Discovery of COVID-19 N-Protein Active Sites for Efficient Antiviral Drug Target Treatment: An Innovative Approach using Torsion Angle Changes in Relation to Functional Activity of Viral N-Proteins

Sreenidhi Sankararaman[1] and Dr. M. Saleet Jafri[1]

[1]School of Systems Biology, George Mason University

(Publishing in scholarly, peer-reviewed *Frontiers in Molecular Biosciences* Journal)

Abstract

COVID-19, a SARS-CoV-2 coronavirus originating from Southeast Asia, has resulted in 13.4 million cases and 580,000 deaths worldwide. The COVID-19 N-protein is responsible for viral replication by assisting in viral RNA synthesis and attaching the viral genome to the replicase-transcriptase complex (RTC). Novel vaccine systems, such as live-attenuated and inactive vaccines, are aimed at suppressing the N-protein by blocking its active sites involved in phosphorylation, oligomerization, and RNA binding. The purpose of this study was to determine active sites of the COVID-19 N-protein for drug targets by identifying torsion angle classifiers for N-protein structural change that correlated with the respective angle's residue inactivation of the N-protein.

In the study, classifiers with a minimum accuracy of 80% determined from NAMD molecular simulation data were analyzed by Principal Component Analysis and cross-validated by Logistic Regression, Support Vector Machine, and Random Forest Classification. Active sites were found at residue 189 for phosphorylation deactivation, residues 252 and 375 for preventing N-protein oligomerization, and residues 252 and 375 for blocking RNA binding. These residues not only were crucial for the aforementioned functions, but they also correlated with torsion angles psi 252 and phi 375 to 100% accuracy. The correlation for the residue matching angles phi/psi 189 was 90.4% accurate. Future applications include virtual drug screening to test the accuracy of drug targets and determining active sites for COVID-19 S-Protein and ACE2 protein, proteins that are used to bind to and invade host cells.

Table of Contents

Introduction

The COVID-19 pandemic has resulted in over 49 million cases and 1.24 million deaths in over 213 countries to date (1). Although isolation measures have been implemented, COVID-19 cases continue to exponentially increase (2). Previous studies have analyzed variations in RNA-binding patterns and sites between coronaviruses, such as HCoV-OC43, and SARS-CoV-2, a family of coronaviruses including COVID-19 (3). Moreover, potential vaccines like the proposed Pfizer vaccine don't provide a permanent solution to COVID-19 suppression and have potential deleterious effects (22). Although scientists have quantified binding patterns and looked at viral genetic codes, they have yet to find specific active sites for the COVID-19 N-Protein, or viral protein. Proposed antiviral drug targeted treatments to suppress the viral protein through these active sites have therefore been ineffective, with no treatment being developed to date (4).

This study discovered COVID-19 N-Protein active sites by (i) determining significant N-Protein torsion angles (ii) establishing significant N-Protein amino acids with molecular dynamics (iii) identifying active sites through significant angle/residue correspondence (iv) testing accuracy through computational cross-validation. Active sites must be found in SARS-CoV-2 N-Proteins, proteins activated by phosphorylation and used for viral assembly, as they package viral RNA and oligomerize viral N-Proteins (5). Active sites determine activity through substrate binding, which results in significant structural and functional change (6).

Structural change is characterized by torsion angles, also known as dihedral angles, which represent protein molecular bond orientations (7). Torsion angles are divided into phi-angles for the N-Ca bond and psi-angles for the Ca-C bond. Their presence is determined with the Ramachandran Plot, which shows the statistical distribution of having both angles in an amino acid (8). Structural changes in proteins are visualized through 3-D protein representations rendered from torsion angle variations and atomic position plots (13). Molecular dynamics, a computational system using force-fields under Newtonian mechanics, determines functional changes by attributing protein function to certain amino acids (10). Specifically, molecular dynamics can be used to provide spatial and temporal images of protein conformations and transitions. Therefore, this study determines a novel approach to identify COVID-19 N-Protein active sites that can be suppressed in efficient drug-targeted treatments by using machine learning and computational cross-validation to discover significant structural torsion angle variations, functional amino acid variations, and the correspondence between the two.

Experimental Design

Methodology and Materials

Materials

NAMD Molecular Simulations were obtained upon request from Dr. Jafri and his lab research team at George Mason University in the School of Systems Biology. In addition, a meta-analysis was conducted across relevant research studies (15, 19, 20, 21) to gather suspected relevant torsion angles in general SARS-CoV-2 N-Proteins. Statistical software packages utilized in this study include the Orange Statistical Software Package, Ramachandran Plots in tandem with Atomic Positions (RMSD and RMSF) Visualization Softwares, and GROMACS Molecular Dynamic Visualization Software.

Overarching Method Outline

This study's goal was to determine COVID-19 N-Protein active sites for suppression in antiviral drug treatment. This was achieved through (i) determining significant N-Protein torsion angle variations (ii) establishing significant N-Protein amino acids residues with molecular dynamics (iii) identifying active sites through significant angle/residue correspondence (iv) testing accuracy through computational cross-validation. The objectives and corresponding materials used are seen in Table 1.

Table 1. Experimental Design Methods and Corresponding Materials

Method Objectives	Materials
Determining Significant Torsion Angles	NAMD Molecular Simulations, Meta-Analysis Article Torsion Angles, Ramachandran Plots, Atomic Position Visualization Softwares, Orange Statistical Software Package
Establishing Significant Residues	GROMACS Molecular Dynamics Visualization Software
Identifying Corresponding Active Sites	Orange Statistical Software Package, NAMD Molecular Simulations, GROMACS Molecular Dynamic Visualization Software
Testing Accuracy through Cross-Validation	Orange Statistical Software

Objective 1: Determining Significant N-Protein Torsion Angles

NAMD Molecular Simulation Data was utilized with Ramachandran Plots and Atomic Position Visualization Softwares to identify allowed torsion angles in the protein. Specifically, the NAMD Molecular Simulation Data provided all numbered phi and psi torsion angles (see Figure 1) in the N-Protein by using a structural calculation model that uses stimulated angle-bond conformations as the degrees of freedom (11). In addition, by conducting a meta-analysis across multiple SARS-CoV-2 peer-reviewed, scientific articles, common torsion angles suspected to be involved in N-Protein function domains were identified.

