Machine Learning for the Identification of Viral Attachment Machinery from Respiratory Virus Sequences

Stepan Demidkin[1#], Maïa Shwarts[1#], Arijit Chakravarty[2], Diane Joseph-McCarthy[1*]

[1] Department of Biomedical Engineering, Boston University, Boston, MA USA

[2]Fractal Therapeutics, Cambridge, MA USA

[#] SD and MS contributed equally to this work

[*]Address correspondence to Diane Joseph-McCarthy, djosephm@bu.edu

Key words: SARS-CoV-2, COVID-19, machine learning models, viral cell entry machinery, structural determinants of host cell recognition

Abstract

At the outset of an emergent viral respiratory pandemic, sequence data is among the first molecular information available. As viral attachment machinery is a key target for therapeutic and prophylactic interventions, rapid identification of viral "spike" proteins from sequence can significantly accelerate the development of medical countermeasures. For five families of respiratory viruses, covering the vast majority of airborne and droplet-transmitted diseases, host cell entry is mediated by the binding of viral surface glycoproteins that interact with a host cell receptor. In this report it is shown that sequence data for an unknown virus belonging to one of the five families above provides sufficient information to identify the protein(s) responsible for viral attachment and to permit an assignment of viral family. Random forest models that take as input a set of respiratory viral sequences can classify the protein as "spike" vs. non-spike based on predicted secondary structure elements alone (with 97.8 % correctly classified) or in combination with N-glycosylation related features (with 98.1 % correctly classified). In addition, a Random Forest model developed using the same dataset and only secondary structural elements was able to predict the respiratory virus family of each protein sequence correctly 89.0 % of the time. Models were validated through 10-fold cross-validation as well as bootstrapping. Surprisingly, we showed that secondary structural element and N-glycosylation features were sufficient for model generation. The ability to rapidly identify viral attachment machinery directly from sequence data holds the potential to accelerate the design of medical countermeasures for future pandemics.

Introduction

The COVID-19 pandemic has underscored the importance of an effective response for emerging viral pathogens that is focused on the rapid deployment of molecular testing and medical countermeasures. Our experiences with the current pandemic have highlighted the vulnerability of the global healthcare infrastructure to respiratory pathogens that, like SARS-CoV-2, are capable of long-range airborne spread via aerosolized particles [1]. In contrast to other pathogens, the window for effective intervention to avert a pandemic resulting from a newly emergent respiratory virus may be very short. Thus, the speed with which molecular diagnostics, therapeutics, and vaccines can be deployed are critical determinants of our ability to contain an outbreak.

The viral attachment machinery (the set of proteins responsible for host cell attachment and cell entry) has served as a historically important focus for the development of molecular tests (for example for influenza [2] and SARS-CoV-2 [3, 4]) as well as medical countermeasures such as vaccines [5-7]. Thus, the accurate and efficient identification of the viral attachment machinery is a critical first step in the design and deployment of biomedical countermeasures. It had been observed for coronaviruses in 2012 (pre-COVID-19) that the tertiary structure of the spike protein is not conserved but that the secondary structure topology is conserved [8]. It was subsequently also noted that the pattern of N-linked glycosylation is highly conserved and may play a role in immune evasion [9].

Automated function prediction (AFP) of novel proteins is a mature field (see [10-13] for reviews). A number of groups have used approaches that leverage structure-based homology, focusing either on the full three-dimensional (3D) protein structure, or on the identification of 3D structural motifs (see, for example, [14-17]). However, 3D structure alone is often insufficient

for functional annotation, as proteins possessing similar global structures can perform very different biological functions (for example, [18]). Computational structural alignment methods, although first pioneered in the 1960s, typically have accuracies on the order of ~90% [19] and at least in the case of coronaviruses as described above the 3D structure is not conserved. Furthermore, 3D structural motifs for viral attachment proteins are often optimized for specifically for enzymes and are not readily able to identify viral attachment machinery. As an alternative, AFP from DNA sequences relies on sequence homology [20-22], or the identification of sequence motifs [23, 24]. A potential weakness of this approach is that novel viruses with low sequence homology to pre-existing pathogens may prove less tractable to homology-based approaches. As a further consideration, during the early days of an emerging pandemic, steps such as multiple sequence alignment, phylogeny reconstruction and 3D structure prediction can add weeks to the timeline for response. An accurate ML model may be able to pinpoint the target within seconds.

With respect to preparedness for potential future pandemics, tools that can aid in the rapid deployment of therapeutic and vaccine countermeasures are clearly needed. Specifically, for viral pathogens originating from the most prevalent respiratory virus families, which are key pathogens of concern, intervening at the localized emergence stage may prevent the transition to a full-blown pandemic. Based on the earlier cited observations, we hypothesized it may be possible to develop a machine learning (ML) model based on predicted secondary structure elements and N-glycosylation features alone capable of identifying viral attachment machinery (the “spike” protein or its equivalent) from an unknown respiratory virus sequence. More generally, we also sought to gain a further understanding of the structural features that may

distinguish viral attachment machinery proteins with a view toward elucidation of key structure-function relationships.

