De novo model building with cryo-EM density and coevolution restraints

Carson Adams

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Science

University of Washington

Winter 2022

Committee:

Frank DiMaio

Neil King

Roland Strong

Ning Zheng

Judit Villen

Program Authorized to Offer Degree:

Biochemistry

©Copyright 2022

Carson Adams

University of Washington

Abstract

De novo model building with cryo-EM density and coevolution restraints

Carson Adams

Chair of the Supervisory Committee:

Frank DiMaio

Biochemistry

Low-resolution limitations have long plagued cryo-EM de novo modeling. To address this challenge, I modified Rosetta de novo to incorporate fragments with tertiary context, or nonlocal fragments. Limitations I identified led me to develop a novel method for low-resolution structure determination, combining density and predicted restraints. I used my pipeline to solve and publish two difficult structures, and show it can handle modeling homologous conformers better than attention-based deep-learning methods, AlphaFold and RosettaFold.

Foreword

Though my graduate school journey did not end as I expected, I wanted to take a moment to thank all of the friends and mentors that have brought me to this point. I have learned so much from this experience, thank you for your support and guidance.

Thank you to my wife, Meg, for keeping me sane the past five years.

Cryo-EM and De Novo Model Building:

Proteins are diverse molecules, performing highly specialized functions which correlate with their three-dimensional structure. Cryo-electron microscopy (cryo-EM) is one method for determining protein structure which has gained a lot of popularity in the last decade. In cryo-EM, the diffraction of a high-voltage electron beam across a sample immobilized in vitreous ice, yields a direct visualization the structure of proteins in two dimensions.[1,2] Projections of these flat images yields a three-dimensional density map, which at low resolution provides insight into the topology of the protein and at high resolution can used to build atomic models.[3] The number of deposited cryo-EM maps in the electron microscopy data bank (EMDB) is rising exponentially year-after-year, with most of the growth in the 3-5 Å resolution range.[4]

At this resolution range, side chain densities are often present but are not sharp enough to be easily distinguished or assigned from the map alone. In many of these cases, a model cannot be built manually from the data. Instead, models are generated either using known homology, or de novo using computational software.[5] Many computational methods exist for de novo modeling of cryo-EM maps.[6-9] The map_to_model module, for example, is provided as part of the Phenix software suite. This method processes the cryo-EM map into smaller segments and identifies points of high density in the map. These points are used to calculate the longest path in the segment, and Cαs are built along the path near identified Cβ positions using optimal distance. However, at resolutions worse than 3.0 Å, the method had only 23% sequence recovery.[10-11]

Rosetta also has a de novo pipeline, which uses fragment-based docking and local side chain refinement to solve structures.[12] Fragments are small regions of protein structure extracted from deposited protein models that closely match the sequence of the protein of interest. Fragments are generated for each window of residues, docked into the map via a translational

and rotational search followed by side-chain conformer optimization. Docked fragments are scored based on their agreement with the map and with each other, and the top scoring placements are assembled using Monte Carlo simulated annealing (MC-SA). The assembled model can be completed through iteration or with alternative Rosetta sampling methods.[13,14]

One of the strengths of Rosetta de novo is assigning sequence to the model as it is built. This allows us to assign accurate sequence where side chains are indistinguishable in the density. On the other hand, to identify the sequence, we must first place native-like fragments in the appropriate region. This is a non-trivial problem for large, low-resolution maps.

In my thesis work, I sought to address this limitation in de novo modeling software. First, I modified Rosetta de novo to incorporate fragments with tertiary context, or nonlocal fragments, to try to improve signal-to-noise in fragment docking. Limitations I identified led me to develop a novel method for low-resolution structure determination, combining density and predicted restraints. I used my pipeline to solve and publish two difficult structures, and show it can handle modeling homologous conformers better than attention-based deep-learning methods, AlphaFold and RosettaFold.

Nonlocal Fragment Docking:

The de novo fragment docking protocol in Rosetta is a powerful method for both backbone modeling and sequence recovery, but it is limited in its ability to identify native fragment placements from false placements at low resolution. Searching for longer fragments could potentially improve this signal, but with each additional search residue, not only does it become computationally more difficult to find sequence matches, but the matches that are found

tend to be more distant, resulting in less likelihood of the fragments matching the native conformation.

