Single and Pairwise Mutations and Their Impact on SARS-CoV-2 Proteins

Danny Higgins
Bellingham, United States
higgind5@wwu.edu

Brandon Vogel
Bellingham, United States
vogelb2@wwu.edu

William Tian
Bellingham, United States
tianw@wwu.edu

Kyle Leibowitz
Bellingham, United States
leibowk@wwu.edu

Gideon Wolfe
Bellingham, United States
wolfeg2@wwu.edu

Othmane Belhoussine
Bellingham, United States
belhouo@wwu.edu

Maxwell Lisaius
Bellingham, United States
lisaium@wwu.edu

Anais Dawson
Bellingham, United States
dawsona6@wwu.edu

Khleo Isaguirre
Bellingham, United States
isaguik@wwu.edu

Garrett Strawn
Bellingham, United States
strawng@wwu.edu

ABSTRACT

UPDATED—June 22, 2020. This paper outlines the experiment performed using a new tool called doubleMutation which takes its predecessor, multiMutant, and expands upon it for creating pairwise mutations. This program was developed to analyze the effects that pairwise mutations have on SARS-CoV-2 and compare them to results produced by single mutations.

Author Keywords

COVID-19; SARS-CoV-2

INTRODUCTION

In December 2019 a new respiratory disease appeared in the city of Wuhan, Hubei province, China (1,2). This disease is caused by a member of the Coronaviridae family which is responsible for the current global pandemic which is termed coronavirus disease 2019 (COVID-19) (1,2). The virus responsible has since been sequenced and named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (1,2). Since the outbreak in December the WHO reports that there have been over 6 million cases globally including over 370 thousand deaths (3). To exacerbate issues, economies have taken a downturn around the world and the global supply chains countries rely on have been compromised resulting in shortages of many basic necessities such as food (4).

A pandemic of such great proportions caused by a virus that was unknown to humans as of the end of last year requires a great deal of research to elucidate possible treatments. As such the scientific community across the globe has been hard at work discovering the sequence of the virus, identifying proteins, and researching the use of currently available and possible new drugs to treat the pandemic. The goal of the research presented here is to elucidate the effects of both single and double mutations on the stability of the main protease (Mpro) and the spike proteins of SARS-CoV-2 so as to inform which sequences of amino acids are of importance to this protein's structure and function.

MOTIVATION AND RELATED WORK

The main protease of SARS-CoV-2 is a 306 residue single chain homodimer that is responsible for cleaving polyproteins necessary for the virus to replicate (5,6). One reason we are targeting this enzyme for double point mutations is because of the aforementioned length and single chain nature of the protein. Other researchers are also targeting the main protease because of the fact that there are no human proteases with similar cleavage specificity, therefore drugs targeting this enzyme are unlikely to be toxic to patients (6). Sequences of interest that we focused on for the double point mutations were those that made up a beta sheet of a canonical beta barrel close to an active site His41 at residues 25-32 (6,7,8). We considered this region a good candidate for double point mutations because such a large perturbation of a highly structured region close to the active site may prove to be highly detrimental to the protein's stability and may even affect catalytic efficiency.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

The spike protein is a homotrimer that is located on the surface of the virus and is necessary for viral entry into the host cell. Viruses often go through mutations overtime depending on different environmental factors, so looking at how mutations can affect proteins necessary for viral survival would shed light on critical positions and functions for that protein. For instance in the study conducted by Korber et al., they created an analysis pipeline that looks at different spike mutations (10). The motivation for this certain study has been to find some kind of "early warning sign" for disease based on mutations that enable the virus to have an advantage in terms of transmission and/or resistance from therapeutics.

In another related study by Pachetti et al., they identified regions in the genome of the SARS-CoV-2 virus that are either prone to mutation or are heavily conserved (11). One such region the authors identified was in the RNA dependent RNA polymerase (RdRp) showing that the efficiacy of therapeutics targeting this protein may differ depending on the exact strain of the virus. Additionally, in another related study by Yuan et al., they have identified a conserved region in the receptor binding domain (RBD) of the Spike protein (S) and have targeted this region of supposed importance with a novel antibody. Out of 28 residues in the epitope, 24 are conserved between SARS-CoV-2 and SARS-CoV (12). Research has been heavily focused on the Spike protein (S) of SARS-CoV-2 as this is the protein that is responsible for the interaction with the human ACE2 receptor initializing infection.

METHOD

For single mutations, a set of high interest proteins was selected for mutation analysis from the 127 currently available PDB structures. Priority was given to spike proteins, as this is the main entry point for the virus. Mutations were done on residues that were located within the receptor binding domain (RBD) of the spike protein. This is the area that comes into contact with the ACE2 receptor found on human lung cells. To generate all possible mutants of the proteins, `ProMute` was used. The EM data and stability predictions were generated by NAMD and SDM.

For every location of interest in the PDB structure, `PromuteBatch` was used to generate `ProMute` commands to exhaustively mutate each target residue. A driver `bash` script was used to generate and run these commands, as well as a mutation script used as input for the SDM calculations.

