Advances In Computer Aided Protein Structure Determination From Sparse Cryo Electron Microscopy Data

Brandon Frenz

A dissertation

Submitted in partial fulfillment of the

Requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Frank DiMaio, Chair

David Baker

Phil Bradley

Program Authorized to Offer Degree:

Department of Biochemistry

University of Washington

Abstract

Advances In Computer Aided Protein Structure Determination From Sparse Cryo Electron Microscopy Data

Brandon Frenz

Chair of the Supervisory Committee:

Professor Frank DiMaio

Department of Biochemistry

Single-particle cryo-electron microscopy (cryoEM) has become a powerful tool for determining macromolecular structures. Thanks to recent advances in direct electron detectors and motion correction algorithms it can frequently deliver electron density maps in the range of 3-5Å resolution. To obtain as much atomic level detail of the structure as possible from this data an accurate atomic model must be built. This can be done manually however, it is laborious and error prone. To resolve this problem modelers have turned to computational tools which can make up for lack of experimental data. Here we describe several tools for modeling with sparse experimental data, including a novel sampling strategy for *de novo* model completion and a novel refinement strategy for glycans with near atomic resolution cryoEM and x-ray crystallography data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acknowledgements

I would like to express my gratitude to my thesis advisor Frank DiMaio for his consistent support and willingness to help me overcome both scientific and software development challenges. His unending patience with me (and other students in the lab) as I developed my skills as computational biochemists will forever be appreciated. In addition his incredible knowledge of the Rosetta software suite has been an invaluable resource throughout the development of the protocols described in this thesis. I has been a pleasure to work with him over these last few years.

I would also like to thank all the members of the DiMaio lab and those have worked closely with us including Ray Wang, Patrick Conway, Ryan Pavlovic, Dan Farrell, and Zibo Chen. I would especially like to mention Ryan and Dan who, in addition to being great coworkers, have also been good friends. I am grateful for their support throughout this process and look forward to working with them in the future.

I would like to thank all of my collaborators but especially those from the Veesler lab including David Veesler, Lexi Walls, and Andrew Borst. We have done many interesting projects together, several of which are described in this thesis. I would like to thank them both for sharing their data with me and also for serving as the beta testers for much of my code. I'm sure at times it was frustrating to use buggy software early in its development but the methods described in this paper are more robust and more error free than they would be without their willingness to do so.

I must also offer my gratitude to all of the sysadmins who maintained the computational systems used throughout the projects described here, including Luki Goldschmidt, Darwin Alonso, and Patrick Vecchiato.

My appreciation should also be extended to the members of my thesis committee David Baker, Phil Bradley, Justin Kollman, and Andy Scharenberg. They have been a source of valuable advice and encouragement. I appreciate their trust in my abilities and willingness to let me be flexible as I changed project directions as opportunities arose.

Thank you also to all of my friends and family who have supported me along the way. Including my parents Kim and Sue Frenz as well as well as my siblings Lexie and Jeremy. I would also like to thank my good friends Dan Rheaume and Johnny Hampton for their constant support. Each of them has played an important role in shaping who I am today.

My family has also recently increased with the addition of new in laws all of whom have been incredibly supportive of me throughout this process including mother in law Kim Morrison, Father in law Bed Ada, and grandparents Dave and Linda Riley. I am grateful to all of them for their support.

Finally I would like to thank my wife Tiffany who has been a constant source of happiness in my life. In addition to being a wonderful partner and dear friend she has also been a great editor for many of the documents I've written during graduate school, work for which I am extremely grateful. Thank you for your constant love and support.

Dedication

To my wife Tiffany

Chapter 1. INTRODUCTION

Recent advances in the field of cryo electron microscopy (cryoEM), particularly the advent of direct electron detectors[1] and motion correction algorithms[2–5] have led to a dramatic improvements in quality of data routinely obtained via the method. It is now fairly common to obtain resolution in the range of 3-5Å referred to as "near atomic" here. As of October 2017 the electron microscopy databank contains 1015 entries in this resolution range. The rate of density maps deposited under 6Å has dramatically increased with 426 new maps deposited under 6Å in 2016 nearly 10x the 43 deposited in 2013 and this trend expected to continue in 2017 and beyond.

Despite these advances determining the precise atomic structure of the protein remains a challenge at resolutions in the range of 3Å and worse. This is because, unlike in high resolutions maps, atomic details are not apparent in the density, with large side chains usually appearing as a only protrusion from the backbone and small side chains often missing entirely. To resolve this issue modelers have turned to the tools developed for computational protein structure prediction to make up for the lack in experimental data.

1.1 Methods for Computational Modeling with Near-Atomic Data

A number of methods are available to aid modelers attempting to determine structures. These includes *de novo* methods originally developed for work with x-ray crystallography that have been adapted for work with cryoEM.[6–10] However, while these methods are useful in certain circumstances, they perform poorly at resolutions 3Å or worse. CryoEM specific methods have also been developed however, while they are useful for assigning Cα traces they struggle to accurately assign sequence at near atomic resolution.[11,12] In addition to *de novo* modeling

methods have also been developed to perform rigid body and flexible fitting[13-15] however these protocols are limited to cases in which an accurate starting model is available.

As one of the premier software packages for protein structure prediction and design Rosetta offers a number of tools for modeling with sparse experimental including integration of electron density as part of the Rosetta score function,[16] a large number of methods for manipulating the protein structure[17] and a pipeline for multi template homology modeling via RosettaCM.[18] This makes Rosetta an ideal development platform for new methods related to modeling with near atomic resolution electron density.

Previously a fully automated method for *de novo* modeling with near atomic resolution cryoEM data was developed in Rosetta[19] however, it's method for model completion, the *de novo* modelling capabilities of RosettaCM, required greater than 70% of the model to be completed in order to obtain reliable results. This limited it's applicability to only cases in which the *de novo* fragment docking step could converge on a fairly complete solution significantly reducing the number of structures which could be determined via the method.

Chapter 2. *DE NOVO* MODEL COMPLETION WITH ROSETTAES

2.1 Introduction to RosettaES

In order to overcome the limitations of the previous *de novo* modeling protocol we developed a novel sampling strategy, called Rosetta enumerative sampling (RosettaES), which uses fragment based assembly to enumerate a pool of possible solutions that are both consistent with the data and possess physically realistic geometry. By taking advantage of the electron density and the Rosetta energy function we are able to prune the set of solutions to a computationally tractable size even when building dozens of missing residues.

RosettaES uses a beam search strategy where a fixed-size ensemble of solutions is maintained throughout sampling (Figure 2.1). The method attempts to complete partial models using the EM density in the sampling process. Modeling starts at the N or C terminus of an internal missing segment and putative solutions are generated by adding 1 additional residue at a time. Conformations are sampled using 3 residue long fragments mined from the pdb which are used to manipulate backbone positions of the newly added residue and the 2 residues proximal, referred to as fragment overlap. The ensemble of solutions is the pruned to remove conformations that are too similar to another solution in the ensemble, energetically unfavorable, or inconsistent with the data. If the ensemble exceeds a user defined size solutions are clustered until the ensemble size matches the defined cap.

2.2 Benchmark Sets

To test this protocol we developed two benchmark sets based on the round 1 models of the *de novo* fragment docking protocol used on the maps described in wang et al.[19] All of these 9 maps are real experimental data and range from 3-5Å resolution. Fragment docking was run on

these maps and models with coverage from 20-80% of the deposited structure were generated. Models were examined and fragments which deviated significantly from the deposited model were removed to ensure it was possible to generate an accurate solution given the input model. These models represented one benchmark set and another was generated by identifying missing segments from these models and removing each of those missing segments independently from the deposited structure. This second set was developed to ensure a fair comparison to Rosetta comparative modeling (RosettaCM),[18] and other methods that complete full structures, could be performed.

2.3 Fragment Sampling

The primary goal of RosettaES is to make up for deficiencies in sampling which exist in other methods, particularly RosettaCM, therefore it is fundamental to the algorithm that it is possible to accurately sample the target structure using the provided set of input fragments. To determine a criteria required to ensure sufficient fragments we ran the RosettaES protocol with a varying number of fragments as input with a range of fragment sizes. The cap on the maximum ensemble size was set to 5 however, rather than evaluate models based on the Rosetta energy and density score, models were selected based on the RMSD to the deposited model. We found a number of fragment strategies were capable of accurately sampling every structure in our benchmark set and, as expected, the number of fragments required to accurate sample the deposited model increased with their length (Figure 2.2). Ultimately a tradeoff exists between the run time, determined by the number of fragments used and the number of residues added at each step, and the accuracy. Although other schemes were capable of completely covering the required conformational space, empirically, we found that a strategy using 100 3 residue long

fragments with a 2 residue overlap generated the most accurate results. This is likely because the finer sampling generated more accurate models with better scores making identification of incorrect conformations more apparent. In addition to the 3 residue long fragments 20 9 residues long fragments were also used to improve sampling in helical regions where the 3 residue long fragments were often insufficient to generate the hydrogen bond patterning necessary to provide clear energy signal.

An important implication of this test is that for every segment in our benchmark set an accurate solution could be generated. This means that, given enough computational power, if one could generate a full set of all possible fragment combinations it would guarantee a solution near the target conformation would exist within that set. Practically this cannot be done however a modest increase in maximum ensemble size is sometimes sufficient to generate an accurate model (Figure 2.3). Modellers are therefore encouraged to increase the cap on the ensemble size when dealing with challenging problems that require additional sampling.

The fact that some combination of fragments exists to form a correct solution suggests that it is therefore the goal of the sampling algorithm, clustering, and electron density to filter out incorrect solutions into a computationally tractable set so that the partial combination of fragments required to generate an accurate model remain ranked enough to persistent throughout the protocol.

2.4 Additional Features

In order to aid the score function and electron density in guiding the sampling a number of additional features were included in the algorithm. Of particular note are a penalty on discontinuous density, a two-tiered filtering strategy to enhance sampling diversity, a beta sheet

sampler designed to orient the hydrogen bonds of beta sheets, and a two tiered approach to modeling side chains in order to efficiently capture and reward correct sequence registration.

2.4.1 Discontinuous Density

Within Rosetta two methods exist for scoring a protein structure in an electron density map, both are based off the idea of creating a synthetic map at a given resolution using the model to be evaluated. This synthetic map is then used to calculate a probability of observing the structure give the experimental map. In the first method, elec_dens_window, a mask is created around a window of residues and the per residue density score is calculated using this mask. This method is slower but more accurate than the second method, elec_dens_fast, which pre computes an atoms worth of density and stores it on a grid. This grid is then used as a look up to rapidly score the structure. This grid based method offers greater speed and scores correlate well with the slower method.

One limitation of both these approaches is that they do not properly account for the continuity of the backbone density that generally occurs in near-atomic resolution data. This means that structures that cross back and forth through different parts of the backbone, such as in a beta sheet, are often viewed as a reasonably good fit despite the fact that these conformations are clearly incorrect by manual inspection.

To deal with this problem we modified our density score function to penalize a window of residues in the structure based on the worst scoring N, C, or Cα atom in that window. This modified score is done using elec_dens_fast to calculate the individual atom score and downscale the residue score of each residue in the window by capping the total density score for

each atom to this value and then adding in a fraction of the original score based on a defined weight, 30% by default.

Using this strategy we greatly reduce the number of solutions viewed as acceptable by our score function and therefore improve the algorithm's ability to identify the correction solution.

2.4.2 Two-tiered Filtering

One characteristic of, particularly cryoEM, electron density maps is variable quality among different regions of the map. This creates a problem when attempting to complete a partial model as sampling is drawn towards regions of higher quality at the expensive of sampling the weaker regions. When dealing with partial models for which multiple segments are missing this variance in region quality results in all of the sampling focusing on only the regions of high quality, which may only correspond to one missing segments, meaning no good solutions are generated for the other segments that need to be completed making the final assembly of the complete model impossible.

To solve this problem we implemented a two-tiered clustering strategy. In the first tier conformations which are overly similar are removed. This process works by first accepting the best scoring solution and then progressively adding new solutions to the pool. If a solution has an RMSD less than 1.5 Å to any solution that has been accepted previously this lower ranking model is rejected.

In the second tier of the filtering first a 0.8 multiplier is applied to the best scoring model and all models scoring worse than this are removed. Conformations are then added to the pool of accepted partial solutions in rank order. If any solution is within 3Å of a previously accepted

model it is passed over until all models above the cutoff have been tested. After this point another pass is made adding 1 more solution to each "cluster" of models within 3Å. This process is repeated round robin style until the cap on the total number of allowed solutions is hit or there are no more solutions better than the cutoff value.

2.4.3 Beta Sheet Sampling

Beta sheets pose a particular challenge when working with incomplete models. This is because, while it is often fairly easy to recognize the general direction of the sheet by eye, from the perspective of an individual conformation within the sheet there are many possible paths that all look reasonably valid. Without the context of the complementary strands it is difficult for the score function to detect the correct path. This is especially true when the resolution is low enough such that side chains often appear as linkage between the two sheets preventing the discontinuous density penalty from fully eliminating these conformations.

To resolve this issue we created a mover within Rosetta called the "Sheet Sampler" this mover takes a structure and a window of residues, default 4, and attempts to add complementary strands on either side. This is done by placing points at ideal positions from hydrogen bond donors and acceptors and aligning standardized anti-parallel sheets to these points. These sheets are then minimized into the density and if they fit the density well, no atoms in the sheet backbones worse than the worst atom in the input residues, and do not clash with the input structure the solution with these sheets receives a score bonus as a fraction of each sheets score, default 50%.

