Bates College
SCARAB

Honors Theses | Capstone Projects

5-2022

AI in Proteomics: A Comparison Between Crystallographic and in silico Methods of Protein Modeling

Alex D. Weissman
Bates College, aweissm2@bates.edu

Follow this and additional works at: https://scarab.bates.edu/honorstheses

Recommended Citation

Weissman, Alex D., "AI in Proteomics: A Comparison Between Crystallographic and in silico Methods of Protein Modeling" (2022). *Honors Theses*. 418.
https://scarab.bates.edu/honorstheses/418

This Restricted: Embargoed [Open Access After Expiration] is brought to you for free and open access by the Capstone Projects at SCARAB. It has been accepted for inclusion in Honors Theses by an authorized administrator of SCARAB. For more information, please contact batesscarab@bates.edu.

AI in Proteomics: A Comparison Between Crystallographic and *in silico* Methods of Protein Modeling

An Honors Thesis

Presented to

The Faculty of the Department of Biology

Bates College

In partial fulfillment of the requirements for the

Degree of Bachelor of Arts

By

Alexander Daiki Weissman

Lewiston, Maine

5/3/2022

Acknowledgements

I would like to thank the Bates College Department of Biology for supporting this research opportunity. I would also like to thank the Google Deepmind team as well as the Yang Lab for providing open-access protein structure prediction software allowing us to perform our work. Research reported in this project was supported by funds from the Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number P20GM103423, and institutional funds from the Biology Department Thesis Fund, New Faculty Start-up, Purposeful Work, and a Roger C. Schmutz Grant for Faculty Research. I am grateful to my thesis advisor Dr. Lori Banks who provided me with numerous research opportunities for the majority of my time at Bates College. One of these opportunities included the Reproducible and FAIR Bioinformatics Analysis of Omics Data training course, which provided me with crucial background information and skills needed for this project. Finally, I would like to personally thank Dr. Peter Schlax, the Science and Data Librarian, for his instrumental contributions to this project. Pete Schlax informed me of the existence of the AlphaFold Colab notebook, ultimately allowing me to pursue my topics of research.

Table of Contents

Abstract

All living cells contain membrane proteins called ion channels. Ion channels regulate the diffusion of ions across both sides of the membrane, allowing for vital functions such as the regulation of cellular content and the transmission of electrical signals between cells. Viroporins are ion channels that are encoded by viruses. Viroporins are found in devastating viral pathogens such as COVID-19, HIV, and rotavirus. NSP4 is a rotavirus viroporin that disrupts cellular Ca^{2+} homeostasis, leading to host cell lysis and the proliferation of more virions. Though the structure of some domains of NSP4 have been determined, the full-length structure of this viroporin is unknown. Until recently, only *in vitro* methods of structure determination, such as x-ray crystallography, were considered to be accurate. However, two recently published algorithms demonstrated a high degree of accuracy when determining the structure of membrane proteins from their amino acid sequences. These algorithms are known as AlphaFold and trRosetta. The goal of this project is to use AlphaFold to predict the molecular structure of the full-length NSP4 from SA11 rotavirus and compare this structure to a recently determined structure from the Banks lab generated via trRosetta, as well as its established crystal structure. This project also includes comparisons between predicted and experimental structures for AlphaFold's intended targets: eukaryotic and prokaryotic proteins.

Introduction

Rotavirus (RV) is a vaccine preventable infectious agent chiefly responsible for viral gastroenteritis in young children across the world. In fact, prior to the introduction of the rotavirus vaccine, it was estimated that 95% of children worldwide were infected with RV gastroenteritis by the age of 5, and caused roughly 450,000 under-5 deaths annually (2, 3, 4, 10). The introduction of the RV vaccine has resulted in this annual death count to drop to roughly 200,000. However, most of these annual deaths originate from low-income countries (1, 3). These deaths are typically met with numerous comorbid conditions including malnutrition, lack of available potable water, and limited access to health care (3). These comorbidities translate to lower vaccine efficacy in low-income countries (5, 10). Uncovering the mechanisms of RV pathogenesis will provide insight into developing drugs to more effectively counter viral gastroenteritis.

Rotaviruses are non-enveloped, icosahedral, double-stranded RNA viruses in the *Sedoreovirinae* subfamily of the *Reoviridae* family. Its three capsids surround a genome of 10-12 segments of dsRNA, varying among the different genera, encoding six structural and six non-structural viral proteins (3, 5). RV structural proteins allow for viral host specificity, cell entry, and enzymatic activity whereas the non-structural proteins are responsible for genome replication and the deactivation of the host's innate immune response (3). RV non-structural protein 4 (NSP4) plays a vital role in RV replication, morphogenesis, and pathogenesis, making it vital for replication. Therefore, NSP4 has proved to be a fascinating target for ongoing drug design.

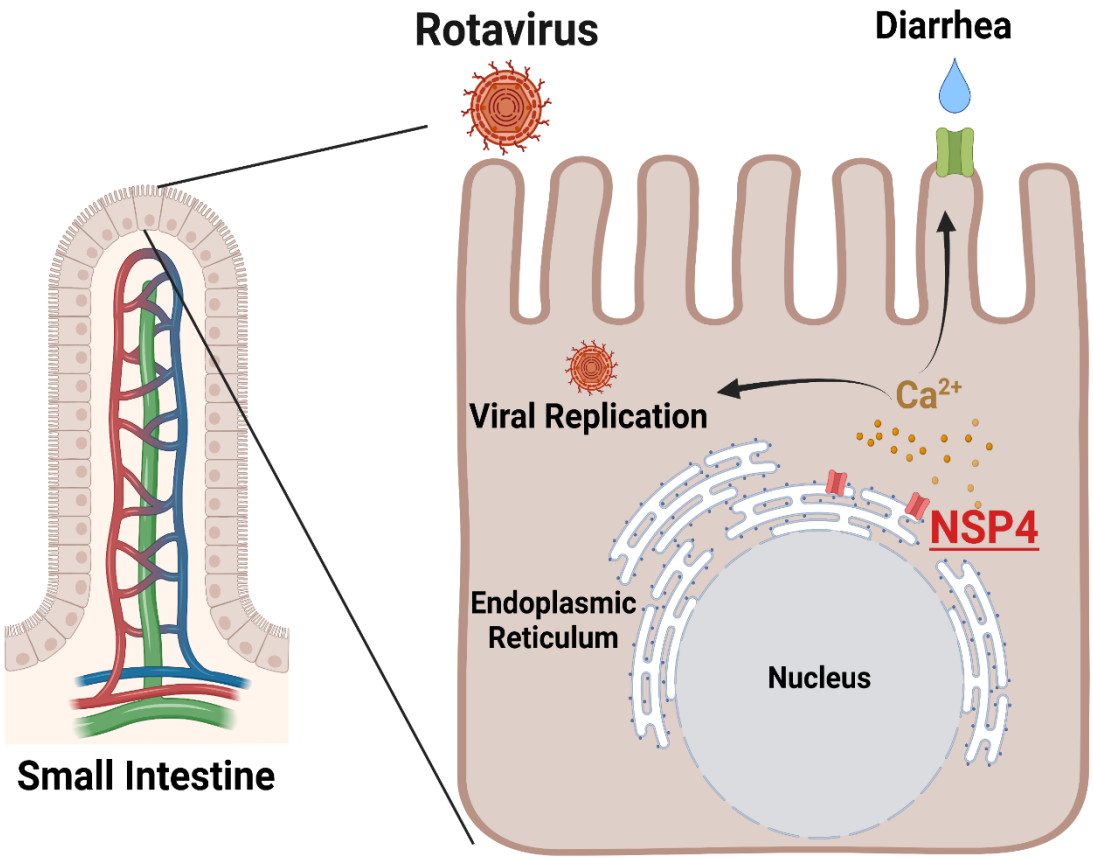

Figure 1. Graphic representation of biochemical pathway by which NSP4 increases cytosolic Ca^{2+}. *When NSP4 inserts itself into the ER membrane, it increases cytosolic Ca^{2+} which activates Ca^{2+} activated Cl^- channels, resulting in the secretion of Cl^- in the lumen of mammals. This is the primary mechanism leading to RV gastroenteritis. This graphic was created with BioRender.com*

NSP4 is a transmembrane glycoprotein which is initially synthesized in the endoplasmic reticulum (ER) (9). In the ER, NSP4 plays a key role in virus maturation. When NSP4 is active, it can release Ca^{2+} from the ER, elevating cytosolic Ca^{2+} in eukaryotic cells (8). This in turn activates Ca^{2+} activated Cl^- channels, resulting in the secretion of Cl^- in the lumen of mammals. This is the primary pathological mechanism which causes RV gastroenteritis. NSP4 is a multifunctional protein consisting of a viroporin domain (VPD, residues 47-90) and a coiled-coil domain (CCD, residues 95-137), motifs that are commonly associated with other virus-encoded ion channel

proteins (viroporins) (8, 9). NSP4 has also been shown to secrete an enterotoxin cleavage product (Enterotoxin, residues 112-175) from infected cells (21). The exact demarcation of each of these domains is not clearly defined. Structurally, viroporins typically consist of 60-120 amino acids. Though viroporins target a wide range of intracellular components, this class of proteins tends to share secondary structural motifs such as amphipathic alpha-helices and clusters of basic residues (6). The electrostatic properties of these common motifs aid in viroporin insertion into the host cell membrane. In NSP4, these common motifs are apparent within the VPD (amino acids 47-90) (9). Within the VPD exists the amphipathic domain (amino acids 71-92) as well as the pentalysine domain (amino acids 63-84), two structures that have been found to be critical for NSP4 transmembrane insertion (9). In addition to the determination of secondary structures within NSP4, crystallographic studies in the past have shown that the CCD of NSP4 has two oligomeric states, a Ca^{2+} bound tetrameric conformation and an ion-free pentameric conformation. The oligomeric state of NSP4 appears to be affected by changes in pH (9). Apart from these crystallographic studies focusing on the several domains of NSP4, the structure of the entire viroporin is largely unknown.

