Protein Complex Structure Determination Guided by Low-Resolution Cryo-Electron Microscopy Maps

Daniel P. Farrell

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2021

Reading Committee:
Frank DiMaio, Chair
Justin M. Kollman
Philip H. Bradley

Program Authorized to Offer Degree:
Biochemistry

University of Washington

Abstract

Protein Complex Structure Determination Guided by Low-Resolution Cryo-Electron Microscopy Maps

Daniel P. Farrell

Chair of the Supervisory Committee:
Frank DiMaio
Biochemistry

Cryo-electron microscopy of protein complexes often leads to moderate resolution maps (4-8 Å), with visible secondary structure elements but poorly resolved loops, making model-building challenging. In the absence of high-resolution structures of homologues, only coarse-grained structural features are typically inferred from these maps, and it is often impossible to assign specific regions of density to individual protein subunits. This dissertation describes a new method for overcoming these difficulties that integrates predicted residue distance distributions from a deep-learned convolutional neural network, computational protein folding using Rosetta, and automated EM-map-guided complex assembly. We will show how this method performs on a diverse benchmarking dataset in addition to describing how it was used to build models for three difficult protein complexes that would have been impossible to solve without this software. We anticipate that our approach will be broadly useful for cryoEM structure determination of large complexes containing many subunits for which there are no homologues of known structure.

Acknowledgements

I would like to thank Frank DiMaio for his mentorship and support throughout my thesis. He has very patiently taught me how to code at the most basic levels and has provided input and direction on numerous aspects of my scientific career. It is rare in life to find a person who always has an answer to your problems, but I feel fortunate to have found that in my thesis advisor.

I would also like to say thank you to the members of the DiMaio lab. I was lucky enough to be coached in the nascency of my thesis by the first generation DiMaio lab members Ryan Pavlovic and Brandon Frenz. Both of you helped jumpstart my academic career, and helped provide a stable foundation for what I would learn throughout the remainder of my thesis. I look forward to staying in touch with both of you and continuing our friendship. I spent the second half of my thesis interacting with the second generation of the DiMaio lab, Carson Adams, Andrew Muenks, Gabriella Reggiano, and Guangfeng Zhou. All four of you have become integral parts of my daily life, and I will miss you all dearly. The unparalleled environment you have all helped build has provided the structure for deep scientific, and *less*-deep unscientific discussion. The discussions and “arguments" we have all had throughout the years will be looked upon fondly, and I can only hope that we have the opportunity to have more in the future. Perhaps one day we can even decide why people think that salted caramels should have salt in them, instead of on them.

I would also like to thank all of my committee members: David Baker, Phil Bradley, Doug Fowler, and Justin Kollman for their direction throughout the course of my thesis.

Finally I would like to thank my fiancé Kayla who has been a part of my life for over 10 years. I will forever be indebted to her for sacrificing the job that she loved to bet on our relationship and to move to Seattle to be with me. I am so happy that I have had you in my life throughout grad school, and am looking forward to continuing our life together! I would also like to especially thank you for your kindness, generosity, and understanding during these last few busy months of grad school, words cannot describe how much I appreciate it.

Dedication

I dedicate this thesis to my fiancé Kayla

Chapter 1. Introduction

Living cells and organisms rely on a myriad of complex biochemical reactions that are driven by a variety of biomolecules and biomolecular structures. How, at the molecular level, these biomolecules work together to perform their duties can be elucidated by experimentally obtaining their atomic structure. Protein atomic structure determination has chiefly been accomplished through X-ray crystallography, which is responsible for approximately 90% of all structures deposited into the protein data bank[1]. X-ray crystallography is a high resolution technique, with the average deposited crystal structure resolution hovering near 2.5 Å. Although widely utilized, this structural determination technique requires proteins to be crystallized. The process of protein crystallization often requires a significant quantity of highly purified protein, and crystallization may be impossible due to intrinsic protein instability, or a variety of other reasons. As an alternative to X-ray crystallography, Cryo electron microscopy (cryoEM) has recently become a substantial contributor of low, medium, and high resolution protein structures.

First conceived in the 1970s, cryoEM has long served as an alternative to X-ray crystallography. In cryoEM, biomolecules or biomolecular complexes are flash frozen in a thin layer of vitreous ice and then imaged using traditional transmission electron microscopy. In contrast to X-ray crystallography, cryoEM structure determination does not require a large quantity of highly purified protein and can be performed on highly dynamic macromolecular complexes. Although there are many advantages of solving protein structures with cryoEM, until recently (2014) the average resolution of a cryoEM structure has lagged far behind X-ray crystallography at 8.9 Å[1]. However, in recent years the average resolution has improved significantly, with the average resolution in 2020 being 3.9 Å. This dramatic increase in cryoEM resolution has been driven by the development of direct electron detector technologies in addition to improvements in image processing softwares.

Structures solved by cryoEM can reach up to resolutions of 1.15 Å[2], but are more typically solved in the near-atomic (3-5 Å) or subnanometer (5-9 Å) resolution ranges. Structures with resolution in the near atomic range typically provide information about the position of the protein backbone but only sparse information regarding the protein sidechains. The lack of side chain information in these structures makes it difficult to accurately assign amino acids to their respective electron density. In the subnanometer resolution range, even more information is lost and only information of the proteins'

secondary structure is distinguishable. Separately, these two tiers provide difficult challenges to overcome, however, cryoEM maps are often significantly heterogeneous. In maps with high levels of resolution heterogeneity, the resolution of various regions in the map can vary significantly[3]. Thus, experimental cryoEM maps of medium resolution that display resolution heterogeneity demand model building software that can overcome the modeling difficulties posed from a wide range of informational ambiguities.

While software for modeling protein into high resolution structures generated from X-ray crystallography has been developed for decades, software to model proteins at the lower resolutions provided from cryoEM has only become necessary in the past 5-6 years. A variety of tools have been developed that are designed to model into near atomic resolution cryoEM structures including: Rosetta *denovo_density*[4] and RosettaES[5], MAINMAST[6], Phenix *map_to_model*[7,8], and DeepTracer[9].

Rosetta *denovo_density* uses a three stage protocol where it first matches predicted backbone conformations to density, then it tries to combine these backbone fits to density in order to find maximally consistent subsets and generate partial models, these partial models are then completed using density guided sampling. RosettaES attempts to improve upon the final stage of *denovo_density*, especially in places where density guided sampling was inadequate at finding near-native backbone conformations. RosettaES does this using a beam search algorithm in addition to building missing sequence one residue at a time iteratively rather than attempting to model all missing sequence at once.

Phenix, MAINMAST, and DeepTracer all attempt to solve near-atomic resolution structures by first identifying Cα atoms, and then each use a different algorithm to expand this information into a final complete model. Phenix *map_to_model* implements a different approach whereby an ensemble of maps is first automatically generated by sharpening the input to maximize the connectivity and various details of the density. Then these sharpened maps are used to trace the density through the high contour density, and identify secondary structure. This information in combination with the side chain density is used to assign sequence to the traced chains. Finally the secondary structure information is then used to generate restraints and the model is refined in the context of the map.

MAINMAST uses similar methodology to Phenix where first Cα atom positions are predicted using a mean-shifting algorithm, then all predicted Cα positions are connected to determine a minimum

spanning tree, and finally a tabu-search algorithm refines the structure of the tree to determine the residue associated with each atom. Finally from these Cα positions a full-atom model can be constructed and then refined with Molecular Dynamics Flexible Fitting (MDFF).

DeepTracer is the newest of these protocols and employs a deep convolutional neural network in order to first predict Cα positions in cryoEM density, then using a custom tailored traveling salesman problem algorithm, these backbone atom positions are traced to determine their sequence identity, finally the backbone atom positions are predicted and side chains are added to maximize fit to density and minimize clashing. These tools all perform well, but their performance is significantly impacted in the presence of lower resolution data. As previously mentioned, as resolution approaches the subnanometer resolution-realm structures lose the high resolution information regarding the protein backbone, which is represented by density that only represents the protein secondary structure. As these methodologies all rely on protein backbone density to predict either backbone conformations or Cα positions none of them are well suited, or designed to work at lower resolutions than 5 Å.

In order to automatically build atomic level detail into 5 Å+ resolution cryoEM data all protocols currently developed require a starting model for each peptide present in the structure. Starting models generally are built using homology modeling if a solved model of similar sequence has already been deposited into the PDB, but without a known homologous model *de-novo* protein structure prediction is the only alternative. Models generated from either method are typically incorrect in some capacity and must be corrected in the context of the cryoEM density, but in some cases[10,11] are correct enough to be trivially docked manually. Once correctly oriented into the density there are many tools to use that exist to automatically refine protein models into medium resolution density such as: Rosetta[12], Molecular Dynamics Flexible Fitting (MDFF)[13], and FlexEM[14]. The Rosetta method of refining protein models in the context of cryoEM density relies on Monte Carlo sampling which drives a search for low-energy model conformations as determined by a fully atomic force field. To fit models into low resolution density MDFF uses all-atom molecular dynamics in concert with an external potential that is derived from the density. FlexEM relies on algorithms present in both the previous softwares that first takes advantage of Monte Carlo rigid fitting, then it uses conjugate-gradient based minimization to discover near native conformations, and finally performs simulated annealing molecular dynamics on subdivisions of the

protein to optimize local interactions. All of these methods perform well given correctly oriented starting models, but only FlexEM is capable of orienting the starting model itself. Depending on the resolution of the map, the quality of the starting model, and the number of peptide chains present manually fitting starting models into low resolution cryoEM density can be laborious, time consuming, and error prone. To solve this, developers have for many years attempted to create software that will perform multi-body protein placement.

Multi-body protein placement into cryoEM density requires an algorithm to accurately orient multiple, usually inaccurate, protein models into medium resolution density. As a maps resolution approachs 10 Å protein secondary structure is no longer distinguishable from the overall topology of the protein or protein complex. This provides another difficult tier of protein complex modeling - where only the approximate topology is known. There is a handful of software that has been developed to address this problem such as: MultiFit[15,16] , Prism-EM[17], and γ-TEMPy[18]. MultiFit is a part of the Integrative Modeling Package (IMP) and works to fit multiple proteins or protein domains into low resolution density by first segmenting the density map with a gaussian mixture model. MultiFit then exhaustively attempts to place all input models in each of the placed gaussians and once the optimal gaussians are identified for each subunit a local fitting procedure optimizes each model's fit to density. Finally, interfaces are optimized with a docking algorithm that is driven via shape complementarity principles. Prism-EM is built off of PRISM (PRotein Interactions by Structural Matching) which uses known protein interfaces to find interface hotspots and then uses these hotspots to drive protein-protein docking (without density). These protein-protein interface pairs are then iterated alongside coarse docking into cryoEM density and after every iteration the algorithm is given the opportunity to add another starting model to the complex. This is repeated until no more starting models are added at the end of the iteration. The newest of these algorithms is γ-TEMPy which was published in 2015 and this program uses a genetic algorithm to optimize starting model fits to density after initially placing them onto areas of strong density. The genetic algorithm works iteratively and each generation it perturbs subunits in all directions in addition to altering their rotation depending. This can be performed indefinitely, but once the populations converge Flex-EM can be applied to the entire complex to optimize the interfaces of all present models. Multi-body protein docking of inaccurate models into low resolution cryoEM density is a difficult task and for software to be

successful it has to do so with very little high resolution information. In order to subsidize this lack of information PRISM-EM and IMP extend their scoring functions to include information such as shape complementarity or experimental distance restraints derived from cross linking mass-spectrometry (XLMS). To properly perform multi-body protein docking all of these approaches utilize a combination of coarse molecular representations in order to optimize each peptide's fit to density and minimize the amount of clashes between fitted peptides. To improve upon these methods, we have developed a new protocol which will introduce pairwise interface optimization with backbone flexibility in combination with an exhaustive specialized rigid body docking protocol to generate accurate protein complex assemblies.

In this dissertation, we introduce a novel end-to-end methodology for determining complex structures guided by subnanometer cryoEM reconstructions. This methodology combines state of the art *de novo* protein domain structure prediction, a novel flexible complex assembly algorithm, and all-atom protein complex refinement. We will show that an atomic representation in addition to backbone flexibility and interface optimization driven by the physically realistic forefield of the Rosetta molecular modeling toolkit will lay the foundation for the next generation of cryoEM modeling. In addition to being benchmarked on both experimental and simulated data, our protocol Rosetta OATS (Optimal Assembly of Transformed Subunits) has been utilized to solve 3 protein complexes that would have been impossible using previously published methodologies. First I will describe the method, algorithms, and inputs to OATS, then I will detail OATSs performance on an extensive cryoEM benchmark dataset, and finally I will describe how OATS was used to model the Bardet-Biedl syndrome complex, the Fanconi Anemia Core Complex and the Homotypic Fusion and Vacuole Protein Sorting complex.

Chapter 2. Methodology

In order to model protein complexes at a multitude of resolutions we have developed a multi-stage algorithm that builds on work originally intended for only near atomic resolution structures. As a general overview of the method (Figure 2.1) we first generate predictions of every domain for every peptide chain using either homology modeling via RosettaCM or *de novo* structure prediction via *trRosetta*. Next we individually dock each model into the cryoEM density, and then we use Rosetta to optimize pairwise interactions between each docked model and use information derived from these optimized model pairs in concert with a Monte-Carlo simulated annealing algorithm to assemble the pairs into an atomic model. Finally, we use RosettaCM in order to complete any remaining loops and optimize each peptide's fit in the context of each other.

2.1 Initial Model building

Generation of initial starting models for multi-body cryoEM complex assembly has traditionally been performed via homology modeling. Protein structures of proteins with similar sequences often have similar structures. In homology modeling this relationship is exploited by first searching the target protein sequence against a database of all sequences with known structures and if any sequences are found to be similar a sequence alignment is generated between the target and template sequences. This sequence alignment can then be used to build a model where the target sequence is mapped onto the template model, which is called a "threaded template". Finally, the homology modeling protocol must close any small gaps in the model caused by insertions or deletions in the sequence alignment and refine the model. Without a solved structure that has a similar sequence to the target sequence, homology modeling is impossible. As an alternative to homology modeling, *de novo* structure prediction has - in the past - lacked accuracy compared to homology modeling, but had moderate success with small (<~200 residue) protein domains[19–21]. However, recent improvements in *de novo* structure prediction driven by deep learning has caused a paradigm shift in the free-modeling of protein structure, and now *de novo* structure prediction can - in some cases - exceed the accuracy of homology modeling[22,23]. Taken together, these two methods for initial model building serve as integral parts of the multi-body cryoEM complex assembly pipeline.

To perform homology modeling we use RosettaCM[24] a comparative modeling protocol distributed through the Rosetta molecular modeling toolkit[25]. RosettaCM incorporates information from multiple threaded templates in addition to 9 residue predicted fragments in order to generate accurate atomic homology models. First, inter-residue distance restraints are built based on the distances between residue Cα atoms in the threaded templates. Then, using these restraints, a random threaded template is selected as a starting model and a full length model is generated using Monte Carlo sampling in the context of the Rosetta *centroid* energy function. Then local geometries are optimized and loops are closed with iterations of fragment superposition and gradient based energy minimization, where the fragments are generated from Rosetta and taken from all threaded templates. Finally the entire structure is subjected to fullatom and backbone refinement with Rosetta *relax*[26]. More details of the model building process are provided in section 2.1.1.

To perform *de novo* structure prediction we rely on *trRosetta*[22] a deep learning based protein folding protocol that was recently developed. *trRosetta* is a two step protocol that first predicts protein distance and residue pair orientation distributions, and subsequently uses those distributions to fold the 3D protein structure. In the first stage of *trRosetta* deep convolutional residual neural networks are used to predict inter-residue orientations and residue distances distributions from protein sequence databases. The neural network predicts this information by exploiting hidden residue-coevolution data in this deep sequencing data. These distance and angle distribution predictions can then be fit to smoothed inter-residue restraints and using these restraints with the Rosetta centroid energy function energy coarse grained models can be made using Rosetta's *MinMover.* The coarse grained models can then be further optimized using Rosetta *relax*[26]. More details on running *trRosetta* can be found in section 2.1.2.

2.1.1 Model building with RosettaCM

Accurate homology models are typically built using RosettaCM[24]. The input to RosettaCM is any number (usually < 10) of threaded homologous protein structures and the desired target protein sequence. To discover homologous protein sequences *hhblits*, and *hhsearch* from the *hh-suite*[27] software suite are used. *Hhblits* is first used to generate a multiple sequence alignment (MSA), and then the MSA can be used with *hhsearch to* attempt to align the target sequence to the sequences of all known protein structures. These sequence alignments (if they are found) are then used with their respective structures

to generate threaded templates. These threaded templates are then input directly into the RosettaCM protocol. Depending on the quality and sequence coverage of the threaded templates the number of models to generate from RosettaCM can vary but most commonly 200 models are generated, and the best scoring models are taken for use in downstream protocols. Models can be generated using the command line:

```
rosetta_scripts -in:file:fasta target.fasta -parser:protocol hybridize.xml -relax:jump_move true
-default_max_cycles 200 -beta_cart -beta -relax:dualspace
```

Where the input XML file (hybridize.xml) contains text similar to that shown in figure 2.3.

2.1.2 Model building with trRosetta

trRosetta model building is a two-step process: in the initial step a deep residual convolutional neural network is used to generate inter-residue distance and orientation predictions from a MSA, and in the second step these predictions are used to model a protein of interest[22]. MSAs for *trRosetta* are made using a two-step procedure. In the first stage, four rounds of iterative *HHblits*[27] (version 3.0.3) searches against the Uniclust30 database (August 2018 version) with gradually relaxed *E*-value cutoffs (10−80, 10−60, 10−40 and 10−20) were used to generate an initial alignment. The resulting alignment was then converted to an HMM profile and additional sequences were collected by a single run of *hmmsearch*[28] (version 3.1b2) against an extensive custom sequence database as described in Wu *et al.*[29]; a bit-score threshold of 0.2 × (protein length) was used to select significant hits. The composite MSAs were filtered with *hhfilter* at 99% sequence-identity and 50% coverage cutoffs. For each domain, 150 centroid models are generated and each model is then refined with the *Rosetta* full-atom FastRelax protocol. The results from this refinement are used to sort the models based on the REF2015 score function, and the top scoring model is selected.