By using this strategy whenever we are in beta sheet ramachandran space we can reduce conformations which do not maintain sheet continuity and favor those which consistently extend

the sheet when possible. Conformations which create hydrogen bond patterning towards the complementary strands are also rewarded making assembly of the completed sheet more likely.

2.4.4 Efficient Use of Side Chain Information

Within Rosetta there are two popular representations of atomic structures. The first all atom representation (full atom) which explicitly represents every atom and the second (centroid) treats the side chain atoms as single centroid extending from the backbone. The centroid representation has several advantages including a more smooth energy landscape which makes sampling less like to be stuck in a false local minima, and it also offers considerable speed advantages over full atom, roughly 10x for our purposes. The disadvantage of the centroid representation is the lack complete side chains and inability to sample rotameric conformations for those side chains.

In order to take advantage of the side chain information of the full atom representation while reducing its downsides a scheme by which all potential solutions are first scored, minimized, and filtered using the centroid representation is used. From these potential solutions the top 2N , where N represents the maximum cap on the ensemble size, have their side chains repacked in full atom. These conformations are then rescored using the side chain information before a second round of filtering reduces the size of the ensemble to N using the filtering strategies described above.

2.4.5 Results of Additional Features on Model Accuracy

To test how these additional features improved sampling accuracy we ran RosettaES on the benchmark set of round 1 fragment docked models progressively adding one new feature at a

time. Backbone and C GDTha was calculated against the deposited model and reported. There was a very clear benefit to using the discontinuous density penalty and the two-tiered sampling strategy. Small improvements were also observed for the sheet sampling and for the inclusion of side chain information, although further tests should be done to determine whether the increased computational time justifies their use (Figure 2.4.b).

2.5 Comparison To RosettaCM

RosettaCM, with its ability to build missing segments a model *de novo*, has been the dominant tool for model completion with low resolution experimental data in Rosetta and also significantly outperforms comparable non-Rosetta tools for this purpose.[20] However the sampling strategy used by RosettaCM is optimized for homology modeling without experimental density which has led to several pathologies in the sampling. The core of the RosettaCM *de novo* sampling strategy involves building the entire missing segment at once, placing it in a in a random orientation, and then attempting to manipulate it into the density. This strategy is problematic due to the fact that when the model is significantly outside of the density small changes, which may be correct locally, do not receive any signal from the experimental data. This results in the sampling strategy wasting a large amount of computational time sampling conformations which are clearly non-viable, final solutions are then often completely outside the proper region of density (Figure 2.4.c).

To compare RosettaES to RosettaCM we used the previously described benchmark set in which the missing segments were removed from the deposited model. In theory this set should favor RosettaCM which should struggle more with false density minima to a greater degree than RosettaES. Nevertheless RosettaES significantly outperformed RosettaCM across the entire

benchmark set. In particular the results on longer segments, up to 100 missing residues, are quite striking, with RosettaES performing far better on that RosettaCM for these problems (Figure 2.4.a and Table 2.1). A direct example of this can be seen in the comparison of the models built for residues 187-267 of FrhA which demonstrate the ability of RosettaES to build models that follow a defined path of density whereas the best scoring model produced by RosettaCM is significantly outside of the map (Figure 2.4.c).

2.6 Comparison of RosettaES to non-Rosetta tools.

In addition to RosettaCM a number of other tools have also been used in an attempt to build *de novo* models using sparse cryoEM data including buccaneer,[21] modeller,[22] privateer,[11] and phenix autobuild.[9] We attempted to use these tools on a subset of our benchmark set and were successfully able to run all but privateer. The percentage of residues assigned to sub 2 Å backbone RMSD to their deposited counterparts were recorded and RosettaES significantly outperformed every other method (Table 2.2).

2.7 Modeling Multiple Missing Segments

RosettaES clearly outperformed RosettaCM on modeling problems involving a single missing segment however models from the *de novo* fragment docking protocol typically contain several missing segments. It is therefore critical to be able to not only sample individual segments but also assemble them into complete models. To do this we introduce several additional strategies including a monte carlo assembly algorithm and taboo sampling.

2.7.1 Monte Carlo Assembly

In order to assemble a final model when multiple missing segments are present two things must occur; 1, an accurate solution must be generated for each segment and 2, from the pool of solutions for each segment the correct solution must be chosen. We've already discussed how an accurate solution can be sampled, to solve the second problem we implemented a monte carlo assembly algorithm that attempts to find a set of non-clashing solutions in the structure from the pool of partial models. To improve runtime the "one body" scores for each solution are kept and the "two body" van der walls clash energies are calculated for each solution pair. These are then stored in a table for efficient lookup.

In the next step the assembly algorithm assigns a random potential solution for each missing segment. The score is summed and a new potential solution is substituted. This change is then accepted or rejected based on the monte carlo metropolis criterion. Initial temperature is set to 200 and halved at each of 6 rounds. 250 moves are made for each temperature. During the sampling the best model visited is stored and output when the sampling is complete. This process is repeated 100 times generating 100 models. The clash score of the least clashing model is reported and this score is used to signal whether to move to the refinement step or generate additional solutions via taboo sampling described below.

2.7.2 Taboo Sampling

The most common way in which the RosettaES protocol fails is when it is unable to accurately sample a particularly difficult segment of the structure. Fortunately this is generally fairly easy to detect as it is rare that the monte carlo assembly algorithm will be able to find an incorrect solution that does not have significant internal clashes. However, when such clashes are

present taboo sampling can be used to generate additional potential solutions for the monte carlo assembly.

Taboo sampling is a search strategy which, when an initial search fails, uses the results of previous rounds of sampling to drive the future rounds towards unexplored paths. In the context of RosettaES it is particularly suited for dealing with cases in which the monte carlo assembly is unable to converge on a non-clashing solution. Taboo sampling is first set up by writing all intermediate solutions that pass the two tiered filters to the hard disk. When additional rounds of taboo sampling are required, there intermediate solutions can be loaded back into Rosetta and future filtering of new solutions will use them to remove conformations which are overly similar and they will also be considered as better scoring models for the purposes of the round robin second filtering tier.

2.7.3 Results of Full Assembly

To test this protocol we ran our full pipeline, with up to 4 rounds of taboo sampling, on our benchmark set of 9 structures, starting from the round 1 *de novo* fragment docking models. In four cases, FrhA, FrhB, FrhG and TMV RosettaES was able to generate accurate models, less than 2 Å backbone RMSD, to the deposited structure. In one case, TRPV1, RosettaES generated a non-clashing model 3.6 Å from the deposited. For the other four structures BPP1, VP6, STIV, and T20S which all contained large beta sheets, we were able to generate accurate solutions for most, but not all, of the missing segments and therefore the monte carlo assembly was unable to converge on a final solution. (Table 2.3).

2.8 Evaluation and Validation of Proposed Models

An important aspect of computational model building with near atomic resolution experimental data is the ability to accurately determine when a correct solution has been generated. There are four major ways in which this can be done with RosettaES.

First, models can be scored using the Rosetta energy function and fit to density. Accurate solutions will be free from clashes and have good Rosetta energies with a strong electron density score.

Second, models should fit together in a complete assembly. As discussed in the previous section when the input model contains multiple missing segments solutions to those segments should be generated in such a way that they can all fit together without creating clashes. The ability to generate a complete model with this feature is a good indication RosettaES has generated an accurate model.

Third, model convergence can be used to provide an indication that RosettaES has generated an accurate solution. By comparing the RMSD of each residue in the solution set to the centroid RMSD for that residue in the ensemble the diversity within the ensemble can be determined. Ensembles which have very little model diversity generally have an accurate solution within that pool, whereas ensembles with high levels of diversity are less likely to contain an accurate conformation (Figure 2.5).

Finally models should always be evaluated manually to ensure good fit to the density and, particularly, that there are not large unexplained volumes in the density. In addition, as will be discussed in the following section, external knowledge, such as the presence of glycosylation sites, can also be used to inform model selection.

2.9 Modeling Novel Structures with RosettaES

In addition to our benchmark sets RosettaES was also used to build several challenging unknown targets. These include two difficult domains of the mouse hepatitis virus, the C-terminus of the human coronavirus NL63, a the flexible c-terminus of the bamboo mosaic virus (BaMV). Each of these is discussed in detail below.

2.9.1 Modeling The Mouse Hepatitis Virus Spike

With the outbreak of severe acute respiratory syndrome coronaviruses (SARS-CoV) in 2002 and Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012 coronaviruses of are increasing medical concern with fatality occurring in 10-37% of SARS/MERS infections. There are no known antiviral treatments for coronaviruses.[23] In addition to SARS and MERS coronaviruses are responsible for 30% of cases of mild respiratory infection and atypical pneumonia worldwide making them an incredibly common human pathogen.[23]

The primary target of coronavirus neutralizing antibodies is the S homotrimer which is used by the virus for cell adhesion and fusion. The trimer contains a S_1 subunit which possesses the receptor binding domain and a S_2 subunit which mediates membrane fusion. The peptide is synthesized as a single chain precursor of about 1300 residues which forms a trimer upon folding. This precursor is cleaved by the host proteases between the S_1 and S_2 in some coronaviruses such as the canonical coronavirus model mouse hepatitis virus (MHV)[24].

Several crystal structures for the coronavirus S post-fusion core exist [25–28] and several structures S receptor-binding domains in complex with their receptors have also been determined[29-32] however the lack of a detailed atomic structure of the S trimer has limited analysis of the infection mechanisms.

To resolve this problem our collaborators produced a MHV S ectoderm trimer with enhanced stability by mutating the S_2 cleavage site and fusing a trimerization motif at the C terminus. Using this construct a cryo-electron microscopy reconstruction to 4.0 Å was generated.

To create a model for this structure we first docked crystal structures into the density of two S1 domains [29-33]. Ultimately all residues from 15-1182 were built with the exception of residues 827-863. Residues 453-535 were poorly resolved so were built using density guide homology modeling and, with the exception of domains C&D, other residues were assigned using a combination of Rosetta *de novo* fragment docking and manual modeling in Coot.

Domains C & D of the S trimer are of importance here as these domains were particularly difficult to model. After initial fragment docking attempts failed a sequence scan was done attempting to only dock one subsection of the missing segment at a time. Eventually a 30 residue fragment was able to be identified in the map with good fit to the density (Figure 2.6.a). In an effort to complete domains C & D, and assign the roughly 150 residues that were resolved, this fragment was used as an anchor for manual model building and RosettaES in parallel. The two models agreed well for the N-terminus however the C-terminal residues posed greater challenge. The RosettaES solutions for the C terminus were well converged for the first 129 residues so the job was split up into two sections where the first 129 residues were added to the structure and used as input for the next round. The remaining residues were built using that model as input adding 25 additional residues to the structure. Comparison of the two models showed strong agreement in the ~30 residues corresponding to the N-terminal segment however significant topological differences were present in the c-terminal region (Figure 2.6.b).

Upon closer examination it became clear that the RosettaES model was correct whereas the manually traced model had significant errors. A number of key features were used to make

this determination including the placement of 6 cysteine residues into the density directly next to each other such that disulfide bonds could be formed that appeared to be supported by the density. These disulfides were placed by RosettaES despite no prior knowledge of their existence or forcefield attempting to form them (Figure 2.6.e). In addition another important features of the structures was identified, the protrusion of density coming from the side chain of ASP 657. ASP 657 is a putative glycosylation site and this density suggests a glycan is present at that position (Figure 2.6.d).

This structure was published in Nature in 2016[34] and in the same issue a homologous structure of the beta coronavirus HKU1 was also published at 4.0 Å[35]. A complete model of HKU1 was not generated over domain D however a partial model was made. This model was able to place the aforementioned disulfide residues which match very closely to those generated by RosettaES, further confirming the accuracy of the RosettaES model (Figure 2.6.f).

2.9.2 Modeling Human Coronavirus NL63

As discussed in the previous section beta coronaviruses represent a significant and growing health concern. Alpha coronaviruses, such as human coronavirus NL63 (HCoV-NL63) also pose a significant risk to human health. HCoV-NL63, first isolated in a 7 month old infant[36] is a common infection during childhood and many adults contain antibodies to the virus[37]. Although common HCoV-NL63 infection is a major cause of pneumonia and bronchitis in infants and can cause severe lower-respiratory-tract infections in immunocompromised patients, the elderly, and children.

In order to identify convergences between the coronaviruses of different genera our collaborators prepared and imaged frozen HCoV-NL63 S ectodomain N-terminally fused to a

GCN4 trimerization motif. Images were obtained on an FEI Titan Krios and a reconstruction was generated to 3.4 Å resolution.

To build a model from this reconstruction we first generated a homology model using the MHV structure described above as the template. The crystal structure of the HCoV-NL63 B domain was docked into the density after which several well resolved regions of the map still did not have a structural assignment. Including a N terminal domain, thought to be the product of a gene duplication event, not present in beta coronaviruses, and a region in the C-terminal domain that was not resolved in the previous MHV reconstruction (Figure 2.6.g).