My solution was inspired by the Grigoryan Lab, which published a study organizing all structural regions of models pulled from the PDB into tertiary structural motifs or TERMs. In this study, models were segmented into 5 residue windows and any residues with rotamer atoms within 3Å of rotamer atoms in the segment were pulled out alongside them. TERMs were generated by clustering of the model segments. Upon clustering these segments, the authors found that over 50% could be represented by just 625 TERMs.[15] Given this high level of conservation, I hypothesized that docking fragments with tertiary context (nonlocal fragments) might improve the signal-to-noise in low-resolution maps and improve our docking accuracy.

Nonlocal fragments start as regular de novo fragments, and are modified based on their predicted secondary structure. If the fragment is predicted to be part of a strand, the corresponding structure is trimmed to the first 5 sequence-matched residues. Nearby strands are identified as a segment of at least 3 continuous residues with all Cβ within 7Å to a Cβ in the trimmed sequence-matched segment. By shortening the strand, I reduce the overall number of residues that I match to the density, thereby reducing the search space. For predicted helical fragments, all the sequence-matched structure is extracted and any predicted-helical residues in the C-terminal direction are also extracted. The helical extension helps to distinguish the native placement from decoys as backbone conformations within just 9 residues are unlikely to be unique enough to find the native placement in large proteins.

In a representative case, I used Rosetta de novo to generate models of an alcohol dehydrogenase monomer with and without nonlocal fragment generation (Figure 1). Nonlocal fragments show improvements in placement of some fragments, however, near-native fragment

placements are lost at other positions, making the results mixed overall. When mapping the missed positions in both fragment searches over the native structure, they encompass a broad range of secondary structures (Figure 1C). The data suggested that the conservation of these tertiary elements may not be tight enough to make them broadly applicable for docking.

Rosetta sampling with density and predicted restraints:

Given the mixed success of nonlocal fragment docking, I wanted to look at using predicted contacts to supplement density information. A recent development at the time, trRosetta is able to use coevolution data to make predictions about the distances and orientations between each pair of residues. Typically, this information is then used for Rosetta-guided folding and minimization to generate a final model . However, as this modeling is not guided by density, it can often produce models that do not match experimental data.

This shortcoming was evident in a collaboration with cryo-EM data for a subunit of phospho-inositol-3-kinase (PI3K). The p110 subunit was built into the density manually, however, the region of the density corresponding to the p101 subunit was too poorly resolved. Our goal was to resolve the interaction between these subunits, which is thought to be important to the function of PI3K.

Initially, I utilized our Rosetta de novo protocol on this structure, but these had limited effectiveness in this case. Fragments were confidently able to fill a strand on the front face of the subunit, but failed to fill in any more of the structure. At this point, I attempted trRosetta, to predict distances and angles between pairs of atoms. The resulting model visibly differed from the topology indicated by the data, however it seemed to indicate a β-strand interactions forming across the whole structure.

To get around this limitation of trRosetta, I developed a novel method for this case. A starting model is given, based on what is known from the data. We then use many trajectories of Monte Carlo sampling to insert and substitute homologous fragments into the growing model. Scoring is based on satisfaction of predicted restraints as well as fit into experimental density (Figure 2). The top-scoring full-length trajectories are analyzed manually, and can be pulled out as starting models for iterative improvements.

The combination of density and predicted restraints was essential to generating a full-length model of p101 (Figure 3). After 10,000 trajectories and 3 100-trajectory iterations using the previous starting model, our trajectories converged onto a structure favored by Rosetta. Manual inspection showed good agreement with the density, and satisfaction of high-confidence restraints remained unchanged. Our p101 structure was crucial to make inferences about the interactions between p110 and p101.[16]

My novel sampling pipeline was also essential in solving the structure of a seipin oligomer's terminal helices. Seipin is an integral membrane protein that forms a 10-mer homo-oligomeric structure, however it has only C5 symmetry as the complex alternates between A and B conformers around the ring.[17] Despite having varied functions, only the structure of the globular domain of the protein has been previously characterized, while the structures of the N- and C- termini which reach into membrane-helices were heretofore unknown.

Initially, I performed trRosetta in this case, but the top-scoring models all closely resembled our starting model and density of the A conformer, while not fitting the B conformer density at all. This result highlights a limitation of trRosetta; without density data, different conformers or states are indistinguishable. I hypothesized that mixed conformer signals make the

trRosetta signal ambiguous, but that density data could help to distinguish the native states from within that data.

