MOLECULAR DYNAMICS IN PROTEIN STRUCTURE QUALITY ASSESSMENT AND REFINEMENT

by

Lyman Monroe

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of Biological Sciences

West Lafayette, Indiana

May 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Daisuke Kihara, Chair
Department of Biological Sciences
Department of Computer Science

Dr. Cynthia Stauffacher
Department of Biological Sciences

Dr. Wen Jiang
Department of Biological Sciences

Dr. Yulia Pushkar
Department of Physics and Astronomy

Approved by:
Dr. Janice P. Evans

For my grandfather

ACKNOWLEDGMENTS

I would like to acknowledge Dr. Genki Terashi, who coauthored the article reproduced in chapter 2. I would also like to acknowledge Brenda Gonzalez who performed the wet lab experiments and electron microscopy related to Phage G in Chapter 4, and Dr. Corey Moore, who performed the wet lab experiments and electron microscopy pertaining to USP7 in Chapter 4.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Proteins are the active biomolecules of the cell. They perform metabolic action, give the cell structure, protect the cell from antigens, give the cell motility, and much more. The function of proteins are intrinsically linked to their structures, so it is therefore necessary to characterize the structure of a protein to fully understand its function and operation. In this research the application of computational methods, primarily molecular dynamics, towards protein structure determination, refinement, and quality assessment were studied. I applied molecular dynamics techniques to four major projects; the determination of relative error of atomic models deposited with electron microscopy maps in the EMDB, solving and refining atomics structure models for the PhageG major capsid proteins, the elucidation of the structure the protein USP7 and the binding pose of a of a candidate therapeutic drug, and the determination of relative stability of candidate protein folds to distinguish near native models from not. Each year an increasing number of protein structures have been solved using electron microscopy (EM). The influx of solved structure has proven to be a boon to the community, but it is necessary to note that the quality EM maps vary substantially. To understand to what extent atomic structure models generated from EM matched their respective maps, two computational structure refinement methods were used to examine how much structures could be refined. The deviation from the starting structure by refinement, as well as the disagreement between refined models produced by the two computational methods, scaled inversely with both the global and local map resolutions. The results suggested that the observed discrepancy between the deposited maps and refined models is due to the lack of resolvable structural data present in EM maps at low to moderate resolutions, and therefore these annotations must be used with caution in further applications. I also successfully implemented molecular dynamics as a method for protein structure quality assessment. Proteins tend towards shapes which minimize their energy. Experimentally, the stability of a protein can be measured through several techniques, one such technique includes the controlled application of tension to proteins in an atomic force microscopy (AFM) framework. This kind of tension-based approach is of interest as it probes the force required to unfold individual domains of a protein rather than a bulk characteristic like molting point or activity. It has been shown that key features observed in an AFM experiment can be well reproduced with molecular dynamics simulation, which has been applied to characterize the mechanisms of unfolding of proteins as well as ligand-protein

interactions. Steered molecular dynamics (SMD) was applied to pull and unfold proteins and determine the force required to unfold them. The relative force required to unfold different models with the same sequence was used to estimate relative model accuracy. This follows from the hypothesis that the structural stability of a given model's conformation would positively correlate with its accuracy, i.e. how close that model is to its native fold. It was found that near-native models could be successfully selected by comparing the forces required to unfold models, indicating that high unfolding forces indeed indicated high model stability, which in turn correlated with model accuracy. I also applied molecular dynamics-based approaches for refinement of protein structures that are determined from cryo-EM density maps. Computational approaches for protein structure refinement are often developed with the design aim of requiring a user input and experimental data. I modeled the atomic structure of the major capsid protein gp27 and the decoration protein gp26 of PhageG to a 6.1Å resolution electron microscopy map. PhageG modeling was done by mapping the sequences to a presumed homolog (Hk97), arranging the subunits into hexamers and trimmers as suggested by mass spectroscopy data, rigid docking to respective map segments, refinement against half maps using MDFF across a range of weights, and then finally refinement to the whole map using the optimized weight. I also modeled the atomic structure of the protein USP7 to an 8.2 Å resolution map. USP7 modeling was done by combining crystalized domains of the whole structure, rigidly docking the model to the EM map by hand, and then refining in a similar manner as PhageG, with the added approach of weight scaling to overcome local minima along the relaxation. The USP7 model was further validated by exhibiting a ligand-protein binding pose, determined by glide, which corresponded to enzymatic activity mutation assays. In summary I applied molecular dynamics, in conjunction with other computational methods, towards protein structure determination, refinement, and quality assessment.

CHAPTER 1. INTRODUCTION

The determination of protein structure at an atomic scale remains one of the most challenging and important problems in the field of biology. Determining the accurate position of the atoms in a protein, both backbone and side chains, can be critical for the understanding of chemical mechanisms, development of drug molecules, the characterization of disease-causing mutations, and much more. The benefits that come from accurate structure determination can have different significance depending on the structure in question, be it a spike or capsid component of a virus, a protein harmful to human or animal health, or a critical enzyme in a pathway we wish to understand or replicate. In the case of the spike and capsid proteins of disease-causing viruses, the structure can give us clues about potential antibody binding epitopes or tell us about the assembly of those viruses, both of which can be used to develop vaccines and therapeutic drugs (Haiyan Zhao, 2016) (Christian G. Noble, 2012). For harmful proteins (such as toxins, prions, or carcinogenic proteins) structure determination can lead to the discovery of druggable regions which can be exploited to denature the harmful protein or cause reduced function (Fuqiang Ban, 2017). For enzymes which we wish to understand or replicate, structure determination can lead to insights of their function on a chemical level, and even reveal possible ways in which site directed mutagenesis can lead to increased function (Baker, 2019).

A natural method for the determination of atomic structure of proteins is to combine computational approaches, which rely on heuristic and theoretical frameworks, with experimental data in a complimentary way. Some experimental methods used to determine the structure of proteins include x-ray crystallography (Savino, 2008), Nuclear Magnetic Resonance (NRM) (Wüthrich, 1990), Electron Microscopy (EM) (Yip, 2020), and Small Angle X-ray scattering (SAXS) (Yang S. , 2014). High resolution methods like x-ray crystallography and NMR can produce highly accurate structures but are limited in the size of structures they can solve (Davis, 2003), and often present significant difficulties when applied to multi-protein complexes (Wang H.-W. a.-W., 2017). EM and SAX, on the other hand, are much more amenable to being used to solve large and complex structures, though at a lower resolution on average than the former two methods (Wang H.-W. a.-W., 2017) (Putnam, 2007). It should be noted that EM can produce resolutions comparable to, or every better than, x-ray cystography (Yip, 2020), but lower resolution maps are still common. According to the statics of the EMDB, as of this writing, of the

15113 maps which have been published, 50% have resolutions worse than 5Å, and 31% of all maps having a resolution worse than 10Å (Catherine L. Lawson, 2016). Low resolution data can be difficult to annotate by eye, but it can, if approached with care, be addressed computationally.

There are many computational approaches for predicting the structure of proteins. Computational approaches can be broadly placed into four categories, ab initio prediction (Yang J. Y., 2015) (Adhikari, Bhattacharya, Cao, & Cheng, 2015) (Carol A. Rohl, 2004), homology modeling (Eswar N., 2008) (Torsten Schwede, 2003), and structure refinement (Aleksander Kuriata, 2018) (Parimal Kar, 2013) (Luca Monticelli, 2008). After candidates structures have been predicted quality assessment (QA) can be applied to select the most promising predicted structures (Mereghetti, 2008) (Bhattacharya, 2008) (Benkert, 2008).

Ab initio methods, such as ROSETTA (Carol A. Rohl, 2004), use only an amino acid sequence as an input to determine a possible structure, and are most useful when any other data is scares. Homology modeling methods, such as MODELLER (Eswar N., 2008), use an amino acid sequence and an already determined structure with a similar sequence, which is hopefully a homolog to the protein whose structure is be solved, to generate a structure. Homology modeling used the assumption that homologs have similar conformations and can therefore be used to approximate the unknown's general fold (Webb & Sali, 2016). Refinement techniques, such as MDFF (Leonardo G. Trabuco E. V., 2009), use a starting structure that is thought to be close to its native structure and frequently, though not always, use complimentary experimental data such as Electron Microscopy maps. Finally, quality assessment approaches, such as QMEAN (Benkert, 2008), use a variety of scoring methods to distinguish high-quality and low-quality structure predictions. These methods can be used to complement one another in a variety of way, for example ab initio approaches can produce several candidate structures, QA can then help to identify promising candidate structures which can then be passed to structure refinement (Terashi G, 2018).

Below we will discuss some of the experimental techniques which can be integrated with computational modeling to produce useable protein structure models. We will then discuss a variety of methods used for flexible modeling and refinement of protein structures, various structure predictions methods, and protein structure quality assessment approaches. We will end with a short discussion of two real applications of combining these techniques. In summary we will be discussing protein structure determination from an experimental framework, then a

theoretical computational framework, and finish with some real-world examples of integrating the two.

1.1 Experimental Structure Determination

There are a variety of methods used to experimentally determine the structure of a protein. They vary in what structures they work best for, how well they handle complexes, and what kinds of resolutions are routinely obtained. In this section, I will briefly describe some of the most popular methods.

1.1.1 Electron Microscopy

To understand the fitting of atomic models to electron microscopy maps, we must first describe the technique. Electron Microscopy (EM) has garnered significant interest in the past decade. This method is an imaging technique for the viewing of matter too small for optical wavelengths of light to visualize. As the name suggests, Electron Microscopy utilizes electrons as a substitute for light in the viewing of specimens, as electrons can interact with much smaller details of matter than optical light. The use of a beam of electrons for microscopy presents several technical changes to the way that a more conventional light microscope is used, but fortunately most of these modifications have a direct analog in optical microscopy.

The Electron Microscope, its components, and their functions

If an Electron Microscope were to be stripped down to the bare minimum required to function, it would consist of four components; an electron source, a limiting aperture, an objective lens, and a means to measure the electrons at the sample such as film or a CCD. This is of course too minimal for any practical application but highlights the similarity to a conventional optical microscope.

The electron source provides electrons which are then selected by emission vector to comprise the beam which will eventually impinge on the sample for viewing. Electron sources come in a variety of forms, such as tungsten filaments and field electron guns. This serves the same purpose as the light bulb, or candle before it, in optical microscopy which generated an abundance of photons for the optical microscope to guide to its target (Cheng, 2018).

The limiting aperture blocks electrons that are not traveling in a generally desirable direction from passing through the microscope towards the sample and imaging medium (Cheng, 2018). This starts the process of controlling the electrons so that they can be focused on to the sample of interest. This functions the same way as an aperture in an optical microscope does, limiting the amount of light that is passed to the focusing lens.

The objective lens in an electron microscope is not made of glass, but rather consists of a magnetic field (Cheng, 2018). The material components of lenses in an electron microscope are coils of wire, or electromagnets, which produce very precise magnetic fields when current passes through them. Electrons interact with magnetic fields in a manner predictable using Maxwell's equations, so the production and tuning of magnetic fields allows for the focusing of electrons onto a sample of interest.

After the electrons have been narrowed down by the aperture, focused by the lens, and have interacted with the sample, it is time to measure the electrons. The electrons which pass through the sample are detected through their interaction with some material; classically this material was film, as the electrons would develop the film where they collided. While film had several benefits over devices like CCDs, the images would then need to be digitized in some way, usually by photo-scanner, if any further processing were required, however, scanning could result in data loss (Yin, 2018). The preferred component in modern EM for this task is a direct electron detector, which mitigates some of the issues of using a CCD, while collecting the data in a digitized format for further analysis (Benjin, 2020).

Single Particle Electron Microscopy

Contemporary Electron Microscopy can be set up in several ways, including Transmission Electron Microscopy (TEM), which can be used to study portions of or all of a cell, and Single Particle Cryo-EM, which can be used to image, among other things, proteins or viruses (Benjin, 2020). Here we will focus on Single Particle Cryo-EM.

The theoretical framework for Single particle EM was developed in the 1970 simultaneously by several labs (Frank, 2018) (Dubochet, 1988). In the early days of the method resolutions were low but, since it did not require a crystal, single particle cryo-EM drew much attention. Production EM images were only 2D projections (particles) of biological molecules until De Rosier and Klug demonstrated that 2D projections could be used to reconstruct 3D

structures in 1968 (DeRosier & Klug, 1968). The next challenge for electron microscopists to overcome was to find a way to keep protein samples hydrated under the high vacuum required by the electron beam. This was solved through the development of a plunge freezing technique by Dubochet et al. in the 1980s (Dubochet, 1988) (M. Adrian, 1984).

Preparation of a protein sample for examination in an electron microscope involves placing a drop of solution containing protein onto a thin grid before, in the case of cryo-EM, flash freezing (Benjin, 2020). This leaves many proteins frozen in suspension in a thin film of vitreous ice. This thin film is then placed in an electron microscope for data collection. The data collected from electron microscopes are called micrographs and contain many images of the proteins frozen in place. The individual protein images are referred to as particles and must be selected, or picked, for further analysis (Benjin, 2020). This particle picking can be semi-automated by using software such as Eman2 or Relion (Guang Tang, 2007) (Scheres, 2012).

Particles in micrographs are often of low resolution, low contract, or both (Cheng, 2018). The low resolution and contrast of the micrographs greatly limits what kind of analysis that can be reliably done with them. One method for dealing with this is known as single particle analysis (Christian Suloway, 2005). Single particle analysis techniques have been developed to overcome the problem of low resolution and contract by combining information from each particles in micrograph data (Christian Suloway, 2005). This can produce a single, representative, high resolution 3-dimensional image of the sample protein. This 3-dimensional image is often called a map or EM map. These maps contain density information about the sample protein, but do not contain any explicit atomic or chemical information. Electron microscopy maps (EM maps) are often the end products of single particle electron microscopy, but are often used for further downstream analysis of protein structure, and can be annotated with an aligned atomic structure (Leonardo G. Trabuco E. V., 2009). A representative workflow for Single Particle Electron Microscopy can be seen in Figure 1.1.

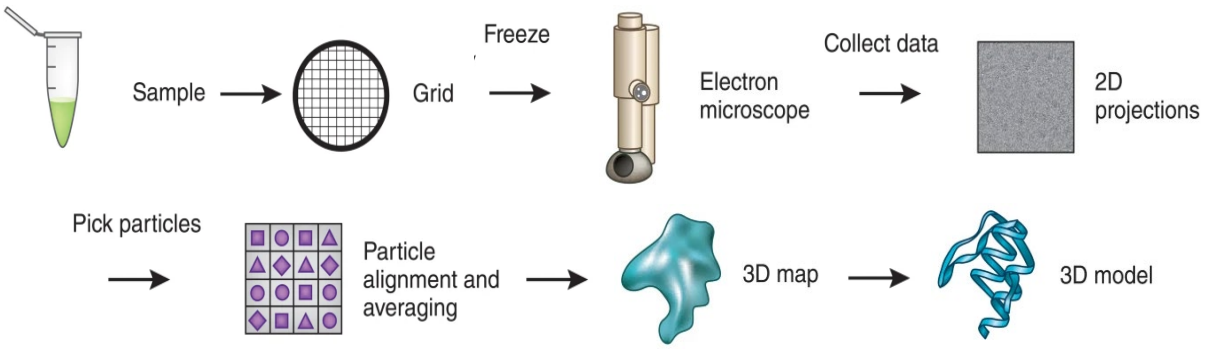

Figure 1.1: Workflow of single-particle cryo-EM. A sample is purified before placing it on a grid. After the sample has ben placed on the grid, and typically dabbed, the sample and grid are plunged into a cryogen to freeze the sample as quickly as possible. The Frozen sample is then loaded into an electron microscope where 2D projections are generated and collected. Within the 2D projections, individual molecules can be seen and are selected. These molecular projections, called particles, are aligned by orientation and averaged, generating what is known as a class average. These class averages are then used to construct a 3D map density, which can subsequently be used to generate a 3D atomistic models (Doerr, 2016).

Electron microscopy can be used on nearly any protein structure but favors larger symmetric proteins (Noble & al., 2018). EM does not always produce resolutions as high as x-ray crystallography or NMR, though that is steady changing (Benjin, 2020), but EM can more readily determine a large structure or multi-meric complex than x-ray crystallography or NMR. In this way, EM compliments x-ray crystallography and NMR, as EM can be used to determine the structure of larger proteins and protein complexes at a moderate to low resolution, while x-ray crystallography and NMR can both be used on smaller subunits of a larger complex. After the acquisition of smaller subunit structures at high resolution and the whole complex at moderate to low resolution, these two data sets can be combined to generate a high-resolution atomic structure of the whole complex that none of the methods alone are conventionally capable of. The final product of single particle electron microscopy before fitting atomic models is an EM-map. EM-maps are three-dimensional representations of the sample structure and take the form of a voxel grid where each point in the grid is assigned a value based on signal density. Resolution of EM maps vary significantly from, and some examples of the protein GroEL at different resolutions is shown in Figure 1.2.

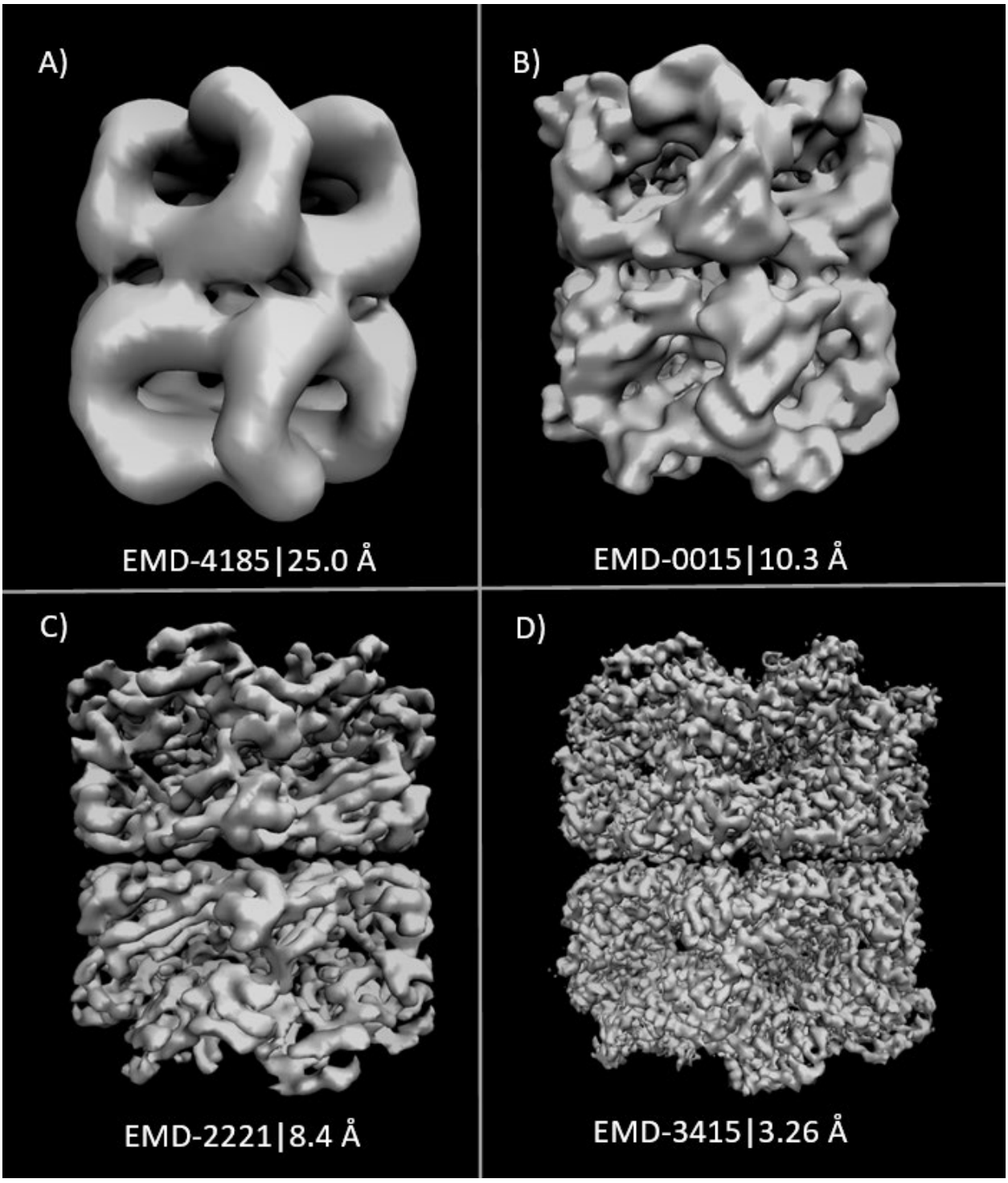

Figure 1.2: Cryo-EM maps of GroEL solved at different resolutions. Panels inset with each map's EMDB ID and It reported resolution. Each map is displayed at the author recommended contour level.

1.1.2 X-ray crystallography

X-ray crystallography has, until relatively recently with the advancement of cryo-EM, been the undisputed favored method for protein structure determination (Smyth & Martin, 2000). The aim of x-ray crystallography is to determine the three-dimensional structure of a protein from a crystal. Crystals of this kind are created from highly purified protein at high concentration (McPherson & Gavira, 2014). The crystal is then exposed to a beam of x-rays which diffracts based on the molecular and atomic geometry of the crystal. The resulting diffraction pattern can then be used to determine the shape of the electron density within the crystal, and an atomic model can then be built from there (Smyth & Martin, 2000). X-ray crystallography often produces high resolutions structures, however there are several factors which make it one of the most challenging techniques for structure determination. Some factors which make x-ray crystallography challenging are the need for a reliable source of protein at high concentrations, purifications of the protein to be crystalized, the crystallization process, and the need to generate multiple crystals if there are not already solved sufficiently similar structures which can be used for molecular replacement (Evans & McCoy, 2008).

Sample preparation

In X-ray crystallography a protein sample must first be purified then crystallized. Purification can be performed in several ways, including expressing the protein in a plasmid vector with a poly-His-tag and then immobilizing the protein on a model ion or using a poly-his antibody (Kim Y, 2011). Rarely is one purification technique sufficient, but when multiple purifications using complimentary selection criteria are used in series, purification can typically be achieved (Kim Y, 2011). After a sufficiently purse sample has been produced, the goal is to crystalize the protein by achieving supersaturation. Supersaturation can be achieved through the addition of salts, modification of pH, changing the ionic strength of the solution, among other approaches. Crystallization is far from a trivial process, and the difficulty of forming a crystal will depend greatly on the character of the protein of interest (McPherson & Gavira, 2014).

Data collection and interpretation

After a crystal has been formed, it is integrated with a beam of x-rays which diffract when interacting with the crystals structure, forming a pattern on a detection screen opposite the x-ray source from the crystal (Smyth & Martin, 2000). The best diffraction patters come from powerful beamlines like the one found at the Argon National laboratory, but preliminary scattering data can be collected on a benchtop device.

The unique character of a diffraction pattern can be used to determine the overall geometry of the electron density in the crystal, referred to as a structure factor, which can in-turn be used to determine the atomic coordinates of the crystal (Smyth & Martin, 2000). An Example diffraction pattern and a segment of the resulting solved electron density is shown in figure 1.3.

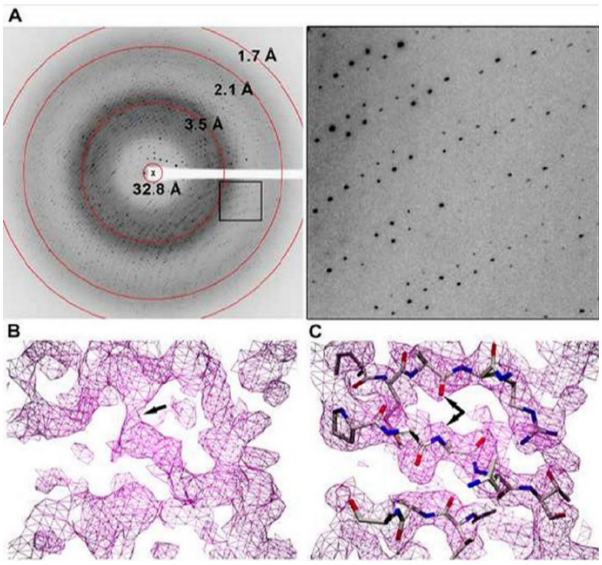

Figure 1.3: Example x-ray diffraction pattern and refined structure (Audrey L. Lamb, 2015). A) a single diffraction image obtained by the authors rotating the crystal through a small angle. Resolution of structural data increases from the center of the diffraction pattern outwards radially. B) a region of the solved electron density. C) The same density as shown in B with a fitted atomic structure. (Lamb, Kappock, & Silvaggi, 2015)

Strengths and weaknesses

X-ray crystallography produces high resolution atomic structures, but is costly both in material and time, and some structures display a high level of recalcitrance to crystallization (Wang H.-W. a.-W., 2017) (McPherson & Gavira, 2014). Aside from its size limitations, x-ray crystallography can also fall short of resolving flexible regions of a protein, and can also suffer from deformation from the native structure caused by crystal contacts (Kim Y, 2011).

1.1.3 NMR

NMR is a method in which a sample in solution is pulsed by a strong magnetic field causing the nuclei of the atoms in the sample to resonate at a characteristic frequency based on their chemical environment (Michael R. Gryk, 2010). From the observed frequencies, the relative position of atoms can be determined to a high degree of precision, granting resolutions in the angstrom range. The general workflow of NMR can be seen in Figure 1.4.

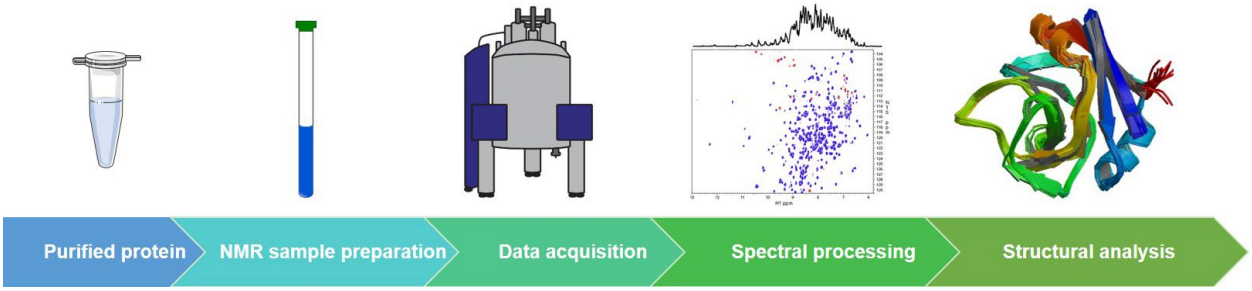

Figure 1.4: Workflow of NMR, (creative-biostructure.com, n.d.). From left to right, protein is purifying as in other techniques, sample is suspended in a deuterated solvent, the sample is then subjected to magnetic fields, data is acquired, and finally a structure is solved from the collected data.

NMR has the added benefit of giving some dynamic information about the protein conformational ensemble (Ishima & Torchia, 2000), that is the variation in the protein's native conformation at a given temperature. NMR provides data useful at an atomic scale, but it suffers from size limitations of the samples that it can be applied to (Michael R. Gryk, 2010).

Sample preparation

Protein samples for NMR are purified and suspended in an aqueous solution, similarly to SAXS, before expose to a strong oscillating magnetic field (Gaetano T. Montelione, 2000). Protein samples can be deuterated for Amide hydrogen-deuterium exchange for determination of regional flexibility and finding binding sites (Gaohua Liu, 2005).

Data collection and interpretation

After a sample has been purified and prepared for NMR, it is loaded into an NMR spectrometer (Figure 7) which pulses the sample with a powerful and precise magnetic field. Magnetic fields are typically multi-Tesla with a pulse frequency of approximately 600-900 Mhz

(Gaetano T. Montelione, 2000). An example NMR device and a schematic representation can be seen in the Figure 1.5.

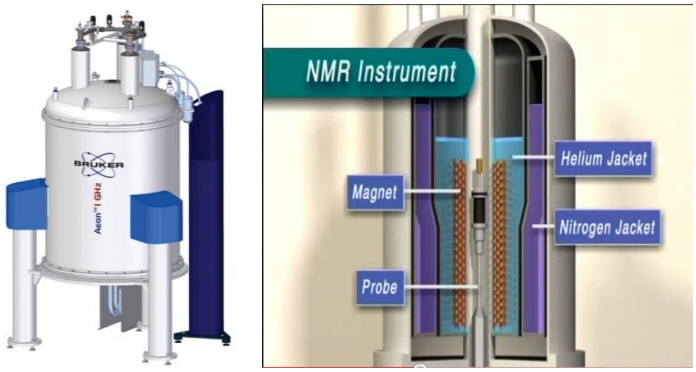

Figure 1.5: Example NMR hardware. NMR spectrometer (left) (Reish, 2015) with a schematic representation (right) (Raja & Barron, 2021). Left, an example NMR device. Right, a cutaway diagram of a typical NMR device, showing the magnetic, probe, and helium and nitrogen jackets.

The data collected from NMR is the resulting resonant magnetic field from the nuclei of the atoms in the protein. The resonance of these atoms will differ depending on the chemical environment of the nuclei and will give characteristic frequency shifts in particular environments. These characteristics shifts, such as H_N, N, C′, H_α, C_α, H_β, and C_β chemical shifts (Michael R. Gryk, 2010), give restraints for structure determining, limiting the number of possible conformations. When there are enough restrains, there remains only one unique possible solution which can satisfy all the given restraints. Example NMR data and a solved structure can be seen in Figure 1.6.

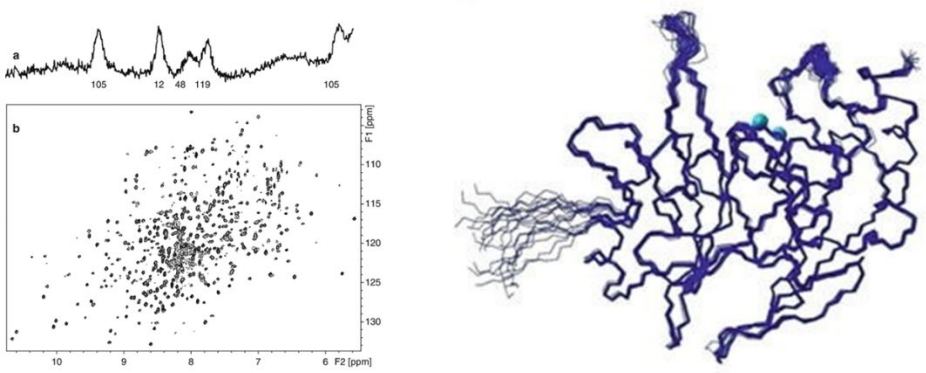

Figure 1.6: Example NMR data (left) and determined structures (right) (Roberts, 2013). Left, a) 100 MHz ^{1}H spectrum with peaks corresponding to histidines in pancreatic ribonuclease. b) 800 MHz ^{1}H-^{15}N neteronuclear single quantum coherence spectrum of NADPH cytochrome P450 reductase. Right, Ensembles of structures of metallo-β-lactamase BcII from *Bacillus cereus.*

Strengths and weaknesses

Because of the resolutions NRM can achieve, and the size regime in which it works best, the most sensible method to compare it to is x-ray crystallography. NMR gives high resolution structures of multiple conformations of a protein in an aqueous environment, which can potentially be more physiologically relevant than a crystal structure. NMR can even provide us with dynamics and binding information. The biggest weakness of NMR is its upper size limit, relegating the method to relatively small molecules (Emwas, 2015).