To extract energy minimization (EM) data, a script was developed to compare the energy state of the wildtype structure to the energy state at each successive pass at EM.

For double mutations a smaller data set is needed as when mutating a section of a protein more than once there quickly becomes a huge number of results. Due to this priority was given to amino acids 25 though 32 of the PDB 6Y2E. This section was analyzed using a python script that uses `ProMute` to generate the first mutation and renames the output, then runs ProMute again on this generated data to create a double mutation. During the operations the -em flag was used in order to generate energy step minimization files. These files were then analyzed against the wildtype similar to a single mutation to create several box plots to represent the spread of the data generated in an attempt to look for outliers.

RESULTS

Single Mutation SDM stability prediction Results:
(See Figures 1-6 below)

Double Mutation Box and Whisper Plot Results:
(See Figures 7-9 below)

CONCLUSION

From the SDM results of the single mutations, we observed that the most destabilizing of the sites of the main protease were W31, L32, L30, and L27. The most disruptive mutations made in these locations were Alanine, Aspartate, Proline, and Serine. Of these combinations the one with the most disruptive prediction was a mutation to Proline at site L30. This produced a $\triangle\triangle$G prediction of -4.3. A probable reason for this is that Proline is a residue that makes a structure much more rigid due to the side chain being incorporated into the peptide backbone. L30 is a residue that is part of a β-sheet found in domain I of the main protease. The incorporation of the Proline within the β-sheet would hinder the formation of necessary hydrogen bonds with the other surrounding residues within that β-sheet. In addition, this particular β-sheet is found near domain II of the main protease so it is probable that there are favorable interactions with domain II residues that are being lost with the substitution of the Proline. However, more investigation needs to be done to see if that is actually the case.

The results of the double mutations show that the two most destabilizing amino acids were Proline as well as Trytophan. From the box plots we can see that the amino acid that had the lowest number of steps to achieve a minimized energy, as well as a low average number of steps, is arginine (Fig. 7) which may be a byproduct of arginine being found in either α-helices, -sheets or turns with a relatively equal probability. Proline is a rigid protein that has a very tight turn due to it's side chain being incorporated into the backbone as stated above thus it is usually found in turns. This is indicated by the high average of steps taken (Fig. 8) to achieve minimized energy it will very readily perturb the stabilization of a β-sheet as it interacts very little with the side chains of nearby proteins but bends the sheet away from the β-barrel at the residues tested. That said tryptophan is typically found in β-sheets but also has a high average number of steps (Fig. 9) to achieve energy minimization thus indicating that it likely had unfavorable steric clashes with nearby proteins due to it's large side chain thus so many steps were taken to find a minimal potential energy of that residue.