We modeled the novel N terminal domain through a process of *de novo* fragment docking, RosettaES, and manual model building. The C-terminal domain was first modelled manually however the last ~30 residues fit the density poorly and contained significant clashes in the manually traced model. To resolve this we deleted these residues and rebuilt them using RosettaES (Figure 2.6.f). RosettaES rapidly converged on a solution with several superior features to the manually traced model. In addition to being free from any major clashes the RosettaES model was also able to place the single large hydrophobic residue, tyr 1227, into a clear pocket of density (Figure 2.6.j). However even more strikingly the RosettaES model placed ASP 1201 and ASP 1218 into the map in such a way that a large pocket of density was protruding from the side chain. These pockets correspond to the glycans attached at these sites, both of which were later modeled into the structure and confirmed with mass spec (Figure 2.6.i). Once again showing that RosettaES is able to generate solutions to modeling problems which are very difficult for even expert microscopists.

2.9.3 Modeling The Bamboo Mosaic Virus

Flexible filamentous plant viruses are single stranded positive sense RNA viruses and are responsible for a large amount of crop damage worldwide.[38] Despite their agricultural harms they are relatively non-toxic to humans and have shown much promise for use in biotechnology, particularly for vaccine production,[39] and protein expression.[40] However, development of these applications has been slowed due to the lack of an atomic model for these viruses despite structural studies that go back at least as far as 1941.[41] This lack of model is due to the fact that these viruses cannot be crystallized and are too flexible to obtain x-ray fiber diffraction at high resolution. On the contrary the first known virus, TMV,[42] is a rigid filamentous virus and has been solved by both x-ray diffraction[43] and cryoEM. TMV contains a right handed pitch about the helical axis and low resolution fiber diffraction data suggested that potexviruses, of which BaMV is a member, share a similar topology with each other. Several low resolution models of potexviruses have been generated[38,44,45] all of which implicitly assume a right handed helical pitch as in TMV.

BaMV is a single stranded RNA virus with genome of about 6.4 Kb[45] and a flexible morphology. It is built mainly from a single coat protein of which the N-terminus is of particular interest due to the fact that up to 35 residues of it can be removed and the virus can maintain replication and assembly.[45–47] These 35 residues can be replaced with foreign peptides in order to serve as a plant expression vector.[48,49]

To solve a structure of a flexible filamentous virus our collaborators in the Egelman lab imaged the wild type BaMV and a virion with a 35 residue deletion on the N-terminus. Helical reconstruction was done and symmetry was determined via trial and error until secondary

structure elements could be observed. Both reconstructions were identical with the respect to symmetry. Two enantiomorphic constructions could be generated, that were consistent with the images, for both the right handed and left handed helical twist. However, when the crystal structure fragment of the 28% homologous papaya mosaic virus was docked into the density it fit only into the left handed reconstruction. Comparing the two reconstructions allowed for a resolution estimate of 5.6Å.

To build a complete model of the BaMV complex we started by building a homology model using the papa mosaic virus crystal structure. This homology model, while fitting the density well, was missing 62 residues of the N-terminus and 42 residues of the C-terminus (Figure 2.6.k). Through manual segmentation we were able to isolate the asymmetric unit of the structure and we attempted to assign these missing residues using RosettaCM. After several iterations we were able to obtain a model of the N-terminus that agreed well with the density. However, RosettaCM was unable to assign a reasonable conformation that fit the density to the C-terminus and the top scoring models were significantly outside the expected region.

To resolve this issue we turned to RosettaES which was able to rapidly generate a set of well converged solutions (Figure 2.6.m). These conformations well explained the density and of particular importance, did not leave significant regions of the map unexplained. The end of the c-terminus was clearly visible in the map and the last residue in the sequence was placed directly into this region. Furthermore when the structure was refined in symmetry the model fit together well with favorable intra subunit contacts being made in the c-terminus. These features made us confident in the models produced by RosettaES despite the lower resolution of the electron density.[50]

2.10 Discussion of RosettaES

Here we have shown that RosettaES is a powerful tool for protein structure determination from sparse cryo electron miscroscopy data. It should expand that range of structures that can be determined automatically to those which include long segments with non-standard secondary structure, extended loops, and sheets. RosettaES can generate accurate atomic models for unassigned segments up to 50 residues and can occasionally generate accurate solutions on missing segments well over 100 residues.

The superior performance of RosettaES stems from it's more refined use of the experimental density to guide sampling. The short, three residue, fragments used to build up the RosettaES models allow for complete coverage of the necessary conformational space and, when combined with the experimental data, allow for much more target sampling of possible solutions. Furthermore the additional features in RosettaES including, the penalty on discontinuous density, two-tiered filtering strategy, beta sheet sampling, and efficient use of side chain information, improve its ability to exclude incorrect conformations increasing accuracy and reducing wasted sampling time.

Failure of RosettaES is almost always the result of insufficient sampling and this problem generally represents itself as the inability to assemble a set of fully assemble solutions. When such failure occurs taboo sampling can be used to generate additional conformations for evaluation. Furthermore diversity of the final solution set can often be an indication of sampling success and the lack of convergence suggests a greater need for sampling.

Ultimately RosettaES should be a valuable tool in aiding microscopists in building challenging structures at near atomic resolution data and it, and other tools like it, will likely play an important role in the revolution in structural biologically happening with cryoEM.

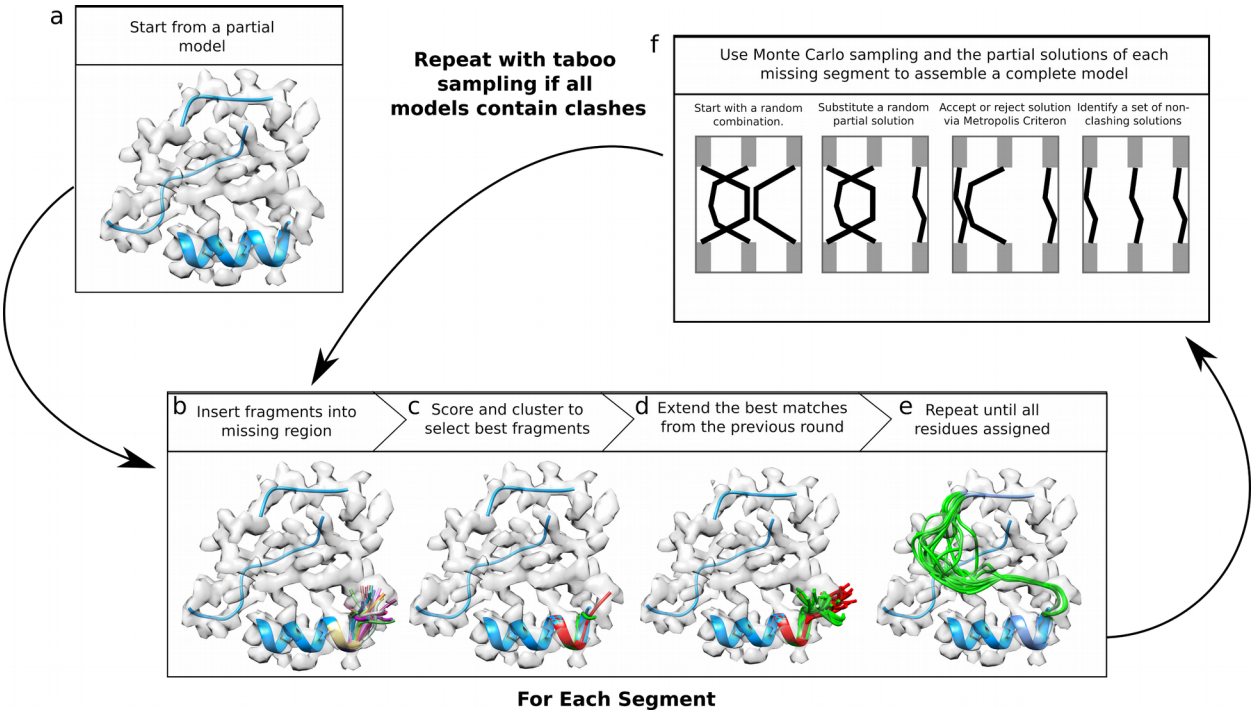

Figure 2.1 Overview of RosettaES[1]

a. RosettaES begins with an incomplete model placed into a density map. **b**. Short fragments are used to grow an ensemble of partial solutions. **c**. The ensemble is scored and pruned to a non-redundant subset consistent with the data. **d**. The remaining structures in the ensemble are again grown with short fragments. **e**. Steps **b** through **d** are repeated until all residues are assigned, yielding an ensemble of solutions consistent with the data. **f**. Given ensembles for each of several unassigned segments, a Monte Carlo assembly algorithm finds a set of non-clashing solutions. If no such solution is found steps **b** through **e** are repeated, enriching for solutions different than previous samples.

1 Figure republished with permission from doi:10.1038/nmeth.4340

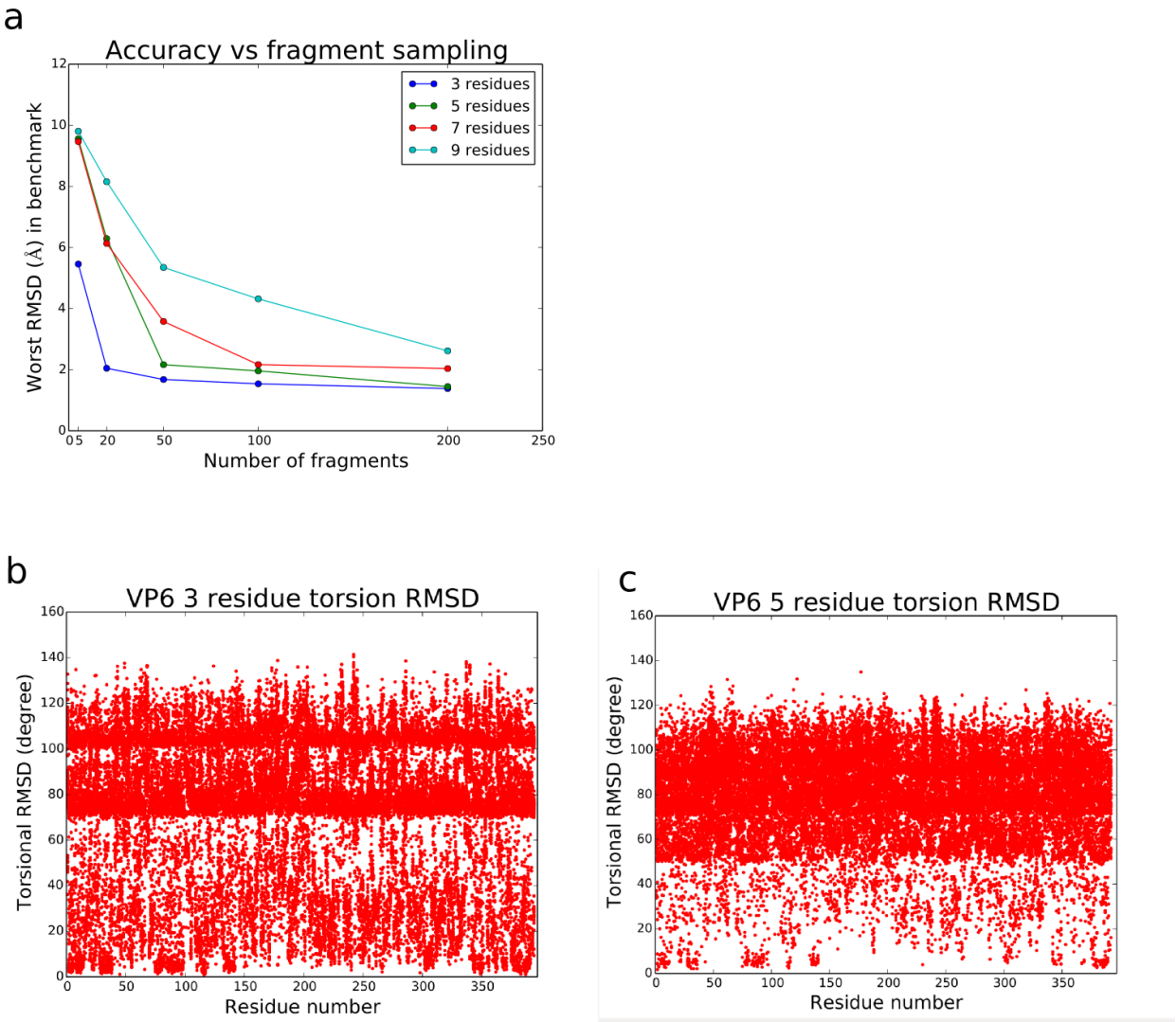

Figure 2.2. Number and size of fragments required to accurately sample all the missing segments in benchmark set from the Rosetta *de novo* models.[2]

a. A scan of different residue lengths and the number of fragments sampled at each step. The RMSD of the worst conformation generated for the missing segments of the round 1 models generated by Rosetta de novo fragment docking on our benchmark is plotted on the y-axis. Twenty, three residue long, fragments (using a two residue overlap for an addition of one residue per step) are sufficient to accurately sample every missing segment in our set. **b & c.** The torsional RMSD of the 100 3 and 5 residue long fragments for the benchmark case VP6. The 3 residue long fragments have much greater coverage near the deposited structure which may explain why near deposited conformations are more easily detected when using them.