In vitro methods of deducing protein structure, which include cryo-EM, NMR, and x-ray crystallography, have several drawbacks. For one, existing *in vitro* methods sometimes struggle to produce atomically accurate structures, especially when there are no known homologous structures (12). Additionally, the cost to generate structure models from these methods is quite high. Furthermore, these methods suffer from exceedingly long turnaround times. This drawback has been exacerbated by supply

chain shortages during the COVID-19 pandemic. Owing to this fact, the original aim of this study, which was to obtain NSP4 structure models via crystallographic methods, has proved to be impossible to complete within the given timeframe. However, our inability to obtain crystallization conditions for NSP4 was also attributable to the fact that membrane proteins are notoriously difficult to crystallize. Since cryo-EM and NMR do not rely on crystallized samples, future studies should attempt these methods to obtain NSP4 structure models. Given the critical role of this viroporin in RV pathogenesis, it is vital to determine the full-length structures of the NSP4 multimers in order to fully understand the various functions of each of its domains. However, previous attempts to do so have been bottlenecked by the protein-folding problem.

First emerging in 1960, the protein folding problem has proven to be among the most elusive mysteries in modern biochemistry. The discovery of the first atomic-resolution protein structures gave rise to the question of how a protein's primary structure dictates its secondary-quaternary structure(s) (11). However, efforts to understand this mystery have been impeded by the previously stated drawbacks of *in vitro* methods. Recognizing that the development of accurate *in silico*-based protein structure modeling methods would allow for modeling from genome sequences, and the potential to replace *in vitro* methods, John Moult created the Critical Assessment of Techniques for Protein Structure Prediction (CASP). CASP is a biennial competition to test the effectiveness of structure-prediction algorithms within the international community of computational biology (11). In this competition, teams are given the sequences of proteins whose structures have already been determined experimentally but have not yet been publicly disclosed. The structures generated by the computational

models are compared to the experimentally determined structures. Last year in the Banks lab, Jeremy Bennett utilized a program called trRosetta, which was among the most accurate structure prediction algorithms featured at CASP13, to generate, to our knowledge, the first full length model of NSP4.

CASP14 marked the most exciting year in the history of this competition, as DeepMind, a Google owned company based in London, released their second iteration of a protein folding algorithm called AlphaFold2. AlphaFold2 is the first algorithm of its kind to regularly predict protein structures with near experimental accuracy even when there is no known homologous structure (12). Interestingly, for many of the predictions where AlphaFold2 disagreed with the experimentally determined structures, the margin of error from both models was so small that it was not actually clear which was closer to the true structure. Although AlphaFold2 had the highest overall accuracy, one weakness of this software was its inability to accurately predict the structure of protein complexes. The AlphaFold team addressed this shortcoming in October of 2021 with the release of AlphaFold-Multimer, a model trained specifically for multimeric inputs which significantly increased the accuracy of predicted multimeric interfaces over single-chain AlphaFold while maintaining high intra-chain accuracy (13). For homomeric interfaces, AlphaFold-Multimer successfully predicted the structure in 69% of cases. These astonishing results are made possible by deep learning.

Neural networks consist of thousands or even millions of densely interconnected processing nodes. Nodes within a neural network are typically organized into layers in which data flows in one direction (14). Deep learning refers to a subset of neural networks with at least three layers of nodes, consisting of an input layer, one or more

succeeding layers, and a final output layer. An individual node is typically connected to several nodes in a layer beneath it and a layer above it, receiving data from one end, and sending data through the other (14). Nodes function by assigning numbers called “weights” to each incoming connection. When a node receives data from an incoming connection, this value is multiplied by its associated weight. If the resulting number exceeds the threshold value, the node “fires” and sends this value to its outgoing nodes. A fully trained neural network consists of properly interacting layers of nodes which can learn to recognize patterns within a training dataset. When neural networks are being trained, all of the weights and thresholds are set to random values. The output is then checked for errors against the training dataset. Data from the errors are used to adjust weights and thresholds over “generations” until the neural network is able to consistently yield accurate outputs (14). With each generation, weights and thresholds associated with proper function emerge, making the algorithm more accurate.

For protein folding neural networks, the training dataset consists of known protein structures from the Protein Data Bank (PDB). AlphaFold is unique from other structure prediction algorithms in its use of a secondary training dataset modeled after an approach called “noisy student self-distillation” (12). The noisy student approach is a type of semi-supervised learning, allowing for the training of deep learning algorithms with the use of labeled and unlabeled data in three main steps: 1) train a teacher model using labeled data, 2) use the teacher model to assign “pseudo labels” to unlabeled data, and 3) train a student model using a combination of labeled and “pseudo labeled” data (16). Using the student model as a teacher to train a new student can result in highly accurate recognition algorithms. The DeepMind team used a trained neural

network to predict the structure of around 350,000 assorted sequences from the Uniclust30 database. A combination of these structure predictions as well as PDB data was used to train AlphaFold2.

The training datasets from both AlphaFold and trRosetta do not include viral proteins. This is because viruses often encode polyproteins, which have exceedingly plastic structures that are often too unpredictable/resource intensive to model. Because NSP4 is not cleaved off of a larger polyprotein, we believe this constraint is not applicable. Furthermore, both AlphaFold and trRosetta have recently been used in addition to molecular dynamics (MD) simulations to model the structure of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) proteins, though these results have yet to be peer reviewed (17,18). Several SARS-CoV-2 proteins, including coronavirus NSP4 have not yet been modeled via *in vitro* methods, making these predictions the only available models.

Here, we generate SA11 rotavirus NSP4 monomer, tetramer, and pentamer models via the Colab notebook version of AlphaFold2/AlphaFold-Multimer. We then compare these models to a trRosetta-generated monomeric model as well as the multimeric models established by previous crystallographic studies. This will allow us to weigh the strengths and weaknesses of each methodology attempted to determine the structure of this viroporin. We also generate AlphaFold2 models of a eukaryotic and prokaryotic protein to compare with their established crystal structures. Lastly, we compare an AlphaFold2 model of SARS-CoV-2 NSP2 that was created and refined via MD simulations by (17) with the recently published corresponding crystal structure (21).

Results

When comparing between the AlphaFold2 model (Figure 2A) and the trRosetta model (Figure 2B), we found several structural differences. Aligning the two models resulted in a staggeringly high root-mean-square deviation (RMSD) value of 23.70 Å (Figure 2C). RMSD is a measure of accuracy in which deviations between different models of a particular dataset are compared. An RMSD value of <3 Å is typical for homologous proteins. The trRosetta model predicted many more coils separated at seemingly random intervals. Furthermore, instead of predicting results consistent with the established motifs of the VPD and CCD, this model instead splits the entire structure into a series of ~26 amino acid long alpha helices (Figure 3). Similarly, the AlphaFold2 model was not consistent with the established VPD (Figure 3) (6). However, the AlphaFold2 model's prediction of the CCD was nearly the same as the established model. This model predicted the CCD to be at residues 92-139, whereas the crystallographic data suggests the CCD to be at residues 95-137.

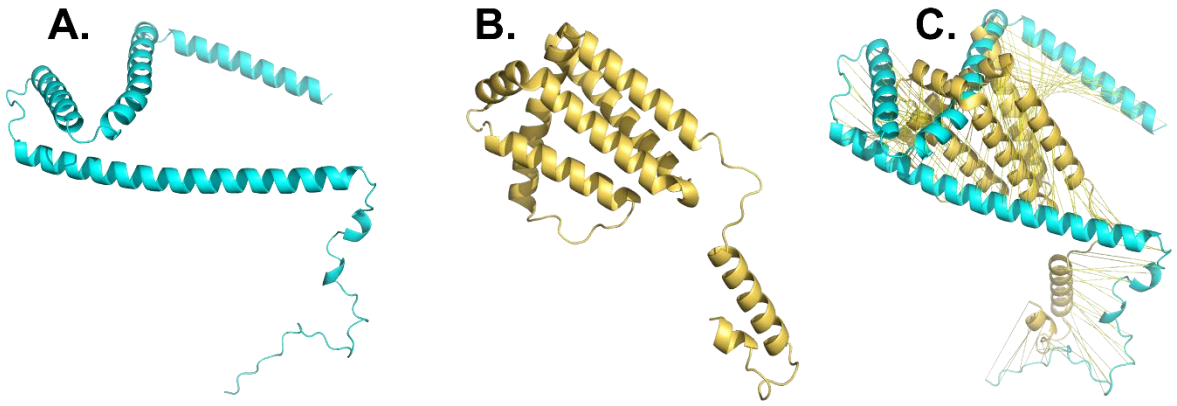

Figure 2. Predicted models of full-length WT rotavirus SA11 NSP4 monomers. *(A) The AlphaFold2 model of NSP4 is colored in cyan. (B) The trRosetta model of NSP4 is colored in gold. (C) Alignment of the two computational models resulted in an RMSD value of 23.70 Å. Homologous molecules are connected by yellow lines. Models were rendered using PyMOL Molecular Graphics System, Version 2.0.*

