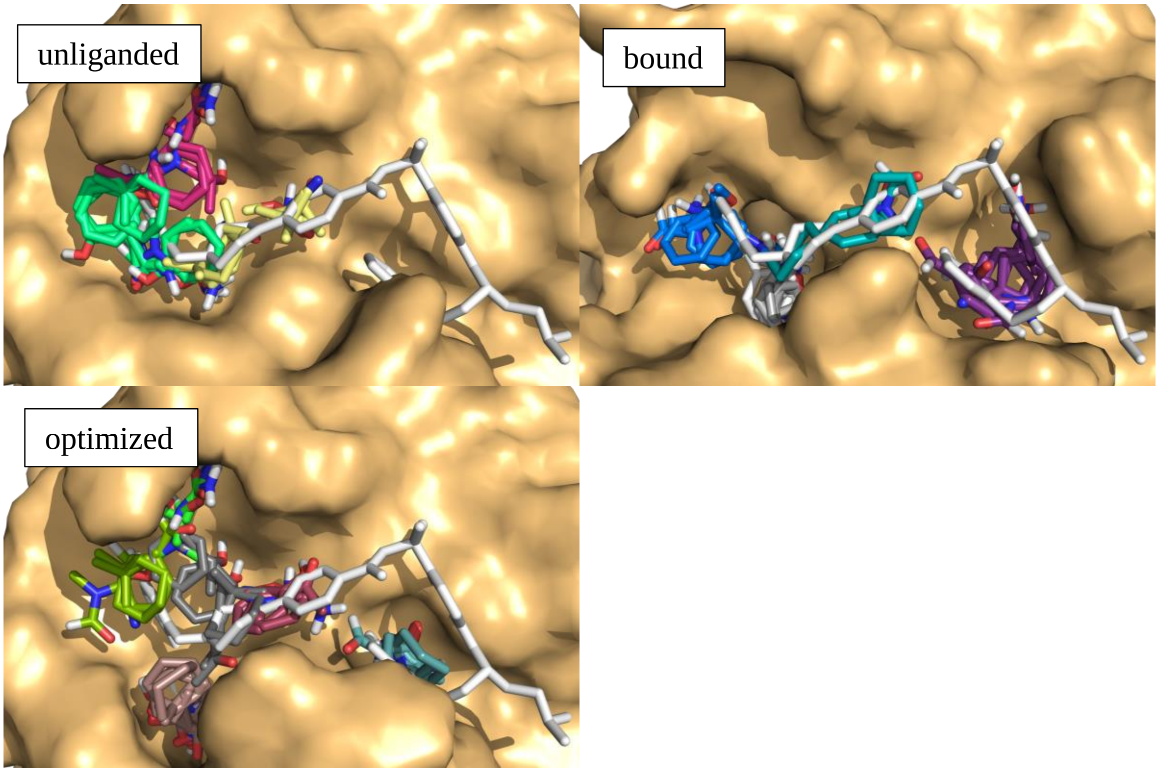

Figure 12: Comparison of mapping results for the Bcl-xl system (PDB ID: 1R2D). The protein is shown as a surface in the background, while the ligand from the bound structure is show in white. Results from FTmap are shown in colored sticks. We can see that the unliganded mapping is missing a consensus cluster on the right side of the molecule that FTmap is able to find when using the bound structure. We recover this consensus cluster after using the structure with sidechain conformers predicted by the FTflex method.

4.4 Discussion

These results clearly show the new algorithm is faster than multiple applications of the old algorithm, while obtaining results of similar quality. We note that the results are not identical to use of the previous algorithm. One possible reason may be because the decomposition of the scoring function to use smaller grids for the sidechains prevents us from smoothing the energy grid as we normally do when applying PIPER. However, this is unlikely to have a large effect on the quality of results. Another limitation of using decomposed energy grids is the inability to use scoring functions which are computed using the global state of the entire protein, as is done for electrostatic component of the scoring function in certain systems. However, in such cases we can use a slightly less accurate Coulombic model for the electrostatic energy function, and still obtain meaningful results. Even with such limitations, the speed of the new program will make it a useful tool for studying many types of systems where the ability to quickly test out many conformers of interface residues is required.

