Protein Structure Determination from Cryo-electron Microscopy Density Maps

Yu-Ruei Wang

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

David Baker, Chair

Frank DiMaio

Liguo Wang

Program Authorized to Offer Degree:

Department of Biochemistry

University of Washington

Abstract

Protein Structure Determination from Cryo-electron Microscopy Density Maps

Yu-Ruei Wang

Chair of the Supervisory Committee:
Professor David Baker
Department of Biochemistry

Single-particle cryo-electron microscopy (cryo-EM) has emerged as a powerful tool in structure determination of macromolecular complexes that are not suitable for crystallographic studies. Recent advances in direct electron detectors and image-processing techniques have allowed cryo-EM to reach near-atomic resolution (3–5 Å) from small particles with low or even no symmetry. However, tools to determine structures from cryo-EM maps have lagged behind. In this dissertation we describe approaches that apply Rosetta high-resolution structure prediction methods to model-building from cryo-EM maps. Using knowledge-based information gathered from proteins of known atomic structure, we describe the development of novel tools for *de novo* model-building from near-atomic resolution cryo-EM maps, and for architecture determination of multi-component macromolecular assemblies from medium- to low-resolution (5–15 Å) cryo-EM maps. We demonstrate that these new computational tools have shown usefulness in determining the structures from newly reconstructed cryo-EM maps, which would otherwise be difficult or impossible for human to do. These methods should enable rapid and reliable structure determination from cryo-EM maps.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express sincere thanks to my Ph.D advisor David Baker for his considerate guidance when I started, generous support when I needed it, and tremendous trust after I gained necessary skills. He has always been around for help and advice. His encouragement, thoughtfulness and enthusiasm have been an inspiration to me. He has dedicated to creating a highly cooperative and intellectually stimulating environment, and to maintaining a fascinating group of scientists, from whom I have learned so much and with whom it was so much fun to work. I sincerely thank him for the opportunity to be in his group.

I would like to thank Frank DiMaio, who previously is my colleague in the Baker lab and later on became my co-advisor on the density-related projects in the last three years of graduate career. All the work I present in this dissertation would not happen without his fundamental pioneer work in Rosetta, and would not be possible without his expertise on computational work. It has been a privilege to work very closely with him over these years, from which I tried to learn and emulate his analytical mindset and his work ethic. He is not only a talented and dedicated scientist, but also a thoughtful mentor, who always respected my ideas, however naïve they might have seemed, and gave me the opportunity to develop my own confidence and independency. I am very grateful to his patience, encouragement, help and friendship.

I would like to acknowledge my collaborators for providing cryo-EM reconstructions to test on the structure determination methods I developed. In this thesis, all the cryo-EM experiments and experimental model validation work were conducted by them. Specifically, in the *de novo* model-building project I would like to thank Yifan Cheng and Xueming Li at UCSF for making the 20S proteasome map available at 4.8 Å resolution, and Mikhail Kudryashev and Marek Basler at University of Basel for making the map of the VipA/VipB complex available, and letting us reveal the target identity in our paper before the publication of their paper that describes the structure, and Edward Egleman at University of Virginia for facilitating the collaboration on the VipA/VipB project; in the multi-component protein assembly project, I

would like to thank Neil Blok, Donyang Tan, Tom Rapoport, and Tom Walz from Harvard University for making the Pex1/Pex6 map available for testing my new method. Stories in my dissertation work could not be more complete without my collaborators. I cannot express how grateful I am to have these wonderful collaborators. I truthfully thank them.

There are many people without whom my graduate career would either not be possible or would not be nearly as exciting and rewarding as it was. I would like to particularly acknowledge James Thompson, who have carefully mentored me since I joined the lab, spent a lot of his already too much occupied time sitting next to me showing me how to do programming, and equipped me with skills to survive in computational projects; Yifan Song, who has taken care of me since the very first day I did my rotation in the lab, been a great support when I messed up things, and given me tremendous help and career advice; Christopher Miles, who patiently and thoughtfully answered all kinds of questions I had about programming and algorithms; Yuan Liu, who first introduced me programming/scripting languages while I was doing my rotation and very patiently helped me solve any computational or mathematical problems I have encountered. They have been great colleagues and friends, from whom I learned a lot professionally and personally. Working with them was the greatest time I had these years. After they left the lab, I miss the period of time when they were around every single day.

I would like to thank my best friends David Kim and Scott Delbecq, from whom I have received enormous support lab-wise and life-wise. It has been a privilege to have loyal friends like them, with whom I have shared all those ups and downs throughout my graduate career. While writing this, I couldn't help but thinking those numerous moments I spent with them, our daily coffee breaks, lunch breaks, fishing trips, and beer nights to name a few. It would not be possible for me to get through the graduate school without their friendships.

I would like to thank folks from the structure prediction group in the Baker lab: Hahnbeom Park and Sergey Ovchinnikov for making my life in the lab so much fun and exciting. I miss all those stimulating and sometimes nonsense discussions during our daily lunchtime.

I would like to thank all the people from whom I received tremendous help in the Baker lab at one time or another during my graduate career. I would like to particularly acknowledge Will Sheffler, Robert Vernon, Mike Tyka, Nikolaos Sgourakis, TJ Brunette, Justin Siegel, Firas

Khatib, Jeremy Mills, Lucas Nivon, Liz Kellogg, Rocco Moretti, Erik Procko, Tim Whitehead, Yakov Kipnis, Christy Tinberg, Vladimir Yarov-Yarovoy, Darwin Alonso, Patrick Conway, Per Jr. Gresien, Lei Shi, Possu Huang, Florian Richter, Vikram Mulligan, Aaron Chevalier, Austin Day, Jiayi Dou, Matt Bick, Chris King, David La, Peiling Lu, Gabe Rocklin, Daneil Silva, Gaurav Bhardwaj, Fabio Parmeggiani, Ryan Pavlovicz, Ratika Krishnamurty, Takahiro Kosugi and Yu-Ru Lin.

I would like to thank the members of my dissertation committee: Ning Zheng, Wenqing Xu and Liguo Wang for their helpful advice on my work. More importantly, when I was confused about which career path to pursue, all of them have given me invaluable advice, sharing with me their experience from the perspectives of successful structural biologists.

I would like to thank my parents. Being the sole breadwinner with a poor graduate student salary, having their financial support was a huge contributor to any success I may have had. I also greatly thank them for letting me decide my own path of career, and for being supportive with it afterward although it was not what they had expected me to do. I am really grateful for their unconditional love.

Finally, I would like to thank my wife Yu-Ting, who has been my love and my partner since we met in the first year of college. She gave up her comfortable life and a career she might have had originally in Taiwan, turning down graduate school offers from other states in the US, and came all the way to Seattle for nothing but because she wanted (or I needed her) to be with me. She has been extremely supportive about what I would like to pursue in my life, and hits me with her words of wisdom whenever I am strayed. She keeps me alive by making me eat and sleep, and makes me laugh by saying something silly from time to time. Above all, she has brought the most beautiful and wonderful creatures into the world – our two daughters, Sophie and Claire. The emotional support and stable family environment she provided has been invaluable to me and my work. If there is any success I have had or will have in life, it must be because of Yu-Ting. For that, I consider myself as the luckiest and happiest man on the face of the earth.

DEDICATION

To my wife Yu-Ting, and my daughters Sophie and Claire

Chapter 1. INTRODUCTION

Cellular processes often involve large macromolecular assemblies, composed of multi-component proteins and other molecules in precise arrangements, to drive the chemical reactions essential to life. To achieve mechanistically understanding of these cellular processes requires protein structural information at atomic-level accuracy. X-ray crystallography has been the method of choice to provide the resolving power that allows positions of each atom within a macromolecule to be assigned accurately. The prerequisite for applying X-ray crystallography to a protein of interest is to obtain crystals. Yet, getting crystals from macromolecular assemblies has proven to be very difficult mostly owing to sample heterogeneity and the intrinsically transient interactions between constituting proteins. Alternatively, single-particle cryogenic electron microscopy (cryo-EM) has emerged to be a powerful method for determining structures of such macromolecular assemblies. Single-particle cryo-EM offers several advantages over X-ray crystallography: (1) it requires no crystals, (2) it does not require large amount of sample with high concentration, and (3) it allows some extent of heterogeneity from sample preparation. Single-particle cryo-EM determines the structure of a molecule by computationally assembling a 3-D electron density map from 2-D profile images of the molecule at different orientations. To achieve the best resolution it typically requires thousands to millions of particle images of the same conformation. However, despite some exceptions of highly symmetric samples such as viruses, the resolution achievable using single-particle cryo-EM typically resides at around 10 Å scale, as of 2012, before the advent of direct electron detectors.

1.1 SINGLE-PARTICLE CRYO-ELECTRON MICROSCOPY AT CRYSTALLOGRAPHIC RESOLUTION

One of the limiting factors for single-particle cryo-EM to reach high-resolution is the signal-to-noise ratio (SNR) in the images. In order to reduce the effects of radiation damage, electron exposures must be extremely low to preserve high-resolution information, which results in images with low signal-to-noise ratio. Almost two decades ago Henderson [1] and Glaeser [2] used theoretically calculations suggested that under idealized conditions where low-dose images produce perfect contrast single-particle cryo-EM should potentially reach near-atomic resolution (3–5 Å) using only ~40,000 particles. However, in practice the resolution achieved by cryo-EM is significantly lower than the predicted physical limit. The problem has generally been attributed to be the blurry of images caused by the electron beam-induced sample movement and charging. This was considered as insurmountable problem until the advent of direct electron detectors.

The introduction of direct electron detectors enables two things: (1) it allows a near noise-free data collection, thus improves the signal-to-ratio enormously, and (2) it allows the fast readout so that the beam-induced movement can be compensated for computationally. The image recording device (i.e. SNR) is now the limiting factor to get cryo-EM to reach high-resolution. As a result, with this technical advance the resolution achievable by single-particle cryo-EM has finally fulfilled what it was predicted two decades ago – near-atomic resolution (3–5Å). Although further improvement on sample preparation has shown that with the current detectors and image processing algorithms cryo-EM is possible to reach a sub-3Å resolution [3,4], in general high-resolution cryo-EM typically yields 3–5 Å resolution density maps. With near-atomic resolution it allows *de novo* model-building for systems with no homologues of known structure. Since then several near-atomic structures of systems with low or even no symmetry

have been revealed. The great leap forward of resolution, which is often referred as “the resolution revolution”, has opened a new era for structural biology, enabling scientists to visualize how macromolecular machines that are impossible to study use X-ray crystallography work at molecular level. However, the abundance of electron density maps of new and large structures also brings up new problems for structure determination in such maps, for example, the model-building problem.

1.2 MODEL-BUILDING PROBLEMS FROM CRYO-EM DENSITY MAPS

Model-building is a process in structure determination, in which coordinates of each atom are obtained by interpreting electron density maps. It is a key process, and yet a time-consuming and tedious task even for the experts. For systems where there are structural homologues available, structure determination from cryo-EM maps typically starts with docking of the detectable atomic models from either X-ray crystallography or NMR to build a pseudo-atomic model. Building a pseudo-atomic model is especially informative when the resolution given by a density map can only allow the molecular envelope to be resolved. Despite lack of all interactions at an atomic level, a pseudo-atomic model can provide detailed understand of the arrangement of the constituting component protein. However, docking of multi-component protein complexes can be a formidable challenge as determination of the positions and orientations for all components can be difficult, especially when the system gets larger. Current structure determination tools have shown to be effective for systems containing no greater than 7 components proteins. With the abundance of cryo-EM maps for large macromolecular assemblies, which sometimes can be composed of 50 chains, these tools apparently to be

insufficient. A new tool is needed to facilitate structure determination of multi-component protein complexes from medium- to low- resolution cryo-EM maps.

When there are no structural homologues detectable, *de novo* model-building must be carried out. A typically approach for both manual and computational model-building is to build backbone trace first, and then assign sequence based on side-chain density. For near-atomic resolution (3–5 Å) electron density maps, although tracing backbone is rather straightforward, assigning sequence relying on side-chain density alone has shown to be difficult and time-consuming. A 300-residue protein would take an expert a month to build the model from 3–4 Å resolution. For resolution at around 4–5 Å, where one can barely see strand separation in the density maps, sometimes it is not even feasible to confidently build a model. Current structure determination tools adapted from X-ray crystallography general fail to build models accurately from cryo-EM maps at this range of resolution. A new tool is needed for *de novo* model-building from near-atomic resolution cryo-EM maps.

Here, we propose to approach these two problems through incorporating Rosetta high-accuracy structure prediction tools, which employ knowledge-based information derived from proteins of known atomic structure, with cryo-EM maps.

1.3 ATOMIC-ACCURACY STRUCTURAL MODELING USING ROSETTA

Atomic-accuracy protein structure prediction is now achievable using Rosetta. Rosetta predicts protein structures by using a library of per-residue specific backbone conformations (fragments) to explore the structure space given a protein sequence, and applying a physical realistic energy function to evaluate and guide the conformational search. The principal idea underlying the fragment approach is based on experimental observations showing that segments

of consecutive residues have preferences for certain backbone conformations. The conformational space of a protein is thus could be largely covered by having representative conformations of each local sequence segment. Rosetta was the first computational approach to successfully obtain atomic models of small proteins (< 100 residues) from sequence information alone [5]. However, for larger proteins it becomes very difficult to achieve the same success because the conformational space to search increases exponentially by protein length. More information is required to efficiently guide conformation sampling. For this respect, Rosetta starts to incorporate experimental data, for example, NMR data and density data from X-ray and cryo-EM. Using backbone-only NMR data to reduce the conformational space to search, Rosetta has shown capability to determine proteins up to 200 residues, often with an RMSD of 2–3 Å over the entire protein [6]. What is more remarkable is that Rosetta is able to get the side-chain packing correct (within 2 Å all-atom RMSD) in the core region of a protein. With this predicting power the Baker lab started to use Rosetta to design proteins, which requires atomic-accuracy structural modeling. Thus far, the Baker lab has shown several successes on designing protein structures with atomic-level accuracy, illuminating the path of using Rosetta high-accuracy structural modeling to derive atomic information from low-resolution data. That is, in the respect of high-resolution cryo-EM where 3–5 Å resolution data are mostly reachable, Rosetta modeling can serve as a bridge to bring out the high-resolution information from near-atomic resolution data.

1.4 BRIDGE THE RESOLUTION GAP USING ROSETTA

With the advance of new detectors, single-particle cryo-EM now is able to achieve near-atomic resolution, determining several structures of macromolecules that would otherwise

impossible to be studied by other methods. However, structure determination tools have been lagged behind. Specifically, *de novo* model-building tools aiming for near-atomic resolution density data are not available. Here, I would like to incorporate Rosetta structure prediction tools with cryo-EM data, developing a Rosetta-based *de novo* model-building tool. Moreover, with Rosetta we can bridge the resolution gap, that is, using knowledge we have learned from proteins of known structure to infer high-resolution information from cryo-EM data. However, unlike NMR data, which provide residue-residue contact information, incorporating density data with Rosetta model tools is not that straight forward. Currently, Rosetta utilizes density data to model a protein it in two ways: (1) using the density map to filter Rosetta *de novo* models after optimal rotation/translation search [7], and (2) using density-guided sampling to refine homology models [8]. However, the first approach only used density information for scoring but not used density to guide conformation sampling, which obviously is the primarily bottleneck of Rosetta method. The second approach requires a starting model. Unlike NMR data type, which gives distance restraints for residue-residue pairs, the density map serves as one Cartesian space restraint for a protein. In order to get properly sampling guidance, the standard Rosetta fold-from-extended-chain protocol hence has to constantly align the model back to the density map while doing folding simulation. This is computationally intractable. To be more effectively using Rosetta or concepts from structure prediction for *de novo* model-building require a new computational scheme. To develop such computational framework to use density data is the major aim of this dissertation.

1.5 DISSERTATION OUTLINE

In this dissertation we describe two novel model-building approaches aiming for determining structures from cryo-EM maps with near-atomic resolution and medium-to-low resolution, respectively. In **Chapter 2**, we describe a new *de novo* model-building approach aiming to determine atomic structures from near-atomic resolution cryo-EM maps. We then applied the method to a previously unsolved cryo-EM map containing the 660-residue contractile sheath protein complex of type VI secretion system from *Vibrio cholera*. In **Chapter 3**, we describe a novel approach adapted from the *de novo* model-building algorithm in the previous chapter to determine structures for multi-component macromolecular assemblies from 5–15 Å cryo-EM maps. We show an application of the method on determining the structure of the peroxisomal Pex1/Pex6 ATPase complex from the 6.5 Å cryo-EM map.

Chapter 2. *DE NOVO* PROTEIN STRUCTURE DETERMINATION

2.1 INTRODUCTION

Model-building is a key step in macromolecular structure determination. While most atomic-resolution structures are solved using X-ray crystallography, single-particle electron cryo-electron microscopy (cryo-EM) has emerged as a powerful tool in determining electron density maps of large and high-symmetry particles to near-atomic resolution (~3–5 Å) [9-13]. Recent advances in electron detector and image-processing techniques even allow it to reach these resolutions from smaller particles with low or even no symmetry [14-20]. Despite these developments in improving image quality and image reconstruction algorithms, little progress has been made in *de novo* model-building into near-atomic resolution cryo-EM density maps.

Structural interpretation of cryo-EM maps typically starts with fitting an atomic X-ray or NMR structure – or a homology model derived from one – into the map [21-23]. Previous work has shown that atomic resolution models are achievable from near-atomic-resolution cryo-EM density, starting from a homologous structure of the correct topology [24]. However, when there are no previously solved structures of homologous proteins, *de novo* model-building must be carried out. Currently, such structure determination requires manually building a backbone model (or *tracing* the backbone) into density and assigning sequence [14-17]. At near-atomic resolution, many detailed structural features are distinguishable in a density map: the pitch of helices, separation of individual β-strands in sheets, and even some bulky side-chains [25]. While tracing the backbone into density at this resolution is often straightforward, manually assigning sequence remains time consuming and error-prone.

Automated protein model-building tools developed for X-ray crystallography [26-28] are widely used in structure determination from maps with resolution better than 3 Å. These methods

generally separate backbone tracing and side-chain assignment, with density features largely guiding side-chain identification. Consequently, at resolutions worse than 3 Å, where much of the side-chain density is indiscernible, these approaches generally fail. Several *de novo* model-building methods targeted to cryo-EM have been developed for maps with resolution ranging from near-atomic (3-5 Å) to medium resolution (5-10 Å) [29-31]. These methods primarily focus on topology determination through estimating optimal connections of either secondary structure elements or Cα atoms assigned in the density map. Although these methods are quite powerful in identifying topology given a map, they have poor recovery, often <50%, of correct sequence registration [29,30].

In this chapter, we describe a novel *de novo* model-building approach for cryo-EM maps at 3–5 Å resolution. Our approach combines sequence-derived backbone conformations with side-chain fit-to-density in order to correctly assign sequence into the maps. On a benchmark set of nine experimental cryo-EM maps at near-atomic resolution with previously determined structure and a previously unsolved map for the 660-residue hetero-dimeric sheath of the type VI secretion system from *Vibrio cholera*, we show that high-accuracy models can be obtained without prior knowledge of detectable structural homologues. Our method should streamline the protein structure determination process from cryo-EM maps at near-atomic resolution.

2.2 METHODS

2.2.1 *Map preparation*

For all benchmark targets, the cryo-EM maps were segmented into single-subunit guided by native structures using UCSF Chimera's "zone" tool at a distance of 4 Å. The cryo-EM maps and the corresponding deposited native structures used are listed in Table 1.