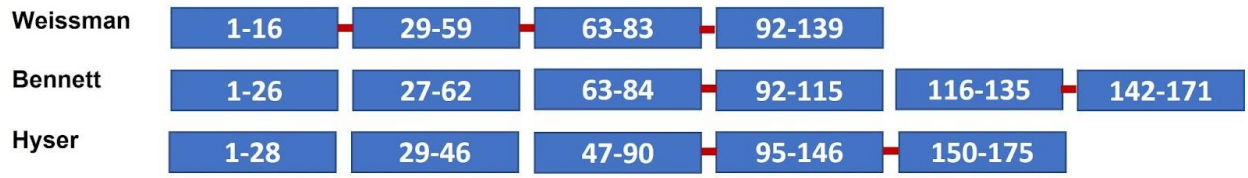

Figure 3. Motif Demarcation of three models of full-length rotavirus SA11 NSP4 monomers. *Blue boxes represent predicted helices, and red lines represent coils/unstructured regions between helices. Bennett, Hyser, and Weissman models were generated via in silico methods (6).*

To properly compare our full-length tetrameric model with the established crystallographic tetramer model of the CCD of the rotavirus SA11 NSP4, we excluded the VPD and residues 162-175 from the AlphaFold-Multimer model. (Figure 4A and B) reveal a discrepancy in the orientation of residues about the Ca^{2+} binding site (E120 and Q123). An initial alignment of the two models resulted in the high RMSD value of 10.79 Å. However, an alignment of the split-state of the two models resulted in an RMSD value of 3.844 Å, suggesting a potentially significant degree of structural similarity between the two models. However, the Ramachandran plot of these two models reveals a high degree of disparity between the torsional angles phi (φ) and psi (ψ) of the backbone of the two models (Figure 7). The AlphaFold-Multimer model has many residues consistent with beta-pleated sheet characteristics, motifs that are not associated with RV NSP4. Conversely, essentially all of the residues from the crystal structure of the tetramer have right-handed helical character, motifs that are heavily associated with RV NSP4. Furthermore, (Figure 7A) shows that 0.6% of the residues in the AlphaFold model are in structurally disallowed regions.

Like the comparison of tetrameric models, we excluded the VPD and residues 162-175 from our full-length pentameric model for proper comparison of the two models. Again, we found a potentially significant degree of similarity between the two models.

Aligning the two models resulted in an RMSD value of 3.848 Å, just 0.004 Å higher than that of the split-state tetramers. Panels A and B of figure 5 demonstrate the preserved rotational symmetry of both models. Despite the structural similarities indicated by the acceptable RMSD value, the Ramachandran plot of these two models told a different story (Figure 8). Again, this plot shows that the AlphaFold-Multimer model has many residues consistent with beta-pleated sheet characteristics despite effectively all the residues from the pentameric crystal structure having right-handed helical character. Furthermore, (Figure 8A) shows that 0.1% of residues in the AlphaFold model were in structurally disallowed regions.

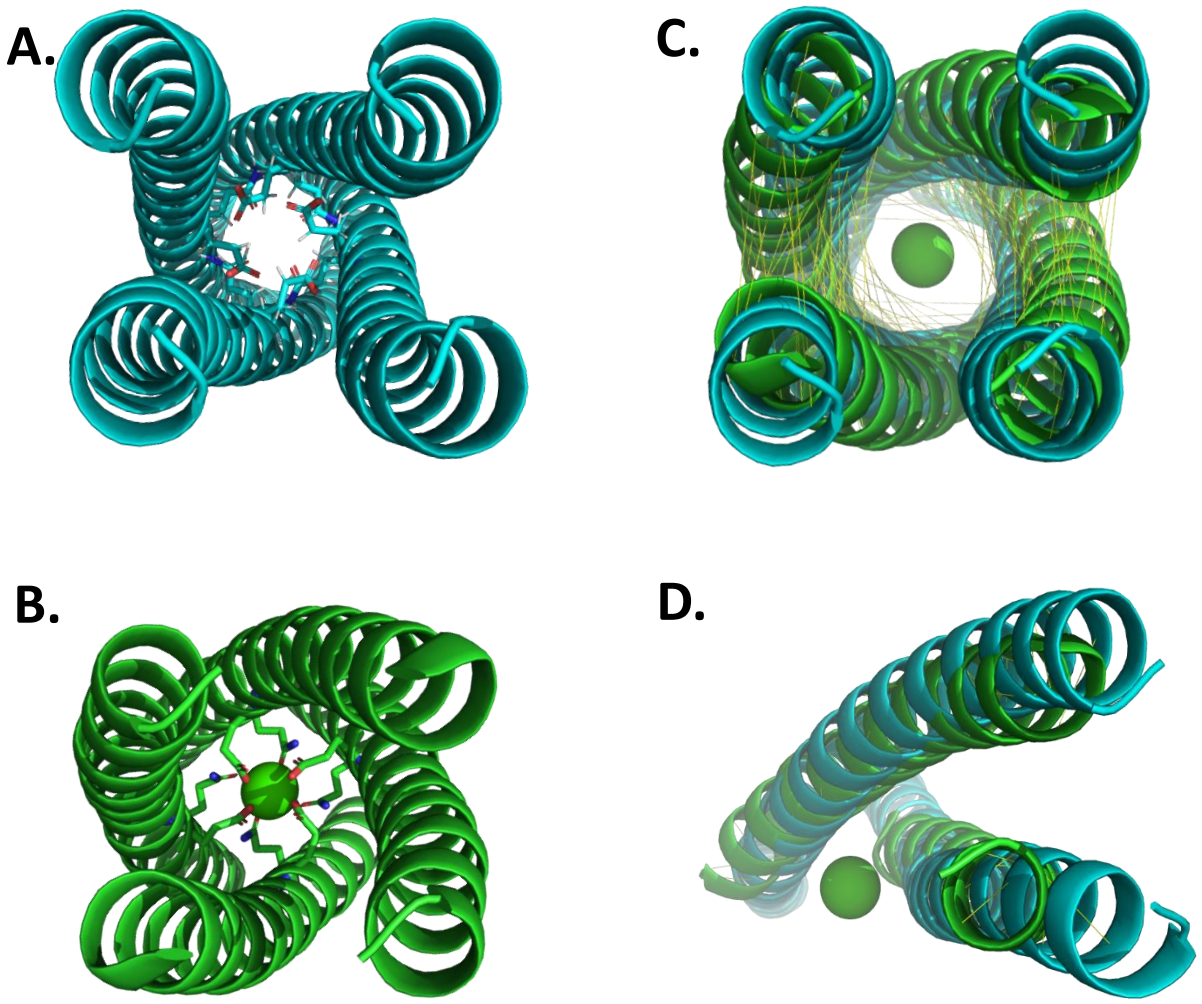

Figure 4. Top-down views of CCD of tetrameric models of WT rotavirus SA11 NSP4. *(A) AlphaFold-Multimer model of NSP4 CCD in cyan. Rendering of predicted residue location/orientation of Ca2+ binding site included. (B) Crystallographic model of NSP4 CCD in green (PDB: 4WB4). Crystallographic models depict a Ca^{2+} ion about the Ca^{2+} binding site shown in green. This panel also includes renderings of the residues at the predicted Ca^{2+} binding site. (C) Alignment of both tetrameric models resulted in an RMSD value of 10.79 Å. Homologous molecules are connected by yellow lines. (D.) Alignment of split-state of tetrameric models resulted in an RMSD value of 3.844 Å. Models were rendered using PyMOL Molecular Graphics System, Version 2.0.*