Using the predicted distances and angles in our novel pipeline, I was able to produce high-confidence models for the terminal helices in both conformers (Figure 4). When I compare the satisfaction of high-probability restraints between the top trRosetta model and both modeled conformers, the restraints remain similarly satisfied in all the models. Though the density for the helices is weak, our model fills the density that is present for both conformers.[18]

Deep-learning methods:

Recently, the attention-based deep learning methods RosettaFold and AlphaFold were published, and shown to make dramatic improvements over previous de novo modeling software.[19,20] One remaining question, however, is the limitations of these methods, and when density information is still necessary. To answer this question, I looked at the performance of both these methods for p101 as well as seipin modeling.

AlphaFold and RosettaFold both produce reasonable models for p101 that validate the model we built into the density (Figure 5). Despite slight differences in less constrained regions, the models are topologically identical. The convergence of these methods increases our confidence in our model, but also indicates that density information is no longer needed to solve this case.

Seipin, however, was unable to be modeled in either AlphaFold or RosettaFold. Due to the conformational heterogeneity, all the top models produced by both methods closely matched the A conformer, while the B conformer remains unmodeled. As the coevolution information

cannot distinguish between conformers, this is a modeling case where density information is absolutely needed.

Discussion:

Overall, my novel pipeline fills a niche in de novo modeling where the predicted information is too ambiguous to identify the native conformation, or when there is more than one predicted state. As cryo-EM is a very effective technique for determining the structure of large complexes, maps that are difficult to interpret are common. Thus, my method continues to be useful alongside the success of new deep learning methods.

The next step for this project is testing AlphaFold and my pipeline across a larger set of structures. Identifying more cases where AlphaFold fails to solve the structure will help to further characterize its limitations, while identifying cases where density information could potentially be helpful.

Figure 1: Nonlocal fragments marginally improve coverage of Alcohol Dehydrogenase
A) Positional coverage of Rosetta de novo fragment placement. Every fragment is plotted as a point according to which position in the sequence was used to generate the fragment and its RMSD to the actual structure in this region. Magenta indicates the lowest-RMSD placement at this position is near-native. B) Positional coverage of Rosetta de novo with nonlocal fragments. Overall, more positions are docked to near native (magenta) compared to local only fragments. C) Comparison of local fragment coverage (left) and nonlocal fragment coverage (right) after de novo docking in Rosetta mapped onto the native structure. The blue color indicates that a near-native dock was achieved at that position, while orange indicates a missed position. Overall, while nonlocal fragments improve fragment docking at some positions, they lose some near-native placements as well.