2 Figure republished with permission from doi:10.1038/nmeth.4340

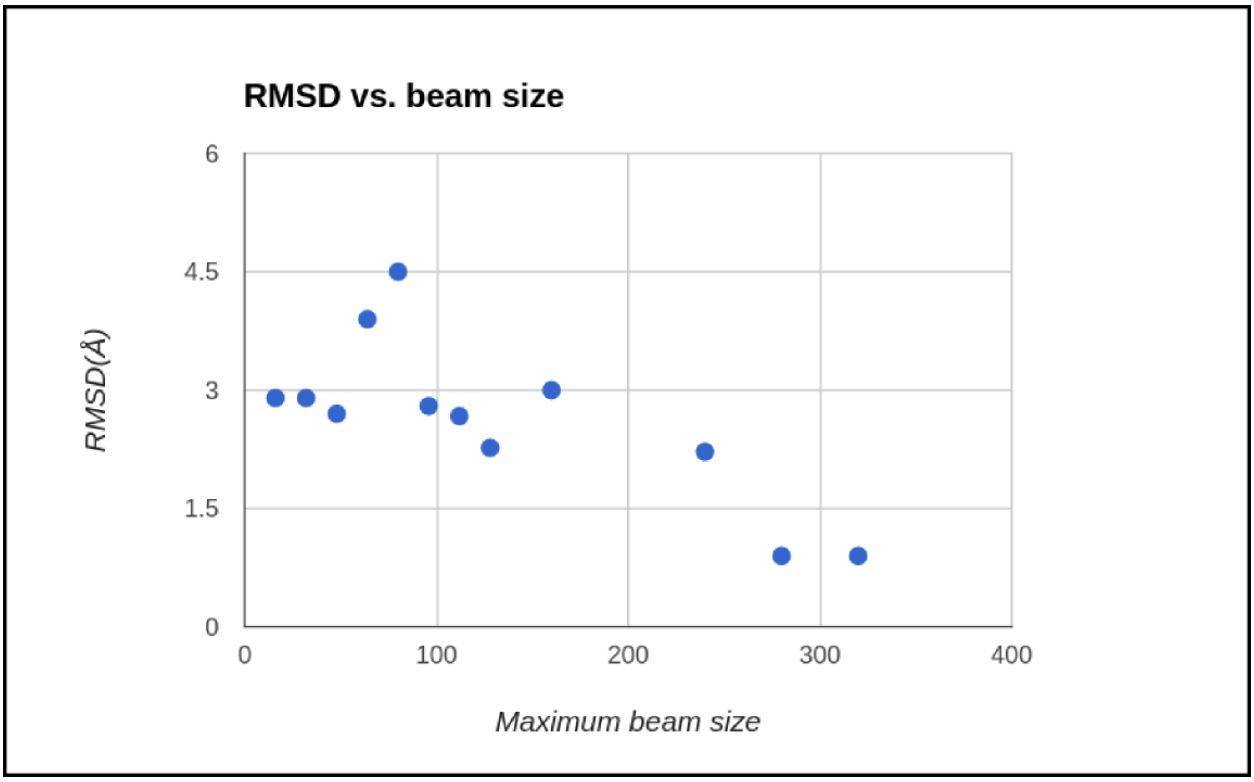

Figure 2.3. Effect of beam size on sampling accuracy
A plot of the best sampled structure of residues 1-65 of FrhG with increasing beam size. When the maximum beam size is set to 280 a 0.9 Å backbone and Cβ RMSD structure is sampled.[3]

3 Figure republished with permission from doi:10.1038/nmeth.4340

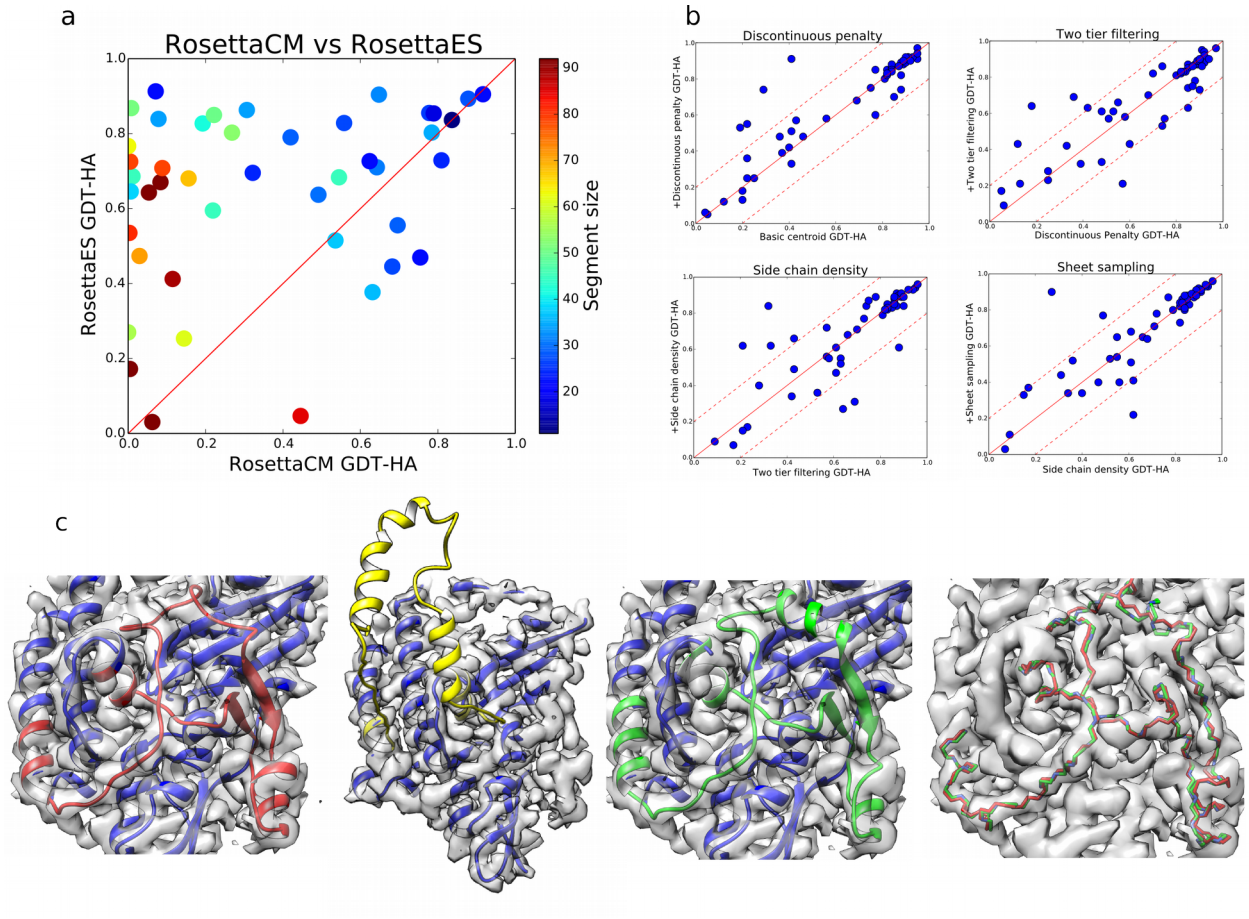

Figure 2.4 Accuracy of RosettaES compared to RosettaCM.[4]

a. A comparison of RosettaES to RosettaCM reporting GDT-HA over the backbone + Cβ atoms on a benchmark set of single missing segments extracted from the deposited model. The x and y axis correspond to the GDT-HA of the model compared to the deposited structures under two conditions. The closer to 1 the more similar the structures. Values above the solid line are improved with RosettaES. **b.** The GDT-HAs over the backbone + Cβ atoms of all the atomic models in the Rosetta *de novo* benchmark set as new features are added. **c**. The deposited structure for FrhA shown in the density. Residues 187-265 (highlighted in red) were removed in the benchmark. The second panel shows how RosettaCM failed to find a solution that fits the density. The top scoring solution is shown in yellow. The last panel shows that RosettaES identified a solution (rendered in green) that fits well into the density. The last panel shows a minimal backbone trace of the deposited model (in red) compared to the one produced by RosettaES. The two have a 1.9Å backbone and Cβ RMSD.

4 Figure republished with permission from doi:10.1038/nmeth.4340

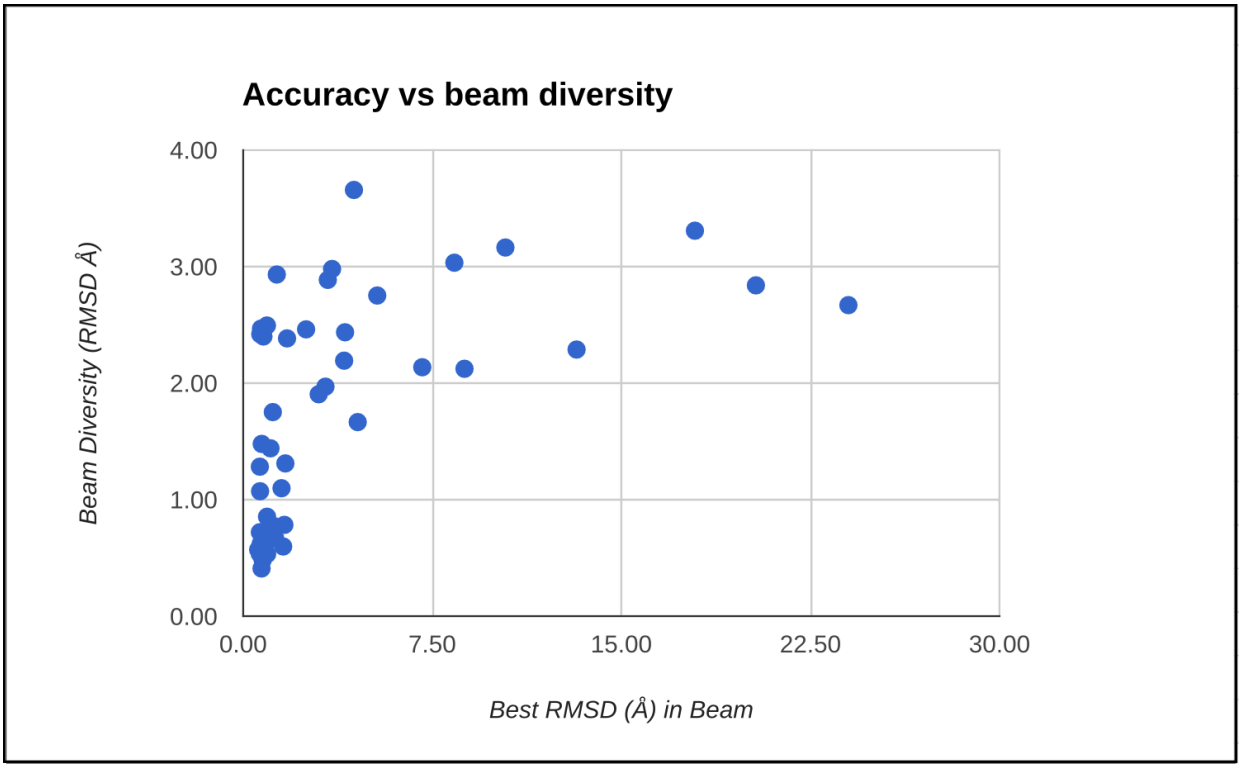

Figure 2.5. The accuracy of the best model in the solution set vs the solution set diversity.[5] A plot of the best backbone and Cβ RMSD to the deposited structure contained in the beam vs the beam diversity, calculated as the RMSD of each residue in every beam compared to the center of mass for that residue.