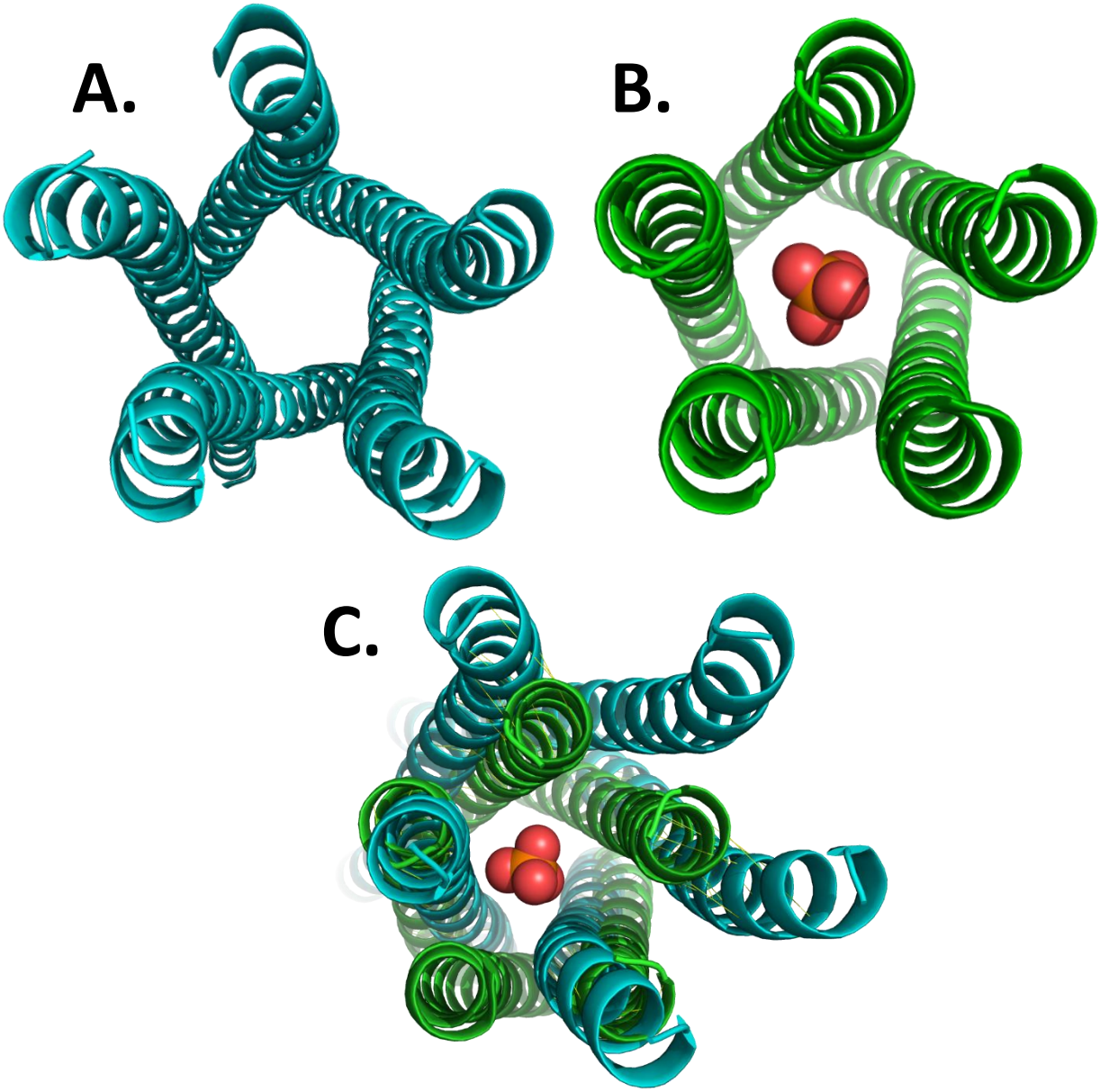

Figure 5. Top-down views of CCD of pentameric models of rotavirus SA11 NSP4. *(A) AlphaFold-Multimer model of WT SA11 NSP4 CCD colored in cyan (B) Crystallographic model of Q/E mutant SA11 NSP4 CCD colored in green (PDB:4WBA). This model depicts a phosphate molecule shown in red, as a phosphate buffer was required to stabilize the pentamer. (C) Alignment of the two models resulted in an RMSD of 3.848. Homologous molecules are connected by yellow lines. Models were rendered using PyMOL Molecular Graphics System, Version 2.0.*

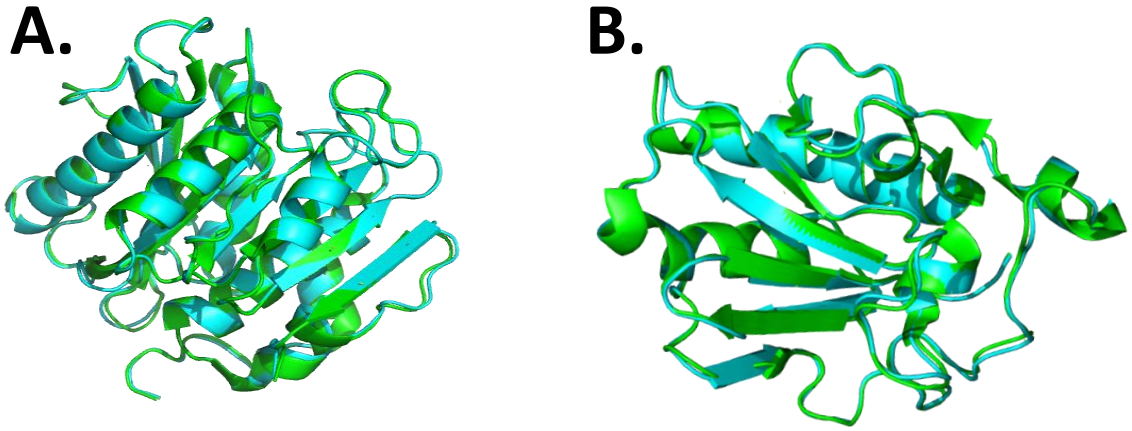

Figure 6. Alignments of AlphaFold models (cyan) and crystal structures (green) of: *(A) polyethylene terephthalate degrading hydrolase (PETase) from Ideonella sakaiensis, RMSD = 0.304 Å, (PDB: 6EQF) and (B) Human Sonic Hedgehog protein, RMSD = 0.326 Å, (PDB: 3M1N). Homologous molecules are connected by yellow lines. Models were rendered using PyMOL Molecular Graphics System, Version 2.0.*

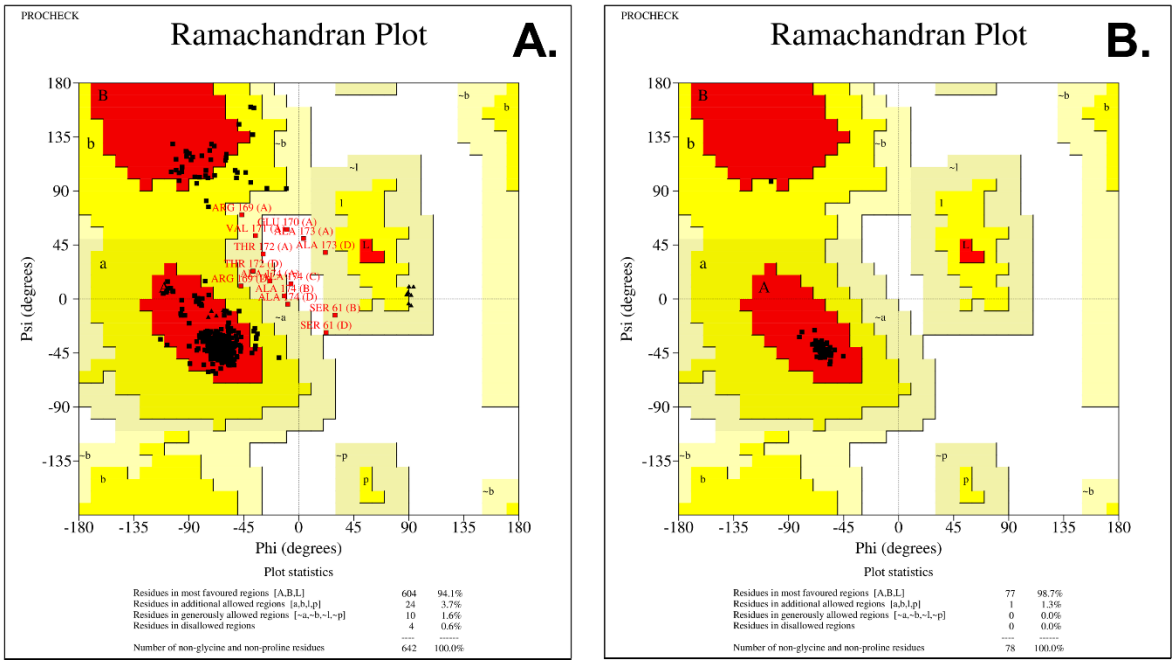

Figure 7. Ramachandran plot of: *(A) AlphaFold-Multimer model of CCD of tetrameric WT SA11 NSP4 and (B) crystal structure of CCD of tetrameric WT SA11 NSP4. Each dot corresponds to a single amino acid within the structure. Triangles represent glycine residues, and squares represent every other type of amino acid. Plots were rendered using PROCHECK with the above listed protein PDB files as input data (24, 25).*

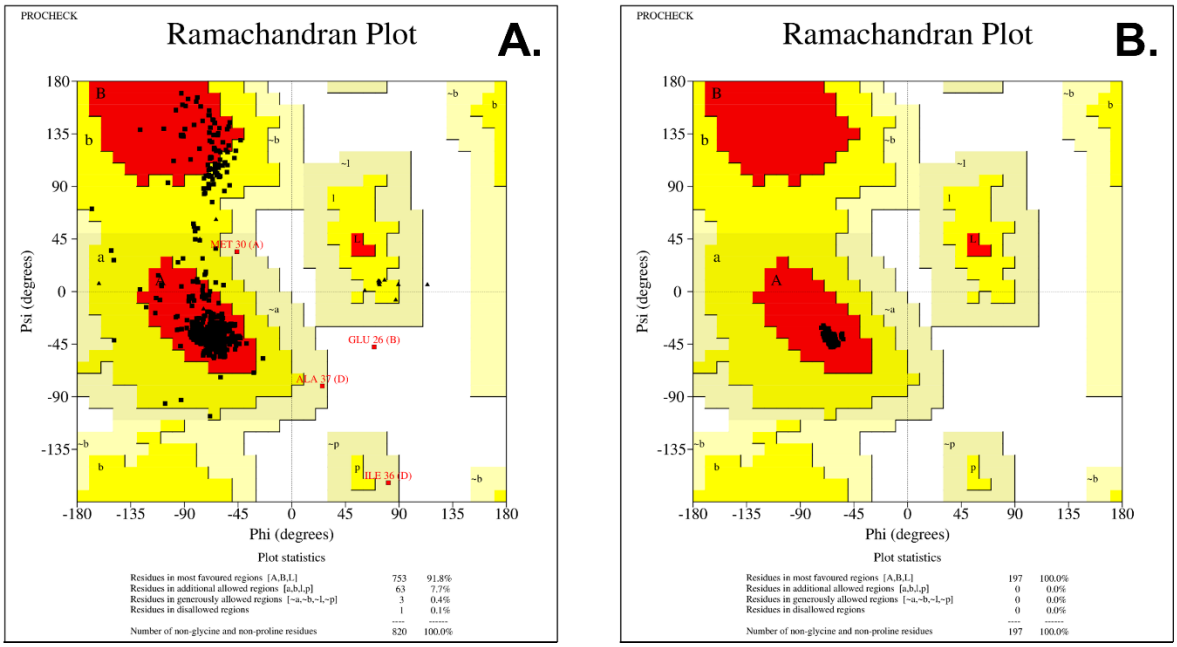

Figure 8. Ramachandran plot of: *(A) AlphaFold-Multimer model of CCD of pentameric WT SA11 NSP4 and (B) crystal structure of CCD of pentameric Q/E mutant SA11 NSP4. Each dot corresponds to a single amino acid within the structure. Triangles represent glycine residues, and squares represent every other type of amino acid. Plots were rendered using PROCHECK with the above listed protein PDB files as input data (24, 25).*