2.2 Docking explanation

Protein models are docked into cryoEM density using the dock_pdb_into_density application in association with the wrapper script dgdp.py. Similar to Wang *et al.* 2015 *dock_pdb_into_density* uses a FFT-accelerated six-dimensional search to find rigid-body placements of the protein model into the cryoEM density map that optimize the overlap of map to model. The algorithm performs FFT convolutions

in rotational space and enumerates over all points of strong density (typically 100,000) that are at least 1.5 Å apart. The 5000 top scoring coarse placements are clustered and filtered until only 1000 models remain. These 1000 fully atomic models are minimized further into the cryoEM density using a masked correlation function[12], then clustered using a rms cutoff of 10 Å and finally filtered based on their elec_dens_fast energy to a maximum of 200 solutions. The docking script additionally permits multiple models of the same peptide to be docked simultaneously and a final step will combine and cluster the results from each model into one ensemble of structures.

2.3 Complex Model Assembly

Given the docked domains from the previous section, we used a modified version of the Monte Carlo simulated-annealing (MC-SA) sampling protocol described in Wang et al. (2016) to build a model of the complex. Briefly, the protocol uses the top 200 placements for each model from our docking protocol, in addition to the cross-linking data, in order to determine a set of domain placements that are most consistent with all available data. This MC-SA domain assembly assigns a placement or 'not found' to each domain to account for the possibility that either all of our predicted models are incorrect or that the domains are correct but not present in the map. Consistency is measured through the function (where d_N is all domains):

$$score_{total}(D = \{d_0, \ldots d_N\}) = w_{dens} \sum_{d_i \in F} score_{dens}(d_i)$$

$$+ \; w_{proximity} \sum_{d_i, d_j \in F} score_{proximity}(d_i, d_j)$$

$$+ w_{centroid\_energy} \sum_{d_i, d_j \in F} score_{centroid\_energy}(d_i, d_j)$$

$$+ w_{distance\_constraints} \sum_{d_i, d_j \in F} score_{distance\_constraints}(d_i, d_j))$$

Where $score_{dens}$ measures the fit of the selected domains to the density and the other terms assess interactions between all domains. The term $score_{proximity}$ validates that when two domains are part of the same peptide chain and not overlapping they are placed within a distance that is closable by a later built peptide linker. The $score_{centroid\ energy}$ term is Rosetta's centroid energy score term which is a coarse

grained representation that is used to verify the quality of domain-domain interfaces, as well as screen for clashing placements. The centroid energy between two domains is evaluated by using Rosetta to combine the two domains into a single system (Pose), evaluate the system's energy, then spatially separating the two domains and again evaluating the energy of the system. The former is subtracted by the latter, and this is used as the unweighted *centroid_energy* score. Finally the $score_{distance\ constraint}$ term serves as a way to incorporate experimental data such as cross linking mass-spectrometry data, and assesses the satisfaction of these constraints. The inter-domain geometry terms are assessed as follows:

$$score_{proximity}(d_i, d_j) = (1/(1 + e^{(-5*(gap\_distance_{i,j} - gap\_N\_residues*3.4+1)}))$$

$$score_{distance\_constraints}(d_i, d_j) = \sum_{\{i,j\} \in distance\_constraints} (1/(1 + e^{(-0.6*(||x_i - x_j|| - 32)}))$$

Weights were determined by fitting on a training set with synthetic 10 Å cryoEM data, and the weights used were w_{dens} = 260, $w_{distance\_constraint}$ = 30,000, $w_{centroid\_energy}$=150, $w_{proximity}$= 1,000.

Using the scoring function above, we evaluate the consistency of the results from docking for all domain-domain pairs. Prior to scorefunction evaluation a custom pairwise interface optimization protocol is applied: domains are slid along an axis through each domains' center of mass to be in contact, but not clashing with each other, moving no more than 5 Å. If after this, domains are still clashing (defined as Rosetta vdw score > 1500), we remove all clashing residues (Rosetta vdw score > 3) with: a) no secondary structure, and b) surface exposure (less than 10 Cα's within 12Å), and rescored. This is then followed by breaking both domains into subdomains (using a reimplementation of DDomainParse [39] in Rosetta), and rigid-body minimizing these domains with respect to the energy function above.

Once all pairs of domains have been refined, and their refined inter-domain energies have been computed, Monte Carlo simulated annealing (MC-SA) sampling is carried out. Each MC-SA move reassigns one domain to either another placement, or "no placement." We carry out 200,000 steps of MC-SA sampling, ending at a temperature of kT=1 . 50,000 independent trajectories are carried out and the top ten scoring assignments are assessed for convergence. Convergence is assessed by manually inspecting the ten domain assignments for domains that were present in a majority of the models. This process can be applied iteratively; after each round of assembly, domains that converged in location are

locked into place, where convergence was assessed, the density occupied by converged domains was removed from the density map, and unassigned domains were re-docked and used as inputs for the next round.

2.4 Simulation of cryoEM maps

Modeling cryoEM density maps in the subatomic-resolution range is a very difficult process and without homology models or near native starting models maps in this resolution have a tendency to be published unmodeled. Without an accurate deposited model, these maps cannot serve as benchmark cases for developers that are trying to develop software to solve this difficult problem. Due to the lack of modeled experimental data in this resolution range it has become common practice to benchmark cryoEM modeling softwares on simulated cryoEM data[6,9,17,18] using solved crystal or cryoEM models as a basis for the simulation. Density maps that are simulated represent ideal representations of perfect cryoEM data, while experimental cryoEM data is significantly more heterogeneous. This heterogeneity owes itself to a myriad of factors including intrinsic protein flexibility, sample heterogeneity, image misalignment, radiation damage, low signal-to-noise ratios, beam induced sample movement, and preferred sample orientation[30–35]. Due to these factors it is common practice to deposit a map that conveys the estimated resolution of individual voxels and their surroundings in addition to the final best reconstruction[3,36]. Simulated cryoEM density maps are generated with a simple gaussian point spread function[37] that is applied to every atom. This method of simulation does not try to imitate any aspects of the heterogeneity commonly present in experimental maps and because of this a protocol's success on these simulated maps does not necessarily correlate with success on experimental maps. To address this gap between experimental and simulated cryoEM data we developed an alternative method for simulating cryoEM data that relies on crudely simulating images - rather than volumes - and performing 3D reconstructions on these images.

To simulate cryoEM maps that are more representative of experimental data the sim_cryo app in Rosetta was developed, which attempts to simulate images rather than reconstructions. These simulated images are passed into cryoSPARC[38] which performs an *ab initio* reconstruction and then homogenous refinement to generate the map. The algorithm crudely simulates cryoEM images by uniformly rotating a pose and then collecting the coordinates of all atoms as viewed from each plane (XY, XZ, YZ). All atom's

coordinates are then randomly perturbed in all directions against the plane and then a gaussian filter is applied to each atom and the sum of all gaussians are stored as an image. Gaussian noise is applied to each image individually and then the images are output as an image stack (.mrcs file) with a .star file. The quality of the resulting cryoEM map can vary on a number of parameters including the number of particles generated, the gaussian sigma applied, the strength of the gaussian noise, if other noise is included, and the pixel size.

The resulting cryoEM maps contain noise and density distribution that is more similar to experimental cryoEM data. (Figure 2.4, A-B). Despite the limited complexity of the image simulation algorithm, the application does generate maps that have reduced density cross correlation to the native models (Figure 2.4, C). The algorithm does not attempt to simulate the more complicated sources of cryoEM map heterogeneity and therefore does not fully recapitulate the diversity of resolution typically present in experimental maps (Figure 2.4, D).

The input to the sim_cryo app is a pdb file, and the output of the app is a stack of images in .mrcs format. A typical run command for the sim_cryo_app might be:

```
Rosetta/bin/sim_cryo -NR 400 -gauss_multiplier 0.8 -resolution 4 -pixel_size 1 -box_size_multiplier 2 -s mypdbfile.pdb -ignore_zero_occupancy false -ignore_unrecognized_res -particle_wt_offset 0.2 -atom_gauss_random_multiplier 3
```

2.5 Benchmarking dataset and results

In order to illustrate the flexibility of the OATS protocol, it was benchmarked across a wide variety of resolutions, protein complexes, and on both experimental and data simulated using our *sim_cryo* protocol. The benchmark data set contains twelve complexes with a variety of complex topology configurations (Figure 2.2, Table 2.1). To ensure that the protocol remained robust, a select number of complexes were benchmarked using a variety of starting models. In some cases additional starting models were generated with trRosetta, and in other cases starting models were split up into domains. This was done to better represent the problems faced in modeling complexes with unknown natives, situations where starting models are only accurate at the domain level, or situations where peptides are only partially resolved in the density. Details on the benchmarking data set are available in Table 2.1 and include information on the dataset pdb ids, the number of unique chains in the model, the number of

unique subunits in the model, the total number of subunits in the model, per component starting model rmsds, the density resolution, the source of the electron density, the starting model sources, and a description of the complex. The performance of OATS on the benchmark set was evaluated based the number of components correctly placed, the number of components incorrectly placed, the total number of components available to place, the correct component average, minimum, and maximum Cα rmsd, of the top scoring assembled complex. Placements were said to be correct if their Cα rmsd's were less than 7 Å. The 7 Å resolution cutoff for success is higher than most typical rmsd cutoffs, but our reported Cα rmsd is a measure of accuracy of both the input model compared to the native structurel, and the placement of the input model relative to the native structure. As an example of this measure, a starting model with an aligned Cα rmsd of 5.5 Å could - with absolute perfect density alignment - receive a Cα rmsd of 5.5 Å. Additionally, as the size of the input model increases, small rotational deviations (such as a 3° tilt) in placement will lead to larger and larger Cα rmsd values. After extensive testing and evaluation we believe that the 7 Å cutoff provides enough margin for small errors in model placement and doesn't miss-classify any incorrect placements.

Of the twelve complexes OATS was able to completely assemble five with high accuracy and no errors (1e6v, 5nd7, 4aod, 6ks6, and 3iyf) (Table 2.2). The average, minimum, and maximum rmsd of each subunit of these assembled complexes were all < 4 Å rmsd to the native. The assembled 1e6v complex had an average rmsd of 1.0 Å, with a minimum rmsd of 0.8 Å and a maximum rmsd of 1.2 Å (Figure 2.5A). The average assembled rmsd of 5nd7 was 2.7 Å, with a minimum rmsd of 2.3 Å, and a maximum rmsd of 3.2 Å. The assembled 4aod complex had an average rmsd of 3.8 Å, with a minimum rmsd of 3.8 Å and a maximum rmsd of 3.9 Å. The average assembled Cα of 6ks6 was 6.6 Å, the minimum subunit rmsd was 5.2 Å, and the maximum was 8.5 Å. The assembled 3iyf complex had an average rmsd of 4.0 Å, a minimum rmsd of 3.9 Å, and a maximum rmsd of 4.0 Å.

Two of the twelve complexes were assembled with only one or two incorrect placements during the assembly (6w18, and 5nem *sim_cryo*) (Table 2.2). The five assembled correct components of 6w18 had an average Cα rmsd of 4.1 Å, a minimum rmsd of 1.7 Å, and a maximum rmsd of 5.0 Å. The F chain starting model of 6w18 had an rmsd of 6.1 Å and was incorrectly placed in the model, while the D chain starting model had an rmsd of 4.6 Å and was not placed at all. The four correctly assembled components

of 5nem *sim_cryo* had an average Cα rmsd of 4.0 Å, with a minimum rmsd of 1.7 Å, and a maximum rmsd of 5.0 Å. Of the incorrect placements in 5nem *sim_cryo* chain 2 had a starting model rmsd of 2.7 Å and was incorrectly assembled with an rmsd of 7.2 Å, and chain 4 had a starting model rmsd of 1.0 Å and was incorrectly assembled with a rmsd of 23.3 Å.

Five complexes were assembled with major (>3) errors in subunit placements, these were 1qle *sim_cryo*, 1tyq *sim_cryo,* 6w18 *sim_cryo,* 5nem emdb ID-3632, and 5xf8 emdb ID-6671 (Table 2.2). 1qle was assembled with two correctly placed domains, and three incorrectly placed domains. The per component Cα rmsds of the assembled 1qle complex was A: 7.8 Å, B 5.6 Å, C 2.6 Å, H: 30 Å, L: 29 Å. The top scoring assembled 1tyq *sim_cryo* complex had three domains placed correctly, and 4 domains placed incorrectly (Figure 2.5B). The per component rmsds of this assembled complex was A: 5.3 Å, B: 4.6 Å, C: 7.1 Å, D: 18.4 Å, E: 2.3 Å, F: 12.9 Å, G: 8.2 Å. The 6w18 benchmark case is similar to 1tyq in that they are both models of the same arp2/3 complex, but are from different organisms and while 1tyq was solved with X-ray crystallography, 6w18 was solved with cryoEM. The top scoring assembled 6w18 *sim_cryo* complex has per component rmsds of A: 3.9 Å, B: 3.0 Å, C: 13.8 Å, D: 4.3 Å, E: 4.8 Å, F: 7.3 Å, G: 7.9 Å. The 5nem emdb ID-3632 assembled complex contained three correctly placed subunits, and three incorrectly placed subunits. The per component rmsds of the assembled complex was 1: 61.8 Å, 2: 3.1 Å, 3: 79 Å, 4: 72 Å, A: 2.0 Å, B: 6.1 Å. The 5xf8 emdb ID-6671 complex was assembled with one correctly placed domain, six incorrectly placed domains. The per component rmsds of this complex was 1: 67.5 Å, 2: 9.6 Å, 3: 8.2 Å, 4: 81.3 Å, 5: 43.7 Å, 6: 8.1 Å, C: 5.9 Å. In all complexes in this category OATS was able to correctly assemble at least a small portion of the complex, however all stand as examples to improve upon in further development.

2.6 OATS discussion

The OATS benchmark set provides many targets with varying difficulties that cryoEM modeling developers should strive to solve. When evaluating the results of the twelve benchmark cases, the results can be separated into three categories; cases where the complex was solved completely and correctly, cases where the complex was mostly solved correctly, and cases where the complex had major (>3) errors in placements of subunuits. The first category, where the complex was assembled correctly

and completely, contains benchmark cases 1e6v, 5nd7, 4aod, 6ks6, and 3iyf. Unexpectedly, this category mainly contains cases with strong levels of pseudosymmetry. We define pseudo-symmetric cases as those where two or more subunits are similar in tertiary structure, which can be quantified with their TMScore which can be evaluated using TM-Align[40]. TM-scores are between 0 and 1.0, and are measures of protein fold similarity. A TM-score of 0.5 or greater indicates that the fold of the two proteins is similar. Modeling cases with pseudosymmetry are especially difficult with inaccurate models, as small deviations in structure can drive the structure to appear more similar to the pseudo-copy, or cases where the cryoEM density is of low resolution, which makes the density for the pseudo-copies more similar to each other. In the case of 1e6v, chain B has a TM-score of 0.62 to chain A, and the *sim_cryo* map resolution of 9.0 Å. The 5nd7 benchmark case requires the software to distinguish between *trRosetta* models of alpha (2.3 Å rmsd) and beta (2.3 Å rmsd) tubulin in addition to a MKLP2 motor domain (2.3 Å rmsd) three times as the map contains three copies of the asymmetric unit. The models of alpha and beta tubulin have a TM-score of 0.97 to each other, and are nearly indistinguishable from each other when structurally aligned. The 6ks6 case is especially interesting as the complex, TRiC, contains 16 chains, 8 of which are unique, that are all extremely similar in sequence and structure to each other, however the resolution of the cryoEM data is the highest in the entire benchmarking set (2.99 Å). This pseudosymmetry already makes multi-body docking of this complex difficult, but in this case *trRosetta* produced relatively inaccurate models (Table 2.1) compounding the difficulty of docking and correctly assembling the complex. Both 4aod and 3iyf are present in the PRISM-EM[17] benchmark set, and are cases where a single subunit must be symmetrically docked into low resolution (6 Å and 8 Å respectively) experimental cryoEM density. Symmetric docking/assembly is not set up for the OATS protocol, and therefore each additional copy of a subunit is treated as its own entity with no ties to the other copies. This means that OATS must correctly identify correct interfaces and placements N number of times.

The second category of the benchmarking results contains cases where OATS was able to mostly correctly assemble the cryoEM complex, but still made a few (<2) errors. This category contains 6w18 emdb ID-21503 and 5nem *sim_cryo.* The 6w18 complex was assembled with five correctly placed subunits (A, B, C, E, G), one subunit placed incorrectly (F), and one subunit not placed at all (D). The F

subunit had the worst starting model to native rms of 6.1 Å which caused the docking protocol, which only rigid body perturbs the molecule, to be unable to find a near native fit within the top 100 solutions. The D subunuit however had a relatively near native rms of 4.6 Å, and the nearest native placement in the top 100 results of docking oriented the subunit with an rmsd of 8.8 Å to the native. In this case, the interface optimization protocol was unable to develop a nonclashing interface between the D subunit and the much-nearer placements of the correctly placed subunuits, which caused it to be left out of the complex and assembled as *null*. The 5nem *sim_cryo* complex was assembled with four correctly placed subunits (1, 3, A, B) and two incorrectly placed (2, 4). All starting model subunits of 5nem are close to the native, and the two incorrectly placed subunits are no exception, with the rmsd of subunit 2 having an rms of 2.7 Å, and subunit 4 having an rms of 1.0 Å. The nearest native dock into the *sim_cryo* density of subunit 2 had an rms of 7.2 Å to the native, and the nearest native dock of subunit 4 was 23 Å. The resolution of this map is very low at 8.0 Å, and the native of subunit 4 only contains 45 residues (two disconnected chains of ~20 residues each) with barely any detectable secondary structure. Correctly placing this at this resolution would be extremely difficult, and it is likely that it was modeled in the 10.8 Å resolution EM structure[41] mainly due to the existence of a known crystal structure of chains 1, 2, 3, and 4. Subunit 2 on the other hand had a very accurate starting model, so to solve this case the OATS protocol likely requires improvements to model docking, or for the interface optimization to be more permissive of models with large rmsds to natives.