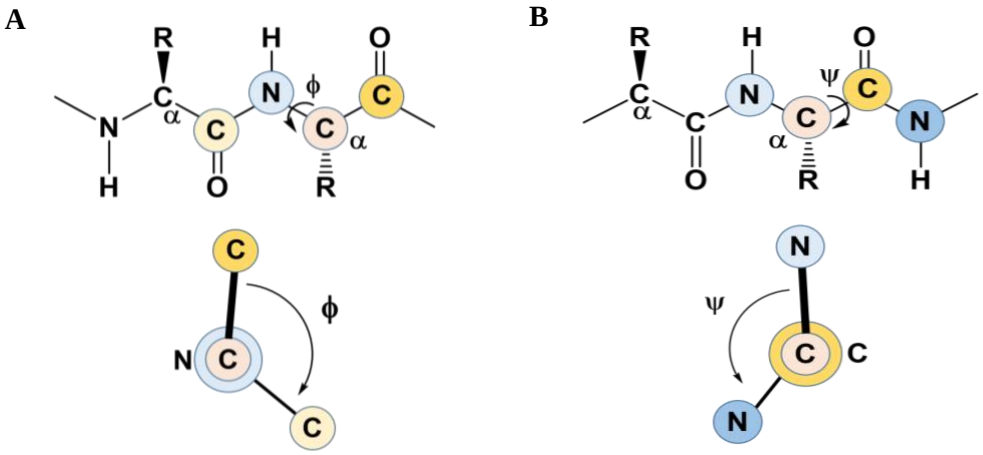

Figure 1: Torsion Angle Biochemical Diagrams. (A) Psi-Angle depiction, seen across two adjacent carbon atoms (B) Phi-Angle depiction, seen across two adjacent nitrogen atoms (12)

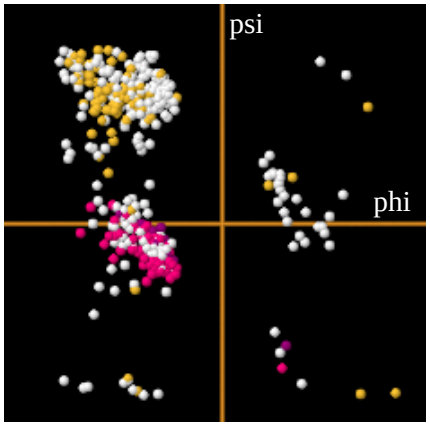

Figure 2: Ramachandran jSMOL Plot for Torsion Angles (18). Pink and yellow regions depicting core/allowed regions

Ramachandran Plots were used to identify permitted torsion angles. Ramachandran Plots have been used in previous studies (8) to plot torsion angles (phi and psi angles) against a residue. Specifically, by plotting torsion angles, core and allowed regions of torsion angle combinations in proteins that are feasible without steric hindrance are visible (see jSMOL plot in Figure 2). Moreover, Atomic Position Visualizations of the N-Protein were gathered for torsion angles insight. Specifically, Normal Distributions, multi-histograms, and trajectories of the RMSD (root-mean-square deviation)

and RMSF (root-mean-square fluctuation) for residues were devised. RMSD and RMSF visualizations of amino acids gave insight in torsion angle calculations as RMSF indicates protein bond flexibility and RMSD indicates the degree of separation, which is used to calculate the angle, between adjacent atoms (13).

After identifying applicable torsion angles that would be sterically-feasible and comparing their relative angle measurements using RMSD/RMSF visualizations, the Orange Statistical Software Package was used to build novel classification networks. Specifically, by training phi/psi angle classifiers to a forward pruning tree algorithm, classification trees were utilized to determine the success of individual angle classifiers in accurately predicting N-Protein inactivity and activity. This prediction was modeled by building a network that determined that a torsion angle resulted in significant structural change if its variations resulted in significant phosphorylation changes that affected protein activity. Using the Orange Modeling Network (seen in Figure 3), classifiers up to at least 80% classification accuracy were determined and distinctions between classifiers were identified by checking if there was separation between the data values in scatter plots and box plots.

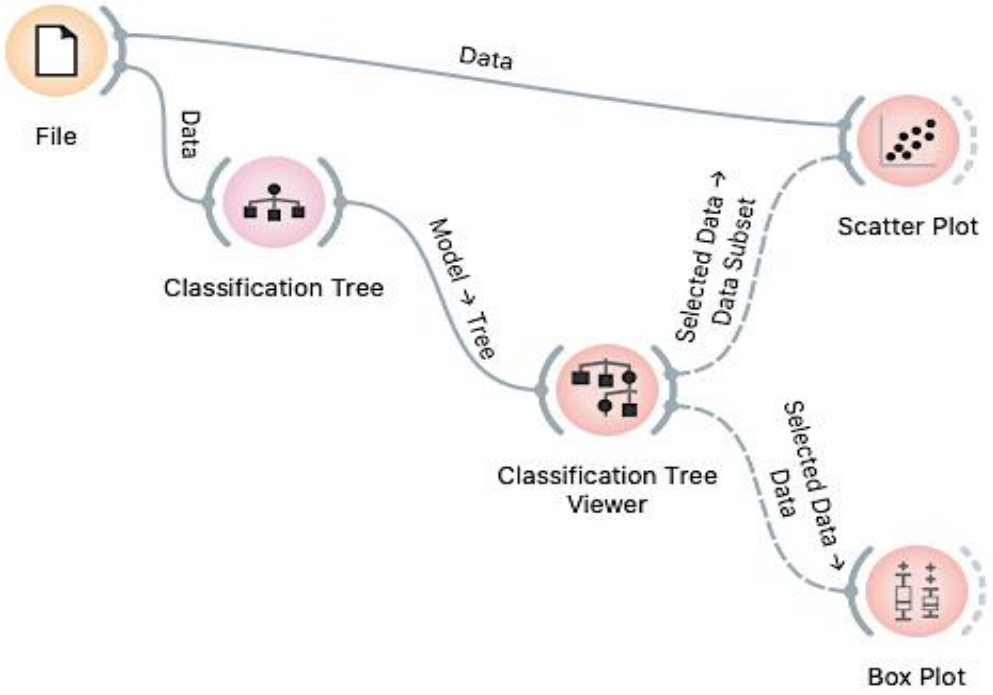

Figure 3: Orange Modeling Network Built for Primary Classification. Utilizes applicable torsion angles from NAMD Molecular Simulations and meta-analysis of relevant articles

After identifying accurate torsion angle classifiers that resulted in significant structural change that predicted the N-Protein activity, a Principal Component Analysis was conducted to streamline the major classifiers.

Objective 2: Establishing Significant N-Protein Residues

After determining the streamlined major classifiers, the GROMACS Molecular Dynamics Visualization Software was used to visualize the N-Protein. Through the molecular dynamic simulations, spatial and temporal realistic models of N-Protein dynamics and flexibility were rendered (10). By manipulating the N-Protein model by removing and altering individualized amino-acids, certain amino acid removals, or even alterations, were found to have resulted in N-Protein inactivity. Although this did not immediately signify that the amino acid residue was a significant active site as it could have been an intermediate amino acid needed for an activation process, such as phosphorylation transduction signaling, significant amino acid removals and alterations that resulted in N-Protein inactivity were documented. This was because these residues were characterized by having significant functional relevance to the N-Protein.