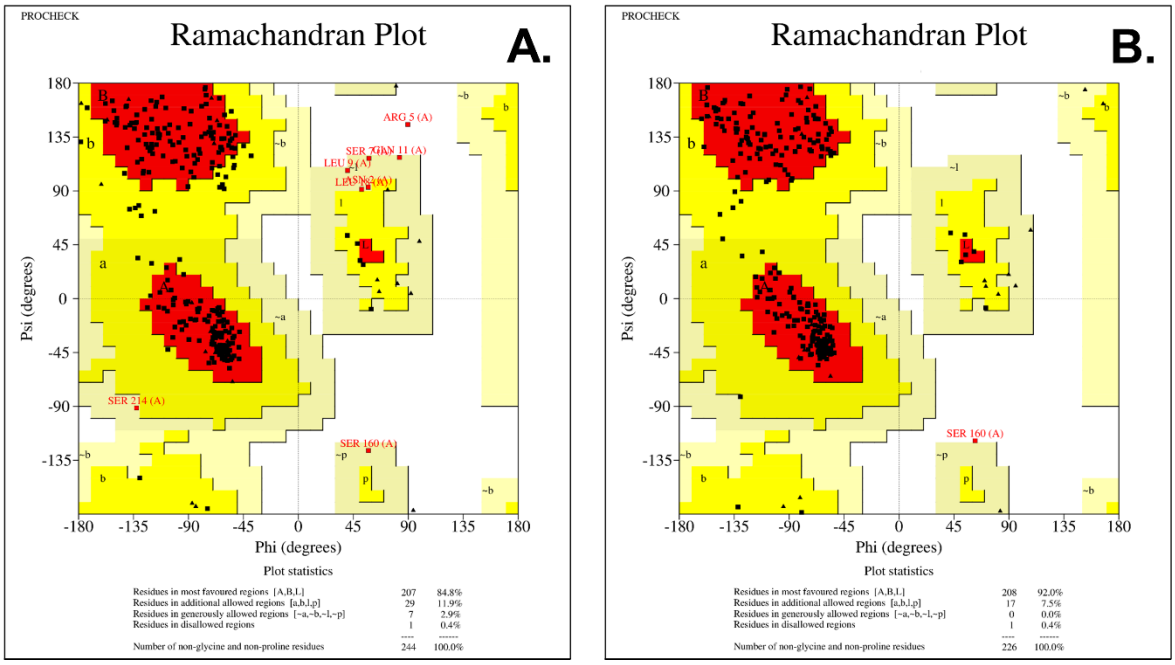

Figure 9. Ramachandran plot of: *(A) AlphaFold2 model of PETase and (B) crystal structure of PETase. Each dot corresponds to a single amino acid within the structure. Triangles represent glycine residues, and squares represent every other type of amino acid. Plots were rendered using PROCHECK with the above listed protein PDB files as input data (24, 25).*

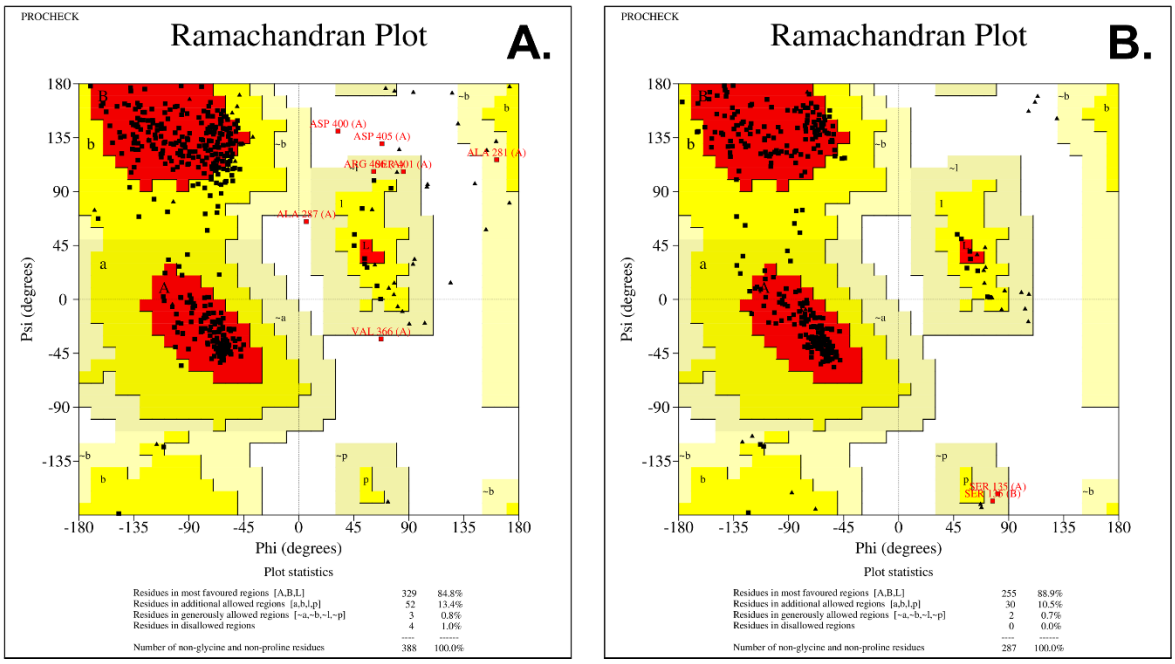

Figure 10. Ramachandran plot of: *(A) AlphaFold2 model of Sonic Hedgehog protein and (B) crystal structure of Sonic Hedgehog protein. Each dot corresponds to a single amino acid within the structure. Triangles represent glycine residues, and squares represent every other type of amino acid. Plots were rendered using PROCHECK with the above listed protein PDB files as input data (24, 25).*

To showcase AlphaFold's remarkable predictive capabilities, AlphaFold models were generated and compared with their crystal structures for a prokaryotic and eukaryotic protein. For this analysis, PETase from *Ideonella sakaiensis* and Human Sonic Hedgehog protein were chosen. The resulting RMSD values for PETase and Sonic Hedgehog were 0.304 Å and 0.380 Å respectively. Though homologous molecules are still connected by yellow lines, discrepancies between the predicted and experimental models are so minute that these lines are effectively invisible. This is supported by the Ramachandran plots of the models, showing an exceptional degree of similarity in both cases (Figures 9, 10). Though the AlphaFold models have more

residues in structurally disallowed regions, some stereochemical violations are to be expected in unrefined structure models.

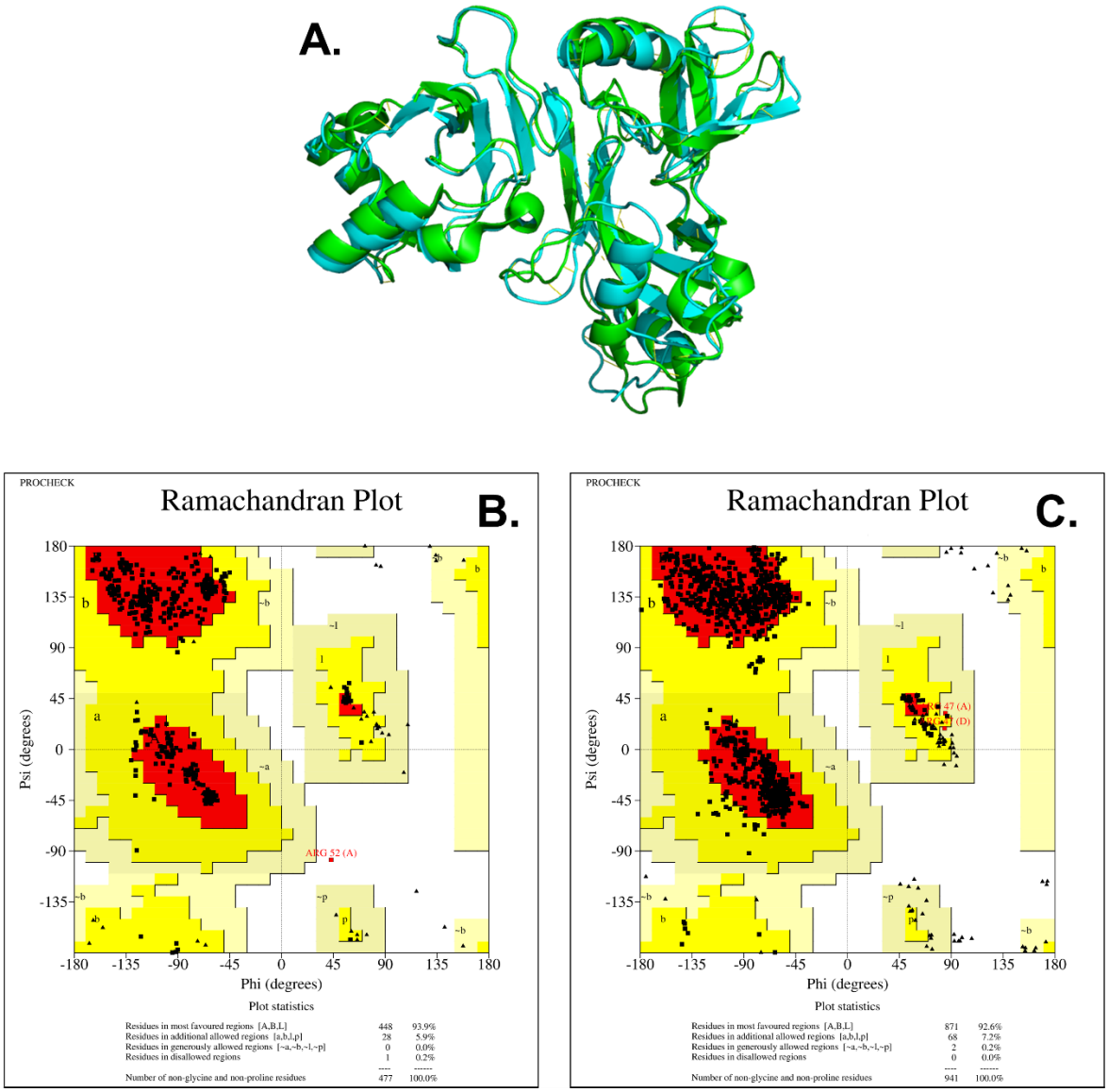

Figure 11. *(A) Alignment of MD-refined AlphaFold2 model (cyan) and crystal structure (green) of N-terminal SARS-CoV-2 NSP2 resulted in an RMSD value of 1.617 Å (PDB: 7EXM). Models were rendered using PyMOL Molecular Graphics System, Version 2.0. Homologous molecules are connected by yellow lines. (B) Ramachandran plot of MD-refined AlphaFold2 model of N-terminal SARS-CoV-2 NSP2. (C) Ramachandran plot of crystal structure of N-terminal SARS-CoV-2 NSP2. Each dot corresponds to a single amino acid within the structure. Triangles represent glycine residues, and squares represent every other type of amino acid. Plots were rendered using PROCHECK with the above listed protein PDB files as input data (24, 25).*