The third and final category of the benchmarking results contains cases where OATS had major errors in subunit placement. This category contains 1qle *sim_cryo*, 1tyq *sim_cryo,* 6w18 *sim_cryo,* 5nem emdb ID-3632, and 5xf8 emdb ID-6671. The assembled 1qle *sim_cryo* complex contained two correctly docked models (B, C), and three incorrectly docked models (A, H, L). The H and L chains are antibody FV fragments and therefore extremely similar in shape. While both H and L placements were found in the docking results for both subunits, the incorrect placements were selected indicating that the interface optimization and scoring was not accurate enough to discriminate between the near native and wrong interfaces. The A subunit was assembled to be 7.8 Å rms to the native. The A subunit is quite large and contains 432 residues, and therefore an rms with a strict cutoff of 7 Å might not be a suitable measure of correctness in this case. The center of mass of the assembled A subunit was off by approximately 4.2 Å,

and the subunit was correctly oriented, but slightly askew leading to a relatively large final rms to native. Due to these relatively minor differences from the native model it is possible that after refinement in the context of the other subunits this deviation might be ameliorated. The 1tyq *sim_cryo* and 6w18 *sim_cryo* cases are both in this category, and both are models of the arp2/3 complex, but at different resolutions and from different organisms. Both were assembled with incorrect placements of domains C, F, and G. The C domain is a β-propellor which has an intrinsic level of pseudosymmetry. When β-propellor domains are placed into low resolution density, it is nearly always impossible to determine the correct orientation using density alone. Therefore in cases with β-propellers the correct orientation must be correctly selected using the interface score as evaluated by OATS. In this case however, OATS was unable to discern between the native and incorrect placements of the β-propellor and this serves as an area where improvement is clearly needed. The F domain contains an extended helix which may have been a cause of its failure in both cases. The fast FFT based electron density docking that is used has a tendency to perform significantly worse on large extended peptides, however the reason for this is not fully understood (see extended helix out of density in Figure 2.5B). A possible reason for this shortcoming is that the default spherical harmonic transform bandwidth is set to 32 for all subunits. Increasing this bandwidth will force the software to search more rotations and lead to increased runtimes, but in cases that contain extended helices this may be necessary as small angular deviations in docking these cases leads to larger Cα rmsds. The cause of the G domain being incorrectly placed in the assembly appears to be a problem the docking part of the protocol as the closest placement in the top 100 placements into the density is 8.2 Å rmsd for 1tyq, and 7.9 Å rmsd for 6w18. The 5nem emdb ID-3632 electron density contains the most heterogeneity in the entire benchmark set. This heterogeneity stems from the map yielding from a viral capsid, which causes the viral capsid proteins to have density that is continuous with separate asymmetric units, and because the proteins bound to the viral capsid appear to be dynamic. In this assembled complex OATS incorrectly places subunits 1, 3, and 4. Interestingly both of the subunits that are correctly placed are in the region with the most heterogeneous density and are models of proteins that are actually bound to the capsid surface. In this case, the protocol seems to have the most difficulty placing the starting models correctly into the viral capsid density. As previously mentioned subunit 4 is a small peptide and is by far the most difficult to place at

this resolution, but both subunits 1 and 3 have zero near native placements in the top 100 docking results. Therefore this is again a case where improvements to docking might lead to noticeable improvements of the protocol on experimental density. The 5xf8 emdb ID-6671 assembly is a special case that contains 6 peptides that are all pseudo-symmetric, and all have TM-scores of 0.8 or higher to each other. To further complicate this problem each peptide encompasses many domains, so subunuits 2, 3, 4, 5, 6, 7, and C were split into domains before being input into the protocol. The protocol ended up incorrectly placing 8 domains, correctly placing 6 domains, and not placing 6 domains. Although the electron density resolution is relatively good at 7.1 Å, 6 of the 20 domains had 0 placements in the top 100 docking results below 7 Å rmsd to the native structure. Although this may be a cause for some of the incorrect placements, the domain interface scores were also not strong enough to pull out the correct domain configuration. Improving the assembly of this complex will require improvements to all parts of OATS, and serves as an incredibly difficult benchmarking target.

Future improvements to OATS will need to address the faults outlined in the previous paragraphs in order to improve its performance on this benchmarking set. Complexes that contain pseudosymmetric domains or chains provide one of the most difficult scenarios for OATS and software like it. For a molecular modeling software to consistently and correctly identify native interfaces of pseudosymmetric domains, the software will likely require a higher resolution atomic representation, and possibly increased domain-domain orientation sampling. A variety of improvements are possible to directly implement into OATS including flexible fitting during docking, utilizing rosetta's full atom scorefunction while optimizing interfaces, and many more. However, both of these suggested improvements will lead to significant runtime increases and will require trade offs, or the utilization of significantly more computing resources. Although many possibilities exist for improvements to the underlying OATS protocol, as protein structure prediction becomes more driven via machine learning driven protein folding or coordinate prediction incorporation of cryoEM data into these pipelines is no doubtedly in the near future. Until then, OATS serves as an intermediary that can exploit the local accuracy of models predicted with machine learning without density, and use medium resolution cryoEM density to assemble these models with better accuracy.

Figure 2.1

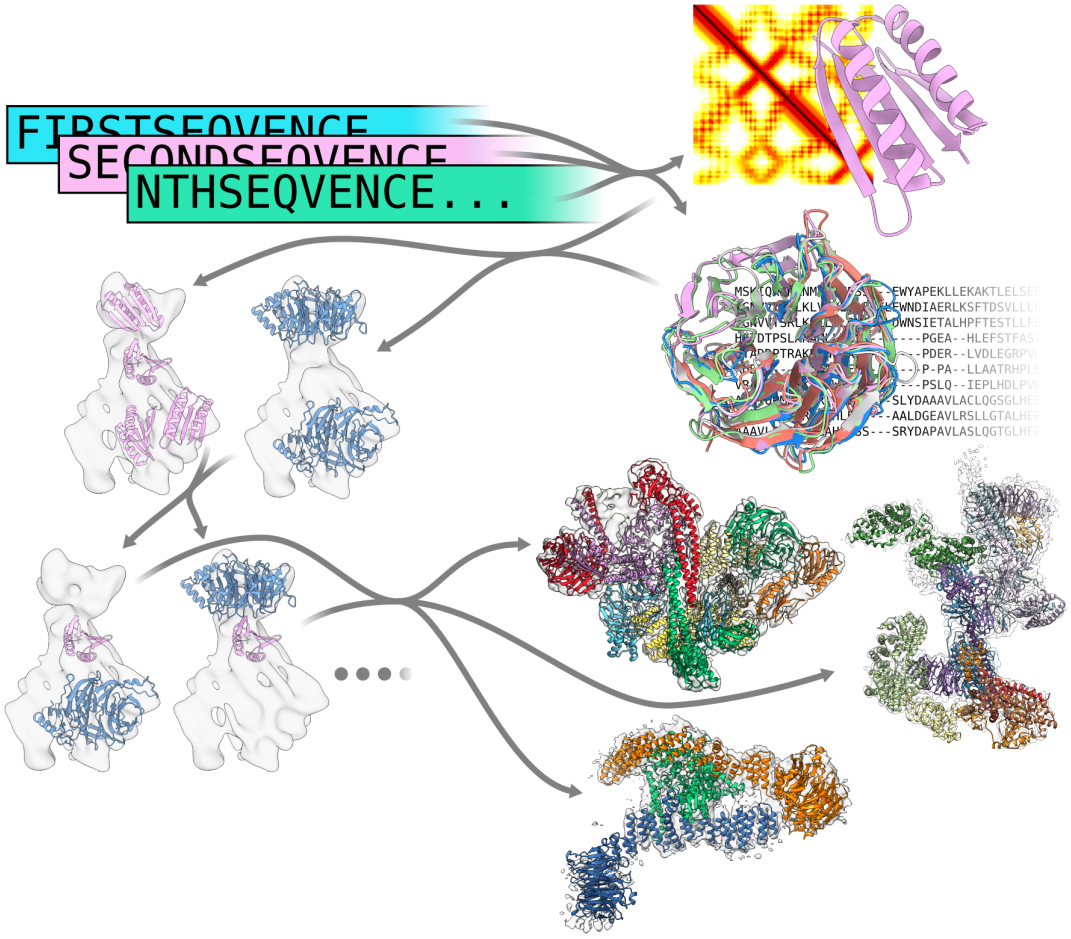

Figure 2.1: An overview of the workflow for modeling with OATS. Initially, multiple sequence alignments (MSAs) of all protein sequences are generated, the sequences are segmented into domains using the MSAs and the individual domains are folded using *trRosetta,* or built using RosettaCM homology modeling. These domains are individually docked into the cryoEM density. All pairwise docked model interfaces are optimized using a novel interface minimization scheme, and then Monte Carlo sampling finds the domain assignment that is maximally consistent with the experimental data (electron density, XL-MS *etc.*) . Finally, linkers between domains are sampled and the entire structure is refined to produce final protein complexes.

Figure 2.2

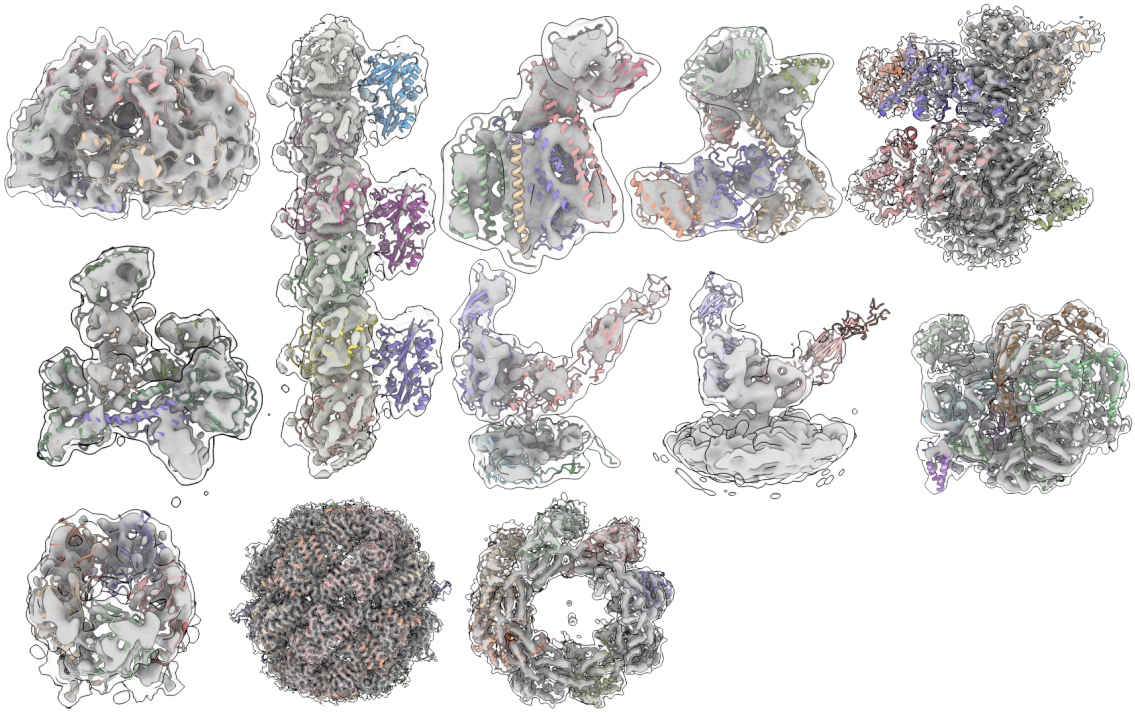

Figure 2.2: All benchmark cases that OATS was trained on. From left to right, all benchmark cases are shown at two density thresholds and with all native chain placements, the cases are: 1e6v *sim_cryo,* 5nd7 emdb ID-3623, 1qle *sim_cryo,* 1tyq *sim_cryo,* 6w18 emdb ID-21503, 6w18 *sim_cryo*, 5nem *sim_cryo*, 5nem emdb ID-3632, 5xf8 emdb ID-6671, 4aod emdb ID-2055, 6ks6 emdb ID-0758, 3iyf emdb ID-5140.

Figure 2.3

```
<ROSETTASCRIPTS>
        <TASKOPERATIONS></TASKOPERATIONS>
        <SCOREFXNS>
                <ScoreFunction name="stage1" weights="score3" symmetric="0">
                        <Reweight scoretype="atom_pair_constraint" weight="0.1" />
                </ScoreFunction>
                <ScoreFunction name="stage2" weights="score4_smooth_cart" symmetric="0">
                        <Reweight scoretype="atom_pair_constraint" weight="0.1" />
                </ScoreFunction>
                <ScoreFunction name="fullatom" weights="beta_cart" symmetric="0">
                        <Reweight scoretype="cart_bonded" weight="2.5" />
                        <Reweight scoretype="atom_pair_constraint" weight="0.1" />
                </ScoreFunction>
        </SCOREFXNS>
        <FILTERS></FILTERS>
        <MOVERS>
                <VirtualRoot name="vrm" />
                <PoseFromSequenceMover name="pfsm" use_all_in_fasta="true" fasta="hybridize.fasta" />
                <Hybridize name="hybridize" stage1_scorefxn="stage1" stage2_scorefxn="stage2" fa_scorefxn="fullatom"
                                stage1_increase_cycles="1.0" stage2_increase_cycles="1.0" hetatm_cst_weight="0" batch="1">
                        <Fragments three_mers="t001_.200.3mers.gz" nine_mers="t001_.200.9mers.gz" />
                        <Template pdb="2iyf_B.pdb" cst_file="AUTO" weight="1.0" />
                        <Template pdb="4lgr_A.pdb" cst_file="AUTO" weight="1.0" />
                        <Template pdb="2xxr_J.pdb" cst_file="AUTO" weight="1.0" />
                </Hybridize>
                <AlignPDBInfoToSequences name="apts" mode="multiple" json_fns="apts.json" />
        </MOVERS>
        <APPLY_TO_POSE></APPLY_TO_POSE>
        <PROTOCOLS>
                <Add mover="pfsm" />
                <Add mover="vrm" />
                <Add mover="hybridize" />
                <Add mover="apts" />
        </PROTOCOLS>
        <OUTPUT scorefxn="fullatom" />
</ROSETTASCRIPTS>
```

Figure 2.3: A typical XML file used by rosettaCM. Depending on the number of available homologous structures, the number of entries (lines starting with “<Template”) can vary to any number of templates, although it is rare to use more than 10 templates at a time.

Figure 2.4

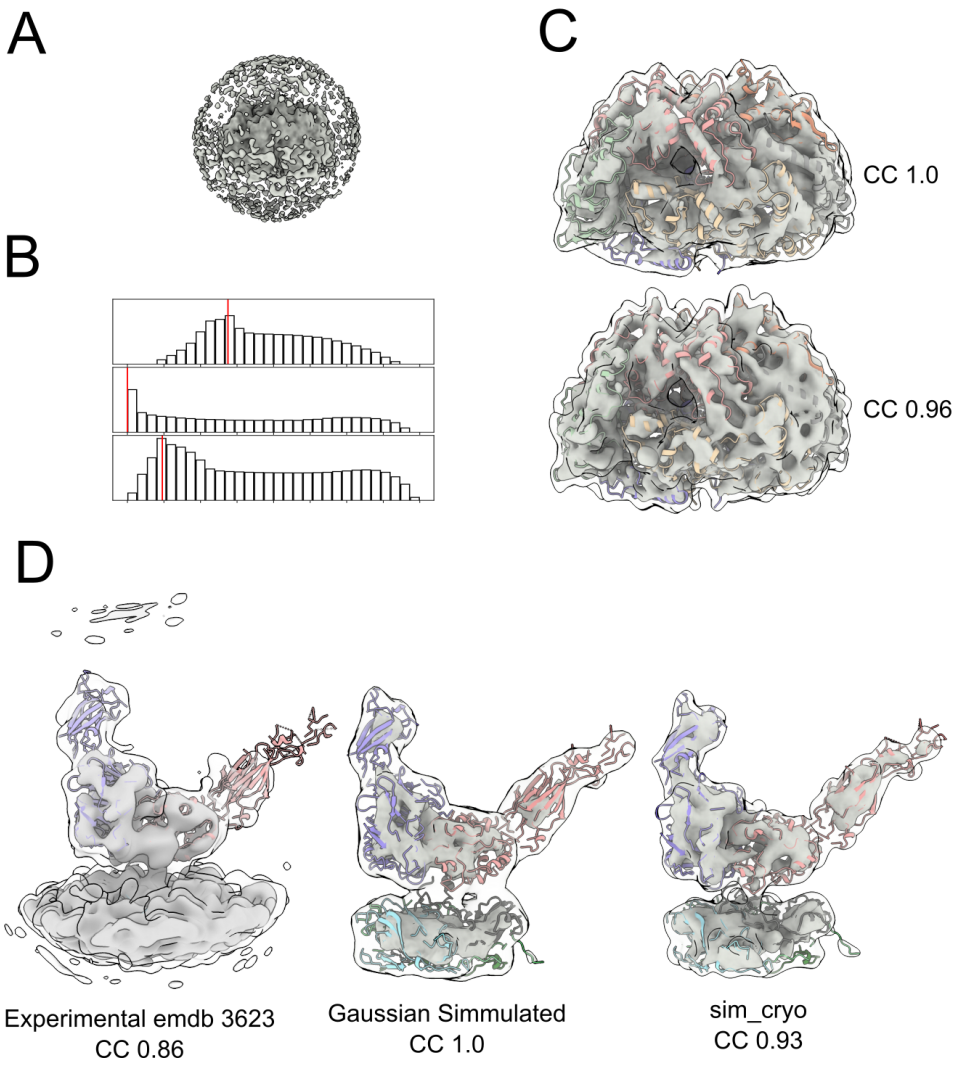

Figure 2.4: Comparison between gaussian simulated, experimental, and *sim_cryo* density maps. A) Maps generated from *sim_cryo* display low resolution noise that is typically seen in experimental density maps. B) A histogram plot shows that maps generated from *sim_cryo* have density distributions that are more similar to experimental maps; the red line on each plot displays the location of 0 on the X axis. Map sources from top to bottom are Experimental map emdb-3623, gaussian simulated, followed by *sim_cryo*. C) A visual comparison between 1E6V maps that have been gaussian simulated (top), or yield from *sim_cryo* images (bottom). While similar to the gaussian simulated map, the map generated with *sim_cryo* has distorted helical density which increases the difficulty of modeling. This can be quantified by comparing the gaussian map to model's cross correlation (1.0) to the map to model cross correlation of the *sim_cryo* map (0.96). D) Dual threshold views of experimental, gaussian simulated, and *sim_cryo* maps for pdb 5nem. Both maps from *sim_cryo* and gaussian simulation are incapable of simulating the flexibility induced resolution heterogeneity displayed in the deposited map. The model to map correlations between each of the three maps is 0.86 for the emdb ID 3623, 1.0 for the gaussian simulated map, and 0.93 for the *sim_cryo* map.