Methods

Virus families, viral sequences, and "spike" proteins: Five families of respiratory viruses were included in this study: Coronaviridae, Paramyxoviridae, Pneumoviridae, Adenoviridae, and Orthomyxoviridae. Each of the viruses within these families has a protein responsible for viral attachment and host cell entry, which will be referred to herein as the "spike" protein (see Fig. 1A). For Coronaviruses, it is the Spike S Glycoprotein which is aptly named because it projects from the surface of the virion (Fig. 1B) as do the other "spike" proteins. Note that for Influenza Virus A within the Orthomyxoviridae family, we selected Hemagglutinin as the equivalent of the "spike" although Neuraminidase is a second antigenic determinant. A total of 39 sequences (ranging from 4 to 12 for each virus family) encoding 316 proteins were utilized (see Table 1). Specifically, we included 7 Coronaviridae sequences representing 7 viruses, 4 Paramyxoviridae sequences representing 4 viruses, 12 Pneumoviridae sequences representing 2 viruses, 8 Adenoviridae sequences representing 1 virus, and 8 Orthomyxoviridea sequences representing 1 virus.

Prediction of secondary structural elements: The Jpred4 [25] secondary structure prediction server was used to predict structural elements for each viral sequence in the dataset. Jpred4 is a server that hosts Jnet, a neural network secondary structure prediction algorithm trained with different representations of multiple sequence alignment profiles for the same sequences [26]. Each residue in a protein sequence is designated as H (helical), E (extended sheet), or other. Since Jpred4 predicts secondary structure on protein sequences up to 800 amino

acids in length, a script (Fig. S1) was written to break protein sequences into 800 residue segments and subsequently concatenated the results. For each protein, the script calculated protein length, % H, and % E, identified the longest helix and the longest sheet and calculated % longest H, and % longest E, where % longest H (E) is the length of the longest H (E) in the protein divided by the length of the protein. Finally, %helix, %sheet, %longest H, and %longest E is output.

Prediction of N-glycosylation sites: For the sequences described above, N-glycosylation sites were predicted for each protein using NetNGlyc [27, 28]. The NetNGlyc method uses artificial neural networks to predicts N-Glycosylation sites in proteins through analysis of the sequence context of Asn-Xaa-Ser/Thr sequons. FASTA format protein sequences were entered on the NetNGlyc 1.0 Server (https://services.healthtech.dtu.dk). Asparagines with overall positive score, denoted by '+', '++', '+++' and '++++' (each counted in their respective category), where '++++' indicates a prediction with highest confidence based on a combination of overall potential score and jury agreement amongst the nine neural networks utilized, were predicted to be glycosylated. The total number of glycosylation sites per protein (total N-sites) was the sum of the number of residues scored '+' or higher. The density was the total sites divided by the number of residues in the protein (as reported by NetNGlyc).

Amino Acid Composition: Protein sequences were obtained from nucleic acid sequences with Bioinformatics Toolbox in MATLAB version 2019b (MathWorks, 2021, Natick, MA, USA), and a letter frequency counter code was used to obtain the occurrence of each amino acid (AA) for each protein. The individual occurrences were divided by the corresponding protein amino acid length and multiplied by 100, giving %AA composition.

Statistical test of association: Two-tailed t-tests for two independent samples were performed using XLSTAT v22.2.3 (Addinsoft, 2020 New York, USA) to assess the association of various features with spike vs. non-spike protein status. Features that showed a statistically significant association (p-value ≤ 0.05) between spike and non-spike groups and thereby rejected the null hypothesis were considered for inclusion in the ML models.

Inputs vectors for ML models: Feature vectors were generated for each of the 316 protein sequences to create the full dataset. For each protein, the following features were calculated as described above: total N-sites, density, %M, %N, %S, %sheet, %helix, %longest sheet, and %longest helix. The designation of spike or non-spike was also included.

Random Forest model development: Weka, an open-source software workbench for ML and data analysis [29], was utilized to develop Random Forest classifiers derived from the dataset described above. Random forest is a supervised ensemble learning method that generates a set of decision trees maximizing the separation of the classes that are sought to be discriminated [30, 31]. Subsets of the data were converted into ARFF format and uploaded to the Weka Explorer version 3.8.4 to generate specific Random Forest models (see Table S1). For each Random Forest model, a ZeroR model was also generated. Ten-fold cross-validation was utilized with both algorithms. The statistical significance of each model result was assessed by performing a Fisher's exact test [32].

Bootstrapping: Bootstrapping datasets were generated using the random sampling with replacement command in MATLAB version 2019b (MathWorks, 2021, Natick, MA, USA). For each model being investigated, 1000 such datasets were generated and saved as CSV files. The CSV files were converted to ARFF format using a modified csv-to-arff Python routine (obtained from github.com/anaavila). For the 50-50 balanced bootstrapping tests, for each dataset, 50% of

the feature vectors were for proteins designated as spike and the other 50% were for those designated as non-spike.