Because (17) has not yet been peer-reviewed, we compared their MD simulation-refined AlphaFold2 model (Figure 11B) of N-terminal SARS-CoV-2 NSP2 with the recently published crystal structure (Figure 11C) (21). An alignment of the two models resulted in an RMSD value of 1.617 Å (Figure 11A). At this resolution, yellow lines connecting homologous molecules are difficult to see, indicating that the two structures are quite similar. This is supported by their respective Ramachandran plots which appear to also be alike. Though 0.2% of residues in the AlphaFold model are in structurally disallowed regions compared to 0% of residues in the crystal structure, the AlphaFold model actually has 1.3% more residues in the most favored regions of the Ramachandran plot (highlighted in red).

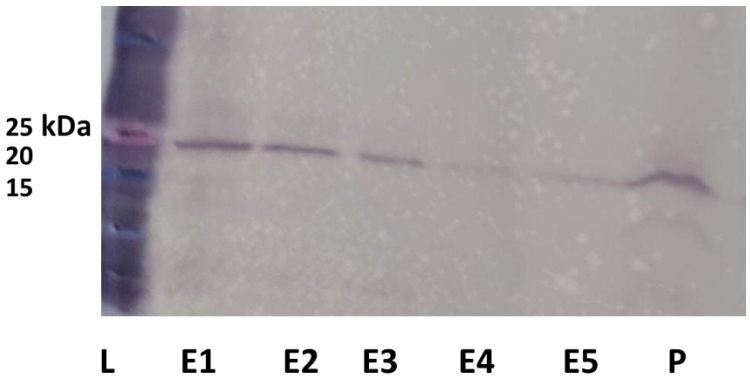

Figure 12. Western blot analysis of whole-cell lysates containing: *10X concentrated samples of full-length WT SA11 NSP4 (13.3 kDa) loaded onto 14-30% gradient SDS-PAGE gels to resolve proteins. Gels were transferred for Western blot analysis where proteins were detected using an a-His6X primary antibody (1:10,000). Bands were visualized with an NBT substrate for the alkaline phosphatase-tagged secondary antibody (1:10,000). Ladder and pellet are labeled L and P respectively. Serial pellet extracts are denoted as "E#" in the order that they were generated.*

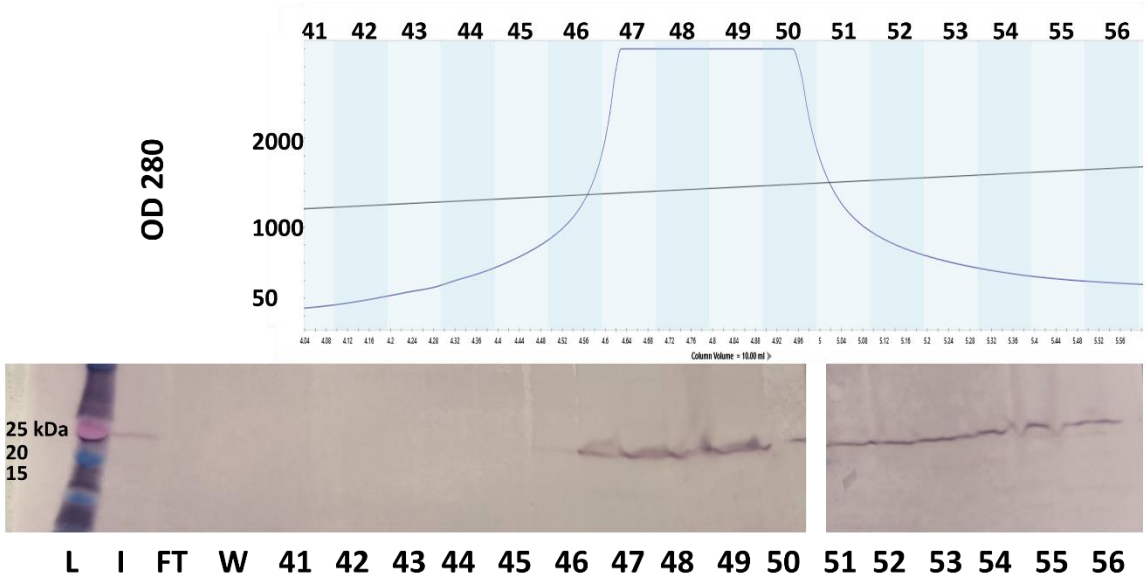

Figure 13. Chromatogram and western blot of purified whole cell lysates containing: *10X concentrated full-length WT SA11 NSP4 (13.3 kDa) samples which were loaded onto 14-30% gradient SDS-PAGE gels to resolve proteins. Gels were transferred for Western blot analysis where proteins were detected using an a-His6X primary antibody (1:10,000). Bands were visualized with an NBT substrate for the alkaline phosphatase-tagged secondary antibody (1:10,000). Fractions of the chromatogram (shown at the top) correspond with the wells on the two pictured western blots. Ladder, Input, Flow Through, and Wash are labeled L, I, FT, and W respectively. Eluate fractions collected post column wash are designated with numbers that correspond to the volume in ml. Fraction numbers from the Western blot correspond to the fraction numbers from the chromatogram above showing total protein elution post wash.*

As mentioned in the introduction, the original aim of this project was to compare SA11 NSP4 structures generated via *in silico* methods with our own crystallographic structures of NSP4. We were unable to determine the crystallization conditions of NSP4 in the given timeframe and have since shifted the goal of this study. Despite this roadblock, there was still a significant amount of progress made towards the crystallization portion of the project. Emily-Claire Duffy of the Banks lab proposed the use of the detergent n-dodecyl-B-D-maltoside (DDM) in our extraction protocol of NSP4. DDM has a long alkyl chain, making it a relatively mild detergent effective at extracting membrane proteins (23). However, it is worth noting that structures solved in DDM are often lower in resolution compared to proteins extracted with short-chain detergents. A new extraction protocol using DDM was devised and scaled up by Osceola Heard, another member of the Banks lab. These findings were essential in extracting quantities of NSP4 required for crystallographic analysis. Using the Duffy-Heard extraction protocol, we were able to obtain cell membrane extracts containing high concentrations of NSP4 (Figure 12.). NSP4 was then purified from these extracts via fast protein liquid chromatography (FPLC). Presence of purified NSP4 in our fractions was confirmed with a chromatogram and western blot analysis (Figure 13.). Due to irregularities in our gel electrophoresis process, fractions eluted at a higher kDa than expected. Our purified samples of NSP4 were then sent to Creative Biostructure for crystallization condition determination.

Discussion

Though the training dataset of both AlphaFold2 and trRosetta exclude viral proteins, the conservation of structural motifs in ion channels across the domains of life can still allow for meaningful modeling of viral proteins excluding polyproteins. We recognize that these algorithms are not intended for use on viral proteins, and that our models represent the limits of what machine-learning based methods are currently capable of. Previous attempts at modeling SARS-CoV-2 proteins using AlphaFold required refinement based on molecular dynamics simulations (17). Any similarities between the experimental and predicted models are indicative of the algorithm's ability to predict protein structures that are not represented in its training dataset.

Only limited insight can be drawn from modeling monomeric NSP4, as this protein is most functional as a tetramer or pentamer. It is likely that NSP4 only exhibits structural stability in these two oligomeric states, as only the tetrameric and pentameric forms have been crystallized. The direct comparison between the AlphaFold2 and trRosetta models shed light on which algorithm is better suited for modeling viroporins. Both machine-learning models deviated from the proposed domains of monomeric NSP4 (Figure 3). However, the trRosetta model was unable to predict the CCD as a coherent motif, and instead split the entire protein into a series of ~26 amino acid long alpha helices. Conversely, the AlphaFold model predicted the CCD to be one long alpha helix spanning from residues 92-139. This is remarkably close to previously the proposed location of the CCD; residues 95-137. This demonstrates AlphaFold's ability to recognize structural motifs within protein sequences that are not included within its training dataset. When aligning the two structures (Figure 2, panel C), the resulting

RMSD structure was a staggering 23.70 Å. RMSD values above 4 Å are generally considered meaningless. Though trRosetta is undoubtedly a cutting-edge tool in protein structure determination, these results speak to this algorithm's relative inability to predict the structure of proteins underrepresented in its training dataset.