Objective 3: Identifying COVID-19 N-Protein Active Sites

Active sites for proteins are locations where proteins bind and conform to substrates (see Figure 4). These sites determine the active state of the protein and, if suppressed, would render an inactive protein. Therefore, these sites are indicators of significant structural and functional changes. In Objective 1, torsion angle variations in residues that resulted in significant structural changes were identified from NAMD data and meta-analysis of articles. In Objective 2, residues that had significant functional value, discovered when said residues were removed, were discovered using GROMACS. By identifying residue numbers that have both a torsion angle variation resulting in significant structural change and have significant functional value, a list of potential active site candidates was devised. Furthermore, active site candidate residue numbers were compared to see if they fall within the numbered N-Protein domains for RNA-binding, phosphorylation cascades, and oligomerization (5, 14, 15). From these comparisons, N-Protein active site residue(s) that were responsible for each domain, showed significant functional change in terms of N-Protein activity

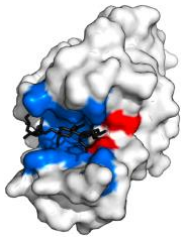

Figure 4: Protein Active Site. (Blue) Binding/conforming region (Red) Catalytic Region (17)

when removed, and showed N-Protein activity prediction capability and significant structural change in the N-Protein when altered were identified.

Objective 4: Testing Accuracy Through Computational Cross-Validation

After determining the COVID-19 N-Protein active sites, a computational cross-validation using Logistic Regression, Random Forest, and Support Vector Machine Models (see Figure 5) was conducted. These computational mathematical models were used as they are most effective for this study that includes a large-scale protein activity and amino acid analysis (16). By generating a confusion matrix, accuracy data was gathered for select active sites that had the highest classification accuracies for their respective N-Protein domains and also matched with a functionally significant residue. Specifically, this accuracy referred to the predictions of determining inactivity and activity of the COVID-19 N-Protein. Using the data from the confusion matrices, the error rate, accuracy, sensitivity, and specificity of the select active sites in determining N-Protein inactivity were calculated.

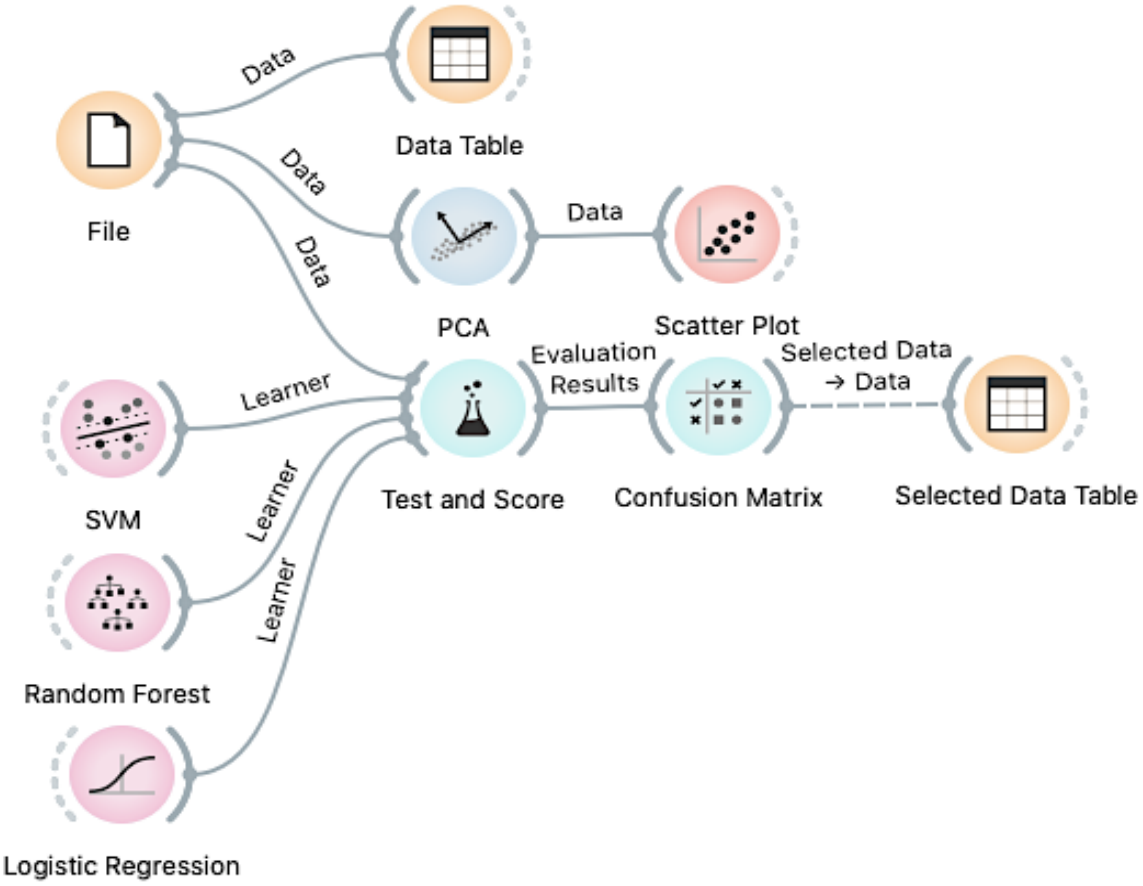

Figure 5: Orange Modeling Network Built for Computational Cross-Validation. Utilizes Support Vector Machine (SVM), Random Forest, and Logistic Regression Models to Build Confusion Matrices to Analyze Accuracy of Select Active Sites.

Results

Objective 1: Determining Significant N-Protein Torsion Angles

Atomic Position Visualization Models were used to first find novel quantifications of acceptable (Ramachandran Plot) torsion angle variations in N-Protein amino acids (see Figure 6).