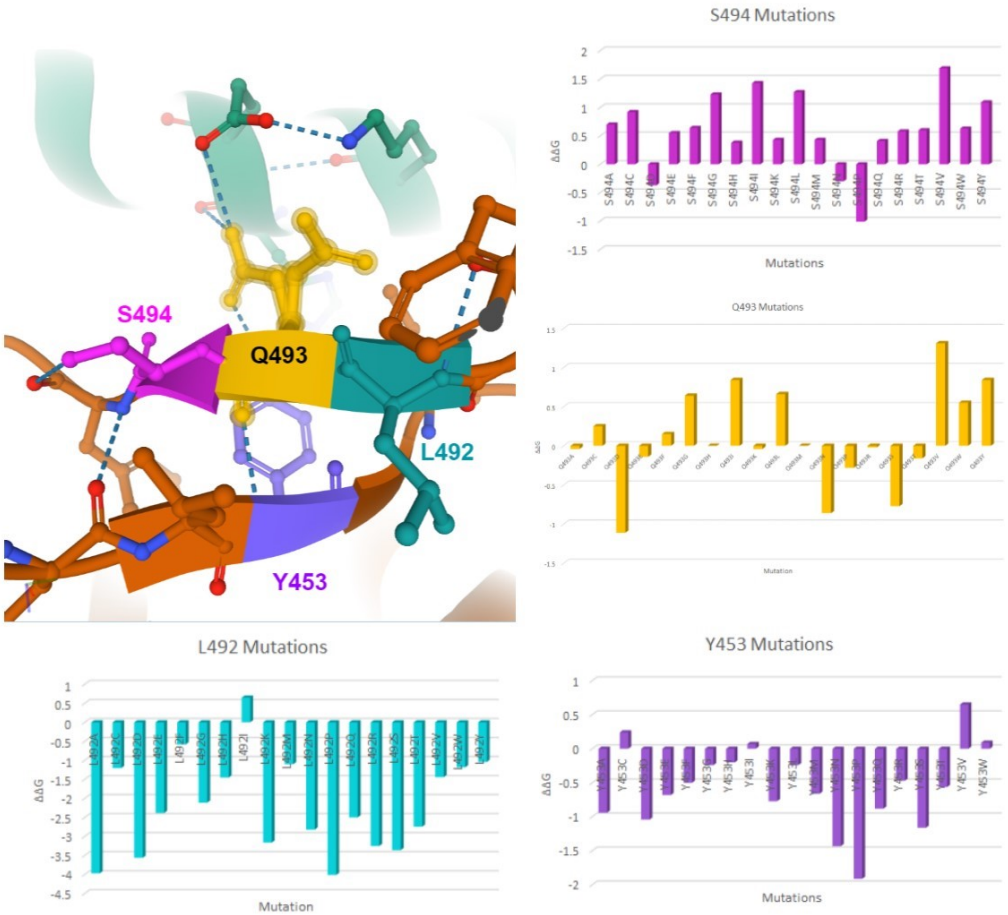

Figure 1. Analysis of exhaustive single amino acid mutations for selected SARS-CoV-2 spike receptor binding domain (RBD) residues located at the interface between the spike RBD and the cell receptor ACE2 (PDB: 6M0J). A). Position of residues within the complex of the spike RBD (orange) and ACE2 (dark green). Relevant side chains are represented by ball and stick and hydrogen bonds between residues are represented by dashed lines. B-E). The effect of each possible amino acid mutation at selected residue on overall complex stability as measured by the change in Gibbs free energy (△△G).

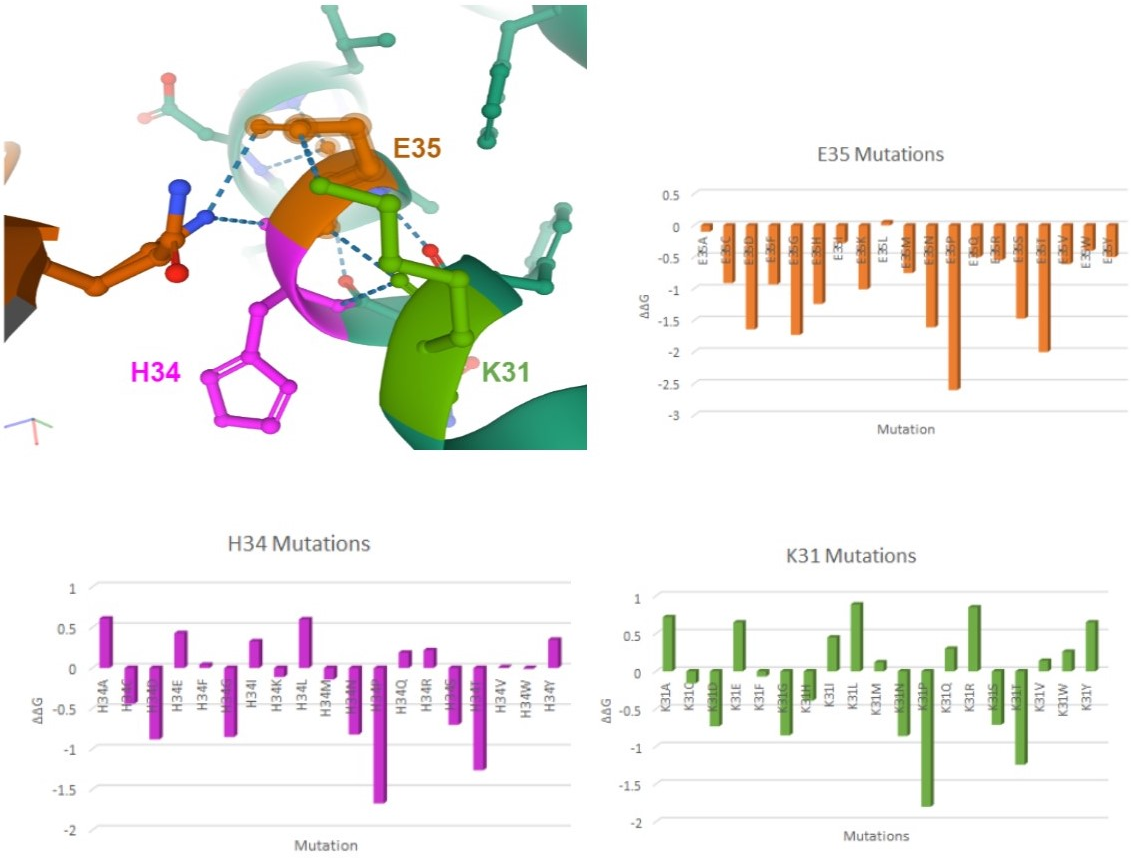

Figure 2. Analysis of exhaustive single amino acid mutations for selected residues in the cell receptor ACE2 located at the interface between the spike RBD and ACE2 (PDB: 6M0J). A). Position of residues within ACE2 (dark green). B-D). The effect of each possible amino acid mutation at selected residue on overall complex stability as measured by the change in Gibbs free energy (△△G).