5 Figure republished with permission from doi:10.1038/nmeth.4340

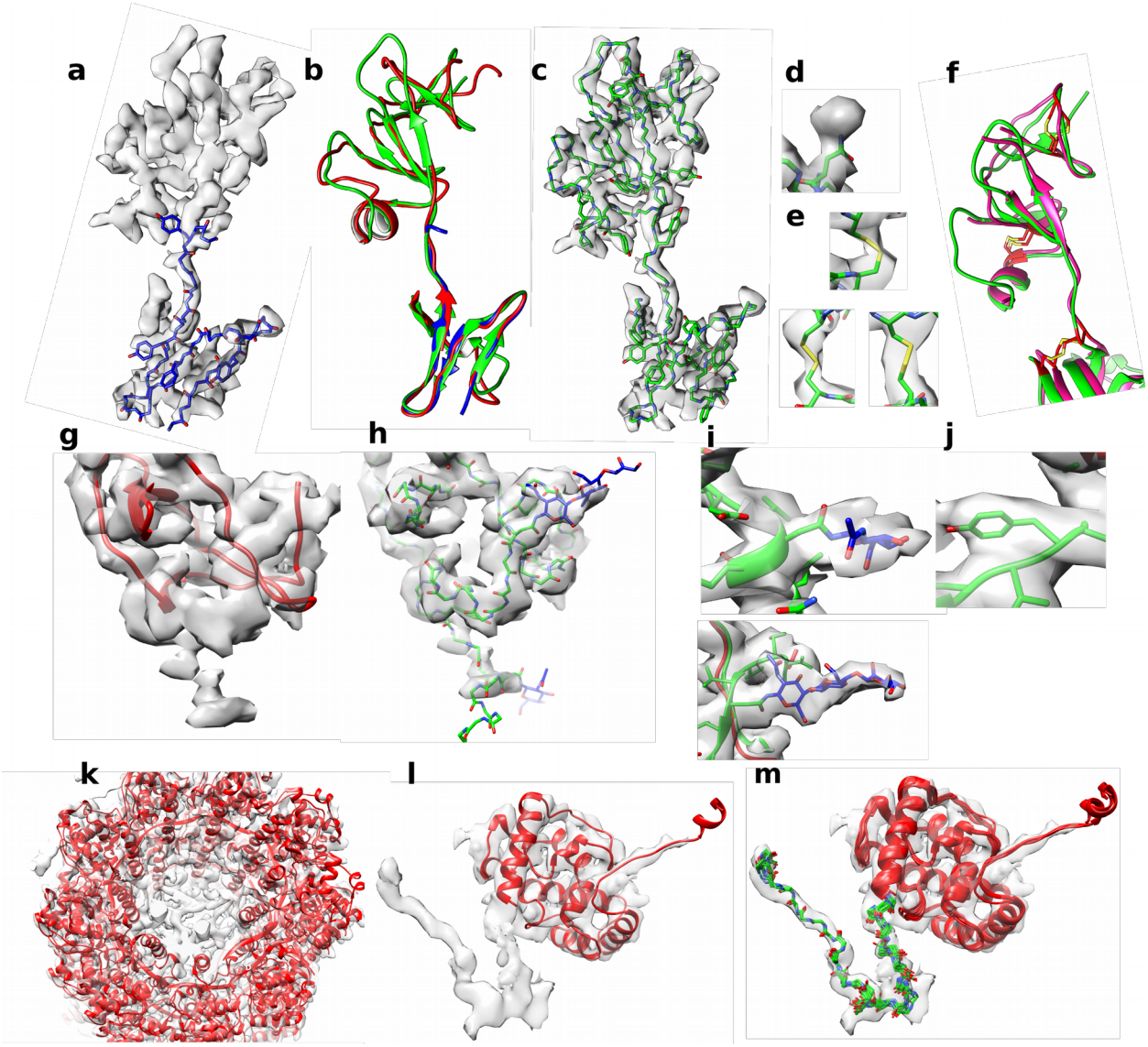

Figure 2.6 RosettaES enables structure determination in challenging cases.[6]
a-f. Building domains C and D of the MHV coronavirus spike with RosettaES. **a**. A 30 residue segment of MHV domain C was placed by Rosetta *de novo* fragment docking. **b.** The model was completed by RosettaES. A hand traced model (red) and the deposited structure generated by our method (green) disagree significantly in domain D, resulting in completely different topologies. **c.** The RosettaES generated model shown in the density. Large aromatic residues are shown as sticks. **d.** Our model is validated by a tube of density at the putatively glycosylated ASN 657 side chain. **e.** The positioning of cysteines to form 3 unique disulfide bonds also suggest correct registration. **f.** A recently determined structure of HKU1 spike protein (magenta) matches the topology obtained by our method (green; cysteines are highlighted in red). **g.** Density for the C-terminal tail of HCoV-NL63 is visible in the map. Docked in red is the partial model built using our structure of MHV. **h.** the final structure, after completion with RosettaES, attachment of

6 Figure republished with permission from doi:10.1038/nmeth.4340

glycans (shown in blue) and refinement. **i.** The model is supported by the placement of Asp 1201 and Asp 1218, positioned with clear density corresponding to glycans, and reasonable sidechain density. **j.** Tyrosine 1227 also fits the density well. **k.** A symmetric homology model of the Papino Mosaic Virus docked docked into the reconstruction of the Bamboo Mosaic Virus. The C-termini in the core are missing from the model. **l.** A closeup view of the asymmetric unit, with the homologous structure shown in red. **l.** Despite the low resolution, the top-scoring models produced by RosettaES, shown in green, are tightly converged and fit the density well.

				RosettaES RMSD (Å)		Rosetta CM RMSD (Å)	
Model ID	Secondary Structure	Residue Range	Total Residues	Best Scoring	Best Sampled	Best Scoring	Best Sampled
BPP1	L&β	256-283	28	0.7	0.6	0.7	0.7
BPP1	L	283-327	45	2.0	1.1	30.5	13.4
BPP1	L&β	207-259	53	0.9	0.7	4.2	4.2
BPP1	L&β	119-207	89	3.4	3	30.8	11.2
BPP1	L&β	1-111	111	2.5	1.7	48.6	37.3
FrhA	L&β	1-22	22	0.5	0.5	6.4	6.4
FrhA	β&α	337-358	22	0.5	0.6	0.5	0.5
FrhA	L&β	24-49	26	0.6	0.5	0.9	0.9
FrhA	L	136-165	30	2.1	1	3.2	3.2
FrhA	L&β	298-339	42	0.8	0.7	10.6	10.6
FrhA	L&β&α	187-265	79	1.9	1.6	11.6	8.1
FrhB	L&β	1-18	18	0.7	0.6	1.4	1.4
FrhB	L&β	87-108	22	1.0	0.6	0.8	0.8
FrhB	L&α	61-87	27	0.8	0.5	2.0	2.0
FrhB	L&β	36-66	31	1.0	0.6	1.5	1.5
FrhB	L&β	110-143	34	0.7	0.6	3.7	3.7
FrhB	L&β&α	179-228	50	0.7	0.7	4.1	4.1
FrhG	L&α	144-172	29	0.9	0.7	1.7	1.7
FrhG	L&β&α	192-228	37	3.0	1.1	1.6	1.6
FrhG	L&β	73-133	61	6.6	4	5.9	5.9
FrhG	L&β&α	1-71	71	3.3	2	28.6	15.1
STIV	L&β	252-319	68	1.8	1.6	11.2	11.2
STIV	L&β&α	161-252	92	2.3	1.3	15.9	15.0
STIV	L&β&α	1-163	163	16.9	14.8	45.9	35.4
T20S	L&β	43-78	36	4.3	2.1	2.1	2.1
T20S	L&β&α	13-50	38	1.6	1.1	14.6	10.2
T20S	L&β&α	159-221	63	1.6	1.4	37.3	27.8
T20S	L&β&α	88-166	79	1.1	1	5.9	5.9
TMV	L&α	1-11	11	0.7	0.6	0.6	0.6
TMV	L	78-106	29	3.2	1.2	1.5	1.5
TMV	L	44-78	35	0.8	0.7	0.8	0.8
TRPV1	L&α	205-226	22	3.4	0.8	0.7	0.7
TRPV1	L&α	111-134	24	1.6	1.1	5.6	4.9
TRPV1	L&α	1-45	45	6.1	4.1	3.0	3.0
TRPV1	L&α	66-110	45	1.4	1.4	1.7	1.6
TRPV1	α	128-183	56	4.0	3.7	4.8	4.8
TRPV1	L&α	226-310	85	7.0	6.5	3.5	3.5
VP6	L&α	349-372	24	1.1	0.8	1.0	1.0
VP6	L&β	243-269	27	2.9	1.9	2.0	2.0
VP6	L&α	87-118	32	0.6	0.7	1.6	1.5
VP6	L&β	266-299	34	0.7	0.9	8.5	8.2
VP6	L&β&α	300-350	51	0.6	0.8	3.6	3.6
VP6	L&β&α	1-81	81	2.6	1.9	43.3	19.1
VP6	L&β&α	115-245	131	4.2	3.8	38.5	21.0

Table 2.1. Results of RosettaCM and RosettaES for each missing segment in the benchmark set with the context of the deposited model.[7]

A table containing a list of the target structures and the missing residues built by both RosettaCM and RosettaES. The backbone + Cβ RMSD in Angstroms of the best scoring and best sampled conformation. The RosettaES is able to sample 38 of 44 segments to 2.0 RMSD or lower whereas RosettaCM is able to sample only 17 to 2.0 RMSD or lower.

7 Table published with permission from doi:10.1038/nmeth.4340

Model ID	Residue Range	Percent Assigned Sub 2.0Å			
		Modeller*	Phenix Autobuild	Buccaneer	RosettaES
BPP1	207-259	9.4	24.5	66	100
FrhA	187-265	1.3	0	61	77
FrhG	1-71	2.8	0	58	58
STIV	252-319	2.9	0	9	75
T20S	43-78	5.5	0	0	44

Table 2.2 A comparison of RosettaES to non-Rosetta modeling software

This table shows the results of running three different model building software packages compared to RosettaES on 5 targets from our benchmark set. RosettaES was run for a single round with a maximum ensemble size of 64 all other protocols were run with default settings. RosettaES outperformed every method across all entries with the exception of FrhG residues 1-71 for which both Buccaneer and RosettaES assigned 58%. However increasing the ensemble size for this set allows RosettaES to find an accurate model for this solution (Figure 2.3).

Model	LoopID	Residues	Round 1 RMSD	Round 2 RMSD	Round 3 RMSD	Round 4 RMSD	Full Model Cα RMSD
BPP1	1	105	11.5	2.7	2.2	2.1	-
BPP1	2	76	7.4	7.7	7.0	7.6	-
BPP1	3	39	0.8	2.2	1.0	1.4	-
BPP1	4	15	0.8	0.8	0.8	1.3	-
BPP1	5	39	6.3	2.9	3.3	4.5	-
FrhA	1	17	0.8	-	-	-	1.4
FrhA	2	13	0.7	-	-	-	1.4
FrhA	3	8	0.6	-	-	-	1.4
FrhA	4	17	0.8	-	-	-	1.4
FrhA	5	66	1.5	-	-	-	1.4
FrhA	6	29	1.5	-	-	-	1.4
FrhA	7	9	0.7	-	-	-	1.4
FrhA	8	10	2.4	-	-	-	1.4
FrhB	1	13	1.7	-	-	-	1.1
FrhB	2	18	0.7	-	-	-	1.1
FrhB	3	14	0.7	-	-	-	1.1
FrhB	4	9	0.8	-	-	-	1.1
FrhB	5	21	0.9	-	-	-	1.1
FrhB	6	37	1.8	-	-	-	1.1
FrhB	7	4	0.9	-	-	-	1.1
FrhG	1	65	4.2	-	-	-	2.0
FrhG	2	48	2.5	-	-	-	2.0
FrhG	3	16	1.7	-	-	-	2.0
FrhG	4	31	1.5	-	-	-	2.0
STIV	1	158	19.2	18.8	21.7	21.4	-
STIV	2	79	9.0	2.1	1.2	1.2	-
STIV	3	55	8.2	3.5	1.6	1.6	-
STIV	4	5	0.8	0.7	0.7	0.7	-
T20S	1	4	1.3	1.2	1.2	1.2	-
T20S	2	25	3.3	2.1	1.6	1.6	-
T20S	3	23	3.0	2.3	1.3	1.8	-
T20S	4	66	2.2	1.4	1.4	1.4	-
T20S	5	57	21.9	21.6	21.8	23.2	-
TMV	1	5	1.0	-	-	-	1.3
TMV	2	8	1.0	-	-	-	1.3
TMV	3	22	0.7	-	-	-	1.3
TMV	4	16	0.9	-	-	-	1.3
TMV	5	8	1.0	-	-	-	1.3
TRPV1	1	39	5.7	-	-	-	3.6
TRPV1	2	32	1.2	-	-	-	3.6
TRPV1	3	11	1.5	-	-	-	3.6
TRPV1	4	43	1.9	-	-	-	3.6
TRPV1	5	9	0.8	-	-	-	3.6
TRPV1	6	79	4.5	-	-	-	3.6
VP6	1	75	4.2	2.3	2.9	2.3	-
VP6	2	19	0.8	0.7	0.9	0.9	-
VP6	3	118	15.5	14.5	6.9	14.5	-
VP6	4	14	1.0	0.9	0.9	0.9	-
VP6	5	21	1.0	0.9	0.9	0.9	-
VP6	6	38	3.8	1.1	1.0	1.0	-
VP6	7	11	1.1	0.9	0.8	0.9	-
VP6	8	5	0.9	0.8	0.8	0.8	-

Table 2.3. Results of RosettaES for the missing segments of the benchmark set that does not have the context of the deposited model.[8]

A table of results for each of the individual missing segments from the round1 fragment docking. Each round is the best sampled backbone and Cβ RMSD reported in Angstroms. Previous round results are used for taboo sampling in subsequent rounds. The last column reports the complete Cα RMSD of the best scoring model after MCA and local refinement.

8 Figure republished with permission from doi:10.1038/nmeth.4340

Chapter 3: GLYCAN REFINEMENT

3.1 Introduction to Modeling Glycoprotein Conjugates

Carbohydrates are some of the most structurally diverse biomolecules in all of nature. They serve a huge range of functions from cell signalling, to energy storage, to viral immune evasion. Glycoprotein conjugates are an important class of these molecules with studies suggesting that more than half of all human proteins are glycoprotein conjugates.[51] Despite the frequency at which protein glycosylation occurs the field of glycan structural biology has faced many challenges.

Due to the highly flexible nature of glycans glycoprotein conjugates have been historically difficult to characterize. Because glycan flexibility is thought to impede crystallization specific removal of glycans through genetic engineering in order to obtain crystal structures is not uncommon[52] as is removal through various glycosidase enzymes. Despite this as of late 2017 over 4000 entries in the protein data bank (PDB) contain N-linked carbohydrates, approximately 5% of the total database, up from ~2.5% in the early 2000s.[53] Furthermore, likely due to the lack of crystallization requirement, the advances in cryoEM have led to an increasing number of structures with glycans being deposited. This trend is expected to continue as many new researchers begin to take advantage of cryoEM technology.