CHAPTER FIVE: Conclusion and Future Directions

This work has increased the accuracy of protein-protein docking, as well as increasing the efficiency of global rigid body sampling with flexible sidechains. Protein-protein docking for many systems is already quite accurate, but is improved when additional experimental data is available and can be incorporated into docking algorithms. The existing ClusPro protein-protein docking algorithm was enhanced by using SAXS data when available to filter rigid body docking results by their fit to the SAXS data, which selects for conformations that better match the shape and size of the complex as determined by SAXS. In addition, ClusPro can now also make use of distance restraints for a protein complex, which can be generated from various types of experimental data. These restraints are used to restrict the region of the global space of rotations and translations which are searched for energy minima. Finally, a novel decomposition of the correlation functions used in PIPER into separate grids for rigid and moving parts of the protein led to significant increases in the efficiency when sampling multiple sidechain conformations. This enhanced method was applied to the existing FTFlex method as a proof of concept. In the future, we hope to use this new fast PIPER program in other applications, such as fragment docking and protein-protein docking.

APPENDIX 1

Description of Restraint JSON File Format

A restraint set is composed of one or more restraint groups, which is composed on one or more restraints. At both the level of the restraint set and restraint group, users can specify how many restraints groups or restraints are required to be satisfied, respectively. Each restraint specifies a residue on both the receptor and the ligand, and specifies a maximum and minimum distance. If the minimum distance between any pair of atoms in the receptor residue and ligand residue is more than the specified minimum, and the maximum distance between any pair of atoms in receptor and ligand residue, then the restraint is considered satisfied. A restraint file would like this example:

```
{
  “required”: 1,
  “groups”: [
    {
      “required”: 1
      “restraints”: [
        {
          “type”: “residue”,
          “dmax”: 1.0,
          “dmin”: 1.0,
          “rec_chain”: “A”,
          “rec_resid”: “1”,
          “lig_chain”: “B”,
          “lig_resid”: “1”
        }
      ]
    }
  ]
}
```

Additional groups and restraints may be added by following the template above.

BIBLIOGRAPHY

Allen, K. N., Bellamacina, C. R., Ding, X., Jeffery, C. J., Mattos, C., Petsko, G. A., & Ringe, D. (1996). An Experimental Approach to Mapping the Binding Surfaces of Crystalline Proteins. *Journal of Physical Chemistry*, *100*(7), 2605–2611.

Beglov, D., Hall, D. R., Brenke, R., Shapovalov, M. V, Dunbrack, R. L., Kozakov, D., … Vajda, S. (2012). Minimal ensembles of side chain conformers for modeling protein-protein interactions. *Proteins*, *80*(2), 591–601.

Bohnuud, T., Kozakov, D., & Vajda, S. (2014). Evidence of conformational selection driving the formation of ligand binding sites in protein-protein interfaces. *PLoS Computational Biology*, *10*(10), e1003872.

Brenke, R., Hall, D. R., Chuang, G.-Y., Comeau, S. R., Bohnuud, T., Beglov, D., … Kozakov, D. (2012). Application of asymmetric statistical potentials to antibody-protein docking. *Bioinformatics*, *28*(20), 2608–2614.

Chen, R., Tong, W., Mintseris, J., Li, L., & Weng, Z. (2003). ZDOCK predictions for the CAPRI challenge. *Proteins*, *52*(1), 68–73.

Chuang, G.-Y., Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2008). DARS (Decoys As the Reference State) Potentials for Protein-Protein Docking. *Biophysical Journal*, *95*(9), 4217–4227.

Comeau, S. R., Kozakov, D., Brenke, R., Shen, Y., Beglov, D., & Vajda, S. (2007). ClusPro: Performance in CAPRI rounds 6-11 and the new server. *Proteins*, *69*(4), 781–785.

Debye, P. (1915). Zerstreuung von Röntgenstrahlen. *Annalen Der Physik*, *351*(6), 809–823.