2.2.2 *Matching fragments into density*

For each 9-residue window of amino-acid sequence, we used standard Rosetta fragment picker[32] to collect libraries of representative backbone conformations from proteins of known structure based on similar sequence and predicted secondary structure. Fragments from proteins of known structure homology (PSI- BLAST e-value < 0.05) to the benchmark proteins were excluded while constructing the fragment libraries. A sequence-derived fragment library given a protein sequence was curated with 25 backbone conformations per sequence position.

We used backbone information given a fragment to first identify the likely locations and orientations in the density map using 6-D translation/rotation search. The density map was subdivided into a regular three-dimensional grid and the search fragment was translated to each grid point in turn. At each grid point, the spherical harmonic decomposition of model and map density was used to rapidly search all rotations of a backbone fragment against regions of experimental density [33]. To further speed up matching, this rotation search was only carried out at regions of high density (mean density Z score > 1 in a sphere around each grid point). For each fragment, the top 2000 placements were collected using the approximated correlation score

between backbone configurations and density [8], giving 50000 candidate placements per sequence position.

Side-chain information was then used to further refine the placements and identify the most likely placements where both backbone and physically realistic side-chain conformations have good agreement to the local density. At each sequence position, the 50000 backbone placements were then further refined with rotamer optimization and rigid-body minimization using Rosetta. After this optimization, 2500 placements for each sequence position are selected for each sequence position using the Rosetta full-atom density correlation score [8]. These fragments were clustered (with 2 Å RMSd cluster radius), and the lowest density score member was taken from each cluster. Finally, if there were more than 50 clusters, only 50 models were carried over to model assembly.

2.2.3 *Evaluating compatible set of fragments*

From these fragment placements, we next want to select a mutually compatible set. We assessed this compatibility using a scoring function with four terms:

$$score_{total}(\boldsymbol{F}) = w_{dens} \sum_{f_i \in \boldsymbol{F}} score_{dens}(f_i) + w_{overlap} \sum_{f_i, f_j \in \boldsymbol{F}} score_{overlap}(f_i, f_j)$$
$$+ w_{close} \sum_{f_i, f_j \in \boldsymbol{F}} score_{close}(f_i, f_j) + w_{clash} \sum_{f_i, f_j \in \boldsymbol{F}} sc_{clash}(f_i, f_j)$$

The term $score_{dens}$ measures the fit of a fragment to density, and is based on the density correlation between the fragment after side-chain rotamer optimization and the experimental map

[8]. The other three terms, $score_{overlap}$, $score_{close}$, and $score_{clash}$, assess the compatibility of a pair of fragments:

$$score_{overlap}(f_i, f_j) = \sum_{\substack{C\alpha_i, C\alpha_j \in f_i, f_j \\ res(C\alpha_i)=res(C\alpha_j)}} \frac{2}{1+\exp\left(-8\cdot\left(\left\|C\alpha_i - C\alpha_j\right\| - 3\right)\right)} - 1$$

$$score_{close}(f_i, f_j) = \begin{cases} -1, & \left\|f_i - f_j\right\| < maxdist(|i-j|) \\ 1, & \left\|f_i - f_j\right\| \geq maxdist(|i-j|) \end{cases}$$

$$score_{clash}(f_j, f_j) = \sum_{\substack{C\alpha_i, C\alpha_j \in f_i f_j \\ |res(C\alpha_i) - res(C\alpha_j)| \geq 3}} \begin{cases} 1, & \left\|C\alpha_i - C\alpha_j\right\| \leq 2.0 \\ 0, & \left\|C\alpha_i - C\alpha_j\right\| > 2.0 \end{cases}$$

The term $score_{overlap}$ gives a bonus to pairs of fragments that place the same residue nearby, with a larger bonus for more overlapping residues; $score_{close}$ penalizes pairs of fragments that put residues close in the sequence further apart than *maxdist,* the maximum observed distance of residues at a particular sequence separation; finally, $score_{clash}$ penalizes fragment pairs with two residues occupying the same place.

2.2.4 *Simulated annealing Monte Carlo sampling*

Simulated annealing Monte Carlo sampling (SA-MC) was used to search for a set of fragments that are mutually compatible. Each sequence position is initially assigned one random (out of 50 possible) fragment placements or a "null placement" which handles the possibility that there may be no good fragment placements at a particular sequence position. Each step in the trajectory replaces the fragment at a particular position subject to the Metropolis criterion using the $score_{total}$. For pairwise score terms, precomputing all pairwise scores allows for fast score evaluation of a fragment assignment. To control precision versus coverage, we assign a density

score, $dens_{null}$, to the null placement; lower values lead to reduced coverage but more precision in fragment placement. All experiments in the paper used $dens_{null} = -150$. Finally, simulated annealing was carried out by slowly reducing the temperature from 500 to 1 in 200 increments with 5000 moves each. Total runtime was approximately 10 minutes per trajectory.

2.2.5 *Iterative assembly of models*

In many cases, there are a few similar fragment assignments with roughly equivalent scores. To identify all of these alternate models, we run 2000 SA-MC trajectories. We use this ensemble to find a high-confidence partial model to carry into the next round. From the lowest scoring 5% of trajectories, we assemble a backbone model by identifying all residues that are placed in the same position (with 3 Å tolerance) and taking the average backbone coordinate at each residue position. If fewer than 70% of backbone residues have been assigned, we iterate fragment matching and SA-MC assembly. The subsequent iteration of fragment matching was carried out by first masking out density which has been assigned in the backbone model from the previous iteration, then placing fragments only from sequence not yet assigned into density.

2.2.6 *Completing models with RosettaCM*

The final step in our approach is to rebuild the final set of unassigned residue positions in the partial models using RosettaCM[34], a comparative modeling method. Unassigned sequence positions in each partial model are rebuilt in the same manner as unaligned regions in comparative modeling. RosettaCM is guided by the cryo-EM density maps in completing partial models, by adding a score term assessing agreement of a model to experimental density during

model-building and refinement with Rosetta's physically realistic all-atom energy function[35]. With 20 percent score cut using Rosetta Score12, 10 lowest density score models were selected from ~1000 models generated by RosettaCM.

2.2.7 *Model-building with Buccaneer*

Model-building with *Buccaneer*[26] used the same segmented maps and was provided the same sequences as was our approach. Reflection data was computed from the cryo-EM maps using *phenix.map_to_structure_factors* [36]. SIGF was set to F/10 for all reflections using *SFTOOLS* from the *CCP4 Program Suite v6.4.0* [37]. A map padding of 5 Å was added to the border to ensure no effects from periodicity on model-building. We ran *Buccaneer* from *the CCP4 Program Suite v6.4.0* with mostly default setting: five cycles of building/refinement were carried out using the correlation target function during model-building, with "use R-free" disabled.

2.3 RESULTS

2.3.1 *Overview of the protocol*

Our approach for *de novo* interpretation of near-atomic-resolution density maps, outlined in Figure 2.1, consists of three steps: (1) matching sequence-based local backbone conformations into the density map; (2) identification of a maximally consistent subset of these fragment matches and assembly into a partial model, and (3) completion of the partial model using density-guided sampling and all-atom refinement.

In the first step, for overlapping 9-residue windows of amino-acid sequence, we identify segments (or *fragments*) of solved protein structures with similar local sequences and predicted secondary structures [32]; this is analogous to the fragments used in Rosetta *de novo* structure prediction [38]. For each fragment, a translation/rotation search identifies placements with good map agreement (after optimizing side-chain conformations). After this first step, only small subsets of these placements are located near the native position (Cα RMSd < 2.5 Å). To identify these correct placements, we look for a mutually compatible subset of fragment placements. Compatibility is assessed with a score function that favors fragment pairs with: (a) the same residue in the same place, (b) residues nearby in sequence nearby in space, and (c) no two residues occupying the same space. Simulated annealing Monte Carlo (SA-MC) guided by this score function finds the maximally consistent subset of fragment placements from this larger set.

Fragment matching and SA-MC assembly are applied iteratively until >70% of the sequence has been assigned into density. Each iteration places fragments from unassigned sequence positions of the sequence into unoccupied regions in density (Figure 2.1 and Figure 2.2). Finally, the partial model from the final iteration is completed through rebuilding and all-atom refinement using RosettaCM [34] guided by the experimental density data.

2.3.2 *Evaluation on 9 cryo-EM maps with known structure.*

We tested our method on a benchmark set of 9 proteins. These proteins range in size from 155 to 397 residues, include different fold types, and have experimental cryo-EM maps varying in resolution range from 3.3 to 4.8 Å (Table 2.1). For each map, a single subunit was first segmented from the entire density map. To simulate true *de novo* modeling, fragments from proteins with similar structures and sequences were excluded while constructing the fragment

libraries. In 7 out of the 9 cases, partial models from the final iteration of the *de novo* building step are within 1.1-2.3 Å Cα RMSD from the experimental structures (Table 2.1 and Figure 2.2), 6 of which are more than 70% complete. These partial models were then completed and refined using RosettaCM, yielding models with 1.3-2.2 Å Cα RMSd (2.0-3.1 Å all-atom RMSd) from the experimentally determined structures. In contrast, Buccaneer [26], a widely used model-building method from X-ray crystallography – while able to trace portions of the backbone for all targets – only correctly identifies more than 5% of the sequence in 3 cases, and never identifies more than 50% (Table 2.2).

Among the proteins in the benchmark set, TRPV1 [16] and FrhB [15,39] were proteins with new folds solved recently by manually building models into cryo-EM density. Our method automatically obtained completed models with 1.4 Å Cα RMSd model for TRPV1 and 1.7 Å Cα RMSd for FrhB. To test the resolution limit at which *de novo* structure determination is possible, a previously unpublished 4.8 Å resolution map from the 20S proteasome α-subunit (20S-α) was used. At this resolution, the α-helix pitch is somewhat visible (Figure 2.2 and 2.3, leftmost columns), however, β-strand separation is only barely resolved (Figure 2.3, leftmost column). Using our approach, the final partial model – after 3 iterations– had 196 out of 221 residues placed, with just 1.3 Å RMSd to the crystal structure (Figure 2.2). Using RosettaCM to build a completed model, we obtained a 1.2 Å Cα RMSd model (2.0 Å all-atom RMSd). Despite the lack of side-chain density details, side-chains in the core of the protein show very good agreement to the deposited crystal structure (Figure 2.3, rightmost column).

2.3.3 *Iterative refinement increases model coverage*

In all cases except one (TMV), more than one iteration was required to obtain a partial model with at least 70% of the sequence placed (Table 2.1). For example, 20S-α took three rounds to reach this level of coverage; the partial model after one round only had 34% of the sequence placed (Table 2.1). The reason for this is that, even though in our fragment libraries there are fragments that adopt near-native conformations for 78% of the sequence (Table 2.3), near-native placements of some fragments do not score well enough initially to be carried over to SA-MC assembly. To address this problem, we iterate fragment matching and assembly. After each round, we assemble a consensus assignment, only containing fragments placed in similar locations in all low-scoring SA-MC trajectories. These regions are locked, the corresponding density is masked out, and another round of fragment search and SA-MC is carried out. This is particularly valuable for accurately placing β-sheets, where density tends to have lower local resolution. For example (Figure 2.3), in 20S-α, sequence positions at S3, S6 and S7 were correctly traced only in the second round, and S1, S2, S5, S9 and S10 only in the third. Moreover, as more fragments are placed accurately, SA-MC shows better convergence (Figure 2.3, rows 2 and 3).

2.3.4 *RosettaCM shows robustness in building full-length models*

The final step in our protocol is completing the partial models using RosettaCM guided by experimental density data [34]. Here, unassigned sequence from the final partial model is sampled through a combination of fragment insertion and minimization (see *Methods*). In four of the cases from our testset, this leads to models that have similar or slightly higher RMSds than the partial model from the final iteration, which is expected since the unbuilt parts are mostly

loops or regions with less resolved density. However, in two cases – FrhA and 20S-α – we see an improvement in overall RMS (Table 2.1). For FrhA, this improvement is particularly striking: the Cα RMSd decreases from 2.3 Å to 1.3 Å. Figure 2.4 illustrates some improvements in the structure: RosettaCM corrected several loop residues incorrectly placed into density from the previous SA-MC assembly step. As indicated in Table 2.1, this rebuilding is consistent and robust, with minimal structural deviation over the 10 lowest scoring models.

2.3.5 *Causes for failure*

In three of the cases in Table 2.1, our approach was unable to automatically determine accurate models to more than 50% sequence coverage. This is clearly identifiable by the poor sequence coverage of the models after a single round of modeling. There are two main reasons why our approach may fail to recover accurate models. First, fragment quality is an intrinsic limiting factor for our method. If a large portion of the protein does not have sufficiently accurate fragment quality, it is not possible to accurately assign positions for these residues into the map. BPP1 is one such case: almost half of the sequence positions have no fragments that adopt native-like conformations (see Table 2.3, Cα RMSd < 1.5 Å). Second, building β-sheets from fragments is difficult due to the conformational variability of sheets compared to helices. STIV and VP6 are mostly β-sheet containing proteins, and our method was only able to build models for the helical regions accurately (Table 2.1 and Figure 2.5). These failures suggest possibilities for future method improvements.

2.4 *DE NOVO* STRUCTURE DETERMINATION OF THE TYPE VI SECRETION SYSTEM CONTRACTILE SHEATH

2.4.1 *Results*

Type VI secretion systems are bacterial virulence-associated nanomachines consisting of proteins that are structurally and functionally related to the components of bacteriophage tails despite shared with very low-sequence identity. With the collaboration with the lab of Basler at the University of Basel, we applied our method on a reconstructed cryo-EM map of the contractile sheath proteins of type VI secretion system (EMD-2699) at a resolution of ~3.5 Å, with no detectable homologues of known structure. The asymmetric unit contained a hetero-dimer with 660 residues total. The helical reconstruction density map was firstly segmented manually to be monomer by naïve guessing using the "Volume Cleaner" tool from UCSF Chimera (Figure 2.6). Using the monomer map, eight iterations of our protocol generated a partial model with 466 residues placed. In parallel, the map was manually traced with the aid of Buccaneer in the lab of Basler. There is good overall agreement between two models: over 394 residues, the Cα RMSd is 1.1 Å (Figure 2.7). However, there are 35 residues where sequence registration is shifted by six positions between the two models (Figure 2.8a and b, residues in between orange and blue arrows). This segment is flanked by disordered residues; this combined with the poor local resolution makes sequence assignment particularly difficult. The sequence assignment made by our method shows better agreement with the density map than the hand-traced model in this region (Figure 2.9). Additionally, our approach was able to assign sequence in regions where the manual model did not (Figure 2.7 and Figure 2.9 c and d). The blind case

shows that our approach is tolerant to errors in segmentation; although our manual segmentation was imperfect, structure determination was still successful.

2.4.2 *Combine the automated and manual models using RosettaCM*

The two independently derived models showed reasonably good agreement: 394 residues were assigned in both models with a Cα RMSD of 1.1 Å. However, there were parts of the protein assigned in each model that were unassigned in the other. Thus, to build and refine the final model, we used RosettaCM, a comparative modeling protocol that assembles protein structures by recombining portions of several models (Figure 2.10); in this case, the inputs were the two independently traced models. RosettaCM was guided by experimental density data, with agreement to the density map as an additional score term while building and refining models. A total of 1,000 models were generated, and a best model was selected based on the all-atom energy plus the ‘‘fit to density’’ energy. Among the low-energy models RosettaCM generated, the segment assigned by the automated method was exclusively chosen, again, suggesting our assignment is more energetically favorable and hence correct. Combining our model with the manual model in RosettaCM, we were able to build a full-length model for the hetero-dimer VipA/VipB complex (Figure 2.11 and Kudryashev et al., 2015 [40]).

2.4.3 *Structure refinement*

Using this model, a final refinement step was carried out in the context of the symmetrical assembly (Figure 2.13 and Figure 2.14), improving model geometry and relieving clashes at the symmetric interfaces [41]. Residues shown as Ramachandran outliers were

specified to refine in the Rosetta iterative backbone rebuilding and refinement protocol (Figure 2.10). The final model shows very good agreement to the density, with 504 of 558 traced residues matching the map with real-space correlations of 0.60 or greater (using *density_tools* in Rosetta), and very good model geometry, with 0.36% Ramachandran outliers, 0% rotamer outliers, a Molprobity clash score of 2.15, and an overall Molprobity score of 1.38 [42]. To test for overfitting during model-building, we uniformly perturbed the final model and refined it against the independently generated EM map. A long refinement cycle (1,000 cycles of backbone rebuilding) was used to ensure the refined model is unbiased from the model fit to the original reconstruction. The resulting model had 0.34 Å Ca RMSd to the original model. Atomic B factors were capped to 600 for heavy atoms and to 720 for H atoms.

2.4.4 *Cross-validation of the model from a homologues protein determined recently*

Along with our publication [40], the Zhou's group from UCLA [43] also reported a structure of the type VI secretion system contractile sheath from *Francisella tularensis,* which shared ~40% sequence identity with the structure we reported. This serves a great opportunity to validate the model determined by the newly developed *de novo* model-building method. We compare our model and the model determined by the Zhou group, focusing the two regions (Figures 2.8A and 2.8C) we had reported that the automated method has advantages over the manual model tracing from the lab of Basler. Although the two sequences are not identical in the structures, the sequence features, for example prolines and aromatic residues (Figure 2.15 and 2.16), assigned in both models suggest the model determined by the automated method is accurate.

2.5 DISCUSSION

We have developed a method for automatic *de novo* protein structure determination from near-atomic-resolution cryo-EM data, and demonstrated its applicability to a wide range of datasets. Our method uses predicted backbone conformation to aid in sequence assignment, allowing determination of structure to atomic-level accuracy without requiring prior knowledge of protein topology from homologous structures or manually traced models [23,24]. It is our hope that this becomes a routine method for determining structures from cryo-EM maps with near-atomic resolution.

The key concept introduced in this chapter -- that local sequences have preferences for certain backbone conformations -- has previously been used to predict structures of small proteins (< 100 residues) [5] *de novo*, and larger proteins through incorporating sparse backbone-only NMR data [6,44,45]. However, no previous approach in protein structure prediction has used this concept in conjunction with experimentally determined local Cartesian-space restraints to restrict conformational space. The method described here should provide a general framework for the use of these types of sparse experimental constraints in protein structure determination.

Several improvements will increase both the applicability and accuracy of our approach. Our tests assumed a map where the asymmetric unit was segmented. While manual segmentation is often straightforward (as in the blind case), it may prove difficult in highly intertwined structures. An obvious future research direction is incorporation of symmetry information in the modeling procedure. Additionally, further improvements of the method on all-β proteins are necessary. Incorporating strand-pairing bonuses in the scoring function combined with more aggressive fragment optimization into density should improve accuracy with all-β proteins. Our

approach is amenable to incorporation of additional structural information: known structures of components are easily incorporated, experimentally derived pairwise distance restraints may guide conformational sampling, and Cα traces provided by users. As techniques in cryo-EM continue to improve, more maps will be available at near-atomic resolution. Our method should contribute to the determination of high accuracy models from such maps, reducing human effort and errors due to human biases.

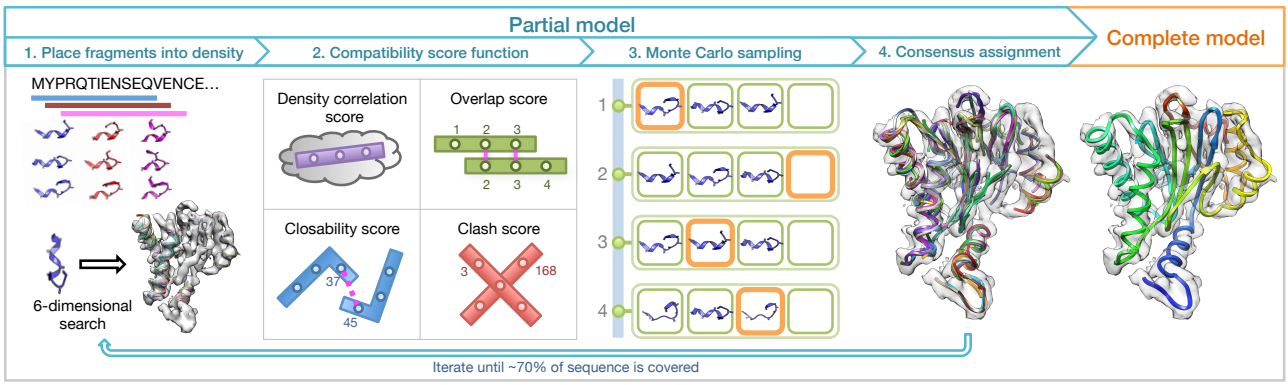

Figure 2.1. Protocol Overview.