1.1.4 SAXS

SAXS (Small Angle X-ray Scattering) is a method in which an x-ray beam is scattered from a material and the density distribution of that material is determined by the resulting scattering pattern to within a nanometer scale. SAXS has the benefit of not requiring a crystal, as x-ray crystallography does, but suffers from the lowest resolution on this list. While SAXS can give an overall geometry of the protein, how spherical the structure is or if it has a bend in it, it does not give data at an angstrom scale, which is usually necessary for functional studies and drug design. SAXS performed well when being used to study bulk motion and protein behavior in solution. SAXS has been used to notable success in studying flexible systems as well as intrinsically disorder proteins (Bernadó, 2012) (Receveur-Brechot & Durand, 2012) (Tainer, 2011) as well as ligand binding (Gareth J Williams, 2011) and the effects of temperature on a proteins ensemble (Sara Ayuso-Tejedor, 2011).

Sample preparation

When preparing a sample for SAXS, it is necessary to purify your protein to a high degree, as SAXS will give data regardless of the quality of the sample used. It is also necessary to use an identical buffer as the zero as the buffer in the sample solution. It insufficient to make a separate buffer solution with the same measured components, as even the slightest difference in composition between the zero and sample buffers will result in substantial error. Buffer for zeros is routinely extracted from the sample solution through means of diffusion (Pauw, 2013).

Data collection and interpretation

Once the sample and buffer are ready for data collection, an x-ray beam can be shot through the sample, and the scattering pattern can be measured. From the scattering pattern two kinds of data are of primary interest, the scattering angle (q) and the x-ray intensity (I). These data are combined into what is known as a Kratky plot, which contains information about the bulk properties of a protein's structure, such as relative flexibility and how globular the protein is. A Kratky schematic is shown in figure 1.7.

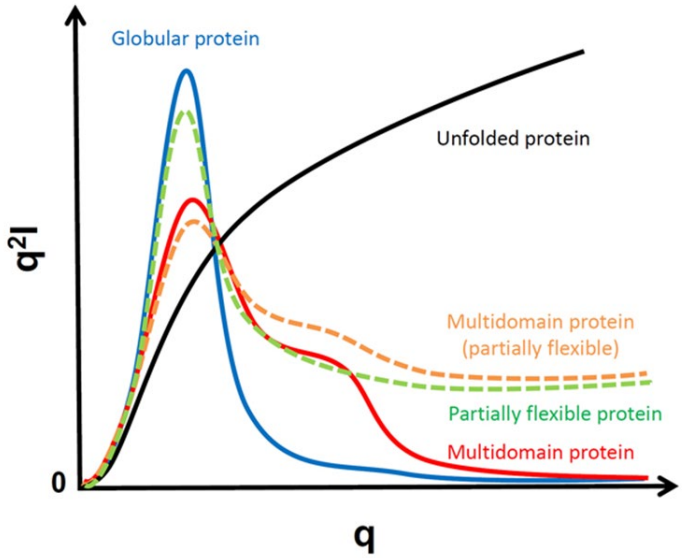

Figure 1.7: SAXS data analysis. A) Krotky plot description. A Krotky plot is a plot of $q^2I(q)$ vs q, where q is the scattering angle, and I is the x-ray intensity. Based on the shape of the curve, general characteristics of the geometry of the protein can be determined (Small Angle X-ray Scattering/Diffraction, n.d.).

After Kratky data has been collected, it is possible to generate a structure with a resolution of 1-100 nm. These structures can take the form of a dummy atom reconstruction like what is made by DAMMIF (Svergun, 1999) (Franke & Svergun, 2009), or an electron density model as is generated by DENSS (Grant, 2018).

Strengths and weaknesses

SAXS performs well relative to other methods when a protein of interest is highly flexible, very large, or difficult to purify; however, the special resolution that SAXS affords us is the most limited of the methods described here.

1.2 Molecular Dynamics

Molecular dynamics is at its core the study of the movement of atoms and molecules. In a computational framework this means the simulation of atoms and molecules to determine how they interact and where they located relative to one another in an environment. Here we will discuss the basics of Molecular Dynamics and some of its relevant applications for this work.

1.2.1 Basics of Molecular Dynamics

Molecular Dynamic simulations can be performed using a variety of software packages. The most popular of these software packages are GROMACS (Van Der Spoel, 2005), CHARMM (Brooks, 2009), and NAMD (J. C. Phillips, 2002). These packages allow for simulations to be modified and applied in a variety of ways, while leaving the core principles intact. The basics of any MD simulation is as follows; start with a system of atoms with their coordinates and characteristics, calculate forces on each atom from their interaction with other atoms, move the atoms based on the net force and the mass of each atom over a short time step Δt, update the time in the simulation by Δt, and repeat until the desired amount of time has passed in the simulation (J. C. Phillips, 2002).

The set of atoms and their coordinates in a molecular dynamics simulation often comes from x-ray crystallography and Nuclear Magnetic Resonance (NRM), in the form of a protein data bank (PDB) file (H.M. Berman, 2000). The way that atoms are characterized and handled varies across methods, but the core loop remains the same.

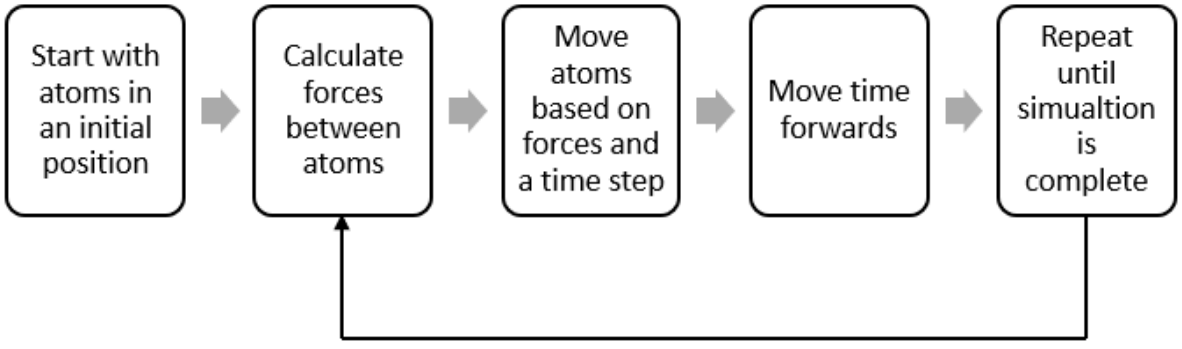

Figure 1.8: Flowchart of Basic Molecular Dynamics. Molecular dynamics simulations start with a set of atoms and their coordinates. From there forces are calculated for every atom in the simulation. Next every atom is moved based on those forces and the time step used in the simulation. After the atom positions have been updated, time is moved forward in the simulation. Finally, the simulation is looped until the desired time has passed in the simulation.

From this basic core molecular dynamics can be added to address specific problems and needs. Modifications can be made by, for example, applying additional forces to the simulation that are not calculated by default, or by forcing part or all of a protein to move in a specific way.

1.2.2 Molecular Dynamics with External Forces

Due to the linear nature of the energy terms in molecular dynamics it is possible to apply additional energy terms to a simulation. The ability to add additional energy terms is something we can take advantage of to sample dynamics or conformations of interest to us.

Here we will discuss two such methods; Molecular Dynamics Flexible Fitting (MDFF) (Leonardo G. Trabuco E. V., 2009) and Steered Molecular Dynamics (SMD).

Molecular Dynamics Flexible Fitting

Molecular Dynamics Flexible Fitting (MDFF) is a modification made to molecular dynamics by Leonardo Trabuco et al. in 2009 which is designed to fit atomic pdb structures to electron microscopy maps. It is desirable to fit atomic pdb models to EM maps because of the shortcomings of x-ray crystallography, NMR, and electron microscopy.

MDFF combines atomic models and EM map densities by converting the map density into a sort of mass potential field. This is done through the following functions.

$$U_{EM}(\boldsymbol{R}) = \sum_j w_j V_{EM}(\boldsymbol{r}_j) \tag{1.1}$$

$$V_{EM}(\boldsymbol{r}) = \begin{cases} \xi \left[1 - \dfrac{\phi(\boldsymbol{r}) - \phi_{thr}}{\phi_{max} - \phi_{thr}}\right] & if\ \phi(\boldsymbol{r}) \geq \phi_{thr} \\ \xi & if\ \phi(\boldsymbol{r}) < \phi_{thr} \end{cases} \tag{1.2}$$

Where $V_{EM}(\mathbf{r})$ is the potential for an atom to be in location **r**, $\phi(\mathbf{r})$ is the density of the EM map at poison **r,** ϕ_{thr} is a threshold density for handling background noise in the map and is set to zero by default. ξ is a scale factor, referred to from here on a the g-scale. w_j is the mass of atom j, and $U_{EM}(\mathbf{Rkl})$ is the potential energy of the whole protein in position **R**. The $U_{EM}(R)$ term is added to the other terms in a standard molecular dynamics simulation of the atomic model that we are trying to fit into the EM density (Leonardo G. Trabuco E. V., 2009). MDFF is discussed in further detail in chapter 2.

Using experimental data to derive an external force in molecular dynamics is a powerful way to merge theory with experiment, as it keeps the theoretical modeling grounded while also helping to control for overfitting to experimental data.

Steered Molecular Dynamics

SMD is a broad class of molecular dynamics in which an external force is applied to some or all the atoms in a simulation to drive spatial or dynamic sampling (al, 1999). A common application of SMD is to pull an atom or small molecule through a protein channel (Mu Gao, 2002) (Hui Lu, 1998) (Alishahi, 2019). When pulling an atom or small molecule in this way, often a dummy atom with not mass or charge is connected to the atom or small molecule by a spring-like bond, and the dummy atom is then moved at a constant velocity through the pore or the protein, while the force between the dummy atom and the atom or small molecule of interest is calculated and recorded throughout the run. This gives a general idea of how much force is required to force the atom or small molecule through the pore, as well as an idea of where the most resistance is met during this translocation. This information can be used to generate a coordinate restraint for umbrella sampling to generate a potential of mean force (PMF), to roughly determine the Gibbs free energy of passing the atom or small molecule through the pore (Alishahi, 2019).

SMD can also be applied to the unfolding of proteins in a manner analogous to atomic force microscopy (AFM) (Hughes & Dougan, 2016). Similarly, to the pore example given above, a dummy atom can be connected to one end of a folded protein, but in this case the other end of the protein is fixed in place. The dummy atom can then move away from the fixed end resulting in tension in the protein which will eventually cause a rearrangement in the structure. If pulled for long enough the tertiary and secondary structure of the protein will be ripped apart until finally the protein is completely linear, assuming no there are no crosslinks in the structure. The force exerted by the dummy atom is collected and by comparing force over time to the change in the protein structure the approximate stability of the different structure motifs can be determined.

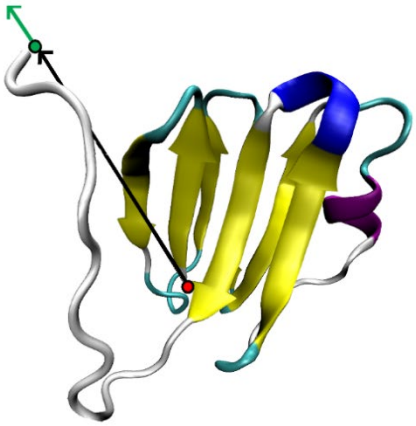

Figure 1.9: Example AFM-type setup for SMD. Here the N-terminus (red dot) is held fixed, while the C-terminus (green dot) is connected to a dummy atom which is moved along the vector represented by the green arrow.

1.2.3 Other Methods for Flexible Protein Modeling

Molecular Dynamics is not the only methodology for modeling protein motion and structure refinement. Other methods implement various heuristics to speed up computation time. Such heuristics include reducing the number of atoms in the simulation to simplifying or outright replacing the physics forcefield with an alternative approach determining optimal atom position.

Course Grained methods

All the Molecular Dynamics we have discussed so far fall under the category of all-atom Molecular Dynamics, where every atom of every molecule in the system is modeled explicitly. One of the most notable alternatives to all-atom Molecular Dynamics is course graining. Coarse grain methods differ from all-atom Molecular Dynamics in two key ways; one, the atoms are combined in various ways to simplify the model, and two, the forcefields which govern the interaction between atoms is often vastly different (Sebastian Kmiecik, 2016). Coarse Grain modeling is a broad subject, but we discuss some examples here.

Elastic Network Models

Elastic Network Models (ENM) are simplified models of proteins used to study slow dynamics (Bastolla, 2014). ENMs are constructed of point particles that are connected by linear hooking spring forces. The original ENM for proteins replaced every atom in the system with a particle (Tirion, 1996), but later ENM approaches reduced entire residues or ~10Å radius spheres around α-carbons into single particles (Ivet Bahar, 1997) (C. Atilgan, 2010). The reduction in the total number of atoms, in conjunction with the simplification of the forces acting on those atoms, trades simulation accuracy for computational speed. The increase in speed that ENM has relative to more conventional all atom MD allows for the convenient study of larger protein complex as well as comparative studies of many proteins (Togashi & Flechsig, 2018).

ENM methods can be divided into two subcategories: Gaussian Network Models (GNM) and Anisotropic Network Models (ANM). GNM and ANM approaches are distinguished from one another by a single assumption. In a GNM, fluctuations around the initial, or reference, structure are assumed to be isotropic, that is the magnitude of the fluctuation does not depend on the direction of that fluctuation. This allows the conformation space to simply be an N-dimensional, where N is the number of particles, array of fluctuation magnitudes. The simplicity of a GNM allows it to be used in normal mode analysis (NMA) to model conformation changes (Togashi & Flechsig, 2018). In an ANM, the direction of displacement from the reference structure is considered. In ANM, the force action of a particle depends on the displacement of that particle, as well as the direction of the springs connected to that particle. ANM is also used in NMA to model conformation changes (Doruker, 2000) (A.R. Atilgan, 2001).

ENMs have been developed to investigate several physiologically relevant systems; including investigation of allosteric conformation change of actin when bound to ATP (Markus Düttmann, 2012), and modeling the operation cycle of the hepatitis C virus helicase motor (Holger Flechsig, 2010).

ENM Examples

CABS (C-alpha, beta, and side chain) (Aleksander Kuriata, 2018), is a coarse-grained medium resolution method. It uses four-beads to represent a residue: the Cα, the Cβ, the center of the side chain, and the center of the peptide bond. To speed up calculations, Cα positions are

restrained to a cubic lattice of length 0.61Å. CABS uses a completely statistical (knowledge-based) force field derived from analysis structures in the Protein Data Bank (PDB). Position sampling is performed using Monte Carlo methods. CABS has been used successfully for structure prediction (Koliński & Bujnicki, 2005), modeling flexibility (Mateusz Kurcinski T. O., 2019), modeling disorder in protein structure (Ciemny MP, 2019), and molecular docking (Mateusz Kurcinski A. K., 2014).

PRIMO (protein intermediate model) is a higher resolution approach than CABS, in which the main chain is represented by three beads per residue, and sidechains are represented by one to four beads. PRIMO uses an adapted physics-based forcefield of the same form as that used for conventional all-atom MD. PRIMO grants a speed up relative to all-atom MD, which allows for the modeling of solvent effects, including the dynamics of membrane proteins (Parimal Kar, 2013).

Martini uses a one-to-four mapping approach, meaning that four heavy atoms and their hydrogens are combined into a single bead. Coarse grain beads are assigned one of four categories, polar, nonpolar, apolar, and charged, and are also assigned hydrogen-bonding capability ('d' for donor, 'a' for acceptor, 'da' for both, and '0' for neither) and level of polarity, ranging from 1 for low polarity to 5 for high polarity). Force fields for MARTINI are physics based, using the coulombic function electrostatic interactions, and a 12-6 Lennard-Jones potential for nonbonded Van der Waals interactions. Bonded interactions are modeled in a similar way to conventional MD using the same form of bond, angle, and dihedral functions. MARTINI was originally used with GROMACS, but has been adapted for use in Desmond, GROMOS, and NAMD.

Because of its simplification of lipid molecules, MARTINI has seen heavy use in the study of lipid membrane characterization (Rajagopal & Nangia, 2019) (Zgorski, Pastor, & Lyman, 2019) (Zhang, et al., 2019) and membrane proteins (Barton, et al., 2016) (Hedger, Shorthouse, Koldsø, & Sansom, 2016). MARTINI also delivers a noticeable speed up relative to conventional MD, which has resulted in its use to characterize protein conformation change (Marrink & Tieleman, 2013).

1.2.4 Protein Structure Prediction

Protein structure prediction aims to produce a three-dimensional structure of a protein from its sequence. Protein structure prediction can be broadly sorted into three categories: homology modeling, threading, and Ab initio modeling.

Homology Modeling

Homology modeling is a class of approached for protein structure prediction which directly takes advantage of previously determined and published protein structures. Generally, the method is to use the sequence of the protein with unknown structure, find a homologous protein which has already had its structure solved through experimental means, and then use the solved structure to generate a structure for the unknown protein.

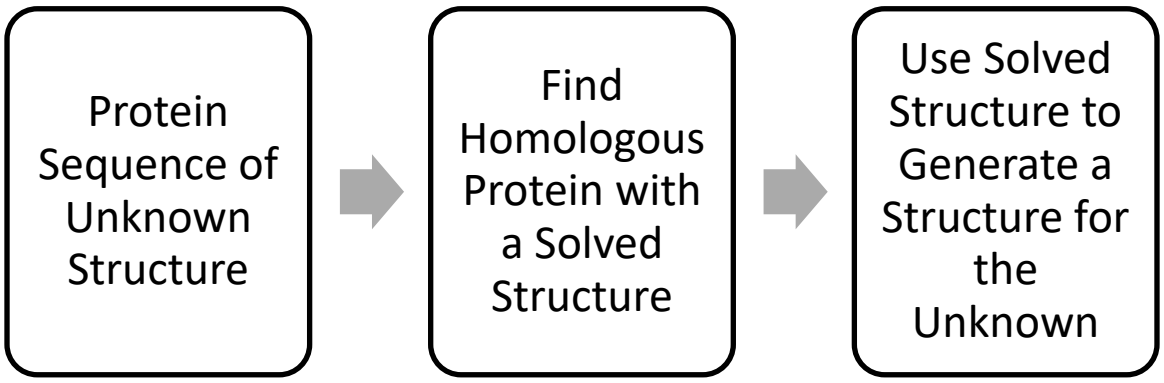

Figure 1.10: Generalized Homology Modeling Workflow. The user starts with a sequence of interest with no known structure. A search is performed, usually by means of sequence similarity, to find a homologous sequence with a structure that is known. If a homologous sequence with a known structure is found, it can then be used to produce a structure for the starting sequence.

Two well-known homology modeling programs are MODELLER and SWISS-MODEL. MODELLER is a homology modeling program. Two things are required by MODELLER; a three-dimensional PDB file of the solved structures template, and a sequence alignment of the template sequence and query sequence, the query sequence being the sequence of the protein whose structure is to be determined. MODELER uses the sequence alignment and fits the query sequence to the 3D model while satisfying spatial restraints. Special restraints are represented as probability density functions for Cα-Cα distances, main chain N-O distances, and mainchain-sidechain dihedrals. The probability density function is optimized using a variable target function (Webb & Sali, 2016). SWISS-MODEL is a homology modeling server which implements a pipeline to generate its 3d models. First, a structure template is identified by using BLAST and HHblits, then a sequence alignment is made. From there, a model is build and minimized using SWISS-MODELs rigid fragment assembly approach. Finally, the model's quality is assessed using a

statistical potential (Waterhouse, et al., 2018) (Bienert, et al., 2017) (Guex, Peitsch, & Schwede, 2009) (Studer, et al., 2020) (Bertoni, Kiefer, Biasini, Bordoli, & Schwede, (2017).

Threading and Fold Recognition

Threading is another method for protein structure prediction which uses already solved protein structures; however, in contrast to conventional Homology modeling, threading uses statistical relationships between sequence and structure in the PDB rather than rely on finding a homolog. Threading can allow for the prediction of structures which exist in a homology island, that is if no closely related proteins have been solved at the time of predicting the structure. One of the best-known threading methods is RaptorX (Peng & Xu, 2011). RaptorX utilizes a regression-tree-based nonlinear scoring function to determine how similar two protein sequences are. This tree uses several rules to determine the probability of a sequence alignment. These rules can relate to mutations scores, solvent accessibility scores, and secondary structure scores. The adaptability of this tree approach allows for different parts of an alignment to be scored by different criteria, which helps to overcome the difficulties which arise from have no near homology models (Källberg, et al., 2012) (Ma J. , Wang, Zhao, & Xu, 2013) (Ma, Peng, Wang, & Xu, 2012) (Peng & Xu, 2011) (Peng & Xu, 2011) (Peng & Xu, 2010).

Ab initio Modeling

Ab initio modeling is the most challenging kind of protein structure prediction. Ab initio modeling aims to predict a three-dimensional protein structure using theoretical principles rather than solved structures. There are several Ab initio modeling techniques, here I will describe four prominent methods: I-TASSER, CONFOLD, ROSETTA, and Alpha Fold. I-TASSER uses templates for protein structure and function prediction. The pipeline for I-TASSER is as follows: first predictions of secondary structure are made. Next, threading by LOMETS is used to produce a template. The template is then broken into fragments, which are then combined with restraints from LOMETS to produce many protein structure candidates. From there the candidates are clustered using SPICKER and the cluster centroid is selected. Finally, a fragment-guided molecular dynamics simulation is performed to generate a final structure, and the function of this structure is annotated using COACH (Roy, Kucukural, & Zhang, 2010) (Yang, et al., 2015) (Yang

& Zhang, I-TASSER server: new development for protein structure and function predictions, 2015). CONFOLD is a method for ab initio protein structure modeling which uses predicted residue-residue contacts and predicted secondary structures to generate distance, dihedral angle, and hydrogen bond restraints. Restraints are then used to build 20 models which are subsequently filtered by unsatisfied contacts and beta-sheet quality (Adhikari, Bhattacharya, Cao, & Cheng, 2015). ROSETTA is a very large software suite for macromolecular modeling. Among its capabilities is ab initio modeling, which uses a fragment-based assembly approach. Fragment based assembly works on the idea that small protein fragments are restricted to a smaller set of conformations than larger fragments, or whole proteins. In ROSETTA, fragments of 3 and then 9 residues are used to build fragments-assembly models. From there energy functions of increasing resolution are used until a final structure is produced (Carol A. Rohl, 2004). Alpha Fold is a deep learning approach to protein structure prediction. Alpha Fold is notable for its exceptional performance in the 2020 CASP. The details of the algorithm have not been released as of this writing.

1.2.5 Quality Assessment

Protein structure prediction methods may produce a structure, but that structure should not be used carelessly. It is essential to determine the quality of a predicted structure before using it for further applications. As a single protein may have thousands of proposed candidate models, it is desirable to be able to measure the quality of a model quickly and accurately. This problem has led to the development of model quality assessment methods, often abbreviated MQA or QA (we will be using QA from here on).

QA methods can be categorized in five ways; if they use a single model or use a consensus model; if they use full atom, backbone, of Cα representations of the model; if the score if global or has local detail; what kind of computational method is used for generating a score; and what features are considered.

Several approaches use only a single input model for quality assessment such as AIDE (Mereghetti, 2008), HOMA (Bhattacharya, 2008), ProQ (Wallner & Elofsson, 2007), and SVMod (Eramian & al., 2006); other methods use an approach in which several models are compared to one another, known as a consensus model approach, such as TASSER-QA (Zhou & Skolnick, 2008); while other methods use both single model metrics combined with consensus approaches,

such as MODCHECK (Sadowski & Jones, 2007), ModFold (McGuffin, 2007), and QMEAN (Benkert, 2008).

Different QA approaches use differing level of structure detail. Some methods, such as AIDE, HOMA, MODCHECK, ProQ, and Qmean consider all of the atoms of a model to determine quality, while other methods, including ModFold, consider only the backbone atoms. Other methods use even fewer atoms when determining the quality of a structure, such as TASSER-QA, which considers only the α-carbons of a protein model.

Most scoring methods, including AIDE, HOMA, MODCHECK, ModFold, SVMod, and TASSER-QA, give a single score for the entire structure, which is what is most important for distinguishing good models from bad. However, some methods, like ProQ and QMEAN, give residue level scores as well, which enables the user to make judgements on different sub-domains of a model.

Finally, many different features of a model can be considered when attempting to determine model quality. These features can be the same sort that MD uses for energy, such as bond length which is used by AIDE and HOMA, or more complex characteristics such as compactness, which is used by ModFold and ProQ.

As the popularity of Machine and Deep leaning has increased in recent years, we have also seen in increase in the number of QA methods which use these techniques. One such example is DeepQA (Cao, Bhattacharya, Hou, & Cheng, 2016). DeepQA approached the QA problem from a machine learning angle. DeepQA applies a deep belief network which implements a number of features which can be associated with model quality, such as physio-chemical characteristics, structure information, and energy. Deep QA was trained on 3113 native protein structure from the PISCES database as well as three models from CASP (CASP8, CASP9, and CASP10) (Cao, Bhattacharya, Hou, & Cheng, 2016). A summary of different QA methods is shown in table 1.

Table 1.1: Summary of different approaches and features of some popular Quality Assessment methods. Table is a recombination of tables found in (Kryshtafovych, Fidelis, & Tramontano, 2011) and (Kihara, Chen, & Yang, Quality Assessment of Protein Structure Models, 2009).

	AIDE	HOMA	MODCHECK	ModFold	ProQ	QMEAN	SVMod	TASSER-QA
Features								
Profile score					X			
Structure similarity of fragments from threading hits								X
Structure terms								
Atom clashed, van der Waals	X	X						
Bond length	X	X						
Bond angles		X						
Main chain torsion angles	X					X	X	
Residue/atom-based, contact potential			X	X	X	X	X	X
Solvation potential, burial preference of residues/atoms	X		X	X	X	X	X	
Hydrophobic residue contacts	X						X	
Compactness				X	X			
Secondary structure content	X							
Agreement of predicted secondary structure	X			X	X	X	X	
Ramachandran plot region	X							
Methods								
Single Model Vs. Consensus	S	S	S + C	S + C	S	S + C	S	C
Full atom/backbone/c-alpha	FA/BB	FA	FA	BB	FA	FA	FA	CA
Global Vs. local	G	G	G	G	GL	GL	G	G
Computational technique	Neural network combining structural parameters	Structure Analysis	Neural network combining features including independent force fields	Neural network that combines multiple scores	Neural network based on structure features	Multiple statistical potentials	SVM combining multiple assessment scores	Structure features, statistical potentials

The rest of this thesis is organized as follows: first there will be a discussion of the reliability of atomic protein models used to annotate electron microscopy data. Then, we will discuss the application of steered molecular dynamics towards the selection of near native atomic models within a diverse candidate model set. Finally, we will discuss two real applications of model fitting to electron microscopy maps.

1.3 Aims and Objectives

The overall aim of this PhD work is to characterize the uncertainty of EM derived atomic models and develop a potential approach for distinguishing reliable models.
This can be subdivided into two primary objectives.

1. Determine the deviation of different modeling methods relative to map resolution:
    1. Root Mean Square Deviation - after refinement of fitted pdb models, calculate the RMSD between the initial fitted structure and the different refinement techniques to determine relative structure variation.
    2. Cross Correlation - after refinement of fitted pdb models, calculate the cross correlation between the different structures to determine relative goodness of fit to the experimental data.
    3. Energetics and Scoring Functions - use reciprocal scoring functions to find structural improvements or overfitting.
2. Develop an approach for the detection of high-quality models:
    1. Apply in silico stresses to models of the same protein in different conformations. Use the energetics of these models to filter for quality.

1.4 References

Atilgan, A.R., Durell, S.R., Jernigan, R.L., Demirel, M.C., Keskin, O., Bahar, I. (2001). Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model. *Biophysical Journal, 80*, 505-515.

Adhikari, B., Bhattacharya, D., Cao, R., & Cheng, J. (2015). CONFOLD: residue-residue contact-guided ab initio protein folding. *Proteins: Structure, Function, and Bioinformatics*, *83*, 1436-1449.

Alishahi, M., & Kamali, R. (2019). A novel molecular dynamics study of CO 2 permeation through aquaporin-5. *The European Physical Journal E*, *42*, 1-8.

Lamb, A. L., Kappock, T. J., & Silvaggi, N. R. (2015). You are lost without a map: Navigating the sea of protein structures. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1854*, 258-268.

Baker, D. (2019). What has de novo protein design taught us about protein folding and biophysics?. *Protein Science*, *28*, 678-683.

Barton, R., Khakbaz, P., Bera, I., Klauda, J. B., Iovine, M. K., & Berger, B. W. (2016). Interplay of specific trans-and juxtamembrane interfaces in plexin A3 dimerization and signal transduction. *Biochemistry*, *55*, 4928-4938.

Bastolla, U. (2014). Computing protein dynamics from protein structure with elastic network models. *WIREs Comput Mol Sci, 4*, 488-503.

Benjin, X. & Ling, L. (2020). Developments, applications, and prospects of cryo-electron microscopy. *Protein Science*, *29*, 872-882.

Benkert, P., Tosatto, S. C., & Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function, and Bioinformatics*, *71*, 261-277.

Bernado, P., & Svergun, D. I. (2012). Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Molecular biosystems*, *8*, 151-167.

Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L., & Schwede, T. (2017). Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology. *Scientific reports*, *7*, 1-15.

Bhattacharya, A., Wunderlich, Z., Monleon, D., Tejero, R., & Montelione, G. T. (2008). Assessing model accuracy using the homology modeling automatically software. *Proteins: Structure, Function, and Bioinformatics*, *70*, 105-118.

Bienert, S., Waterhouse, A., de Beer, T. A., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2017). The SWISS-MODEL Repository—new features and functionality. *Nucleic acids research*, *45*, D313-D319.

Brooks, B. R., Brooks III, C. L., Mackerell Jr, A. D., Nilsson, L., Petrella, R. J., Roux, B., ... & Karplus, M. (2009). CHARMM: the biomolecular simulation program. *Journal of computational chemistry*, *30*, 1545-1614.

Atilgan, C., Gerek, Z. N., Ozkan, S. B., & Atilgan, A. R. (2010). Manipulation of conformational change in proteins by single-residue perturbations. *Biophysical journal*, *99*, 933-943.

Cao, R., Bhattacharya, D., Hou, J., & Cheng, J. (2016). DeepQA: improving the estimation of single protein model quality with deep belief networks. *BMC bioinformatics*, *17*, 1-9.

Lawson, C. L., Patwardhan, A., Baker, M. L., Hryc, C., Garcia, E. S., Hudson, B. P., ... & Chiu, W. (2016). EMDataBank unified data resource for 3DEM. *Nucleic acids research*, *44*, D396-D403.

Noble, C. G., & Shi, P. Y. (2012). Structural biology of dengue virus enzymes: towards rational design of therapeutics. *Antiviral research*, *96*, 115-126.

Suloway, C., Pulokas, J., Fellmann, D., Cheng, A., Guerra, F., Quispe, J., ... & Carragher, B. (2005). Automated molecular microscopy: the new Leginon system. *Journal of structural biology*, *151*, 41-60.