Figure 2.5

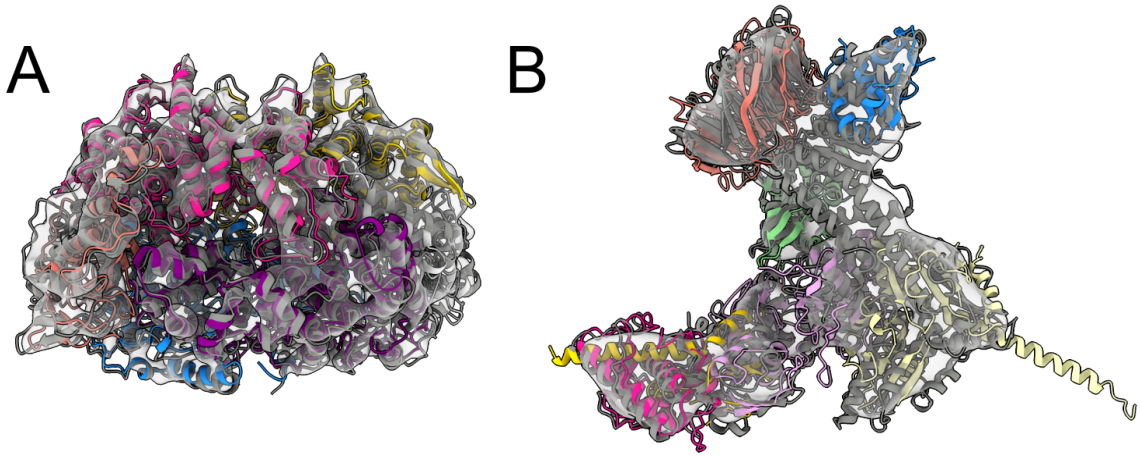

Figure 2.5: A comparison of native models, and models generated from OATS. A) Comparison of the 1.0 Å rmsd model generated from OATS (chains various colors) and the native 1e6v model (dark grey) in the context of the electron density. B) Comparison of the partially correct model for 1tyq *sim_cryo* (chains various colors) generated by OATS and the native 1tyq model (dark grey).

Table 2.1

Pdb file ID	# unique chains	# components (# unique components)	Component SM RMSDs (Å)	density map source	resolution (Å)*	Starting Model Source	complex description
1e6v	3	6, (3)	[A: 1.2, B: 0.8, F: 1.1]	*sim_cryo*	7.4	homology modeling	Methyl-coenzyme M reductase (1E6V)
5nd7	3	9, (3)	[A: 2.3, B: 3.2, C: 2.3]	3623	7.9	trRosetta	Microtubule-bound MKLP2
1qle	5	5, (5)	[A: 5.1, B: 4.1, C: 1.0, H: 0.7, L: 1.2]	*sim_cryo*	8.5	homology modeling	Cytochrome C oxidase
1tyq	7	7, (7)	[A: 5.0, B: 1.4, C: 6.3, D: 1.3, E: 0.3, F: 6.5, G: 1.8]	*sim_cryo*	11.5	homology modeling	Arp2/3 complex
6w18	7	7, (7)	[A: 3.7, B: 2.9, C: 4.2, D: 4.6, E: 4.4, F: 6.1, G: 4.4]	21503	4.2	trRosetta	"
"	"	"	"	*sim_cryo*	9.5	"	"
5nem	6	6, (6)	[1: 1.7, 2: 2.7, 3: 2.0, 4: 1.0, A: 1.5, B: 2.5]	*sim_cryo*	8	homology modeling	Alpha V Beta 6 bound to hand foot and mouth disease virus
"	"	"	"	3632	10.8	"	"
5xf8	7	7, (7)	[2: 11.6, 3: 9.1, 4: 8.3, 5: 7.3, 6: 6.1, 7: 7.8, C: 5.6]	6671	7.1	homology modeling	minichromosome maintenance complex 2-7 & Cdt1
4aod	1	5, (1)	[A: 1.9]	2055	6	PRISM-EM paper	Biomphalaria glabrata Acetylcholine-binding protein type 1
6ks6	8	16, (8)	[A: 5.3, B: 6.3, D: 6.5, E: 7.1, G: 5.4, H: 7.8, Q: 8.4, Z: 5.6]	0758	2.99	trRosetta	TRiC at 0.2 mM ADP-AlFx Conformation 1
3iyf	1	8, (1)	[A: 2.6]	5140	8	PRISM-EM paper	Lidless Mm-cpn in the Open State

Table 2.1: A description of all cases that OATS was benchmarked on. The pdb IDs, electron density map sources, electron density map resolution, and starting model sources are provided. The second column describes the number of unique peptide chains present

in the electron density map, and the third column describes how many total subunits there are, and how many of those subunits are unique. The fourth column provides information on the quality of the subunit starting models and displays the aligned Cα RMSDs for each of them. The final column provides a brief description of the complexes.

Table 2.2

Complex Description	map resolution (source)	# components (correct, incorrect, expected)	correct component Cα RMSD (Å) [average (range)]
1e6v, *sim_cryo* map	9 Å	6, 0, 6	[1.0, (0.8-1.2)]
5nd7, emdb ID-3623	7.9 Å	9, 0, 9	[2.7, (2.3-3.2)]
1qle, *sim_cryo* map	14 Å	2, 3, 5	[5.4, (5.6-2.6)]
1tyq, *sim_cryo* map	13 Å	3, 4, 7	[4.1, (2.3-5.3)]
6w18, emdb ID-21503	4.2 Å	5, 1, 7	[4.1, (2.9-4.6)]
6w18, *sim_cryo* map	9.5 Å	4, 3, 7	[4.1, (2.9-4.6)]
5nem, *sim_cryo* map	8 Å	4, 2, 6	[4.0, (1.7-5.0)]
5nem, emdb ID-3632	10.8 Å	3, 3, 6	[3.7, (2.0-6.1)]
5xf8, emdb ID-6671	7.1 Å	6, 8, 20	[4.1, (3.0-5.1)]
4aod, emdb ID-2055	6 Å	5, 0, 5	[3.8, (3.8-3.9)]
6ks6, emdb ID-0758	2.99 Å	16, 0, 16	[6.6, (5.2-8.5)]
3iyf, emdb ID 5140	8 Å	8, 0, 8	[4.0, (3.9-4.0)]

Table 2.2: A table representation of the OATS performance on the benchmark set. The first column provides information on pdb ID that has been deposited and was used as the native structure when modeling, in addition to the source of the electron density map. The second column provides information on the resolution of the electron density map. The third and fourth columns display information on the performance of OATS on each benchmarking case. The third column shows the number of components correct, incorrect, and expected in the complex, while the fourth column expresses the average, minimum, and maximum Cα RMSDs of the assembled complex's correct components.

Chapter 3. Modeling of Novel Structures

Over the course of its development OATS was successfully applied to three blind modeling tasks. We first used this new software in order to solve the Bardet-Biedl syndrome complex (BBSOME)[42]. A short time later we then again applied the software in order to solve the Fanconi anemia core complex, and finally shortly after we solved the structure of the HOmotypic fusion and Protein Sorting (HOPS/VPS) complex. Between solving the BBSome and the FACC, progress in protein structure prediction was accelerated after the publication of AlphaFold[23]. The use of machine learning to predict protein residue pair distance, and residue pair orientation distributions using deep residual-convolutional neural networks generated a source of models that we found to generally be locally accurate, but inaccurate in terms of long range residue pair distances. These models when combined with subnanometer cryoEM density and our software allows for the density to filter correct submodels out of these *de novo* predictions and obtain highly accurate atomic models without the need for homologous information.

3.1 Modeling of the Bardet-Biedl Syndrome Complex

3.1.2 Bardet-Biedl Syndrome Complex Introduction

The Bardet-Biedl syndrome (BBS) complex (BBSome) is a ciliopathy characterized by obesity, retinal degeneration, polydactyly, and kidney cysts. The BBSome, an evolutionarily conserved complex of eight BBS proteins ferries GPR161, SMO, and other ciliary membrane proteins across the TZ and out of cilia[43,44]. The BBSome acts in concert with another conserved BBS gene product, the ARF-like GTPase, ARL6/BBS3, which recruits the BBSome to membranes[45] and enables TZ crossing of GPR161[46]. The BBSome also associates with intraflagellar transport (IFT) trains comprised of microtubule motors, and IFT-A and IFT-B complexes. IFT trains undergo processive intraciliary transport and the BBSome functions as an adaptor complex between IFT complexes and cargoes[43,44]. First, the BBSome directly recognizes cytoplasmic determinants on signaling receptors such as SMO and GPR161[46,47] and is required for the intraciliary movements of the cargo phospholipase D in Chlamydomonas[48]. Second,

BBSome subunits consist mostly of domains characteristic of coat adaptors (a solenoids, b propellers, and appendages, Figure 1A). Finally, the polymerization of a membrane-apposed BBSome/ARL6 coat[45] is reminiscent of the clathrin coat adaptor AP-2, which polymerizes onto membranes[49,50].

A major open question is how the BBSome in complex with GTP-bound ARL6 (ARL6GTP) enables cargo exit from cilia[51]. A better understanding of the molecular architecture of the BBSome may provide some answers to this question. Although some progress has been made in determining the high-resolution structure of individual BBSome subunits[52,53] and in analyzing the subunit organization of BBSome sub-complexes[47], structural information on the intact complex is currently lacking.

In this chapter we utilized OATS in combination with cryo-electron microscopy and single-particle reconstruction to arrive at a near-complete Ca model of the BBSome. Unexpectedly, the map of the BBSome shows that its predominant conformation does not permit binding of ARL6GTP. We conclude that the BBSome, like the coat adaptor complexes AP-1, AP-2, and COPI, exists mostly in an autoinhibited, closed conformation in solution and becomes activated as it is recruited to membranes.

3.1.2 Bardet-Biedl Syndrome Complex Peptide and Map analysis

The recurrence of a solenoids, b propellers, and appendage domains within BBSome subunits posed considerable challenges for subunit assignment (Figure 3.1). BBS2, 7, and 9 share an identical domain organization consisting of a b propeller (bprop) followed by a cc, an appendage domain subdivided into a g-adaptin ear (GAE) domain and a platform (pf) domain, and a C-terminal helical bundle (CtH); BBS1 is very similar but truncated after the GAE domain. The crystal structures of BBS1bprop and BBS9bprop have revealed that these two b propellers are more closely related to one another than to any other b propeller in the PDB[52,53]. Meanwhile, both BBS4 and BBS8 consist of 12 tetratricopeptide (TPR) repeats. Individual TPR repeats are known to fold into a stacked pair of helices, and stacking of TPR repeats results in the formation of a right-handed superhelix. BBS5 is composed of two pleckstrin homology (PH) domains followed by a small C-terminal three-helix bundle[54], while the micropeptide BBS18 is predicted to be a mixture of α helices and unstructured regions[55].

Altogether, 29 domains distributed in 8 proteins spanning 4,375 amino acids (aa) had to be located (Figure 3.1.1). Since the resolution was not sufficient to identify residues, a first approach used

differences in 3D classes, known binary interactions, and structural predictions to manually assign each domain to a specific region in the map. The entire process of subunit assignment is detailed below. In addition, we subjected the BBSome to chemical crosslinking MS (XLMS) and identified 42 intersubunit and 34 intra-subunit crosslinks with high confidence (Figure 3.1.2A).

3.1.4 Bardet-Biedl Syndrome Complex Model Building and Refinement

In a first stage, we attempted to model the structure of the 29 domains that make up the BBSome: 4 for BBS1 (BBS1bprop, BBS1ins, BBS1link and BBS1GAE), 6 for BBS2 (BBS2bprop, BBS2cc, BBS2GAE, BBS2pf, BBS2hp and BBS2CtH), 1 for BBS4, 3 for BBS5 (BBS5PH1, BBS5PH2 and BBS5CtH), 6 for BBS7 (BBS7bprop, BBS7cc, BBS7GAE, BBS7pf, BBS7hp and BBS7CtH), 2 for BBS8 (BBS8TPR1-2 and BBS8TPR3-12), 6 for BBS9 (BBS9bprop, BBS9cc, BBS9GAE, BBS9pf, BBS9hp and BBS9CtH), and 1 for BBS18.

The models of 10 domains could be built using high-resolution structures of homologous proteins as guides (and using alignments from hhpred (So¨ ding, 2005)). These 10 domains are BBS1bprop, BBS1GAE, BBS2bprop, BBS4, BBS5PH1, BBS5PH2, BBS7bprop, BBS8TPR3-12, BBS9bprop and BBS9GAE[24]. For each of these domains, 100 to 2000 models were built using RosettaCM[24] (the number of models generated was based on convergence of low-energy structures). While models for all other domains were built without the use of the density map, models for BBS4 and BBS8TPR3-12 were built guided by the density of the two possible regions representing them in the map (half the models were generated from each of the two regions) to ensure that the curvature of the two subunits was consistent with the density map. For each of the 10 domains, the top 20 models obtained with Rosetta all-atom energy[56] were saved for the next step. Of the remaining 19 domains, modeling with Rosetta ab initio[57] produced converged models for 15 domains. These domains are BBS1ins, BBS2cc, BBS2GAE, BBS2pf, BBS2hp, BBS2CtH, BBS7cc, BBS7GAE, BBS7pf, BBS7hp, BBS7CtH, BBS8TPR1-2, BBS9pf, BBS9hp, BBS9CtH (Figure 3.1.1A). 100,000 models were generated for these domains, which all have no more than 150 residues and consist mostly of helical bundles or coiled coils (except for the three platform domains). Overall convergence for BBS7hp, BBS7CtH and BBS9CtH was poor, but the models contained a consistent core that could be placed into the map, allowing the missing regions to be re-modeled based

on their density in the map. The models for two other domains, BBS2GAE and BBS7GAE, only converged when metagenomic data were used to predict contacts from co-evolution[58]. The convergence of these 15 domains suggests that they represent near-native configurations of the domains[58]. As we did for the domains modeled with RosettaCM, the top 20 models of these 15 domains were saved for the next step. Of the remaining four domains, three domains, BBS1link, BBS5CtH and BBS9cc, could not be modeled with any confidence, and BBS18 was modeled at the very end (see below).

In a second stage, we assembled the complete complex using the domain models built as described above and guided by the cryo-EM density map and the crosslinking data. We used a modified version of a protocol used for de novo model building guided by cryo-EM data[4] . This modified protocol was previously employed to determine the domain architecture of the Pex1/Pex6 ATPase complex[59]. Briefly, the protocol first matches each domain individually to the density. Then, Monte Carlo assembly of the entire complex is carried out with a simple score function favoring: a) placement of as many domains as possible, b) maximizing agreement of the placed models with density, c) placement of domains within the same chain such that their termini are close enough to connect, and d) placements that maximize agreement with the crosslinking data. Since the protocol does not allow backbone movement, the crosslink distance was extended by 10 A˚ from ideal to account for possible backbone flexibility (note that changing this distance in between 5 A˚ and 20 A˚ did not change the final domain assignment). Following 10,000 independent trajectories, our Monte Carlo assembly yielded two low-energy solutions: one that was consistent with the manual domain assignment, and one in which the positions of BBS2bprop and BBS7bprop and their respective coiled-coil linker domains were reversed. A final flexible-backbone refinement of both configurations clearly identified the manual domain assignment as the correct one once the maximum crosslink distance between Cb atoms (27 A˚ for BS3 and 30 A˚ for DSSeb) was taken into account, with 44 crosslinks satisfied in the manual assignment compared to 41 for the flipped configuration. Flexible-backbone refinement was performed using RosettaCM guided by the cryo-EM density map and experimental crosslinks.

In the final step, BBS18, was visible as a mostly extended peptide snaking through the core of the complex. Since the configuration of this subunit appeared to be defined by contacts with the other subunits, we modeled this peptide in the context of the complete assembly. A fourteen-residue

poly-alanine helix was initially docked into the clearest helical density region. Every possible fourteen residue stretch of amino acids were threaded, and each was extended with RosettaES[5]. All resulting models were refined, and the lowest-energy configuration was accepted. For this lowest-energy configuration, the RosettaES ensemble was quite well converged. In addition to the 44 crosslinks that were fully compatible with our final model, 15 corresponded to amino acids that were not present in the final model and 17 were in conflict. Given the conformational flexibility of the BBSome, it is likely that crosslinks that are in conflict with the model built into Map 1 are satisfied in other conformations.

3.1.5 Bardet-Biedl Syndrome Complex Model analysis

To independently validate our subunit assignment and gain an understanding of the BBSome at amino acid resolution, we derived a backbone model of the entire complex using OATS and constraints from XLMS (Figure 3.1.3B). Building a Ca model of the assembly was challenging owing to the similarity of many domains in the complex, the lack of homologous structural information for many of the domains present, and the limited resolution of the density map. When this complex was solved, structure prediction using distance distributions from deep learning did not exist, so in absence of homologous information we could only predict short (~<150 amino acids) sequences at a time using Rosetta *Abinitio.* Modeling was performed in three steps. In a first step, models were built for individual domains (Figure 3.1.3A). Of the 29 domains found in the BBSome subunits, 25 could be modeled, 10 with RosettaCM (guided by available high-resolution structures of homologs), and 15 with Rosetta ab initio (the models of BBS2GAE and BBS7GAE also made use of sequence co-evolution data) (Figure 3.1.4). Only three domains could not be modeled with confidence (BBS1link, BBS5CtH, and BBS9cc). In a second step, the 25 individual domain models were assembled using Monte Carlo domain assembly based on our previous de novo model-building method[4], Wang et al., 2015. This process was guided by both density and crosslinking data and resulted in a structure in which 23 domains (except BBS2GAE and BBS7GAE, see below) were uniquely placed. In a last step, RosettaES was used to model the micropeptide BBS18 into the EM density in the context of the 23 already placed domain models. The hybrid modeling approach fitted 3,522 residues of the Ca backbone through 24 different domains (Figure 3.1.3B). Of the 76 crosslink pairs mapped with high confidence, 61 pairs had both positions present in the model, and 44 of these 61

crosslinks were satisfied by the model (Figure 3.1.3C). Although the modeling approach converged to a single model, it is conceivable that very similar domains (e.g., BBS2bprop and BBS7bprop) may be swapped in the actual structure. The future determination of a BBSome structure at near-atomic resolution will help solidify the domain assignment.