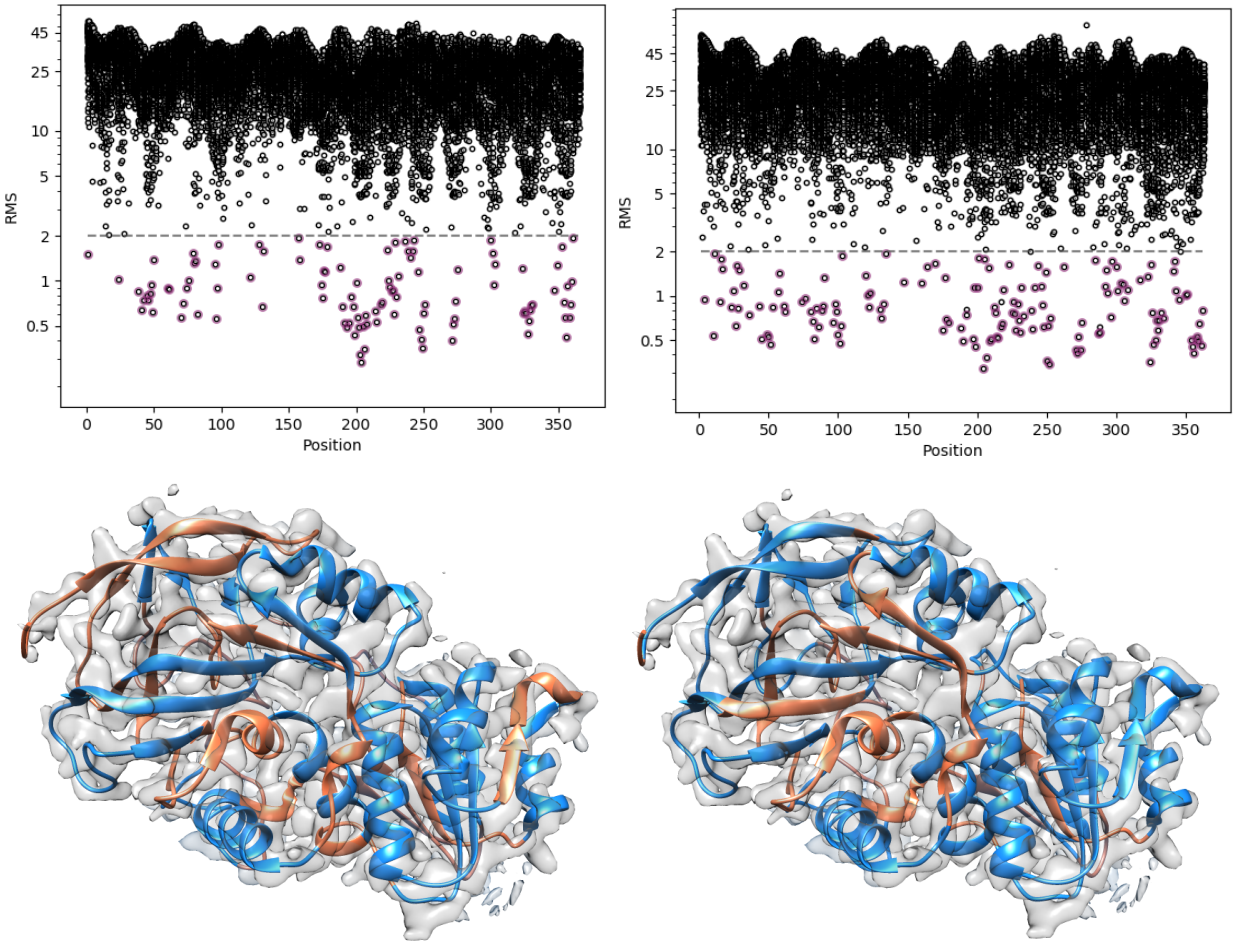


Figure 2: Novel pipeline combines density and predicted restraints in Rosetta-guided modeling. Starting from an initial model, our pipeline uses 10,000 trajectories of sampling from which models are built via combinations of fragment insertions. Each trajectory is color-coded below, and example structures resulting from a fragment insertion are shown in 2-D cartoon form. Each insertion is scored based on its agreement with the density, as well as the predicted restraints, exemplified by the colored dotted line. Red indicates unsatisfied, while green indicates satisfied restraints. The Monte Carlo simulation then accepts or rejects each fragment, and the resulting model is used for repeated rounds of insertions, until a full-length model is completed.