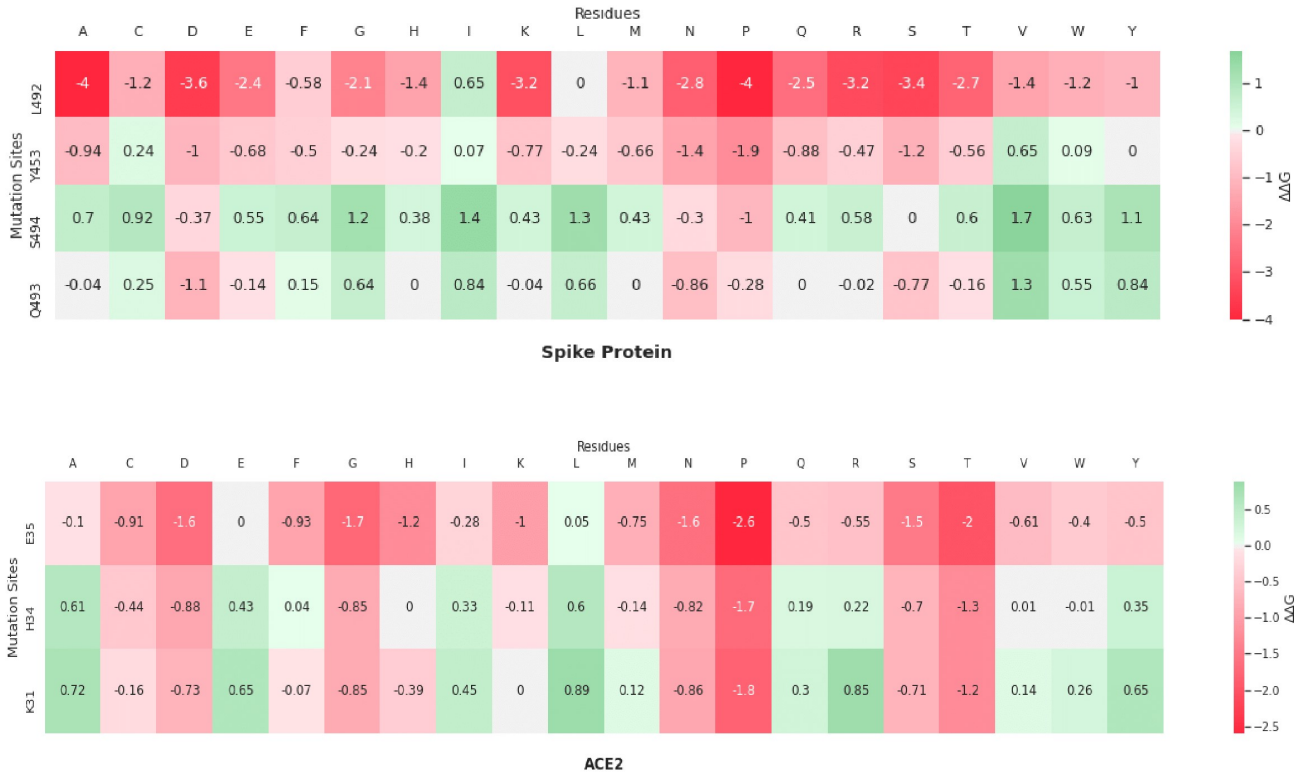

Figure 3. Heatmap visualization of exhaustive single amino acid mutations for all selected mutation sites (Y-axis) in the SARS-CoV-2 spike receptor and the cell receptor ACE2 (PDB: 6M0J). It represents the effect of each possible amino acid mutation at selected residue on protein stability as measured by the change in Gibbs free energy (△△G).