Results and Discussion

To examine the feasibility of using a machine learning model trained on viral sequences, data set was assembled consisting of 316 protein sequences for 39 respiratory viruses from five virus families, with each protein classified as "spike" (viral attachment machinery) or non-spike. Next, the associations between various features and the classification of "spike" vs. non-spike for the coronaviruses in the dataset were examined to look for signals indicating that certain feature types may help to differentiate "spike" vs. non-spike.

It has previously been shown that across coronaviruses, prior to the emergence of SARS-CoV-2, the tertiary structure of the spike protein is not conserved but the connectivity of secondary structure elements is [8]. As evidenced in Fig. 1A, the tertiary structure of the "spike" protein is clearly not conserved across different respiratory families. For the coronavirus sequences, two-tailed t-tests were performed looking at the association of %helix, %sheet, %longest sheet, %longest helix, respectively, with spike vs. non-spike status. A statistically significant association was observed for %sheet (p-value = 0.001), whereas none was for %helix (p-value = 0.087), %longest helix (p-value = 0.083) and %longest sheet (p-value = 0.208). The %longest helix was examined because when predicted secondary structure topology was examined across the SARS-CoV-2 sequence (NC_045512.2) the spike region appeared to have more longer helical segments than the other regions of the sequence; %longest sheet was added for completeness.

The pattern of N-linked glycosylation of the spike protein is highly conserved (ref) and may play a role in immune evasion [9, 33]. Again, for the coronavirus sequences, t-tests were performed examining the correlation of total N-sites and density, respectively, for spike vs. non-spike. A significant statistical difference was found for the total N-sites (p-value < 0.0001) and density (p-value = 0.010). The %AA was also examined over the coronaviruses dataset to determine if there were significant differences in amino acid composition for spike vs. non-spike. Of the 20 %AAs, a significant difference was observed for %N (p-value = 0.008), %S (p-value = 0.030), and %M (p-value = 0.032).

Based on these preliminary findings, we developed Random Forest machine learning classifiers with a feature vector that consisted of glycosylation, amino acid composition, and secondary structure element related features. To place these results in context, we compared classifier accuracy in each case to the ZeroR Scores for the same dataset. The ZeroR, which consists of a simple classification rule which simply predicts the majority category (class), provides a benchmark for classification performance. We also performed a test of association between the predicted and actual classes, using Fisher's Exact Test (ref).

Our first set of Random Forest models were developed based on the coronavirus dataset (see Table S1). All but one classified the proteins correctly 100% of the time with a ZeroR Score of 86.8% and Fisher's Exact Test of 0.006. A comparison of these five models suggests that only total N-sites and density may contribute significantly to the models. The other model (**A.1**) involving only secondary structure—%sheet, %helix, %longest sheet, %longest helix—yielded 96.2% correctly classified; that same set of features was then used to develop a model separately for each of the other four virus families. For each of these models the % correctly classified ranged from 96.2% to 100% with a sensitivity ranging from 0.86 to 1.0, and a specificity ranging

from 0.98 to 1.0. To place these results in context, the ZeroR Scores for these datasets ranged from 86.8% to 88.5%. For each of the classifiers, there was a strong association between the actual classes and the class predicted by the Random Forest, with *p*-values ranging from 0.004 to 0.065. Models based on combining total N-sites, density, %sheet, %helix, and % longest helix were also generated for each virus family (**B.1**), respectively; in this case, the % correctly classified ranged from 93.5% to 100% (compared with ZeroR Scores from 86.5% to 87.5%), and Fisher's exact test *p*-values ranging from 0.004 to 0.238. These two feature sets (associated with the A.1 and B.1 models, respectively) were then used to create cross respiratory virus family models (**A** and **B**, respectively) using the full dataset, yielding %correctly classified of 97.8% and 98.1%, respectively. A cross virus family model not including secondary structure elements (**C**) yielded significantly poorer results with a % correctly classified of 92%. These data taken together point to the robustness of the models overall.

Cross virus family models **A** and **B** are described in detail in Table 2. As a crosscheck against overfitting, we carried out bootstrapping with 1000 datasets, using resampling with replacement to generate synthetic datasets with 316 data points each. For each dataset, we built a new Random Forest model using 10-fold cross validation and evaluated its accuracy. Ninety-five percent of the Random Forest models built for the bootstrapped datasets showed an accuracy of greater than 98%, indicating that the models were in fact not overfitted to the original dataset. For model **A**, the mean and upper confidence intervals of the %correctly classified at the 95% level were 98.86% and 98.90%; while for model **B**, they were 98.82% and 98.86%.

Next, bootstrapping was performed with 1000 datasets that were 50-50 balanced for "spike" vs. non-spike to eliminate the possibility that the accuracy of the models could be due to the fact that non-spike was overrepresented in the database (although obviously the proportion

"spike" vs. non-spike is reflective of distribution). In this 50-50 balanced bootstrapping exercise, each dataset was comprised of 158 "spike" and 158 non-spike proteins, randomly sampled with replacement. In the balanced case, for model **A**, the mean and upper confidence intervals of the %correctly classified at the 95% level were 98.89% and 98.92%; while for model **B**, they were 99.68% and 99.84% (Table S2). This can be compared against the ZeroR score of 50% that is expected in a 50-50 balanced dataset. Thus, models **A** and **B** can successfully differentiate "spike" from non-spike respiratory virus sequence without specifying the viral family.