Ciemny, M. P., Badaczewska-Dawid, A. E., Pikuzinska, M., Kolinski, A., & Kmiecik, S. (2019). Modeling of disordered protein structures using monte carlo simulations and knowledge-based statistical force fields. *International journal of molecular sciences*, *20*, 606.

creative-biostructure.com. (n.d.). (creative biostructure) Retrieved May 5, 2021, from https://www.creative-biostructure.com/nmr-services_28.htm

Davis, A. M., Teague, S. J., & Kleywegt, G. J. (2003). Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angewandte Chemie International Edition*, *42*, 2718-2736.

De Rosier, D. J., & Klug, A. (1968). Reconstruction of three dimensional structures from electron micrographs. *Nature*, *217*, 130-134.

Doruker, P., Atilgan, A. R., & Bahar, I. (2000). Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to α-amylase inhibitor. *Proteins: Structure, Function, and Bioinformatics*, *40*, 512-524.

Adrian, M., Timmins, P. A., & Witz, J. (1992). In vitro decapsidation of turnip yellow mosaic virus investigated by cryo-electron microscopy: a model for the decapsidation of a small isometric virus. *Journal of general virology*, *73*, 2079-2083.

Emwas, A. H. M. (2015). The strengths and weaknesses of NMR spectroscopy and mass spectrometry with particular focus on metabolomics research. In *Metabonomics* (pp. 161-193). Humana Press, New York, NY.

Eramian, D., Shen, M. Y., Devos, D., Melo, F., Sali, A., & Marti-Renom, M. A. (2006). A composite score for predicting errors in protein structure models. *Protein science*, *15*, 1653-1666.

Eswar, N., Eramian, D., Webb, B., Shen, M. Y., & Sali, A. (2008). Protein structure modeling with MODELLER. In *Structural proteomics* (pp. 145-159). Humana Press.

Evans, P., & McCoy, A. (2008). An introduction to molecular replacement. *Acta Crystallographica Section D: Biological Crystallography*, *64*, 1-10.

Frank, J. (1975). Averaging of low exposure electron micrographs of non-periodic objects. In *Single-Particle Cryo-Electron Microscopy: The Path Toward Atomic Resolution: Selected Papers of Joachim Frank with Commentaries* 69-72.

Franke, D., & Svergun, D. I. (2009). DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *Journal of applied crystallography*, *42*, 342-346.

Ban, F., Dalal, K., Li, H., LeBlanc, E., Rennie, P. S., & Cherkasov, A. (2017). Best practices of computer-aided drug discovery: lessons learned from the development of a preclinical candidate for prostate cancer with a new mechanism of action. *Journal of chemical information and modeling*, *57*, 1018-1028.

Montelione, G. T., Zheng, D., Huang, Y. J., Gunsalus, K. C., & Szyperski, T. (2000). Protein NMR spectroscopy in structural genomics. *nature structural biology*, *7*, 982-985.

Liu, G., Shen, Y., Atreya, H. S., Parish, D., Shao, Y., Sukumaran, D. K., Xiao, R., Yee, A., Lemak, A., Bhattacharya, A., Acton, T. A., Arrowsmith, C. H., Montelione, G. T. & Szyperski, T. (2005). NMR data collection and analysis protocol for high-throughput protein structure determination. *Proceedings of the National Academy of Sciences*, *102*, 10487-10492.

Williams, G. J., Williams, R. S., Williams, J. S., Moncalian, G., Arvai, A. S., Limbo, O.,

Guenther, G., SilDas, S., Hammel, M., Russell, P. & Tainer, J. A. (2011). ABC ATPase signature helices in Rad50 link nucleotide state to Mre11 interface for DNA repair. *Nature structural & molecular biology*, *18*, 423-431.

Grant, T. D. (2018). Ab initio electron density determination directly from solution scattering data. *Nature methods*, *15*, 191-193.

Tang, G., Peng, L., Baldwin, P. R., Mann, D. S., Jiang, W., Rees, I., & Ludtke, S. J. (2007). EMAN2: an extensible image processing suite for electron microscopy. *Journal of structural biology*, *157*, 38-46.

Guex, N., Peitsch, M. C., & Schwede, T. (2009). Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: A historical perspective. *Electrophoresis*, *30*, S162-S173.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, *28*, 235-242.

Zhao, H., Fernandez, E., Dowd, K. A., Speer, S. D., Platt, D. J., Gorman, M. J., Govero, J., Nelson, C. Pierson, T. C., Diamond, M. S. & Fremont, D. H. (2016). Structural basis of Zika virus-specific antibody protection. *Cell*, *166*, 1016-1027.

Hedger, G., Shorthouse, D., Koldsø, H., & Sansom, M. S. (2016). Free energy landscape of lipid interactions with regulatory binding sites on the transmembrane domain of the EGF receptor. *The Journal of Physical Chemistry B*, *120*, 8154-8163.

Flechsig, H., & Mikhailov, A. S. (2010). Tracing entire operation cycles of molecular motor hepatitis C virus helicase in structurally resolved dynamical simulations. *Proceedings of the National Academy of Sciences*, *107*, 20875-20880.

Hughes, M. L., & Dougan, L. (2016). The physics of pulling polyproteins: a review of single molecule force spectroscopy using the AFM to study protein unfolding. *Reports on Progress in Physics*, *79*, 076601.

Izrailev, S., Stepaniants, S., Isralewitz, B., Kosztin, D., Lu, H., Molnar, F. & Schulten, K. (1999). Steered molecular dynamics. In *Computational molecular dynamics: challenges, methods, ideas* (pp. 39-65). Springer, Berlin, Heidelberg.

Kurcinski, M., Oleniecki, T., Ciemny, M. P., Kuriata, A., Kolinski, A., & Kmiecik, S. (2019). CABS-flex standalone: a simulation environment for fast modeling of protein flexibility. *Bioinformatics*, *35*, 694-695.

Lu, H., Isralewitz, B., Krammer, A., Vogel, V., & Schulten, K. (1998). Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophysical journal*, *75*, 662-671.

Ishima, R., & Torchia, D. A. (2000). Protein dynamics from NMR. *Nature structural biology*, *7*, 740-743.

Bahar, I., Atilgan, A. R., & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding and Design*, *2*, 173-181

Phillips, J. C., Zheng, G., Kumar, S., & Kalé, L. V. (2002, November). NAMD: Biomolecular simulation on thousands of processors. In *SC'02: Proceedings of the 2002 ACM/IEEE Conference on Supercomputing,* 36-36.

Dubochet, J., Chang, J. J., Freeman, R., Lepault, J., & McDowall, A. W. (1982). Frozen aqueous suspensions. *Ultramicroscopy*, *10*, 55-61.

Källberg, M., Wang, H., Wang, S., Peng, J., Wang, Z., Lu, H., & Xu, J. (2012). Template-based protein structure modeling using the RaptorX web server. *Nature protocols*, *7*, 1511-1522.

Chen, H., & Kihara, D. (2008). Estimating quality of template-based protein models by alignment stability. *Proteins: Structure, Function, and Bioinformatics*, *71*, 1255-1274.

Kihara, D., Chen, H., & Yang, Y. D. (2009). Quality assessment of protein structure models. *Current Protein and Peptide Science*, *10*, 216-228.

Kim, Y., Babnigg, G., Jedrzejczak, R., Eschenfeldt, W. H., Li, H., Maltseva, N., Hatzos-Skintges, C., Gu, M., Makowska-Grzyska, M., Wu, R., An, H., Chhor, G. & Joachimiak, A. (2011). High-throughput protein purification and quality assessment for crystallization. *Methods*, *55*, 12-28.

Kim, H., & Kihara, D. (2014). Detecting local residue environment similarity for recognizing near-native structure models. *Proteins: Structure, Function, and Bioinformatics*, *82*, 3255-3272.

Koliński, A., & Bujnicki, J. M. (2005). Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. *Proteins: Structure, Function, and Bioinformatics*, *61*, 84-90.

Kryshtafovych, A., Fidelis, K., & Tramontano, A. (2011). Evaluation of model quality predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics*, *79*, 91-106.

Lamb, A. L., Kappock, T. J., & Silvaggi, N. R. (2015). You are lost without a map: Navigating the sea of protein structures. *Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics*, *1854*, 258-268.

Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B., & Schulten, K. (2009). Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods*, *49*, 174-180.

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., & Marrink, S. J. (2008). The MARTINI coarse-grained force field: extension to proteins. *Journal of chemical theory and computation*, *4*, 819-834.

Adrian, M., Dubochet, J., Lepault, J., & McDowall, A. W. (1984). Cryo-electron microscopy of viruses. *Nature*, *308*, 32-36.

Ma, J., Peng, J., Wang, S., & Xu, J. (2012). A conditional neural fields model for protein threading. *Bioinformatics*, *28*, i59-i66.

Ma, J., Wang, S., Wang, Z., & Xu, J. (2015). Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics*, *31*, 3506-3513

Düttmann, M., Mittnenzweig, M., Togashi, Y., Yanagida, T., & Mikhailov, A. S. (2012). Complex intramolecular mechanics of G-actin—an elastic network study.

Marrink, S. J., & Tieleman, D. P. (2013). Perspective on the Martini model. *Chemical Society Reviews*, *42*, 6801-6822.

Kurcinski, M., Kolinski, A., & Kmiecik, S. (2014). Mechanism of folding and binding of an intrinsically disordered protein as revealed by ab initio simulations. *Journal of Chemical Theory and Computation*, *10*, 2224-2231.

Kurcinski, M., Oleniecki, T., Ciemny, M. P., Kuriata, A., Kolinski, A., & Kmiecik, S. (2019). CABS-flex standalone: a simulation environment for fast modeling of protein flexibility. *Bioinformatics*, *35*, 694-695.

McGuffin, L. J. (2007). Benchmarking consensus model quality assessment for protein fold recognition. *BMC bioinformatics*, *8*, 1-15.

McPherson, A., & Cudney, B. (2014). Optimization of crystallization conditions for biological macromolecules. *Acta Crystallographica Section F: Structural Biology Communications*, *70*, 1445-1467.

Mereghetti, P., Ganadu, M. L., Papaleo, E., Fantucci, P., & De Gioia, L. (2008). Validation of protein models by a neural network approach. *BMC bioinformatics*, *9*, 1-11.

Gryk, M. R., Vyas, J., & Maciejewski, M. W. (2010). Biomolecular NMR data analysis. *Progress in nuclear magnetic resonance spectroscopy*, *56*, 329.

Gao, M., Craig, D., Vogel, V., & Schulten, K. (2002). Identifying unfolding intermediates of FN-III10 by steered molecular dynamics. *Journal of molecular biology*, *323*, 939-950.

Noble, A. J., Dandey, V. P., Wei, H., Brasch, J., Chase, J., Acharya, P., Tan, Y. Z., Zhang, Z., Kim, L. Y., Scapin, G., Rapp, M., Eng, E. T., Rice, W. J., Cheng, A., Negro, C. J., Shapiro, L., Kwong, P. D., Jeruzalmi, D., Georges, A. D., Potter, C. S. & Carragher, B. (2018). Routine single particle CryoEM sample and grid characterization by tomography. *Elife*, *7*, e34257.

Kar, P., Gopal, S. M., Cheng, Y. M., Predeus, A., & Feig, M. (2013). PRIMO: a transferable coarse-grained force field for proteins. *Journal of chemical theory and computation*, *9*, 3769-3788.

Pauw, B. R. (2013). Everything SAXS: small-angle scattering pattern collection and correction. *Journal of Physics: Condensed Matter*, *25*, 383201.

Peng, J., & Xu, J. (2009, May). Boosting protein threading accuracy. In *Annual International Conference on Research in Computational Molecular Biology* (pp. 31-45). Springer, Berlin, Heidelberg.

Peng, J., & Xu, J. (2010). Low-homology protein threading. *Bioinformatics*, *26*, i294-i300.

Peng, J., & Xu, J. (2011). A multiple-template approach to protein threading. *Proteins: Structure, Function, and Bioinformatics*, *79*, 1930-1939.

Peng, J., & Xu, J. (2011). RaptorX: exploiting structure information for protein alignment by statistical inference. *Proteins: Structure, Function, and Bioinformatics*, *79*, 161-171.

Peterson, L. X., Kang, X., & Kihara, D. (2014). Assessment of protein side-chain conformation prediction methods in different residue environments. *Proteins: Structure, Function, and Bioinformatics*, *82*, 1971-1984.

Putnam, C. D., Hammel, M., Hura, G. L., & Tainer, J. A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Quarterly reviews of biophysics*, *40*, 191-285.

Raja, P., & Barron, A. (2021, March 21). *NMR Spectroscopy*. Retrieved May 20, 2021, from https://chem.libretexts.org/@go/page/55887

Rajagopal, N., & Nangia, S. (2019). Obtaining Protein Association Energy Landscape for Integral Membrane Proteins. *Journal of chemical theory and computation*, *15*, 6444-6455.

Receveur-Bréchot, V., & Durand, D. (2012). How random are intrinsically disordered proteins? A small angle scattering perspective. *Current Protein and Peptide Science*, *13*, 55-75.

Reish, M. S. (2015, June 19). *Chemical & Engineering News*. Retrieved May 10, 2021, from https://cen.acs.org/articles/93/i26/NMR-Instrument-Price-Hikes-Spook.html

Roberts, G. C. (2013). *Protein NMR – Introduction. In: Roberts G.C.K. (eds) Encyclopedia of Biophysics.* Berlin, Heidelberg: Springer.

Roy, A., Kucukural, A., & Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nature protocols*, *5*, 725-738.

Sadowski, M. I., & Jones, D. T. (2007). Benchmarking template selection and model quality assessment for high-resolution comparative modeling. *Proteins: Structure, Function, and Bioinformatics*, *69*(3), 476-485.

Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, *234*, 779-815.

Ayuso-Tejedor, S., García-Fandiño, R., Orozco, M., Sancho, J., & Bernadó, P. (2011). Structural analysis of an equilibrium folding intermediate in the apoflavodoxin native ensemble by small-angle X-ray scattering. *Journal of molecular biology*, *406*, 604-619.

Ilari, A., & Savino, C. (2008). Protein structure determination by x-ray crystallography. *Bioinformatics*, *452*, 63-87.

Scheres, S. H. (2012). RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology*, *180*, 519-530.

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., & Kolinski, A. (2016). Coarse-grained protein models and their applications. *Chemical reviews*, *116*, 7898-7936.

Shen, M. Y., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein science*, *15*, 2507-2524.

Shin, W. H., Kang, X., Zhang, J., & Kihara, D. (2017). Prediction of local quality of protein structure models considering spatial neighbors in graphical models. *Scientific reports*, *7*, 1-14.

Small Angle X-ray Scattering/Diffraction. (n.d.). (Stanford Synchrotron Radiatioon Lightsource) Retrieved May 17, 2021, from https://www-ssrl.slac.stanford.edu/smb-saxs/content/data-analysis-primer

Smyth, M. S., & Martin, J. H. (2000). X Ray Crystallography. *Mol Pathol., 53*, 8-14.

Studer, G., Rempfer, C., Waterhouse, A. M., Gumienny, R., Haas, J., & Schwede, T. (2020). QMEANDisCo—distance constraints applied on model quality estimation. *Bioinformatics*, *36*, 1765-1771.

Svergun, D. I. (1999). Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophysical journal*, *76*, 2879-2886.

Rambo, R. P., & Tainer, J. A. (2011). Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers*, *95*, 559-571.

Terashi, G., & Kihara, D. (2018). Protein structure model refinement in CASP12 using short and long molecular dynamics simulations in implicit solvent. *Proteins: Structure, Function, and Bioinformatics*, *86*, 189-201.

Tirion, M. M. (1996). Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Physical review letters*, *77*, 1905.

Togashi, Y., & Flechsig, H. (2018). Coarse-grained protein dynamics studies using elastic network models. *International journal of molecular sciences*, *19*, 3899.

Schwede, T., Kopp, J., Guex, N., & Peitsch, M. C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic acids research*, *31*, 3381-3385.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. (2005). GROMACS: fast, flexible, and free. *Journal of computational chemistry*, *26*, 1701-1718.

Wallner, B., & Elofsson, A. (2007). Prediction of global and local model quality in CASP7 using Pcons and ProQ. *Proteins: Structure, Function, and Bioinformatics*, *69*, 184-193.

Wang, G., & Dunbrack Jr, R. L. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, *19*, 1589-1591.

Wang, H. W., & Wang, J. W. (2017). How cryo-electron microscopy and X-ray crystallography complement each other. *Protein Science*, *26*, 32-39.

Wang, Z., Zhao, F., Peng, J., & Xu, J. (2010, December). Protein 8-class secondary structure prediction using conditional neural fields. In *2010 IEEE international conference on bioinformatics and biomedicine* 109-114

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., Beer, T., Rempfer, C., Bordoli, L., Lepore, R., & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research*, *46*, W296-W303.

Webb, B., & Sali, A. (2016). Comparative protein structure modeling using MODELLER. *Current protocols in bioinformatics*, *54*, 5-6.

Wüthrich, K. (1990). Protein structure determination in solution by NMR spectroscopy. *Journal of Biological Chemistry*, *265*, 22059-22062.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature methods*, *12*, 7-8.

Yang, J., & Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. *Nucleic acids research*, *43*, W174-W181.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., & Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. *Nature methods*, *12*, 7-8.

Yang, S. (2014). Methods for SAXS-based structure determination of biomolecular complexes. *Advanced materials*, *26*, 7902-7910.

Yang, Y. D., Spratt, P., Chen, H., Park, C., & Kihara, D. (2010). Sub-AQUA: real-value quality assessment of protein structure models. *Protein Engineering, Design & Selection*, *23*, 617-632.

Yin, C. C. (2018). Structural biology revolution led by technical breakthroughs in cryo-electron microscopy. *Chinese Physics B*, *27*, 058703.

Yip, K. M., Fischer, N., Paknia, E., Chari, A., & Stark, H. (2020). Atomic-resolution protein structure determination by cryo-EM. *Nature*, *587*, 157-161.

Zgorski, A., Pastor, R. W., & Lyman, E. (2019). Surface shear viscosity and interleaflet friction from nonequilibrium simulations of lipid bilayers. *Journal of chemical theory and computation*, *15*, 6471-6481.

Zhang, S., Fu, L., Wan, M., Song, J., Gao, L., & Fang, W. (2019). Peripheral antimicrobial peptide gomesin induces membrane protrusion, folding, and laceration. *Langmuir*, *35*, 13233-13242.

Zhao, F., Peng, J., & Xu, J. (2010). Fragment-free approach to protein folding using conditional neural fields. *Bioinformatics*, *26*, i310-i317.

Zhao, F., Peng, J., Debartolo, J., Freed, K. F., Sosnick, T. R., & Xu, J. (2010). A probabilistic and continuous model of protein conformational space for template-free modeling. *Journal of Computational Biology*, *17*, 783-798.

Zhou, H., & Skolnick, J. (2008). Protein model quality assessment prediction by combining fragment comparisons and a consensus Cα contact potential. *PROTEINS: Structure, Function, and Bioinformatics*, *71*, 1211-1218.

Zhou, H., & Skolnick, J. (2011). GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophysical journal*, *101*, 2043-2052.

CHAPTER 2. VARIABILITY OF PROTEIN STRUCTURE MODELS FROM ELECTRON MICROSCOPY

Chapter reproduced from a manuscript previously published in the journal *Structure* in 2017

2.1 Abstract

An increasing number of biomolecular structures are solved by electron microscopy (EM). However, the quality of structure models determined from EM maps vary substantially. To understand to what extent structure models are supported by information embedded in EM maps, we used two computational structure refinement methods to examine how much structures can be refined using a dataset of 49 maps with accompanying structure models. The extent of structure modification as well as the disagreement between refinement models produced by the two computational methods scaled inversely with the global and the local map resolutions. A general quantitative estimation of deviations of structures for particular map resolutions are provided. Our results indicate that the observed discrepancy between the deposited map and the refined models is due to the lack of structural information present in EM maps and thus these annotations must be used with caution for further applications.

2.2 Introduction

Electron microscopy, particularly, cryo-electron microscopy (cryo-EM) is an emerging technique in structural biology for determining 3D structures of large biological macromolecules. Its notable advantage of solving large macromolecular assemblies is complementary to conventional structural biology techniques, such as X-ray crystallography and nuclear magnetic resonance, bridging atomic-detailed structures of molecules with a higher level of structure information of molecular machinery and interactions in a cell. Recent technical breakthroughs in cryo-EM, both in its hardware (Faruqi and Henderson, 2007) and software (Scheres, 2012), have enabled determining 3D structures to nearly atomic-level resolutions (Kuhlbrandt, 2014a, 2014b), which has further attracted biologists to apply this new technology to their biological systems. Cryo-EM, together with other types of EM, has now become a key technique in structural biology. The number of structures solved by EM is increasing rapidly, resulting in over 3,600 EM density

maps deposited and stored in the EM Data Bank (EMDB) (Velankar et al., 2016), the primary repository of EM density maps.

Although near-atomic resolution structures have been reported frequently in recent years, about 90% of released maps are solved at a resolution of 5 A˚ or less (Lawson et al., 2011). Typically, a structure model is built partly manually with assistance from some computational structure building methods (Esquivel-Rodriguez and Kihara, 2013) and visualization software (Pettersen et al., 2004). Homology models are commonly used in this process if a homologous protein structure is available as a template for modeling (Yang et al., 2012; Zhu et al., 2010). Various types of structure building methods exist, which range from rigid-body docking (structure fitting) (Ceulemans and Russell, 2004; Esquivel-Rodriguez and Kihara, 2012; Rossmann, 2000; Woetzel et al., 2012), local structure identification (Dror et al., 2007; Jiang et al., 2001), to flexible fitting (McGreevy et al., 2014; Wang et al., 2015), to meet the needs of different situations of modeling. In the EMDB, even EM maps of a low-resolution are often accompanied with structure models. Structure models are fit in EM maps determined at a resolution of as low as 40 A˚ . Even among EM maps of 20 A˚ or less, models were built for 6.4% of the maps. Structure models built from EM maps were used as the basis of discussion on mechanisms of biological functions of the macromolecules, often without careful consideration about the extent to which the modeled structure is supported by the EM maps. However, the quality of structure models would vary substantially depending on the global and local resolution of EM maps as well as the methods used for building structure models. Currently, in the EMDB, the fit of a structure model to an EMDB map is validated through visual inspection, and the stereochemistry of a model deposited in the PDB is checked in the same way as regular PDB entries; however, standards of validation are still yet to be developed (Lawson et al., 2011). The lack of a standardized method for assessing model quality brings the reliability of some structure models into question.

In this work, to investigate to what extent structure models are supported by the electron density of the EM maps, we used two structure modeling/refinement methods to determine if structure models had an increased fit to their map after refinement. More concretely, for 49 EM maps of a wide range of resolutions, we ran two structure fitting/refinement protocols, Molecular Dynamics Flexible Fitting (MDFF) (McGreevy et al., 2014, 2016) and Rosetta (Wang et al., 2015), starting from the deposited structure models and observed changes of cross-correlation of the models to the EM maps and the energy of those structures. These two programs were chosen

because they are among the most popular structure modeling and refinement tools used for protein structure determination of EM maps. They are capable of refining structures based on sound physical principles. The energy of the structures was tracked throughout the refinement to ensure that the structure was not undergoing undue stress to over-fitting to the density map. In cases where the energy of the structure decreased or stayed the same while the cross-correlation increased, the refined structure was considered to be better supported than the original annotation. We observed that, for over 60% of the cases tested, structures were further refined from the deposited structures. The extent of structure modification scaled inversely with the global and the local resolutions of the maps that these structures annotate. That is, the refinement protocols did not move structures much if they were derived from high-resolution EM maps, while large movements were observed for the models of low-resolution EM maps. More quantitatively, the extent of the structure change in terms of the Ca root-mean-square deviation (RMSD) of models after the refinement was roughly about 30% or the map resolution. We also observed that the refined models by the two refinement protocols do not agree for those with low-resolution maps. The amount of discrepancy of the models also inversely correlated with the resolution of their maps. These indicate that the reliability of a structure model from an EM map critically depends on the resolution of the map and the reliability substantially decreases as the resolution of a map decreases. We conclude therefore that, unless the models were derived from high-resolution maps, it is critical that structure models are used with caution for further analysis and discussion. It has been discussed that EM maps at lower resolutions contain less structure information (Henderson et al., 2012). Different types of computational structure modeling methods have been developed to address the different levels of structure information contained in maps at various resolutions (Esquivel-Rodriguez and Kihara, 2013; Villa and Lasker, 2014; Wriggers and Chacon, 2001). In developing a modeling tool, it is common to test the tool on maps at different resolutions (Jolley et al., 2008; Singharoy et al., 2016). Concerning bias and misinterpretation of densities of EM maps, validation methods have been developed for checking each step of the structure modeling process from maps (Falkner and Schroder, 2013; Rosenthal and Rubinstein, 2015). Egelman (2008) discussed problems and potential errors in structure modeling with EM maps, focusing on cases of high-resolution structure fitting to maps. Although the problem of the model reliability from EM maps has been known, each of the earlier works addressed this problem on a small number of particular proteins. In contrast to the earlier recognition of this problem, the current work shows a comprehensive and

general view on the reliability of structure models using a dataset of EM maps determined at a wide range of resolutions. Furthermore, the current work provides a general quantitative estimation of deviations of structures for particular map resolutions. Given the current situation where an increasing number of protein structures are being solved by EM and rapidly accumulated and reused, it is crucial that the reliability of structure models in the EMDB is well and widely understood.

2.3 Results

For a dataset of 49 EMDB entries (Table 1), structure models modified by MDFF or Rosetta are examined relative to the deposited structure models in the EMDB along with the EM maps, which were the starting structures of the refinement protocols. The 49 maps were selected from 688 maps available in the EMDB that were associated with fitted protein structures in the PDB. Maps were removed from the initial pool if the fitted structures did not have sufficient overlap and cross-correlation. The table lists a dataset of 49 EM maps from EMDB and associated protein structure models used in this study. EMDB ID, name of the complex, map resolution, molecular mass calculated from the PDB file, PDB ID of the structure model that was fit to the map, the author-recommended contour level, and the cross-correlation between the maps and their models are provided to the map, which indicated that maps have a large empty region or did not contain much structure information and refinement was not expected to work properly. The data selection procedure is detailed in the STAR Methods. First we will discuss overall changes of the models in terms of the energy of the structure and the cross-correlation to the EM density map. Subsequently, the amount of structural deviations is discussed relative to the global and local map resolutions. Finally, some illustrative examples are presented. The overall flowchart of the analyses performed in this work is illustrated in Figure S1.

Table 2.1: The EM map dataset. The table lists a dataset of 49 EM maps from EMDB and associated protein structure models used in this study.

Map ID	Complex Name	Resolution	Mass (MDa)	PDB ID	Contour	Cross Correlation
1046	GroES-ADP7-GroEL-ATP7 from E.coli	23.5	0.678	1gru	0.029	0.965
1047	GroEL-ATP from E.coli	14.9	0.621	2c7e	0.084	0.896
1149	Recombinant protein Acr1 From M.Tuberculosis	16.5	0.109	2byu	0.012	0.8
1180	GroEL-ATP7-GroES	7.7	0.674	2c7c	0.608	0.842
1181	GroEL-ADP7-GroES	8.7	0.674	2c7d	1.9	0.849
1202	GroEL-ADP-gp31	8.2	0.666	2cgt	0.452	0.831
1494	Saf pilus	17	0.024	3cre	2.95	0.768
1495	Saf pilus	17	0.024	3crf	1.59	0.915
1505	DegP dodecamer	28	0.43	2zle	0.013	0.755
1654	RbcL8-X8	17	0.488	2wvw	3	0.757
1655	Rubisco assembly intermediate	9	0.488	2wvw	5.6	0.84
1871	Needle complex from Salmonella typhimurium	8.3	0.189	2y9k	0.06	0.717
1894	Human alphaB crystallin	9.4	0.387	2ygd	0.004	0.764
1932	R. sphaeroides CbbX	21	0.153	3zuh	0.005	0.794
1940	Nicotiana tabacum Rubisco Activase (R294V)	20	0.144	3zw6	0.02	0.721
1960	Bovine TRiC/CCT in the nucleotide-free (apo) state	10.5	0.65	4a0o	1.13	0.845
1961	bovine TRiC/CCT in the AMP-PNP state	10.7	0.682	4a0v	1.13	0.89
1962	bovine TRiC/CCT	13.9	0.693	4a0w	1	0.921
1963	bovine TRiC/CCT in the ADP state	11.3	0.686	4a13	1.08	0.873
2001	GroEL-ATP14 Rd1-Rd3	8.5	0.618	4aau	0.2	0.7
2003	GroEL-ATP14 Rd5-Rdopen	8.5	0.618	4ab3	0.2	0.77
2325	GroEL variant EL43Py capped by GroES	8.9	0.678	3zpz	1.3	0.93
2327	Non-native RuBisCO substrate protein encapsulated in GroEL cavity	15.9	0.678	3zq1	1	0.91
2365	Bacteriophage MS2 bound to its receptor, the E. coli F-pilus	39	1.987	4bp7	1.45	0.815
2526	MloK1 with cAMP	7	0.118	4chv	2.3	0.964
2548	Fv antibody domain bound beta-galactosidase	13	0.45	4ckd	0.13	0.946
2807	rabbit RyR1 in complex with its modulator FKBP12	3.8	1.177	3j8h	0.04	0.73
2856	Dynactin complex from pig brain	4	0.429	5adx	0.088	0.808
2924	Recombinant human APC/C-Cdh1-Emi1 ternary complex	3.6	0.779	4ui9	0.07	0.867
2984	E. coli beta-galactosidase	2.2	0.371	5a1a	0.05	0.456
5169	E. coli RNA polymerase	11	0.289	3lu0	11	0.845
5186	human apoptosome with bound procaspase-9 CARD	9.5	0.798	3j2t	1.3	0.869
5258	Lidless D386A Mm-cpn variant	8	0.674	3j02	0.02	0.836

Table 2.1 continued

Map ID	Complex Name	Resolution	Mass (MDa)	PDB ID	Contour	Cross Correlation
5395	rATcpn-beta in apo state	8.3	0.794	3j1e	5	0.795
5450	YiiP from Shewanella oneidensis in DOPG lipids	13	0.052	3j 1z	3.33	0.91
5466	Enterovirus 71 empty capsid	9.2	3.719	3j23	1	0.834
5607	MecA-ClpC(E280A)	9	0.477	3j3t	1.5	0.866
5608	MecA-ClpC(E618A)	11	0.477	3j3s	1.5	0.866
5609	MecA-ClpC (E280A,E618A) with ATP	10	0.477	3j3u	1.5	0.857
5610	MecA-ClpC(E280A,E618A)with ADP	9.4	0.477	3j3r	1.5	0.878
5679	Aquaporin-0 bound to Calmodulin	25	0.107	3j41	4.96	0.798
5776	Rat TRPV1 in complex with DkTx and resiniferatoxin	3.8	0.217	3j5q	7	0.678
5925	MAVS filament	3.64	0.073	3j6j	0.3	0.868
5995	Escherichia coli beta-galactosidase	3.2	0.371	3j7h	0.0224	0.764
6272	13-fold average of VP6 trimer from full rotavirus reconstruction	2.6	0.108	3j9s	0.0198	0.684
6337	Vesicular Stomatitis Virus L-Protein	3.8	0.183	5a22	1.2	0.707
6344	Zebra fish alpha-1 glycine receptor bound with strychnine	3.9	0.158	3jad	7	0.693
6345	Zebra fish alpha-1 glycine receptor bound with glycine	3.9	0.158	3jae	6.5	0.7
6346	Zebra fish alpha-1 glycine receptor bound with glycine/ivermectin	3.8	0.16	3jaf	7	0.714

2.3.1 Changes in Energy and Cross-Correlation

Figure 1 shows the change observed in the energy and crosscorrelation to their EM maps of structure models before and after the refinement protocols. The energy change of a structure, namely, the CHARMM potential energy for MDFF and the Rosetta free energy for the Rosetta protocol, indicates how well the structure is refined without considering its fit to the EM map. The difference of the CHARMM potential energy for MDFF was computed against the initial structure at the start of the refinement, i.e., after the initial energy minimization is applied, which removes atomic clashes, and the temperature is raised to 300 K. On the other hand, cross-correlation of a structure to its EM map indicates how well the electron density of the protein structure model agrees with the EM electron density map. If the cross-correlation increases after refinement, it is an indication that the refined structure is globally in better agreement with the map data. For MDFF, results for refinements for different g-scale values, 0.1, 0.3, 0.5, and 0.7 are shown (Figure 1A). The g-scale controls how much the fit to the EM map contributes to the overall potential function

used in MDFF (see STAR Methods). Refinements with large g-scale values cause their structure to fit more to their map, and the opposite with a small g-scale value.