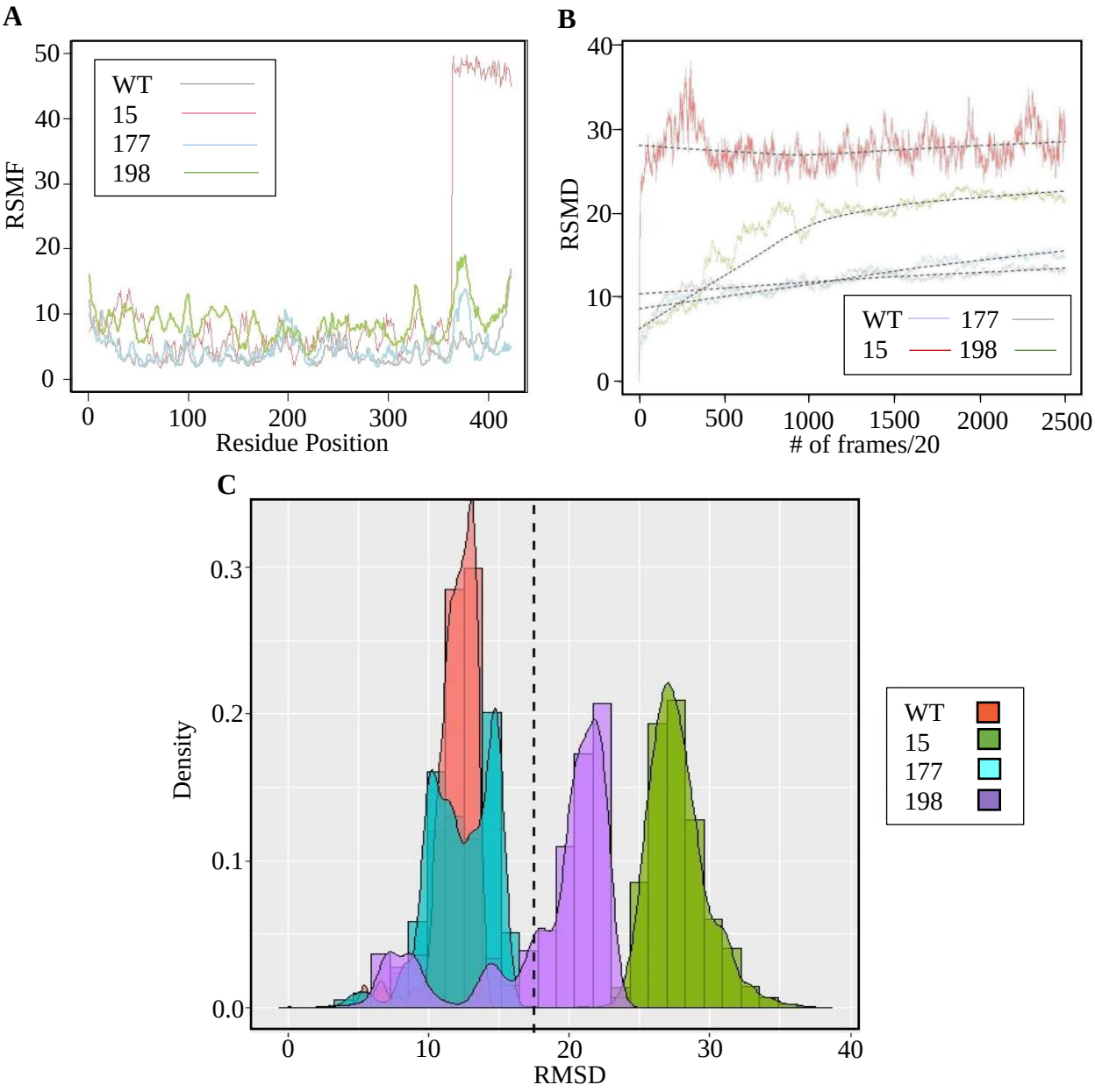

Figure 6: Atomic Position Visualization for Torsion Angle Quantifications. (A) Projection of RSMF (root-mean-square fluctuations) of residues to indicate protein flexibility (B) Projection of RSMD (root-mean-square deviations) of all residues to indicate degree difference between adjacent atoms (C) Prevalence of RSMD values across residues

Upon a meta-analysis of articles, torsion angles suspected to play a significant role in SARS-CoV-2 N-Proteins in general were also included. These torsion angles (psi/phi 69, 290, and 328), along with the sterically-acceptable, quantified torsion angles from the atomic position simulations conducted were classified through a novel computational network built using the Orange Statistics Software Package. Classifiers that had at least 80% accuracy in determining the activity of the COVID-19 N-Protein were determined (see Figure 7).

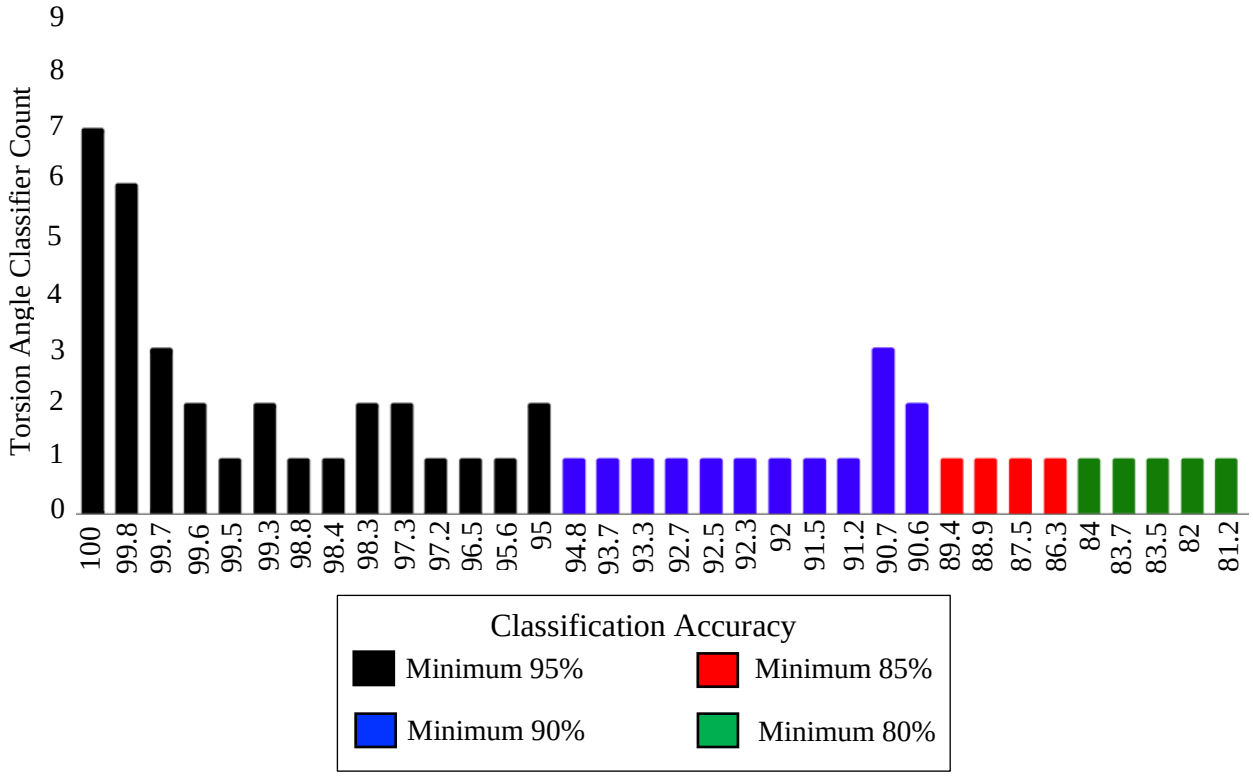

Figure 7: Determination of Classification Accuracy Ranges for Phi/Psi Torsion Angles. Orange Classification Trees use decision-based machine learning to identify the minimum 95%, 90%, 85%, and 80% accuracy torsion angles classifiers. Classification trees determine accuracy based on accurately predicted inactive and active states for the respective phi/psi angle. Classifiers shown do not include branched classifiers.