First, for a 9-residue window centered on each position in the sequence, representative backbone conformations (fragments) are collected and docked into the density map. Second, the resulting fragment placements are then evaluated using a score function consisting of four terms: a density correlation term assessing the agreement of fragment and map; an overlap term favoring fragment pairs assigning the same residue to the same location; a closability term favoring fragment pairs close in sequence that are close in space; and a clash term preventing two residues from occupying the same place. Third, from the candidate placements (square green blocks), simulated annealing Monte Carlo finds a set of fragments (square orange blocks) optimizing the score function; a null placement (empty blocks) may be assigned in positions where no good placements have been identified. Fourth, a partial model is assembled by combining fragment placements from multiple Monte Carlo trajectories. Steps 1–4 are carried out iteratively until ~70% of sequence is covered. Finally, unassigned regions in the final partial model are completed using density-guided loop sampling followed by all-atom refinement.

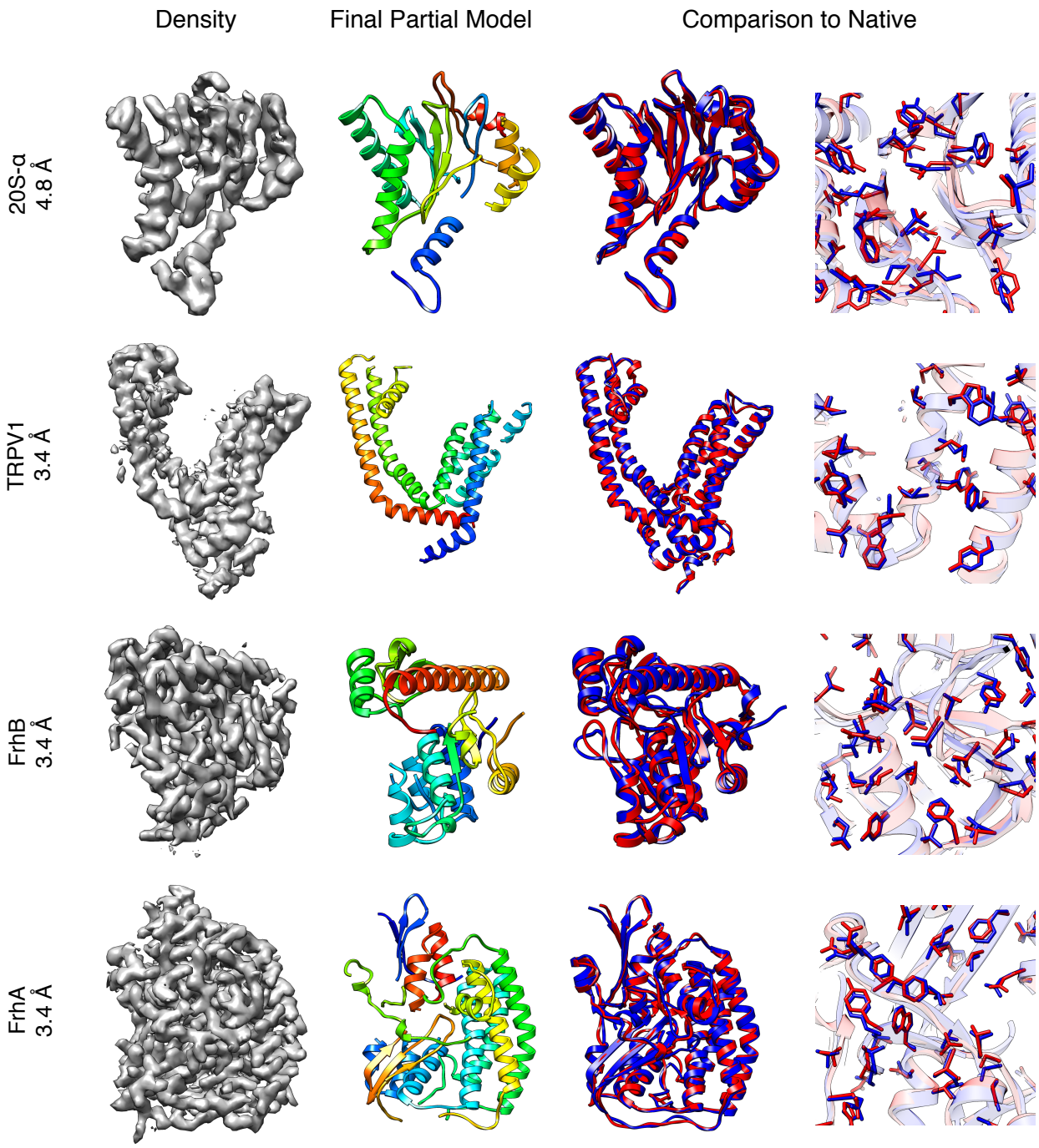

Figure 2.2. High-accuracy model-building in near-atomic resolution cryo-EM maps. **(Leftmost column)** The density maps used for *de novo* model-building on 20S-α at 4.8 Å, TRPV1-TM at 3.4 Å, FrhB at 3.4 Å, and FrhA at 3.4 Å (Row 1 to 4, respectively). (**Column2**) The partial model at the final iteration. (**Column 3 and 4**) Full-length RosettaCM models (red) are superimposed with the native structure (blue). Each sub-figure shows the lowest-RMSd structure from 10 lowest-electron-density-score models (left) with a close-up of the core showing that native core packing is recovered (right).

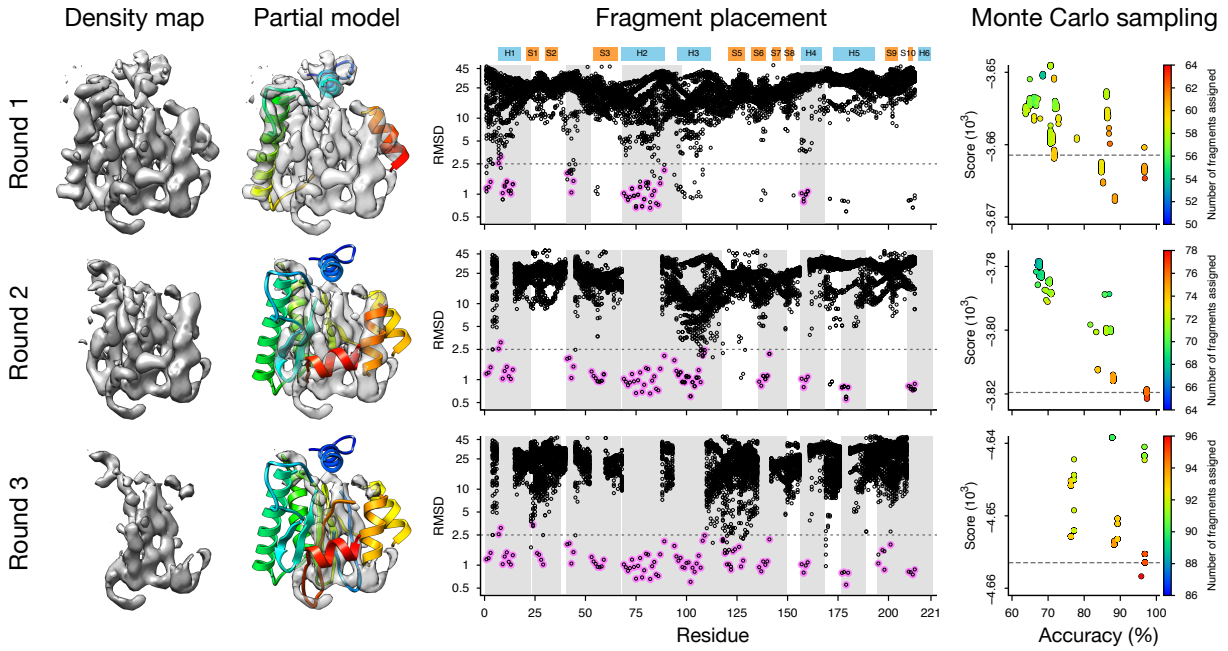

Figure 2.3. Iterative approach.

Model-building for the 20S α-subunti in a 4.8 Å resolution cryo-EM map required three iterations, illustrated in the three rows in the figure. In leftmost column**,** the density map used for the corresponding iteration, after masking out density from the previous round's partial model. In column 2**,** the assembled partial models after Monte Carlo sampling (colored blue at the N-terminus to red at the C-terminus). In column 3, fragment placement results after translation and rotation search. The x-axis covers the sequence of the protein, and each black point represents a single fragment placement; the y-axis indicates the distance of the fragment placement to the native conformation. Pink points indicate fragments chosen to assemble the partial model, and the grey shading shows residues covered in the partial model. Secondary structural elements in the native are indicated above the plot, where H indicates helix and S indicates strand. In rightmost column, convergence of Monte Carlo trajectories. Each point represents the fragment assignment of an independent search trajectory, colored by number of total fragments placed. The X-axis indicates the percentage of fragments placed within 2.5 Å RMSd to the native configuration, while the Y-axis shows the score with the fragment compatibility function. The horizontal dashed line shows the score cut used for partial model generation.

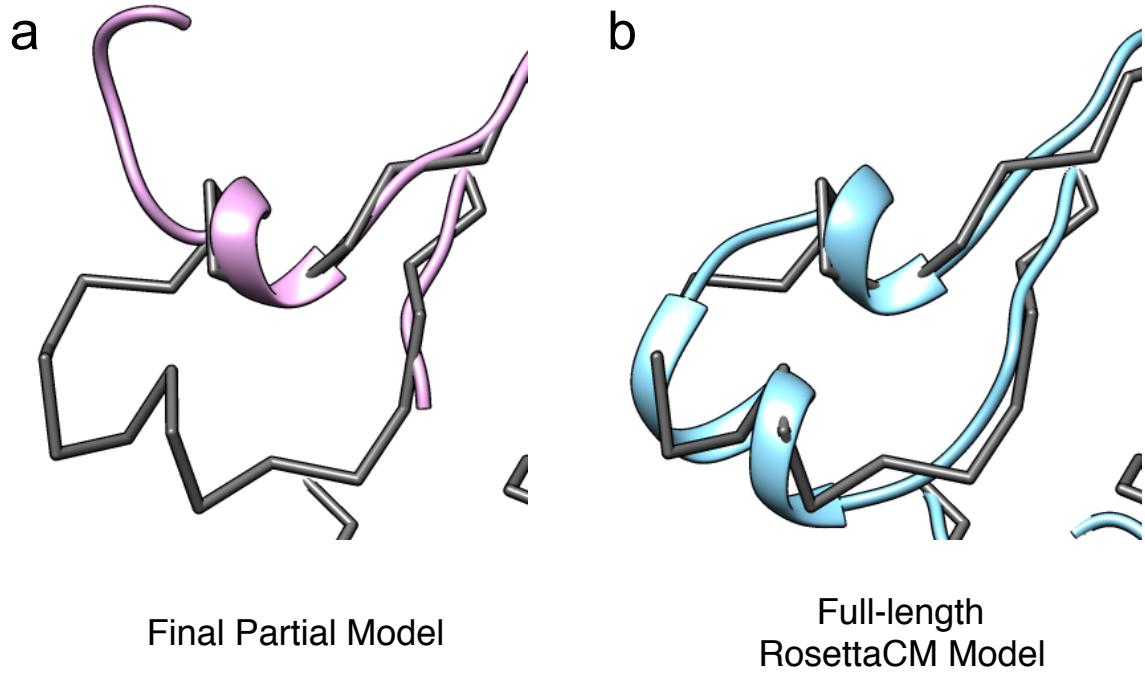

Figure 2.4. RosettaCM fixes placement errors in the final partial model of FrhA.
(a) One case where there is an error in fragment placement. **(b)** Density-guided rebuilding in RosettaCM is able to correct this error as the missing residues are rebuilt. The backbone trace in grey indicates the native structure (4ci0).

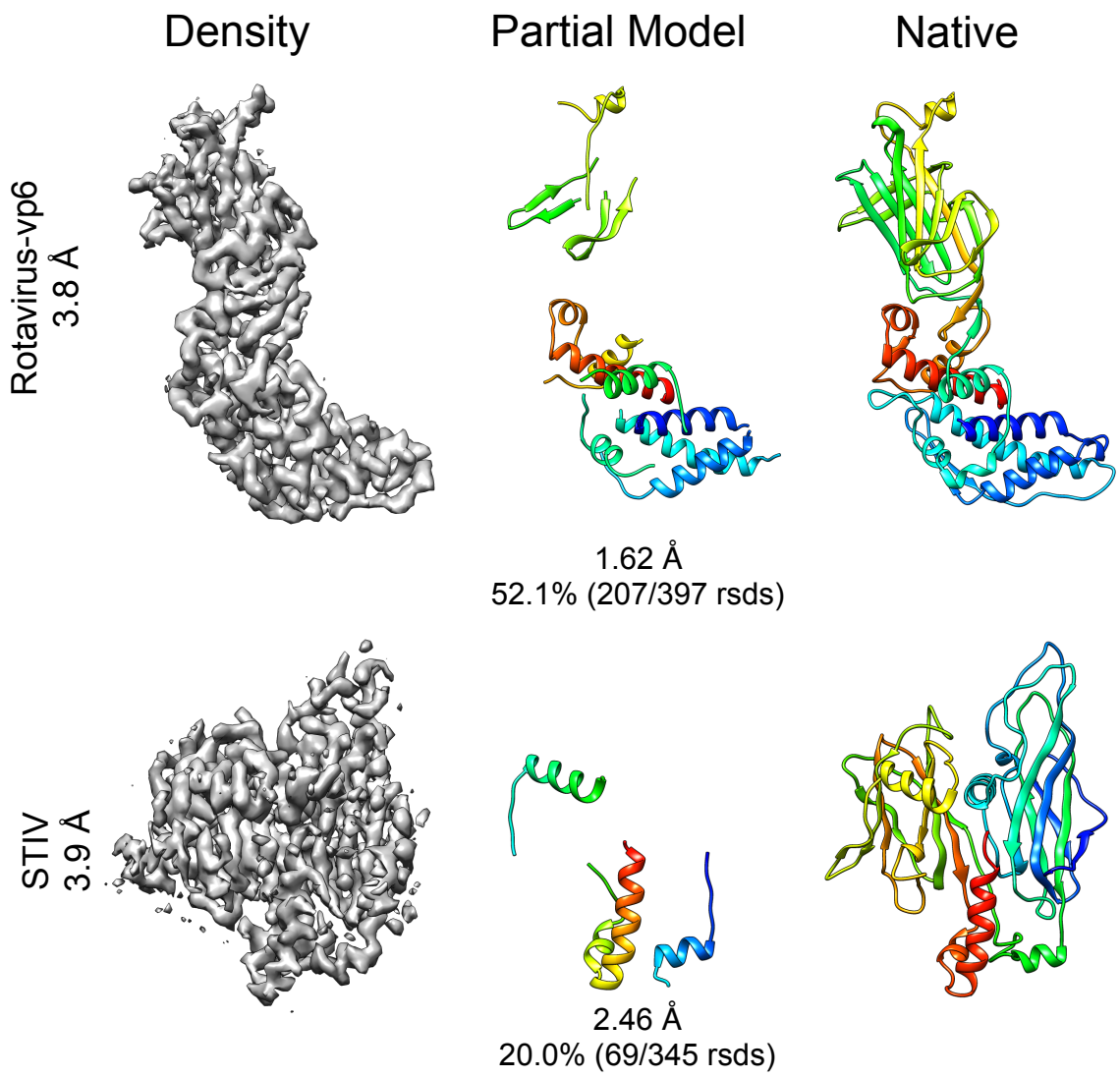

Figure 2.5. Problems in building β-sheets.
Each row indicates a case where our automatic approach failed to determine the structure: VP6 at 3.8 Å (top) and STIV at 3.9 Å (bottom). Illustrated are the density map (left), the partial model at the iteration with highest accuracy (middle), and the native structure (right).