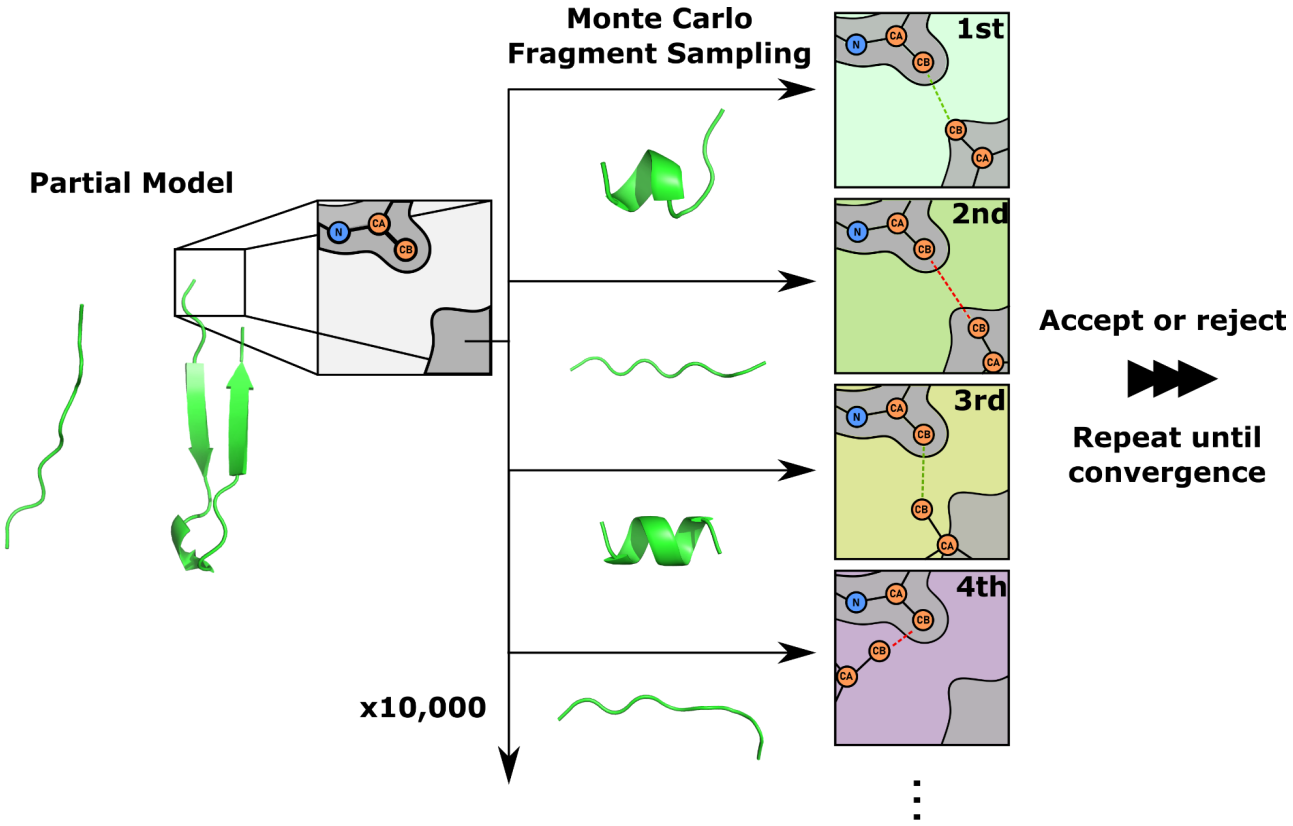
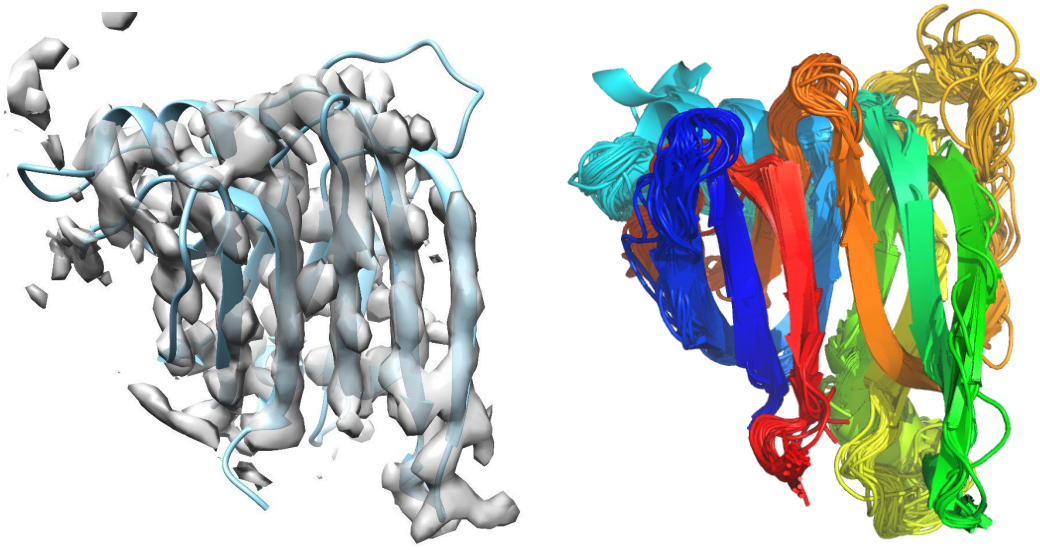


Figure 3: Density and coevolution-guided pipeline identifies strands in PIK3CG p101. A) Our final p101 model shown in the density. Overall, the density agrees well with the model, outside of errant loops B) An overlay of the ensemble of 100 models from our novel pipeline. The closely overlapping chain trace indicates that all trajectories converged on the same solution, validating the model. C) The predicted and actual distances between beta-carbons of residues in the p101 nanobody-interacting domain. The color of each pixel corresponds to the distance in angstroms between these atoms. Plotted on the left is the least distance predicted by our improved trRosetta pipeline with >95% probability for each pair of CB atoms. On the right are the actual distances between these atoms. We correctly predict beta-strand interactions between residues 720 and 740, 750 and 770, as well as 770 and 800, among others.