Dennis, S., Kortvelyesi, T., & Vajda, S. (2002). Computational mapping identifies the binding sites of organic solvents on proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(7), 4290–5.

Graewert, M. A., & Svergun, D. I. (2013). Impact and progress in small and wide angle X-ray scattering (SAXS and WAXS). *Current Opinion in Structural Biology*, *23*(5), 748–754.

Grove, L. E., Hall, D. R., Beglov, D., Vajda, S., & Kozakov, D. (2013). FTFlex: accounting for binding site flexibility to improve fragment-based identification of druggable hot spots. *Bioinformatics*, *29*(9), 1218–1219.

Hwang, H., Vreven, T., Janin, J., & Weng, Z. (2010). Protein-protein docking benchmark version 4.0. *Proteins*, *78*(15), 3111–4.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., & Vakser, I. A. (1992). Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences of the United States of America*, *89*(6), 2195–2199.

Kozakov, D., Beglov, D., Bohnuud, T., Mottarella, S. E., Xia, B., Hall, D. R., & Vajda, S. (2013). How good is automated protein docking? *Proteins*, *81*(12), 2159–2166.

Kozakov, D., Brenke, R., Comeau, S. R., & Vajda, S. (2006). PIPER: an FFT-based protein docking program with pairwise potentials. *Proteins*, *65*(2), 392–406.

Kozakov, D., Grove, L. E., Hall, D. R., Bohnuud, T., Mottarella, S. E., Luo, L., … Vajda, S. (2015). The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. *Nature Protocols, 10*(5), 733–755.

Kozakov, D., Hall, D. R., Beglov, D., Brenke, R., Comeau, S. R., Shen, Y., … Vajda, S. (2010). Achieving reliability and high accuracy in automated protein docking: Cluspro, PIPER, SDU, and stability analysis in CAPRI rounds 13-19. *Proteins, 78*(15), 3124–3130.

Kozakov, D., Hall, D. R., Jehle, S., Jehle, S., Luo, L., Ochiana, S. O., … Vajda, S. (2015). Ligand deconstruction: Why some fragment binding positions are conserved and others are not. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(20), E2585-94.

Landon, M. R., Lancia, D. R., Yu, J., Thiel, S. C., & Vajda, S. (2007). Identification of Hot Spots within Druggable Binding Regions by Computational Solvent Mapping of Proteins. *Journal of Medicinal Chemistry*, *50*(6), 1231–1240.

Lensink, M. F., Méndez, R., & Wodak, S. J. (2007). Docking and scoring protein complexes:CAPRI 3rd Edition. *Proteins*, *69*(4), 704–718.

Lensink, M. F., & Wodak, S. J. (2010). Docking and scoring protein interactions: CAPRI 2009. *Proteins*, *78*(15), 3073–3084.

Lensink, M. F., & Wodak, S. J. (2013). Docking, scoring, and affinity prediction in CAPRI. *Proteins*, *81*(12), 2082–2095.

Leysen, S., Vanderkelen, L., Weeks, S. D., Michiels, C. W., & Strelkov, S. V. (2013). Structural basis of bacterial defense against g-type lysozyme-based innate immunity.

Cellular and Molecular Life Sciences, *70*(6), 1113–1122.

Mattos, C., & Ringe, D. (1996). Locating and characterizing binding sites on proteins. *Nature Biotechnology*, *14*(5), 595–599.

Rajamani, D., Thiel, S., Vajda, S., & Camacho, C. J. (2004). Anchor residues in protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(31), 11287–11292.

Schneidman-Duhovny, D., Hammel, M., & Sali, A. (2011). Macromolecular docking restrained by a small angle X-ray scattering profile. *Journal of Structural Biology*, *173*(3), 461–71.

Schneidman-Duhovny, D., Hammel, M., Tainer, J. A., & Sali, A. (2013). Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. *Biophysical Journal*, *105*(4), 962–974.

Yang, S. (2014). Methods for SAXS-Based Structure Determination of Biomolecular Complexes. *Advanced Materials*, *26*(46), 7902–7910.

CURRICULUM VITAE

[redacted]

[redacted]