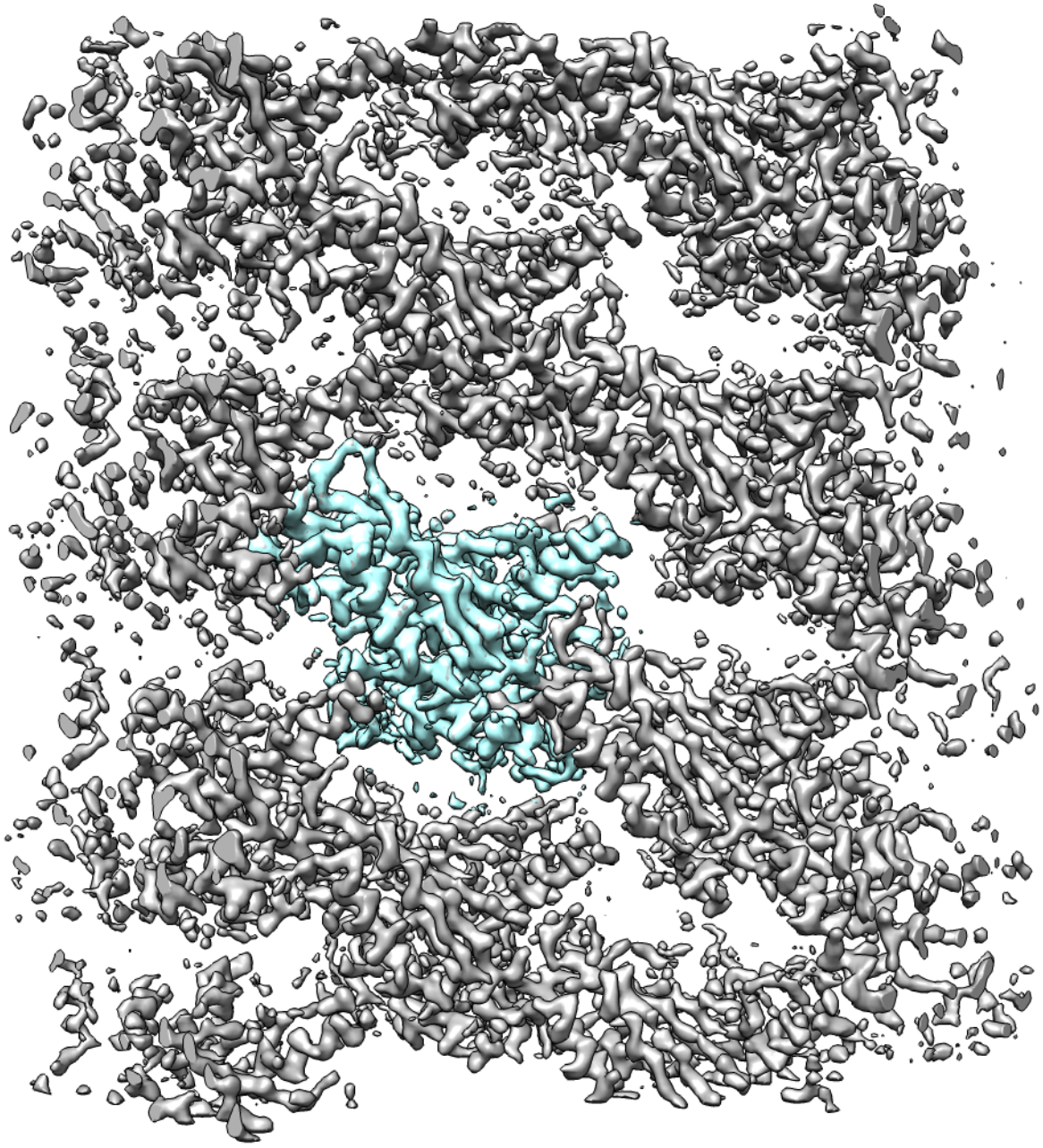

Figure 2.6. Manually truncation of the asymmetric unit of the VipA/VipB complex

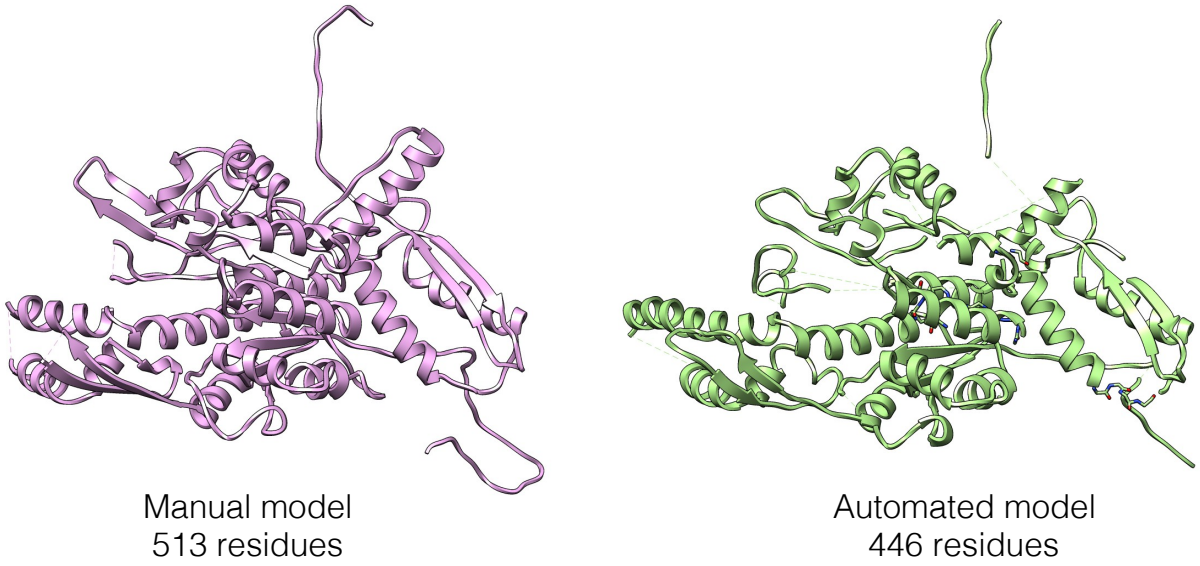

Figure 2.7. Comparison of automated model-building (right) and manual model-building (left).

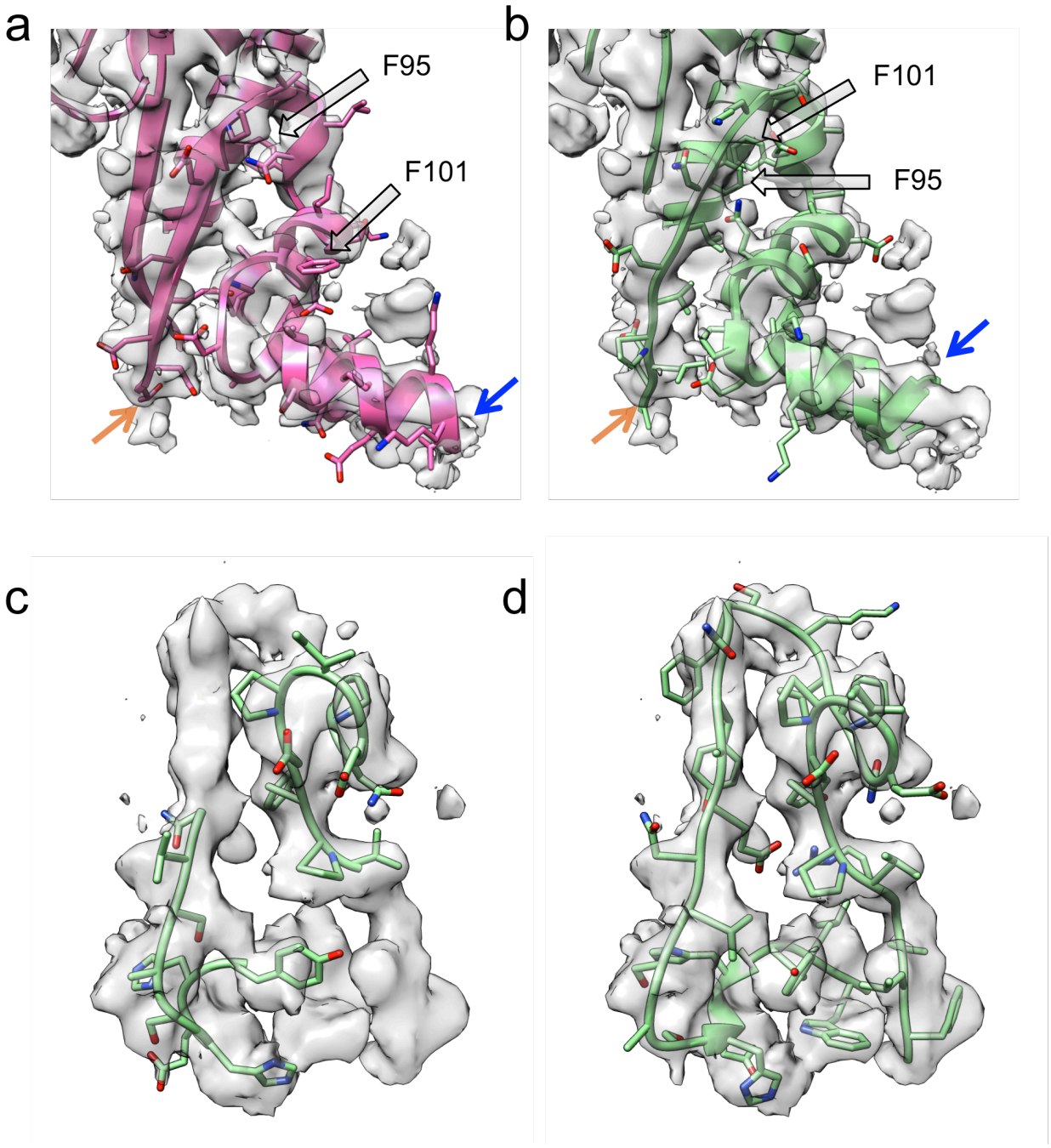

Figure 2.8. Blind structure determination.

An error in the manually traced model (pink, A) is corrected by our method (green, B). The arrows in black show the positions of two residues in both models (F95 and F101), highlighting the six-residue registration shift between the models. Orange and blue arrows in A and B indicate the beginning and end of the region with the sequence registration discrepancy. (C) A partial trace generated by our method in a region where manual tracing was impossible. (D) The full-length RosettaCM model at the same region shows good agreement with the map.

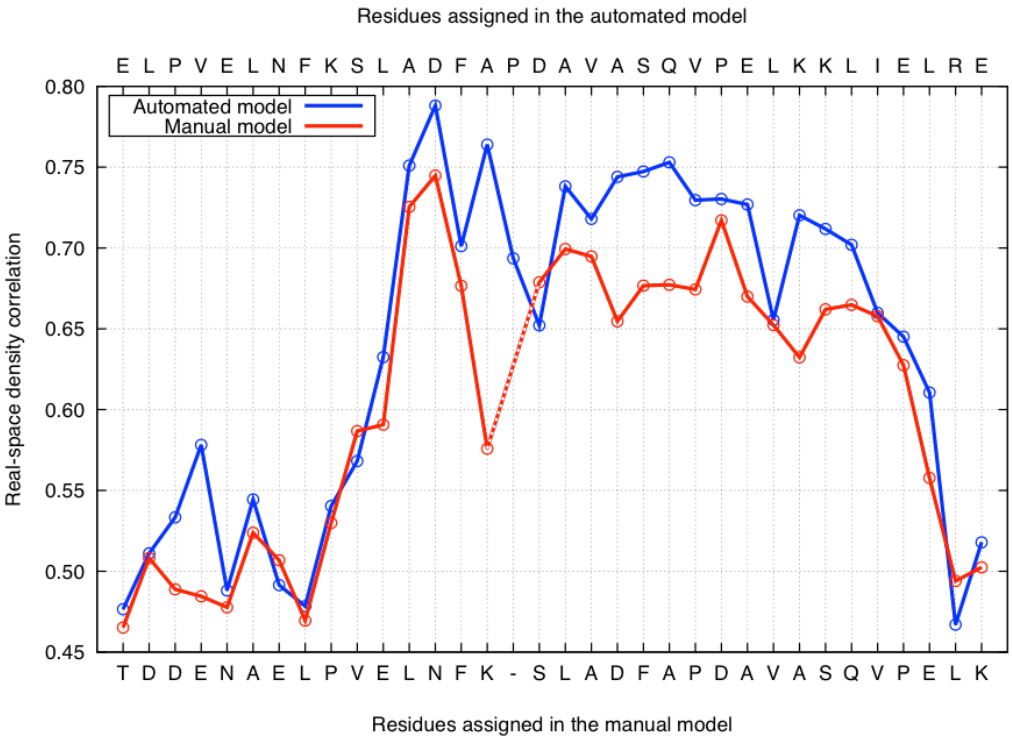

Figure 2.9. Real-space density correlation over ambiguous region in the map.
A structural alignment of the residues with a discrepancy in sequence registration is indicated on the x-axis (the residues in between the orange and blue arrows in Figure 3a and 3b). The sequence assignment of the automated and manual models is labeled on the upper and lower axes, respectively. The blue line (automated model) and the red line (manual model) show the real-space density correlation for each assignment (y-axis). Real-space density correlation was calculated using *density_tools* in Rosetta with a 2.5Å mask.

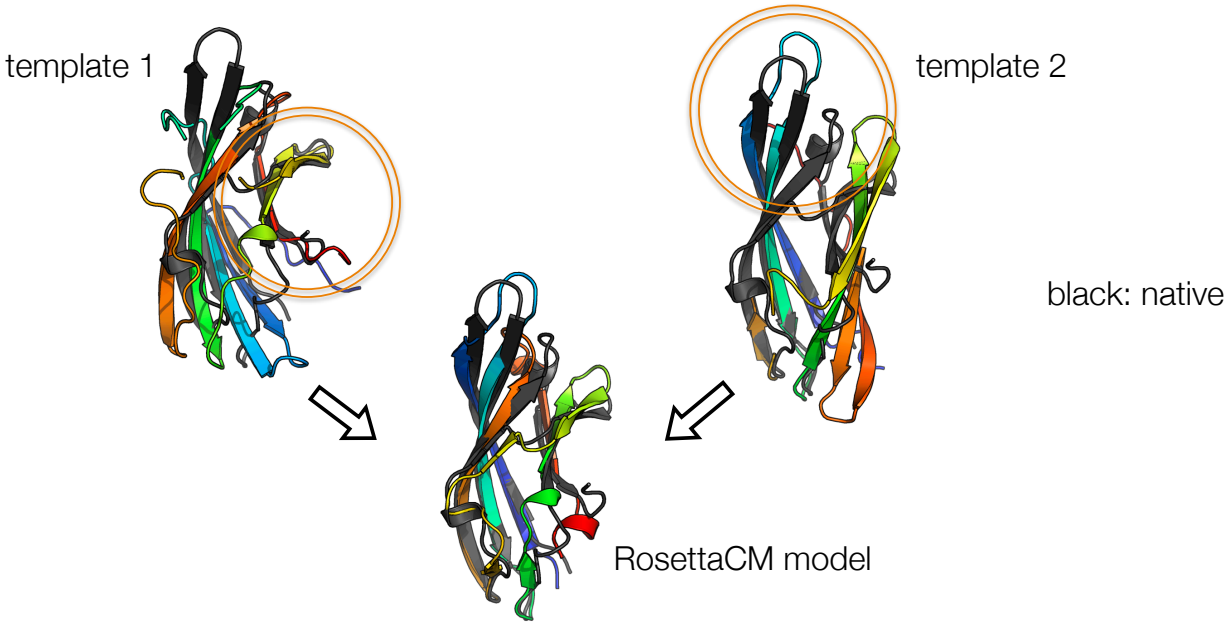

Figure 2.10. The concept of RosettaCM – combining conformation information from multiple templates.

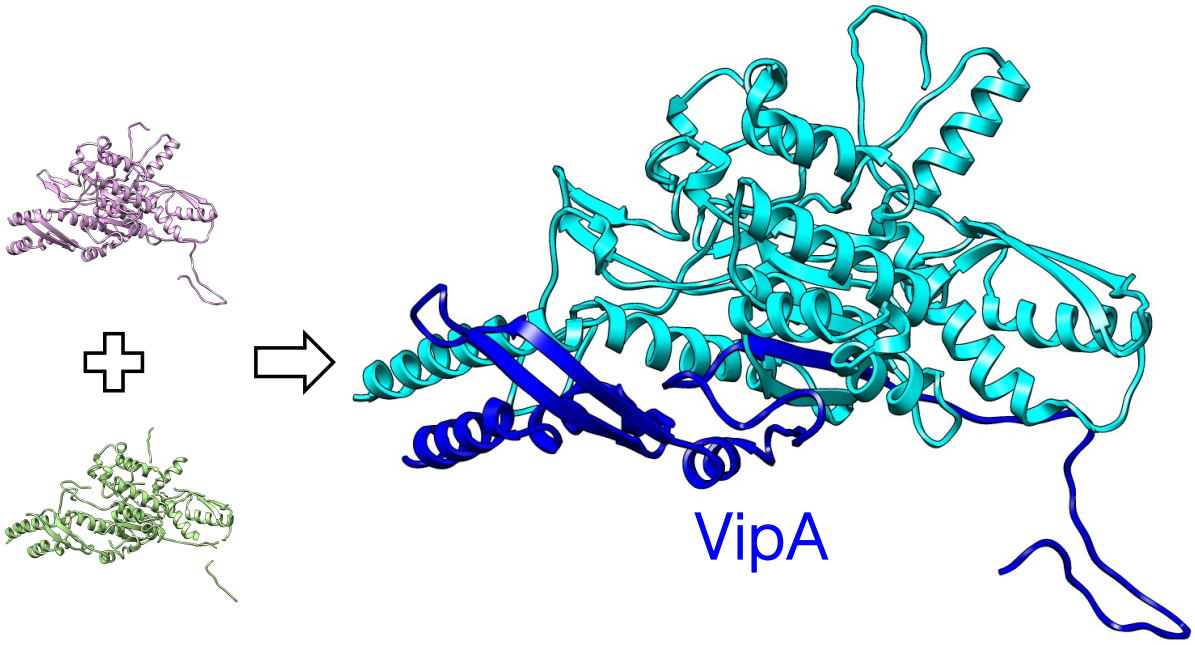

Figure 2.11. The complete model was determined using information from both automated and manual models using RosettaCM.

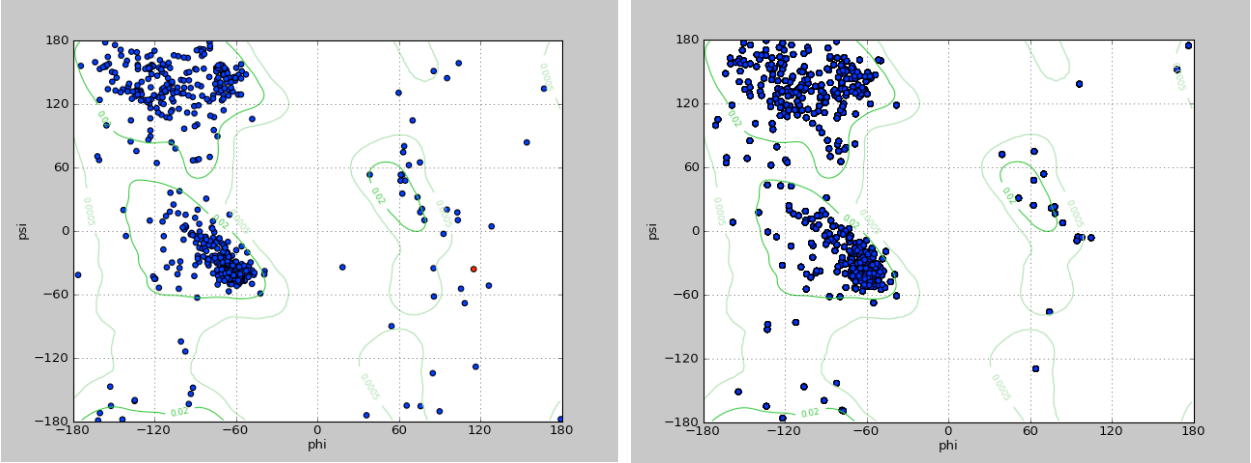

Figure 2.12. Ramachandran plots to show the improvement of backbone geometry before (left) and after (right) structure refinement using Rosetta iterative backbone rebuilt and all atom refinement.