Finally, the capability of a ML model to identify the virus family from the sequence using the same feature vectors was explored. The Random Forest model generated was 86% accurate in predicting the virus family (see Table 2). This result is particularly impressive given that the ZeroR baseline performance indicator was only 22.36%.

In summary, the models developed by us in this work can correctly identify viral "spike" proteins from viruses (within the five viral families examined here) within seconds. The ability to utilize ML models to predict the protein responsible for cell entry (the "spike") from a viral sequence as well as to predict the virus family of a novel viral sequence may in the future expedite the development of biomedical interventions for respiratory pandemics. In addition, the predictiveness of the models points to the underlying importance of secondary structure and N-glycosylation in viral host cell recognition.

Acknowledgments

We would like to thank the Boston University College of Engineering STARS program for support for SD and MS.

References

1. Greenhalgh, T., et al., *Ten scientific reasons in support of airborne transmission of SARS-CoV-2.* Lancet, 2021. **397**(10285): p. 1603-1605.
2. Ravina, et al., *A changing trend in diagnostic methods of Influenza A (H3N2) virus in human: a review.* 3 Biotech, 2021. **11**(2): p. 87.
3. Thomas, E., S. Delabat, and D.M. Andrews, *Diagnostic Testing for SARS-CoV-2 Infection.* Curr Hepatol Rep, 2021: p. 1-9.
4. Benda, A., et al., *COVID-19 Testing and Diagnostics: A Review of Commercialized Technologies for Cost, Convenience and Quality of Tests.* Sensors (Basel), 2021. **21**(19).
5. Almehdi, A.M., et al., *SARS-CoV-2 spike protein: pathogenesis, vaccines, and potential therapies.* Infection, 2021. **49**(5): p. 855-876.
6. Jin, D., J. Wei, and J. Sun, *Analysis of the molecular mechanism of SARS-CoV-2 antibodies.* Biochem Biophys Res Commun, 2021. **566**: p. 45-52.
7. Zieneldien, T., et al., *COVID-19 Vaccines: Current Conditions and Future Prospects.* Biology (Basel), 2021. **10**(10).
8. Li, F., *Evidence for a common evolutionary origin of coronavirus spike protein receptor-binding subunits.* J Virol, 2012. **86**(5): p. 2856-8.
9. Watanabe, Y., et al., *Site-specific glycan analysis of the SARS-CoV-2 spike.* Science, 2020. **369**(6501): p. 330-333.
10. Sleator, R.D. and P. Walsh, *An overview of in silico protein function prediction.* Arch Microbiol, 2010. **192**(3): p. 151-5.
11. Grant, M.A., *Integrating computational protein function prediction into drug discovery initiatives.* Drug Dev Res, 2011. **72**: p. 4-16.

12. Cruz, L.M., et al., *Protein Function Prediction.* Methods Mol Biol, 2017. **1654**: p. 55-75.

13. Loewenstein, Y., et al., *Protein function annotation by homology-based inference.* Genome Biol, 2009. **10**(2): p. 207.

14. Aloy, P., et al., *Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking.* J Mol Biol, 2001. **311**(2): p. 395-408.

15. Gligorijevic, V., et al., *Structure-based protein function prediction using graph convolutional networks.* Nat Commun, 2021. **12**(1): p. 3168.

16. Li, S., et al., *A fast protein binding site comparison algorithm for proteome-wide protein function prediction and drug repurposing.* Proteins, 2021. **89**(11): p. 1541-1556.

17. Piovesan, D. and S.C.E. Tosatto, *INGA 2.0: improving protein function prediction for the dark proteome.* Nucleic Acids Res, 2019. **47**(W1): p. W373-W378.

18. Nagano, N., C.A. Orengo, and J.M. Thornton, *One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions.* J Mol Biol, 2002. **321**(5): p. 741-65.

19. Naeem, A., et al., *The accuracy of protein structure alignment servers.* Electronic Journal of Biotechnology, 2016. **20**: p. 9-13.

20. Chitale, M., et al., *ESG: extended similarity group method for automated protein function prediction.* Bioinformatics, 2009. **25**(14): p. 1739-45.

21. Jain, A. and D. Kihara, *Phylo-PFP: improved automated protein function prediction using phylogenetic distance of distantly related sequences.* Bioinformatics, 2019. **35**(5): p. 753-759.

22. Martin, D.M., M. Berriman, and G.J. Barton, *GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes.* BMC Bioinformatics, 2004. **5**: p. 178.

23. Cozzetto, D., et al., *Protein function prediction by massive integration of evolutionary analyses and multiple data sources.* BMC Bioinformatics, 2013. **14 Suppl 3**: p. S1.

24. You, R., et al., *GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank.* Bioinformatics, 2018. **34**(14): p. 2465-2473.

25. Drozdetskiy, A., et al., *JPred4: a protein secondary structure prediction server.* Nucleic Acids Res, 2015. **43**(W1): p. W389-94.