The torsion angles determined from these classifiers were differentiated using classification trees. These trees identified their classification accuracy based upon how accurate the angle was in predicting the active or inactive state of the N-Protein (seen in Figure 8). This activity was inferred from the phosphorylation changes predicted to occur in correspondence with the structural changes associated with the torsion angles.

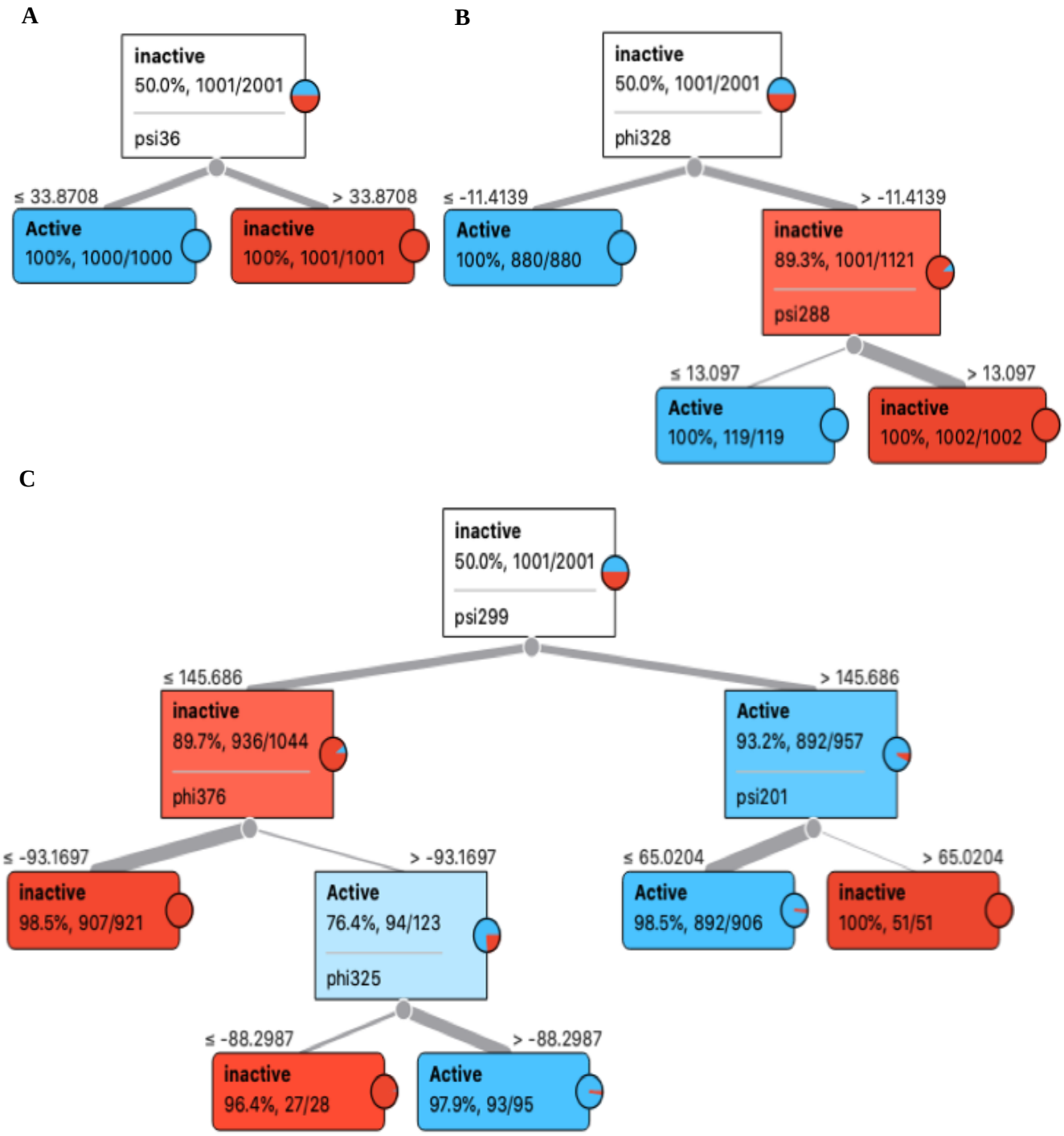

Figure 8: Classification Tree Visualization for Torsion Angles. Orange Classification Trees used decision-based machine learning to use inactivity/activity predictions based on phosphorylation changes attributed to torsion angle structural changes to identify the percentage of classified and misclassified states. (A) Single-parent trees in which greatest child percentage is used as classification accuracy (B) Limited branching (3 layers) tree shows greater than 80% classifiers put into consideration (C) Increased Branching (4 layers) full tree indicates increased minor classifiers identified above 80% accuracy

Upon classification, angle classifiers with accuracy of at least 80% were identified. Further classifiers were noted but were not considered in determining primary/supplementary active sites. These classifiers were deemed significant based on the significant predicted functional change associated with their protein structures. The classifiers were further divided into groups for efficient analysis through Principal Component Analysis (seen in Figure 9).