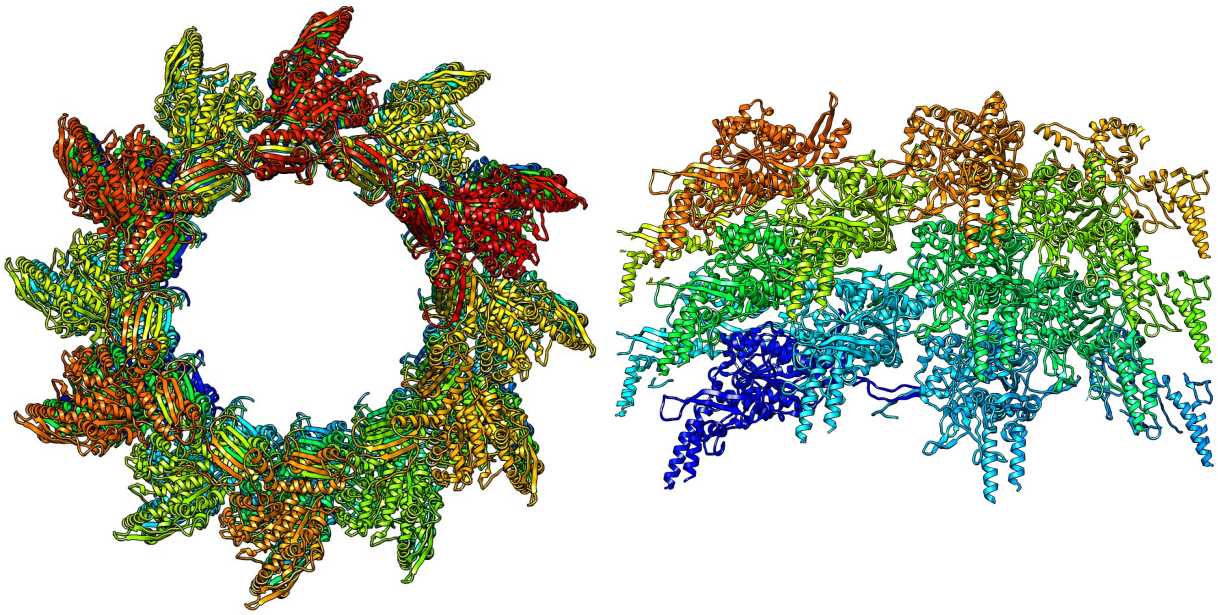

Figure 2.13. The overview of the assembly of VipA/VipB complex revealed by cryo-EM

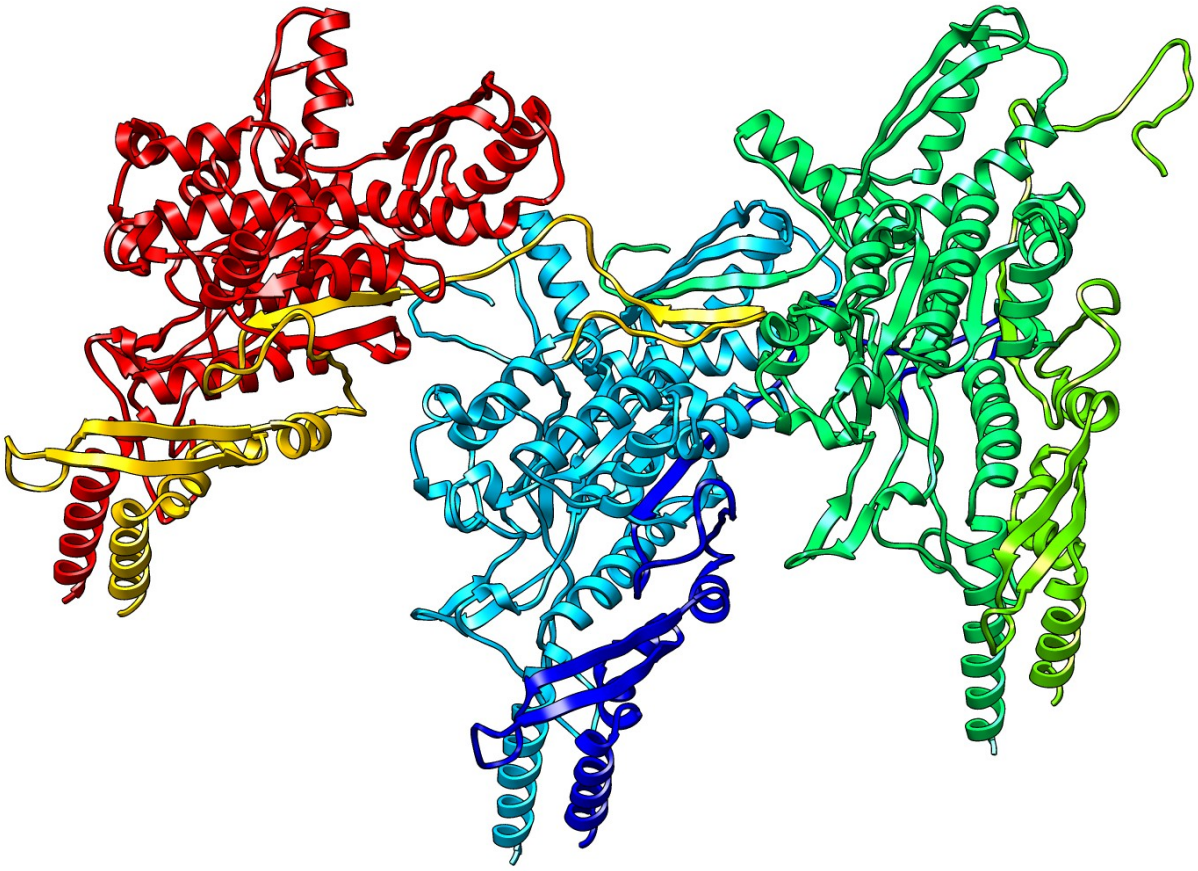

Figure 2.14. The “handshake” domain of the VipA/VipB assembly.

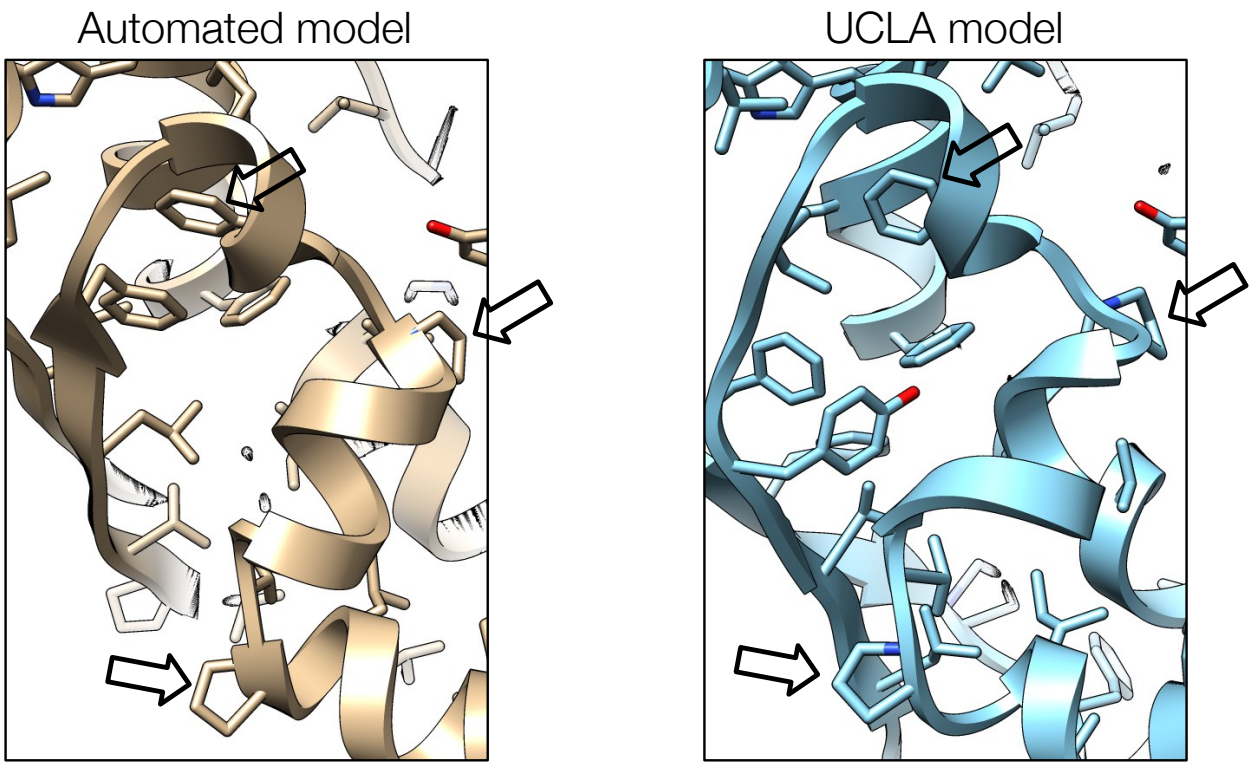

Figure 2.15. Comparison of the structure determined by the automated method and by the homologous structure determined by a UCLA group.

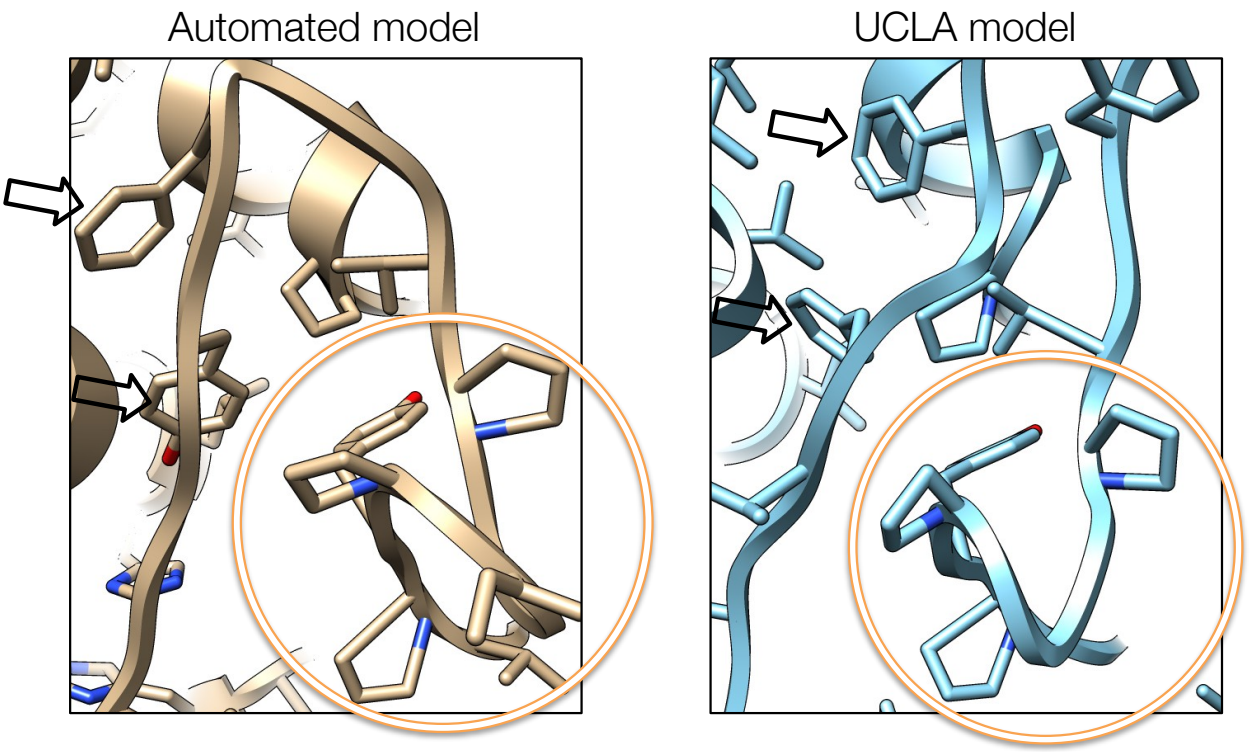

Figure 2.16. Comparison of the structure determined by the automated method and by the homologous structure determined by a UCLA group.

Table 2.1. Model-building in near-atomic-resolution cryo-EM maps

Target	PDB ID (chain)	EMDB ID	Fold	Reported resolution (Å)	Length (aa)	Partial model		Completed model		
						Round1 Cα RMSd [Å] (%)	Final Cα RMSd [Å] (%)	Lowest Cα RMSd [a] [Å]	All-atom RMSd [b] [Å]	Mean Cα RMSd [c] [Å] (σ [d])
TMV	3j06 (A)	5185	α	3.3	155	1.3 (81)	1.3 (81) [e]	1.7	2.6	2.1 (0.2)
TRPV1 [f]	3j5q (A)	5778	α	3.4	310	1.1 (46)	1.1 (76)	1.4	2.2	1.8 (0.3)
FrhA	4ci0 (A)	2513	α/β	3.4	385	1.0 (74)	2.3 (91)	1.3	2.1	1.7 (0.2)
FrhB	4ci0 (C)	2513	α/β	3.4	280	1.2 (74)	1.4 (85)	1.7	2.5	2.0 (0.2)
FrhG	4ci0 (B)	2513	α/β	3.4	228	1.4 (46)	1.6 (73)	2.2	3.1	2.6 (0.3)
BPP1	3j4u (A)	5764	α/β	3.5	327	17.2 (42)	17.2 (42)	- [g]	-	-
VP6	1qhd (A)	1461	α/β	3.8	397	1.2 (37)	1.6 (52)	-	-	-
20S-α	1pma (A)	TBD	α/β	4.8	221	1.3 (34)	1.3 (88)	1.2	2.0	1.2 (0.0)
STIV	3j31 (A)	5584	β	3.9	344	2.5 (20)	21.9 (26)	-	-	-

Numbers in parentheses from all RMSD columns represent the percentage of sequence covered in the partial model.

[a] The model is selected among 10 lowest density score models.

[b] All-atom RMSD is calculated from the model reported for lowest Cα RMSD using all non-hydrogen atoms.

[c] Mean values of Cα RMSD from the 10 lowest density score models.

[d] Standard deviation.

[c] Only one iteration of model-building was run.

[f] Only the transmembrane domain (residue 381-719) was used here.

[g] Full-length modeling is not carried out.

Table 2.2. Model-building results using Buccaneer

Target	Length (aa)	Cα-atom placed	Sequence registered	Correctly registered [a]
TMV	155	145	56	0
TRPV1	315	257	190	0
FrhA	386	382	367	185
FrhB	281	192	186	126
FrhG	228	242	190	63
BPP1	327	339	162	0
VP6	397	405	155	0
20S-α	221	224	135	7
STIV	345	553	259	0

[a] Residues within 2.5 Å to the deposited structures are considered correctly registered.

Table 2.3. Fragment quality for the benchmark proteins

Target	Total fragment positions [count]	Number of positions with a fragment adopting a native-like conformation			
		0.5 Å Cα RMSd [count](%)	**1.0 Å Cα RMSd [count] (%)**	**1.5 Å Cα RMSd [count] (%)**	**2.0 Å Cα RMSd [count](%)**
TMV	147	50 (34.0)	76 (51.7)	95 (64.6)	111 (75.5)
TRPV1	307	145 (47.2)	204 (66.4)	243 (79.2)	269 (87.6)
FrhA	378	154 (40.7)	245 (64.8)	293 (77.5)	329 (87.0)
FrhB	273	83 (30.4)	163 (59.7)	198 (72.5)	237 (86.8)
FrhG	220	62 (28.2)	114 (51.8)	148 (67.3)	182 (82.7)
BPP1	319	49 (15.4)	117 (36.7)	170 (53.3)	219 (68.7)
VP6	389	107 (27.5)	183 (47.0)	237 (60.9)	298 (76.6)
20S-α	213	73 (34.3)	132 (62.0)	165 (77.5)	194 (91.1)
STIV	337	45 (13.4)	128 (38.0)	210 (62.3)	270 (80.1)

Chapter 3. STRUCTURE DETERMINATION OF MULTI-COMPONENT PROTEIN COMPLEXES

3.1 INTRODUCTION

Molecular assemblies and machines play crucial roles in a wide variety of biological structure and function. Structures of such assemblies solved by X-ray crystallography have increased our understanding of how biological systems work. However, due to their complexity and large size, many protein complexes cannot be crystallized. Cryo-EM and improved image processing techniques have made it possible to carry out structural studies of such macromolecular complexes. Three-dimensional single particle reconstructions typically produce electron density maps with low resolution, where the molecular envelope of the complex is visible. By fitting atomic-level structures of components accurately into these maps, pseudo-atomic models of the assembly can be obtained. The components of assemblies may be placed accurately with the aid of rapidly growing and mature experimental methods such as NMR and cross-linking data [46], as well as evolutionary information such as predicted residue-residue contacts from protein sequence co-evolution [47,48]. By using computational modeling methods that use this data together with low-resolution cryo-EM maps, modeling the structure of molecular machines with atomic level accuracy may be possible.

Fitting multiple components into a density map is a considerable challenge, especially when symmetry operations cannot be applied to reduce the number of degrees of freedom. Several studies have approached this problem [49,50], however, none were able to handle more than seven components, and none were able to produce high-resolution models of assemblies. Many molecular assemblies and machines are composed of more than 20 subunits. Rosetta is capable of producing atomic-level models of symmetrical molecular machines with and without

using sparse experimental data [51,52]. Our goal is to develop a generalized protocol to determine the structures for multi-component complexes using medium- to low-resolution cryo-EM maps with atomic accuracy.

In this chapter we describe a new approach to assemble complexes from component structures in medium-resolution (5–15 Å) cryo-EM maps. The approach starts by docking likely configurations of each component structure in the density map through a six-dimensional search. Monte Carlo sampling guided by a score function finds solutions of component placements that are mutually compatible. The protocol is similar to the place-and-assemble scheme from Chapter 2, and involves two steps (Figure 3.1): first, subunits are placed into density, and second, subunits that are compatible with each other are identified. For each subunit, many alternative placements are saved into a list of possible placements for each subunit, and then configurations with multiple subunits will be searched using Monte Carlo sampling to optimize a score function that considers the following: (1) the fit of each subunit to the density map, (2) loop closability of termini between subunits (if applicable), and (3) clashes between subunits. We applied the method to a 6.5 Å resolution cryo-EM map of the peroxisomal Pex1/Pex6 ATPase complex, a trimeric double-ring AAA+ molecular machine comprised of eight domains with only two unique topologies.

Pex1 and Pex6 are members of the AAA family of ATPase, which contain two ATPase domains in a single polypeptide chain and form hexameric double rings. These two Pex proteins are involved in biogenesis of peroxisomes, and mutations in them frequently cause diseases. To determine the structure for Pex1/Pex6 complex, the labs of Rapoport and Walz used single-particle cryo-EM to get a ~7 Å resolution density map (Figure 3.2A). However, structure determination of Pex1/Pex6 complex served a certain extent of challenge although sharing

similar double-ring architecture with p97. First, there are only two distinct topologies shared by eight domains in the asymmetric unit of the Pex1/Pex6 complex (Figure 3.2 B). Homology detection using HHsearch [53] suggests that both Pex1 and Pex6 proteins contain two N-terminal domains (designated N1 and N2 sequentially) that share the double Ψ-β barrel fold to the single N-terminal domain of p97 and NSF, and two C-terminal domains (designated D1 and D2 sequentially) that share the AAA-ATPase fold to the two C-terminal domains of p97 and NSF (Figure 3.4 B) This poses challenging problem of determining positions and orientations of 8 domains of two protein folds for the Pex1/Pex6 complex. Second, the low-sequence identity alignments of two N domains of Pex1/Pex6 proteins pose a challenge of modeling the proteins accurately using detected templates. The sequence alignments with homologues of known structure indicate the N2 domain of Pex1 and the N1 domain of Pex6 have only templates with low-sequence (<10%) identified. Third, in the asymmetric unit of the complex only 7 out of 8 domains are present in the density map, where one double Ψ-β barrel domain is missing. From the 2D class averages of negative-stain images (Figure 3.2 C), it appears that there is one domain showing flexible. Sharing with the same protein fold, the missing domain could be any of the four N domains from Pex1/Pex6 proteins. Together, to build a high-resolution model for the Pex1/Pex6 complex, we need to build individual models from low-sequence identity templates using comparative modeling, as well as assemble the constituting models. To overcome these challenges, we applied a hybrid approach that employed a newly developed complex assembly algorithm with the component models built from a density-guided comparative modeling method (outlined in Figure 3.3). We uniquely identified the placement of domains, unambiguously determining the structure of the complex. Further experiments carried out by our collaborators validated that my domain placement was correct.

3.2 METHODS

3.2.1 *Template identification and domain placement*

We first used three different sequence-alignment packages (HHsearch [53], RaptorX [54], and SPARKS-X [55]) to find structural templates for each domain of Pex1 and Pex6. This procedure identified the domain organization of each protein, with the D1 and D2 domains having AAA ATPase folds and the N1 and N2 domains having Ψ-β barrel folds. We initially used two model template structures, 1iy2A for the AAA domain and 1wlfA for the Ψ-β barrel fold, to identify the likely placements of each domain in the density map. The search used the spherical harmonic decomposition [33] of both model and experimental density maps to search over rotational space; the rotational matching was carried out over all possible translations. After the search, the 10,000 top-scoring placements for each template were selected. In a second step, the placements were clustered and ranked by using a more precise masked density correlation to replace the spherical harmonic decomposition. This protocol resulted in five high-confidence placements for the Ψ-β barrel fold and four high-confidence placements for the AAA domain. Next, we attempted to place individual homology models for the domains of Pex1 and Pex6 into the preliminary positions of the Ψ-β barrel or AAA-ATPase folds. The RosettaCM comparative modeling pipeline [34] was employed to generate threaded templates (partial threads that include only the aligned residues for each template) for each of the domains of Pex1 and Pex6. Each of the partial threads was superimposed on the corresponding topology placements from the previous step, and was further refined using rigid-body minimization into the local density. For each domain of the Pex1/Pex6 complex, the ~100 best positions for the partial threads were retained for assembly of the full complex.