26. Cuff, J.A. and G.J. Barton, *Application of multiple sequence alignment profiles to improve protein secondary structure prediction.* Proteins, 2000. **40**(3): p. 502-11.

27. Blom, N., et al., *Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence.* Proteomics, 2004. **4**(6): p. 1633-49.

28. Gupta, R. and S. Brunak, *Prediction of glycosylation across the human proteome and the correlation to protein function.* Pac Symp Biocomput, 2002: p. 310-22.

29. Frank, E., Hall, M.A., Witten, I.H., *The WEKA Workbench,* in *Data Mining: Practical Machine Learning Tools and Techniques,* I.H. Witten, Frank, E., Hall, M.A., Pal, C.J., Editor. 2016, Morgan Kaufmann: Burlington, MA USA.

30. Ho, T.K., *The random subspace method for constructing decision forests.* IEEE transactions on pattern analysis and machine intelligence, 1998. **20**(8): p. 832--844.

31. Tin Kam, H., *Random decision forests.* 1995. **1**: p. 278--282 vol.1.

32. Fisher, R.A., *On the interpretation of χ2 from contingency tables, and the calculation of P.* Journal of the Royal Statistical Society, 1922. **85**(1): p. 87–94.

33. Zhou, D., et al., *Identification of 22 N-glycosites on spike glycoprotein of SARS-CoV-2 and accessible surface glycopeptide motifs: Implications for vaccination and antibody therapeutics.* Glycobiology, 2021. **31**(1): p. 69-80.

34. Hu, B., et al., *Characteristics of SARS-CoV-2 and COVID-19.* Nat Rev Microbiol, 2021. **19**(3): p. 141-154.

35. Cevik, M., et al., *SARS-CoV-2, SARS-CoV, and MERS-CoV viral load dynamics, duration of viral shedding, and infectiousness: a systematic review and meta-analysis.* Lancet Microbe, 2021. **2**(1): p. e13-e22.

36. Xiao, S., et al., *A study of the probable transmission routes of MERS-CoV during the first hospital outbreak in the Republic of Korea.* Indoor Air, 2018. **28**(1): p. 51-63.

37. St-Jean, J.R., et al., *Human respiratory coronavirus OC43: genetic stability and neuroinvasion.* J Virol, 2004. **78**(16): p. 8824-34.

38. Wong, A.H.M., et al., *Receptor-binding loops in alphacoronavirus adaptation and evolution.* Nat Commun, 2017. **8**(1): p. 1735.

39. Linster, M., et al., *Clinical and Molecular Epidemiology of Human Parainfluenza Viruses 1-4 in Children from Viet Nam.* Sci Rep, 2018. **8**(1): p. 6833.

40. Battles, M.B. and J.S. McLellan, *Respiratory syncytial virus entry and how to block it.* Nat Rev Microbiol, 2019. **17**(4): p. 233-245.

41. Yi, L., et al., *Epidemiology, evolution and transmission of human metapneumovirus in Guangzhou China, 2013-2017.* Sci Rep, 2019. **9**(1): p. 14022.

42. Hong, J.Y., et al., *Lower respiratory tract infections due to adenovirus in hospitalized Korean children: epidemiology, clinical features, and prognosis.* Clin Infect Dis, 2001. **32**(10): p. 1423-9.

43. Biggs, H.M., et al., *Adenovirus-Associated Influenza-Like Illness among College Students, Pennsylvania, USA.* Emerg Infect Dis, 2018. **24**(11): p. 2117-2119.

44. Kajon, A.E., et al., *Adenovirus Type 4 Respiratory Infections among Civilian Adults, Northeastern United States, 2011-2015(1).* Emerg Infect Dis, 2018. **24**(2): p. 201-209.

45. Bruckova, M., et al., *An outbreak of respiratory disease due to a type 5 adenovirus identified as genome type 5a.* Acta Virol, 1980. **24**(3): p. 161-5.

46. Lamson, D.M., et al., *Detection and Genetic Characterization of Adenovirus Type 14 Strain in Students with Influenza-Like Illness, New York, USA, 2014-2015.* Emerg Infect Dis, 2017. **23**(7): p. 1194-1197.

47. Sun, B., et al., *Emergent severe acute respiratory distress syndrome caused by adenovirus type 55 in immunocompetent adults in 2013: a prospective observational study.* Crit Care, 2014. **18**(4): p. 456.