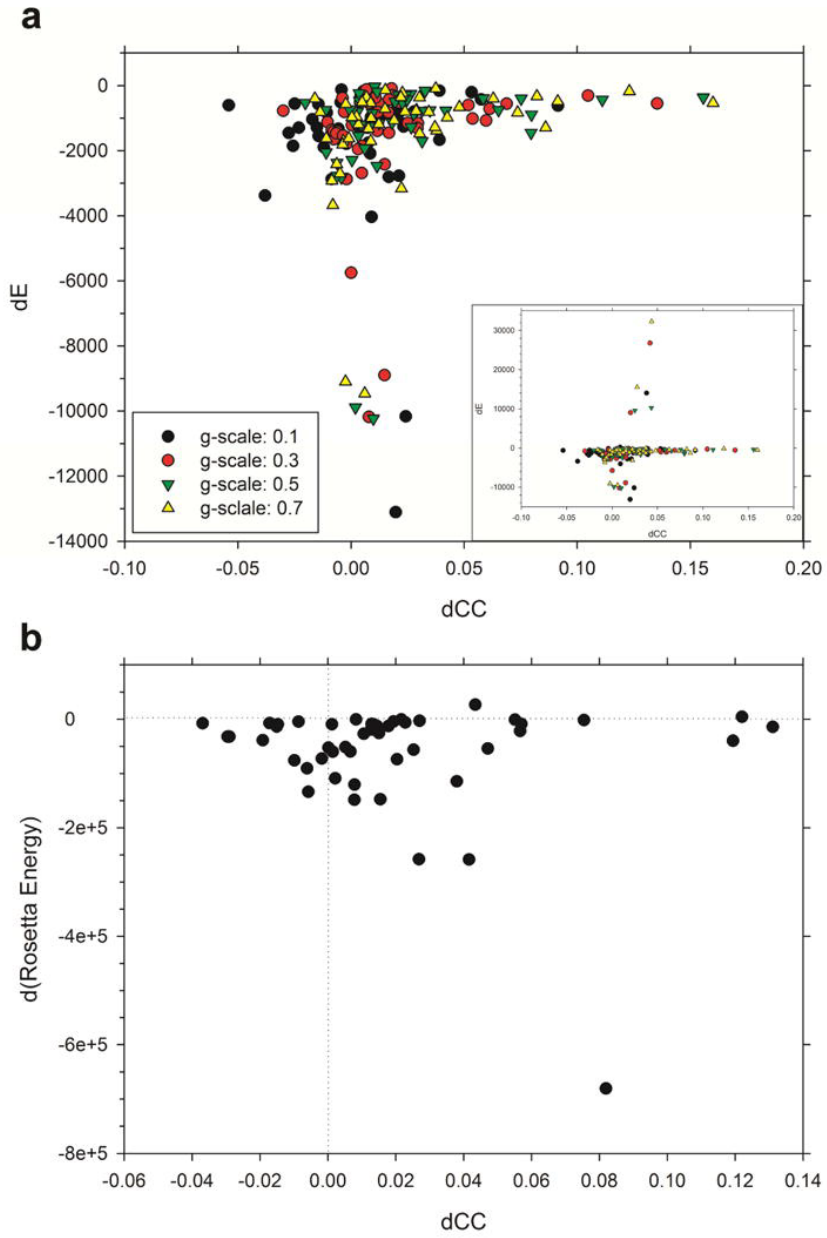

Figure 2.1: Change in the potential energy and cross-correlation to EM maps of the refined protein structure models. The two values were computed after the refinement in comparison with the initial structure. dCC, the difference of the cross-correlation; dE, difference of the energy of structure models. (a), Results using MDFF with four different g-scale values, 0.1 (filled circles), 0.3 (red circles), 0.5 (green triangles), 0.7 (yellow triangles). The energy was evaluated with the CHARMM potential energy used in MDFF, excluding the map fitness term. The figure shows results for 47 EM maps excluding two virus capsids, EMD-2365 and EMD-5466, which showed exceptionally large positive dE (see text). The inset figure includes al 49 maps. (b), results using Rosetta. The Rosetta free energy was used.

MDFF results (Figure 1A) show that the energy of almost all the models were lowered after the refinement. There were two exceptions, both virus capsid structures, EMD-2365 and EMD-5466, which showed exceptionally large positive energy after running MDFF (Figure 1A, inset). These two EM maps were solved at a relatively low-resolution, 39 and 9.2 A˚ , respectively, and the associated crystal structures, which are slightly larger than the EM maps, were fit to the maps by rigid-body fitting. Therefore, the structures were compressed by the MDFF runs to better fit into the maps, which caused a small positive energy for many atoms in the capsids that accumulated into a large positive energy. The results are summarized in Table S1 by classifying the cases into four categories, cases where the cross-correlation increased by sacrificing potential energy by MDFF (i.e., an increase of cross-correlation, i.e., dCC > 0, with an increase of the energy, dE > 0), cases where both energy and cross-correlation became worse (i.e., dE > 0 and dCC < 0), cases where an energy decrease was accompanied with an increase of cross-correlation (i.e., dE < 0 and dCC > 0), and cases where an energy decrease occurred by reducing the global fit to the EM map (i.e., dE < 0 and dCC < 0). With regard to cross-correlation, an increase is seen for more than 60% of the proteins after their refinement for all the g-scale values used. Naturally, in general an increase of cross-correlation is observed when a large g-scale value is used. When the g scale was set to 0.1, cross-correlation increased for 32 cases, while such cases increased to 42 and 37 for g scales of both 0.5 and 0.7. Turning our attention to the change of potential energy of structures, lowering of energy (dE < 0) was observed for all but two of the cases, EMD-2365 and EMD-5466. Lowering of energy was accompanied by the decrease cross-correlation for 17 cases with a g scale of 0.1, which decreased as a larger g scale was used. Finally, both lowering energy (dE < 0) and increase of cross-correlation (dCC > 0) was observed for more than 60% of the cases, for 61.2% (30/49) with a g scale of 0.1, and the largest at 81.6% (40/49) with a g scale of 0.5. The results are essentially the same with the Rosetta refinement protocol; improvement in both Rosetta free energy and cross-correlation was observed Figure 1. Change in the Potential Energy and Cross-Correlation to EM Maps of the Refined Protein Structure Models The two values were computed after the refinement in comparison with the initial structure. dCC, the difference of the cross-correlation; dE, difference of the energy of structure models. (A) Results using MDFF with four different g-scale values, 0.1 (filled circles), 0.3 (red circles), 0.5 (green triangles), and 0.7 (yellow triangles). The energy was evaluated with the CHARMM potential energy used in MDFF, excluding the map fitness term. The figure shows results for 47 EM maps excluding two virus

capsids, EMDB: 2365 and EMDB: 5466, which showed exceptionally large positive dE values (see text). The inset figure includes all 49 maps. (B) Results using Rosetta. The Rosetta free energy was used for 69.4% (34/49) of the cases (Figure 1B). In both Figures 1A and 1B, dCC did not show clear correlation to the change of the structure energy, mainly because each data point in the plots is from a different protein structure. Models that underwent the MDFF protocol were cross-evaluated with the Rosetta free energy in Table S2 and vice versa in Table S3. As shown in Table S2, for the majority of the cases the MDFF protocol also lowered their Rosetta energy. Similarly, for almost all of the cases, the Rosetta protocol also lowered the CHARMM energy (Table S3).

In addition to the CHARMM potential energy and the Rosetta free energy, we further examined structural change by the refinements with the MolProbity Score (MPScore) (Chen et al., 2010) (Figure 2 and Table 2). MPScore is a structure validation score used to evaluate the quality of a protein structure solved by experiments typically before submission to public databases. MPScore considers atom clashes, outliers of rotamers, and main-chain dihedral angles, and exhibits a high value if a structure has many such unfavorable features. Thus, if the local quality of a structure was improved by a refinement protocol, its MPScore decreases. We used MPScore as an additional independent structure evaluation, which was not the target of optimization by the two refinement methods. When MDFF was used, the MPScore improved (i.e., decreased) for almost all the models, except for a couple of cases, regardless of the g-scale value used (Figure 2A and Table 2). Of the structures, 61.2% (30/49 with 0.1 g scale) to 81.6% (40/49 with a 0.5 g scale) showed improvement in both MPScore and cross-correlation. In the case of Rosetta, an MPScore improvement was observed for 36 structures, among which 25 were associated with an increase of cross-correlation (Figure 2B). With the Rosetta protocol, more models (13) had a deterioration in their MPScores than with MDFF.

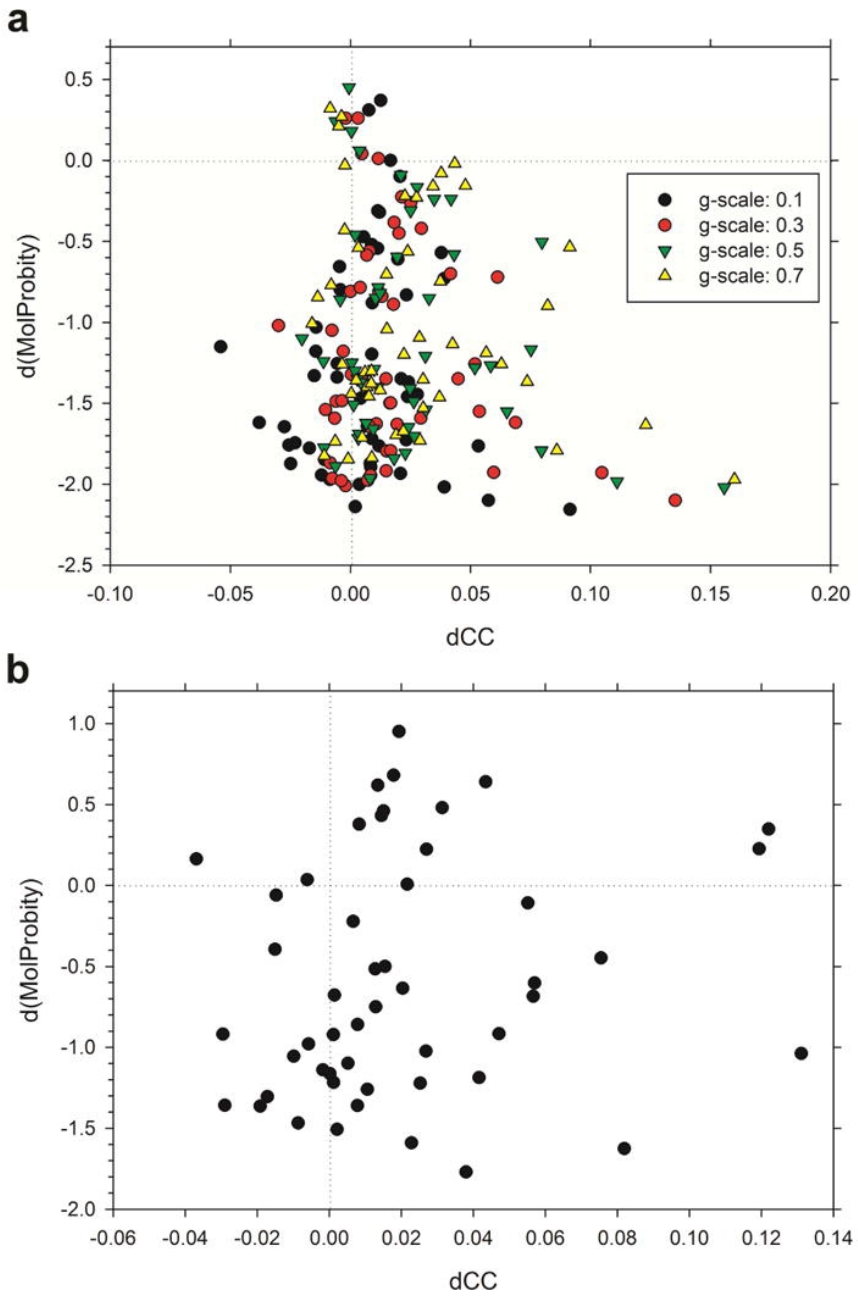

Figure 2.2: Change in MolProbity score (MPScore) and cross correlation between initial models and final refined models. (a), MDFF refinement results; (b), models refined with Rosetta.

Table 2.2: Classification of refined structure models using the MolProbity score. 49 structure models are classified in terms of the direction of the change in term of MolProbity score and cross-correlation to its EM maps. dM, the difference of the MolProbity score of a structure model after the refinement relative to the initial structure; dCC, the difference of cross-correlation of a structure model to its EM map after the refinement relative to the initial structure. The numbers in the parentheses for MDFF are the g-scale value used.

Method	**dM>0 & dCC>0**	**dM>0 & dCC<0**	**dM<0 & dCC>0**	**dM<0 & dCC<0**
MDFF (0.1)	2	0	30	17
MDFF (0.3)	3	1	35	10
MDFF (0.5)	2	2	40	5
MDFF (0.7)	0	3	37	9
Rosetta	13	0	25	11

To summarize this section, the majority of atomic-detailed structures deposited in the EMDB can have a better fit to their EM maps, and almost all among such cases also showed better structural energy. These results are consistent for three scores/energies: CHARMM, Rosetta energy, and MPScore. In particular, it is worth mentioning that the consistent trend was observed with an independent scoring function, MPScore.

2.3.2 Structure change relative to global map resolution

Next, we examined how much structures were changed by the refinement protocols in consideration with their associated EM map resolution. Fig. 3a shows Cα RMSD between structures before and after the refinement using MDFF with different g-scale values relative to the EM map resolution. Fig 3b shows results when Rosetta was used. In all of the plots, an inverse correlation was observed between map resolution and model-refinement RMSD. Structures tend to move more by MDFF when a larger g-scale value is used. It is apparent that the extent of the structure modification became larger as the map resolution decreases (i.e. larger resolution values on the x-axis). Results for MDFF with a g-scale of 0.5 fit to a weighted linear regression of RMSD = 0.528 + 0.247*(map resolution) with an r2 of 0.541. Since the distribution shows

heteroscedasticity (i.e. a fan-like distribution where the variance of RMSD becomes larger as the map resolution values increases), the reciprocal predicted RMSD value was used as weights. According to this regression line, estimated structural deviation (RMSD) for a model is about 26–31% (e.g. for a resolution of 8.0 Å, estimated RMSD is 2.50 Å, which is 31.3% of 8.0 Å) of the map's resolution. For the Rosetta refinement (Fig. 3b), a consistent relationship between the map resolution and the extent of model modification was observed. Compared with the MDFF refinement, Rosetta tended to make a larger modification than MDFF with a 0.5 g-scale. The RMSD relative to the map resolution fits to a weighted regression of −0.046 + 0.343 * (map resolution) with r2 of 0.504. The regression line is shown in Fig. 3b. According to this regression, the modification (RMSD) by Rosetta is roughly 33% of the map resolution.

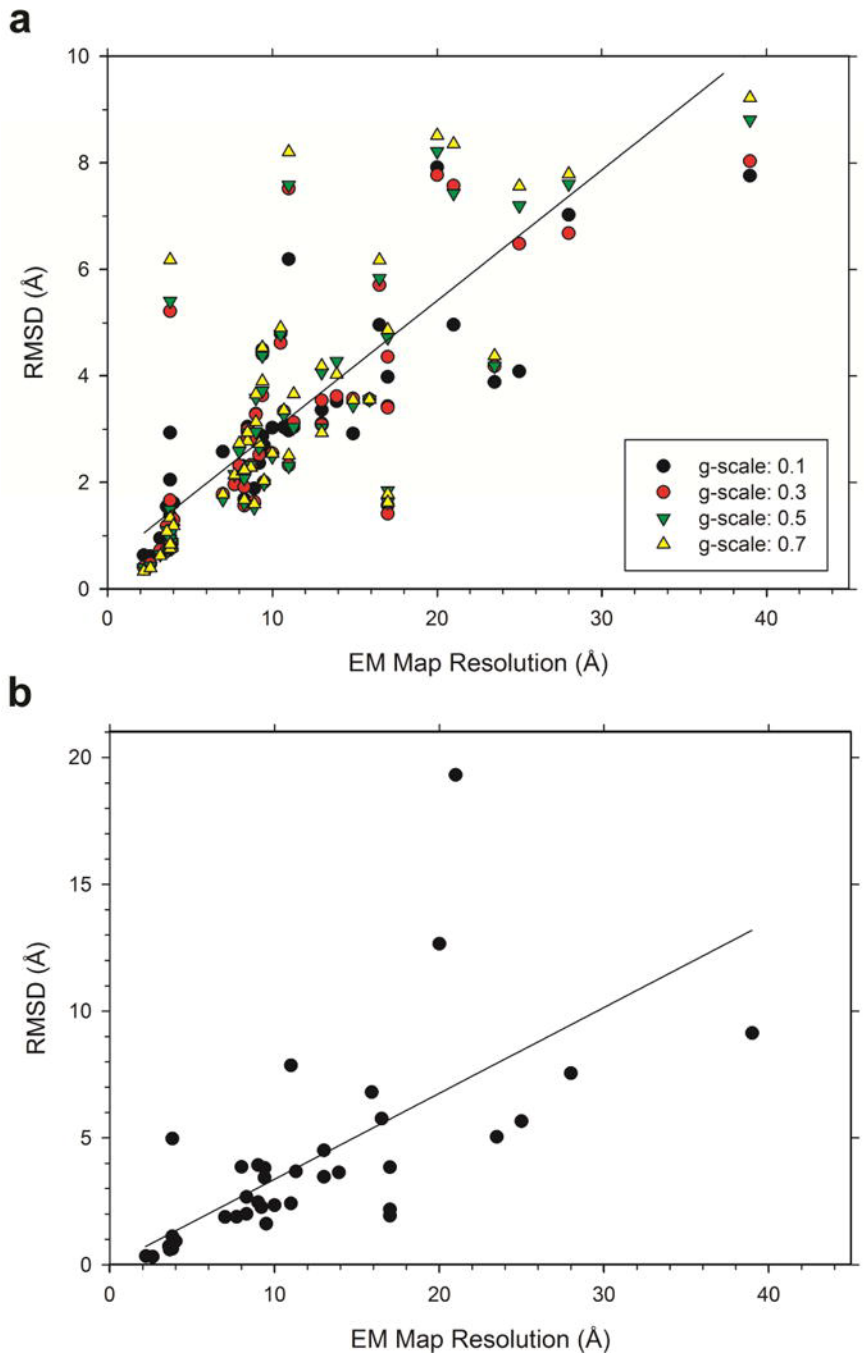

Figure 2.3: RMSD between the initial fitted protein model and the final structure after refinement against the resolution of their respective maps. (a), Results for MDFF with four different g-scale values, 0.1 (filled circles), 0.3 (red circles), 0.5 (green triangles), 0.7 (yellow triangles). The line shown is a weighted regression line for a g-scale of 0.5: RMSD = 0.528 + 0.247 *(map resolution). The reciprocal predicted value was used for weights. r2 is 0.541. (b), results for Rosetta. A weighted regression line using the reciprocal predicted value is shown: RMSD = −0.046 + 0.343*(map resolution). r2 is 0.504. To compute the regression lines, redundant entries of the same proteins with a similar map resolution and RMSD values were excluded. Those excluded were (GroEL: 2c7d, 2cgt, 4aau, 4ab3, 3zpz; α-1 glycine receptor: 3jad, 3jae; MacA-ClpC complex: 3j3r, 3j3s, 3j3u; β-galactosidase: 3j7h; TriC: 4a0v).

We further compared refined structures by MDFF and Rosetta to determine the discrepancy between the end results of these refinement protocols. Figure 4a shows RMSD between the refined structures produced by these two protocols. It shows that the discrepancy (i.e. RMSD) scaled inversely with map resolution, providing a similar picture as Figure 4. Four different symbols indicate comparison against MDFF results with the four different g-scale values. The overall trend does not change by using different g-scale values, but when examined closely, the discrepancy of MDFF-refined models against Rosetta-refined models are largest when a g-scale of 0.1 was used and smallest when 0.7 was used for the g-scale (Fig. 4b, Table S4 in Supplemental Information). This is reasonable because a larger g-scale makes the model refinement by MDFF more biased to the EM map, and influence of the different potential functions for Rosetta and MDFF is minimized. To summarize Figure 3 and Figure 4a, as a map resolution lowers, the two structure refinement protocols modify an initial structure with a larger extent, but the deviation of the resulting models by the two methods also increases. This result clearly indicates that maps with a low resolution do not have sufficient structure information to lead to a single model solution.

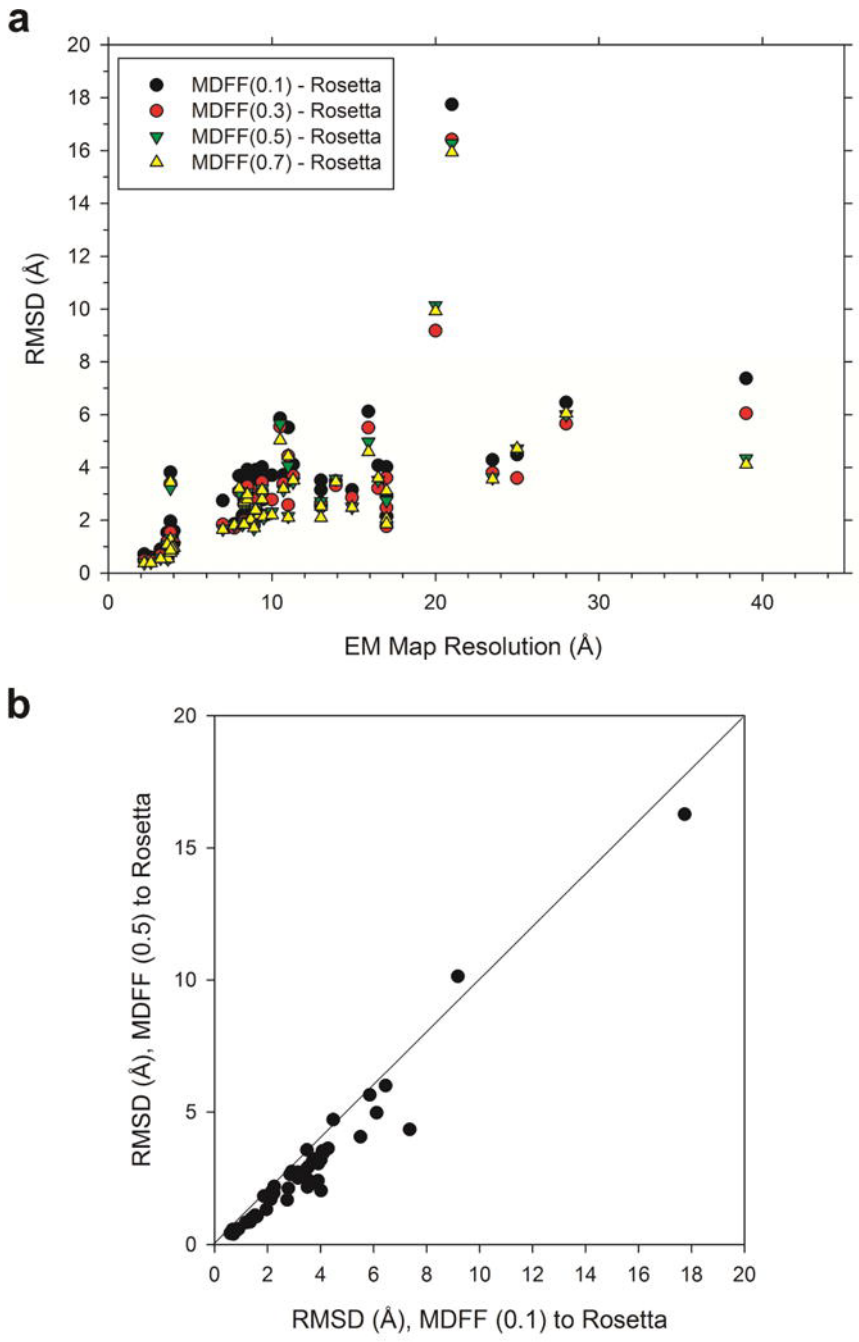

Figure 2.4: RMSD between the refined models using MDFF and Rosetta. (a), RMSD between Rosetta and MDFF refined models relative to the map resolutions. For MDFF, the four different g-scales were used. The color code is the same as Figs. 1 and and3.3. (b), RMSD between refined models by Rosetta and refined models by MDFF with a g-scale of 0.5 relative to the cases when a g-scale of 0.1 was used for MDFF.

2.3.3 Comparison with other crystal structures

In Figure 5, we compared refined models with other crystal structures of the same protein. Two contrasting examples of EM maps and their associated structure models were used. The first example shown in Figure 5a and 5b is an EM map of beta-galactosidase, EMD-5955 (PDB ID: 3j7h) in comparison with five other crystal structures (see the figure caption). This structure was solved at a high resolution, 3.2 Å. In this case, the energy difference of the refined model relative to the initial structure in terms of the CHARMM (Fig. 5a) and Rosetta (Fig. 5b) (the empty circle) energies was in the same range as the other crystal structures. The RMSD of the refined model from the initial structure (3j7h) was 0.65 Å for MDFF and 0.53 for Rosetta, slightly smaller than the RMSD of other crystal structures to 3j7h, which ranged from 0.69 to 0.75 Å. In the next example is a GroE structure, 1gru, associated with an EM map determined at a low resolution, 23.5 Å (EMD-1046). This structure, 1gru, was a result of a rigid-body fitting of a crystal structure, 1aon, to the EM map. Since the map was determined at a low resolution, there was relatively more room for the refinement protocols to move the structure, which resulted in a larger RMSD of 4.19 Å for MDFF (Fig. 5c) and 5.04 Å for Rosetta (Fig. 5d) to the initial structure than the other five crystal structures, whose RMSD values ranged from 2.68 to 2.79 Å. This is consistent with the other EM maps in our analysis (Fig. 3). In terms of energy, the absolute energy values of the refined models by MDFF and Rosetta were in the same range as the other crystal structures (Fig. 5c and 5d).

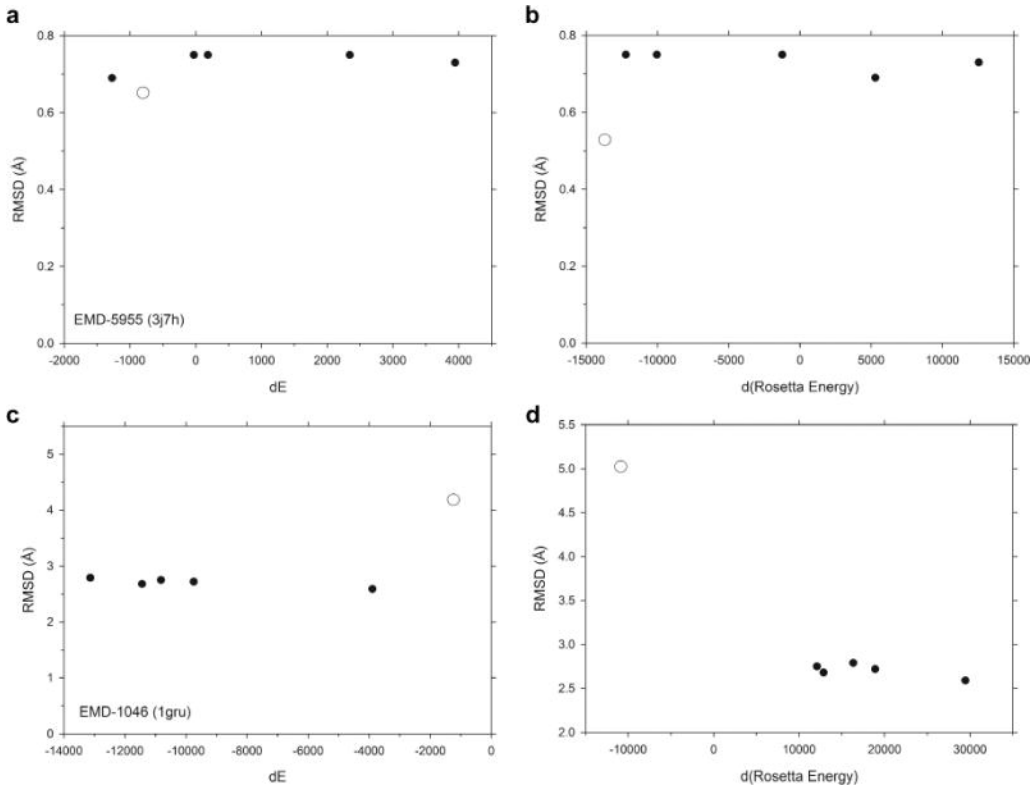

Figure 2.5: Comparison with other crystal structures. For two examples of EM maps with associated PDB entry, RMSD and the energy difference with other crystal structures were computed. (a), (b), beta-galactosidase, EMD-5955 (PDB ID: 3j7h), solved at 3.2 Å. (a) shows the CHARMM energy difference; and (b) shows the difference in term of the Rosetta energy with RMSD between 3j7h and five other crystal structures of the same protein, 1f4h, 1hn1, 1jz2, 3iaq, and 3t2o (solid circles). The open circle is the refined structure by (a) MDFF (g-scale 0.5) and (b) Rosetta, compared with 3j7h. For the CHARMM energy, structures were evaluated at the start of the refinement, after the initial energy minimization was applied and the temperature is raised to 300K, in the same way as the earlier figures. (c), (d), GroEL, EMD-1046 (PDB ID: 1gru), solved at 23.5 Å. Crystal structures used were 1aon, 1pcq, 1pf9, 1svt, and 1sx4 (solid circles).

2.3.4 Residue displacement relative to local map resolution

We have also examined the extent of displacement of each residue by the refinement protocols relative to the local map resolution (Fig. 6). Local map resolution was computed using the Resmap program (Kucukelbir et al., 2014), which provides a local resolution to each grid point in a map. In this analysis, the Cα atom of each residue in the initial structure was assigned to the nearest grid point in the EM map, and the displacement of the Cα atom by the refinement protocols was recorded. Then, the displacements of Cα atoms for grid positions were averaged over all grid

points with the same local resolution in the map. Data for a resolution was discarded if fewer than 10 Cα atoms belonged to the resolution in the map.