3.2.2 *Assemble placements of partial threads using Monte Carlo sampling*

In the next step, all eight domains of Pex1 and Pex6 were placed into the density map simultaneously. The placement was done with a simulated annealing Monte Carlo (MC) sampling procedure, in which each combination of domain placement is assigned a score. The scoring function assesses the agreement of an assignment to the density data ($score_{density}$), as well as the consistency of a domain assignment. Consistency assesses whether neighboring domains clash with one another ($score_{clash}$) and whether they can be connected with the known linker lengths ($score_{closability}$). The scoring function thus has three terms:

$$score_{total}(\boldsymbol{D}) = w_{density} \sum_{d_i \in \boldsymbol{D}} score_{density}(d_i) + w_{closability} \sum_{d_i, d_j \in \boldsymbol{D}} score_{closability}(d_i, d_j) + w_{clash} \sum_{d_i, d_j \in \boldsymbol{D}} score_{clash}(d_i, d_j)$$

Here, $\boldsymbol{D}$=[$d_1,\ldots,d_8$] is a combination of placements of the eight domains into the density. We included the possibility that a domain might not be represented in the density map.

The term $score_{density}$ assesses the agreement of a domain placement to the experimental map, and is a function of the real-space correlation between the density expected from the model and the observed density [8].

The term $score_{closability}$ gives a bonus when sequence-adjacent domains are placed close to each other. It is only nonzero for pairs of domains adjacent in sequence, and is dependent on the linker length connecting the two domains as well as the distance that a linker must cover in a particular model:

$$score_{closability}(d_i, d_j) = \begin{cases} S(|i-j|) \cdot (d_i - d_j), & \|d_i - d_j\| < maxdist(|i-j|) \\ 100, & \|d_i - d_j\| \geq maxdist(|i-j|) \end{cases}$$

Here, *maxdist* is the maximum distance that a linker of known length can cover, based on distances observed in known structures. Placing domain pairs such that their termini are farther apart than this distance incurs a large penalty (100). For gap distances less than *maxdist*, there is a weaker penalty that favors domain termini placed closer together. The penalty takes the form $S \cdot dist$, where the coefficient S depends on the number of residues in the linker. Its value is generally near 1 and somewhat larger for linkers with fewer residues.

The term $score_{clash}$ penalizes clashes between domain pairs. The clash penalty is relatively small since some overlap is possible because we are starting with partial threads rather than actual structures.

$$score_{clash}(d_i, d_j) = \sum_{C\alpha_i, C\alpha_j \in d_i d_j} \begin{cases} 1, & \|C\alpha_i - C\alpha_j\| \leq 2.0 \\ 0, & \|C\alpha_i - C\alpha_j\| > 2.0 \end{cases}$$

Simulated annealing Monte Carlo (MC) optimization [56] was then used to find the best domain assignment. The MC trajectory samples domain assignments, uses $score_{total}$ to evaluate each, and accepts or rejects each assignment using the Metropolis criterion . The temperature of the trajectory was reduced from 500 to 1 in 100 increments of 1000 steps each; the runtime for each trajectory was approximately 30 sec.

Weights were initially tuned using synthetic density data derived from multiple protein complexes that have known crystal structures. This analysis showed that $w_{density}$=1, $w_{closability}$=10, and w_{clash}=10 yielded a large score gap between the correct solution and highest-scoring incorrect solution (unpublished data). As each term basically behaves as a soft step function, we expect that the results of the method are largely insensitive to the exact values on the individual weight terms; a wide variety of weights will give the same results below and above the step.

To better illustrate the contribution of the experimental data term ($score_{density}$) and the modeling gemoetry terms ($score_{clash}$ and $score_{closability}$), we plot the scores for all the visited MC trajectories, as opposed to the final converged trajectories shown in Table 3.1). Shown in Figure 3.7, the dynamic range from $score_{density}$ is small (~100 unit). As the result of that, the discrimination power with this term alone is weaker given that there are only two topologies shared by eight domains. Indeed, as shown in Figure 3.8, when we take the second and third solutions, ranked using $score_{density}$ alone, the solutions have linkers that are physically almost not closable. The problem inherited from the difficulty of determining structure for the Pex1/Pex6 system could therefore be alleviated by using the $score_{closability}$ to reward placement solutions with linkers having shorter distances to close.

3.2.3 *Second iteration of Monte Carlo sampling using refined models*

The first application of MC optimization was carried out using partial threads of individual domains as input. This initial assembly converged upon four possible domain assignments. Next, we used the full sequences of the individual domains (not just the partial threads). For each of the four possible domain assignments, we used RosettaCM to rebuild unaligned residues and refine each domain into the density map (modeling was done per domain for tractability). RosettaCM builds protein structures by combining segments from multiple homologs, with an additional term assessing agreement of a model to density [8,34]. The resulting models were used as input for a second iteration of MC optimization. This second iteration gave a much better separation between the best and next best domain placements (Tables 3.1 and 3.2). In the best domain placement, the N1 domain of Pex1 is omitted.

Finally, full-length models were constructed in RosettaCM by rebuilding the linkers connecting individual domains, and refining the symmetrical assembly into the symmetrized density map. While this gave good agreement over most of the structure, the D2 domains poorly fit the symmetrized map. We then further refined the D2 domains into the unsymmetrized map of the Pex1/Pex6 complex in ATPγS. Initially this refinement was carried out using a model of only the six D2 domains from the symmetric model. Next, the loops connecting D1 and D2 were rebuilt (using RosettaCM) and the complete complex was refined using both RosettaCM and a Rosetta iterative local rebuilding protocol [57]. This model was used as a starting point for allatom refinement of the Pex1/Pex6 structure in ADP using the Rosetta *relax* protocol.

3.3 RESULTS

The HHsearch algorithm [53] predicted that both Pex1 and Pex6 have two N-terminal domains, designated N1 and N2, which have a double Ψ-β barrel fold, followed by tandem AAA ATPase domains, D1 and D2. The D2 domains of Pex1 and Pex6 have structural homologs of high sequence identity (~40%), whereas the N and D1 domains have only low sequence identity templates (<20%). However, even for the least conserved domain by sequence identity, the N2 of Pex1, the identified structural homologs all share the same overall fold (Figure 3.2B), although these templates are as different from one another by sequence identity as they are from the Pex1 domain.

To place the domains of Pex1 and Pex6 into the density map, we developed a method for the assembly of protein complex models. We first used a known representative Ψ-β barrel and a known AAA-ATPase domain structure from the Protein Data Bank (PDB) to identify the location of these domains in the density map. To this end, a full rotation and translation search was performed in the asymmetric unit of the symmetrized cryo-EM map, identifying candidate placements for each fold (Figure 3.2B). This preliminary positioning into the cryo-EM density map resulted in an average real-space correlation of 0.7, calculated per domain with the "fit to density" tool in UCSF Chimera. The agreement to the density map suggests that the domain fold identification results from HHsearch are reliable.

Next, we replaced these representative placements with the actual, unique domains of Pex1 and Pex6 from homology search. We used the RosettaCM pipeline [34] to generate ~25 structural homology models for each domain. In each case, nonaligned residues were excluded, resulting in partial models. Positioning each partial model in the identified candidate placements from the previous step resulted in ~100 placements of partial model for each domain. We then

sought to identify mutually compatible domain placements from the combinations of these placed partial models for all domains. Compatibility was assessed with a scoring function that evaluates the quality of the fit into the density map, the degree of clashes between domains, and consistency with known linker lengths between adjacent domains in a polypeptide chain. Four possible domain assignments emerged. To distinguish among these, for each domain assignment we added the unaligned residues originally omitted in the partial models and used RosettaCM to rebuild and refine each domain to better fit the EM density map. Using these refined homology models, we again looked for mutually compatible domain placements through optimizing the compatibility scoring function. A single solution stood out, from which the N1 domain of Pex1 was excluded from the domain assignment. Although the same domain placement emerged as the best solution using partial and complete amino acid sequences, the score difference from alternative solutions was significantly increased (Table 3.2). This domain placement is distinguished from the next best solutions by a large score gap (Figure 3.7 and Table 3.1). Furthermore, the top solution fits the density map better than the next best (Figure 3.6). Even if we consider fit-to-density or geometry criteria individually, the identified model has the best score (Figure 3.7). When the fit to the experimental density map was used as the sole criterion, the two best alternative solutions have the polypeptide chain termini in consecutive domains so far apart that they are very difficult to be connected by the corresponding linker; they are therefore highly disfavored by the geometry terms (Figure 3.8). Taken together, these results show that the domains can be unambiguously placed into the density map.

We completed the structure by building the linkers between domains using RosettaCM, and performed density-guided all-atom refinement in the context of the full assembly [34]. The resulting symmetric model of the Pex1/Pex6 complex showed good agreement with the density

map, particularly in the N and the D1 domains. To improve the model for the D2 ring, we used the unsymmetrized density map, which contains better resolved density in the D2 ring than the symmetrized map. Further refinement of individual D2 domains, followed by reassembly into the full-length proteins and refinement of the full-length proteins, resulted in a model for the entire complex. This final model fits well into the entire unsymmetrized EM density map (Figure 3.10). Using the novel multi-component complex assembly method, our models determine the alternative assignment of Pex1 and Pex6, and predict that the N1 domain of Pex1 is missing.

3.3.1 *Confirmation of the structures determined by the novel method*

With the models of the Pex1/Pex6 in hands, our collaborators, the labs of Walz and Rapoport, carried out a series of experiments to validate the models. They noticed that treatment of the Pex1/Pex6 complex with elastase or trypsin truncated Pex1, but left Pex6 intact (Figure 3.9A). They carried out mass spectrometry on fragments generated with these proteases, and found that Pex1 was truncated at the N terminus, resulting in a fragment that lacks only the N1 domain. They therefore purified the elastase-treated Pex1/Pex6 complex by gel filtration (Figure. 3.9A, the green arrow) and used it for cryo-EM analysis. A low-resolution map, generated from 5,000 particles, was completely superimposable onto a map of the full-length complex low-pass filtered to the same resolution (Figure 3.9). The similarity of the structures of the proteolyzed and nonproteolyzed samples provides strong evidence that our model correctly assigned the domains of Pex1 and Pex6. Through analyzing negative-stain EM images using 2D clustering with the iterative stable alignment and clustering procedure, they discovered that, in many of the classified images displayed additional density at varying locations on the side of the triangular complex (Figure 3.2C). The position of the domain is consistent with the expected location of the

linker to the N2 domain of Pex1. Altogether, the cryo-EM structures of the nonproteolyzed and proteolyzed Pex1/Pex6 complex, the 2D class averages of negative-stain images, and the modeling all support the conclusion that the N1 domain of Pex1 is attached through a flexible linker to the rest of the complex. In addition to it, our assignment of Pex1 and Pex6 is also in agreement with a recent low-resolution structure determined by negative-stain EM, in which the position of Pex6 was determined by fusion to the maltose-binding protein [58]. Finally, all the experimental data provide strong validation for the computational domain placement procedure.

3.4 CONCLUSION

We have developed a novel method for structure determination of multi-component macromolecular complexes from medium- to low- resolution cryo-EM maps. Using Pex1/Pex6 as an example, where experts had problems to determine the structure unambiguously, we establishes a precedent for generating a molecular model of a protein complex from a intermediate-resolution EM map. Whereas previous modeling approaches placed known crystal structures into the density map and used flexible fitting to improve the fit, our methodology uses density-guided comparative modeling of putative domain placements, followed by Monte Carlo sampling to identify the correct configuration of domains. This novel method allows modeling in cases where only structures of distantly related proteins are known or where domains of similar, but nonidentical, structure need to be placed into the density map. We anticipate that this approach will also be applicable to many other cases in which atomic resolution has been impossible to achieve.

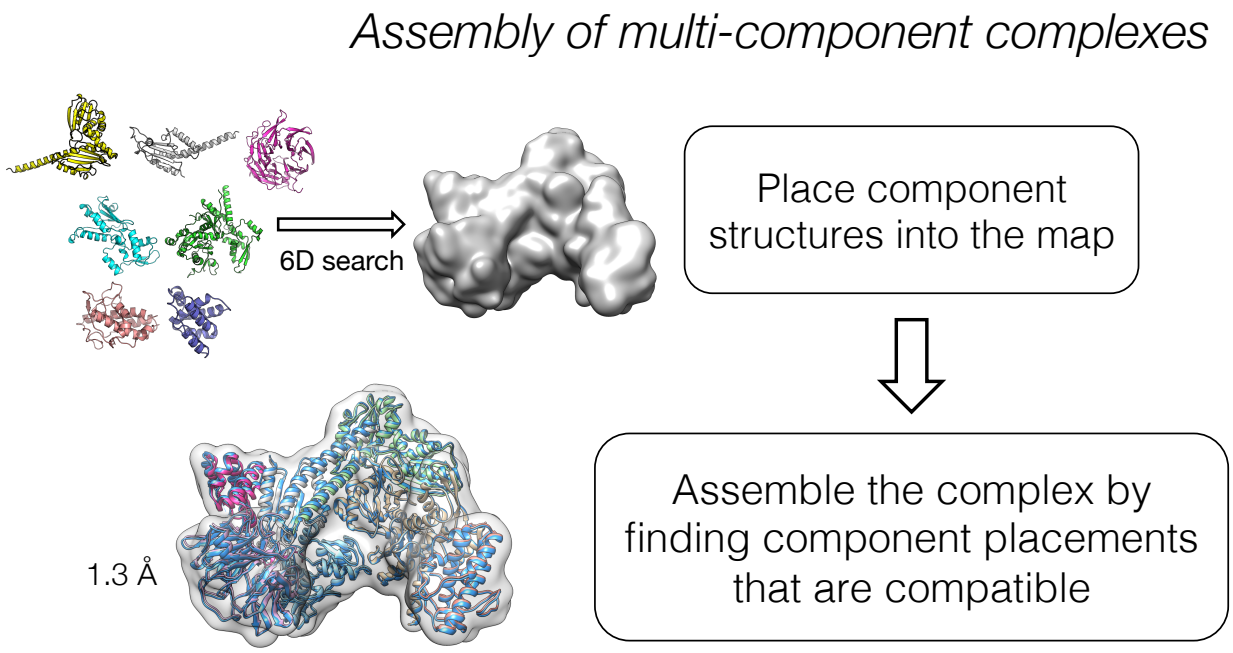

Figure 3.1. Concept of the algorithm used in the assembly of multi-component complexes.

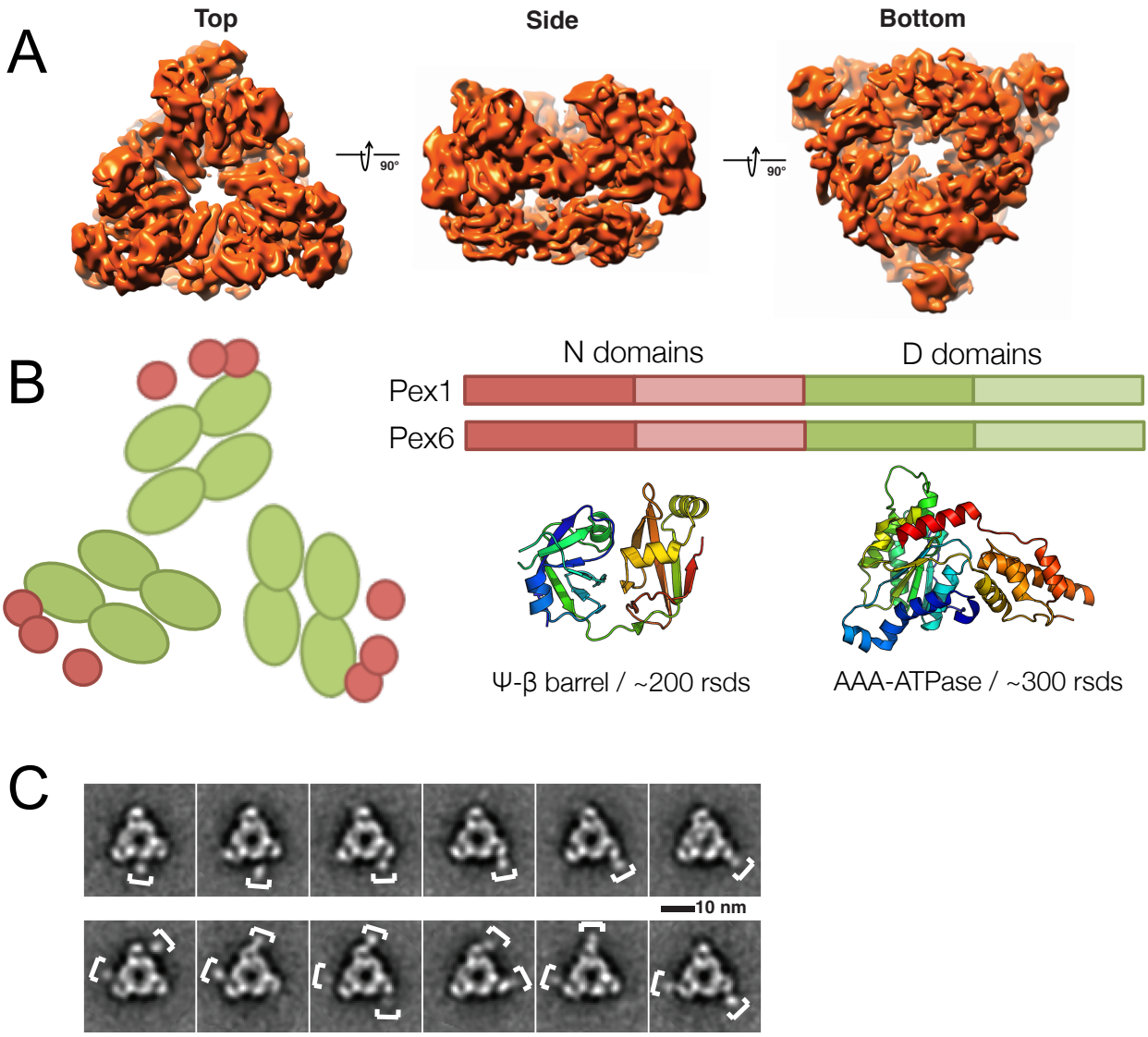

Figure 3.2. Challenges of structure determination for the Pex1/Pex6 protein complex. (A) Overview of the architecture of the Pex1/Pex6 complex. (B) Domain arrange of the Pex1/Pex6 complex with a cartoon representation of the trimer of heterodimeric Pex1/Pex6 complex. Pex1 and Pex6 proteins share a similar domain arrangement with two N domain at N-termius predicted to be a psi-beta barrel scaffold, and D domains at C-terminus predicted to be a AAA-ATPase fold. In the asymmetric unit, although predicted to contain 8 domains, there are only 7 domains present in the 3D cryo-EM reconstruction, containing three N-domains and four D-domains. One of the N domains from either Pex1 or Pex6 protein is missing. (C) 2D cluster images indicate that there is a flexible domain at the apex of the triangle-like Pex1/Pex6 complex.