48. Krammer, F., et al., *Influenza.* Nat Rev Dis Primers, 2018. **4**(1): p. 3.

Table 1. Respiratory Virus Sequences Used in Model Development

Virus Family	Virus[a]	Strain	Sequence Identifier[b]
Coronaviridae	SARS-CoV-2 [34]	Wuhan-Hu-1	NC_045512.2
Coronaviridae	SARS-CoV-1 [35]	Tor2	NC_004718.3
Coronaviridae	MERS [36]	HCoV-EMC/2012	NC_019843.3
Coronaviridae	hCoV-OC43 [37]	ATCC VR-759	NC_006213.1
Coronaviridae	hCoV-HKU1 [38]	HKU1	NC_006577.2
Coronaviridae	hCoV-NL63 [38]	Amsterdam I	NC_005831.2
Coronaviridae	hCoV-229E [38]	299E	NC_002645.1
Paramyxoviridae	HPIV 1 [39]	Washington 1964	NC_003461.1
Paramyxoviridae	HPIV 2 [39]	VIROAF10	KM190939.1*
Paramyxoviridae	HPIV 3 [39]	GP	NC_001796.2
Paramyxoviridae	HPIV 4a [39]	M-25	NC_021928.1
Pneumoviridae	HRSV [40]	Subgroup A	NC_038235.1
Pneumoviridae	HRSV	CA-17	LC385004.1*
Pneumoviridae	HRSV	CA-15	LC385003.1*
Pneumoviridae	HRSV	KW-15	LC385002.1*
Pneumoviridae	HMPV [41]	PER/FPP00726/2011/A	KJ627437.1*
Pneumoviridae	HMPV	Isolate 00-1	NC_039199.1
Pneumoviridae	HMPV	PER/IPE00957/2012/A	KJ627433.1*
Pneumoviridae	HMPV	Seattle/USA/SC0380/2019	MN306028.1*
Pneumoviridae	HMPV	01/KEN/2015	MK588634.1*
Pneumoviridae	HMPV	USA/NM013/2016	KY474543.1*
Pneumoviridae	HMPV	BuenosAires/ARG/001/2016	MG773272.1*
Pneumoviridae	HMPV	AUS/183219938/2004/B	KF530178.1*
Adenoviridae	HAdV [42]	Type 2	J01917.1*
Adenoviridae	HAdV [43]	Type 3	DQ086466.1*
Adenoviridae	HAdV [44]	Type 4	KF006344.1*
Adenoviridae	HAdV [45]	Type 5	AC_000008.1
Adenoviridae	HAdV [43]	Type 7	AC_000018.1

Adenoviridae	HAdV [46]	Type 14	AY803294.1*
Adenoviridae	HAdV [42]	Type 35	AC_000019.1
Adenoviridae	HAdV [47]	Type 55	MG905110.1*
Orthomyxoviridae	Influenza Virus A [48]	A/chicken/Morocco/SF5/2016 (H9N2)	LT598501.1* LT598506.1* LT598511.1* LT598516.1* LT598521.1* LT598526.1* LT598531.1* LT598536.1*
Orthomyxoviridae	Influenza Virus A	A/California/07/2009 (H1N1)	YP_009118626.1 YP_009118628.1 CY121687.1* KU933483.1* CY121682.1* CY121684* KU933488.1* CY121683.1*
Orthomyxoviridae	Influenza Virus A	A/Berlin/3/1964 (H2N2)	ACD85187.1* ACD85195.1* ACD85197.1* ACD85194.1* ACD85190.1* ACD85192.1* ACD85188.1* ACD85191.1*
Orthomyxoviridae	Influenza Virus A	A/Shanghai/02/2013 (H7N9)	NC_026425.1 NC_026423.1 NC_026422.1 NC_026424.1 NC_026429.1 NC_026428.1 NC_026427.1 NC_026426.1
Orthomyxoviridae	Influenza Virus A	A/ruddy turnstone/Delaware Bay/262/2006 (H7N3)	ACO95657.1* ACO95665.1* ACO95667.1* ACO95664.1*

			ACO95660.1* ACO95662.1* ACO95658.1* ACO95661.1*
Orthomyxoviridae	Influenza Virus A	A/Chicken/Hong Kong/715.5/01 (H5N1)	AF509025.1* AF509178.2* AF509152.2* AF509204.2* AF509100.2* AF509075.1* AF509049.1* AF509126.2*
Orthomyxoviridae	Influenza Virus A	A/swine/France/IIIeetVilaine-0346/2011 (H1N2)	KC894804.1* KR701484.1* KR701483.1* KR701485.1* KC894807.1* KR701488.1* KR701487.1* KR701486.1*
Orthomyxoviridae	Influenza Virus A	A/swine/Texas/4199-2/1998(H3N2))	AEK70342.1 AAD51248.1 AEK70339.1 AEK70341.1 AEK70343.1 AEK70344.1 AEK70345.1 AEK70347.1

[a] MERS = Middle East Respiratory Syndrome, HPIV = human parainfluenza virus, HRSV = human respiratory syncytial virus, HMPV = human metapneumovirus, HAdV=human adenovirus; references indicate that the virus is responsible for respiratory disease.
[b] A * indicates that the sequence is an NCBI Reference Sequence.

Table 2. ML Models Successfully Differentiate Spike from Non-Spike[a, b, c]

Model	Features	ZeroR Score	Correctly Classified	Fisher's Exact Test	Bootstrapping Mean/Upper CI (95%)
A	%sheet, %helix, %longest sheet, %longest helix	87.66%	97.78%	0.0000003	98.86%/98.90%
B	total N-sites, density, %sheet, %helix, %longest helix	87.66%	98.10%	0.00000009	98.82%/98.86%

[a] For 5 virus families, 39 viral sequences, 316 individual proteins
[b] Random forest (RF) scores calculated in Weka with 10-fold cross-validation (CV)
[c] 1000 resampled with replacement bootstrapping datasets as input into Weka RF tests with 10-fold CV yielding 10,000 values of "% Correctly Classified";

Table 3. ML Also Identifies Virus Families from Sequence [a]

Virus Families	Features	ZeroR Score	Correctly Classified	Fisher's Exact Test
5	total N-sites, density, %sheet, %helix, %longest helix	32.91%	87.97%	< 0.0000001

[a] Family class used as the output attribute (instead of "spike vs. non-spike").