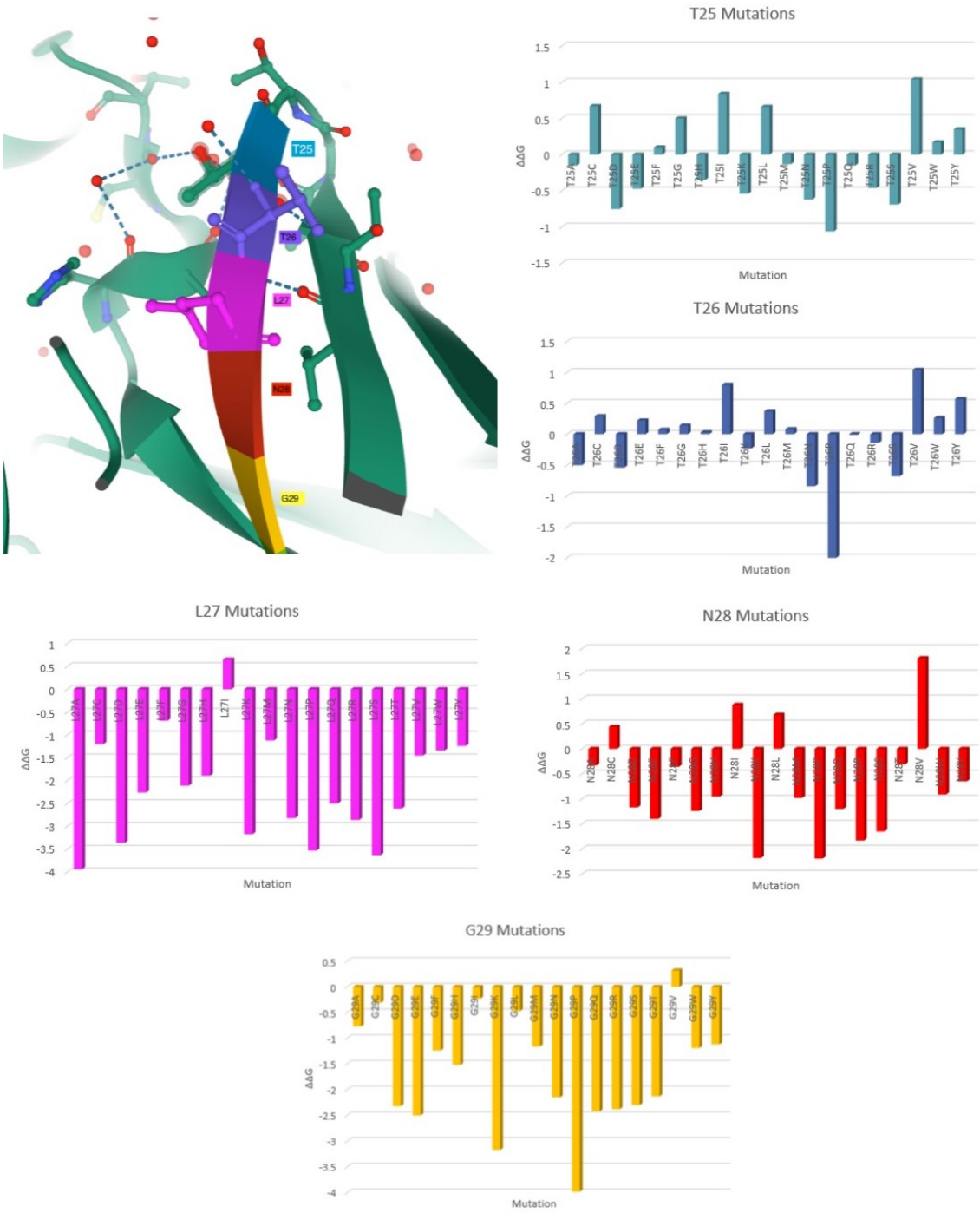

Figure 4. Analysis of exhaustive single amino acid mutations for selected residues in the SARS-CoV-2 main protease (Mpro) (PDB: 6Y2E). A). Position of residues within Mpro (dark green). B-F). The effect of each possible amino acid mutation at selected residue on protein stability as measured by the change in Gibbs free energy (△△G).