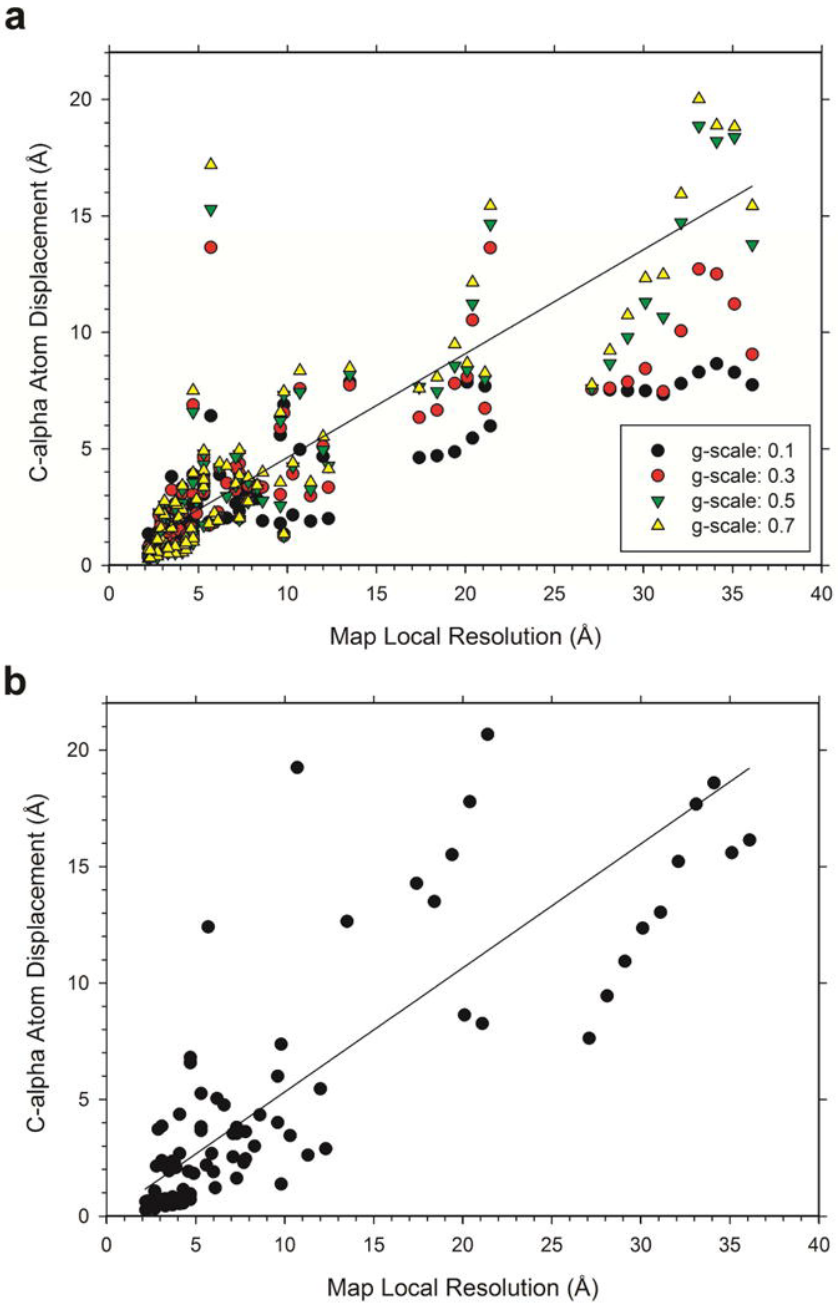

Figure 2.6: Distances of Cα atoms moved by the refinement relative to the local resolution maps. Local map resolution was computed with ResMap. For an EM map, the displacements of Cα atoms for grid positions with the same local resolution were averaged. Data for a resolution was discarded if less than 10 Cα atoms belonged to the resolution. (a), Results for MDFF with four different g-scale values, 0.1 (filled circles), 0.3 (red circles), 0.5 (green triangles), 0.7 (yellow triangles). The line shown is a weighted regression line for a g-scale of 0.5: Cα displacement = 0.124 + 0.446 *(local map resolution). The reciprocal predicted value was used for weights. r2 is 0.597. (b), results for Rosetta. A weighted regression line using the reciprocal predicted value is shown: RMSD = −0.03 + 0.533*(map resolution). r2 is 0.567.

Results in Figure 6 show a similar trend as we observed for the correlation between the global RMSD and the overall map resolution in Figure 3. Fig. 6a is for the MDFF protocol with the four g-scale values while the Rosetta results are shown in Fig. 6b. The extent of the residue displacements inversely correlated with the local map resolution (Fig. 6a and 6b) and a larger displacement was observed with a larger g-scale value for MDFF (Fig. 6a). The Cα displacement relative to the local map resolution fits to a weighted regression of 0.124 + 0.446 * (local map resolution) with r2 of 0.597 for MDFF with a g-scale of 0.5 (Fig. 6a) while the results for Rosetta fit to −0.03 + 0.533*(map resolution) with r2 of 0.567 (Fig. 6b). The slopes observed in the regression lines for the local resolution were larger than those for the global map resolution in Figure 3. Consistent with the global RMSD analysis in Figure 3, Rosetta moved residues in models more than MDFF. Following the regression lines, the estimated deviation of a residue in a model is roughly 46% and 53% of the map's local resolution for MDFF with a g-scale of 0.4 and Rosetta, respectively.

2.3.5 Examples of refined models

We discuss examples of refined structures by MDFF and Rosetta for models constructed for maps of four different resolutions. The first example (Fig. 7a) is L-protein of vesicular stomatitis virus determined from an EM map of 3.8 Å (EMDB ID: 6337, PDB ID: 5a22) (Liang et al., 2015). This model was manually constructed using programs O (Jones et al., 1991) and Coot (Emsley et al., 2010) by tracing the main-chain and placing helix segments guided by known homologous structures and secondary structure prediction, which is finally followed by loop and side-chain refinement. For this model, MDFF with g-scale 0.5 and Rosetta modified the structure to the extent of 0.944 Å and 0.762 Å RMSD from the deposited model, respectively. Slight deviation at loop regions was observed but no substantial structural changes occurred by the refinement.

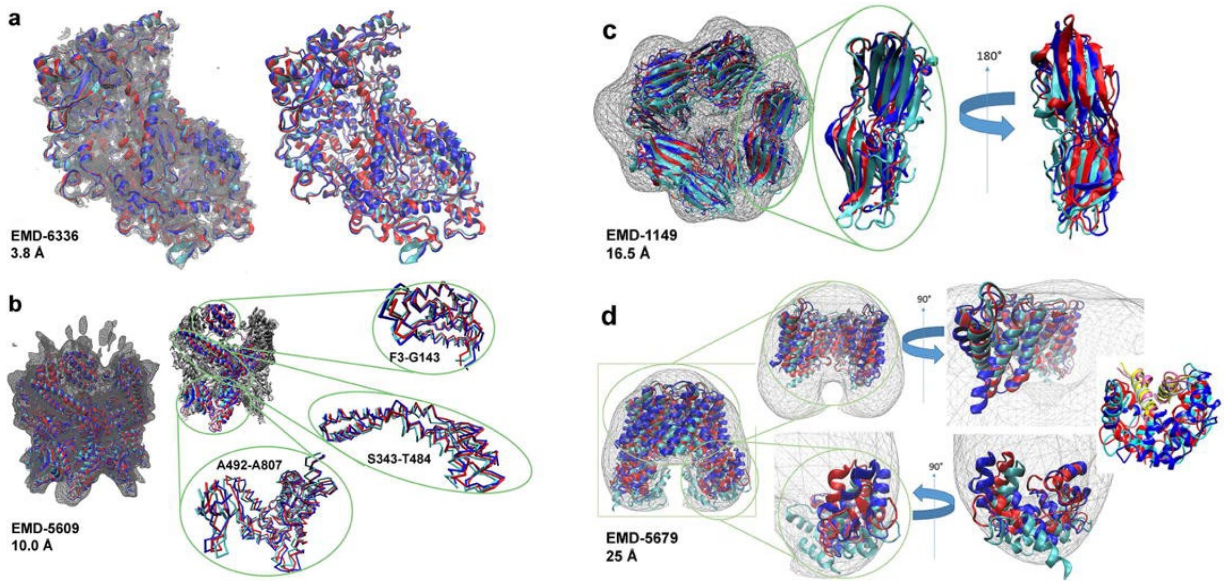

Figure 2.7: Examples of structure refinement. The overlay of selected initial and refined structures produced by MDFF (using a g-scale of 0.5) and Rosetta are colored cyan, blue, and red respectively. Density maps for these structures are shown as gray wire frames. (a), The 3.8 Å resolution map of L-protein of vesicular stomatitis virus (EMDB ID: 6337) and its atomic model (PDB ID: 5a22) (left), as well as the atomic model shown without the wire frame map for visual clarity (right). (b), The 10.0 Å resolution map, EMD-5609, and its structure model (PDB ID: 3j3u) of MecA-ClpC complex (left). The structures with the A chain shown in color, while the rest of the complex is shown in white (center). Selected domains are isolated and magnified for visual clarity (right). The residue range of these domains are included as insets near each image. (c), The 16.5 Å resolution map (EMDB: 1149) and its structure model (PDB ID: 2byu) of small heat shock protein Arc 1 (left). An isolated subunit of the structure magnified for visual clarity (center). A 180-degree rotated view of the isolated subunit (right). (d), A 25 Å resolution map (EMDB: 5649) and its structure model (PDB ID: 3j41) of aquaporin-O/calmodulin complex (left). Map and structure of the core region of the complex with lobe domains and front half of core removed for ease of viewing (top center). A rotated view of the core domain (top right). Magnification of a single lobe calmodulin domain with core domain removed (bottom center) and a rotated view of the calmodulin domain (bottom right). Interaction with the calmodulin domain with two helices (chain C, D: 225–241) (shown in yellow, light blue, and pink for the original structure model, the model from MDFF, and the Rosetta model, respectively) are highlighted in far right.

In the next example, a structure model of MecA-ClpC complex solved from an EM map of 10.0 Å resolution is shown (EMDB ID: 5609; PDB ID: 3j3u) (Liu et al., 2013). The model is based on a crystal structure of this complex, where some loops were built using Modeller (Sali and Blundell, 1993). The model was subsequently fit to the map with MDFF. By our refinement, MDFF with g-scale 0.5 produced a refined model of an RMSD of 2.484 Å from the initial structure, similarly Rosetta produced the model that deviated by 2.196 Å RMSD. As shown in Figure 7b, in addition to larger deviation of loop regions, modification of helical (e.g. residue 343 to 484) and β-sheet regions (e.g. residue 738 to 807) are clearly observed.

In the third example, a model of a dodecameric structure of the small heat shock protein Arc1 determined with an EM map of a 16.5 Å resolution is shown (EMDB ID: 1149; PDB ID: 2byu) (Kennaway et al., 2005). The map was solved by negative staining. This model was built by rigid body fitting taking symmetry into account. MDFF with g-scale 0.5 and Rosetta moved the overall structure by 5.835 Å and 4.349 Å, respectively. As evident from the figure (Fig. 7c), the whole subunits shifted relative to the map, indicating that the initial model only contained approximate subunit arrangement information of the complex.

When the map resolution is even lower, as illustrated in the final example, a model for aquaporin-O/calmodulin complex (EMDB ID: 5679; PDB ID: 3j41) solved from a 25 Å negative stain EM map (Reichow et al., 2013), the refinement caused large domain modification. The deposited model was constructed using a crystal structure of the transmembrane domain as a base, on which the rest of the structure was added through several steps of manual building of helices and loops. The lobe calmodulin domain saw compression of secondary structure elements into the density map and the transmembrane domain showed substantial rearrangement of orientation of helices even at the core of the complex. In both MDFF and Rosetta refined models, the calmodulin domain maintained its interaction with two helices from aquaporin-O as shown in the far right panel in Fig. 7d. The RMSD between the refined model and the initial model was 7.197 Å and 7.625 Å by MDFF and Rosetta, respectively.

2.4 Discussion

Here we used two popular EM structure fitting methods, MDFF and Rosetta, to investigate how well structure models derived from EM maps are supported by those maps. We refined structure models deposited in EMDB further using the two methods independently, and RMSD between structures before and after the refinement as well as discrepancy between resulting structures by the two methods were examined. It turned out that for both methods more than 60% of the cases structures can be refined to have a higher cross-correlation to associated EM maps, almost all of which were with a decrease of the energy of the structures. The extent of the observed structural change by the structural refinement increased as map's resolution declined, indeed it scaled inversely to map resolutions. According to the weighted regression lines that correlate structural change (RMSD) and map resolutions, model structures changed up to an RMSD of roughly 30% of the map resolution. A similar trend of structural change was observed relative to

local map resolution, too. As the discrepancy of refined models by MDFF and Rosetta also scaled inversely to map resolution, it is suggested that the observed structural changes by the refinement methods are due to the lack of structural information in EM maps but not because the deposited structures missed the optimal solution.

Recently, the Schulten group has proposed two structure modeling protocols for a high-resolution EM map using MDFF (Singharoy et al., 2016). The main idea behind the protocols is to start running MDFF with a blurred EM map so that the structure can explore a large conformation space avoiding local energy minima and then to gradually change the resolution of the map higher back to its original (high) resolution. The new procedures were tested on two actual high-resolution EM maps determined at 3.2 Å and 3.4 Å, one for beta galactosidase and another one for TRPV1 respectively. The new protocols would change the results for high resolution maps slightly if used in this study, but the overall trend would not change because the new protocols are designed for high resolution maps of higher than 4–5 Å and our dataset contains maps determined at a wide range of resolutions. Actually, our results using the standard MDFF and Rosetta on the 3.2 Å map of beta galactosidase (EMD-5955) are consistent with their results. RMSD values of our results were 0.65 Å and 0.53 Å with the standard MDFF and Rosetta, respectively, while the two new protocols obtained 0.7 Å and 0.9 Å (Table 1 in their paper). It is also interesting that their analysis on the root mean square fluctuation of MDFF-refined models relative to the local resolution of a map (Figure 4 in their paper) shows a similar trend to the residue displacement relative to the local map resolution, which we showed in Figure 6.

A structure model is built for an EM map not only from the electron density information in the map but also in consideration of other biological information of the proteins, such as known structures of homologous proteins and results of biochemical assays. Thus, a deposited structure model for an EM map would be the result of best effort in considering various sources of information of the protein with the EM maps as a piece of information. Nevertheless, it is advised that map resolution be critically considered when one uses EM-derived structures for further analysis; be that validation of a structure prediction or refinement methodology, the structure based design of a drug-like molecule, or the analysis of the biochemical or energetic character of these structures. Considering that an increasing number of structures are solved by EM and that EMDB is becoming a valuable source of biomolecular structural analysis, it is important that users of structures determined by EM are made well aware of the limitations of structure models. One

possible solution would be for the EMDB to include the local resolution information of the map in the structure analysis report called Visual Analysis, which is associated with each EM map entry in EMDB. It is so important that the local resolution information is linked to the associated PDB entry of the map, either in the wwPDB EM Map/Model Validation Report, which is provided for each entry in PDB, or even in the PDB file itself in an explicit way, for example, by providing resolution information for each atom in a structure model analogous to the B-factor of X-ray crystallography.

2.5 References

Bradley, P., Misura, K. M., & Baker, D. (2005). Toward high-resolution de novo structure prediction for small proteins. *Science*, *309*, 1868-1871.

Ceulemans, H., & Russell, R. B. (2004). Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization. *Journal of molecular biology*, *338*, 783-793.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Kapral, L. W., Richardson, J. S., & Richardson, D. C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, *66*, 12-21.

DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W., & Baker, D. (2009). Refinement of protein structures into low-resolution density maps using rosetta. *Journal of molecular biology*, *392*, 181-190.

Dror, O., Lasker, K., Nussinov, R., & Wolfson, H. (2007). EMatch: an efficient method for aligning atomic resolution subunits into intermediate-resolution cryo-EM maps of large macromolecular assemblies. *Acta Crystallographica Section D: Biological Crystallography*, *63*, 42-49.

Egelman, E. H. (2008). Problems in fitting high resolution structures into electron microscopic reconstructions. *HFSP journal*, *2*, 324-331.

Emsley, P., Lohkamp, B., Scott, W. G., & Cowtan, K. (2010). Features and development of Coot. *Acta Crystallographica Section D: Biological Crystallography*, *66*, 486-501.

Esquivel-Rodríguez, J., & Kihara, D. (2012). Fitting multimeric protein complexes into electron microscopy maps using 3D Zernike descriptors. *The journal of physical chemistry B*, *116*, 6854-6861.

Esquivel-Rodríguez, J., & Kihara, D. (2013). Computational methods for constructing protein structure models from 3D electron microscopy maps. *Journal of structural biology*, *184*, 93-102.

Falkner, B., & Schröder, G. F. (2013). Cross-validation in cryo-EM–based structural modeling. *Proceedings of the National Academy of Sciences*, *110*, 8930-8935.

Faruqi, A. R., & Henderson, R. (2007). Electronic detectors for electron microscopy. *Current opinion in structural biology*, *17*, 549-555.

Henderson, R., Sali, A., Baker, M. L., Carragher, B., Devn.kota, B., Downing, K. H., Egelman, E. H., Feng, Z., Frank, J., Grigorieff, N., Jiang, W., Ludtke, S. J., Medalia, O., Penczek, A. P., Rosenthal P. B., Rossmann, M. G., Schmid, M. F., Schröder G. F., Steven, A. C., Stokes, D. L., Westbrook, J. D., Wriggers, W., Yang, H., Young, J., Berman, H. M., Chiu, W., Kleywegt G. J. & Lawson, C. L. (2012). Outcome of the first electron microscopy validation task force meeting. *Structure*, *20*, 205-214.

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics*, *14*, 33-38.

Jiang, W., Baker, M. L., Ludtke, S. J., & Chiu, W. (2001). Bridging the information gap: computational tools for intermediate resolution structure interpretation. *Journal of molecular biology*, *308*, 1033-1044.

Jolley, C. C., Wells, S. A., Fromme, P., & Thorpe, M. F. (2008). Fitting low-resolution cryo-EM maps of proteins using constrained geometric simulations. *Biophysical journal*, *94*, 1613-1621.

Jones, T. A., Zou, J. Y., Cowan, S. T., & Kjeldgaard, M. (1991). Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallographica Section A: Foundations of Crystallography*, *47*, 110-119.

Kennaway, C. K., Benesch, J. L., Gohlke, U., Wang, L., Robinson, C. V., Orlova, E. V., Saibi, H. R. & Keep, N. H. (2005). Dodecameric structure of the small heat shock protein Acr1 from Mycobacterium tuberculosis. *Journal of Biological Chemistry*, *280*, 33419-33425.

Kucukelbir, A., Sigworth, F. J., & Tagare, H. D. (2014). Quantifying the local resolution of cryo-EM density maps. *Nature methods*, *11*, 63-65.

Kühlbrandt, W. (2014). The resolution revolution. *Science*, *343*, 1443-1444.

Kühlbrandt, W. (2014). Microscopy: cryo-EM enters a new era. *Elife*, *3*, e03678.

Lawson, C. L., Baker, M. L., Best, C., Bi, C., Dougherty, M., Feng, P., ... & Chiu, W. (2010). EMDataBank. org: unified data resource for CryoEM. *Nucleic acids research*, *39*, D456-D464.

Liang, B., Li, Z., Jenni, S., Rahmeh, A. A., Morin, B. M., Grant, T., ... & Whelan, S. P. (2015). Structure of the L protein of vesicular stomatitis virus from electron cryomicroscopy. *Cell*, *162*, 314-327.

Liu, J., Mei, Z., Li, N., Qi, Y., Xu, Y., Shi, Y., Wang, F., Lei, J. & Gao, N. (2013). Structural Dynamics of the MecA-ClpC Complex: A TYPE II AAA+ PROTEIN UNFOLDING MACHINE. *Journal of Biological Chemistry*, *288*, 17597-17608.

McGreevy, R., Singharoy, A., Li, Q., Zhang, J., Xu, D., Perozo, E., & Schulten, K. (2014). xMDFF: molecular dynamics flexible fitting of low-resolution X-ray structures. *Acta Crystallographica Section D: Biological Crystallography*, *70*, 2344-2355.

McGreevy, R., Teo, I., Singharoy, A., & Schulten, K. (2016). Advances in the molecular dynamics flexible fitting method for cryo-EM modeling. *Methods*, *100*, 50-60.

Modi, V., & Dunbrack Jr, R. L. (2016). Assessment of refinement of template-based models in CASP11. *Proteins: Structure, Function, and Bioinformatics*, *84*, 260-281.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, *25*, 1605-1612.

Reichow, S. L., Clemens, D. M., Freites, J. A., Németh-Cahalan, K. L., Heyden, M., Tobias, D. J., Hall, J., & Gonen, T. (2013). Allosteric mechanism of water-channel gating by Ca 2+–calmodulin. *Nature structural & molecular biology*, *20*, 1085-1092.

Rosenthal, P. B., & Rubinstein, J. L. (2015). Validating maps from single particle electron cryomicroscopy. *Current opinion in structural biology*, *34*, 135-144.

Rossmann, M. G. (2000). Fitting atomic models into electron-microscopy maps. *Acta Crystallographica Section D: Biological Crystallography*, *56*, 1341-1349.

Šali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, *234*, 779-815.

Scheres, S. H. (2012). RELION: implementation of a Bayesian approach to cryo-EM structure determination. *Journal of structural biology*, *180*, 519-530.

Singharoy, A., Teo, I., McGreevy, R., Stone, J. E., Zhao, J., & Schulten, K. (2016). Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *Elife*, *5*, e16105.

Velankar, S., van Ginkel, G., Alhroub, Y., Battle, G. M., Berrisford, J. M., Conroy, M. J., Dana, J. M., Gore, S. P., Gutmanas, A., Haslam, P., Hendrickx, P. M. S., Lagerstedt, I., Mir, S., Montecelo, M. A. F., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Sanz-García, E., Sen, A., Slowley, R. A., Wainwright, M. E.,Deshpande, M. S., Ludin, A., Sahni, G., Torres, J. S., Hirshberg, M., Mak, L., Nadzirin, N., Armstrong, D. R., Alice, R. C., Smart, O. S., Korir, P. K., & Kleywegt, G. J. (2016). PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic acids research*, *44*, D385-D395.

Villa, E., & Lasker, K. (2014). Finding the right fit: chiseling structures out of cryo-electron microscopy maps. *Current opinion in structural biology*, *25*, 118-125.

Wang, R. Y. R., Kudryashev, M., Li, X., Egelman, E. H., Basler, M., Cheng, Y., Baker, D., & DiMaio, F. (2015). De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nature methods*, *12*, 335-338.

Woetzel, N., Karakaş, M., Staritzbichler, R., Müller, R., Weiner, B. E., & Meiler, J. (2012). BCL:: Score—knowledge based energy potentials for ranking protein models represented by idealized secondary structure elements. *PloS one*, *7*, e49242.

Wriggers, W., & Chacon, P. (2001). Modeling tricks and fitting techniques for multiresolution structures. *Structure*, *9*, 779-788.

Wriggers, W., & He, J. (2015). Numerical geometry of map and model assessment. *Journal of structural biology*, *192*, 255-261.

Yang, Z., Lasker, K., Schneidman-Duhovny, D., Webb, B., Huang, C. C., Pettersen, E. F., ... & Ferrin, T. E. (2012). UCSF Chimera, MODELLER, and IMP: an integrated modeling system. *Journal of structural biology*, *179*, 269-278.

Zhu, J., Cheng, L., Fang, Q., Zhou, Z. H., & Honig, B. (2010). Building and refining protein models within cryo-electron microscopy density maps based on homology modeling and multiscale structure refinement. *Journal of molecular biology*, *397*, 835-851.

CHAPTER 3. USING STEERED MOLECULAR DYNAMIC TENSION FOR ASSESSING QUALITY OF COMPUTATIONAL PROTEIN STRUCTURE MODELS

Chapter reproduced from a manuscript previously published in the journal *Journal of Computational Chemistry* in 2022

3.1 Abstract

The native structures of proteins, except for notable exceptions of intrinsically disordered proteins, in general take their most stable conformation in the physiological condition to maintain their structural framework so that their biological function can be properly carried out. Experimentally, the stability of a protein can be measured by several means, among which the pulling experiment using the atomic force microscope (AFM) stands as a unique method. AFM directly measures the resistance from unfolding, which can be quantified from the observed force-extension profile. It has been shown that key features observed in an AFM pulling experiment can be well reproduced by computational molecular dynamics simulations. Here, we applied computational pulling for estimating the accuracy of computational protein structure models under the hypothesis that the structural stability would positively correlated with the accuracy, i.e. the closeness to the native, of a model. We used in total 4,929 structure models for 24 target proteins from the Critical Assessment of Techniques of Structure Prediction (CASP) and investigated if the magnitude of the break force, i.e. the force required to rearrange the model's structure, from the force profile was sufficient information for selecting near-native models. We found that near-native models can be successfully selected by examining their break forces suggesting that high break force indeed indicates high stability of models. On the other hand, there were also near-native models that had relatively low peak forces. The mechanisms of the stability exhibited by the break forces were explored and discussed.

3.2 Introduction

It is generally understood that the native structure of a protein adopts the most thermodynamically stable fold in the protein's conformation space (Anfinsen, 1973) except for some notable examples including of intrinsic disordered proteins and prions (Pan et al., 1993). An

implication of this postulation, which is called the thermodynamic hypothesis, is that the thermodynamically stable structures of proteins are dictated by the amino acid sequence under physiological conditions, leading to the development of methods for protein structure prediction from amino acid sequence. In structure prediction, a method often generates tens to thousands of possible conformations for a single protein sequence. This required the development of a sub-field of structure prediction, which is known as protein quality assessment (QA) (Kihara et al., 2009). QA is aimed at predicting the accuracy of computational structure models or rank models in a model pool based on their expected accuracy. Strategies of QA include application of statistical and physical potential functions (Uziela et al., 2016)(Manavalan & Lee, 2017)(Luthy et al., 1992), evaluating consistency to stereochemistry of models to known structures (Kim & Kihara, 2014)(Laskowski et al., 1993), consistency of models with predicted local structures and alignments to known structures (Chen & Kihara, 2014)(Wallner & Elofsson, 2006), consensus with alternative models (Cao et al., 2015)(Wang et al., 2011), and machine learning approaches that combine various structural features of models (Cao et al., 2017)(Uziela et al., 2017)(Shin et al., 2017)(Yang et al., 2010).

Here, we evaluated computational protein structure models by directly measuring their stability in virtual pulling experiment of protein models. We used Steered Molecular Dynamics (SMD), which mimics pulling experiments with Atomic Force Microscopy (AFM). AFM conventionally scans a sample by a probe and can measure the force between the probe and the sample surface or can be applied to image the sample shape (Hersam & Chung et al., 2010). With regard to proteins, AFM has been used to measure the binding force between a protein and a ligand (Li et al., 2006), to characterize enzyme activity (Arredondo et al., 2018), to study unfolding pathways (Bujalowski & Oberhauser, 2013)(Carrion-Vazquez et al., 1999)(Rounsevell et al., 2004) and mechanical stability of proteins(Valbuena et al., 2009). Early applications of SMD used to mimic AFM on biological systems were performed by the Schulten group to study the unbinding of the avidin-biotin complex (Izrailev et al., 1997) and the unfolding pathway of titin IG domains(Lu et al., 1998), which showed good agreement with contemporary experiments using AFM (Rief et al., 1997). Recently, simulated AFM was applied to investigate unfolding of the cold shock protein B (Csp) (Schonfelder et al., 2016) and the Src SH3 domain (Zhuravlev et al., 2016). SMD has also been used to probe mechanical resistance of peptide-receptor interactions[29].

In this study we hypothesized that, under MD-simulated stress, models with a high structural similarity to their native conformation would be more resistant to unfolding than models with more inaccurate folds. Using a set of computational models submitted to the Critical Assessment of Techniques for protein Structure Prediction (CASP) (Moult et al., 2014)(Moult et al., 2016)(Moult et al., 2018), a community-wide experiment of protein structure prediction, we investigated if models that are close to the native would be selected by examining the peak of the forces measured in the simulated pulling experiment. We found that near-native models can be successfully selected by examining the magnitude of their break forces, that is the force required to rearrange the structure, suggesting that high break forces indeed indicates high stability of models. On the other hand, there were also models that were similarly close to the native yet had relatively lower break forces. By comparing these two groups, it was determined that among near-native models, the force required for structural rearrangement was dependent on three primary factors; bonded interactions, electrostatic interactions, and solvent interactions.

3.3 Materials and Methods

3.3.1 Data Set

We used the submitted prediction models for CASP10, 11, and 12, which are available at the CASP website (Taylor et al., 2014)(Kinch et al., 2016)(Abriata et al., 2018). The files were available to download at https://predictioncenter.org for CASP10 files, and at the corresponding locations for targets from CASP 11 and 12. From the CASP data, target sets were selected for use if the sets included models with GDT-TS (global distance test total score) (Zemla, 2003) below 50.0 and above 85.0. GDT-TS computes the average of the percentage of C□ atoms in a model that are modelled within 1.0, 2.0, 4.0, and 8.0 Å and ranges from 0 to 100. After filtering by GDT-TS scores, we further removed any target set which were either part of a larger complex or were associated with lipid molecules, as the stabilizing effects of either a binding partner or lipid environment would not be present in the simulation. Any models with disulfide bonds were removed, as the molecular dynamics would not break these disulfide bonds and would rapidly become unphysical. We also checked knotted proteins (Sulkowska et al., 2008) by comparing with the KnotProt 2.0 database (Dabrowski-Tumanski, et al., 2019) in the intention of removing it from the dataset. The only knotted proteins found was T0826 (PDB: 5fgn), a slipknot type knot protein.

However, T0826 was not considered for our dataset because the decoy set did not include high accurate models. The max GDT_TS of 63.46, which was below our 85.0 cut off.

Filtering at this point left us with 24 target sets with a total of 4,929 models to test. The average number of models for a target was 205.38. The average minimum, maximum, and average GDT_TS of models for a target were, 15.84, 92.17, and 75.48, respectively. The distribution of models across GDT_TS ranges for each target is shown in Table 1. GDT-TS. values of models were taken from the tables that associated with the model files from the CASP website.

Table 3.1**:** The dataset of protein structure models.