Figure Legends

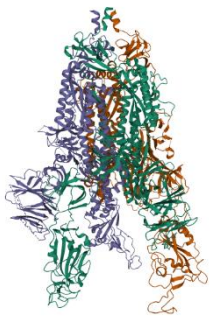
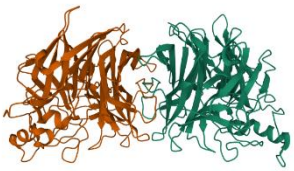
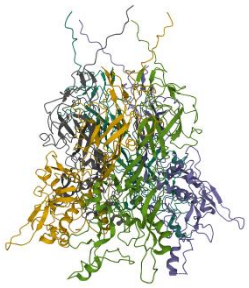
FIG 1. Five families of respiratory viruses and their "spike" proteins. In (A) the identity and representative structure of the "spike" protein (gene name given in parentheses) is shown for each of the virus families studied. PDB identifiers for structures 1-5 are also listed with the corresponding virus indicated. Shown in (B) is a schematic of the coronavirus SARS-CoV-2 structure indicating the prominence of the spike.

FIG 2. Overall model development workflow. The procedure for the development of ML models to differentiate Spike from non-Spike in a sequence

FIG 3. Random forest inputs vs. outputs.

FIG 4. Schematic of bootstrapping process for cross validation of selected models. In this case, each of the 1000 bootstrapped datasets contains feature vectors for 316 protein sequences.

A

1. **Coronaviridae:** Spike Glycoprotein (S2)		2. **Paramyxoviridae:** Hemagglutinin Neuraminidase Glycoprotein (HN)	
3. **Pneumoviridae:** Fusion Glycoprotein F2 (F)		4. **Adenoviridea:** Penton Protein (L2)	
5. **Ortho-myxoviridae:** Hemagglutinin Glycoprotein (HA)		PDB identifier: Organism 1. 7KJ4: SARS-CoV-2 2. 1V3E: Human parainfluenza virus 3 3. 6OUS: Human respiratory syncytial virus A2 4. 3IZO: Human adenovirus 5 5. 2WRG: Influenza virus A	

B

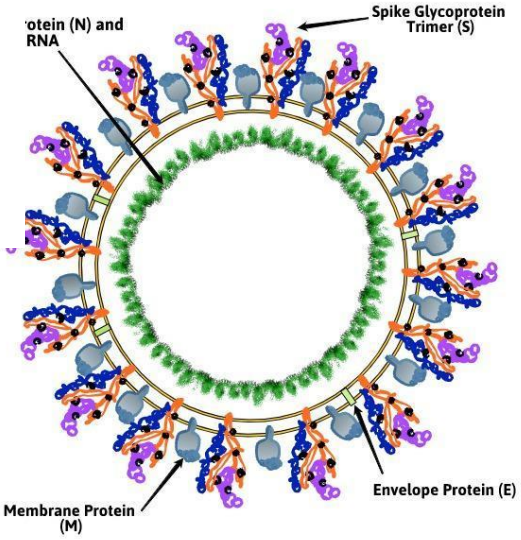


FIG 1. Five families of respiratory viruses and their "spike" proteins. In (A) the identity and representative structure of the "spike" protein (gene name given in parentheses) is shown for each of the virus families studied. PDB identifiers for structures 1-5 are also listed with the corresponding virus indicated. Shown in (B) is a schematic of the coronavirus SARS-CoV-2 structure indicating the prominence of the spike.