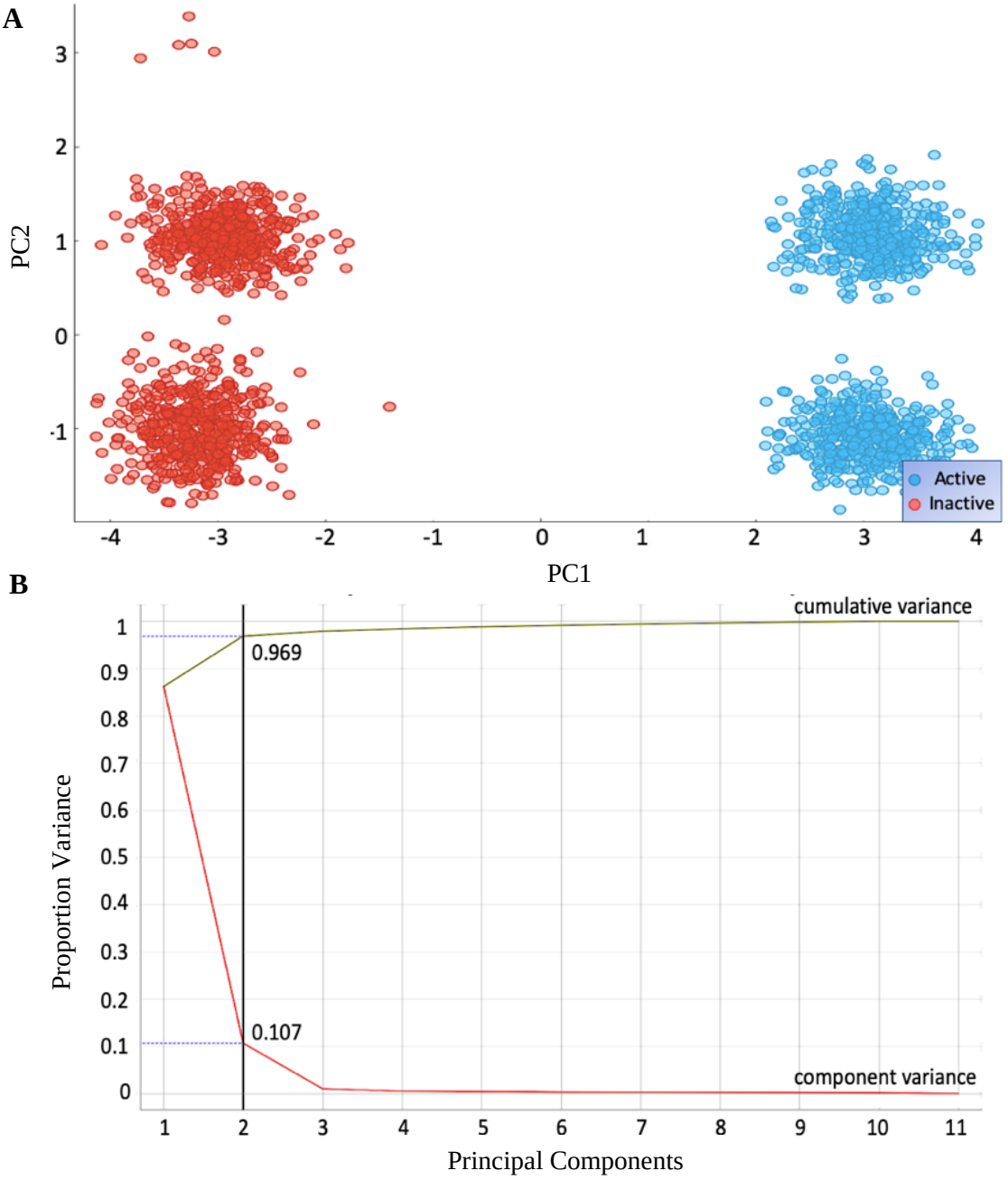

Figure 9: Principal Component Analysis of Major Angle Classifiers. Principal Component Analysis (PCA) streamlined major angle classifiers of 100% accuracy and 90.4% accuracy (A) Scatter Plot Graph after PCA division showing clearly defined groups (B) PCA Analysis explains 96% variance

Through this innovative, quantitative approach, significant torsion angles that had the ability to predict inactivity of N-Protein and had significant structural relevance to the N-Protein were determined.

Objective 2: Establishing Significant N-Protein Residues

After the torsion angles were divided into smaller groups for simplified analysis, molecular simulations were utilized to determine corresponding residues (seen in Figure 10). The corresponding residues had to show significant functional change and had to correspond with a primary angle classifier that predicted significant structural change.

Figure 10: Molecular Simulation Residue Classification. Molecular Dynamic Simulations were used to identify significant amino acid residue. Residues were removed/modified to determine significant functional change in terms of inactivity and activity. Regions highlighted in blue highlight the identified significant residues.

Through harnessing the power of molecular dynamics, significant N-Protein residues that, when manipulated resulted in N-Protein inactivity, were discovered.

Objective 3: Identifying COVID-19 N-Protein Active Sites

Following the determination of significant residues in correspondence with the angle classifiers previously identified, three primary residues were identified as active site candidates. These residues were residues 252, 375, and 189, which corresponded with angles psi 252, phi 375, and psi 189. Torsion angles 252 and 375 have a 100% classification accuracy and torsion

angle 189 has a 90.4% classification accuracy. Based on comparing domain residue numbers of the N-Protein (5, 19, 20), the primary active site residues for RNA-binding and N-Protein oligomerization are residues 252 and 375. The supplementary active site responsible for the phosphorylation cascade to activate aforementioned functions is residue 189. Cross-validation was used to see the accuracy of the residues under statistical models (seen in Figure 11).