Target	PDB	Length	Total	100-80	80-60	60-40	40-20	>20
T0644	4fr9	158	549	156 (28.4)	208 (37.9)	24 (4.4)	118 (21.5)	43 (7.8)
T0650	5fmz	333	241	103 (42.7)	71 (29.5)	40 (16.6)	17 (7.1)	10 (4.1)
T0659	4esn	89	253	193 (76.3)	10 (4.0)	6 (2.4)	42 (16.6)	2 (0.8)
T0689	4fvs	225	240	158 (65.8)	23 (9.6)	1 (0.4)	28 (11.7)	30 (12.5)
T0700	4hfx	76	585	106 (18.1)	174 (29.7)	288 (49.2)	17 (2.9)	0 (0.0)
T0709	6mm4	33	552	439 (79.5)	41 (7.4)	64 (11.6)	8 (1.4)	0 (0.0)
T0711	2m7t	33	565	294 (52.0)	146 (25.8)	79 (14.0)	46 (8.1)	0 (0.0)
T0712	4gbs	203	248	162 (65.3)	26 (10.5)	16 (6.5)	0 (0.0)	44 (17.7)
T0714	2lvc	88	271	165 (60.9)	82 (30.3)	18 (6.6)	5 (1.8)	1 (0.4)
T0716	2ly9	70	276	237 (85.9)	26 (9.4)	13 (4.7)	0 (0.0)	0 (0.0)
T0731	2lz1	78	259	196 (75.7)	30 (11.6)	19 (7.3)	11 (4.2)	3 (1.2)
T0738	4is2	239	264	195 (73.9)	48 (18.2)	1 (0.4)	19 (7.2)	1 (0.4)
T0749	4gl3	429	254	194 (76.4)	8 (3.1)	0 (0.0)	16 (6.3)	36 (14.2)
T0752	4gb5	148	262	67 (25.6)	163 (62.2)	12 (4.6)	19 (7.3)	1 (0.4)
T0755	4h1x	248	240	9 (3.8)	9 (3.8)	52 (21.7)	152 (63.3)	18 (7.5)
T0757	4gak	243	236	31 (13.1)	165 (69.9)	22 (9.3)	10 (4.2)	8 (3.4)
T0762	4q5t	272	178	121 (68.0)	46 (25.8)	8 (4.5)	1 (0.6)	2 (1.1)
T0766	4q52	127	181	128 (70.7)	34 (18.8)	16 (8.8)	0 (0.0)	3 (1.7)
T0768	4oju	163	186	7 (3.8)	56 (30.1)	113 (60.8)	8 (4.3)	2 (1.1)
T0773	2n2u	66	544	104 (19.1)	197 (36.2)	132 (24.3)	107 (19.7)	4 (0.7)
T0797	4ojk	32	469	278 (59.3)	133 (28.4)	45 (9.6)	13 (2.8)	0 (0.0)
T0798	4ojk	189	422	386 (91.5)	9 (2.1)	8 (1.9)	10 (2.4)	9 (2.1)
T0801	4piw	367	188	129 (68.6)	42 (22.3)	10 (5.3)	5 (2.7)	2 (1.1)
T0811	- [a)]	180	183	158 (86.3)	18 (9.8)	0 (0.0)	5 (2.7)	2 (1.1)
T0815	4u13	106	177	132 (74.6)	42 (23.7)	0 (0.0)	1 (0.6)	2 (1.1)
T0861	5j5v	324	181	143 (79.0)	15 (8.3)	3 (1.7)	8 (4.4)	12 (6.6)
T0865	2n64	55	175	11 (6.3)	65 (37.1)	77 (44.0)	21 (12.0)	1 (0.6)
T0891	4ymp	125	172	124 (72.1)	24 (14.0)	8 (4.7)	4 (2.3)	12 (7.0
T0903	5a7d	368	182	76 (41.8)	50 (27.5)	17 (9.3)	24 (13.2)	15 (8.2)

The model sets were taken from CASP 10-12. The number of models was counted for five GDT_TS ranges, 100 to 80, 80-60, 60-40, 40-20, and less than 20. The percentage of the total is shown in parentheses. a), The PDB entry ID is not available for this target. GDT_TS values of all the models from all the targets were taken from the data file that associated with the model structure files downloaded from the CASP website.

For each model, we produced two additional variant models with identical backbone structure but different sequences. These different sequence models consisted of one in which all residues were converted to alanine, and one where the sequence was reversed. This was done by aligning an all alanine and a reversed sequence with the original model's sequence, and then producing a structure using MODELLER (Webb & Sali, 2017). The purpose of these additional models was to introduce a sequence variable to the sequence-structure relationship. Converting the sequence of each protein to all alanine removed the specific sequence characteristics and unique sidechain interactions that contributed to the energetic profile of the fold. Reversing the sequence was for the same purpose but retaining the same set of atoms and total mass as the original model.

3.3.2 Pulling a structure model using Molecular Dynamics

All molecular dynamics (MD) simulations were carried out using NAMD(Philips et al., 2005). The MD protocol consisted of four phases; first, a psf file that contains atom bonds and angle information of a target protein structure was generated using VMD's autopsf function. and the models were then solvated with TRIP3P water molecules using the VMD autopsf plug-in, solvate(Humphrey et al., 1996). Second, the models were minimized using a step size of 1 femto seconds (fs) for 100 pico seconds (ps) to remove atomic clashes. Subsequently, the models were equilibrated with the temperature increasing from 10 to 300 Kelvin over 30 ps and then further equilibrated for 0.5 nano seconds (ns). Finally, the structures were simulated under a steered MD (SMD) framework for 1 ns.

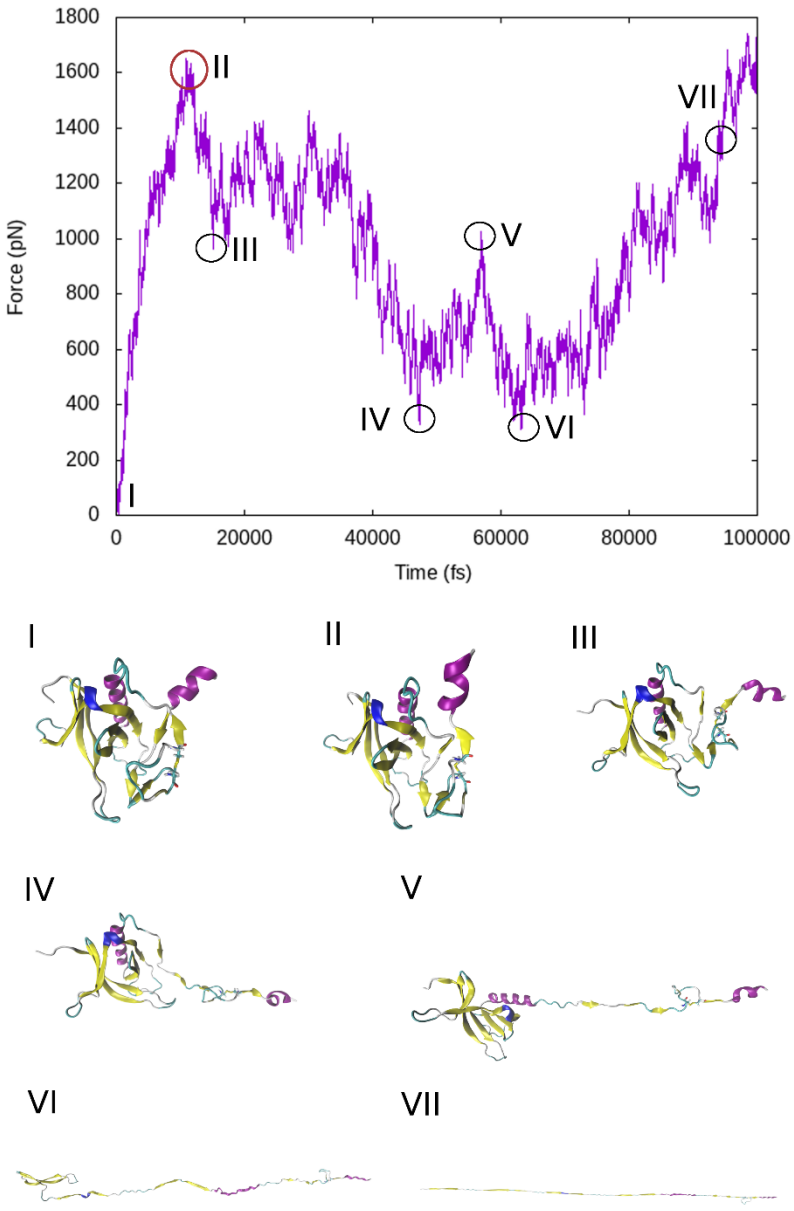

Figure 3.1: Force overtime with snap shots of the structure model T0659TS439_2 through its SMD pulling trajectory. Corresponding conformation of the model at their respective points in the trajectory are shown. Large conformational shifts are seen from II to III. The peak selected here is denoted by a red circle.

In the production run, an SMD atom with no mass or charge was placed in the simulation with a spring with a spring constant of 7 kcal/mol/Å^2 connecting the dummy atom with the C-terminal alpha-carbon. This dummy atom was them moved at a constant velocity of 0.001 Å per fs along the vector connecting the N-terminal alpha-carbon and the C-terminal alpha carbon.

Meanwhile, the N-terminal alpha-carbon was fixed in place. This caused the dummy atom to pull the protein until it unfolded and eventually linearized. By calculating the tension on the simulated spring, we determined the force required to pull apart and rearrange domains of the protein throughout the simulation. **Figure 3.1** shows an example of a force profile of a protein model. From the profile, the force required for rearrangement was taken as the first substantial peak in a trajectory, highlighted with a red circle in **Figure 3.1**. Along the trajectory, the structure of the model is shown below the profile panel. Stage I shows the started structure before the simulation begins. Stage II shows the structure at a high tension just before rearrangement. At this point, the only significant modification to the structure is the terminal coil being pulled tight rather than loose as in stage I. At stage III the structure is relaxed after much of the tension has been relieved. At this point, the two □ sheet domains have slid relative to each other and are no longer strongly connected. Stage IV shows a peeled structure where the □ sheet region has been split in two and pulled to either side of the protein. Stage V shows a further stretched structures. At stage VI, only a couple of secondary structures have remained, which are finally completely stretched at stage VII.

The peaks were selected by our script, which is made available at ***http://kiharalab.org/SMD/***. For automated peak selection, we first smoothed the data using a moving average with a window of 10, the equivalent of 1ps. Using this smoothed data, we then detected the force peak *F(t)* at time *t* that satisfies the following criteria:

$$t > 10ps$$

$$AND$$

$$F(t) > F(t-1ps) > F(t-2ps) > F(t-3ps) > F(t-4ps) > F(t-5ps)$$

$$AND$$

$$F(t) > F(t+1ps) > F(t+2ps) > F(t+3ps) > F(t+4ps) > F(t+5ps)$$

Thus, the force at that time step was greater than its neighbors of within 5 ps and that appeared after 10 ps

3.3.3 Logistic regression to predict GDT_TS

We also predicted GDT_TS of models from break forces observed in the native, all-alanine, and reversed-sequence models of the same conformations using logistic regression. For prediction of models in a target, all the models from other targets were used. Parameters of the regression

models and the mean square error of the predictions of the target proteins are provided in Supplementary Table 1.

3.4 Results

3.4.1 Examples of the break force relative to the quality

The main idea of this work is to use the break force from steered MD to select high quality protein structure models. In **Figure 3.2**, we show examples of structure models with varying model qualities as defined by GDT_TS. Panel a shows the force profile of four models of T0659, a 89 residue-long single domain protein of a b-sandwich fold. The highest quality model among them, the one with GDT_TS of 94.59, clearly has the highest peak. The other three models, with a GDT_TS of 79.73, 62.84, and 28.72, (panel c) have lower break force than the highest quality model. The second highest break force was observed for the model with 28.72 GDT_TS. Thus, interestingly, the break force does not have a clear overall correlation to GDT_TS of models.

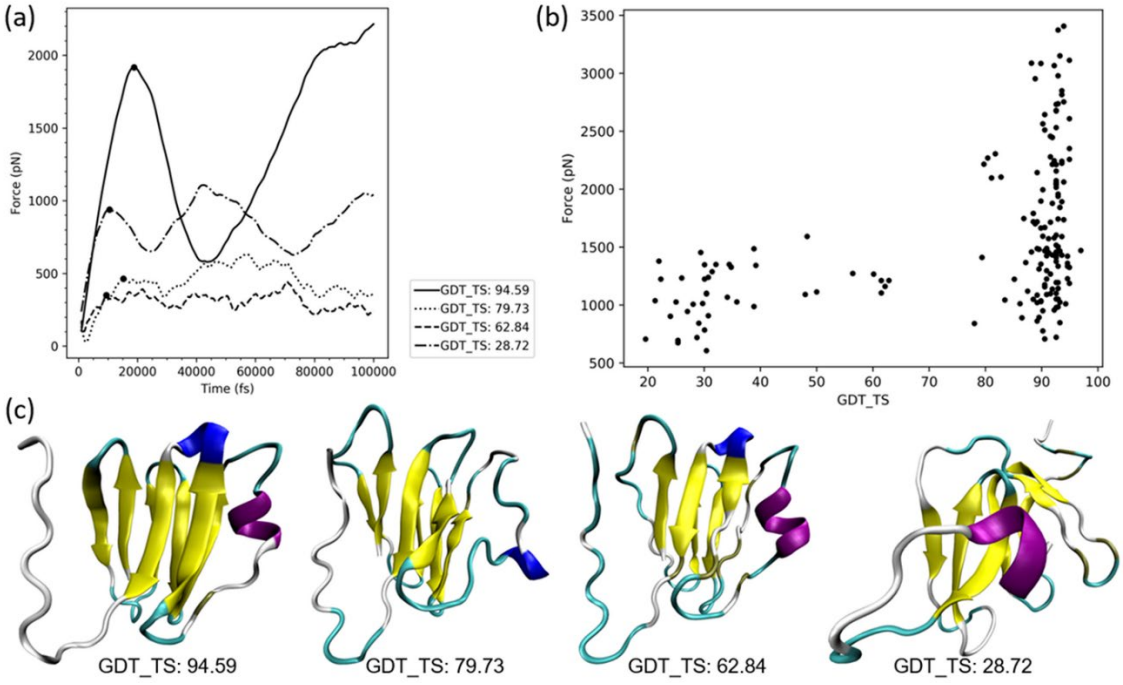

Figure 3.2**:** Examples of force curves of models with different GDT_TS. **(a)** Force curves of four structure models for T0659: T0659TS222_1, T0659TS035_5, T0659TS114_3, and T0659TS179_4, which are with GDT_TS scores of 94.69, 79.73, 62.84, and 28.72, respectively. The peak force selected for a force curve is indicated with a black dot. **(b)** Peak forces of individual models of the targetT0659 plotted against starting GDT_TS. **(c)** the starting models corresponding to the curves in the panel (a) with GDT_TS scores inset.

Panel b shows break force distribution of all 253 models of this target. It can be seen that models have a high break force have high GDT_TS, around 90. Thus, if models are sorted by their break force, top ranked models all have a GDT_TS around 90. However, there are high quality models with a lower break force, e.g. below 1500 pN, the same level of break force as models of 30 GDT_TS. Later we will discuss difference of high-quality models with relatively high and low break force.

3.4.2 Predictive Capabilities

The underlying question of this work was if the physical stress given by pulling the protein chain can be used to determine relative model quality of computational models. To answer the question, for each target in the model set (**Table 3.2**), we selected five models with the highest break force values and counted the number of high-quality models with a GDT_TS score of 80 or higher among them (**Figure 3.3**). Also shown are the results from the same analysis with all alanine sequence models and reversed sequence models. We evaluated the model selection performance by choosing five models following the CASP structure prediction evaluation, where participants are asked to submit five models for a target protein. As reference, selection by break force was compared with random selection of five models. The random selection was performed 1000 times, and the average counts from the1000 selections were reported.

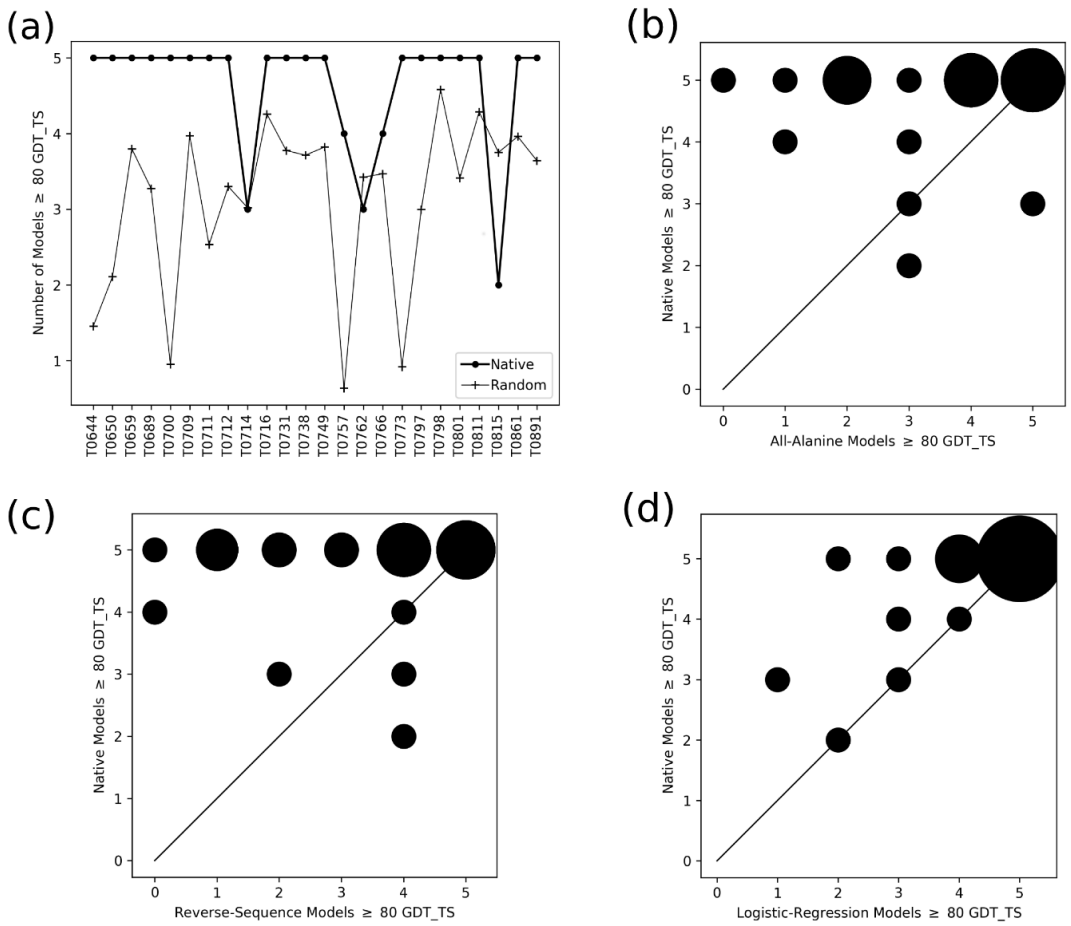

Figure 3.3**:** Performance of model selection with first peak force. For a model pool of a target protein, five models with the highest peak forces were selected. **(a)** For each target, the number of high-quality models that have a GDT_TS score of 80 or higher among the top five selections were plotted. The results of the selection (black circles connected by bold lines) are compared with random selection (crosses connected by thin lines). **(b)** Comparison of the model selection performance for all-alanine models (x-axis) and the native sequence models (y-axis). For both cases, the number of high-quality models with a GDT_TS of 80 or higher among the top five selections was counted. The area of bubbles is proportional to the number of target sets at the same coordinates. **(c)** Model selection performance comparison for the reversed-sequence models (x-axis) and the native sequence models (y-axis). **(d)** Model selection performance comparison for the logistic regression (x-axis) and the native sequence models (y-axis).

Figure 3.3a compares the model selection results using break force values with random selection. Out of 24 targets, using break force showed better performance than random for 21 targets (87.5%). Among the 21 targets, in 19 cases all five selections, thus, 100%, had over 80 GDT_TS score. In **Figure 3.3b** and **3.3c**, we compared the selection on the native sequence models with all-alanine models and reversed sequence-models, respectively. For both cases, the model

selection worked better for the native sequence models, indicating that specific amino acid interactions local structures in the native sequence models contributed to stabilize the near-native structures relative to other models. In **Figure 3.3d**, we compared the model selection from the native sequence models and the logistic regression models, which combines the break force measured on the native sequence model, the all-alanine model, and the reversed sequence-model. The regression models tied on 16 models and underperformed against the native sequence model force selection on the remaining eight. This underperformance by the regression model appears to be a result of at least two factors; the high performance of the native alone leaves little room for improvement by the regression model in the 80 and over GDT_TS category and also a larger spread of highly stable models across lower break-force models in the all-alanine model simulation (data not shown).

Comparison between using all-alanine and reversed-sequence models against random selection is provided in **Table 3.2**. As shown, using these two models worked better than random for the majority of the targets. These results imply that near-native conformation have features that can increase break force even for all-alanine or reversed sequences. The most frequently observed characteristic of highly stable reverse sequence models and especially all alanine models were the presence of a closed-off hydrophobic core. It was observed that native models were often stabilized by hydrophobic cores, and even destabilized by initial models that opened such cores to solvent in otherwise near-native conformations, so when there was conservation of a hydrophobic core in reverse sequence and all alanine models similar stability would be understandable.

Table 3.2. Comparison of model selection performance with random selection.

	Better	Worse
Native	21	3
All-Alanine	16	8
Reversed-Sequence	15	9
Logistic Regression	20	4

The number of target sets where each model performed better or worse in selecting high quality models of 80 or larger GDT_TS among the top five selections are listed. There are 24 targets in total.

While most native target sets showed strong enrichment, there were some exceptions. The target set T0815, for instance, had two models with a GDT_TS of 80 or above in the top 5 highest peak force model selection, while random selection would produce nearly twice that. However, the mean GDT_TS of the top 5 selected for native models was higher than the mean GDT_TS of the dataset, 80.236 and 73.792 respectively, showing reasonable selectivity. Similarly, the mean GDT_TS of the other two underperforming targets in terms of top 5 selection with GDT_TS of 80 or above also had higher mean GDT_TS than the dataset: T0714 has a mean native selected GDT_TS of 83.582 against the dataset mean of 71.957, and T0762 having a native selected mean GDT_TS of 82.566 against the dataset mean of 72.8544. Thus, even when native peak force elevation did not produce as many models with GDT_TS of 80 or above and random selection, the average quality of a model in the top 5 using native peak selection was higher.

3.4.3 Physical Characteristics

To determine what caused the variation of peak force values among models with high GDT_TS values we explored physical characteristics of these high scoring models. First, we noticed that among many target sets there was a noticeable variation in exposed hydrophobic surfaces across models with similar GDT_TS. To quantify this, we calculated the solvent accessible surface areas of hydrophobic residues, using the Kyte Doolittle hydrophobicity index as reference. For each model we then multiplied the exposed area of each hydrophobic residue of the structure after equilibration by its respective hydrophobicity index value and summed then together. We refer to the resulting value as the initial Hydrophobic Solvent Accessible Surface Area (iHSASA). We compared iHSASA to the break force value of each model with a GDT_TS greater than or equal to 80 in each target set. This showed strong correlation for many of the target sets (**Figure 3.4**).

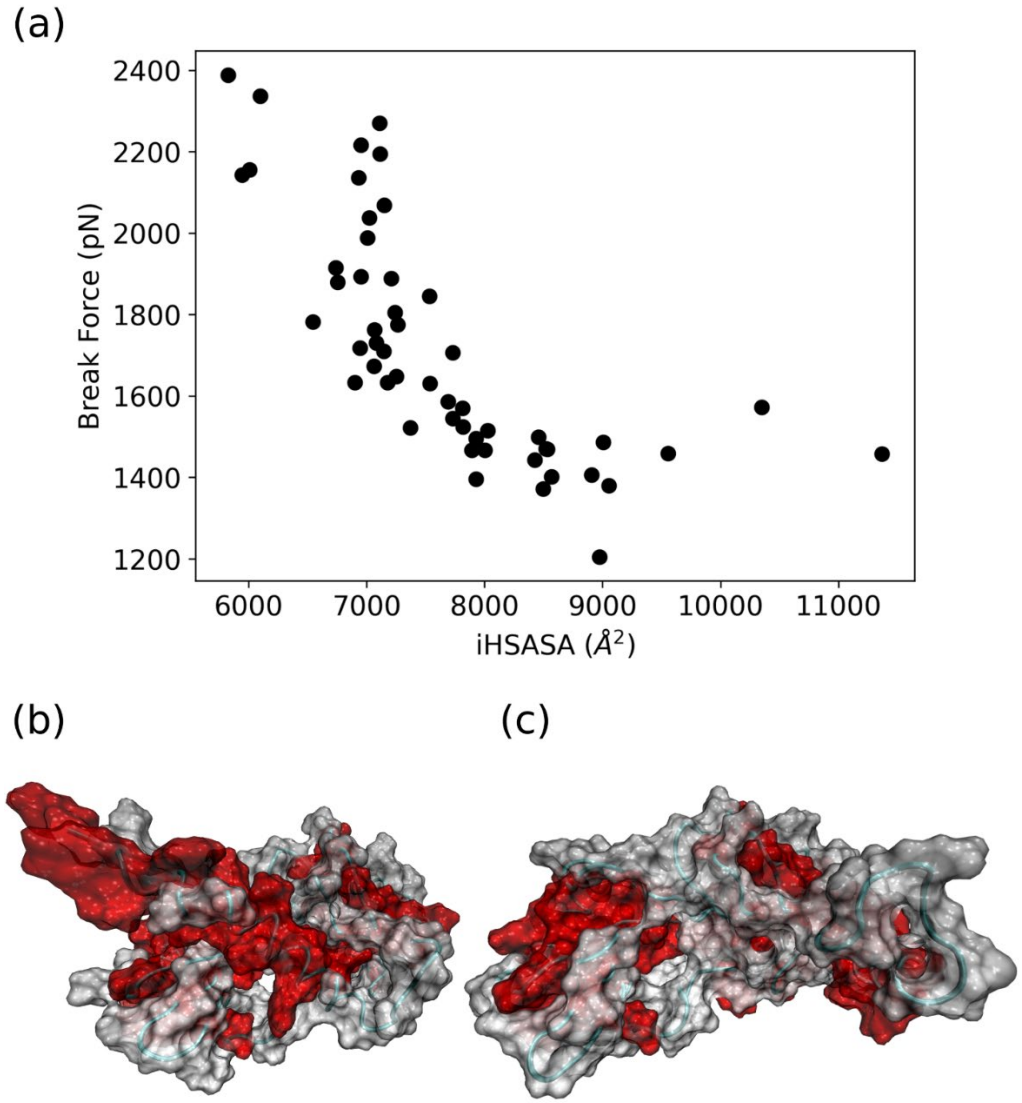

Figure 3.4**:** Initial hydrophobic solvent accessible surface area (iHSASA) and break force. Solvent accessible hydrophobic resides were calculated from the final frame of the equilibration before the pulling production run. (a) comparison of iHSASA and break force for all 54 models in the T0644 target set with GDT_TS of 80 or higher. (b) T0644TS079_5, the model with the lowest break force in the T0644 target set with a GDT_TS greater than 80. T0644TS079_5 had a break force peak of 1205.138 pN, an iHSASA of 8975.3 $Å^2$, and a GDT_TS of 82.62. The backbone is shown in cyan, and a surface representation is shown with hydrophobic surfaces shown in red. (c) T0644TS405_4, the model with the highest break force in the T0644 target set with a GDT_TS greater than 80. T0644TS405_5 had a break force peak of 2388.648 pN, an iHSASA of 5826.2 $Å^2$, and a GDT_TS of 82.45. structures in (a) and (b) are aligned by all residues excluding those in the terminal helix which in a different position in the two models. In (a) the helix if on the face of the structure while in (b) it if folded to the right.

Variation in break force was not completely accounted for by variation in iHSASA. As iHSASA accounts for protein-solvent interaction, we turned our attention towards protein-self interactions. To do this, we calculated the change in electrostatic and conformation energies of the protein by itself before and after the break force peak. This was done by selecting the protein with the atomselect command in VMD and then used the namdenergy plugin. To find shifts in energy from before and after a trajectory's break force we used a variable window ranging from one to ten picoseconds before and after the time of the peak force. a Pearson correlation was then calculated for each energy term for each window for each target set. The window that had the largest Pearson correlation for any single energy term was kept and used for principal component analysis.

We performed principal component analysis (PCA) using the Pearson correlations for the changes in energy for each bonded term and electrostatic energies and the iHSASA Pearson correlation for each target set. PCA appeared to show that there were three categories of characteristics that correlated with the variation in break forces within target sets; iHASA as described above, self-electric interaction energy, and conformational energy (**Figure 3.5**)

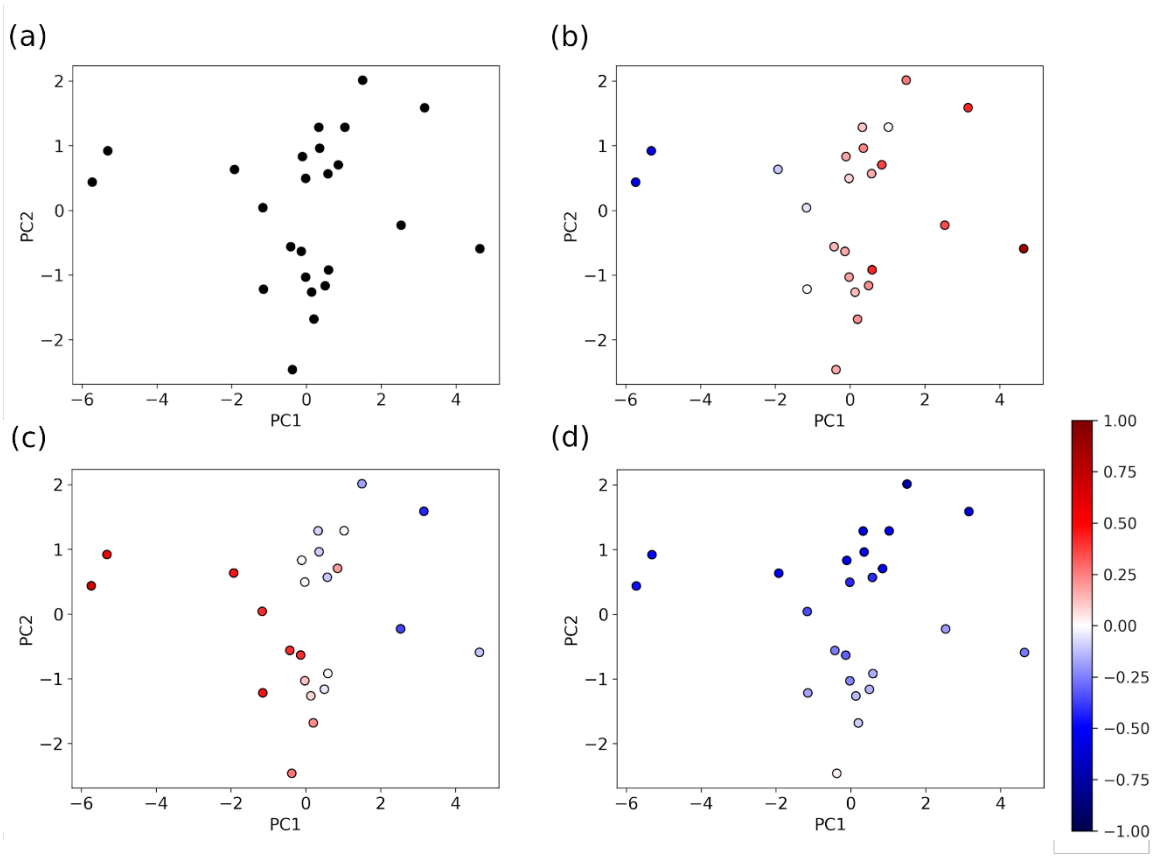

Figure 3.5**:** Principle component analysis of physiochemical properties of models with GDT_TS of 80 or higher. **(a)** Plot of principle component 1(PC1) against principal component 2(PC2) generated from using Pearson correlations of iHSASA and energy components against break force peak magnitude. **(b)** PC1 plotted against PC2 colored according to the Pearson correlation of change in conformational energy against the magnitude of the break force peak of each model with GDT_TS of 80 or greater. **(c)** PC1 plotted against PC2 colored according to the Pearson correlation of change in electrostatic energy against the magnitude of the break force peaks. **(d)** PC1 plotted against PC2 colored according to the Pearson correlation of the initial hydrophobic SASA against the magnitude of the break force peaks. Color legend inset on lower right for Pearson values.