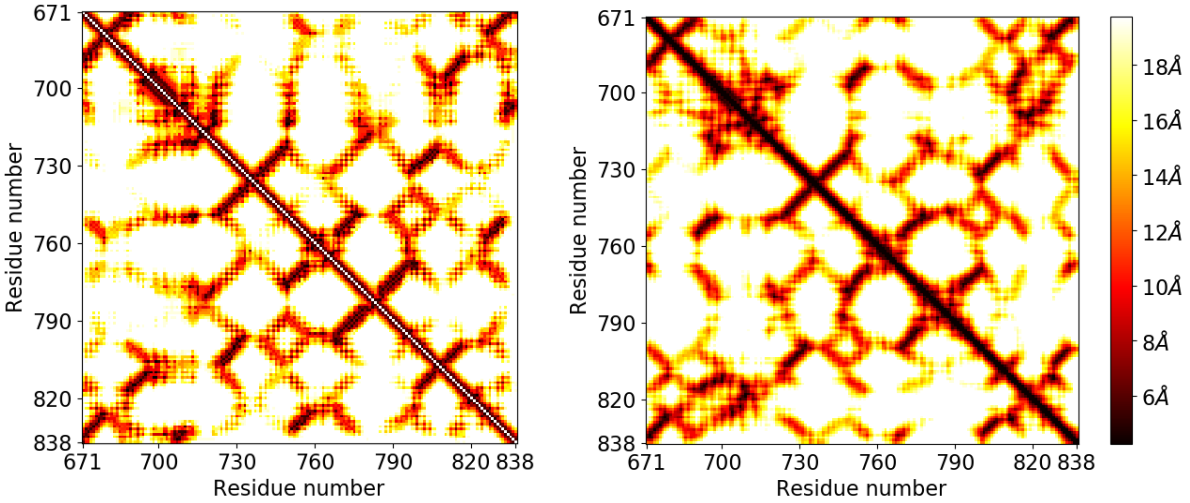


Figure 4: Density and coevolution-guided sampling models both seipin conformers.
A) Final model built for each conformer by our pipeline into the density. Though inconsistent, the density that is present agrees with the orientation of helices, both in conformer A (blue) and conformer B (pink). B) Overhead view of the conformers built into the full 10-mer density. C) Distances predicted by trRosetta on the left and actual distances for the A and B conformer respectively to the right. The pairwise distance maps for both conformers match the predictions across the helical regions (lower-left and upper-right corners).