Unfortunately structural refinement tools for work with glycans have not kept up with these technological advances.[53] This has caused significant problems throughout the field. In 2003 it was shown that roughly 30% of all glycan containing structures in the PDB contain errors in either their nomenclature or their chemistry.[54] Furthermore, in 2015 it was shown that a

number of other anomalies exist in the PDB. Including a phenomenon where high energy sugar conformations start to occur at alarming frequency as the resolution of the experimental data decreases.[55] Suggesting that, rather than modelers assigning high energy conformations only when there is strong support in the experimental data for their presence, they are being poorly fit into weak data and refinement programs are not resolving these mistakes. With cryoEM producing an increasing number of reconstructions precisely at these resolutions there is an urgent need for better tools for modeling glycans with experimental data of this quality.

3.2 Introduction To The Rosetta Glycan Framework

The Rosetta software suite is a powerful tool for protein structure determination and refinement. Because the origins of Rosetta are in purely computational structure prediction and design it is highly optimized for use without experimental data. It includes such features as an all atom physically realistic energy function[56] and a large number of ways to manipulate the conformation of a structure. In addition a number of novel methods that make use of experimental data, including the fragment docking and RosettaES protocols described in chapter 1 and 2 and the Phenix-Rosetta tool for crystal structure refinement,[57] have been developed. In addition to these tools a framework for glycan modeling and design has been implemented in Rosetta.[58] This framework includes torsional based glycan related score terms and a number of tools for glycan conformational manipulation.

It is therefore reasonable to suppose that Rosetta's realistic energy function, refinement tools, and glycan modeling framework could come together to form a powerful tool for modeling glycans with near atomic resolution data. However, a number of major limitations exist within the existing glycan framework which prevent that from being the case. These include the lack of

ability to write glycans in standard pdb format, the extreme syntax requirements of input PDB files (especially LINK records), a lack of cartesian based score terms. Here we will described the steps taken to address these issues as well as describe results that demonstrate the utility of this new Rosetta glycan refinement protocol for refinement with both cryoEM and x-ray crystallography data in the range of 1.5-5Å resolution.

3.3 Cartesian Scoring Of Glycans

Within Rosetta there are two common formats by which a structure can be manipulated, either by using torsional space or cartesian space. In torsional space bond lengths and angles are kept fixed and only atom torsions are allowed to move. This reduction in the degrees of freedom makes sampling and scoring faster and more computationally tractable while also allowing for large domain based motions to occur more easily. With the introduction of the original glycan framework a number of glycan related score terms were implemented for torsional based scoring however no modifications were made to allow for scoring in cartesian space leaving it insufficient to resolve many of the bond length and error issues common in refinement targets. Given the large number of the problems involving bond lengths and bond angles it is critical that cartesian space of the glycans work reliably in Rosetta. To accomplish this goal a number of modifications were made to the cartesian energy function.

3.3.1 Scoring of Glycan Rings in Cartesian Space

In order to ensure glycan rings were constrained to known conformations functions were added to cartesian scoring which match the current positions to the closest conformer, via nu angles. Constraints are then placed on the ring torsions to drive the ring towards this

conformation. To handle the weights on these constraints a new score type, cart_bonded_ring, was added to the cartesian score function.

While this design of constraining the rings to the closest conformation of the input structure offers the flexibility to model high energy states of the rings when dealing with sparse experimental data one should never build anything but the most favorable ring conformations as the experimental data cannot support modeling the glycan in a high energy state. To force this behavior a flag was added, -ideal_sugars, forcing Rosetta to always assign constraints only to the low energy conformation for each pyranose moiety. These constraints are then applied during cartesian minimization to resolve ring conformation anomalies.

3.3.2 Cartesian Scoring of Anomers

In order to resolve issues with discrepancies between the nomenclature and the geometry of a structure (alpha sugars in beta positions and vice versa) additional torsional constraints were added to the cart bonded ring score term. These torsional constraints occur between the atom of the glycosidic bond from the lower residue and the C1, C5, and C6 atoms of the sugar. However, while these torsional constraints are sufficient to drive the heavy atoms into the correct positions in order for the hydrogens to be placed correctly the chirality at the anomeric position must change. To resolve this issue a new mover was implemented in Rosetta which rebuilds the anomeric hydrogen based on the ideal coordinates of the residue. This mover is added to the Rosetta xml script after a minimization or relax cycle and is sufficient to resolve these nomenclature/geometry discrepancies. In total these changes to cartesian scoring are sufficient to resolve most of the common structural flaws of glycoprotein conjugates.

3.4 Reading and Writing Structures with Glycans

As part of the original glycan framework in Rosetta a custom nomenclature system was developed for Rosetta glycans. While Rosetta can read both PDB and glycam formatted structure files, internally a custom naming scheme was developed to represent all the possibilities of glycan structural biologically and avoid the limitation of unambiguous structures from the pdb names. While this naming scheme is useful for its intended use of *de novo* glycan design within Rosetta it is poorly suited for dealing with refinement problems which frequently involve the use of multiple software packages whether they be molecular graphics or other refinement tools. This especially critical for pipelines such as the Rosetta-Phenix crystal refinement protocol in which tight integration between Rosetta and Phenix at all steps of the protocol is required

Reading of PDB (and glycam) formatted files and converting them into Rosetta nomenclature is done by taking advantage of an internal database that matches three letter pdb codes to the Rosetta base name for the glycan, however this database does not match the base name in a one to one ratio instead multiple pdb codes can correspond to the same Rosetta residue base name. This problem is resolved by the Rosetta residue matcher which applies patches to the base residue type depending on the atoms in the residue.

Because Rosetta does not maintain the original pdb code for the sugar, and to ensure naming PDB naming could be used regardless of the input source, it is therefore necessary to map the full Rosetta name (including patch names) back to the original pdb codes if Rosetta is to be capable of output in PDB format. To resolve this problem we added additional data to the Rosetta pdb naming database which now stores the name of each patch added to the base residue type in order to create the desired residue. Additional functionality was also added to the Rosetta nomenclature manager which will attempt to match base names and patch names to the proper

pdb three letter code. Using this system for naming is particularly valuable because it means Rosetta is able to take files in any format and convert them into the pdb standard ready to be used for deposition in the database. It also resolves the issues of software compatibility given that most refinement packages that handle glycans accept them in standard pdb format.

3.4.1 Connectivity and LINK Records

Another major limitation of working with glycans in Rosetta is the extreme syntax requirements for the input pdb files. Two areas that cause particular issue are the requirements that all glycans be numbered in such a way that the numbering completes each chain before the next begins. This is not a requirement in standard pdb format meaning that Rosetta is unable to load many of the valid structures stored in the database. The other major requirement is that all link records in the pdb file occur in such an order as to create a valid Rosetta fold tree. This imposes very tight restrictions on the order of which the LINK records can occur and results in a load failure whenever such requirements are not met. These two issues, combined with cryptic error messages, mean that if often requires an expert Rosetta developer spend hours manually manipulating the glycan containing pdb file into a state which can be read by Rosetta making it nearly unusable as a refinement tool.

Two recent additions to Rosetta have dramatically improved this problem. The first is a method for auto detection of glycan connections added by our collaborator Sebastian Rämisch. This code uses distances to map the connectivity of all the glycans in the structure and renumber them internally into acceptable format. This resolves the chain numbering issue allowing Rosetta to accept chains numbered in any format and instead relying on geometry to find the correct connectivity. Second a refactor of the connectivity code in Rosetta, performed by Rocco Moretti,

fixed the issues involving the LINK record formatting allowing for LINK records to be written in any pdb compatible format and treated appropriately by Rosetta.

We took advantage of both these changes to implement a combined system that uses explicit link records, when present, combined with auto detection code to find any links that may have been missed. This system is far more robust allowing Rosetta to load and treat most structures reasonably. Furthermore the use of explicit LINKs in addition to auto detection means modellers can now utilize the common strategy of rigid body fitting glycans from deposited models into their structure and, as long as the original glycan has reasonable geometry, they need only worry about explicitly declaring the connection to the side chain and Rosetta will handle the rest of the connectivity appropriately.

3.5 Refining Low Resolution Glycoprotein Crystal Structures

In order to test Rosetta glycan refinement with low resolution crystal data we collected a benchmark set 12 glycan containing crystal structures from the PDB. Resolutions of the data ranged from 1.9 to 3.5 Å and in total this set contained 133 sugar moieties. Through a combination privateer and manual analysis we were able to identify a number of issues in these structures. Including 3 problems with incorrect anomers and 23 problems with high energy ring conformations in one case, 5k65, one of the glycans also did not form the correct glycosidic bond with the side chain.

To refine these structures we used a modified version of the phenix-rosetta low resolution refinement strategy with several rounds of minimization starting from low repulsive weights and ramping to full weight across 4 cycles being performed before the normal protocol. This protocol was able to resolve all 3 of the incorrect anomers and all 25 of the energy ring conformations as

reported by privateer. Examples of these fixes can be seen for entries 1c1z, where the incorrect anomeric state for Mannose 397 has been corrected, and 5n92 where the high energy ring conformation of Fucose 597 is resolved (Figure 3.1 A-D). For the missing connection identified between the glycan and ASN 297 of entry 5k65 an explicit link record was added before refining bringing closed and forming the glycosidic bond which becomes visible in the density after rephasing (Figure 3.1.E-F). Furthermore these improvements to the geometry come at almost no cost in fit to density with almost all models, save one, showing improved R and Rfree after refinement (Table 3.1 Figure 3.2).

3.6 CryoEM Glycan Refinement

In addition to the refinement of crystal structures we also used Rosetta glycan refinement in the modeling process of two recent structures, the HCoV-NL63 structure described in chapter 1 and the HIV trimer (unpublished).

3.6.1 Modeling The Glycans of HCoV-NL63

To model the glycans of HCoV-NL63 sugar moieties were taken from the PDB and were docked into the density approximately near the the asparagine to which they were bound. Little care was taken to ensure correct glycan geometry with many glycans docked haphazardly creating glycosidic bonds often several angstroms apart, this was by design as instead Rosetta glycan refinement was used to resolve any of the issues resulting from poor docking. In total 31 sugar moieties were modeled for each monomer, providing valuable insight into the structure of the viral glycan shield (Figure 3.3).

Comparison of the starting model to the deposited one show several major improvements to glycan geometries. Several examples of these errors in the input model are ASN 1218 which fails to form a correct linkage to the NAG 1469 which also fits the density poorly (FIgure 3.4.A), mannose 1428 which is docked in a high energy boat conformation (Figure 3.4.B) and ASN 218 which contains significant clashes with the attached glycan and an extremely long glycosidic bond (Figure 3.4.C). All of these anomalies are resolved during Rosetta refinement resulting in a significantly improved structure demonstrating the power of this method to resolve even very poorly fit glycans.

3.6.2 Modeling The Glycans Of The HIV Trimer

The HIV trimer is one of the most heavily glycosylated proteins in nature and the glycans have drawn much attention for the role they play in HIV immune evasion. To aid in the understanding of these interactions our collaborators used cryoEM to generate a 3.4 Å resolution reconstruction of an antibody bound HIV trimer. In total 60 glycan moieties were resolved well enough that they could be modeled in the data. Similarly to HCoV-NL63 glycans were first docked into the density using coot and the glycans were then refined using Rosetta by first minimizing in cartesian space and idealizing anomeric hydrogens followed by local relax with the glycan framework enabled. Again significant improvements over the starting model were made, including resolving atomic clashes, incorrect anomers, and long bond distances (Figure 3.5).

3.7 Discussion Of Glycan Refinement

Here we have described a novel refinement strategy for work with glyco-protein conjugates with sparse cryoEM and x-ray crystallography data. This method is able to resolve many of the common errors found throughout the PDB, especially at lower resolutions including problems with incorrect anomeric forms, high energy ring formations, and unrealistic glycosidic bonds. This protocol comes at a time when, thanks to the advances in cryoEM, an increasing number of glycosylated structures at near atomic resolution are expected to be solved at near atomic resolution precisely where existing refinement protocols perform so poorly. We anticipate that going forward this tool will be a valuable resource for crystallographers and microscopists looking to build accurate models of glycoprotein conjugates and we hope it will begin to reduce the number of new structures being deposited with significant errors in their glycan conformations.