Both of our AlphaFold-Multimer models were trimmed, excluding the VPD and residues 162-175. This was done in order to reduce computational cost and to compare the predicted vs. established structures of the CCD. However, as the results from Figure 2 may have indicated, these excluded regions also contained the most problematic regions of the predicted models. The predicted VPD of both multimeric forms of NSP4 included numerous stereochemical violations mainly consisting of overlapping helices. Residues 162-175 were also excluded as AlphaFold predicted this region to be largely unstructured. Unstructured regions are generally meaningless in determining structure/function relationships. The initial alignment of the two models resulted in a discouragingly high RMSD value of 10.79 Å (Figure 4, panel C). However, comparing the split-state of these two models resulted in a much more reasonable RMSD value of 3.844 Å (Figure 4, panel D). Split-state models are another method to reduce the computational cost of prediction and allow for more refined domain analysis. Though this RMSD value is not within the optimal range of <3 Å, these results show that the prediction is somewhat similar to the experimental model, albeit with many sidechains placed with the wrong rotamer. This is made apparent in panels A and B of Figure 4, in which we see residues in the binding site in the same location, but in different orientations. This is also supported by the Ramachandran plot of the two models (Figure 7). Because both models are comprised by alpha-helices (Figure 4), we should

expect to see exclusively helical character in these plots. Although this is the case for the crystal structure, the AlphaFold-Multimer model has a significant number of residues with beta-pleated sheet characteristics. Furthermore, 0.6% of residues in the AlphaFold model are placed in disallowed regions compared to 0% of the residues in the crystal structure. This suggests that the AlphaFold-multimer model struggles to predict accurate torsional angles of the backbone for many of its residues. However, since most of the residues have right-handed helical characteristics, the predicted model is clearly not entirely wrong.

When aligning the pentameric models of NSP4, we obtained an RMSD value of 3.848 Å, a mere 0.004 Å higher than that of the split-state tetramers (Figure 5C). These results indicate a high level of precision in the AlphaFold framework. This slight discrepancy in RMSD values is possibly attributable to the different mutant of NSP4 used in the crystallization of pentameric NSP4. The Q/E mutant of NSP4 was required for this crystallization process. Therefore, this demonstrates AlphaFold's potential to predict the structure of WT proteins without the need for mutagenesis. That being said, the pentameric AlphaFold model also struggled to predict accurate torsional angles of the backbone (Figure 8).

The modest RMSD values and problematic Ramachandran plots of the AlphaFold-Multimer models reveal that there is much work to be done before machine-learning models alone can challenge crystallographic analysis for viral proteins. This is hardly surprising given that predicting the structure of viral proteins is not AlphaFold's intended purpose (yet). To demonstrate AlphaFold's predictive capabilities for eukaryotic and prokaryotic proteins, "celebrity" proteins petASE from *Ideonella*

sakaiensis and Human Sonic Hedgehog protein were selected. The resulting RMSD values were a stunning 0.304 Å and 0.326 Å respectively. It is widely accepted that RMSD values of 0.5-1.5 have such high resolution that there are unlikely to be many errors. To obtain RMSD values under this range puts into question which structure is closer to the proteins true native state. The Ramachandran plots of the predicted and experimental models sheds some light on this question (Figures 9 and 10). Though these pairs of plots are doubtlessly similar to one another, the AlphaFold models tend to have more residues in disallowed regions and less residues in the most favorable regions when compared to crystal structures. However, there may be a solution to AlphaFold's inaccuracies.

The methodology laid out by (17) in which AlphaFold2 models are refined with MD simulations has been devised so recently that its findings have not yet been peer reviewed. To test these findings, we compared the MD-refined AlphaFold2 model of SARS-CoV-2 N-terminal NSP2 with the recently published corresponding crystal structure. An alignment of the two models resulted in an RMSD value of 1.617 Å, suggesting a high degree of structural similarity between them (Figure 11A). This is supported by the Ramachandran plots of the two models (Figure 11B and C), suggesting that the torsional angles of their backbones are consistent with one another. Interestingly, the MD refined AlphaFold model (Figure 11B) had 1.3% more residues in the most favorable regions than the crystal structure (Figure 11C). This suggests that properly refined AlphaFold models significantly reduce incorrectly placed residues, and can place residues in the most favored regions more so than crystal structures.

This demonstrates the near experimental-grade accuracy of MD-refined AlphaFold2 protein models.

Machine learning based methods are a highly enticing alternative to slower, more expensive experimental methods. However, there remain some constraints to machine learning models. These methods rely on template-based modeling. As a result, predictions largely focus on the average structure from a library of homologous sequences. This approach results in an inability to accurately capture differences in structure packing because of varying sidechains. This is reflected in figures 9 and 10. Physics-based protein model refinement methods can address these shortcomings using conformational sampling around the predicted structures. Molecular dynamics, a subset of protein model refinement, is one of the most successful methods for refining dubious structural features (18). MD-refined AlphaFold2 structures allowed for accurate modeling of SARS-CoV-2 proteins early in the COVID-19 pandemic (17). Unfortunately, until our methodology includes some variation of protein model refinement, our models cannot hold up to the established crystal structures of NSP4.

The advent of accurate and precise structure prediction algorithms marks a new era of proteomics. Here, we demonstrate a methodology in which anyone with a computer and stable internet connection can feasibly generate experimental-grade protein structures in the span of a single afternoon. We also demonstrate that when used properly, structure prediction algorithms can potentially challenge experimentally obtained protein structures. This will result in increased competition between companies capable of these expensive and inaccessible methods of structure determination, hopefully resulting in lower costs and more rapid turnarounds.

Ongoing studies should attempt to refine our models of NSP4. This can first be done by utilizing the full version of AlphaFold. Though the Colab notebook version of AlphaFold is user friendly, it is a simplified version of the full algorithm. Future studies should resume our attempts to run the full version of AlphaFold via the Bates Leavitt HPCC. Our NSP4 models can also be improved by using protein model refinement, namely molecular dynamics simulations. Ongoing studies should also attempt other *in vitro* methods of structure determination such as cryo-EM or NMR due to their relative success in modeling membrane proteins. They may also re-attempt the more challenging task of crystallizing NSP4. Future studies should also explore generating models with RoseTTAFold, another highly accurate open-access algorithm featured at CASP14.

Methods

Computational Modeling

The amino acid sequence of SA11 rotavirus NSP4 was obtained from NCBI, and the sequences of PETase and Sonic Hedgehog protein were obtained from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (PDB). These sequences were inserted into the Colab notebook version of AlphaFoldv2.1.0 and AlphaFold-Multimer in FASTA format. The outputs of the algorithm were downloaded as PDB files. All protein structure models were rendered, and RMSD values generated, using PyMOL Molecular Graphics System, Version 2.0 (Schrödinger, LLC) for detailed analysis. All Ramachandran plots were generated using PROCHECK, a free tool developed by UCLA-DOE that checks the stereochemical quality of a protein structure (24, 25).

The AlphaFold-generated NSP4 models were compared with a monomeric model created last year in the Banks lab via trRosetta, as well as the established crystallographic tetramer and pentamer models (Jeremy Bennett, Unpublished) (9). The PDB codes for NSP4 the tetramer and pentamer are 4WB4 and 4WBA respectively. The AlphaFold-generated PETase and Sonic Hedgehog protein models were compared with their established crystal structures. The PDB codes for the crystal structures of PETase and Sonic Hedgehog protein are 6EQF and 3M1N respectively (19) (20).

Overexpression of NSP4

We completed bacterial cell culturing as previously described by (6). Plasmids encoding WT SA11 NSP4 residues 47-146 (accession number AF087678.1) were

transformed into BL21 (DE3) pLysS *E. coli* cells (Promega) via heat shock in a 42°C hot water bath for 30 seconds, and then recovered at 37°C for 1 hour prior to plating onto LB broth (Research Products International) containing 1% glucose, 50 mg/mL carbenicillin, and 37 mg/mL chloramphenicol. Single colonies were picked, and grown overnight at 37°C in LB containing 1% glucose, 50 mg/mL carbenicillin, and 37 mg/mL chloramphenicol plates and shaken at 200 rpm overnight. The overnight culture was diluted into fresh media and grown until the optical density (OD) at 600 nm reached the range of 0.4-0.6. At this point, the culture was induced with 1M Isopropyl ß-D-1-thiogalactopyranoside (IPTG) for 2-3 hours. The culture was then divided into 250 mL bottles and centrifuged at 11,000 rpm for 10 minutes. The supernatant was discarded, and the pellets were stored at -20°C for detergent extraction of NSP4 from membrane fragments using a method initially developed by Dr. Lori Banks and optimized by Emliy-Claire Duffy and Osceola Heard (other Banks Lab members).

Purification of NSP4

Where specified, cell pellets were resuspended in 1X Phosphate Buffered Saline (PBS) containing 10 mM imidazole 1X DDM and stirred on ice. Suspensions were centrifuged at 10,000 rpm for 15 minutes at 4˚C, and the extracted proteins in the supernatant were collected. The extraction process was repeated two more times. Lysate samples were centrifuged at 10,000 rpm for 10 minutes and combined for loading onto a nickel-charged IMAC column run on our NGC Protein Purification System (BioRad). Later purification runs employed the same protocol with the exception that 20 mM Tris pH 7.4, containing 10 mM imidazole and 1X DDM, with either 100 mM or 500 mM NaCl, as the extraction buffer. Pooled extracts were filtered and then loaded onto a

5mL nickel-charged IMAC column at 5mL/min, washed with 20 mM Tris pH 7.4, containing 10 mM imidazole and 1X DDM, with either 100 mM or 500 mM NaCl. Proteins were eluted using a linear gradient of 20 mM Tris pH 7.4, containing 500mM imidazole and 1X DDM, with either 100mM or 500 mM NaCl, at a flow rate of 5mL/min. 1 mL fractions were collected, and samples taken for Coomassie and Western blot analysis to determine NSP4-containing fractions and protein purity.