The models of BBS2GAE and BBS7GAE were the only ones that could not be satisfactorily placed into the cryo-EM density. Although the application of evolutionary contacts led to well converged models of the individual GAE domains, these could only be approximately placed into the density, and only their relative position and approximate orientation could be determined (Figure 3.1.3B). Considering the lack of a clearly defined domain boundary in the density representing the GAE domains of BBS2 and BBS7, and the failure to incorporate BBS2 and BBS7 into a recombinant BBSome in insect cells[47], we surmise that strand exchange between BBS2GAE and BBS7GAE may preclude folding of the combined BBS2GAE/ BBS7GAE domain in silico and in heterologous systems. In support of this interpretation, assembly of the BBSome requires a specialized machinery that comprises the chaperonin-like proteins BBS6, BBS10, and BBS12. The canonical group II chaperonin TCP-1 ring complex (TRiC) temporarily holds hydrophobic b strands to fold complex structures such as actin, tubulin and some b propellers[60]. BBS6, 10, and 12 substitute for some of the TRiC subunits, forming an alternate TRiC/BBS complex that binds BBS2 and BBS7[61] and incorporates BBS2/7 into the BBSome[62]. The intimate association of BBS2GAE and BBS7GAE may explain the requirement for a specialized folding machinery, because hydrophobic b edges in one GAE domain may need to be chaperoned until the partner polypeptide becomes available for co-folding. The binding of BBS2cc to BBS6 and the position of BBS2cc immediately before BBS2GAE in the polypeptide chain suggest that the alternate TRiC/BBS complex may become recruited to the nascent BBS2 polypeptide immediately before the GAE emerges from the ribosome.

3.1.6 Cα model mutational analysis

At present, 89 missense pathogenic variants in BBSome subunits have been found to cause monogenic disorders ranging from non-syndromic retinal degeneration or obesity to fullfledged BBS (Figure 3.1.4A). The most common and best-studied missense variant is BBS1[c.1169T > G,

p.Met390Arg], which affects a buried residue inside BBS1bprop and interferes with folding of the b propeller[53]. Consequently, BBS1 levels are drastically reduced in the BBS1[p.Met390Arg] mouse mode[63]. Mapping each of the 78 BBS pathogenic variants onto the Ca model reveals that most of the affected positions are likely to be buried (Figures 3.1.4B and 3.1.4C) and predicts that these variants affect folding of a specific domain, thereby resulting in BBSome assembly defects. We surmise that most BBS alleles converge on defective BBSome assembly.

The distribution of pathogenic variants is non-even among BBSome domains, with BBS1bprop and BBS2bprop harboring nearly half of all missense variants (Figure 3.1.4A), suggesting considerable functional importance for these two domains. While BBS1bprop is known to bind ARL6GTP, no specific interactions have yet been described for BBS2bprop. Finally, despite being nearly entirely contained within the core of the BBSome, BBS8 is affected by only two pathogenic variations, both facing toward BBS1GAE. Similarly, missense variants on BBS4 are concentrated at the sites of interactions with BBS18 (BBS4TPR810) or BBS9 (BBS4TPR23).

Interestingly, visual inspection suggests that, while BBS-associated pathogenic variants appear to affect residues predicted to be buried, pathogenic variants associated with non-syndromic retinal degeneration or obesity tend to be on residues that are predicted to be solvent exposed (Figure 3.1.4B and 3.1.4C). This finding suggests that the latter types of variants may affect specific interactions with cargoes or other partners, whereas BBS variants tend to disrupt the entire structure of the complex. A correlation between relatively mild variants and retinal degeneration is in agreement with the tremendous transport rate in photoreceptors that may heighten the sensitivity of this cell type to slight alterations in BBSome activity.

Figure 3.1.1

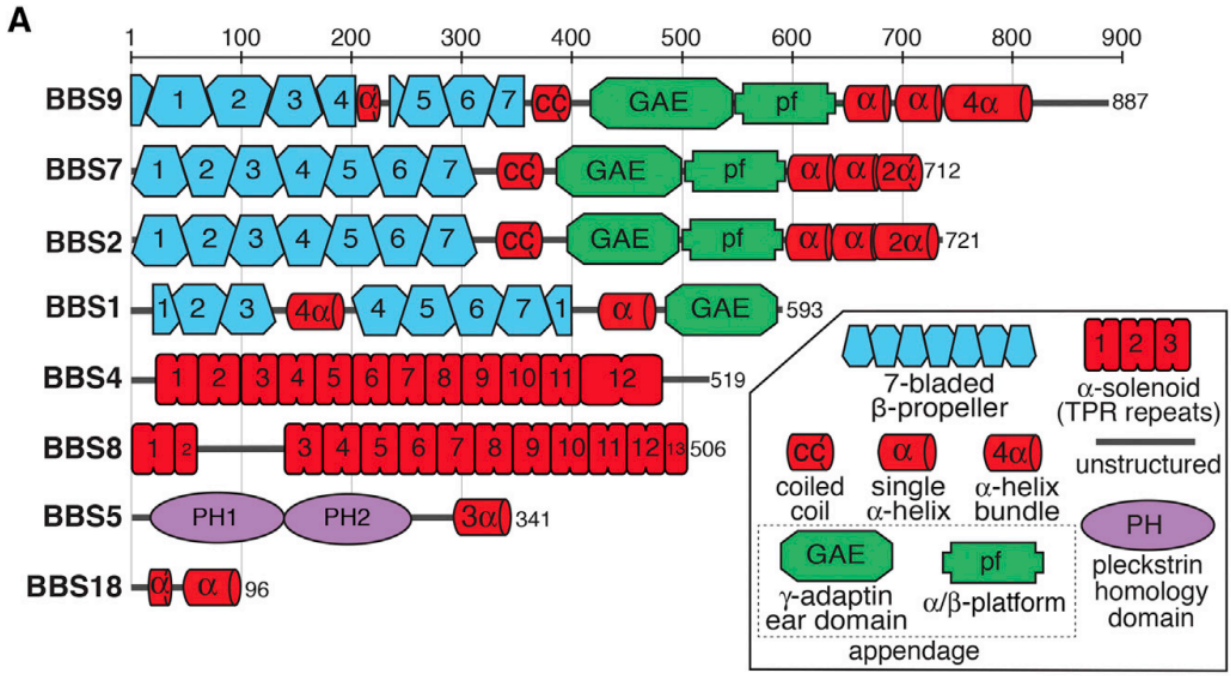

Figure 3.1.1: BBSome Subunits and Cryo-EM Density Map of the BBSome (A) Domain organization of the eight BBSome subunits. The 29 domains making up the BBSome subunits are 4 b propellers (BBS1/2/7/9), one 4- helix bundle inserted into a b propeller (BBS1), 4 connector helices between the b propellers and the GAE domains (some of which are predicted to form coiled coils) (BBS1/2/7/9), 4 GAE domains (BBS1/2/ 7/9), 3 platform domains (BBS2/7/9), 3 hairpins (BBS2/7/9), 3 helical bundles (BBS2/7/9), 3 a solenoids (BBS4/8TPR12 /8TPR313), 2 PH domains (BBS5), 1 3-helix bundle (BBS5), and 1 helical micropeptide (BBS18).

Figure 3.1.2

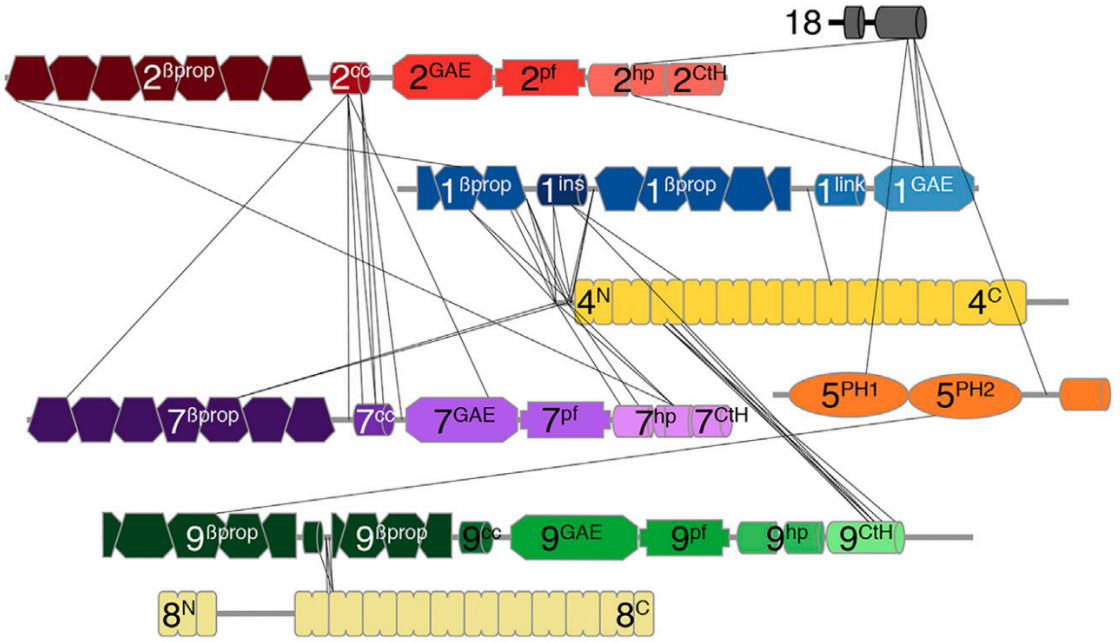

Figure 3.1.2: BBSome identified interactions and domain architecture (A) Inter-subunit crosslinks identified by mass spectrometry mapped onto the subunits. Each subunit is drawn to the scale of its length. The numbers identify the subunit, and the superscripts denote the specific domain: bprop, b propeller; cc, coiled coil; GAE, g-adaptin ear; pf, platform; hp, hairpin; CtH, C-terminal helix bundle; ins, insert; link, linker; PH, pleckstrin homology.

Figure 3.1.3

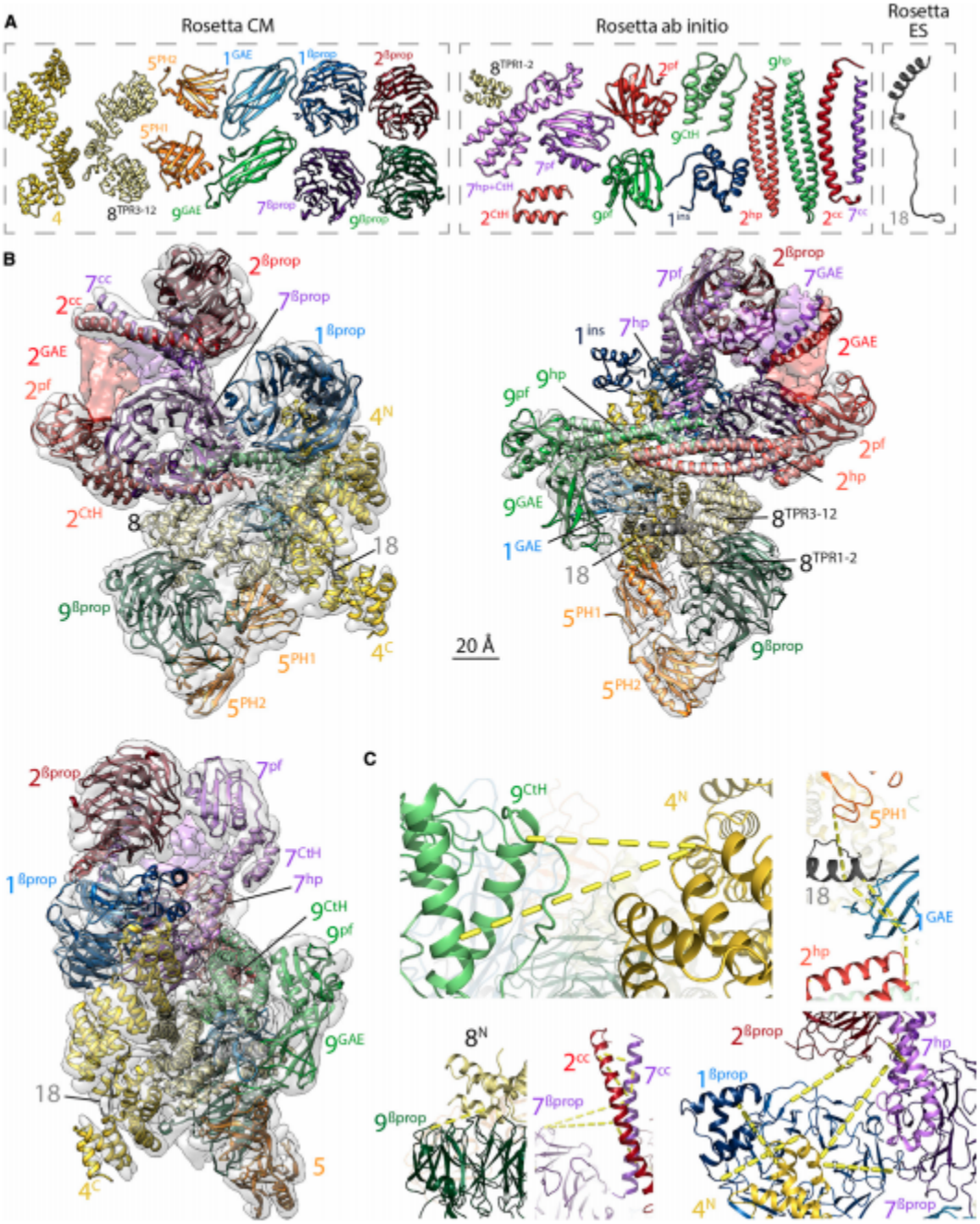

Figure 3.1.3: Rosetta-Generated Ca Model of the BBSome (A) Ca models of the 24 domains from BBSome subunits that were obtained with three different Rosetta modeling protocols (CM, ab initio, and ES; see the STAR Methods for details) and could be assembled into the Ca model of the BBSome. Although the GAE domains of BBS2 and BBS7 could be modeled using co-evolutionary data (Figure 3.1.4), they are not shown because they could not be satisfactorily built into the cryo-EM density map. The colors and labels are as in Figure 3.1.2. (B) Nearly complete Ca model of the BBSome obtained using Rosetta to assemble the 24 domains into a complex, guided by the cryo-EM density map and XLMS data. The GAE domains of BBS2 and BBS7 are not included in the final Ca model, but their general placement is indicated by coloring the density map. (C) Magnified views of crosslink clusters in the final BBSome model. The yellow dotted lines indicate crosslinks that were satisfied by the final Rosetta

molecular model. For clarity, only selected crosslinks of each cluster are shown. Depicted crosslinks are: top left, 9CtH[K789]-4N[K116] and 9CtH[K810]-4N[K116]; top right, 5PH1[K87]-18[K90], 18[K90]-1GAE[K553], 18[K93]-1GAE[K553], and 1GAE[K553]-2hp[K638]; bottom left, 9bprop[K218]-8N[K181]; bottom middle, 7bprop[K56]- 7cc[K352], 7bprop[K56]-2cc[K345], 2cc[K360]-7cc[K359], 2cc[K360]-7cc[K352], 7cc[K359]-7cc[K352], 2cc[K345]-7cc[K352], and 2cc[K345]-7cc[K338]; bottom right, 2 bprop[K9]-1bprop[K69], 2 bprop[K13]-7hp[K658], 4N[K20]-7hp[K659], 4N[K20]-7bprop[K222], 4N[K5]-1bprop[K143], and 1bprop[K192]-4N[K25].

Figure 3.1.4

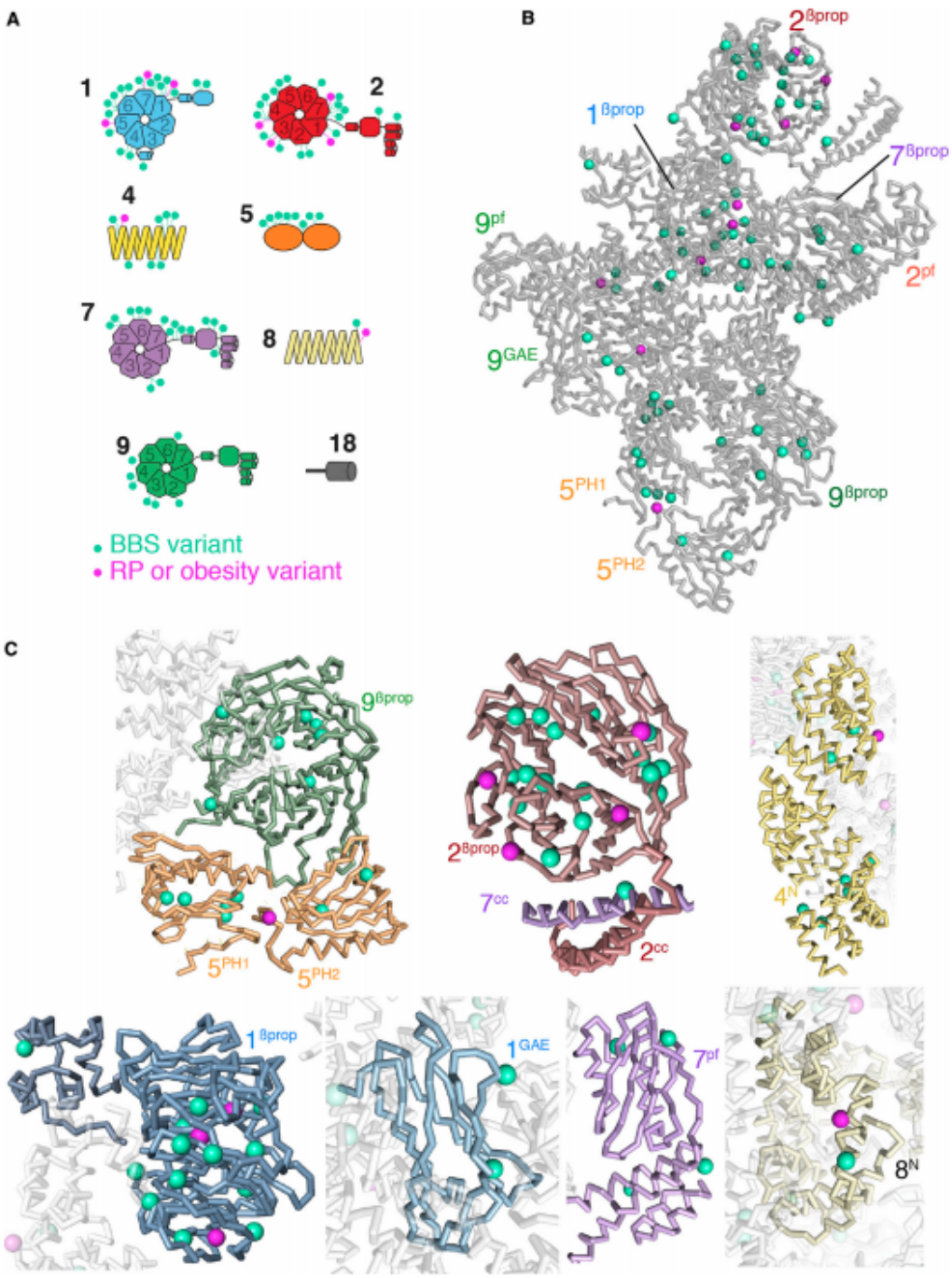

Figure 3.1.4: Mapping of Missense Pathogenic Variants onto the Ca Model of the BBSome Missense variants causing Bardet-Biedl syndrome are shown in cyan, and variants causing less severe disease phenotypes in magenta. (A) Variants were placed on diagrams of each BBSome subunit. RP, retinitis pigmentosa, i.e., retinal degeneration. (B) All variants were mapped onto the Ca model of the BBSome to show the spatial distribution of the variants. (C) Close-up views of variants present in specific domains, with BBSome subunits colored as in Figure 3.1.3.