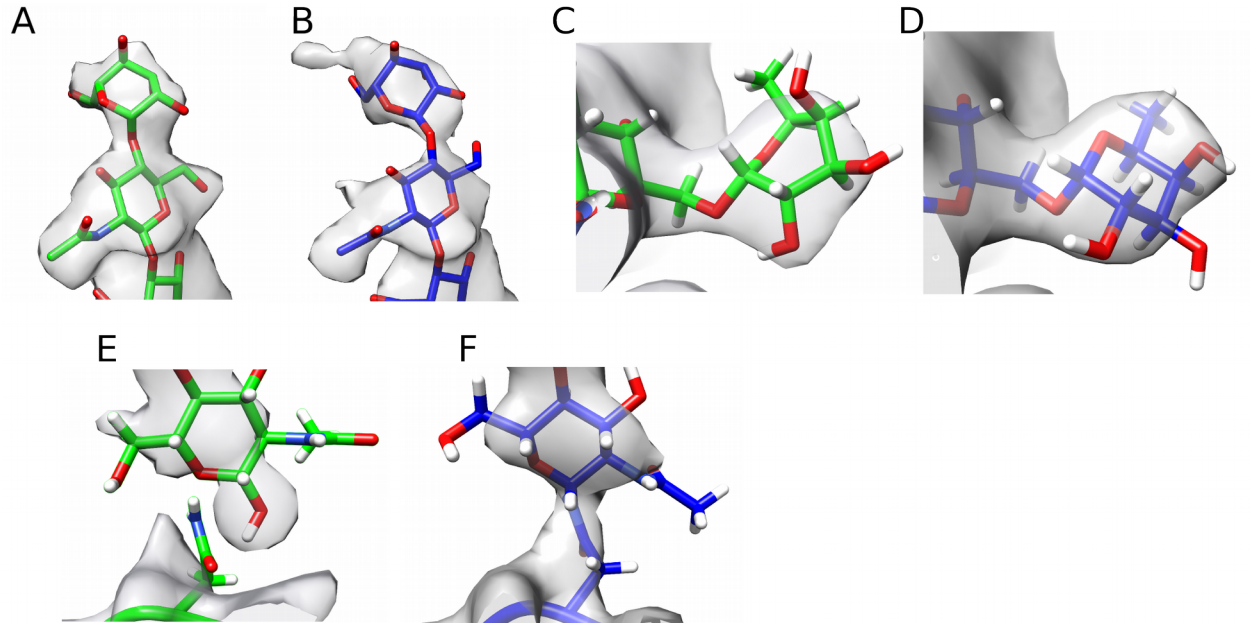

Figure 3.1 Anomalies in deposited crystal structures Resolved by Rosetta-Phenix Refinement

A. Mannose 337 from pdb entry 1c1z is in the incorrect anomeric conformation. **B.** Rosetta refinement has resolved the incorrect anomeric attachment assigning the correct alpha conformation to the sugar. **C.** Fucose 597 of entry 5n92 is in a high energy boat conformation. **D.** 5n92 fucose 597 has been moved to the low energy chair conformation by Rosetta glycan refinement. **E.** The N-acetylglucosamine residue 501 of chain B for entry 5k65 does not form the glycosidic bond to the ASN 297 of the same chain. **F.** After explicitly assigning the linkage Rosetta refinement is able to form the correct bond and density for the connection is visible upon rephasing.

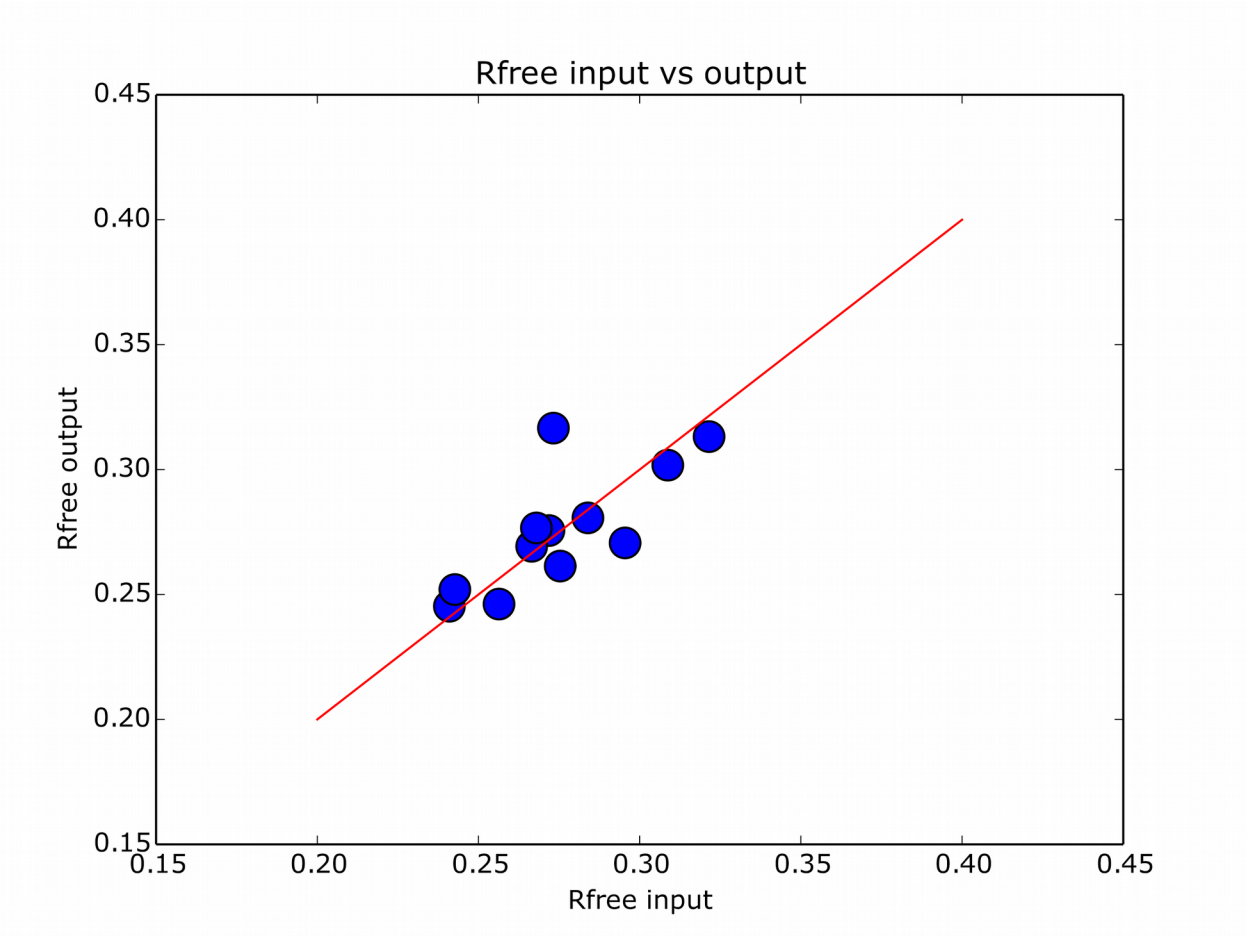

Figure 3.2 Rfree before and after glycan refinement

Figure 3.2 shows Rfree before and after gylcan refinement in Rosetta. With the exception of once case Rfree remains unchanged despite resolution of the glycan anomalies.

]

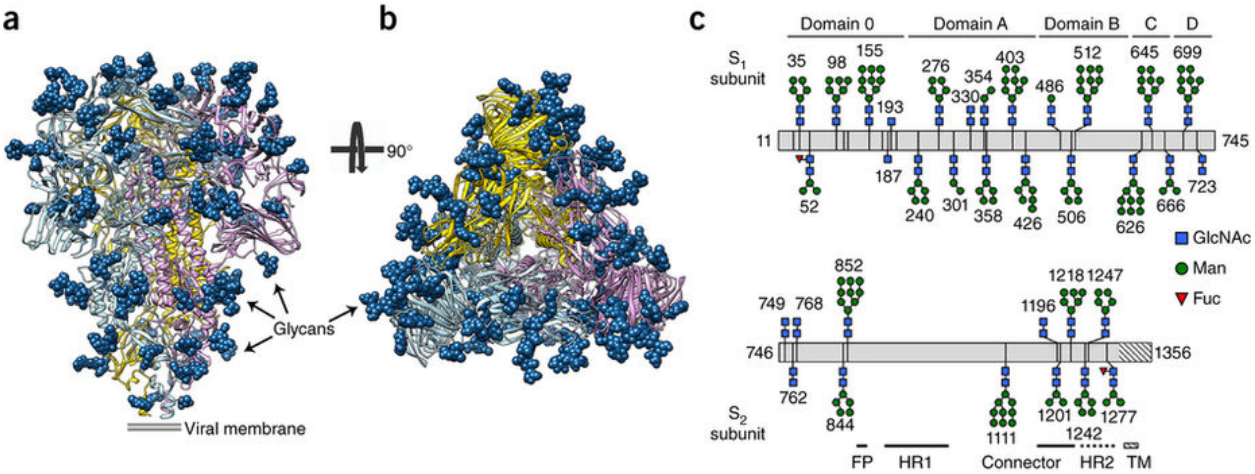

Figure 3.3 Architecture and glycans for HCoV-NL63[9]**(a,b)** Ribbon diagrams showing two orthogonal views of the S trimer, from the side (**a**) and from the top (**b**), facing toward the viral membrane. Glycans are shown as dark-blue spheres. (**c**) Residue-level schematic of N-linked glycans. The most extensive glycan structure detected by MS at each site is represented except for glycans observed only by cryo-EM, for which the resolved sugar moieties are shown. FP, fusion peptide, HR1, heptad-repeat 1 region; HR2, heptad-repeat 2 region (shown with a dashed line because it is not resolved in the map); TM, transmembrane domain (the striated texture indicates regions that are not part of the construct); GlcNac, *N*-acetylglucosamine; Man, mannose; Fuc, fucose.

9 Figure published with permission from doi:10.1038/nsmb.3293

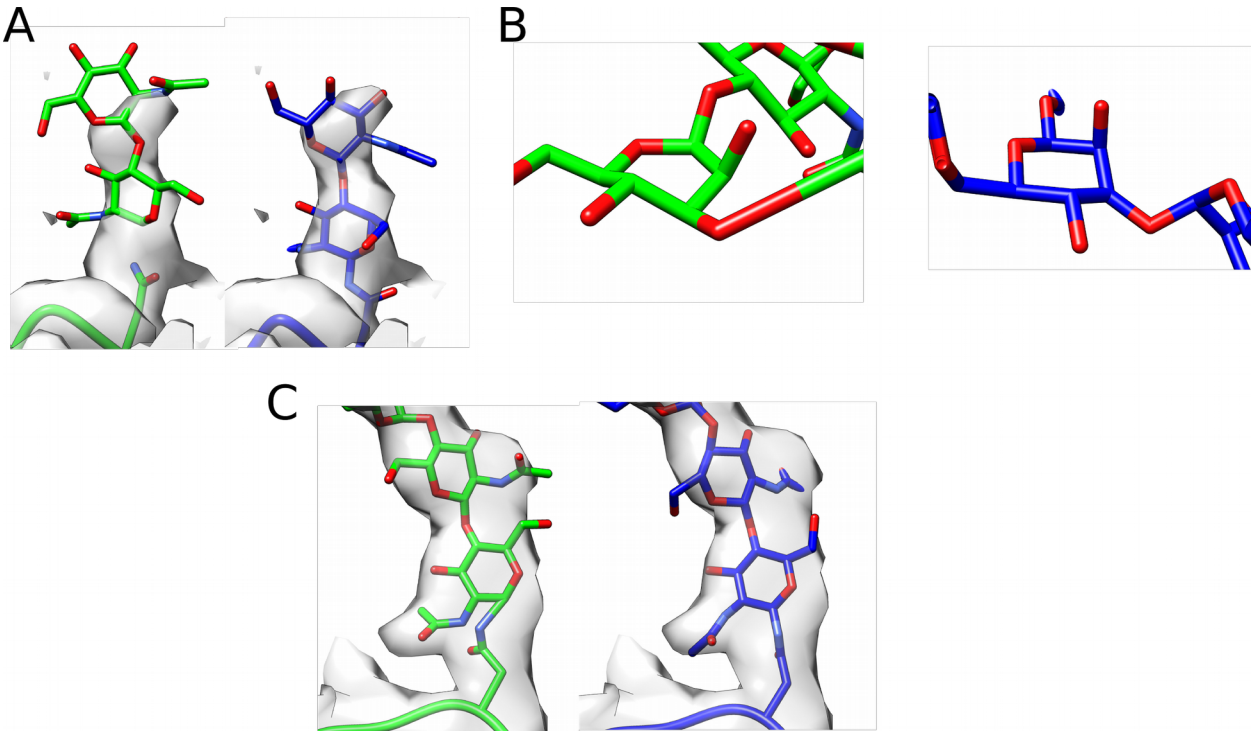

Figure 3.4 Improved glycan geometries for HCoV-NL63

A. The glycosidic linkage between ASN 1218 and NAG 1469 is not made in the input structure (green) and the input fits the density poorly. In the deposited model (blue) the glycosidic bond has been formed and the model fits the density well. **B.** Mannose 1428 in the input structure exists in a high energy boat conformation (green). This mannose is converted to a chair form by Rosetta refinement (blue). **C.** ASN 218 of the input model (green) has significant clashes with the attached glycan as well as a highly unrealistic glycosidic bond. These issues are resolved by Rosetta in the output model (blue).