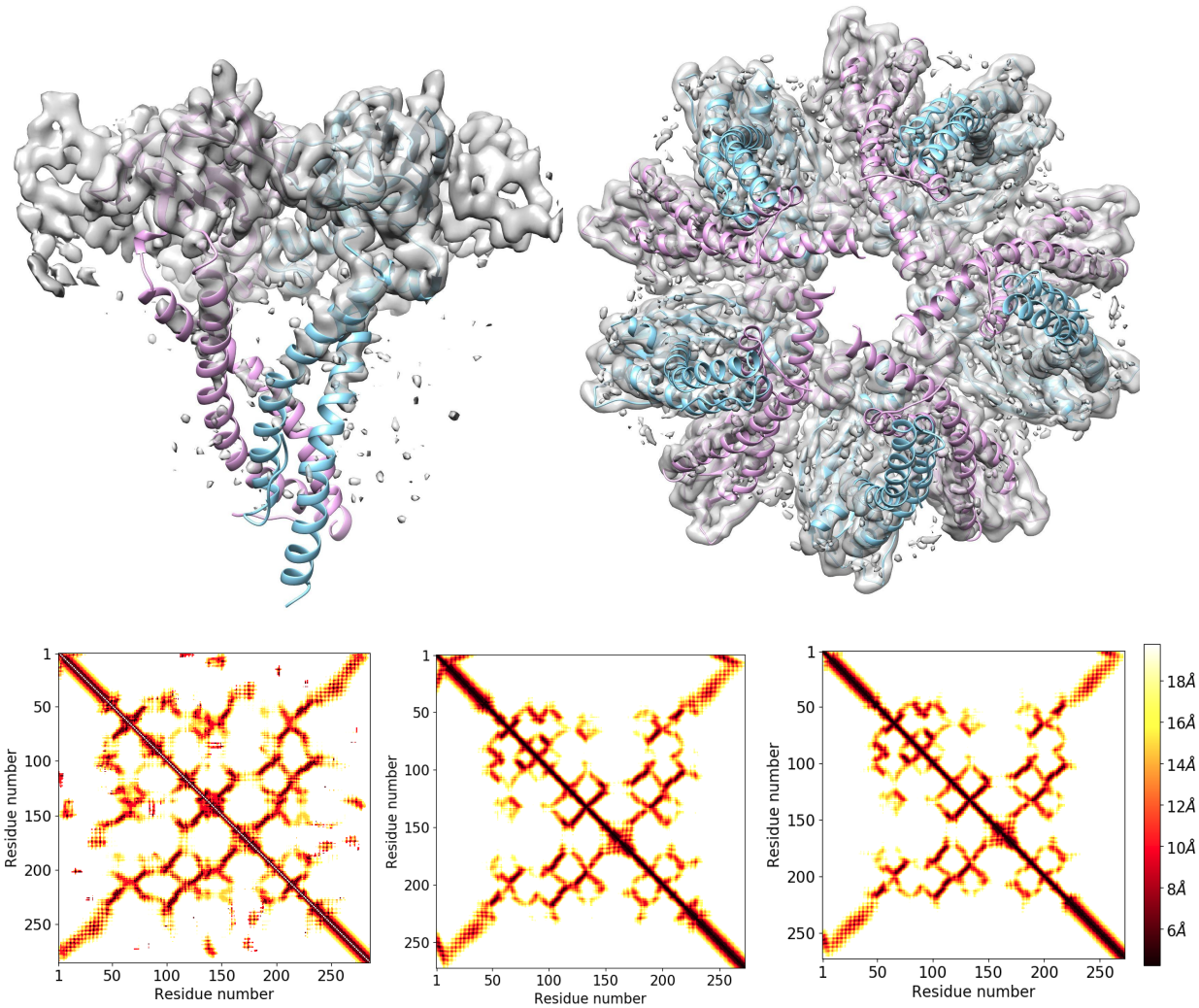
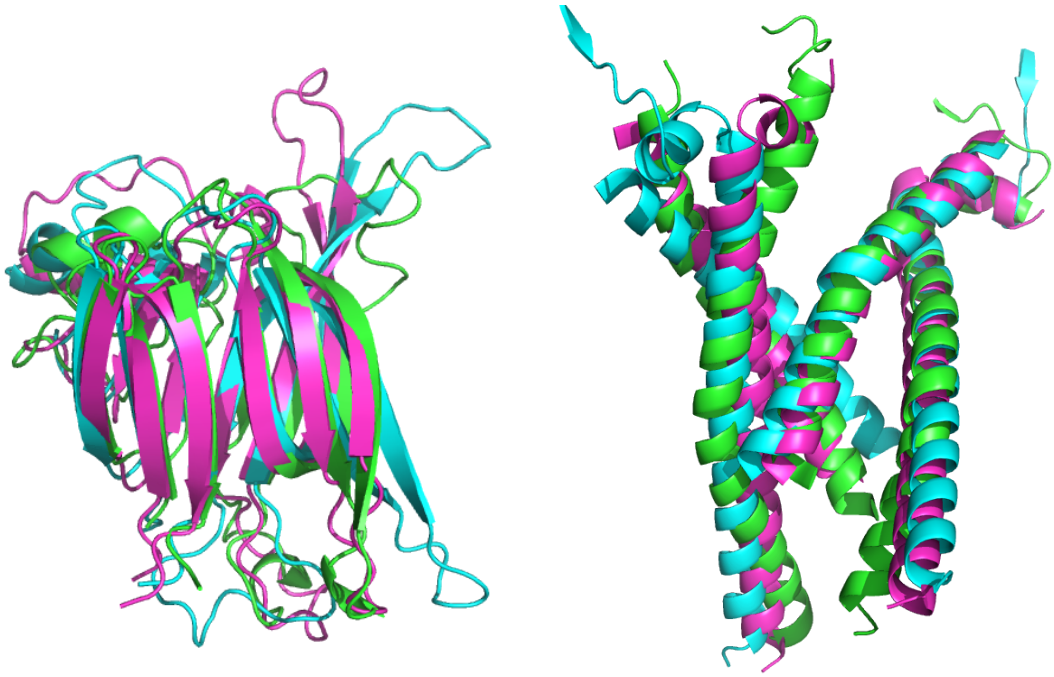


Figure 5: Deep learning validates PIK3CG p101 model, but fails in seipin modeling. A) AlphaFold (cyan) and RosettaFold (magenta) models of p101 compared to our final structure (green). Both methods correctly identify the topology of the strands, helping to validate the model we deposited but also solving the problem outright. B) AlphaFold and RosettaFold models of seipin, compared to the native, generally fail to reproduce the conformers. The AlphaFold model had a global RMSD of 6.15 Å and the RosettaFold model had a global RMSD of 12.63 Å compared to our deposited model.