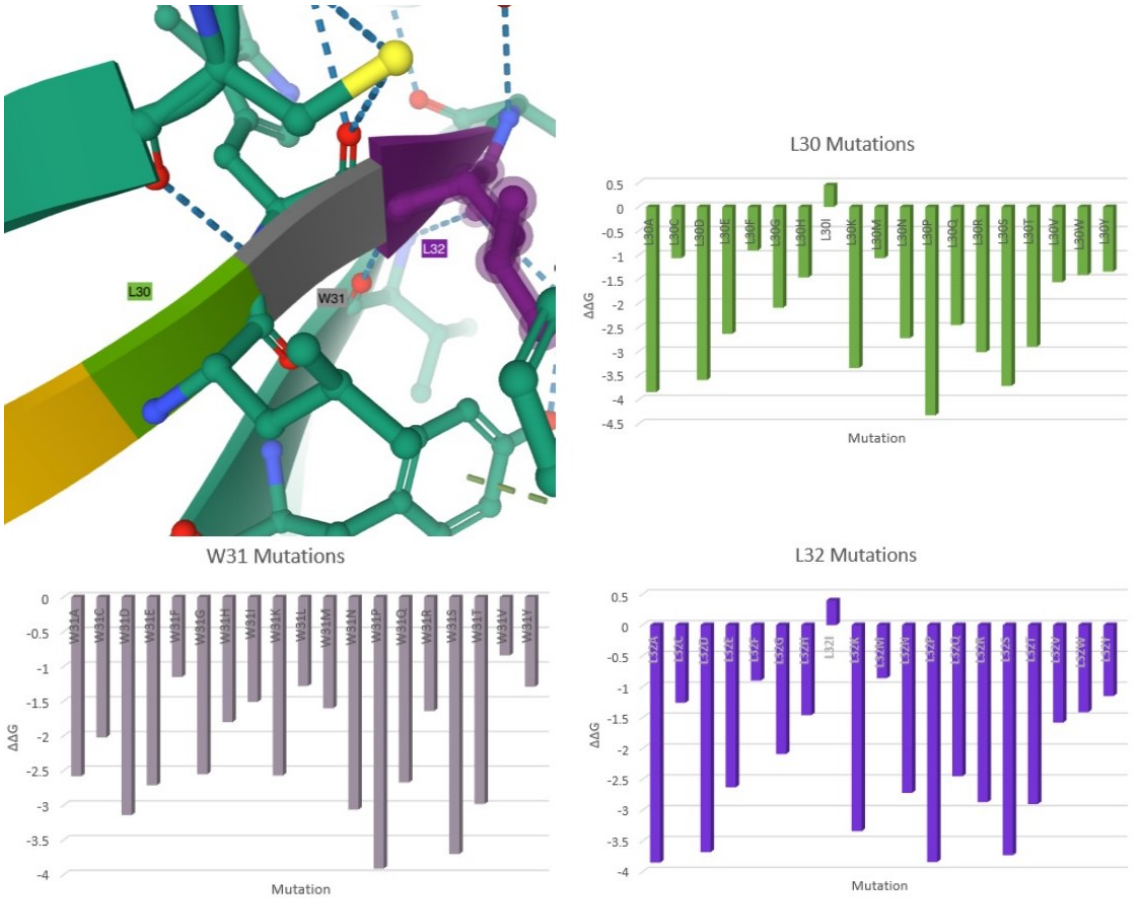

Figure 5. Analysis of exhaustive single amino acid mutations for selected residues in the SARS-CoV-2 main protease (Mpro) (PDB: 6Y2E). A). Position of residues within Mpro (dark green). B-F). The effect of each possible amino acid mutation at selected residue on protein stability as measured by the change in Gibbs free energy (△△G).

Figure 6. Heatmap visualization of exhaustive single amino acid mutations for all selected mutation sites (Y-axis) in the SARS-CoV-2 main protease (Mpro) (PDB: 6Y2E). It represents the effect of each possible amino acid mutation at selected residue on protein stability as measured by the change in Gibbs free energy (△△G).

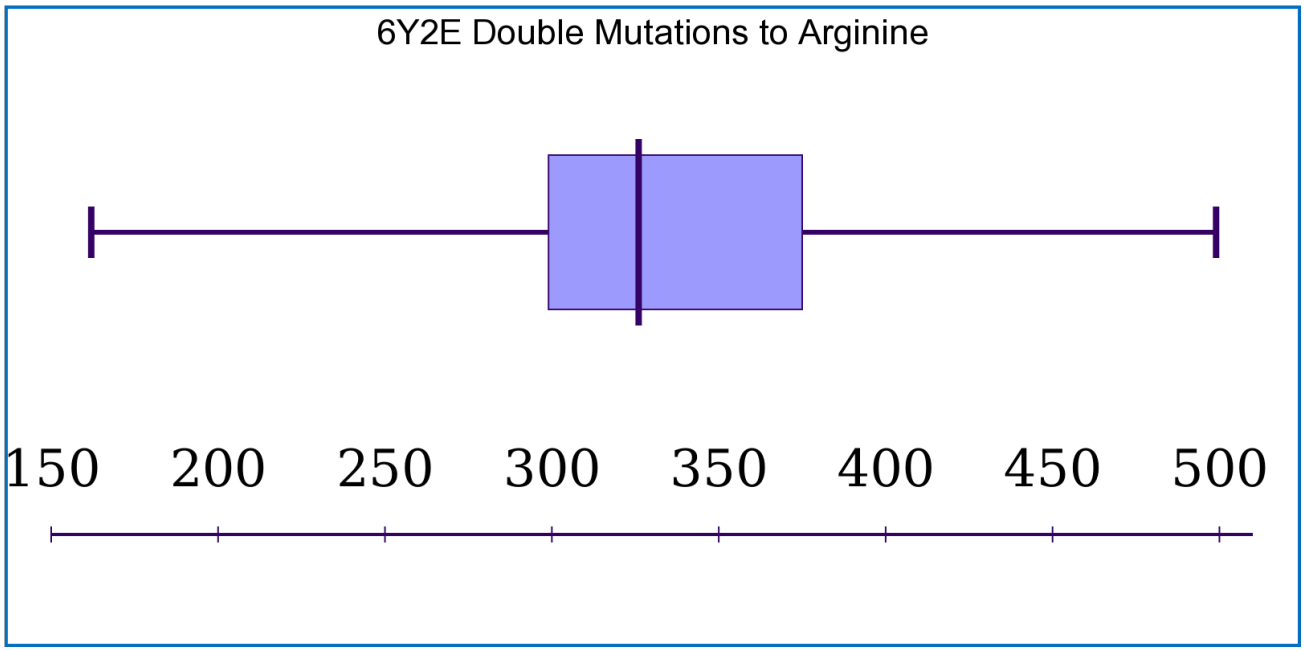

Figure 7. Box and Whisker plot of any double mutation which included Arginine. This was by far, the most stable amino acid.