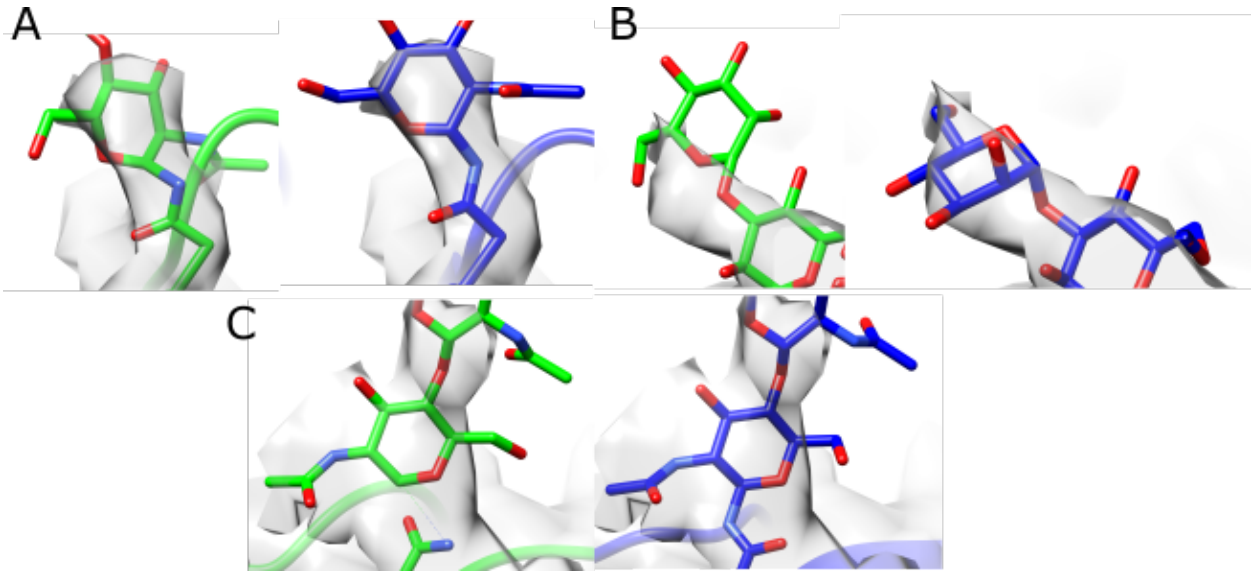

Figure 3.5 Glycans of the HIV trimer fixed with Rosetta glycan refinement.
A. Asp 241 has a stretched glycosidic bond and poor bond geometry in the input model, green, which is resolved in refinement, blue. **B.** Man 1200 has poor fit to density and bad glycosidic bond geometry in the input. These problems have been fixed by Rosetta glycan refinement, blue. **C.** The glycosidic bond distance of NAG 1301 is stretched and the nitrogen of the asp is oriented away from the glycan in the input. This is resolved in the output.

PDBID	Reported Resolution	Reported R/Rfree	PDB redo R/Rfree	Rosetta Starting	Rosetta Final
1c1z	2.87	0.2380/0.2440	0.2310/0.2420	0.2385/0.2531	0.2198/0.2462
1uzg	3.5	0.2840/0.3240	0.2886/0.3075	0.2374/0.2748	0.1889/0.3166
2i69	3.11	0.2060/0.2680	0.2551/0.3039	0.2290/0.2850	0.2097/0.2706
5ezj	1.95	0.1980/0.2470	0.2242/0.2497	0.2430/0.2738	0.2317/0.2756
5gz4	2.55	0.2350/0.2880	0.2150/0.2991	0.2772/0.3159	0.2422/0.3132
5hy9	1.97	0.2120/0.2720	0.2370/0.2617	0.2454/0.3072	0.2323/0.3018
5k65	2.5	0.2210/0.2460	0.2316/0.2626	0.2357/0.2665	0.2169/0.2693
5la4	1.9	0.1850/0.2270	0.1651/0.2016	0.2136/0.2455	0.2041/0.2453
5n92	2.3	0.2050/0.2410	0.2075/0.2481	0.2355/0.2694	0.2292/0.2766
5nsc	2.3	0.2250/0.2650	0.2107/0.2497	0.2399/0.2602	0.2065/0.2614
5vem	2.6	0.1960/0.2270	0.1944/0.2113	0.2151/0.2476	0.1925/0.2520
5wbe	2.75	0.1950/0.2290	0.1926/0.2213	0.2446/0.2858	0.2107/0.2807

Table 3.1 R and Rfree of reported, PDB redo, and before and after Rosetta-Phenix refinement
This table reports the R and Rfree of all 12 models in the benchmark set. R and Rfree improve in every case between the Rosetta-Phenix starting and final models. All water and non-glycan ligands are have been stripped from the input model, possibly explaining any major differences between PDB and PDBredo reported values and those of Rosetta-Phenix.

Bibliography

1. McMullan, G., Chen, S., Henderson, R. & Faruqi, A. R. Detective quantum efficiency of electron area detectors in electron microscopy. *Ultramicroscopy* **109,** 1126–1143 (2009).
2. Levitt, M. & Zhang, J. Faculty of 1000 evaluation for Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *F1000 - Post-publication peer review of the biomedical literature* (2013). doi:10.3410/f.718021192.793479770
3. Scheres, S. H. Beam-induced motion correction for sub-megadalton cryo-EM particles. *Elife* **3,** e03665 (2014).
4. Vinothkumar, K. R. Faculty of 1000 evaluation for Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. *F1000 - Post-publication peer review of the biomedical literature* (2012). doi:10.3410/f.717965034.793466194
5. Campbell, M. G. *et al.* Movies of ice-embedded particles enhance resolution in electron cryo-microscopy. *Structure* **20,** 1823–1828 (2012).
6. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47,** 110–119 (1991).
7. Terwilliger, T. C. Automated main-chain model building by template matching and iterative fragment extension. *Acta Crystallogr. D Biol. Crystallogr.* **59,** 38–44 (2003).
8. Depristo, M. A., de Bakker, P. I. W., Johnson, R. J. K. & Blundell, T. L. Crystallographic refinement by knowledge-based exploration of complex energy landscapes. *Structure* **13,** 1311–1319 (2005).
9. Terwilliger, T. C. *et al.* Iterative model building, structure refinement and density

modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64,** 61–69 (2008).

10. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66,** 213–221 (2010).
11. Baker, M. R., Rees, I., Ludtke, S. J., Chiu, W. & Baker, M. L. Constructing and validating initial Cα models from subnanometer resolution density maps with pathwalking. *Structure* **20,** 450–463 (2012).
12. Lindert, S. *et al.* EM-fold: de novo atomic-detail protein structure determination from medium-resolution density maps. *Structure* **20,** 464–478 (2012).
13. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16,** 673–683 (2008).
14. Orzechowski, M. & Tama, F. Flexible fitting of high-resolution x-ray structures into cryoelectron microscopy maps using biased molecular dynamics simulations. *Biophys. J.* **95,** 5692–5705 (2008).
15. Schröder, G. F., Brunger, A. T. & Levitt, M. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure* **15,** 1630–1641 (2007).
16. 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).
17. Kaufmann, K. W., Lemmon, G. H., DeLuca, S. L., Sheehan, J. H. & Meiler, J. Practically Useful: What the RosettaProtein Modeling Suite Can Do for You. *Biochemistry* **49,** 2987–2998 (2010).

18. Song, Y. *et al.* High-resolution comparative modeling with RosettaCM. *Structure* **21,** 1735–1742 (2013).

19. Wang, R. Y.-R. *et al.* De novo protein structure determination from near-atomic-resolution cryo-EM maps. *Nat. Methods* **12,** 335–338 (2015).

20. 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).

21. Cowtan, K. The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallogr. D Biol. Crystallogr.* **62,** 1002–1011 (2006).

22. Šali, A. & Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **234,** 779–815 (1993).

23. Coleman, C. M. & Frieman, M. B. Coronaviruses: important emerging human pathogens. *J. Virol.* **88,** 5209–5212 (2014).

24. Bosch, B. J., van der Zee, R., de Haan, C. A. M. & Rottier, P. J. M. The Coronavirus Spike Protein Is a Class I Virus Fusion Protein: Structural and Functional Characterization of the Fusion Core Complex. *J. Virol.* **77,** 8801–8811 (2003).

25. Xu, Y. *et al.* Structural basis for coronavirus-mediated membrane fusion. Crystal structure of mouse hepatitis virus spike protein fusion core. *J. Biol. Chem.* **279,** 30514–30522 (2004).

26. Duquerroy, S., Vigouroux, A., Rottier, P. J. M., Rey, F. A. & Bosch, B. J. Central ions and lateral asparagine/glutamine zippers stabilize the post-fusion hairpin conformation of the SARS coronavirus spike glycoprotein. *Virology* **335,** 276–285 (2005).

27. Gao, J. *et al.* Structure of the Fusion Core and Inhibition of Fusion by a Heptad Repeat Peptide Derived from the S Protein of Middle East Respiratory Syndrome Coronavirus. *J.*

Virol. **87,** 13134–13140 (2013).

28. Supekar, V. M. *et al.* Structure of a proteolytically resistant core from the severe acute respiratory syndrome coronavirus S2 fusion protein. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 17958–17963 (2004).
29. Lu, G. *et al.* Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. *Nature* **500,** 227–231 (2013).
30. Li, F. Structure of SARS Coronavirus Spike Receptor-Binding Domain Complexed with Receptor. *Science* **309,** 1864–1868 (2005).
31. Peng, G. *et al.* Crystal structure of mouse coronavirus receptor-binding domain complexed with its murine receptor. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 10696–10701 (2011).
32. Wu, K., Li, W., Peng, G. & Li, F. Crystal structure of NL63 respiratory coronavirus receptor-binding domain complexed with its human receptor. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 19970–19974 (2009).
33. Peng, G. *et al.* Crystal structure of bovine coronavirus spike protein lectin domain. *J. Biol. Chem.* **287,** 41931–41938 (2012).
34. Walls, A. C. *et al.* Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature* **531,** 114–117 (2016).
35. Kirchdoerfer, R. N. *et al.* Pre-fusion structure of a human coronavirus spike protein. *Nature* **531,** 118–121 (2016).
36. van der Hoek, L. *et al.* Identification of a new human coronavirus. *Nat. Med.* **10,** 368–373 (2004).
37. Hofmann, H. *et al.* Human coronavirus NL63 employs the severe acute respiratory syndrome coronavirus receptor for cellular entry. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 7988–

7993 (2005).

38. Kendall, A. *et al.* Structure of flexible filamentous plant viruses. *J. Virol.* **82,** 9546–9554 (2008).
39. Canizares, M. C., Carmen Canizares, M., Nicholson, L. & Lomonossoff, G. P. Use of viral vectors for vaccine production in plants. *Immunol. Cell Biol.* **83,** 263–270 (2005).
40. Pogue, G. P., Lindbo, J. A., Garger, S. J. & Fitzmaurice, W. P. Making an ally from an enemy: plant virology and the new agriculture. *Annu. Rev. Phytopathol.* **40,** 45–74 (2002).
41. Bernal, J. D. & Fankuchen, I. X-RAY AND CRYSTALLOGRAPHIC STUDIES OF PLANT VIRUS PREPARATIONS. III. *J. Gen. Physiol.* **25,** 147–165 (1941).
42. Fraenkel-Conrat, H. Tobacco Mosaic Virus The History of Tobacco Mosaic Virus and the Evolution of Molecular Biology. in *The Plant Viruses* 5–17 (1986).
43. Namba, K. & Stubbs, G. Structure of tobacco mosaic virus at 3.6 A resolution: implications for assembly. *Science* **231,** 1401–1406 (1986).
44. Kendall, A. *et al.* A common structure for the potexviruses. *Virology* **436,** 173–178 (2013).
45. Yang, S. *et al.* Crystal structure of the coat protein of the flexible filamentous papaya mosaic virus. *J. Mol. Biol.* **422,** 263–273 (2012).
46. Lin, N.-S. *et al.* Nucleotide sequence of the genomic RNA of bamboo mosaic potexvirus. *J. Gen. Virol.* **75,** 2513–2518 (1994).
47. Lan, P., Yeh, W.-B., Tsai, C.-W. & Lin, N.-S. A unique glycine-rich motif at the N-terminal region of Bamboo mosaic virus coat protein is required for symptom expression. *Mol. Plant. Microbe. Interact.* **23,** 903–914 (2010).
48. Chen, T.-H. *et al.* Induction of protective immunity in chickens immunized with plant-made chimeric Bamboo mosaic virus particles expressing very virulent Infectious bursal disease

virus antigen. *Virus Res.* **166,** 109–115 (2012).

49. Yang, C.-D. *et al.* Induction of protective immunity in swine by recombinant bamboo mosaic virus expressing foot-and-mouth disease virus epitopes. *BMC Biotechnol.* **7,** 62 (2007).
50. DiMaio, F. *et al.* The molecular basis for flexibility in the flexible filamentous plant viruses. *Nat. Struct. Mol. Biol.* **22,** 642–644 (2015).
51. Apweiler, R., Hermjakob, H. & Sharon, N. On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochim. Biophys. Acta* **1473,** 4–8 (1999).
52. Derewenda, Z. The use of recombinant methods and molecular engineering in protein crystallization. *Methods* **34,** 354–363 (2004).
53. Agirre, J., Davies, G. J., Wilson, K. S. & Cowtan, K. D. Carbohydrate structure: the rocky road to automation. *Curr. Opin. Struct. Biol.* **44,** 39–47 (2017).
54. Lütteke, T., Frank, M. & von der Lieth, C.-W. Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. *Carbohydr. Res.* **339,** 1015–1020 (2004).
55. Agirre, J., Davies, G., Wilson, K. & Cowtan, K. Erratum: Carbohydrate anomalies in the PDB. *Nat. Chem. Biol.* **11,** 532 (2015).
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. DiMaio, F. *et al.* Improved low-resolution crystallographic refinement with Phenix and Rosetta. *Nat. Methods* **10,** 1102–1104 (2013).

58. Labonte, J. W., Adolf-Bryfogle, J., Schief, W. R. & Gray, J. J. Residue-centric modeling and design of saccharide and glycoconjugate structures. *J. Comput. Chem.* **38,** 276–287 (2017).