Electric self-interaction energy target sets were observed to have variation of the relative proximity of acidic and basic residues, and models with high break force peaks tended to have these charged residues pull away from one another at or shortly after the time of the break force peak (**Figure 3.6**).

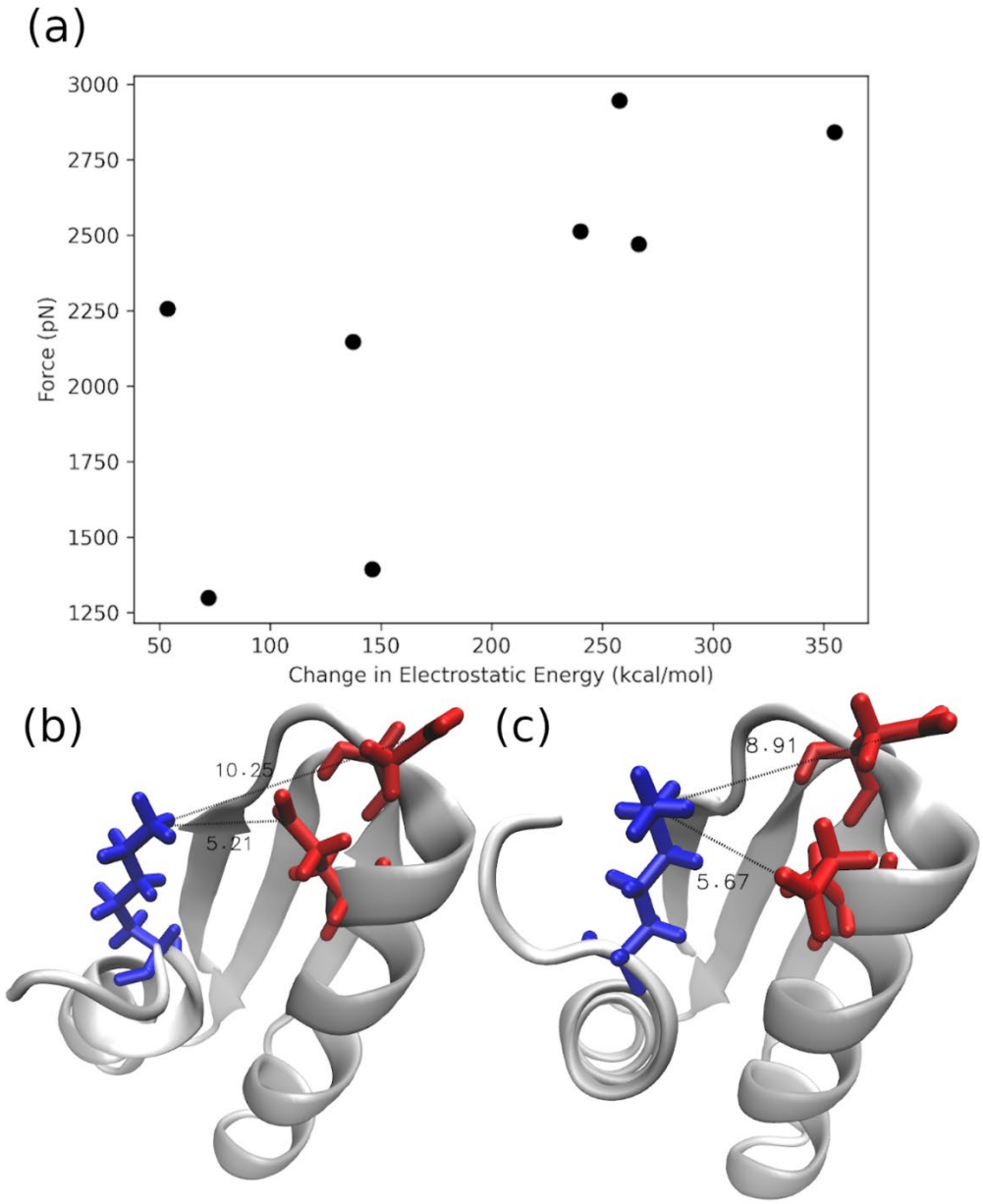

Figure 3.6**:** Change in Electrostatic energy relative to break force.

(a) comparison of Delta Electrostatic energy before and after break force peak time and break force peak magnitude for the 8 model in the T0773 target set with GDT_TS of 80 or higher. (b) T0773TS457_4, the model with the lowest break force in the T0773 target set with a GDT_TS greater than 80. T0773TS457_4 had a break force peak of 1299.229 pN, a change in energy of 72.180 kcal/mol, and a GDT_TS of 86.36. The backbone is shown in cyan, and a surface representation is shown with hydrophobic surfaces shown in red. (c) T0773TS479_1, the model with the highest break force in the T0773 target set with a GDT_TS greater than 80. T0773TS479_1 had a break force peak of 2945.653 pN, a change in energy of 257.710 kcal/mol, and a GDT_TS of 80.52.

Conformation energy target sets were observed to involve significant rearrangement of ordered structures near the time of their respective break force. In some models with high break force peaks this rearrangement was observed to be cause in some cases by strong interactions in different regions of the protein model than the portion that was rearranged (**Figure 3.7**).

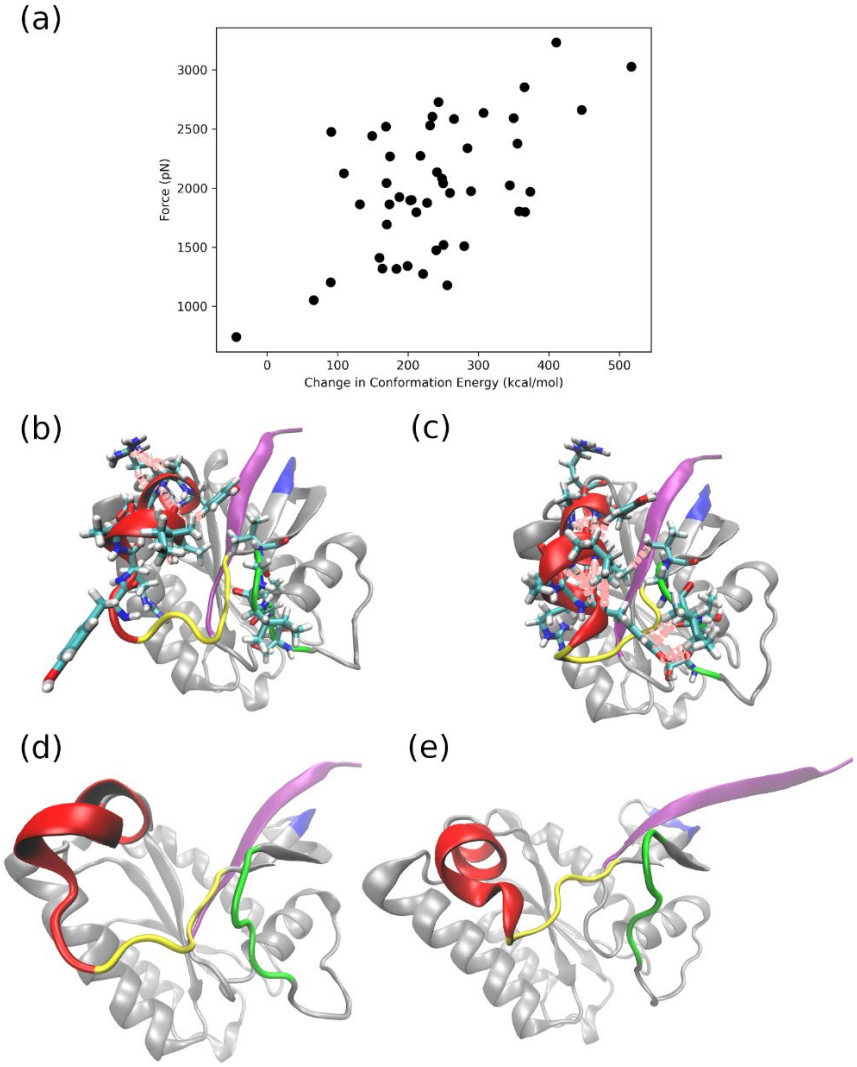

Figure 3.7**:** Peak Force and Conformation energy.

(a) Change in conformation energy of each model in the T0798 target set with GDT_TS of 80 or higher before and after their respective force peak against the magnitude of the break force peak for each model. (b) Structure of T0798TS117_1, break force of 1204.220 pN and GDT_TS of 92.15, after equilibration. Residues 1 through 16 are shown in purple. Residues 52 through 56 are shown in blue Residues 66 through 75 are shown in in red with licorice representations for sidechains. Residues 37 through 39 are shown in green with a a licorice representation for sidechains. Hydrogen bonds are shown in pink. The coil from residue 61 to 65 is highlighted in yellow. (c) Structure of T0798TS430_1, break force of 3231.310 pN and GDT_TS of 91.86, after equilibration. Color code is the same as in (b). (d) T0798TS117_1 at step 100 after significant unfolding. (e) T0798TS430_1 at step 100 after significant unfolding.

The most visually striking difference in the trajectories between the strong (T9798TS430_1) and weak (T0798TS117_1) is the deformation of the strong model's beta sheets (residues 52 through 56), while in the weak model there is a gentle unfolding around a hinge-like coil domain (residue 61 to 65). In the strong model, it was noticed that the helical region 66-75 is closer to the loop region 37 to 39 than in the weaker model. This helix and loop region in the strong model appear to interact with 56 atoms within 5Å or each other and 6 hydrogen bond as defined by the VMD hbonds plugin using a cutoff of 4Å and 60º. In contrast, the weak model had 0 atoms within 5Å and 0 hydrogen bond between the helix and loop. This higher interaction between the helix and loop of the strong model appears to prevent movement of the 61-65 hinge, forcing unfolding through a more energetically steep pathway. In the crystal structures 40jk (reference structure for this casp model), 2f9m, and 2f9n this helix and loop region (66-75 and 37-39) are in the binding pocket for GTP.

3.5 Conclusion

Protein structure prediction remains one of the most challenging problems in modern science. Similarly, protein quality assessment of predicted structures remains an unsolved problem. In this work we hypothesized that the stability of predicted structures would correlate with a prediction's accuracy. We tested this hypothesis by subjecting structure predictions from CASP to tensile stress SMD. Models which required the most force to rearrange tended to have near-native conformations, though not all near-native conformations required larger amounts of force to rearrange.

Multiple physical characteristics were observed among models with high quality. First, SASA of hydrophobic residues inversely correlated with stability. This stabilizing mechanism would come from forcing rearrangement of both protein and water molecules in such a way that water would end up in an unfavorable hydrophobic environment. However, models that already had water in these environments, i.e. in contact with hydrophobic patches on the protein, would not have this kind of environmental change to go through so would not be so resistant to rearrangement.

Target sets that did not show strong correlation between hydrophobic sasa and break force among high GDT_TS models did show a correlation between break force and the change in self-interacting electrostatic energy or conformation energy before and after the break force. Self-

interaction electrostatic energy, electrostatic energy of the protein interacting with itself, was observed in models where basic and acid sidechains were near one another at the start of the production pulling runs and did not pull apart until break force peak occurred, which models with similar GDT_TS but lower break forces started with the same sidechains further apart. Large changes in conformational energy were observed to occur in models which had some interaction short-circuiting the tension and preventing the unfolding of a coil region and forcing the rearrangement of a more structured region of the model, as in **Figure 3.7**.

We have demonstrated that by stressing models in molecular dynamics simulations, high quality models can be detected based on their stability, supporting the initial hypothesis. In addition, we have characterized some of the underlying mechanisms determining the stability of these models.

This work is in revision (Februrary, 2022) in the Journal of Computational Chemistry.

3.6 Acknowledgements

The authors would like to acknowledge Amitava Roy for consulting on the molecular dynamics simulation setup and interpretation. This work was partly supported by the National Institutes of Health (R01GM123055), the National Science Foundation (DMS1614777, CMMI1825941), and the Purdue Institute of Drug Discovery.

3.7 References

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181***,** 223-230.

Pan, K. M., Baldwin, M., Nguyen, J., Gasset, M., Serban, A., Growth, D., Mehlhorn, I., Huang, Z., Fletterick, R. J., Cohen, F. E., et al. (1993). Conversion of alpha-helices into beta-sheets features in the formation of the scrapie prion proteins. *Proc Natl Acad Sci U S A*, *90*, 10962-10966.

Kihara, D., Chen, H,. Yang, Y. D. (2009). Quality assessment of protein structure models. *Curr Protein Pept Sci*, *10*, 216-228.

Uziela, K., Shu, N., Wallner, B., Elofsson, A. (2016). ProQ3: Improved model quality assessments using Rosetta energy terms. *Sci Rep*, *6*, 33509.

Manavalan, B., Lee, J. (2017). SVMQA: support-vector-machine-based protein single-model quality assessment. *Bioinformatics*, *33*, 2496-2503.

Luthy, R., Bowie, J. U., Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*, *356*, 83-85.

Kim, H., Kihara, D. (2014). Detecting local residue environment similarity for recognizing near-native structure models. *Proteins*, *82*, 3255-3272.

Laskowski, R. A., MacArthur, M. W., Moss. D. S., Thornton, J. M. (1993). Procheck - A Program to Check the Stereochemical Quality of Protein Structures. *Journal of Applied Crystallography*, *26*, 283.

Chen, H., Kihara, D. (2008). Estimating quality of template-based protein models by alignment stability. *Proteins*, *71*, 1255-1274.

Wallner, B., Elofsson, A. (2006). Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci*, *15*, 900-913.

Cao, R., Bhattacharya, D., Adhikari, B., Li, J., Cheng, J. (2015). Large-scale model quality assessment for improving protein tertiary structure prediction. *Bioinformatics*, *31*, i116-123.

Wang, Q., Vantasin, K., Xu, D., Shang, Y. (2011). MUFOLD-WQA: A new selective consensus method for quality assessment in protein structure prediction. *Proteins*, *79*, 185-195.

Cao, R., Adhikari, B., Bhattacharya, D., Sun, M., Hou, J., Cheng, J. (2017). QAcon: single model quality assessment using protein structural and contact information with machine learning techniques. *Bioinformatics*, *33*, 586-588.

Uziela, K., Menendez Hurtado, D., Shu, N., Wallner, B., Elofsson, A. (2017). ProQ3D: improved model quality assessments using deep learning. *Bioinformatics*, *33*, 1578-1580.

Shin, W. H., Kang, X., Zhang, J., Kihara, D. (2017). Prediction of Local Quality of Protein Structure Models Considering Spatial Neighbors in Graphical Models. *Sci Rep*, *7*, 40629.

Yang, Y. D., Spratt, P., Chen, H., Park, C., Kihara, D. (2010). Sub-AQUA: real-value quality assessment of protein structure models. *Protein Eng Des Sel*, *23*, 617-632.

Hersam, M. C., Chung, Y. W. (2003). Detecting elusive surface atoms with atomic force microscopy. *Proc Natl Acad Sci U S A*, *100*, 12531-12532.

Li, G., Xi, N., Wang, D. H. (2006). Probing membrane proteins using atomic force microscopy. *J Cell Biochem*, *97*, 1191-1197.

Arredondo, M., Stoytcheva, M., Morales-Reyes, I., Batina, N. (2018). AFM and MFM techniques for enzyme activity imaging and quantification. *Biotechnology & Biotechnological Equipment*, *32*, 1065-1074.

Bujalowski, P. J., Oberhauser, A. F. (2013). Tracking unfolding and refolding reactions of single proteins using atomic force microscopy methods. *Methods*, *60*, 151-160.

Carrion-Vazquez, M., Oberhauser, A. F., Fowler, S. B., Marszalek, P. E., Broedel, S. E., Clarke, J., Fernandez, J. M. (1999). Mechanical and chemical unfolding of a single protein: a comparison. *Proc Natl Acad Sci U S A*, *96*, 694-3699.

Rounsevell, R., Forman, J. R., Clarke, J. (2004). Atomic force microscopy: mechanical unfolding of proteins. *Methods*, *34*, 100-111.

Valbuena, A., Oroz, J., Hervas, R., Vera, A. M., Rodriguez, D., Menendez, M., Sulkowska, J. I., Cieplak, M., Carrion-Vazquez, M. (2009). On the remarkable mechanostability of scaffoldins and the mechanical clamp motif. *Proc Natl Acad Sci U S A*, *106*, 13791-13796.

Izrailev, S., Stepaniants, S., Balsera, M., Oono, Y., Schulten, K. (1997). Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys J*, *72*, 1568-1581.

Lu, H., Isralewitz, B., Krammer, A., Vogel, V., Schulten, K. (1998). Unfolding of titin immunoglobulin domains by steered molecular dynamics simulation. *Biophys J*, *75*, 662-671.

Rief, M., Gautel, M., Oesterhelt, F., Fernandez, J. M., Gaub, H. E. (1997). Reversible unfolding of individual titin immunoglobulin domains by AFM. *Science*, *276*, 1109-1112.

Schonfelder, J., Perez-Jimenez, R., Munoz, V. (2016). A simple two-state protein unfolds mechanically via multiple heterogeneous pathways at single-molecule resolution. *Nat Commun*, *7*, 11777.

Zhuravlev, P. I., Hinczewski, M., Chakrabarti, S., Marqusee, S., Thirumalai, D. (2016). Force-dependent switch in protein unfolding pathways and transition-state movements. *Proc Natl Acad Sci U S A, 113*, E715-724.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A. (2014). Critical assessment of methods of protein structure prediction (CASP)--round x. *Proteins*, *82*, 1-6.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A. (2016). Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins, 84*, 4-14.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., Tramontano, A. (2018). Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins*, *86*, 7-15.

Taylor, T. J., Tai, C. H., Huang, Y. J., Block, J., Bai, H., Kryshtafovych, A., Montelione, G. T., Lee, B. (2014). Definition and classification of evaluation units for CASP10. *Proteins*, *82*, 14-25.

Kinch, L. N., Li, W. , Schaeffer, R. D., Dunbrack, R. L., Monastyrskyy B, Kryshtafovych A, Grishin N. V. (2016). CASP 11 target classification. *Proteins, 84***,** 20-33.

Abriata, L. A., Kinch, L. N., Tamo, G. E., Monastyrskyy, B., Kryshtafovych, A., Dal Peraro, M. (2018).

Definition and classification of evaluation units for tertiary structure prediction in CASP12 facilitated through semi-automated metrics. *Proteins*, *86*, 16-26.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures. *Nucleic Acids Res*, *31*, 3370.

Sulkowska, J. I.; Sulkowski, P.; Szymczak, P.; Cieplak, M. (2008). Stabilizing effect of knots on proteins. *PNAS*, 105, 19714-19719.

Dabrowski-Tumanski, P., Rubach, P., Goundaroulis, D., Dorier, J., Sułkowski, P., Millett, K. C., Rawdon E. J., Stasiak, A., Sulkowska J. I. (2019). KnotProt 2.0: a database of proteins with knots and other entangled structures, *Nucleic Acids Res*, 47, D367–D375.

Webb, B., Sali, A. (2017). Protein Structure Modeling with MODELLER. *Methods Mol Biol*, *1654*, 39-54.

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

Humphrey, W., Dalke, A., Schulten, K. (1996). VMD: visual molecular dynamics. *JMolGraph*, *14*, 33.

CHAPTER 4. PHAGE G AND USP7: CASE STUDIES IN ATOMIC MODEL PRODUCTION AT INTERMEDIATE RESOLUTIONS

4.1 Abstract

I applied molecular dynamics-based approaches to the refinement of protein structures that were determined from cryo-EM density maps. I modeled the atomic structure of the major capsid protein gp27 and the decoration protein gp26 of PhageG to a 6.1Å resolution electron microscopy map. PhageG modeling was performed by mapping the sequences to a presumed homolog structure (Hk97), arranging the subunits into hexamers and trimmers as suggested by mass spectroscopy data, rigid docking to respective map segments, refinement against half maps using MDFF across a range of weights, and then finally refinement to the whole map using the optimized weight. I also modeled the atomic structure of the protein USP7 to an 8.2 Å resolution map. USP7 modeling was done by combining crystalized domains of the whole structure, rigidly docking the composite model to the EM map by hand, and then refining in a similar manner as PhageG, with the added approach of weight scaling to overcome local minima along the relaxation. The USP7 model was further validated by exhibiting a ligand-protein binding pose, determined by glide, which corresponded to enzymatic activity mutation assays.

4.2 Introduction

In chapter 1, we discussed some of the characteristics of electron microscopy data and the importance of atomic modeling. In chapter 2, we illustrated the variability in atomic models deposited with electron microscopy maps. In chapter 3 we demonstrated a method for selecting near native models if many candidate models exist. In this chapter, we will be discussing two case studies involving the fitting of atomic models to two very different electron microscopy maps. In these cases models were initially generated using experimental data such as mass spectrometry, enzymatic assays, and other data, and were then refined to their respective maps in a manner that reduced overfitting a produced a representative structure. The models produced were for the hexomeric complex of the major capsid protein gp27 and the trimeric complex of the decoration protein gp26 of PhageG, and the full structure of the protein Ubiquitin Specific Protease 7 (USP7).

4.2.1 Phage G

The first modelling case study we will discuss is the Phage G capsid. Phage G is notable for its large size of ~180 nm in diameter. This large size, in conjunction with its T=52 icosahedral symmetry and relatively low flexibility make it a promising candidate for electron microscopy data acquisition. Unfortunately, the large size of this structure poses some challenges when modeling. The primary challenge is the low speed of molecular dynamics as the number of atoms in a model is very high, and the size of the map file (~22 Gb) preclude the use of GPU acceleration on site. The sheer scale of the atomic model, when all atoms and all subunits are included, requires the use of GPU hardware to accelerate the molecular dynamics so that it can complete the modeling in an acceptable time frame. The necessity of GPU usage, however, results in a new limitation, and that is available memory. When these simulations were conducted, the GPUs available to us for use had memory on the order of 6 Gb. Recall that the map file alone, which is necessary for refinement, was approximately 22 Gb on its own. The Phage G capsid consists of major capsid proteins which form repeating hexamers. At the corners of these hexamers, where three major capsid subunits meet, there are trimeric decoration proteins. This character repeats around the icosahedral capsid of the virus. By taking advantage of repeating structure, we were able to sample smaller subsections of the structure and use them in refinement, allowing us to infer the structure of the whole capsid from a refined subunit. We will discuss this process in detail later in this chapter.

4.2.2 Ubiquitin Specific Protease 7

On the other end of the spectrum from PhageG is Ubiquitin Specific Protease 7 (USP7). USP7 is a small protein produced by human cells which breaks down ubiquitin. USP7 is associated with several varieties of cancer, though to be cause by higher than acceptable activity. USP7 consists of three major domains; a TRAF-like domain, a catalytic domain which binds ubiquitin, and a hubble domain which consists of ubiquitin-like subunits.

The hubble domain of USP7 was observed to be highly flexible, likely due in part to the links between the ubiquitin-like domains, which results in appreciable conformational heterogeneity in cryo-EM micrographs. This heterogeneity either causes blurring in the reconstructed 3D map or requires the classification of conformations which results in a reduced

data set. Both of these things can cause a decrease in resolution of the EM model, making imaging and modeling of an atomic structure a challenge.

The flexibility of USP7 is not the only characteristic that makes USP7 a challenging protein to image. One obstacle for cryo-EM when used to image USP7 is that USP7 is asymmetrical, meaning that all sides of the molecule must be viewed independently, unlike symmetric molecules. Acquiring enough data for every view can be a problem if the molecule has a biased orientation in the thin film. USP7 is a very small protein for use in cryo-EM, having a mass of only 135 Kd. The typical lower limit of cryo-EM is approximately 50 KDa, below the mass of USP7, but the relatively low mass of USP7 combined with its asymmetry and flexibility all combine to make imaging non-trivial. These factors; high flexibility, asymmetry, and low mass, result not just in lower resolutions, but variable resolutions across the model. These aspects of the final map must be consistent when fitting an atomic model. In this chapter, I will discuss a procedure used to address these issues and produce a reliable atomic model for use in drug screening.

4.3 Modeling of Phage G

Phage G is an unusually large bacteriophage with a capsid around 180nm in diameter(Donelli, 1976; Donellie, 1968). This makes phage G's capsid the largest of all observed tailed phages(G. Donelli, 1975).

Some structural commonalities are shared by all dsDNA, tailed phages. All known phages have major capsid proteins with an HK97 fold(Sun & Serwer, 1997). Many dsDNA, tailed phages contain decoration proteins which locate on the external surface of the capsid, and may play a role in stabilizing the overall structure(Rader et al., 2005). These decoration proteins frequently adopt trimeric oligomerization states(Sae-Ueng et al., 2014; Wang et al., 2018). as in Lambda and TW1(Sathaliyawala et al., 2010). The trimeric decoration proteins position themselves at the 3-fold symmetry axes between three major capsid protein hexamers(Sathaliyawala et al., 2010; Wang et al., 2018).

During the reconstruction a new decoration capsid protein, gp26, was identified, and was proposed to be responsible for the stabilization of the gp27 hexamers. The final map produced for the Phage G capsid used in the following refinement was at an intermediate resolution of 6.2Å, making it too low for automated modeling approaches.

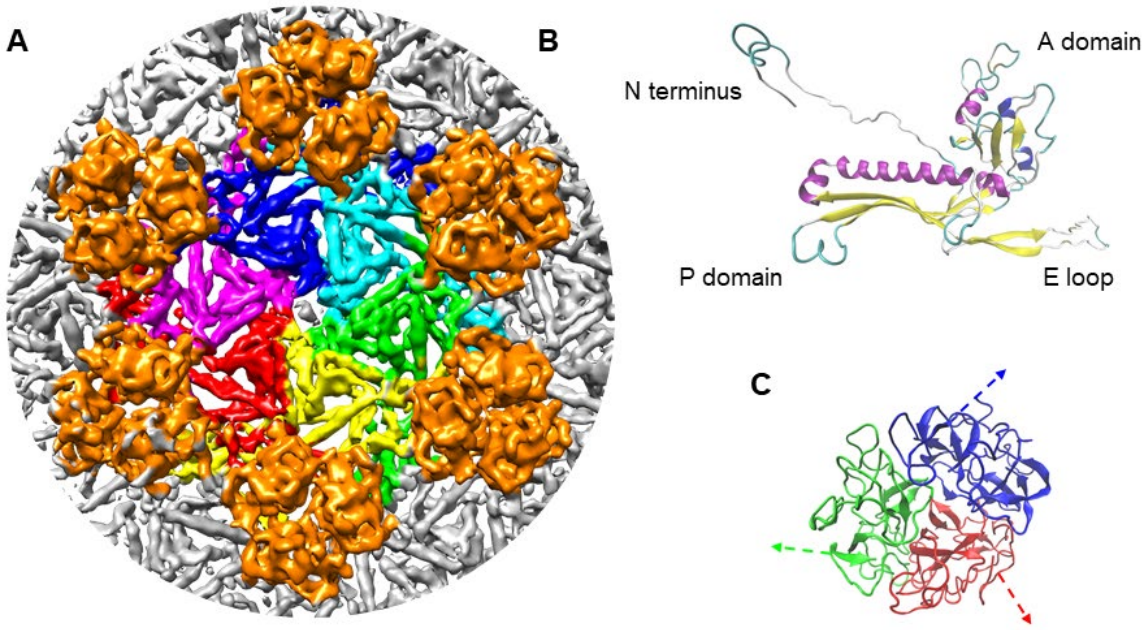

Figure 4.1**:** Phage G major capsid protein hexamer and decoration protein trimer arrangement. A) Phage G's major capsid protein, gp27, and its decoration protein, gp26 arrangement. The structure shows gp27 hexamers and gp26 trimers positioned at the 3-fold axes around the hexamers. B) An overview of phage G's gp27 homology model. All 3 domains of the structure are consistent with the domains described in HK97's major capsid protein including the A domain, E loop, P domain. C) The phage G decoration, gp26, protein oligomerizes into trimers. The first 15 amino acids were omitted in modeling because of their flexibility. Arrows indicate the direction the N-terminus extends in the capsid density contacting neighboring gp27 subunits.

4.3.1 Producing an Initial Atomic Model

The Phage G capsid consists of repeating gp27 hexamers and gp26 trimers organized into an icosahedron with T=52 symmetry (Figure 2A). A structure model of gp27 (Figure 2B), the 282 residue long major capsid protein(Sae-Ueng et al., 2014) was produced using modeller with HK97's major capsid protein, gp5 as a structure template(Serwer et al., 2014). Since all known phages have an HK97 fold, this was the natural choice for a template, despite only having a sequence identity of 23%.

Next, the decoration proteins were modeled. The decoration proteins had been experimentally determined to be gp26 were and organized into trimers on the outside of the capsid. The trimers were observed to be structurally similar in arrangement to those of phage Lambda and TW1's decoration proteins(Chang et al., 2006; Jiang et al., 2006). Phage G's gp26 is 55 residues

longer than Lambda's gpD and 17 residue longer than the gp56 of TW1(Guo & Jiang, 2014). Because TW1's gp56 was closer in length to gp26 it was chosen as a template for generating a structure model.

4.3.2 Fitting and Refinement of Phage G Capsid Proteins to Cryo-EM density

After generating initial atomic models it was time to place the atomic models into the reconstructed density of the Phage G capsid. To do this, the density map was segmented into its hexameric and trimeric subunits using segger in chimera. The initial segmentation had a tendency to segment hexameric subunits into more than one fragment, so manual selection and merging was required. Once each hexamer and trimmer was separated, each segment was saved as its own map. The hexameric and the trimeric models of gp27 and gp26, respectively, were rigidly docked into the density segments using the *collage* program in Situs (Thomas et al., 2007). The cross-correlation coefficients were calculated for each set of docked structures and their respective map segments. By using the cross-correlation values as a filter, I was able to efficiently select usable docking poses. By combining all the acceptable poses I was able to model the entirety of the phage G capsid. However, this model required additional refinement, as rigid docking does not produce an ideal fit to the density data.

To refine the model to the density map I used Molecular Dynamics Flexible Fitting, MDFF(Lander et al., 2013). It was at this point that the hardware limitations of running a model at this scale became apparent. To get around this problem, I took advantage of the capsid's symmetry. By only using one of the icosahedral faces, or facets, I was able to reduce the number of atoms in the model to less than a twentieth of the full capsid, and reduce the size of the map needed from 22Gb to nearly 1Gb.