Figure 3.1.5

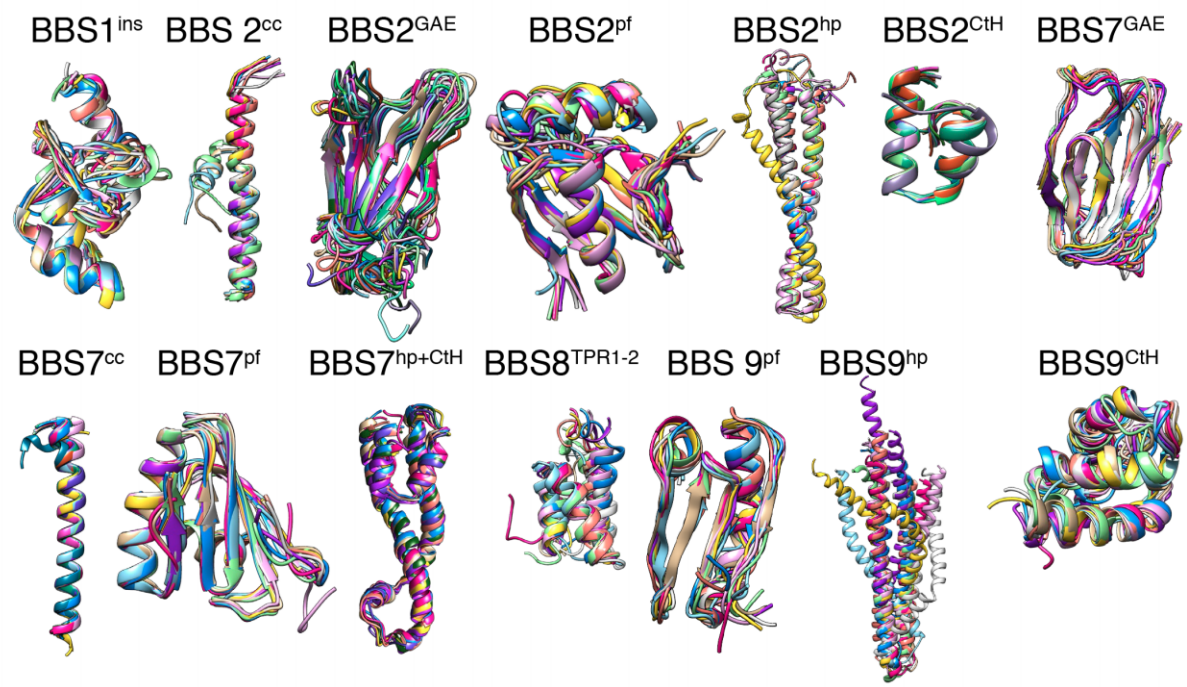

Figure 3.1.5: Details of Rosetta modeling. Top 10 scoring models for the 15 domains that could be modeled with the Rosetta ab initio protocol. For BBS7hp and BBS9CtH, the ab initio modeling only converged for a core segment. The top 10 models shown for these two domains were obtained after the consistent core segments were placed into the cryo-EM map and the unassigned segments were modeled in the context of the density with Rosetta CM.

3.2 Modeling of the Fanconi Anemia Core Complex

We illustrate the effectiveness of this approach by building an atomistic model of the *Gallus gallus* (chicken) Fanconi anemia core complex (FAcc), guided by a recently published heterogeneous 4.6 Å resolution single-particle cryoEM reconstruction and cross-linking mass-spectrometry data. In previous work, crystal structures of FANCF, FANCE and FANCL were docked and secondary-structural elements were placed into the map[64]. In contrast, here we are able to generate an atomistic model for nearly all of the complex. This method overcomes the limitations of direct interpretation of the cryoEM map, including a lack of recognizable homology to proteins of known structure for the majority of the subunits, and the relatively low resolution of substantial portions of the complex. The novel structural information provided by *trRosetta*-predicted distance distributions enables accurate topology-level predictions for domains and subunits with no recognizable homology. By combining these *trRosetta* predictions (and *Rosetta* density-guided modeling tools[65]) with subnanometre-resolution cryoEM data, we are able to infer a nearly complete FAcc model, providing key insights into the function and organization of this complex.

We illustrate the power of *trRosetta* predictions by applying this approach to build an atomic model into the recently determined cryoEM reconstruction of the Fanconi anemia core complex (FAcc[64]). These data were obtained from a fully recombinant complex after the overexpression of eight protein subunits (FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL and FAAP100) in insect cells. A 3D reconstruction at an overall resolution of 4.6 Å and cross-linking mass-spectrometry data were obtained. Secondary structure elements were previously identified within the majority of the cryoEM map and fitted with homology models (Figure 3.2.1 A). Using cross-linking, native and hydrogen–deuterium exchange mass spectrometry, as well as EM of purified subcomplexes, the general locations of all components were identified, except FANCA. However, in this previous work, residue assignments were confidently determined only for FANCL. To gain further insight into the molecular mechanisms of FAcc, atomic models of all subunits are required.

Using *trRosetta*-predicted distance distributions, we were able to determine a complete sequence assignment of the full FAcc (Figure 3.2.1 B), encompassing 5182 residues out of an expected total of 6154 residues, or 84% of the sequence, with very little unexplained density. Modeling did not make use

of the domain assignments or the backbone trace of the prior work. Our model validates many of the putative subunit assignments from the prior study (with minor differences) and provides residue-level detail of subunit locations and interactions. The next several sections describe the modeling process, followed by an analysis of our final model.

3.2.1 Fold trRosetta models

Our protocol uses multiple sequence alignments (MSAs) for individual proteins as the input to a deep residual convolutional network which predicts the relative distances and orientations of all residue pairs in the protein. These predictions are applied to a restrained minimization using a Rosetta model-building protocol. For FAcc, MSAs were generated for every chain without known homologous structures [homology models were available for portions of FANCE, FANCF, FANCL and FAAP100; see Figure 3.2.1A. Although homologous structures to FANCG also exist, there was significant structural variability within the family, and therefore we modeled FANCG with trRosetta in addition to building homology models.

From the MSAs, domains were manually parsed, and models were built using trRosetta (in regions with no known homologs) or comparative modeling (in regions with known homologs). Modeling yielded converged structures for almost all domains (Figure 3.2.2A), with typical maximal r.m.s.d.s over the top models of 2–4 Å. Several of the domains that showed poor convergence (two of the domains in FANCB and two of the domains in FAAP100) still contained subregions (`converged cores') with small deviations (2–4 Å) between models; for these cases, unconverged or poorly packed segments of the models were manually trimmed. Three of the domains (the coiled-coil domains of FANCB and FAAP100 and the α/β and CTH domains of FAAP100) were poorly converged with no `converged core'; a modified version of trRosetta (unpublished work) in which structural information on distant homologues was used as input to the neural network led to well converged models. In total, trRosetta was able to build all 42 attempted domains, which were used in the subsequent stages of the model-building protocol.

3.2.2 Assembling domains into cryoEM density

While we found FANCL, FANCF and FANCENtD straightforward to manually place into the map, ambiguity in the placement of the other subunits necessitated a more robust automated assembly procedure. Initially, the top five models for each domain were docked using an FFT-accelerated 6D search of the map. A modified version of the MC-SA sampling protocol described in Wang *et al.*[4] was then used to identify the non-clashing placement of models that maximized the overall fit of the complex model to the density. This MC-SA domain assembly assigns a placement or `not found' to each domain to account for the possibility that either all of our predicted models are incorrect or that domains are correct but not present in the map. In this way, the map serves not only to orient domains but also as validation for the *trRosetta* predictions. Some examples of model validation with the map are shown in Figure 3.2.2B. Two examples of incorrect predictions (subsequently fixed by splitting models into two domains) are shown in Figure 3.2.2C and 3.2.2D. For several domains (the aforementioned coiled-coil domains of FANCB and FAAP100 and the FAAP100 α/β and CtH domains), manual docking was necessary.

In order to model FAcc in its entirety, this Monte Carlo simulated-annealing assembly process was applied iteratively: in each round, the converged domains from the previous round were frozen, and all unassigned domains were redocked and reassembled. Convergence was assessed by manually inspecting the ten domain assignments with the best overall agreement with the density and the XL-MS data. Once the iterative process had converged (after five rounds), with the vast majority of the density occupied, the connections between domains were built and refined in the context of the cryoEM density with *RosettaCM*[24]. Additionally, the placed domains were individually inspected and poorly placed segments were also rebuilt in *RosettaCM*.

When refining the final assembled model we found that most *trRosetta* models were quite accurate, often requiring only modest (<6 Å r.m.s.d.) modifications throughout the refinement process Figure 3.2.2E. Only one placed domain required significant movement: the β-propeller domain of FANCB. To refine this domain, the model was automatically segmented into subdomains and was redocked and assembled using the same Monte Carlo procedure before refinement.

Finally, for FANCG, a repeat protein for which homologous structures were available, we additionally used *trRosetta* for modeling, as predicting changes in repeat geometries can prove

challenging. *trRosetta* yielded models that contained two long adjacent helices between residues 416 and 491, while homology modeling generated models which contained four shorter helices. In assembly, both *trRosetta* and homology models were considered, and we found that the *trRosetta* models led to a much better agreement between the model and the map. In contrast, in previous work homology modeling was used for FANCG resulting in the placement of only ~280 residues into the map[64].

3.2.3 Analysis of the final model

Using our protocol, we were able to build and assign 5182 residues (out of 6557 in the full complex); in previous work only 337 residues were assigned. Still, the protein and domain identities assigned previously were largely consistent with the models obtained with this new method: we found similar placements of FANCB, FANCF, FAAP100 and FANCL, as well as of one of the two copies of FANCG. While we were unable to identify any density associated with FANCA, *trRosetta* provided well converged models. Combining these models with the cross-linking data, we speculate that a region of unassigned density in the middle of the complex corresponds to FANCA. However, owing to the poor quality and incompleteness of the density in this region, we were not able to confidently dock the model into the map.

Our final model reveals that the `bottom lobe' (Figure 3.2.1) contains FANCBβprop, FANCL, FANCE, FANCF, FANCG and FAAP100βprop [using the domain terminology of Fig. 3.2.1A]. In contrast, in previous work FANCC and FANCE were identified within a region of density that we assign to FANCG. The `middle lobe' of our model consists of two copies of FANCBβsand+α/β+CtH, a second copy of FANCG and two copies of FAAP100βsand+α/β+CtH, all of which are consistent with the previously proposed domain assignment[64]. Finally, the `top lobe' was found to contain a second copy of FANCBβprop, FANCLELF and a second copy of FAAP100βprop. This also is consistent with the hypothesized model from the prior work. Finally, both the top and bottom lobes were connected to the middle lobe through a FANCB and FAAP100 intermolecular coiled coil. Thus, in addition to validating much of the domain assignment of previous work, our new model now provides accurate positioning of all protein residues.

3.2.4 Model Validation

One potential source of model validation arises from the cross-linking data. However, as these data were used in domain assembly, they do not serve as independent validation data. As a measure of confidence, we can still use these data by analyzing the gap between the satisfied cross-links in our model and the number satisfied by the second-best domain arrangement. In our final model, we see good agreement between the cross-links and the model (144 of 188 in total; most of the 834 cross-links in the full data set involve FANCA, which is not present in our model). Of the inter-domain cross-links, 39 of 59 (66%) are satisfied to a CA–CA distance of 30 Å, which is regarded as an acceptable distance given the usage of the BS3 cross-linker[66]. Freezing the unambiguously placed domains and redocking the remaining potentially ambiguous domains finds that a second-best arrangement, which replaces FANCC with the Ct-helices of FANCE, satisfies only 33 of 63 (52%) inter-domain crosslinks. This loss of inter-domain cross-link satisfaction provides fairly strong confidence in our final model. Further analysis of the unsatisfied cross-links reveals that most of the unsatisfied cross-links (14 of 19) occur between the C-terminus (residues 103 onwards) of FANCL. Our model suggests that one of the two copies of FANCL in the complex has a disordered C-terminal domain, strongly suggesting that most of the unsatisfied cross-links come from this disordered (and possibly dynamic) region.

One particularly strong criteria for model validation is the agreement of the maps with individual domain models. The *trRosetta* models of individual domains were predicted without using density data at all, so rigid-body fitting of these domains into density can be seen as `independent validation'. Aside from domains that exhibit internal symmetry or pseudo-symmetry (FANCBβprop and FAAP100βprop), we found that the *trRosetta* predictions all matched with real-space correlations of 0.72 or better (FANCC, 0.82; FANCENtD, 0.75; FAAP100βsand, 0.75; FAAP100α/β, 0.72), while the second-best solution (the best `wrong' solution) has a correlation that is worse by at least 0.05 in all cases. Subjectively, these second-best, incorrect placements look significantly worse. For our placed domains, this gap between the best and second-best solutions is quite large and strongly suggests that these domains are unlikely to match this well by random chance.

The overall agreement between the refined model and map is consistent with what we would expect given the resolution of the data. We were able to assess the quality of our model by segmenting it

against the three individual focused classification maps (used to generate the composite map used in modeling). We find that the model–map correlation for the bottom and middle reconstructions crosses an FSC of 0.5 at about 7.2 Å, while the top reconstruction crosses an FSC of 0.5 at about 7.1 Å. The overall model–map FSC curves show that the model–map agreements are worse at higher resolutions for the `bottom' reconstruction than the other two, which is consistent with local resolution estimates.

Additionally, we can validate models by mapping human mutation data onto the final structure. Using the Fanconi Anemia Mutation Database (http://www2.rockefeller.edu/fanconi/), we identified 30 mutations that were not identified as benign throughout the complex. While most (22) of these are in the core of protein subunits (and are likely to destabilize individual subunits), we identified four (of the remaining eight) at protein–protein interfaces in our model of the FAcc complex. Mutations of FANCB residues 230 and 236 would appear to disturb the interface between FANCBβprop and FANCGHR, while a mutation at FANCB residue 336 would disturb the interface between FANCBβprop and FAAP100βprop. Additionally, a mutation of FANCC residue 295 would be likely to disturb the interface between FANCC and FANCE. All interface mutations are marked as magenta spheres in Figure 3.2.3A, while non-interface mutations are marked with tan-colored spheres.

3.2.5 Fanconi Anemia Core Complex Discussion

Here we report a new computational method for determining atomistic models of protein complexes, guided by a subnanometre cryoEM map and cross-linking mass spectrometry data. Using distance distributions predicted from deep residual neural networks, we built accurate models of 42 domains of the FAcc, obviating the necessity for homologous high-resolution structures for interpretation of intermediate resolution maps. This provides a complete picture of the full FAcc, while previous efforts had resulted in atomic models for only three subunits (FANCL, FANCE and FANCF) in the map. The strong agreement between *RosettaTR*-predicted models and density (not used in prediction) provides validation of our predictions, as does the model's consistency with biochemical data, including cross-linking mass spectrometry and mutational studies. Our all-atom model provides molecular insight into the underlying mechanisms of previously reported disease-causing mutations, and illustrates the potential of combining intermediate resolution cryoEM density and cutting-edge de novo structure prediction.

The challenges faced when determining a model of the FAcc are not unique[42,67]. As microscopists pursue larger, more difficult, and more dynamic complexes, we will need more computational techniques that are able to build models of subnanometre resolution data with little to no homologous structure information available. While tools have been developed for integrative modelling of structures into subnanometre resolution density, all of these tools require either the existence of homologous structures for domains, or are necessarily low-resolution `domain level' models. Previous attempts to model FAcc resulted in only 387 residues being assigned to the cryoEM data, while the methods described in this paper – making use of 42 deep-learning guided domain predictions, and a protocol able to infer their arrangement – were able to increase the number of assigned residues to 5182.

Our approach shows that, while maps at these resolutions are not of sufficient quality to build models by direct chain tracing, the resolution is sufficient to assess the tertiary structure and accuracy of predicted models. In the absence of high-resolution homologous structures, the method is able to determine structures to an atomic level of detail. In addition to cryoEM data, we have recently shown that a similar approach can be applied to solve low-resolution crystal structure data where traditional molecular replacement techniques were unsuccessful[68]. We expect that the modeling power of *trRosetta* and related techniques will continue to improve in the future as the number of known sequences increases, coupled with improvements in deep-learning methodologies. We anticipate that this combined approach will be an important tool for determining atomic models of protein complexes, particularly when combined with low-resolution data sources, enabling accurate protein complex structure determination without the requirement of high resolution data.