*(A) and (C) are figures reproduced from Blok et al. *PNAS* (2015)[59]

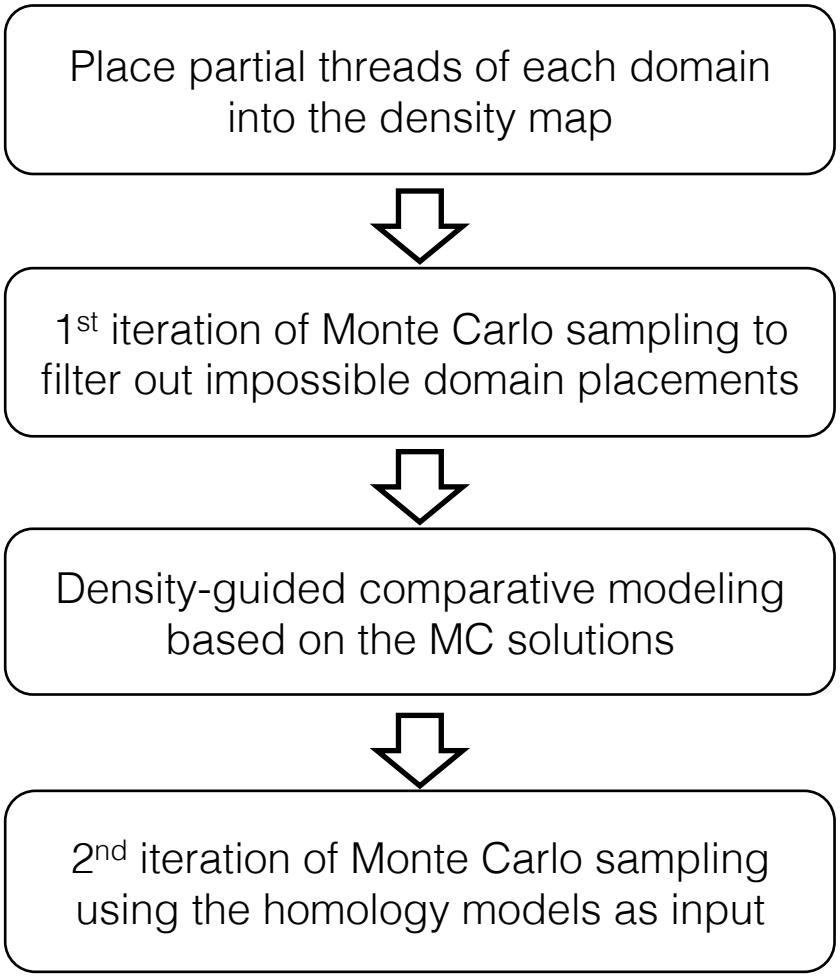

Figure 3.3. Protocol used in determining the structure for the Pex1/Pex6 complex.

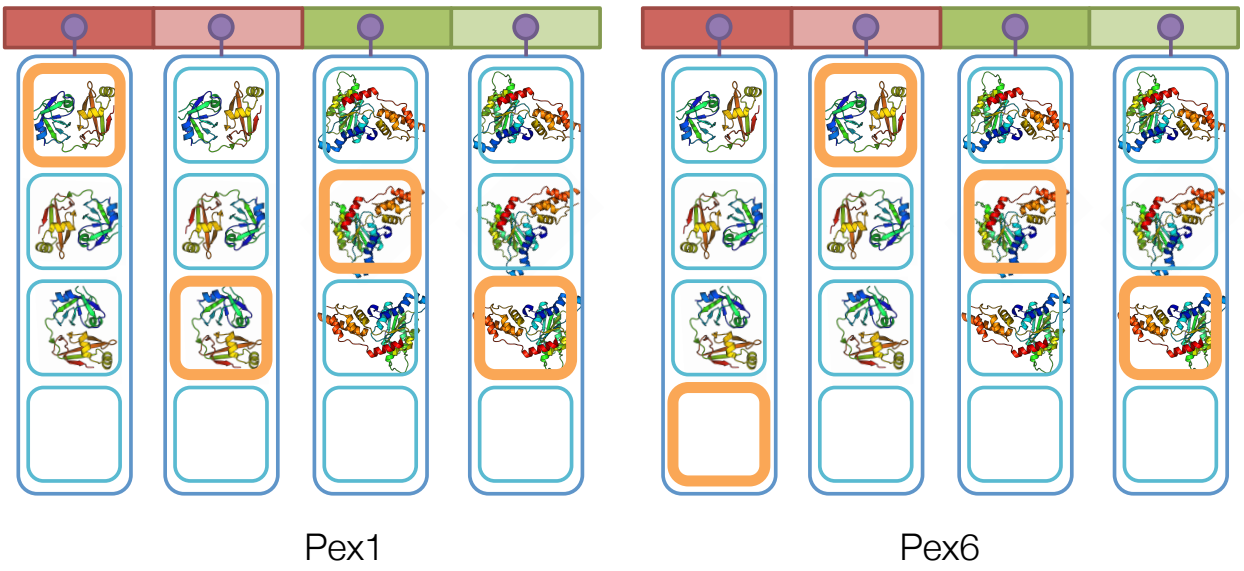

Figure 3.4. Monte Carlo sampling of structure placements.

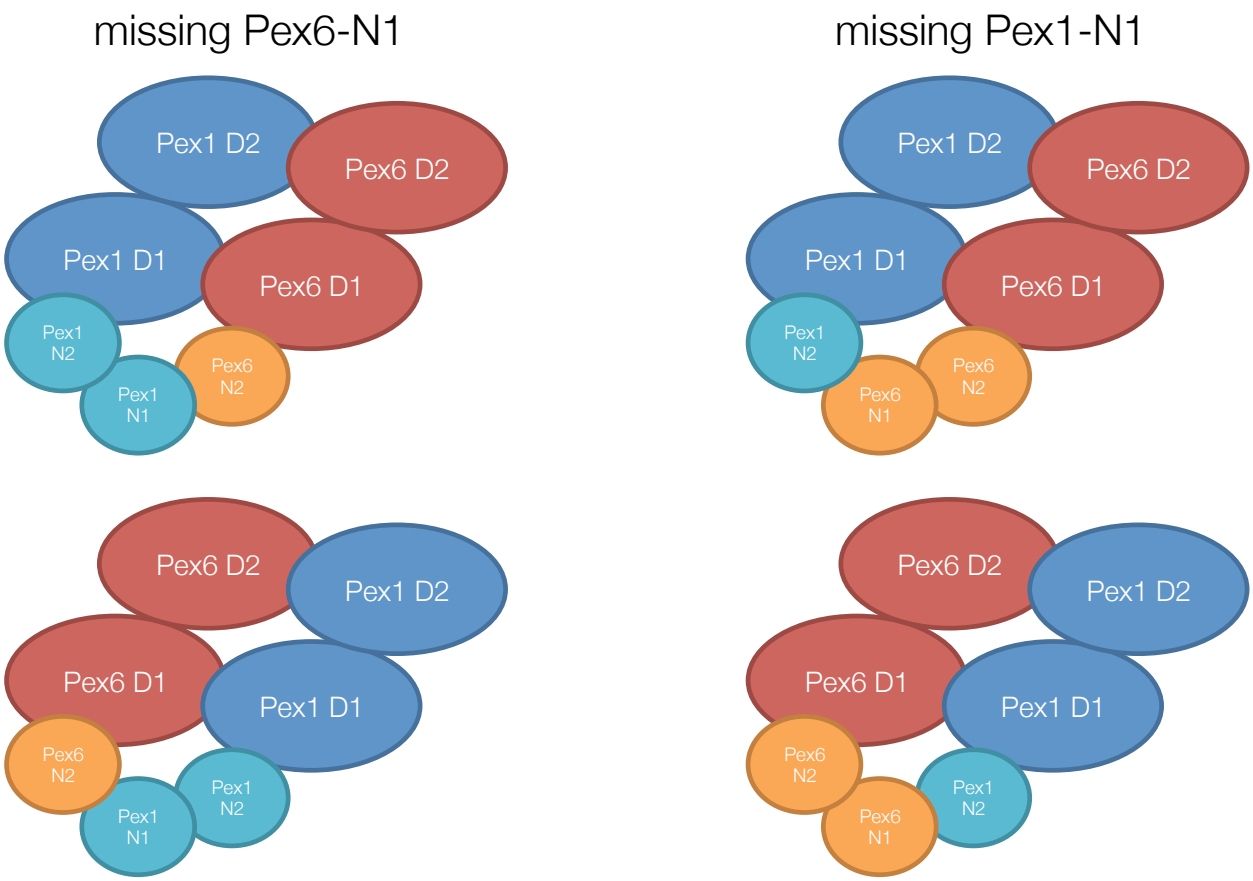

Figure 3.5. Four solutions emerged from the first round of Monte Carlo sampling using placements of partial thread as input.

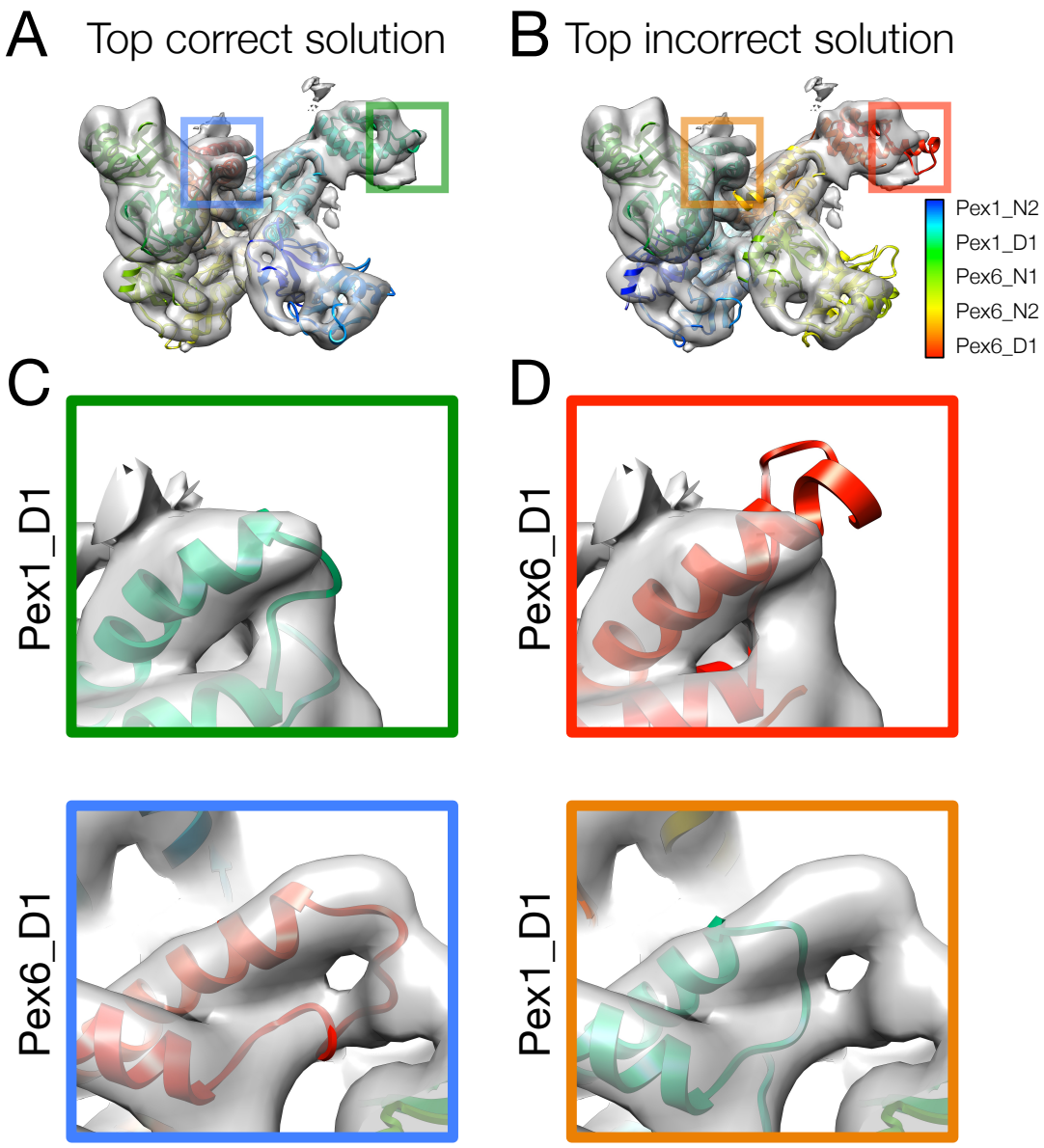

Figure 3.6.. Fit to the density map of the best domain placement and the top scoring incorrect placement.

(A) Fit of the best solution of domain placement into the overall density map of the Pex1/Pex6 complex. The domains are colored from blue to red as indicated in the scale. (B) As in A, but for the top incorrect solution. (C) A close-up view of two regions of the density map, demonstrating that the best solution fits the density map. (D) As in (C), but for the top incorrect solution. Note that in this case, the model is outside of density in one region (top) and does not fully account for all the observed density in the other (bottom).

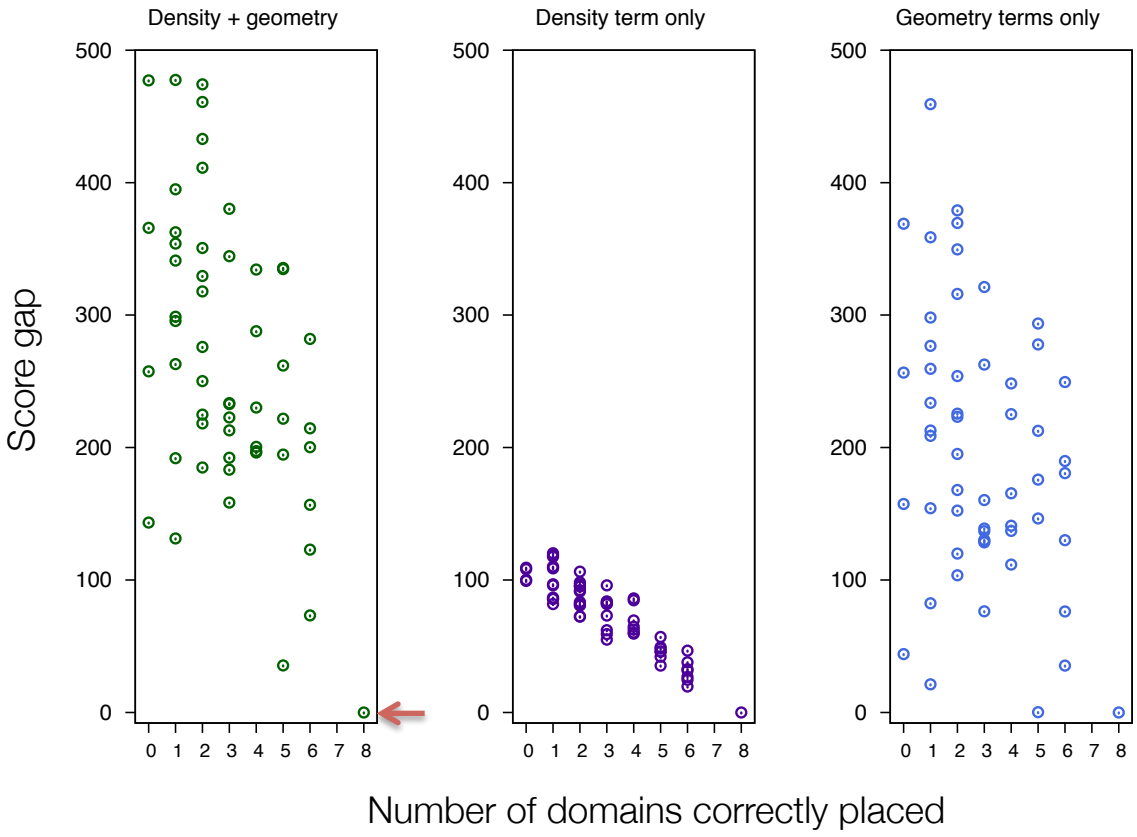

Figure 3.7. Score contribution of density and geometry terms in domain assembly.
The left panel shows the score difference between the best solution (highlighted with a red arrow) and all other solutions visited by Monte Carlo trajectories before convergence (converged states shown in Table 3.2). The middle and right panels shows for the same set of models plots of the score contribution of the density and geometry terms alone ($score_{density}$ and $score_{closability}$ plus $score_{clash}$, respectively).

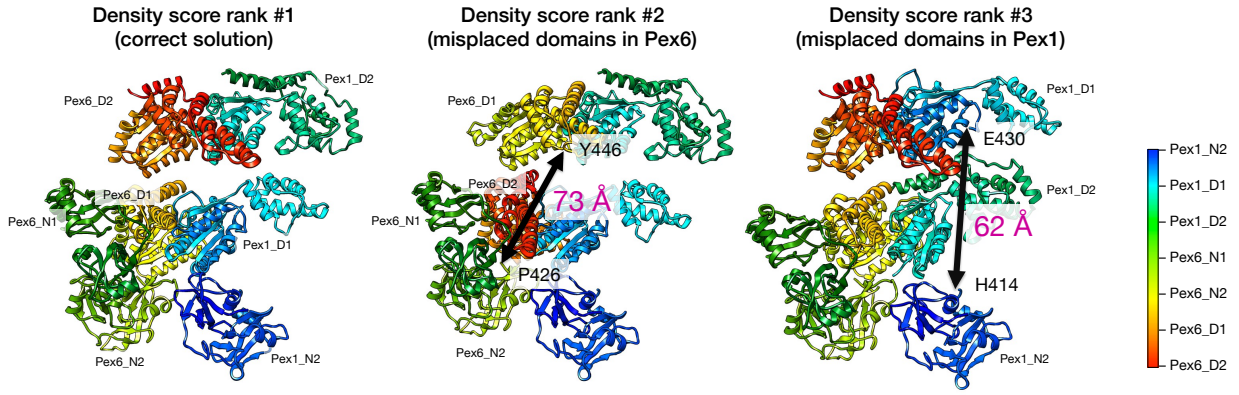

Figure 3.8. Comparison of the top three sampled domain placements using only fit-to-density as criterion.

The top three sampled solutions are shown using the density score term alone ($score_{density}$), without geometry terms. The domain placements in the asymmetric unit of the Pex1/Pex6 complex are shown with a rainbow color scheme indicated on the far right. The left panel shows the determined solution. The right two panels show the top two incorrect solutions by density-fit alone. Both of these solutions feature termini placements that are impossible to close with linkers of the given length. In the structure shown in the middle, the distance from the C-terminus of the Pex6 N2 domain (P426) to the N-terminus of the D1 domain (Y446) is ~73 Å. In the right structure, the distance from the C-terminus of the Pex1 N2 domain (H414) to the N-terminus of the D1 domain (E430) is ~62 Å.

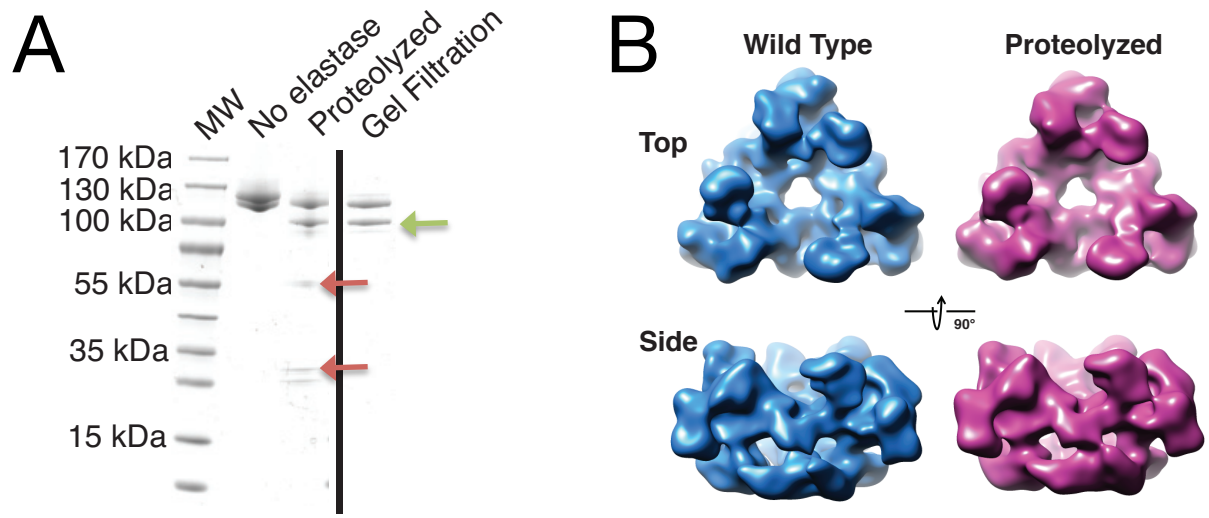

Figure 3.9. Experimental verification of the model for the Pex1/Pex6 complex.
(A) SDS-PAGE of the sample of the Pex1/Pex6 complex under treatments of proteases. After proteolysis experiment, fragments are shown on the lane labeled as "Proteolyzed". Mass spectroscopy analysis indicated that the fragments belong to the N1 domain of Pex1 (red arrows), which is consistent with the model determined by the novel computational method. The sample with N1 domain of Pex1 cleaved was subjected to gel-filtration (green arrow), and was examined using single-particle reconstruction. (B) Comparison of the cryo-EM reconstructions of the wild-type (blue) and proteolyzed (purple) Pex1/Pex6 complex.