Western Blot Analysis of NSP4

Extract and pellet fractions from the above protocol were then prepared for western blot analysis to specifically detect the fractional location of our NSP4 construct. 10X concentrated whole-cell lysates containing WT SA11 NSP4 residues 47-146 (predicted monomer size, 13.3 kDa) were loaded onto 4-20% Tris-glycine gradient SDS-PAGE gels to resolve proteins. Gels were either stained with BioSafe Coomasie solution to see total protein, or transferred for Western blot analysis using a semi-dry method, where proteins were detected using an a-His6X primary antibody (1:10,000). Bands were visualized with a 5-bromo-4-chloro-3-indolyl phosphate (BCIP)/nitro blue tetrazolium (NBT) substrate for the alkaline phosphatase-tagged secondary antibody (1:10,000).

Works Cited

1. Troeger, C., Blacker, B. F., Khalil, I. A., Rao, P. C., Cao, S., Zimsen, S. R. M., Albertson, S. B., Stanaway, J. D., Deshpande, A., Abebe, Z., Alvis-Guzman, N., Amare, A. T., Asgedom, S. W., Anteneh, Z. A., Antonio, C. A., Aremu, O., Asfaw, E. T., Atey, T. M., Atique, S., Avokpaho, E. F., Awasthi, A., Ayele, H. T., Barac, A., Barreto, M. L., Bassat, Q., Belay, S. A., Bensenor, I. M., Bhutta, Z. A., Bijani, A., Bizuneh, H., Castañeda-Orjuela, C. A., Dadi, A. F., Dandona, L., Dandona, R., Do, H. P., Dubey, M., Dubljanin, E., Edessa, D., Endries, A. Y., Eshrati, B., Farag, T., Feyissa, G. T., Foreman, K. J., Forouzanfar, M. H., Fullman, N., Gething, P. W., Gishu, M. D., Godwin, W. W., Gugnani, H. C., Gupta, R., Hailu, G. B., Hassen, H. Y., Hibstu, D. T., Ilesanmi, O. S., Jonas, J. B., Kahsay, A., Kang, G., Kasaeian, A., Khader, Y. S., Khalil, I. A., Khan, E. A., Khan, M. A., Khang, Y.-H., Kissoon, N., Kochhar, S., Kotloff, K. L., Koyanagi, A., Kumar, G. A., Magdy Abd El Razek, H., Malekzadeh, R., Malta, D. C., Mehata, S., Mendoza, W., Mengistu, D. T., Menota, B. G., Mezgebe, H. B., Mlashu, F. W., Murthy, S., Naik, G. A., Nguyen, C. T., Nguyen, T. H., Ningrum, D. N., Ogbo, F. A., Olagunju, A. T., Paudel, D., Platts-Mills, J. A., Qorbani, M., Rafay, A., Rai, R. K., Rana, S. M., Ranabhat, C. L., Rasella, D., Ray, S. E., Reis, C., Renzaho, A. M. N., Rezai, M. S., Ruhago, G. M., Safiri, S., Salomon, J. A., Sanabria, J. R., Sartorius, B., Sawhney, M., Sepanlou, S. G., Shigematsu, M., Sisay, M., Somayaji, R., Sreeramareddy, C. T., Sykes, B. L., Taffere, G. R., Topor-Madry, R., Tran, B. X., Tuem, K. B., Ukwaja, K. N., Vollset, S. E., Walson, J. L., Weaver, M. R., Weldegwergs, K. G., Werdecker, A., Workicho, A., Yenesew, M., Yirsaw, B. D., Yonemoto, N., El Sayed Zaki, M., Vos, T., Lim, S. S., Naghavi, M., Murray, C. J. L., Mokdad, A. H., Hay, S. I., and Reiner, R. C. (2018) Estimates of the Global, regional, and national morbidity, mortality, and aetiologies of diarrhoea in 195 countries: A systematic analysis for the global burden of disease study 2016. *The Lancet Infectious Diseases*. **18**, 1211–1228

2. Bernstein, D. I. (2009) Rotavirus overview. *Pediatric Infectious Disease Journal*. 10.1097/inf.0b013e3181967bee

3. Crawford, S. E., Ramani, S., Tate, J. E., Parashar, U. D., Svensson, L., Hagbom, M., Franco, M. A., Greenberg, H. B., O'Ryan, M., Kang, G., Desselberger, U., and Estes, M. K. (2017) Rotavirus infection. *Nature Reviews Disease Primers*. 10.1038/nrdp.2017.83

4. Parashar, U. D., Gibson, C. J., Bresee, J. S., and Glass, R. I. (2006) Rotavirus and severe childhood diarrhea. *Emerging Infectious Diseases*. **12**, 304–306

5. Baker, M., and Prasad, B. V. V. (2013) Rotavirus Cell Entry. in *Cell entry by non-enveloped viruses*, Springer Berlin, Berlin

6. Hyser, J. M., Collinson-Pautz, M. R., Utama, B., and Estes, M. K. (2010) Rotavirus disrupts calcium homeostasis by NSP4 viroporin activity. *mBio*. 10.1128/mbio.00265-10

7. Gonzalez, M. E., and Carrasco, L. (2003) Viroporins. *FEBS Letters*. **552**, 28–34

8. Pham, T., Perry, J. L., Dosey, T. L., Delcour, A. H., and Hyser, J. M. (2017) The rotavirus NSP4 viroporin domain is a calcium-conducting ion channel. *Scientific Reports*. 10.1038/srep43487

9. Sastri, N. P., Viskovska, M., Hyser, J. M., Tanner, M. R., Horton, L. B., Sankaran, B., Prasad, B. V., and Estes, M. K. (2014) Structural plasticity of the coiled-coil domain of rotavirus NSP4. *Journal of Virology*. **88**, 13602–13612

10. Bennett, A., Bar-Zeev, N., and Cunliffe, N. A. (2016) Measuring indirect effects of rotavirus vaccine in low income countries. *Vaccine*. **34**, 4351–435

11. Dill, K. A., Ozkan, S. B., Shell, M. S., and Weikl, T. R. (2008) The protein folding problem. *Annual Review of Biophysics*. **37**, 289–316

12. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021) Highly accurate protein structure prediction with alphafold. *Nature*. **596**, 583–589

13. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J., and Hassabis, D. (2021) Protein complex prediction with Alphafold-Multimer. 10.1101/2021.10.04.463034

14. Larry Hardesty | MIT News Office Explained: Neural networks. *MIT News | Massachusetts Institute of Technology*. [online] https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414 (Accessed March 28, 2022)

15. Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. (2020) Improved protein structure prediction using potentials from deep learning. *Nature*. **577**, 706–710

16. Xie, Q., Luong, M.-T., Hovy, E., and Le, Q. V. (2020) Self-training with noisy student improves ImageNet Classification. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10.1109/cvpr42600.2020.01070

17. Heo, L., and Feig, M. (2020) Modeling of severe acute respiratory syndrome coronavirus 2 (SARS-COV-2) proteins by machine learning and physics-based refinement. 10.1101/2020.03.25.008904

18. Heo, L., Arbour, C. F., Janson, G., and Feig, M. (2021) Improved sampling strategies for protein model refinement based on molecular dynamics simulation. *Journal of Chemical Theory and Computation*. **17**, 1931–1943

19. Austin, H. P., Allen, M. D., Donohoe, B. S., Rorrer, N. A., Kearns, F. L., Silveira, R. L., Pollard, B. C., Dominick, G., Duman, R., El Omari, K., Mykhaylyk, V., Wagner, A., Michener, W. E., Amore, A., Skaf, M. S., Crowley, M. F., Thorne, A. W., Johnson, C. W., Woodcock, H. L., McGeehan, J. E., and Beckham, G. T. (2018) Characterization and engineering of a plastic-degrading aromatic polyesterase. *Proceedings of the National Academy of Sciences*. 10.1073/pnas.1718804115

20. Pepinsky, R. B., Rayhorn, P., Day, E. S., Dergay, A., Williams, K. P., Galdes, A., Taylor, F. R., Boriack-Sjodin, P. A., and Garber, E. A. (2000) Mapping sonic hedgehog-receptor interactions by steric interference. *Journal of Biological Chemistry*. **275**, 10995–11001

21. Ma, J., Chen, Y., Wu, W., and Chen, Z. (2021) Structure and function of N terminal zinc finger domain of SARS-COV-2 NSP2. *Virologica Sinica*. **36**, 1104–1112

22. Boshuizen, J. A., Rossen, J. W., Sitaram, C. K., Kimenai, F. F., Simons-Oosterhuis, Y., Laffeber, C., Büller H. A., and Einerhand, A. W. (2004) Rotavirus enterotoxin NSP4 binds to the extracellular matrix proteins laminin-β3 and fibronectin. *Journal of Virology*. **78**, 10045–10053

23. Sonoda, Y., Cameron, A., Newstead, S., Omote, H., Moriyama, Y., Kasahara, M., Iwata, S., and Drew, D. (2010) Tricks of the trade used to accelerate high-resolution structure determination of membrane proteins. *FEBS Letters*. **584**, 2539–2547

24. Laskowski, R. A., MacArthur, M. W., Moss, D. S., and Thornton, J. M. (1993) Procheck: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*. **26**, 283–291

25. Laskowski, R. A., Rullmann, J. A. C., MacArthur, M. W., Kaptein, R., and Thornton, J. M. (1996) Aqua and Procheck-NMR: Programs for checking the quality of protein structures solved by NMR. *Journal of Biomolecular NMR*. 10.1007/bf00228148