References:

1. Adrian, M., Dubochet, J., Lepault, J., McDowall, A. W. (1984). Cryo-electron microscopy of viruses. Nature, 308(5954), 32.
2. Dubochet, J., Adrian, M., Chang, J. J., Homo, J. C., Lepault, J., McDowall, A. W., Schultz, P. (1988). Cryo-electron microscopy of vitrified specimens. Quarterly reviews of biophysics, 21(2), 129-228.
3. Frank, J., Goldfarb, W., Eisenberg, D., Baker, T. S. (1978). Reconstruction of glutamine synthetase using computer averaging. Ultramicroscopy, 3(3), 283-290.
4. Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., ..., Moriarty, N. W. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallographica Section D: Structural Biology, 75(10), 861-877.
5. Wang, G., Porta, C., Chen, Z. et al. (1992). Identification of a Fab interaction footprint site on an icosahedral virus by cryoelectron microscopy and X-ray crystallography. Nature 355, 275–278
6. Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallographica Section D: Biological Crystallography, 62(9), 1002-1011.
7. Baker, M. L., Baker, M. R., Hryc, C. F., Ju, T., & Chiu, W. (2012). Gorgon and pathwalking: macromolecular modeling tools for subnanometer resolution density maps. Biopolymers, 97(9), 655-668.
8. Chen, M., Baldwin, P. R., Ludtke, S. J., Baker, M. L. (2016). De Novo modeling in cryo-EM density maps with Pathwalking. Journal of structural biology, 196(3), 289-298.
9. Rotkiewicz P, Skolnick J. (2008). Fast procedure for reconstruction of full‐atom protein models from reduced representations. J Comput Chem 29:1460–1465.
10. Thomas C. Terwilliger, Paul D. Adams, Pavel V. Afonine, O. V. S. (2018). A fully automatic method yielding initial models from high-resolution electron cryo-microscopy maps. BioRxiv Biochemistry, 15(November). https://doi.org/10.1101/267138
11. Terwilliger, T. C., Adams, P. D., Afonine, P. V., Sobolev, O. V. (2019). Cryo‐EM map interpretation and protein model‐building using iterative map segmentation. Protein Science.
12. Wang, R. Y. R., Kudryashev, M., Li, X., Egelman, E. H., Basler, M., Cheng, Y., … DiMaio, F. (2015). De novo protein structure determination from near-atomic-resolution cryo-EM maps. Nature Methods, 12(4), 335–338.
13. Song, Y., DiMaio, F., Wang, R. Y. R., Kim, D., Miles, C., Brunette, T. J., ... Baker, D. (2013). High-resolution comparative modeling with RosettaCM. Structure, 21(10), 1735-1742.
14. Frenz, B., Walls, A. C., Egelman, E. H., Veesler, D., DiMaio, F. (2017). RosettaES: a sampling strategy enabling automated interpretation of difficult cryo-EM maps. Nature methods, 14(8), 797.
15. Mackenzie, C. O., Zhou, J., & Grigoryan, G. (2016). Tertiary alphabet for the observable protein structural universe. Proceedings of the National Academy of Sciences, 113(47), E7438–E7447.
16. Rathinaswamy, M. K., Dalwadi, U., Fleming, K. D., Adams, C., Stariha, J. T., Pardon, E., ... & Burke, J. E. (2021). Structure of the phosphoinositide 3-kinase (PI3K) p110γ-p101

complex reveals molecular mechanism of GPCR activation. *Science Advances*, *7*(35), eabj4282.
17. Binns, D., Lee, S., Hilton, C. L., Jiang, Q. X., & Goodman, J. M. (2010). Seipin is a discrete homooligomer. *Biochemistry*, *49*(50), 10747-10755.
18. Arlt, H., Sui, X., Folger, B., Adams, C., Chen, X., Remme, R., ... & Walther, T. C. (2022). Seipin forms a flexible cage at lipid droplet formation sites. *Nature Structural & Molecular Biology*, 1-9.
19. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583-589.
20. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., ... & Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871-876.