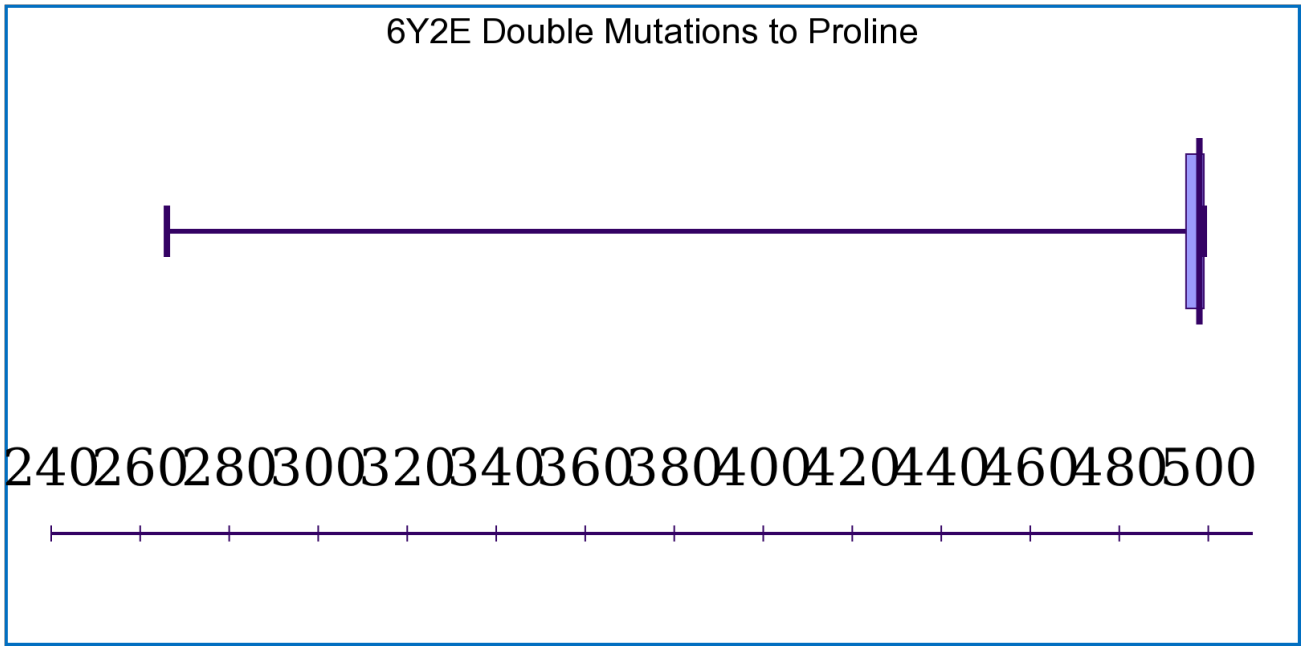

Figure 8. Box and Whisker plot of any double mutation which included Proline. This was the amino acid that was the most unstable.

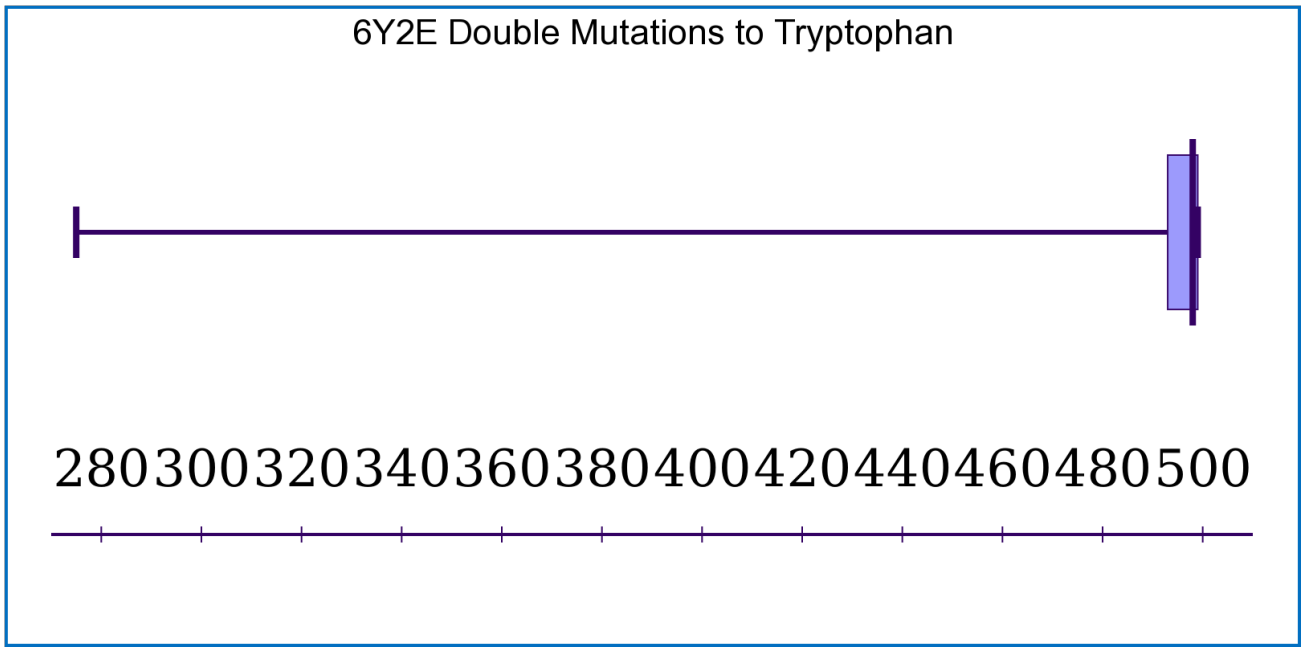

Figure 9. Box and Whisker plot of any double mutation which included Tryptophan. This was the amino acid that was the 2nd most unstable.