* The entire figures were reproduced from Blok et al. *PNAS* (2015) [59]

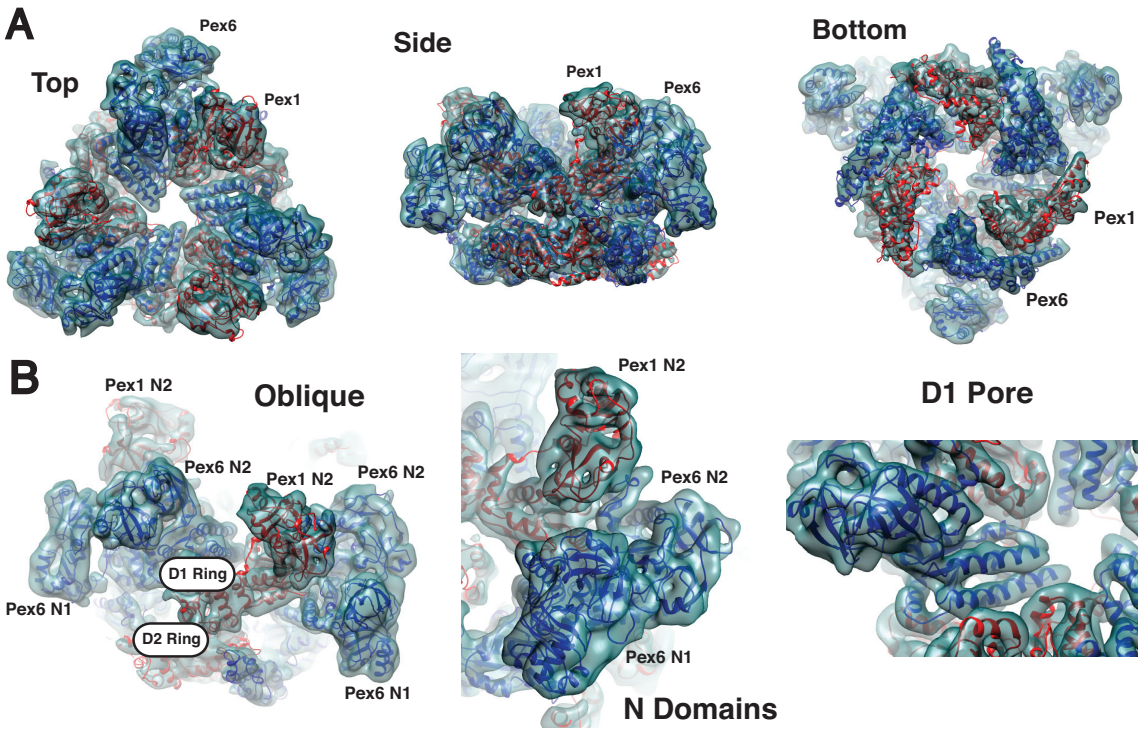

Figure 3.10. The asymmetric structure of the Pex1/Pex6 complex.
(A) The overall structure of the Pex1/Pex6 determined at the asymmetric state. (B) The quality of the structure determined using Rosetta. As shown in the figure, almost all the density can be explained using Rosetta structural modeling techniques.

* The entire figure is adopted from Blok et al. PNAS (2015) [59]

Table 3.1. Comparison of scores of different domain placements after the second round of Monte Carlo assembly.

Placement	Score	Number of decoys	Score – Score (best)
A:0 B:5 C:6 D:7 E:1 F:3 G:8 H:9	**-912**	**3489**	**0**
A:0 B:3 C:8 D:9 E:1 F:5 G:6 H:7	-806	3214	106
A:0 B:3 C:8 D:7 E:1 F:5 G:6 H:9	-769	3180	143
A:0 B:5 C:8 D:9 E:1 F:3 G:6 H:7	-752	1980	160
A:0 B:5 C:8 D:7 E:1 F:3 G:6 H:9	-715	1991	197
A:0 B:5 C:6 D:9 E:1 F:3 G:8 H:7	-709	2028	203

The letters A-D refer to the N1, N2, D1, and D2 domains of Pex1, and the letters E-H to the corresponding domains of Pex6. The numbers refer to locations in the density map, with 0 indicating that the domain remained unplaced. All sampled domain placements are listed. The top solution is shown in bold. The score gap between a given domain placement and the best solution is also shown. The last column gives the number of trajectories sampling a given domain placement in the MC procedure.

Table 3.2. Total scores of the best and next best domain placements after the first and second rounds of Monte Carlo assembly.

Placement	Stage 1 Partial model assembly	Stage 2 Full-length model assembly
Correct solution	-1162	-912
Top incorrect solution	-1121	-806

The total scores for the best solution and the top incorrect solution were calculated after the first round of MC assembly using partial models for each domain and after the second round of assembly using full-length, refined homology models for these domains. Note that the score gap between the two solutions is significantly larger after the second round of assembly, although the same solution emerges as the best.

BIBLIOGRAPHY

1. Henderson R: **Realizing the potential of electron cryo-microscopy.** *Quarterly Reviews of Biophysics* (2004) **37**(1):3-13.

2. Glaeser RM: **Review: Electron crystallography: Present excitement, a nod to the past, anticipating the future.** *Journal of structural biology* (1999) **128**(1):3-14.

3. Fischer N, Neumann P, Konevega AL, Bock LV, Ficner R, Rodnina MV, Stark H: **Structure of the e. Coli ribosome-ef-tu complex at <3 a resolution by cs-corrected cryo-em.** *Nature* (2015) **520**(7548):567-570.

4. Bartesaghi A, Merk A, Banerjee S, Matthies D, Wu X, Milne JL, Subramaniam S: **2.2 a resolution cryo-em structure of beta-galactosidase in complex with a cell-permeant inhibitor.** *Science* (2015) **348**(6239):1147-1151.

5. Bradley P, Misura KM, Baker D: **Toward high-resolution de novo structure prediction for small proteins.** *Science* (2005) **309**(5742):1868-1871.

6. Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA *et al*: **Nmr structure determination for larger proteins using backbone-only data.** *Science* (2010) **327**(5968):1014-1018.

7. Baker ML, Jiang W, Wedemeyer WJ, Rixon FJ, Baker D, Chiu W: **Ab initio modeling of the herpesvirus vp26 core domain assessed by cryoem density.** *PLoS computational biology* (2006) **2**(10):e146.

8. DiMaio F, Tyka MD, Baker ML, Chiu W, Baker D: **Refinement of protein structures into low-resolution density maps using rosetta.** *Journal of molecular biology* (2009) **392**(1):181-190.

9. Zhang X, Jin L, Fang Q, Hui WH, Zhou ZH: **3.3 a cryo-em structure of a nonenveloped virus reveals a priming mechanism for cell entry.** *Cell* (2010) **141**(3):472-482.

10. Zhang X, Settembre E, Xu C, Dormitzer PR, Bellamy R, Harrison SC, Grigorieff N: **Near-atomic resolution using electron cryomicroscopy and single-particle reconstruction.** *Proceedings of the National Academy of Sciences of the United States of America* (2008) **105**(6):1867-1872.

11. Veesler D, Ng TS, Sendamarai AK, Eilers BJ, Lawrence CM, Lok SM, Young MJ, Johnson JE, Fu CY: **Atomic structure of the 75 mda extremophile sulfolobus turreted icosahedral virus determined by cryoem and x-ray crystallography.** *Proceedings of the National Academy of Sciences of the United States of America* (2013) **110**(14):5504-5509.

12. Grigorieff N, Harrison SC: **Near-atomic resolution reconstructions of icosahedral viruses from electron cryo-microscopy.** *Current opinion in structural biology* (2011) **21**(2):265-273.

13. Hryc CF, Chen DH, Chiu W: **Near-atomic-resolution cryo-em for molecular virology.** *Current opinion in virology* (2011) **1**(2):110-117.

14. Li X, Mooney P, Zheng S, Booth CR, Braunfeld MB, Gubbens S, Agard DA, Cheng Y: **Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-em.** *Nature methods* (2013) **10**(6):584-590.

15. Allegretti M, Mills DJ, McMullan G, Kuhlbrandt W, Vonck J: **Atomic model of the f420-reducing [nife] hydrogenase by electron cryo-microscopy using a direct electron detector.** *eLife* (2014) **3**(e01963.

16. Liao M, Cao E, Julius D, Cheng Y: **Structure of the trpv1 ion channel determined by electron cryo-microscopy.** *Nature* (2013) **504**(7478):107-112.

17. Amunts A, Brown A, Bai XC, Llacer JL, Hussain T, Emsley P, Long F, Murshudov G, Scheres SH, Ramakrishnan V: **Structure of the yeast mitochondrial large ribosomal subunit.** *Science* (2014) **343**(6178):1485-1489.

18. Bai XC, Fernandez IS, McMullan G, Scheres SH: **Ribosome structures to near-atomic resolution from thirty thousand cryo-em particles.** *eLife* (2013) **2**(e00461.

19. Lu P, Bai XC, Ma D, Xie T, Yan C, Sun L, Yang G, Zhao Y, Zhou R, Scheres SH, Shi Y: **Three-dimensional structure of human gamma-secretase.** *Nature* (2014) **512**(7513):166-170.

20. Scheres SH: **Beam-induced motion correction for sub-megadalton cryo-em particles.** *eLife* (2014) **3**(e03665.

21. Wriggers W, Milligan RA, McCammon JA: **Situs: A package for docking crystal structures into low-resolution maps from electron microscopy.** *Journal of structural biology* (1999) **125**(2-3):185-195.

22. Rossmann MG, Bernal R, Pletnev SV: **Combining electron microscopic with x-ray crystallographic structures.** *Journal of structural biology* (2001) **136**(3):190-200.

23. Trabuco LG, Villa E, Mitra K, Frank J, Schulten K: **Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics.** *Structure* (2008) **16**(5):673-683.

24. DiMaio F, Song Y, Li X, Brunner M, Xu C, Conticello V, Egelman E, Marlovits T, Cheng Y, Baker D: **Atomic accuracy models from 4.5 å cryo-electron microscopy data with density-guided iterative local rebuilding and refinement.** *Nature methods* (2014) **(In Press)**(

25. Zhou ZH: **Towards atomic resolution structural determination by single-particle cryo-electron microscopy.** *Current opinion in structural biology* (2008) **18**(2):218-228.

26. Cowtan K: **The buccaneer software for automated model building. 1. Tracing protein chains.** *Acta crystallographica Section D, Biological crystallography* (2006) **62**(Pt 9):1002-1011.

27. Terwilliger TC, Grosse-Kunstleve RW, Afonine PV, Moriarty NW, Zwart PH, Hung LW, Read RJ, Adams PD: **Iterative model building, structure refinement and density modification with the phenix autobuild wizard.** *Acta crystallographica Section D, Biological crystallography* (2008) **64**(Pt 1):61-69.

28. Langer G, Cohen SX, Lamzin VS, Perrakis A: **Automated macromolecular model building for x-ray crystallography using arp/warp version 7.** *Nature protocols* (2008) **3**(7):1171-1179.

29. Baker MR, Rees I, Ludtke SJ, Chiu W, Baker ML: **Constructing and validating initial calpha models from subnanometer resolution density maps with pathwalking.** *Structure* (2012) **20**(3):450-463.

30. Lindert S, Staritzbichler R, Wotzel N, Karakas M, Stewart PL, Meiler J: **Em-fold: De novo folding of alpha-helical proteins guided by intermediate-resolution electron microscopy density maps.** *Structure* (2009) **17**(7):990-1003.

31. Baker ML, Abeysinghe SS, Schuh S, Coleman RA, Abrams A, Marsh MP, Hryc CF, Ruths T, Chiu W, Ju T: **Modeling protein structure at near atomic resolutions with gorgon.** *Journal of structural biology* (2011) **174**(2):360-373.

32. Gront D, Kulp DW, Vernon RM, Strauss CE, Baker D: **Generalized fragment picking in rosetta: Design, protocols and applications.** *PloS one* (2011) **6**(8):e23294.

33. DiMaio FP, Soni AB, Phillips GN, Jr., Shavlik JW: **Spherical-harmonic decomposition for molecular recognition in electron-density maps.** *International journal of data mining and bioinformatics* (2009) **3**(2):205-227.

34. Song Y, DiMaio F, Wang RY, Kim D, Miles C, Brunette T, Thompson J, Baker D: **High-resolution comparative modeling with rosettacm.** *Structure* (2013) **21**(10):1735-1742.

35. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D: **Design of a novel globular protein fold with atomic-level accuracy.** *Science* (2003) **302**(5649):1364-1368.

36. Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, Headd JJ, Hung LW, Kapral GJ, Grosse-Kunstleve RW, McCoy AJ *et al*: **Phenix: A comprehensive python-based system for macromolecular structure solution.** *Acta crystallographica Section D, Biological crystallography* (2010) **66**(Pt 2):213-221.

37. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, Keegan RM, Krissinel EB, Leslie AG, McCoy A, McNicholas SJ *et al*: **Overview of the ccp4 suite and current developments.** *Acta crystallographica Section D, Biological crystallography* (2011) **67**(Pt 4):235-242.

38. Rohl CA, Strauss CE, Misura KM, Baker D: **Protein structure prediction using rosetta.** *Methods in enzymology* (2004) **383**(66-93.

39. Mills DJ, Vitt S, Strauss M, Shima S, Vonck J: **De novo modeling of the f420-reducing [nife]-hydrogenase from a methanogenic archaeon by cryo-electron microscopy.** *eLife* (2013) **2**(e00218.

40. Kudryashev M, Wang RY, Brackmann M, Scherer S, Maier T, Baker D, DiMaio F, Stahlberg H, Egelman EH, Basler M: **Structure of the type vi secretion system contractile sheath.** *Cell* (2015) **160**(5):952-962.

41. DiMaio F, Leaver-Fay A, Bradley P, Baker D, Andre I: **Modeling symmetric macromolecular structures in rosetta3.** *PloS one* (2011) **6**(6):e20450.

42. Chen VB, Arendall WB, 3rd, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC: **Molprobity: All-atom structure validation for macromolecular crystallography.** *Acta crystallographica Section D, Biological crystallography* (2010) **66**(Pt 1):12-21.

43. Clemens DL, Ge P, Lee BY, Horwitz MA, Zhou ZH: **Atomic structure of t6ss reveals interlaced array essential to function.** *Cell* (2015) **160**(5):940-951.

44. Lange OF, Baker D: **Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation.** *Proteins: Structure, Function, and Bioinformatics* (2012) **80**(3):884-895.

45. Lange OF, Rossi P, Sgourakis NG, Song Y, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D: **Determination of solution structures of proteins up to 40 kda using cs-rosetta with sparse nmr data from deuterated samples.** *Proceedings of the National Academy of Sciences of the United States of America* (2012) **109**(27):10873-10878.

46. Kalisman N, Adams CM, Levitt M: **Subunit order of eukaryotic tric/cct chaperonin by cross-linking, mass spectrometry, and combinatorial homology modeling.** *Proc Natl Acad Sci U S A* (2012) **109**(8):2884-2889.

47. Ovchinnikov S, Kamisetty H, Baker D: **Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information.** *eLife* (2014) **3**(e02030.

48. Kamisetty H, Ovchinnikov S, Baker D: **Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era.**

Proceedings of the National Academy of Sciences of the United States of America (2013) **110**(39):15674-15679.

49. Lasker K, Topf M, Sali A, Wolfson HJ: **Inferential optimization for simultaneous fitting of multiple components into a cryoem map of their assembly.** *Journal of molecular biology* (2009) **388**(1):180-194.

50. Zhang S, Vasishtan D, Xu M, Topf M, Alber F: **A fast mathematical programming procedure for simultaneous fitting of assembly components into cryoem density maps.** *Bioinformatics* (2010) **26**(12):i261-268.

51. Zhang J, Ma B, DiMaio F, Douglas NR, Joachimiak LA, Baker D, Frydman J, Levitt M, Chiu W: **Cryo-em structure of a group ii chaperonin in the prehydrolysis atp-bound state leading to lid closure.** *Structure* (2011) **19**(5):633-639.

52. Loquet A, Sgourakis NG, Gupta R, Giller K, Riedel D, Goosmann C, Griesinger C, Kolbe M, Baker D, Becker S, Lange A: **Atomic model of the type iii secretion system needle.** *Nature* (2012) **486**(7402):276-279.

53. Soding J: **Protein homology detection by hmm-hmm comparison.** *Bioinformatics* (2005) **21**(7):951-960.

54. Peng J, Xu J: **A multiple-template approach to protein threading.** *Proteins* (2011) **79**(6):1930-1939.

55. Yang Y, Faraggi E, Zhao H, Zhou Y: **Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates.** *Bioinformatics* (2011) **27**(15):2076-2082.

56. Kirkpatrick S, Gelatt CD, Jr., Vecchi MP: **Optimization by simulated annealing.** *Science* (1983) **220**(4598):671-680.

57. DiMaio F, Song Y, Li X, Brunner MJ, Xu C, Conticello V, Egelman E, Marlovits TC, Cheng Y, Baker D: **Atomic-accuracy models from 4.5-a cryo-electron microscopy data with density-guided iterative local refinement.** *Nature methods* (2015) **12**(4):361-365.

58. Gardner BM, Chowdhury S, Lander GC, Martin A: **The pex1/pex6 complex is a heterohexameric aaa+ motor with alternating and highly coordinated subunits.** *Journal of molecular biology* (2015) **427**(6 Pt B):1375-1388.

59. Blok NB, Tan D, Wang RY, Penczek PA, Baker D, DiMaio F, Rapoport TA, Walz T: **Unique double-ring structure of the peroxisomal pex1/pex6 atpase complex revealed by cryo-electron microscopy.** *Proceedings of the National Academy of Sciences of the United States of America* (2015) **112**(30):E4017-4025.

VITA

Yu-Ruei Wang was born in Taipei, Taiwan. He received a Bachelor of Science degree in 2007 from National Taiwan University in Biochemistry. He carried out his undergraduate research in the lab of Dr. Andrew H.-J. Wang at Institute of Biological Chemistry, Academia Sinica, Taiwan, where he used X-ray crystallography to study enzymes and have thus developed a keen interest in structural biology. Spending a year in the mandatory military service after graduating from college, in 2008 he took a technician job in the lab of Dr. Che Ma at Genomic Research Center, Academia Sinica, Taiwan, where he conducted X-ray crystallography studies on membrane proteins and started applying to graduate schools in the States. He got admitted and decided to enter the graduate program in Biomolecular Structure and Design (now Biological Physics, Structure and Design) at the University of Washington in Seattle in 2009 and joined Dr. David Baker's group in 2010. In 2015, he received his Doctoral of Philosophy degree from University of Washington in Biochemistry.