Objective 4: Testing Accuracy through Computational Cross-Validation

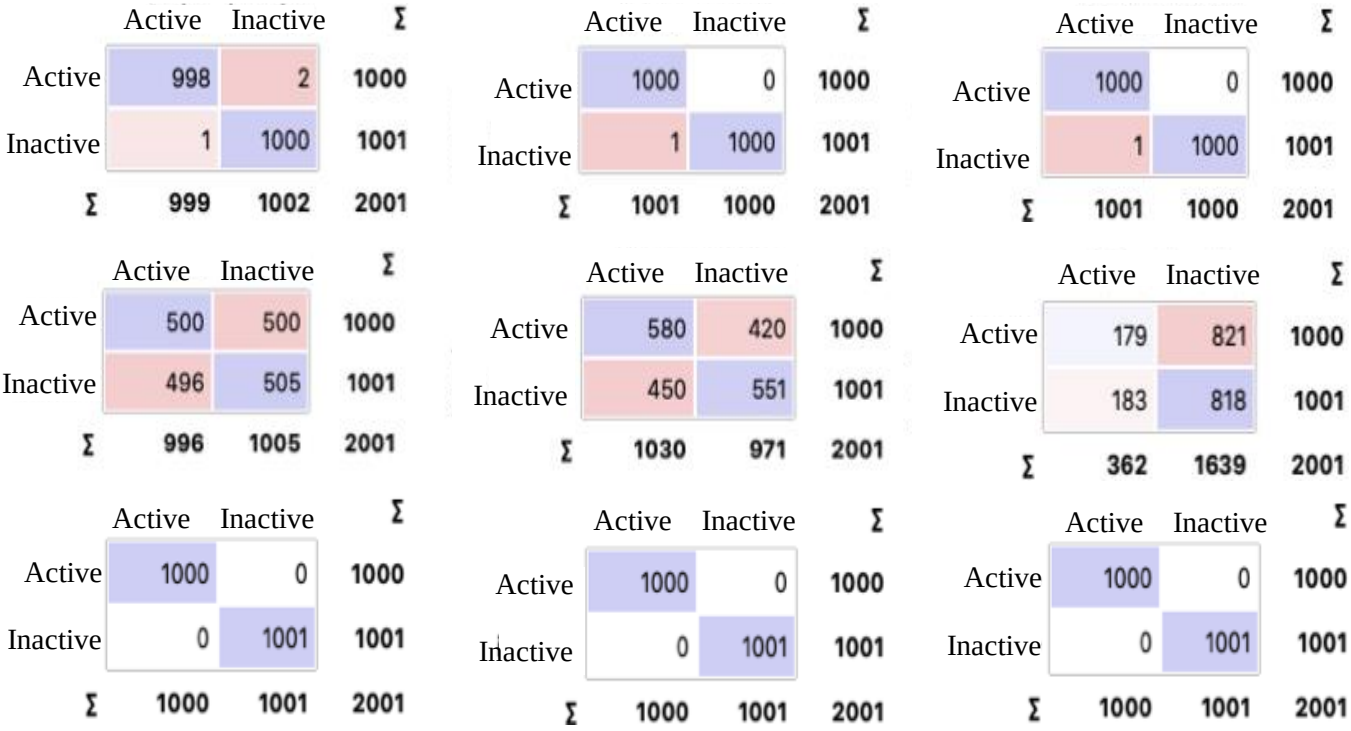

Figure 11: Confusion Matrices for Logistic Regression, Random Forest Classification, and SVM Models. Confusion Matrices show the cross-validation results for statistical models regarding major classifiers with left activity states (actual) and upper activity states (predicted) (A) Confusion Matrices for psi 252 for Logistic Regression, Random Forest, and SVM (L-R) (B) Confusion Matrices for phi 375 for Logistic Regression, Random Forest, and SVM (L-R) (C) Confusion Matrices for psi 189 for Logistic Regression, Random Forest, and SVM (L-R)

The Confusion Matrices provide support for the identified active sites as psi 252 and phi 375 are highly accurate in determining activity as they are primary active sites, while the 50% accuracy of psi 189 is expected as it is a supplementary active site. In addition, Random Forest Classification was the most accurate model. The differential performances of the three residues under the Random Forest model was visualized through scatter plots (seen in Figure 12).

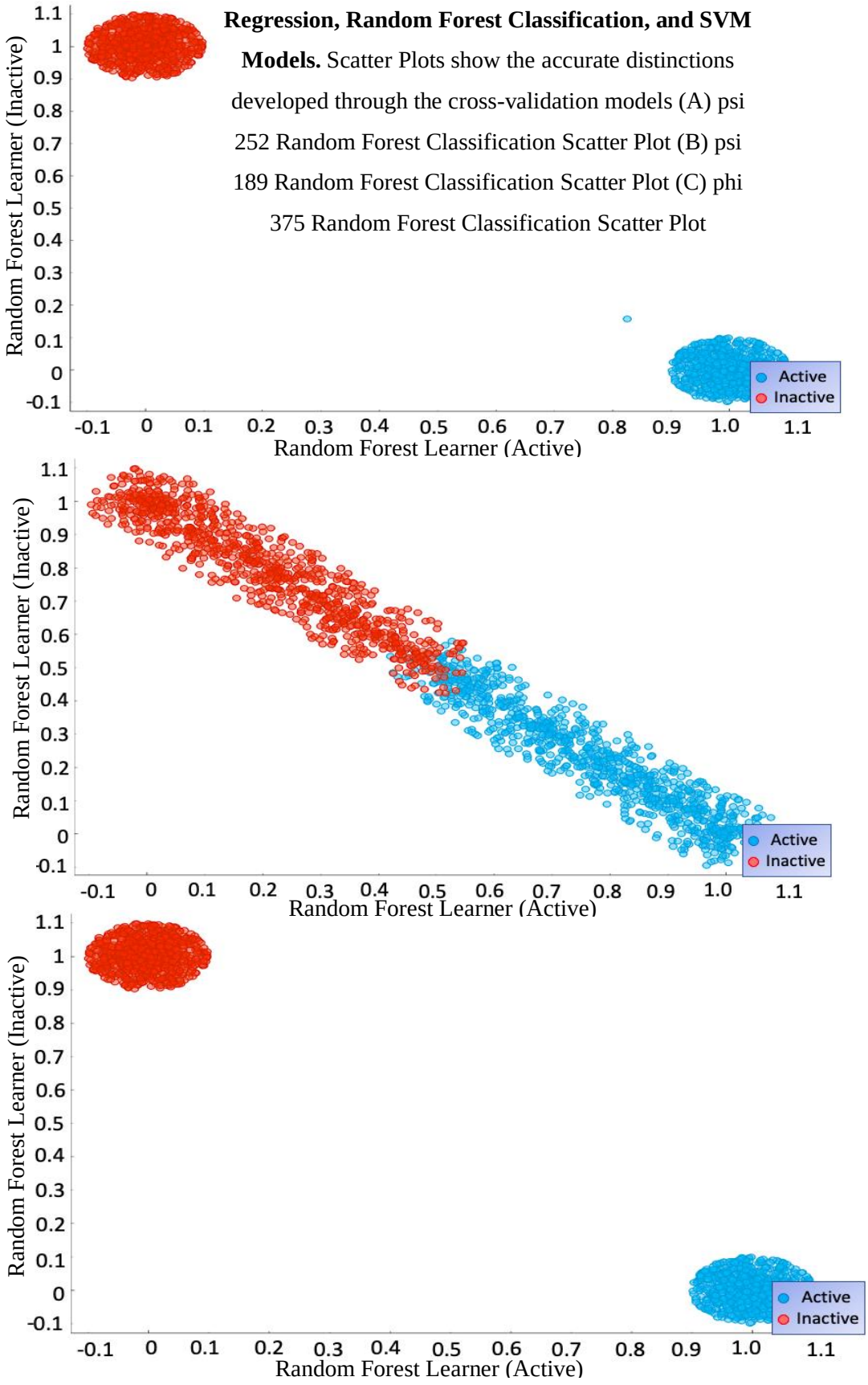

Figure 12: Scatter Plot Graphs for Logistic Regression, Random Forest Classification, and SVM Models. Scatter Plots show the accurate distinctions developed through the cross-validation models (A) psi 252 Random Forest Classification Scatter Plot (B) psi 189 Random Forest Classification Scatter Plot (C) phi 375 Random Forest Classification Scatter Plot

These scatter plots further support the assertion of the primary active site residues 252 and 375 and the supplementary active site 189 by showing distinctions to lack of distinctions respectively under the same model. To further quantify accuracy of determined active sites, confusion matrices provided data on the true positives, true negatives, false positives, and false negatives for calculating sensitivity, specificity, error rate, and accuracy of residues (seen in Figure 13).