NEXT STEPS

To expand upon this project, there could be the further analysis and effects of additional residues, in the proteins that were analyzed (i.e. spike protein and the main protease). There could also be experiments done on additional proteins pertaining to SARS-CoV-2. In terms of the research on single point mutations, more analysis could have been done on more residues that make up the receptor binding domain (RBD) of the spike protein, since only 6 positions were analyzed. In addition, since the spike protein exists as a trimer, mutations at positions that enable the subunits of the spike protein to come together would be interesting to analyze to see what residues are critical for spike formation. As for the single mutations done on the main protease, mutations were only done on 8 residues that are located near the substrate binding pocket. For future work, more mutations can be done for residues that make up the binding pocket of the SARS-CoV-2 main protease to see how critical these residues are for functionality of the main protease.

When considering where the research on double point mutations should focus next, one area of interest would be looking into the effect that specific pairs of amino acids have on the stability of the protein. In addition, there's some interest in investigating the regions within the main protease protein, investigating mutations of the spike protein suite, and investigating mutations in some other proteins in the SARS-CoV-2 proteome. To begin with, two sequences that we identified as areas of interest within the main protease are the residues 135-145 which contained a catalytic Cys145 and the oxyanion hole at residues 143-145 and the residues around the catalytic His41, perhaps 36-46 (6,7). These sequences have been of interest to researchers attempting to find possible drug candidates that can bind the active site and as such inhibit the catalytic effects of main protease (6,7,8). Investigating double point mutations on these two regions may elucidate information about which residues are of the most importance to the structural stability of the protein, thus informing future drug design. Another important sequence of interest may be the C-terminus residues from 297-306. Upon inspection of the homodimeric form of main protease using pdb ID: 6Y2E in Pymol, it appears as though there may be a hydrogen bond that is involved with interfacing between the dimers located at the Gln306 of one and the Ser121 of the other. Disruption of this region could potentially lead to an overall disruption of the interface of the dimers thus leading to an inhibition of catalytic efficiency based on potential allosteric effects of these changes. More target proteins of interest for double point mutations include the spike protein investigated here, especially the interface between the S protein and the ACE2 receptor. A final protein of interest would be that of RNA dependent RNA polymerase (RdRp) or either nonstructural proteins (nsp) 8 and 12 which are important proteins that form part of a complex with RdRp (9). These proteins are responsible for the replication of the SARS-CoV-2 genome and are the current target of the highly promising antiviral drug remdesivir (9).

REFERENCES

1. Hirano, T., Murakami, M. (2020). COVID-19: A New Virus, but a Familiar Receptor and Cytokine Release Syndrome. Immunity, 52(5), 731–733. doi: 10.1016/j.immuni.2020.04.003

2. Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., … Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 579(7798), 270–273. doi: 10.1038/s41586-020-2012-7

3. WHO, Coronavirus disease (COVID-19) Situation Report – 134 . (2020, June 2). Retrieved June 3, 2020, from https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200602-covid-19-sitrep-134.pdf?sfvrsn=cc95e5d5$_2$

4. Torero, M. (2020). Without food, there can be no exit from the pandemic. Nature, 580(7805), 588–589. doi: 10.1038/d41586-020-01181-3

5. Mengist, H. M., Fan, X., Jin, T. (2020). Designing of improved drugs for COVID-19: Crystal structure of SARS-CoV-2 main protease Mpro. Signal Transduction and Targeted Therapy, 5(1). doi: 10.1038/s41392-020-0178-y

6. Linlin Zhang et al. 2020. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved -ketoamide inhibitors. Science (2020). DOI:http://dx.doi.org/10.1126/science.abb3405

7. Jin, Z., Du, X., Xu, Y. et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature 582, 289–293 (2020). https://doi.org/10.1038/s41586-020-2223-y

8. Zhenming Jin et al. 2020. Structural basis for the inhibition of SARS-CoV-2 main protease by antineoplastic drug carmofur. Nature Structural Molecular Biology (July 2020). DOI:http://dx.doi.org/10.1038/s41594-020-0440-6

9. Gordon, Calvin J., et al. "Remdesivir Is a Direct-Acting Antiviral That Inhibits RNA-Dependent RNA Polymerase from Severe Acute Respiratory Syndrome Coronavirus 2 with High Potency." Journal of Biological Chemistry, vol. 295, no. 20, 2020, pp. 6785–6797., doi:10.1074/jbc.ra120.013679.

10. Korber, B., et al. "Spike mutation pipline reveals the emergence of a more transmissible for of SARS-CoV-2." bioRxiv, Apr. 2020, doi:10.1101/2020.04.29.069054

11. Pachetti, Maria et al. "Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant." Journal of translational medicine vol. 18,1 179. 22 Apr. 2020, doi:10.1186/s12967-020-02344-6

12. Yuan, Meng et al. "A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV." Science (New York, N.Y.) vol. 368,6491 (2020): 630-633. doi:10.1126/science.abb7269