A possible error one can make when refining an atomic model to a density is to overfit that model. This can occur when using MDFF if the g-scale, map weight, is set too high; however, there is not an ideal map weight to use for all refinements. Typical g-scale weights range from 0.1 to 0.5. While applying too large a scale factor can result in overfitting, applying too small of a weight does not make full use of the experimental data. To select an optimal scale factor, fitting was performed first on the even half-map. Weights of 0.1, 0.3, 0.5, and 0.7 were used over a 1 ns MDFF refinement, after minimization. The final frame of the refinement trajectory was used to calculate a cross-correlation with the odd half-map reconstruction. The scale factor which

produced the highest cross-correlation was taken to be the optimal weight. Using the reciprocal half-map for validation in this way minimizes model fitting to noise from sources such as solvent or misaligned frames. This is because the random noise will not be rewarded in the cross-correlation of the reciprocal map the way it might be using the same map for validation. The weight value which resulted in the highest cross-correlation was then found to be 0.5 and was used for the final refinement using the full reconstructed density. The resulting gp27 and gp26 atomic models are available in the Protein Data Bank (PDB) under the accession ID 6WKK.

4.4 Modeling of USP7

Human ubiquitin-specific protease 7 (USP7) is a particularly potent oncoprotein that has a demonstrated role in many cellular functions (Cheng et al., 2015a; Colland et al., 2009; Faesen, Dirac, et al., 2011a). USP7 has been described as a contributing factor in the progression of numerous diseases including prostate cancer, multiple myeloma, and colon cancer. Due to its oncogenic properties USP7 has generated interest as a drug target(Cheng et al., 2015a; Colland et al., 2009). A major caveat to the development of cancer therapeutics in the USP family is that structure-activity-relationship studies of inhibitor optimization have relied mostly on the conserved catalytic domain, rarely accounting for the unique ancillary domains and how they regulate the activity of each protease(Cheng et al., 2015a; Faesen, Dirac, et al., 2011a; Pfoh et al., 2015a; Zhang et al., 2015a). Although this line of inquire has resulted in several therapeutic leads as inhibitors of USP7, these inhibitors are generally targeted at the active site, or at an adjacent site within the catalytic domain. Consequently, it is anticipated this approach may lead to off-target effects in closely related enzymes.

The ancillary domains surrounding the catalytic domain are thought to provide each USP with an individual ubiquitin-cleavage fingerprint, defining substrate specificity and regulation(Cheng et al., 2015; Song et al., 2008; van der Horst et al., 2006; Zhang et al., 2015). USP7 is a unique member of the USP family due to its 6 ancillary domains, some of which lack significant homology with related enzymes(Zhang et al., 2015). The domain architecture includes a catalytic domain flanked by a TRAF domain, and five HUBL domains (termed 1-5) that are hypothesized to function as a tethered-rheostat activator(Chen et al., 2015). Biochemical and structural studies have uncovered a role for the TRAF domain in protein-protein interactions with substrates such as tumor suppressor p53, ubiquitin ligase MDM2, and viral DNA-binding protein EBNA1(Chauhan et al.,

2012; Cummins & Vogelstein, 2004; Pfoh et al., 2015; Saridakis et al., 2005). HUBL1-5 have been found to extensively control the activity of USP7, truncations of which result in an 80-fold reduction in K_m and k_{cat}(Kon et al., 2010; Li et al., 2004).

Kinetic evaluation of USP7 within the context of the ancillary domains have revealed information suggesting a tethered-rheostat mechanism of intramolecular activation(Kon et al., 2010; Li et al., 2004; Meulmeester et al., 2005). TRAF domain deletion in USP7 results in little change to the K_m and k_{cat}, suggesting that the sole function of the TRAF domain is to provide substrate specificity by recognizing specific peptides(Faesen, Dirac, et al., 2011a; Rougé et al., 2016). However, truncation of HUBL4-5 reduces catalytic efficiency like a HUBL1-5 truncation, suggesting that HUBL1-3 may not be necessary for activation(Faesen, Luna-Vargas, et al., 2011; Rougé et al., 2016).

This difference in catalytic efficiency attributed to the loss of HUBL4-5, remains to be fully understood, but thorough biochemical analysis by Faesen *et al*. defined the requirement of HUBL4-5 in USP7 catalysis as the c-terminal 19 residues(Kon et al., 2010). They hypothesized an interaction between this c-terminal peptide of HUBL domain 5 ($USP7_{1084-1102}$) and a switching loop of the catalytic domain ($USP7_{285-291}$) was responsible for this activation of USP7(Kon et al., 2010). Crystallographic structures solved by Rougé *et al.* built upon this model by revealing a density for the C-terminal peptide of HUBL domain 5 bound within the newly defined *activation cleft* of the catalytic domain, resulting in movement of this switching loop and access of the catalytic residues (Rougé et al., 2016). This study provided the first structural evidence for *in trans* activation of USP7 by the HUBL5 peptide; however, it is important to note the USP7 construct used was an artificial one. HUBL4-5 was tethered directly to the catalytic domain by a flexible glycine-serine linker, in lieu of HUBL1-3(Rougé et al., 2016). Kim *et al*. have recently suggested a model of *in cis* activation by the HUBL1-5 rheostat supported by NMR data, but a structural density of this intramolecular binding has yet to be observed(Colland et al., 2009).

To elicit this degree of activation by the HUBL1-5 rheostat, several groups have described the large degree of conformational flexibility that is supposedly maintained by the rheostat(Colland et al., 2009). Kim, *et al*. have determined the importance of the residues in the α-helical linker connecting the catalytic domain and HUBL1-3 for rheostat function, indicating amino acids that are required for conformational flexibility(Li et al., 2004). Likewise, Pfoh, *et al*. have described the existence of potential hinge regions between HUBL2-3 and HUBL3-4 that are hypothesized to

facilitate this movement of the rheostat to potentially support *in cis* and *in trans* activation of USP7(Cheng et al., 2015).

HUBL1-3 were originally described to serve primarily as a binding platform for protein-protein interactions with DNMT1, UHRF1, and ICP0, as well as the USP7 allosteric activator guanosine 5'- monophosphate synthase (GMPS) (Cheng et al., 2015a; Faesen, Dirac, et al., 2011a; Pfoh et al., 2015a; Zhang et al., 2015a). Although, our kinetic evaluation has uncovered a second role for HUBL1-3 as part of the rheostat which regulates the level of activation that HUBL4-5 impose upon USP7. Our data suggests that this rheostat function is negatively affected – at least in part – by the presence of the TRAF domain, likely indicating an interfacial region where an interaction may occur.

4.5 Electron Microscopy Map Generation of USP7

USP7 was imaged by cryo-TEM using a Volta Phase Plate (VPP) allowing us to visualize USP7 particles, resulting in an 8.2 Å electron density from single particle. According to the averaging and 3D projections obtained from the density, the HUBL domains appear to occupy a range of conformational states suggesting that they remain flexible, moving via the hypothesized HUBL2-3 and HUBL3-4 hinges. Next, I use MDFF to flexible dock the atoms from crystallographic structures into the low-resolution volume to determine the position of domains with a high degree of agreement.

Comparison of the position of the TRAF domain to the crystallographic structures published by Hu, *et al*. (2006) showed the TRAF domain appears to adopt an alternative conformation, supporting the idea that there is a flexible linker between the catalytic domain and TRAF. This more packed conformation of TRAF positions the domain near HUBL1-3, creating a potential interfacial region in the mono-ubiquitin bound state.

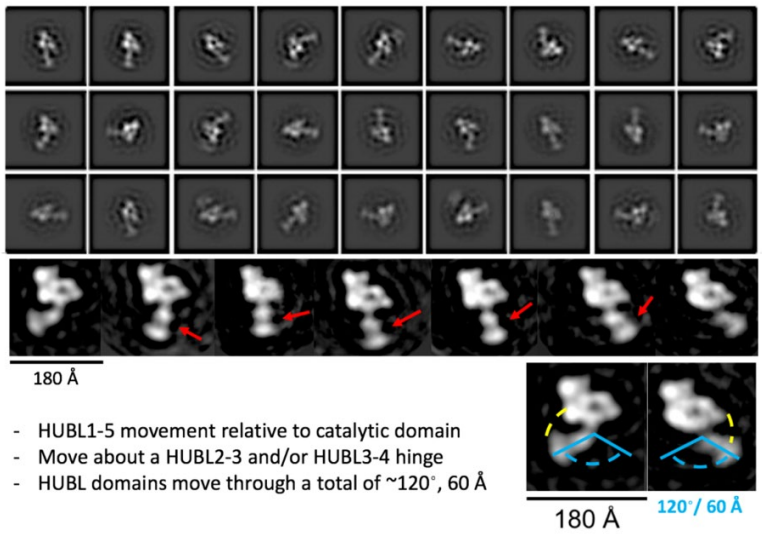

Figure 4.2**:** USP7 displays flexibility under electron microscopy. Multiple conformations were observed among the particles. Class averages display a 120-degree variation on the hubble domain.

4.6 Generation and fitting of an atomic model

To investigate this potential interfacial region I fit an atomic model into this newly determined density. In order to model an atomic structure from a cryo-EM density map I first needed to produce a starting structure. Fortunately, each of the three domains of USP7 had already been independently crystalized, so what I combined these structure end-to-end. The breaks between each domain were mid coil, based on the crystal structures, so refinement would be able to correct small errors in placement of the domains. Using the shape of the EM-density as a guide fragments were placed by hand relative to one another to resemble the overall conformation of the map. All domains were then combined into a single PDB file, and residues and atoms were then renumbered to make a coherent file.

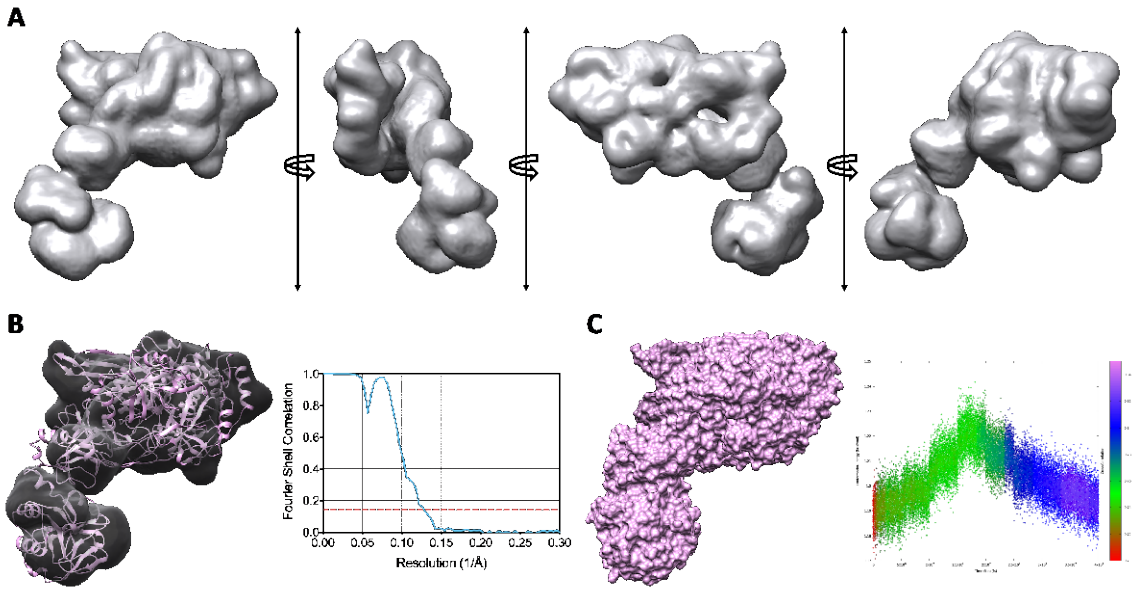

Figure 4.3: USP7 images acquired with VPP on Titan Krios at 130,000x nominal mag. (Left) Representative image of USP7~Ub-PA on pyrene-graphene oxide (pGO) coated gold grids. Red circles are representative particles that would be manually selected for auto-picking templates. Orange scale bar represents ~100 nm distance. (Right) 2D power spectrum of the image to the left with CTF estimation done by CTFfind4.1 through Relion. This is an image of one of the higher signal power spectra. Use of the VPP resulted in most showing an absence of Thon Rings.

In the EM-density, the hubble domain was blurry and relatively low resolution due to the flexibility of that domain. However, the catalytic subunit was clear enough to interpret. Using the catalytic domain's density we were able to place the atomic model into the map. Because of the flexibility of the hubble domain, this proved to be close enough to model. Using MDFF, the catalytic domain aligned further with its relevant density, and the flexible hubble domain curled into place. The initial weight used in the docking was 0.1 and was scaled up by 0.1 every 500ps until reaching a value of 0.5. From here the map density term was gradually reduced to 0, reducing by 0.1 every 500ps, so that the model could move freely. This yielded a better cross coorilation and conformational energy as shown in Figure 4.3c.

4.7 Glide Docking to USP7

Drug docking of the inhibitor shown in figure 4.4 was performed using glide. The inhibitor docked into the binding pocket illustrated below with a free energy of -4.1 kcal/mol. This was one of several orientations in this pocket that resulted in favorable free energies. The search window

for the drug was expanded to encompass the entire protein, yet the drug appeared to have no significant or meaningful interaction elsewhere. The residues highlighted for the interaction below correspond to the TRAF domain (N10, T12, V13, M14, R16), the catalytic domain (N351), and the HUBL3 domain (R730).

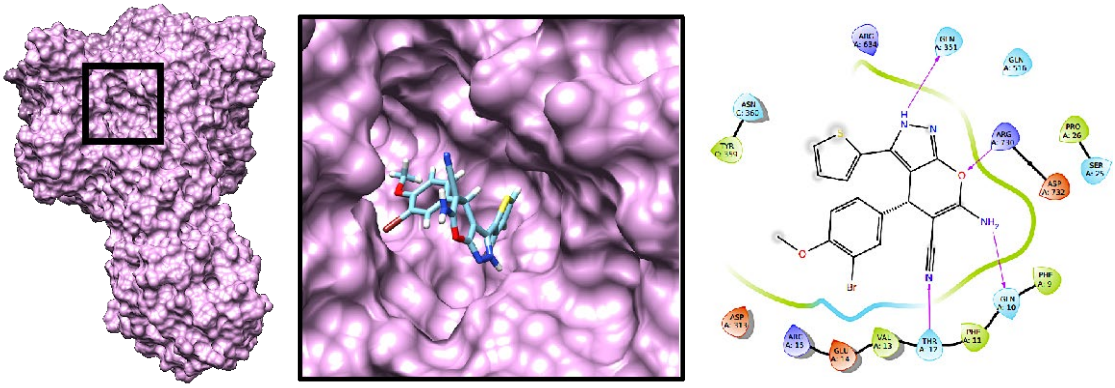

Figure 4.4: TRAF-HUBL1-3 interfacial region and corroborated by drug binding. (Left) Cryo-EM-derived model of USP7 with drug-binding pocket highlighted by black frame. (Middle) Close-up of binding pocket with APII-004 compound modeled in. (Right) Glide docking model of APII-004 with residues Gln10, Thr12, Val13, Glu14, Arg15 from the TRAF domain, Gln351 of the catalytic domain, and Arg730 from H3 domain.

4.8 Glide docking validated by differential scanning fluorimetry (DSF)

Both kinetics experiments and the Glide docking results agree with the differential scanning fluorimetry (DSF) data for USP7. DSF experiments were performed with 1 mg/mL final concentrations of USP7 and saturating concentrations of inhibitor (100 μM) to measure differences in the thermal stability of the enzyme with the inhibitor present. The experiment was performed to observe differences in thermal stability of each of the enzymes domains – both apo and mono-ubiquitin-bound forms – to determine which were affected by the presence of the drug. Only when the TRAF domain and catalytic domain, HUBL domains 1-3 and catalytic domain, or HUBL domains alone were present did the drug bind (Figure 4.5 A & B). The drug had no measurable effect on the catalytic domain alone, or the TRAF domain alone. Although there was a weakly statistically significant decrease in thermal stability for the TRAF-CD construct ($p = 0.046$), the enzymatic activity of this construct did not appear to be affected by the presence of the inhibitor. It is thought that the drug binds the domains to hold the enzyme in an unfavorable conformation that drives down the k_{cat} – acting as a noncompetitive inhibitor.

According to the results summarized in Figure 4.5 C below, both apo and substrate-bound forms of USP7 had significant destabilization affects from the inhibitor. This corroborates the idea that the inhibitor works as a mixed, competitive inhibitor. It is thought that all domains must be present for the drug to coordinate and have the observed effects on catalysis.

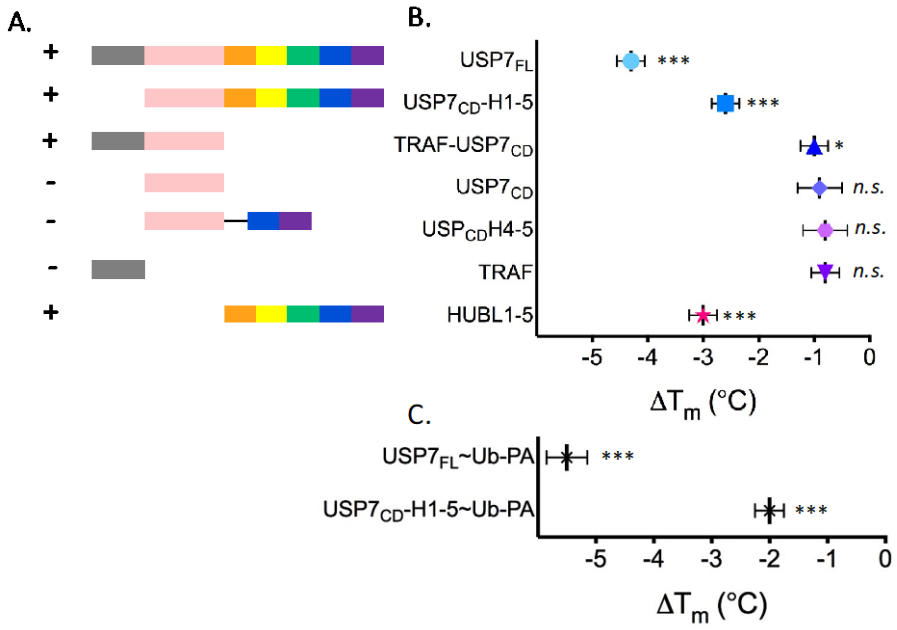

Figure 4.5: DSF results for APII-USP7 pyrazole-derived inhibitors of USP7. (A) Schematic od USP7 constructs used for the experiment; grey = TRAF domain, pink = catalytic domain, orange = H1, yellow = H2, green = H3, blue = H4, purple = H5. (B) Thermal stability changes for the 100 μM inhibitor-treated construct vs the untreated construct. The untreated constructs were in buffer-DMSO composition identical to the inhibitor-treated samples. (C) Ub-PA conjugated samples. For all experiments: points are mean ± SD; n = 3. Statistical significance was calculated with a paired t-test between treated and untreated melting temp: *** = <0.001, ** = <0.01, * = <0.05.

4.9 Conclusion

Both PhageG and USP7 presented their own challenges to atomic modeling. Both maps were of resolutions that would be considered low in contemporary computational circles, but such resolutions will continue to be present when applying cryo-EM to molecules on the edges of what cryo-EM's capabilities. Structures at either edge of single particle cryo-EM's range, or molecules that are highly flexible and therefore present significant conformational heterogeneity on a micrograph, will continue to be challenging for the foreseeable future. Meanwhile, the focus of theorists has seemed to shift towards the highest resolution maps. It is therefore important that methodologies be developed for addressing these edge cases, and the work described here may be

a strong place to start. Questions remain; what is the minimum amount of additional experimental data needed to reliably produce an accurate atomic structure? Does weight scaling reliably result in conformational improvements? What is the lowest resolution that we can hope to use to reliably produce usable structures? It is my opinion that on these edges are where many of the most interesting problems in structural biology can be found.

4.10 References

Chang, J., Weigele, P., King, J., Chiu, W., & Jiang, W. (2006). Cryo-EM asymmetric reconstruction of bacteriophage P22 reveals organization of its DNA packaging and infecting machinery. *Structure*, *14*, 1073-1082.

Chauhan, D., Tian, Z., Nicholson, B., Kumar, K. S., Zhou, B., Carrasco, R., Carrasco, R., Mc Dermott, J. L., Leach, C. A., Fulcinniti, M., Kodrasov M. P., Weinstock, J., Kingsbury, W. D., Hideshima, T., Shah, P. K., Minvielle, S., Altun, M., Kessler B. M., Orlowski, R., Richardson, P. & Anderson, K. C. (2012). A small molecule inhibitor of ubiquitin-specific protease-7 induces apoptosis in multiple myeloma cells and overcomes bortezomib resistance. *Cancer cell*, *22*, 345-358.

Chen, S. T., Okada, M., Nakato, R., Izumi, K., Bando, M., & Shirahige, K. (2015). The deubiquitinating enzyme USP7 regulates androgen receptor activity by modulating its binding to chromatin. *Journal of Biological Chemistry*, *290*, 21713-21723.

Cheng, J., Yang, H., Fang, J., Ma, L., Gong, R., Wang, P., Li, Z. & Xu, Y. (2015). Molecular mechanism for USP7-mediated DNMT1 stabilization by acetylation. *Nature communications*, *6*, 1-11.

Colland, F., Formstecher, E., Jacq, X., Reverdy, C., Planquette, C., Conrath, S., Trouplin, V., Bianchi, J., Aushev, V. N., Camonis, J., Calabrese, A., Borg-Capra, C., Sippl, W., Collura, V., Boissy, G., Jean-Christophe Rain, J., Guedat, P., Delansorne, R. & Daviet, L. (2009). Small-molecule inhibitor of USP7/HAUSP ubiquitin protease stabilizes and activates p53 in cells. *Molecular cancer therapeutics*, *8*(8), 2286-2295.Cummins, J. M., & Vogelstein, B. (2004). HAUSP is required for p53 destabilization. *Cell Cycle*, *3*, 687–690.

Donelli, G., Griso, G., Paoletti, L., & Rebessi, S. (1976). Capsomeric arrangement in the bacteriophage G head. In *Sixth Eur. Reg. Conf. Electron Microsc.(Jerusalem) 2*, 502-503.

Donelli, G. (1968). Isolation of a bacteriophage of exceptional dimensions active in R Megatherium. *Atti della accadamia nazionale dei lincei rendiconti-classe di scienze fisiche-matematiche & naturali*, *44*, 95.

Faesen, A. C., Dirac, A. M., Shanmugham, A., Ovaa, H., Perrakis, A., & Sixma, T. K. (2011). Mechanism of USP7/HAUSP activation by its C-terminal ubiquitin-like domain and allosteric regulation by GMP-synthetase. *Molecular cell*, *44*, 147-159.

Donelli, G., Dore, E., Frontali, C., & Grandolfo, M. E. (1975). Structure and physico-chemical properties of bacteriophage G: III. A homogeneous DNA of molecular weight 5× 108. *Journal of molecular biology*, *94*, 555-565.

Guo, F., & Jiang, W. (2014). Single particle cryo-electron microscopy and 3-D reconstruction of viruses. In *Electron Microscopy* 401-443. Humana Press, Totowa, NJ.

Jiang, W., Chang, J., Jakana, J., Weigele, P., King, J., & Chiu, W. (2006). Structure of epsilon15 bacteriophage reveals genome organization and DNA packaging/injection apparatus. *Nature*, *439*, 612-616.

Kon, N., Kobayashi, Y., Li, M., Brooks, C. L., Ludwig, T., & Gu, W. (2010). Inactivation of HAUSP in vivo modulates p53 function. *Oncogene*, *29*, 1270-1279.

Lander, G. C., Johnson, J. E., Rau, D. C., Potter, C. S., Carragher, B., & Evilevitch, A. (2013). DNA bending-induced phase transition of encapsidated genome in phage λ. *Nucleic acids research*, *41*, 4518-4524.

Li, M., Brooks, C. L., Kon, N., & Gu, W. (2004). A dynamic role of HAUSP in the p53-Mdm2 pathway. *Molecular cell*, *13*, 879-886.

Meulmeester, E., Maurice, M. M., Boutell, C., Teunisse, A. F., Ovaa, H., Abraham, T. E., Dirks, R. W. & Jochemsen, A. G. (2005). Loss of HAUSP-mediated deubiquitination contributes to DNA damage-induced destabilization of Hdmx and Hdm2. *Molecular cell*, *18*, 565-576.

Pfoh, R., Lacdao, I. K., Georges, A. A., Capar, A., Zheng, H., Frappier, L., & Saridakis, V. (2015). Crystal structure of USP7 ubiquitin-like domains with an ICP0 peptide reveals a novel mechanism used by viral and cellular proteins to target USP7. *PLoS pathogens*, *11*, e1004950.

Rader, A. J., Vlad, D. H., & Bahar, I. (2005). Maturation dynamics of bacteriophage HK97 capsid. *Structure*, *13*, 413-421.

Rougé, L., Bainbridge, T. W., Kwok, M., Tong, R., Di Lello, P., Wertz, I. E., Maurer, R., Ernst, J. A. & Murray, J. (2016). Molecular understanding of USP7 substrate recognition and C-terminal activation. *Structure*, *24*, 1335-1345.

Sae-Ueng, U., Liu, T., Catalano, C. E., Huffman, J. B., Homa, F. L., & Evilevitch, A. (2014). Major capsid reinforcement by a minor protein in herpesviruses and phage. *Nucleic acids research*, *42*, 9096-9107.

Saridakis, V., Sheng, Y., Sarkari, F., Holowaty, M. N., Shire, K., Nguyen, T., Zhang, R. G., Liao, J., Weontae, L., Edwards, A. M., Arrowsmith, C. H. & Frappier, L. (2005). Structure of the p53 binding domain of HAUSP/USP7 bound to Epstein-Barr nuclear antigen 1: implications for EBV-mediated immortalization. *Molecular cell*, *18*, 25-36.

Sathaliyawala, T., Islam, M. Z., Li, Q., Fokine, A., Rossmann, M. G. & Rao, V. B. (2010). Functional analysis of the highly antigenic outer capsid protein, Hoc, a virus decoration protein from T4-like bacteriophages. *Molecular microbiology*, *77*, 444-455.

Serwer, P., Wright, E. T., Liu, Z. & Jiang, W. (2014). Length quantization of DNA partially expelled from heads of a bacteriophage T3 mutant. *Virology*, *456*, 157-170.

Song, M. S., Salmena, L., Carracedo, A., Egia, A., Lo-Coco, F., Teruya-Feldstein, J. & Pandolfi, P. P. (2008). The deubiquitinylation and localization of PTEN are regulated by a HAUSP–PML network. *Nature*, *455*(7214), 813-817.

Sun, M., & Serwer, P. (1997). The conformation of DNA packaged in bacteriophage G. *Biophysical journal*, *72*, 958-963.

Thomas, J. A., Hardies, S. C., Rolando, M., Hayes, S. J., Lieman, K., Carroll, C. A., Weintraub, S. T., & Serwer, P. (2007). Complete genomic sequence and mass spectrometric analysis of highly diverse, atypical Bacillus thuringiensis phage 0305ϕ8–36. *Virology*, *368*, 405-421.

Van der Horst, A., de Vries-Smits, A. M., Brenkman, A. B., van Triest, M. H., van den Broek, N., Colland, F., Maurice, M. M., & Burgering, B. M. (2006). FOXO4 transcriptional activity is regulated by monoubiquitination and USP7/HAUSP. *Nature cell biology*, *8*, 1064-1073.

Wang, Z., Hardies, S. C., Fokine, A., Klose, T., Jiang, W., Cho, B. C., & Rossmann, M. G. (2018). Structure of the marine siphovirus TW1: Evolution of capsid-stabilizing proteins and tail spikes. *Structure*, *26*, 238-248.

Zhang, Z. M., Rothbart, S. B., Allison, D. F., Cai, Q., Harrison, J. S., Li, L., Wang, Y., Strahl, B. D., Wang, G. G. & Song, J. (2015). An allosteric interaction links USP7 to deubiquitination and chromatin targeting of UHRF1. *Cell reports*, *12*, 1400-1406.

CHAPTER 5. CONCLUSION

5.1 Remaining Challenges

In the regime is protein modeling, we are far from finished. Though recent developments in machine learning, including the advent of alphafold2, have made considerable gains in the prediction of single chain protein structures, the larger challenge of understanding how proteins interact to perform their functions and how they combine into large complex machines remains an unsolved problem. With the development of electron microscopy, in parallel with the advancement of computational techniques for model generation and quality assessment, we creep ever closer towards interrogating the complexities of living things, beyond the single molecule view.

5.2 Future Work

There is no shortage of new things to be done. The investigation of the motion of single molecules is becoming more routine, added by the development of accurate structure prediction methodologies. The motion of a single molecule allows us to begin to understand what its function might be, and how it carries out that function.

Another exciting goal for us to reach towards is the understanding of complex molecular machines alluded to above. Many years of research have gone into the field of protein-protein interaction, but with increasing computational power available in modern and GPUs, increasingly sophisticated machine learning techniques, and modern advancements in experimental techniques such as single particle cryo-EM, it has become possible to bring to bear an unprecedented and continually increasing level of knowledge and technology to the field.

In closing, among the greatest challenges for researching working in the theoretical and computational realm of this field (or any other field, for that matter) is to reconcile theory with experiment. Theory will always have inaccuracies that are either simplifications there by design, or errors resulting from the unknown or poorly understood. Experiment will always contain eccentricities, noise from known and unknown effects, deviation from what a theorist might call ideal conditions necessary for the experiment to be possible, as well as various other artifacts in

data that may not be common knowledge outside of experimentalist circles. Understanding the limitations of both theory and experiment is necessary to combine the two.

VITA

Lyman Monroe

Education

PhD., Biological Sciences, 2022, Purdue University, West Lafayette, Indiana, USA

B.Sc., Chemistry, 2013, Purdue University, West Lafayette, Indiana, USA

B.Sc., Physics, 2012, Purdue University, West Lafayette, Indiana, USA

Awards

Teaching Academy Graduate Teaching Award, Purdue Teaching Academy, 2019

Biology Best Teaching Assistant Award, Purdue Department of Biological Sciences, 2019

Bilsland Dissertation Fellowship, 2021

PUBLICATIONS

1. **Monroe L.**, Kihara D. (2022) Using Steered Molecular Dynamic Tension for Assessing Quality of Computational Protein Structure Models. *J Comp Chem*. Accepted.

2. Gonzalez B., **Monroe L.**, Li K., Wright E., Walter T., Kihara D., Weintraub S. E., Thomas J. A., Serwer P., Jiang W (2020). Phage G capsid structure studies reveal toroid-like DNA density and lead to correction of host identity as *Lysinibacillus* species. *J Mol Biol*. 4139-4153

3. Xusi Han, Wong-Hee Shin, Charles W. Christoffer, Genki Terashi, **Lyman Monroe**, & Daisuke Kihara (2019). Study of the variability of the native protein structure. *Encyclopedia of Bioinformatics and Computational Biology*, 3: 606-619

4. **Monroe L.**, Terashi G., Kihara D. (2017). Variability of Protein Structure Models from Electron Microscopy. *Structure*. 25, 592-602.

5. Gardner N., **Monroe L.**, Daisuke K., Park C. (2016). Energetic Coupling between Ligand Binding and Dimerization in Escherichia coli Phosphoglycerate Mutase. *Biochemistry*. 55, 1711-23.

6. Hu B.,Zhu X., **Monroe L.**, Bures M., Kihara D. (2014). PL-PatchSurfer: a novel molecular local surface-based method for exploring protein-ligand interactions. *International Journal of Molecular Sciences*. 15, 15122-45.