Figure 3.2.1

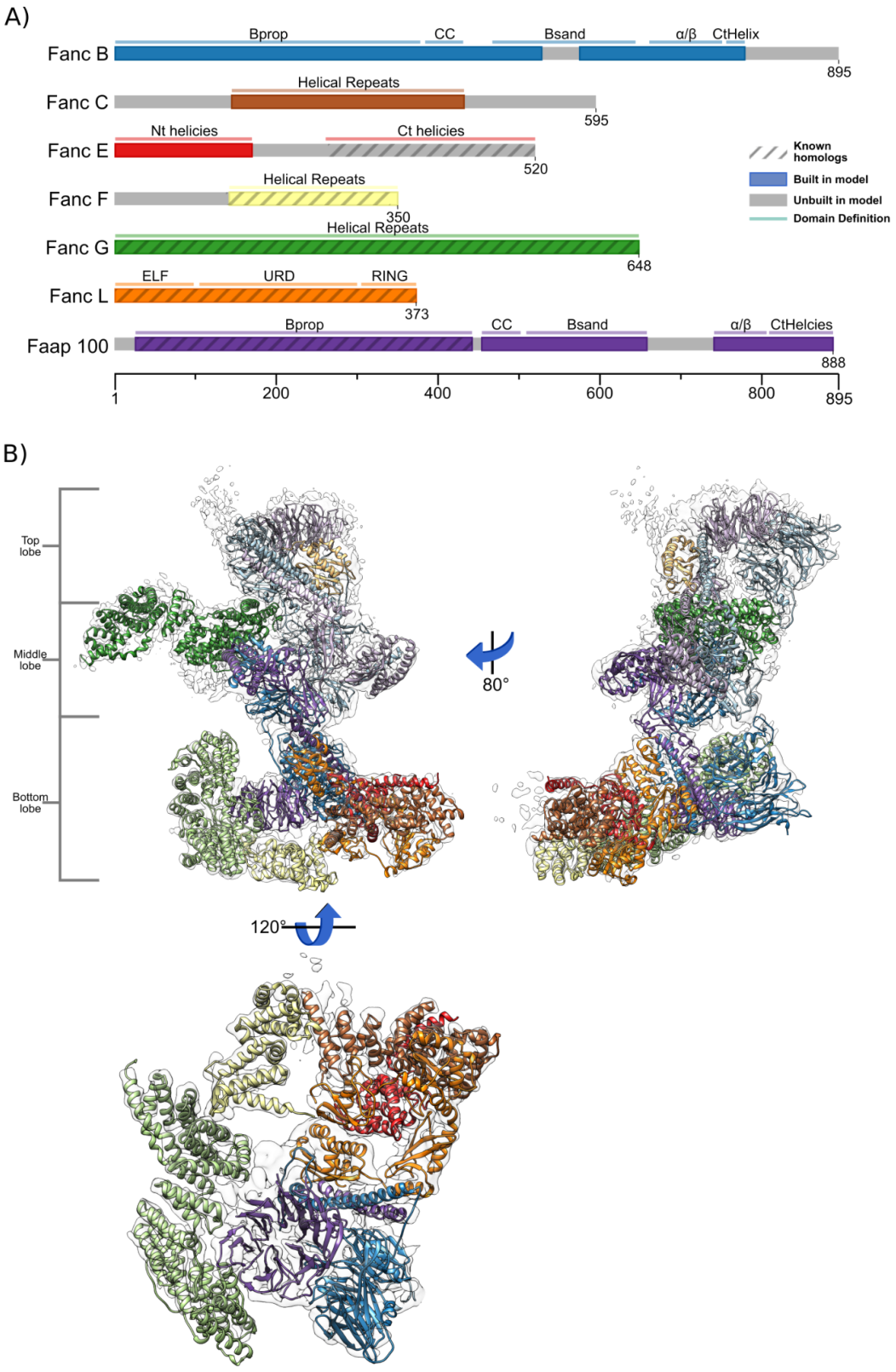
A)
Fanc B
Bprop
CC
Bsand
α/β
CtHelix
895
Fanc C
Helical Repeats
595
Fanc E
Nt helicies
Ct helicies
520
Fanc F
Helical Repeats
350
Fanc G
Helical Repeats
648
Fanc L
ELF
URD
RING
373
Faap 100
Bprop
CC
Bsand
α/β
CtHelcies
888
1
200
400
600
800
895
Known homologs
Built in model
Unbuilt in model
Domain Definition
B)
Top lobe
Middle lobe
Bottom lobe
80°
120°

Figure 3.2.1: An overview of FAcc. (*a*) Domain organization of the seven subunits of FAcc. Based on our modeling, we find that the complex consists of 18 domains, indicated with narrow bars. FANCB and FAAP100 have the same domain organization, with a β-propeller (βprop) followed by a long coiled coil (CC), a β-sandwich (βsand) and then an α/β domain, finally followed by a C-terminal helical region. FANCC, FANCF and FANCG are all comprised of a single helical-repeat domain, while we find FANCE to have two separated helical-repeat domains (one N-terminal and one C-terminal). Finally, FANCL is organized as an ELF domain followed by a URD domain and then lastly a RING domain. Also indicated is the availability of known structures or homologous proteins throughout the modeling process with striations. Domains with known structures or available homologous proteins used include the C-terminal helices of FANCE, the helical repeats of FANCF, all of FANCG and FANCL, and the β-propeller of FAAP100. (*b*) Three views of the complete model of FAcc as determined by our modeling protocol. Colors are matched to the diagram in (*a*), with those that have multiple copies (FANCB, FANCG and FAAP100) having different shades of the coloring. The orientations of the top, middle and bottom lobes are indicated.

Figure 3.2.2

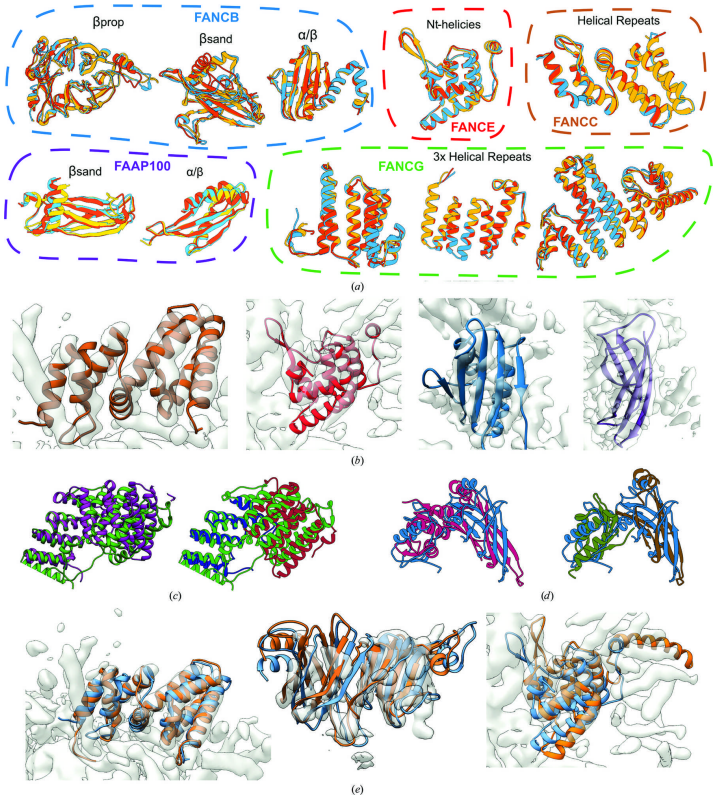

Figure 3.2.2: An overview of trRosetta-predicted domains. (a) The top three models from trRosetta for ten representative domains indicate a tight convergence of modeling. The identity of the domains follows the coloring in Fig. 2[link](a). Domains from FANCB, FANCE and FANCC are shown in the top row, while those from FAAP100 and FANCG are shown in the bottom row. (b) Several examples of trRosetta models docked into density before refinement, showing the role that the map plays in the validation and selection of models. From left to right [the colors match those in Fig. 2[link](a)]: the helical repeats of FANCC, the N-terminal repeats of FANCE, the α/β domain of FANCB and the β-sandwich of FAAP100. (c, d) Two examples illustrating the importance of domain segmentation when docking trRosetta-generated models. (c) The trRosetta model of FANCG (magenta) poorly matches the final structure (green); segmenting this model into two domains (red and blue) shows a much better match, as the individual domain structures are accurate, even though their relative orientation is not. (d) Similarly, a trRosetta prediction of the FANCB β-sandwich–α/β domain (pink) is dissimilar from the final structure (blue); splitting it into domains (brown and green) shows good overall agreement. (e) trRosetta models (blue) generally fit the map well, although some refinement was necessary to maximize agreement with the density (orange).

Figure 3.2.3

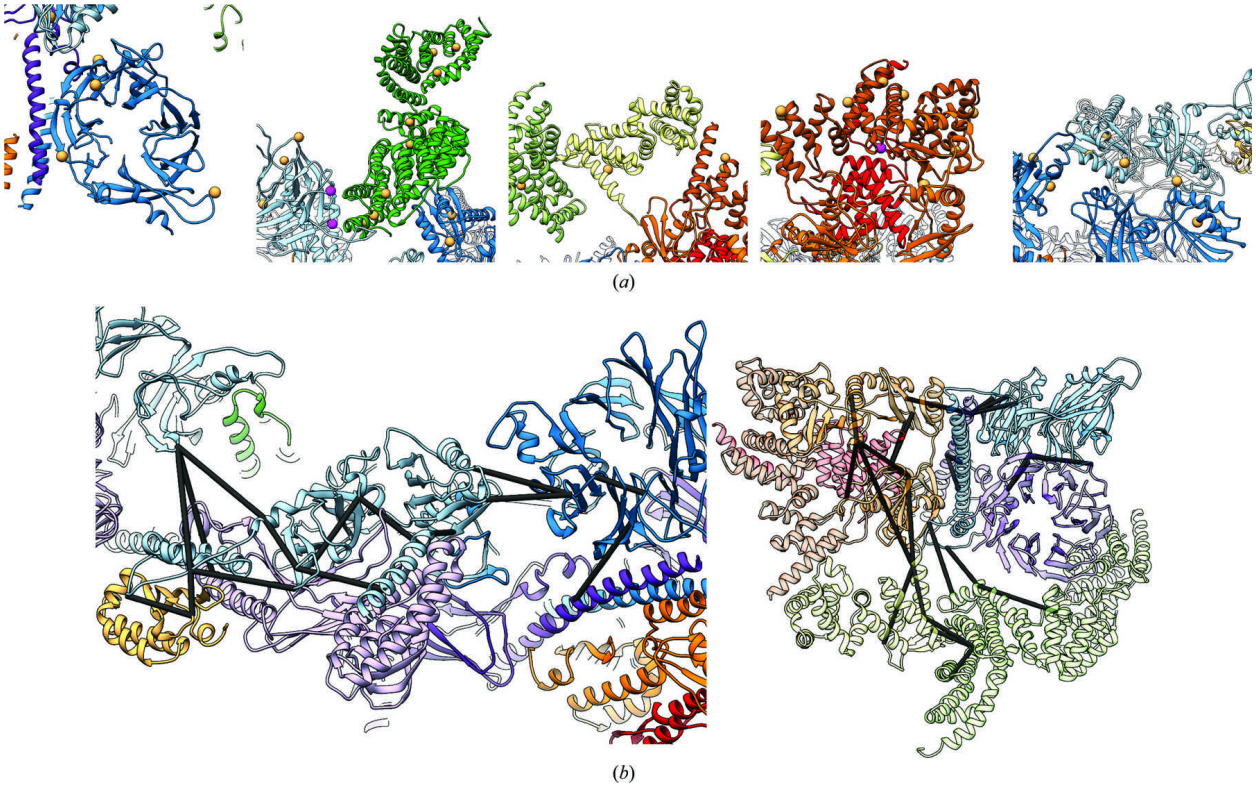

Figure 3.2.3: Model validation by mutational and cross-linking data. (a) 30 nonbenign human mutations mapped to our model of FAcc. All interface mutations are marked with magenta spheres; non-interface mutations are marked with tan spheres. (b) Close-up renderings of cross-links throughout the FAcc model. Black lines indicate cross-links that are satisfied (<30 Å) by the final refined structure. Representatives from each cross-link cluster are shown for the middle lobe (left) and the bottom lobe (right).

3.3 Modeling of the Vacuolar Protein Sorting Complex

3.3.2 Model building of the Vacuolar Protein Sorting Complex

To build the HOPS/VPS complex we employed a combination of OATS, trRosetta, and previously deposited crystal structures. First trRosetta was used to generate initial starting models for every chain in the complex. The input to trRosetta is a multiple sequence alignment (MSA) for every desired sequence. Using the methodology previously described[69] MSAs were generated for each of the sequences of VPS 16, 18, and 33 using *HHblits*[27], *hmmsearch*[28] and a custom sequence database[29]. These MSAs were fed into *trRosetta* to generate full length predicted atomic models for each sequence. As previously noted[69] models generated from *trRosetta* have a tendency to be locally accurate, but inaccurate with increasing sequence space. To take advantage of this local accuracy, the *trRosetta* models were manually segmented into domains using the previously described methodology[69]. The final domain definitions were: VPS 16, 1-323, 337-464, 466-571, 572-684, 688-833; VPS 18, 1-399, 400-555, 556-724, 726-801, 865-968; VPS 33, 38-172, 173-696. All generated domains are displayed in Figure 3.3.1.

Multiple crystal structures of VPS 16 and VPS 33 have been deposited into the PDB[1]. PDB ID 5BV0 was selected and threaded with the desired VPS 16 and 33 sequences using the *partial_thread*[24] app in Rosetta and *HHblits+HHpred*[27]. Initial attempts to manually rigid body fit both chains of the threaded 5BV0 pdb using UCSF Chimera[70] were somewhat successful. While VPS 33 could be easily fit into the cryoEM density, VPS 16 could not fit the density with the VPS 33 conformation present in the crystal structure. To fit both chains into the map, The chains were approximately placed into the density using UCSF Chimera, and then VPS 33 was split into the following domains: 5-140, (141-248, 469-543, 556-582, 599-659), (249-272, 296-315), 316-380, 381-468. Then, rigid body centroid minimization, followed by a full atom relax was performed on the entire split complex. Once the domains had been properly fit into the density together, the loops were completed, and the entire complex was refined using *RosettaCM*.

As inputs to OATS, models generated from the previous two pragraphs were utilized. While VPS 16 was easily extended using the VPS 16-33 refined model, there were initially no anchor points for VPS

16. OATS was applied iteratively, and in the first round VPS 18 domain 556-724, and VPS 16 466-571 was placed. In the second round, VPS 18 was extended C-terminally with domain 400-555, and the VPS 16 beta propeller (1-323) was placed. Due to ambiguity in the fit to density of VPS 16 337-464 this region could not be built with OATS. To properly fit this region to the density a new model was generated using the domain definition of 1-464. This was manually aligned to the placed VPS 16 beta propeller and then relaxed into the density using a “dualspace” rosetta relax[71]. In a similar fashion, a “dualspace” relax was applied to the vps16 688-833 domain after being manually aligned to the already existing model. Finally, using the *trRosetta* constraints as a guide, the VPS 18 beta propellor was manually aligned into the low resolution cryoEM density, and relaxed using a dualspace relax with constraints that would ensure the model would satisfy the *trRosetta* residue pair distance predictions. This final model was then re-refined with a combination of RosettaCM[24] and Rosetta’s *CartesianSampler* based cryoEM refinement protocols[65].

Figure 3.3.1

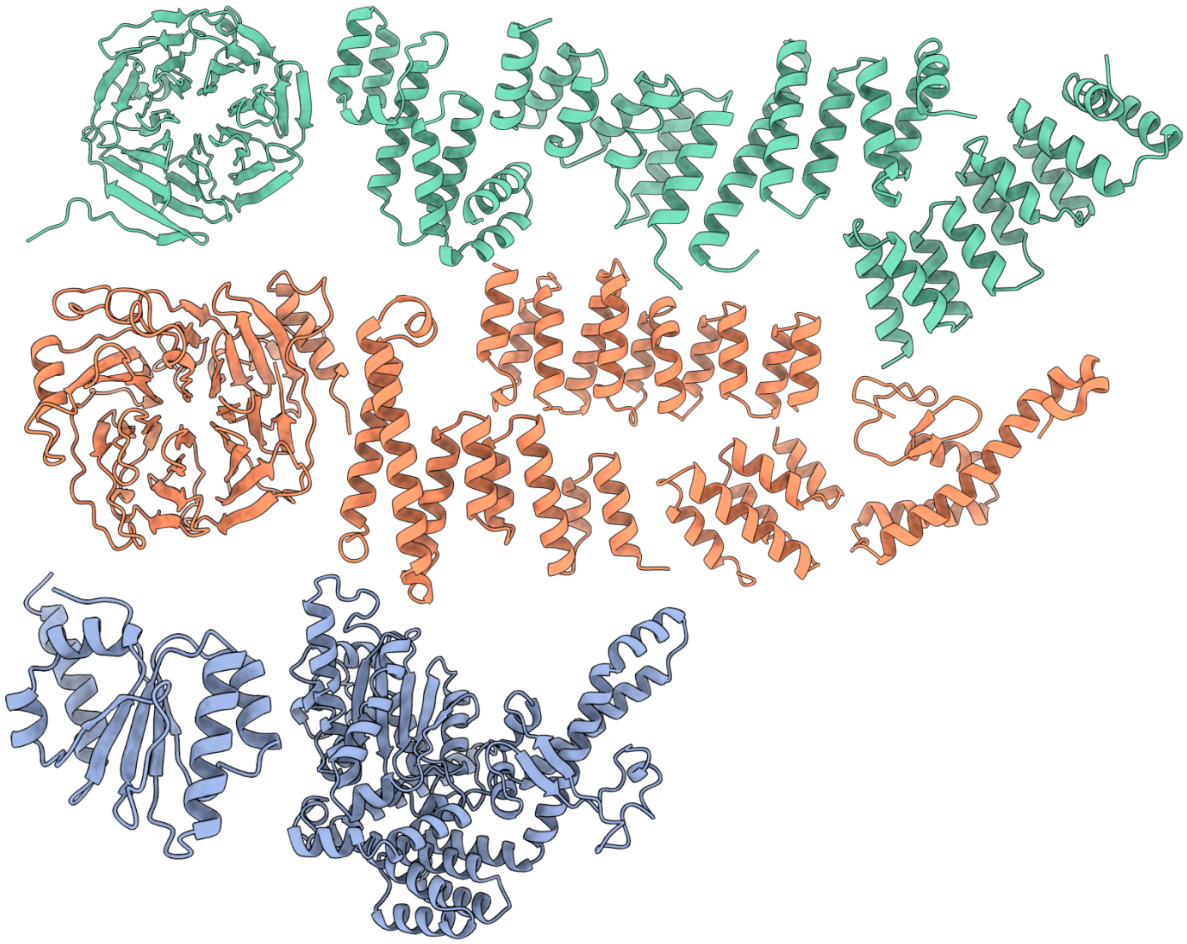

Figure 3.3.1 A depiction of all domains that were used as subunit inputs for OATS. The teal models are manually segmented subunits of VPS16, the orange models are manually segmented subunits of VPS18, and the navy models are of VPS33. The VPS33 models were not used during docking as the crystal structure was trivally docked manually into the cryoEM density. The VPS16 domains from left to right: 1-323, 337-464, 466-571, 572-684, 688-833. The VPS18 domains from left to right are: 1-399, 400-555, 556-724, 726-801, 865-968, and the VPS33 domains from left to right are: 38-172, 173-696.