A

Logistic Regression	
Error Rate	0.001
Accuracy	0.998
Sensitivity	0.998
Specificity	0.998

Random Forest	
Error Rate	0.000
Accuracy	1.000
Sensitivity	0.999
Specificity	1.000

SVM	
Error Rate	0.000
Accuracy	1.000
Sensitivity	0.999
Specificity	1.000

B

Logistic Regression	
Error Rate	0.000
Accuracy	1.000
Sensitivity	1.000
Specificity	1.000

Random Forest	
Error Rate	0.000
Accuracy	1.000
Sensitivity	1.000
Specificity	1.000

SVM	
Error Rate	0.000
Accuracy	1.000
Sensitivity	1.000
Specificity	1.000

C

Logistic Regression	
Error Rate	0.498
Accuracy	0.502
Sensitivity	0.502
Specificity	0.502

Random Forest	
Error Rate	0.435
Accuracy	0.565
Sensitivity	0.563
Specificity	0.567

SVM	
Error Rate	0.502
Accuracy	0.498
Sensitivity	0.494
Specificity	0.499

Figure 13: Sensitivity, Specificity, Error Rate, and Accuracy Analysis of Classifiers. Data gathered from confusion matrices of statistical models to determine sensitivity, specificity, error rate, and accuracy analysis of classifiers (A) Statistical Model Analysis for psi 252 (B) Statistical Model Analysis for phi 375 (C) Statistical Model Analysis for psi 189

This computational cross-validation, in addition to research support of the N-Protein domain characteristics, support the determination of the active site residues 189, 252, and 375 for the COVID-19 N-Protein.

Discussion

The study's findings suggest the COVID-19 N-protein primary active sites as being at residues 252 and 375 for both N-protein oligomerization and RNA-binding and the supplementary active site as residue 189 for phosphorylation cascades within the protein. Active sites are known to both characterize the activity of a protein and result in significant structural and functional change when interacted with, either through substrates or through external manipulation (6). These residues fulfill all three characteristics required to be an active site. To begin with, all 3 residues correspond with torsion angles that have high classification accuracy when determining N-Protein structural change through activity, as seen in Objective 1. Residues 252 and 375 have a 100% classification accuracy while residue 189 has a 90.4% accuracy. In addition to being significant indicators of structural change, they also demonstrate significant functional change. Specifically, when removed in molecular dynamic simulations in Objective 2, all 3 residues resulted in N-Protein inactivity. Moreover, as seen in the computational cross-validation, primary sites 252 and 375 had nearly 100% accuracy, sensitivity, specificity and a nearly negligible error rate when determining the activity of N-Proteins through models, with Random Forest surprisingly being the most successful.

Supplementary active site 189 had approximately 50% accuracy in determining activity of N-proteins through statistical models. However, this serves to support that residue 189 is a supplementary active site as it alone does not determine activity, but rather, is a causation factor for the functioning of the other primary sites. This is because, as supported through previous research on N-Protein domains (5, 19, 20), residues 252 and 375 fall within the general RNA-binding and oligomerization ranges while residue 189 falls within the phosphorylation cascade domain, which is a transduction signaling pathway that serves to activate other functions. Hence, as supported by previous research and by the findings in this study, residues 252, 375, and 189 serve as the COVID-19 N-protein active sites for its viral functions. Potential limitations for this study may arise from using only the phi and psi torsion angles. Using omega torsion angles, which are a minor subset of torsion angles, may provide further insight into COVID-19 N-protein characteristics. However, as the NAMD dataset provides ample data, the lack of omega torsion angles in this study is not concerning. In essence, in addition to determining COVID-19 N-Protein active sites for antiviral drug treatments to potentially save millions of lives, this study also provides relevant insights to previous studies on N-protein domain characteristics.

Conclusion

The purpose of this study was to discover COVID-19 N-Protein active sites for antiviral drug treatments. Specifically, identifying these active sites that have remained unknown to date is crucial in stopping the global COVID-19 pandemic as it allows for immunosuppression of the virus through drug treatments that suppress the active sites identified. This study found the COVID-19 primary active sites to be residues 252 and 375 for both N-protein oligomerization and RNA-binding, viral functions of the N-Protein. In addition, it found the supplementary active site in charge of facilitating phosphorylation cascades for the activation of the aforementioned functions to be residue 189. Implication of this novel study in terms of scientific fields include giving insight to scientists about previously unknown active site locations, allowing for researchers worldwide to build off of this finding in devising treatments targeting to suppress these specifically identified residues, and adding value to previous research on general N-Protein domain characteristics. In addition, global societal and economic implications of having newly-identified active sites to base once-unfeasible antiviral treatments off of include reinsuring people's livelihoods, restoring trade balance between once-isolated countries, and restoring valuable in-person education for our nation's youth. Most importantly, global health implications of this study include facilitating efficient, antiviral treatments from the discovery of these active sites, providing permanent COVID-19 suppression, stopping the spread of worldwide cases, and potentially saving tens of millions of lives from the global COVID-19 health crisis that persists to date.

Recommendations

Future, related courses of study include conducting virtual drug screening to train and test the accuracy of antiviral drug targets in identifying and suppressing active sites of the N-Protein. In addition, COVID-19 S-Protein and ACE2, minor supplementary proteins to the COVID-19 N-Protein, can also be analyzed for characteristics that can help in identifying external interactions of the N-Protein and how this might affect its viral propagation. Computational docking of drug libraries to these sites is also possible during future studies.

Bibliography

1. https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/
2. https://journals.sagepub.com/doi/full/10.1177/2516602620938542
3. https://www.biorxiv.org/content/10.1101/2020.03.06.977876v2.full
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7192075/
5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4147684/
6. https://courses.lumenlearning.com/boundless-biology/chapter/proteins/
7. https://www3.cmbi.umcn.nl/wiki/index.php/Torsion_angle
8. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/ramachandran-plot
9. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3794089/
10. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5321087/
11. https://science.sciencemag.org/content/220/4598/671
12. https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/chapter-2-protein-structure/
13. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3317885/
14. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7094638/
15. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7106479/
16. https://ieeexplore.ieee.org/document/8455066
17. https://www.rcsb.org/structure/9LYZ
18. https://proteopedia.org/wiki/index.php/User:Karl_Oberholser/Ramachandran_Plots
19. https://www.biorxiv.org/content/10.1101/2020.06.28.176248v1.full
20. https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0065045&type=printable
21. https://reader.elsevier.com/reader/sd/pii/S2211383520305505?token=985C7673F5BF891B2CB4A70266FC8573B4BCD3E3FE8F9311E1B8727A592624F93C9A116C5E984120013F84FE9D9A2882
22. https://www.pfizer.com/news/press-release/press-release-detail/pfizer-and-biontech-announce-vaccine-candidate-against