Figure 3.3.2

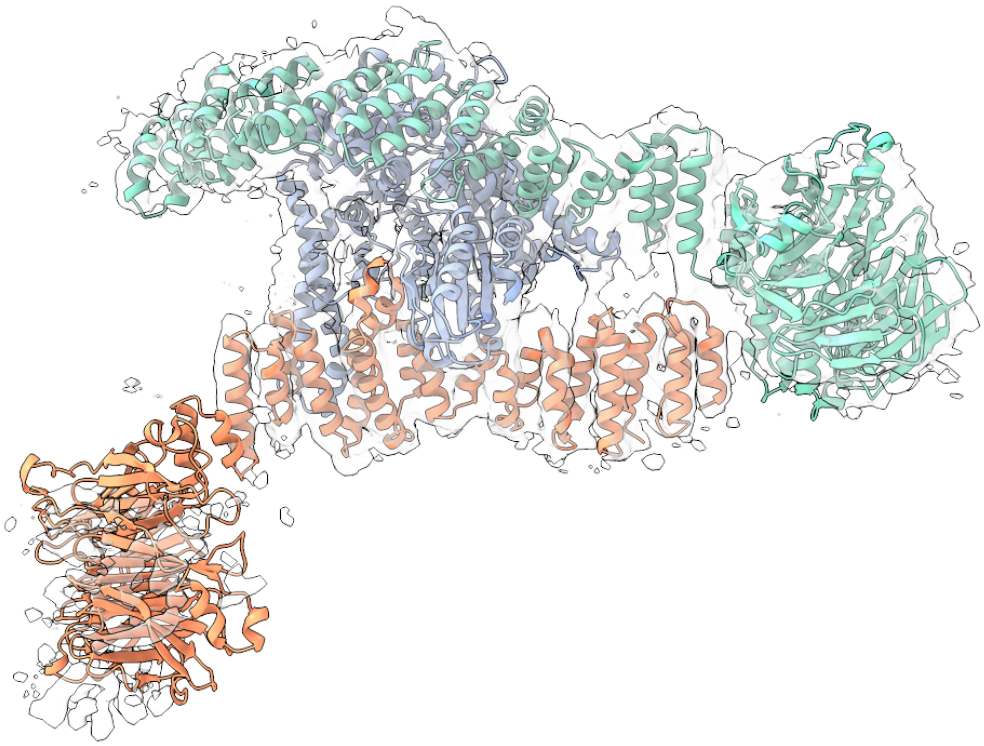

Figure 3.3.2 A depiction of the final model within the structure of the HOPS/VPS complex. The colors match the subunit coloring of Figure 3.3.1.

Figure 3.3.3

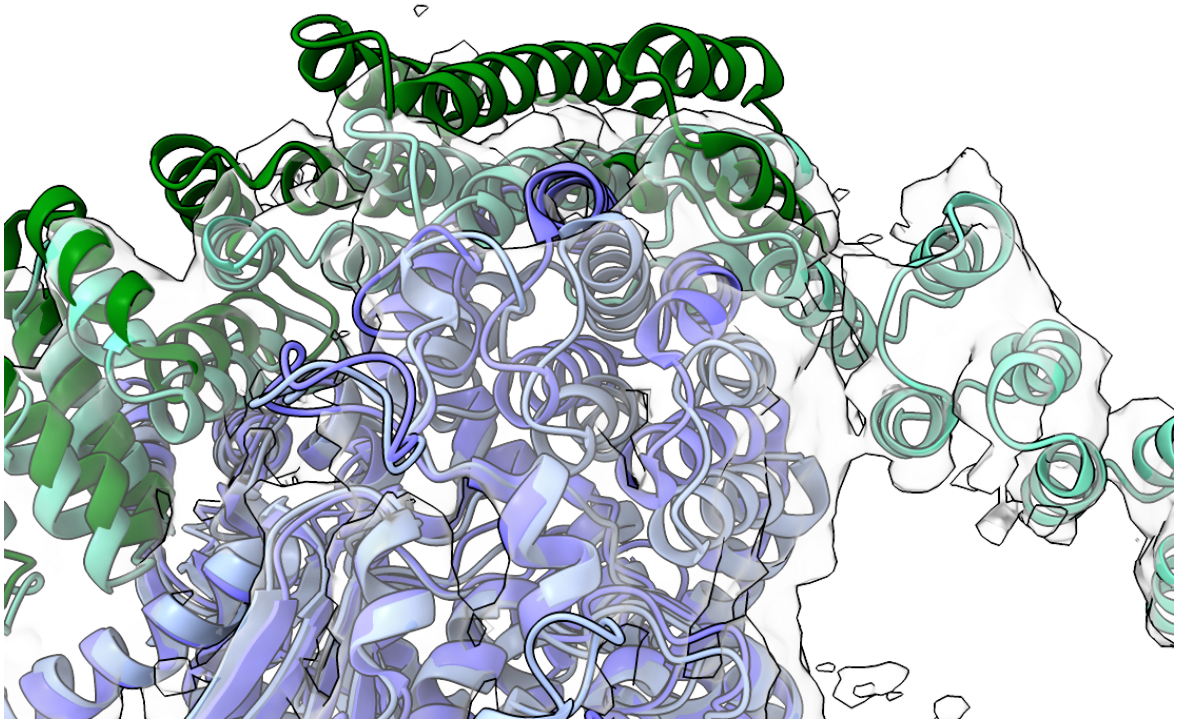

Figure 3.3.3: A comparison between the crystal structure 5BV0 and the model built using OATS into the cryoEM density. The 5BV0 structure is aligned onto the OATS model using VPS33 as a reference, and each chain is represented as a darker shade of its matching component in the cryoEM model. The chain coloring of VPS16 is shades of green, and the chain coloring of VPS33 is shades of blue.

3.4 Discussion of OATS

Developing software that can quickly, accurately, and automatically interpret cryoEM maps is of paramount importance as cryoEM becomes more widely utilized. The protocols described in this manuscript serve to fulfill that goal by combining Monte-Carlo Simulated Annealing with pairwise component interface optimization within Rosetta. Based on the benchmarking data, optimized the parameters associated with the OATS protocol, and demonstrated that OATS is widely applicable to a variety of experimental and simulated cases with a plethora of protein complexes. In addition to being benchmarked, OATS was applied to three blind cases that other structural biologists had failed to build using the currently available modeling tools.

Bibliography

1. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **10**, 980–980 (2003).
2. Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **587**, 157–161 (2020).
3. Cardone, G., Heymann, J. B. & Steven, A. C. One number does not fit all: Mapping local variations in resolution in cryo-EM reconstructions. *J. Struct. Biol.* **184**, 226–236 (2013).
4. Wang, R. Y.-R. *et al.* De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nat. Methods* **12**, 335–338 (2015).
5. Frenz, B., Walls, A. C., Egelman, E. H., Veesler, D. & DiMaio, F. RosettaES: a sampling strategy enabling automated interpretation of difficult cryo-EM maps. *Nat. Methods* **14**, 797–800 (2017).
6. Terashi, G. & Kihara, D. De novo main-chain modeling for EM maps using MAINMAST. *Nat. Commun.* **9**, (2018).
7. Afonine, P. V. *et al.* Real-space refinement in *PHENIX* for cryo-EM and crystallography. *Acta Crystallogr. Sect. Struct. Biol.* **74**, 531–544 (2018).
8. Terwilliger, T. C., Adams, P. D., Afonine, P. V. & Sobolev, O. V. A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps. *Nat. Methods* **15**, 905–908 (2018).
9. Pfab, J., Phan, N. M. & Si, D. DeepTracer for fast de novo cryo-EM protein structure modeling and special studies on CoV-related complexes. *Proc. Natl. Acad. Sci.* **118**, e2017525118 (2021).
10. Abdella, R. *et al.* Structure of the human Mediator-bound transcription preinitiation complex. *Science* **372**, 52–56 (2021).
11. Lynch, E. M. *et al.* Human CTP synthase filament structure reveals the active enzyme

conformation. *Nat. Struct. Mol. Biol.* **24**, 507–514 (2017).

12. DiMaio, F., Tyka, M. D., Baker, M. L., Chiu, W. & Baker, D. Refinement of Protein Structures into Low-Resolution Density Maps Using Rosetta. *J. Mol. Biol.* **392**, 181–190 (2009).

13. Murshudov, G. N. *et al. REFMAC* 5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* **67**, 355–367 (2011).

14. Topf, M. *et al.* Protein Structure Fitting and Refinement Guided by Cryo-EM Density. *Structure* **16**, 295–307 (2008).

15. Russel, D. *et al.* Putting the Pieces Together: Integrative Modeling Platform Software for Structure Determination of Macromolecular Assemblies. *PLoS Biol.* **10**, e1001244 (2012).

16. Lasker, K., Sali, A. & Wolfson, H. J. Determining macromolecular assembly structures by molecular docking and fitting into an electron density map. *Proteins Struct. Funct. Bioinforma.* **78**, 3205–3211 (2010).

17. Kuzu, G., Keskin, O., Nussinov, R. & Gursoy, A. *PRISM-EM* : template interface-based modelling of multi-protein complexes guided by cryo-electron microscopy density maps. *Acta Crystallogr. Sect. Struct. Biol.* **72**, 1137–1148 (2016).

18. Pandurangan, A. P., Vasishtan, D., Alber, F. & Topf, M. γ-TEMPy: Simultaneous Fitting of Components in 3D-EM Maps of Their Assembly Using a Genetic Algorithm. *Structure* **23**, 2365–2376 (2015).

19. Jauch, R., Yeo, H. C., Kolatkar, P. R. & Clarke, N. D. Assessment of CASP7 structure predictions for template free targets. *Proteins Struct. Funct. Bioinforma.* **69**, 57–67 (2007).

20. Tramontano, A. & Morea, V. Assessment of homology-based predictions in CASP5. *Proteins Struct. Funct. Genet.* **53**, 352–368 (2003).

21. Tress, M., Ezkurdia, I., Graña, O., López, G. & Valencia, A. Assessment of predictions submitted for the CASP6 comparative modeling category. *Proteins Struct. Funct. Bioinforma.* **61**, 27–45 (2005).

22. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* **117**, 1496–1503 (2020).

23. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

24. Song, Y. *et al.* High-Resolution Comparative Modeling with RosettaCM. *Structure* **21**, 1735–1742 (2013).

25. Alford, R. F. *et al.* The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **13**, 3031–3048 (2017).

26. Khatib, F. *et al.* Algorithm discovery by protein folding game players. *Proc. Natl. Acad. Sci.* **108**, 18949–18953 (2011).

27. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, (2019).

28. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).

29. Wu, Q. *et al.* Protein contact prediction using metagenome sequence data and residual neural networks. *Bioinformatics* **36**, 41–48 (2020).

30. D'Imprima, E. & Kühlbrandt, W. Current limitations to high-resolution structure determination by single-particle cryoEM. *Q. Rev. Biophys.* **54**, e4 (2021).

31. Lyumkis, D. Challenges and opportunities in cryo-EM single-particle analysis. *J. Biol. Chem.* **294**, 5181–5197 (2019).

32. Murata, K. & Wolf, M. Cryo-electron microscopy for structural analysis of dynamic biological macromolecules. *Biochim. Biophys. Acta BBA - Gen. Subj.* **1862**, 324–334 (2018).

33. Tan, Y. Z. *et al.* Addressing preferred specimen orientation in single-particle cryo-EM through tilting. *Nat. Methods* **14**, 793–796 (2017).

34. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).

35. Hattne, J. *et al.* Analysis of Global and Site-Specific Radiation Damage in Cryo-EM.

Structure **26**, 759-766.e4 (2018).

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

37. Vasishtan, D. & Topf, M. Scoring functions for cryoEM density fitting. *J. Struct. Biol.* **174**, 333–343 (2011).

38. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

39. Zhou, H., Xue, B. & Zhou, Y. DDOMAIN: Dividing structures into domains using a normalized domain–domain interaction profile. *Protein Sci. Publ. Protein Soc.* **16**, 947–955 (2007).

40. Zhang, Y. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).

41. Kotecha, A. *et al.* Rules of engagement between αvβ6 integrin and foot-and-mouth disease virus. *Nat. Commun.* **8**, 15408 (2017).

42. Chou, H.-T. *et al.* The Molecular Architecture of Native BBSome Obtained by an Integrated Structural Approach. *Structure* **27**, 1384-1394.e4 (2019).

43. Nachury, M. V. The molecular machines that traffic signaling receptors into and out of cilia. *Curr. Opin. Cell Biol.* **51**, 124–131 (2018).

44. Wingfield, J. L., Lechtreck, K.-F. & Lorentzen, E. Trafficking of ciliary membrane proteins by the intraflagellar transport/BBSome machinery. *Essays Biochem.* **62**, 753–763 (2018).

45. Jin, H. *et al.* The Conserved Bardet-Biedl Syndrome Proteins Assemble a Coat that Traffics Membrane Proteins to Cilia. *Cell* **141**, 1208–1219 (2010).

46. Ye, F., Nager, A. R. & Nachury, M. V. BBSome trains remove activated GPCRs from cilia by enabling passage through the transition zone. *J. Cell Biol.* **217**, 1847–1868 (2018).

47. Klink, B. U. *et al.* A recombinant BBSome core complex and how it interacts with ciliary cargo. *eLife* **6**, e27434 (2017).

48. Liu, P. & Lechtreck, K. F. The Bardet–Biedl syndrome protein complex is an adapter expanding the cargo range of intraflagellar transport trains for ciliary export. *Proc. Natl. Acad. Sci.* **115**, E934–E943 (2018).

49. Hinrichsen, L., Meyerholz, A., Groos, S. & Ungewickell, E. J. Bending a membrane: How clathrin affects budding. *Proc. Natl. Acad. Sci.* **103**, 8715–8720 (2006).

50. Elkhatib, N. *et al.* Tubular clathrin/AP-2 lattices pinch collagen fibers to support 3D cell migration. *Science* **356**, eaal4713 (2017).

51. Nachury, M. V. & Mick, D. U. Establishing and regulating the composition of cilia for signal transduction. *Nat. Rev. Mol. Cell Biol.* **20**, 389–405 (2019).

52. Knockenhauer, K. E. & Schwartz, T. U. Structural Characterization of Bardet-Biedl Syndrome 9 Protein (BBS9). *J. Biol. Chem.* **290**, 19569–19583 (2015).

53. Mourão, A., Nager, A. R., Nachury, M. V. & Lorentzen, E. Structural basis for membrane targeting of the BBSome by ARL6. *Nat. Struct. Mol. Biol.* **21**, 1035–1041 (2014).

54. Nachury, M. V. *et al.* A Core Complex of BBS Proteins Cooperates with the GTPase Rab8 to Promote Ciliary Membrane Biogenesis. *Cell* **129**, 1201–1213 (2007).

55. Loktev, A. V. *et al.* A BBSome Subunit Links Ciliogenesis, Microtubule Stability, and Acetylation. *Dev. Cell* **15**, 854–865 (2008).

56. Park, H. *et al.* Simultaneous Optimization of Biomolecular Energy Functions on Features from Small Molecules and Macromolecules. *J. Chem. Theory Comput.* **12**, 6201–6212 (2016).

57. Bradley, P. *et al.* Free modeling with Rosetta in CASP6. *Proteins Struct. Funct. Bioinforma.* **61**, 128–134 (2005).

58. Ovchinnikov, S. *et al.* Protein structure determination using metagenome sequence data. *Science* **355**, 294–298 (2017).

59. Blok, N. B. *et al.* Unique double-ring structure of the peroxisomal Pex1/Pex6 ATPase complex revealed by cryo-electron microscopy. *Proc. Natl. Acad. Sci.* **112**, E4017–E4025

(2015).

60. Yam, A. Y. *et al.* Defining the TRiC/CCT interactome links chaperonin function to stabilization of newly made proteins with complex topologies. *Nat. Struct. Mol. Biol.* **15**, 1255–1262 (2008).

61. Seo, S. *et al.* BBS6, BBS10, and BBS12 form a complex with CCT/TRiC family chaperonins and mediate BBSome assembly. *Proc. Natl. Acad. Sci.* **107**, 1488–1493 (2010).

62. Sinha, S., Belcastro, M., Datta, P., Seo, S. & Sokolov, M. Essential Role of the Chaperonin CCT in Rod Outer Segment Biogenesis. *Investig. Opthalmology Vis. Sci.* **55**, 3775 (2014).

63. Zhang, Q., Yu, D., Seo, S., Stone, E. M. & Sheffield, V. C. Intrinsic Protein-Protein Interaction-mediated and Chaperonin-assisted Sequential Assembly of Stable Bardet-Biedl Syndrome Protein Complex, the BBSome. *J. Biol. Chem.* **287**, 20625–20635 (2012).

64. Shakeel, S. *et al.* Structure of the Fanconi anaemia monoubiquitin ligase complex. *Nature* **575**, 234–237 (2019).

65. Wang, R. Y.-R. *et al.* Automated structure refinement of macromolecular assemblies from cryo-EM maps using Rosetta. 22.

66. Merkley, E. D. *et al.* Distance restraints from crosslinking mass spectrometry: Mining a molecular dynamics simulation database to evaluate lysine–lysine distances. *Protein Sci. Publ. Protein Soc.* **23**, 747–759 (2014).

67. Kim, S. J. *et al.* Integrative structure and functional anatomy of a nuclear pore complex. *Nature* **555**, 475–482 (2018).

68. Bhargava, H. K. *et al. Structural basis for autophagy inhibition by the human Rubicon-Rab7 complex*. http://biorxiv.org/lookup/doi/10.1101/2020.04.18.048462 (2020) doi:10.1101/2020.04.18.048462.

69. Farrell, D. P. *et al.* Deep learning enables the atomic structure determination of the Fanconi Anemia core complex from cryoEM. *IUCrJ* **7**, 881–892 (2020).

70. Pettersen, E. F. *et al.* UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

71. Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E. & Baker, D. Relaxation of backbone bond geometry improves protein energy landscape modeling: Relaxation of Backbone Bond Geometry. *Protein Sci.* **23**, 47–55 (2014).