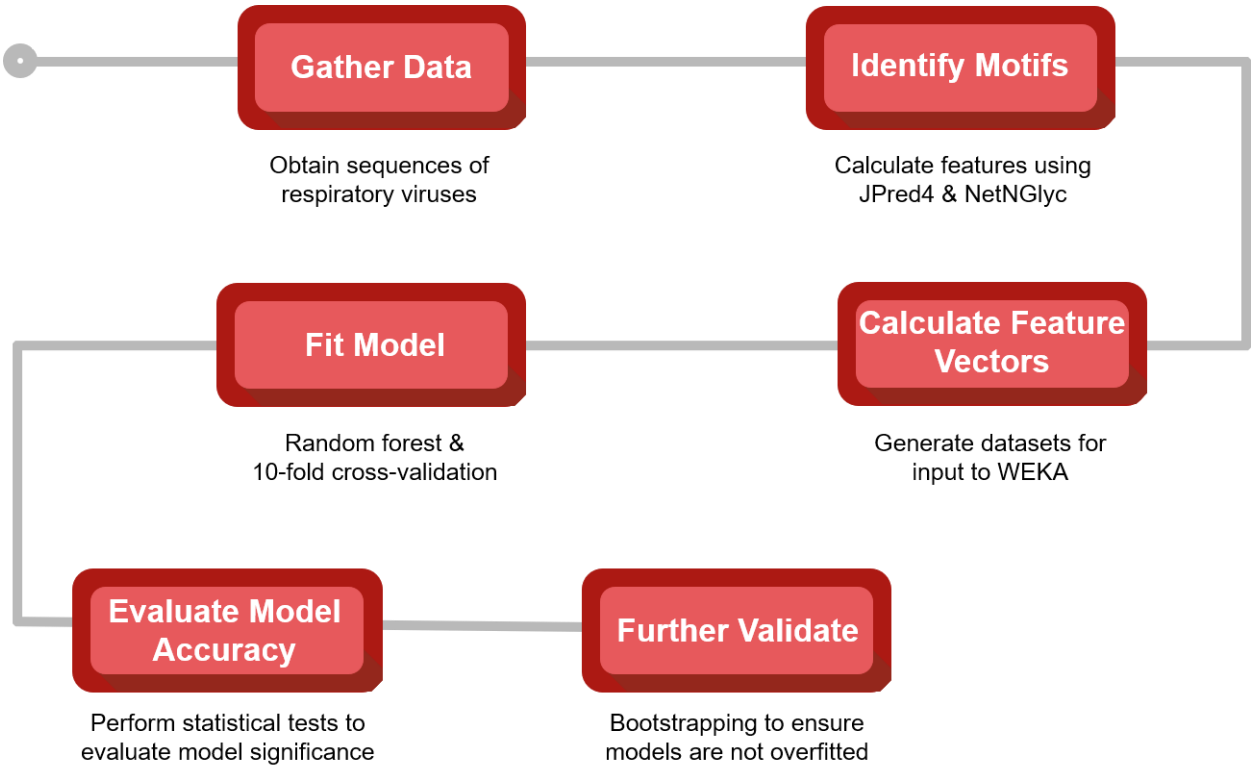


FIG 2. Overall model development workflow. The procedure for the development of ML models to differentiate Spike from non-Spike in a sequence.

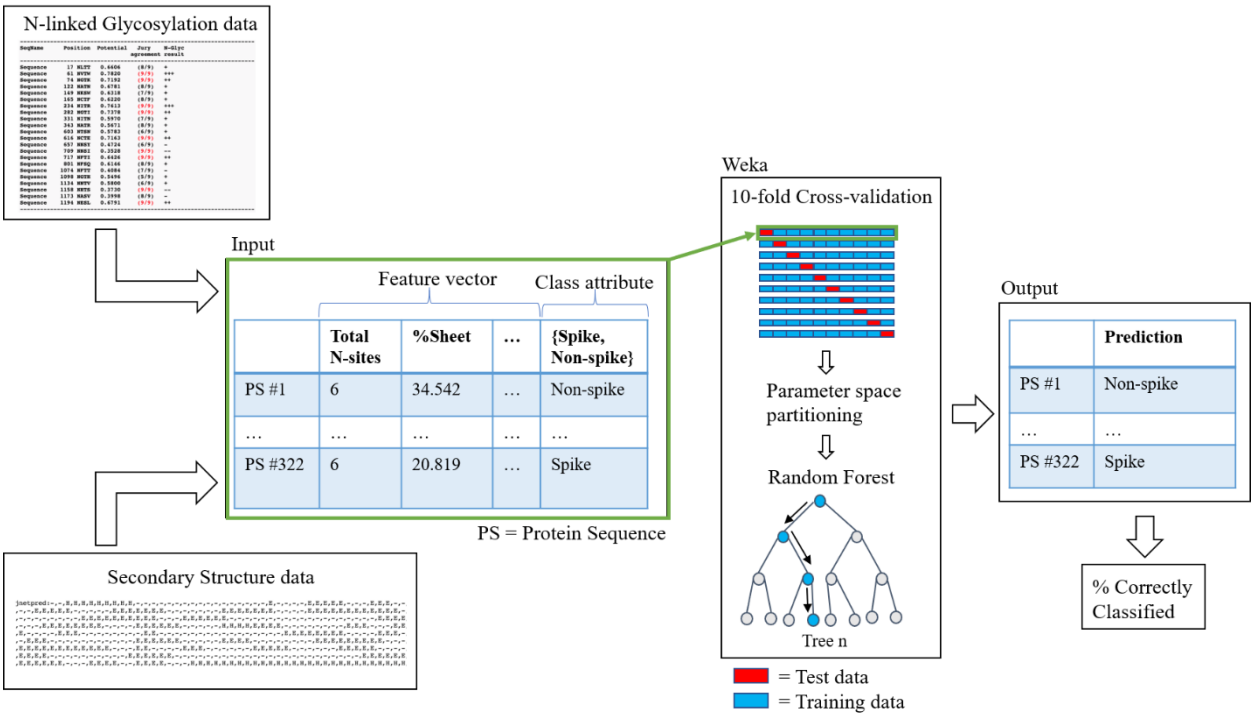


FIG 3. Random forest inputs, cross validation, and outputs. Data was input for 316 protein sequences.

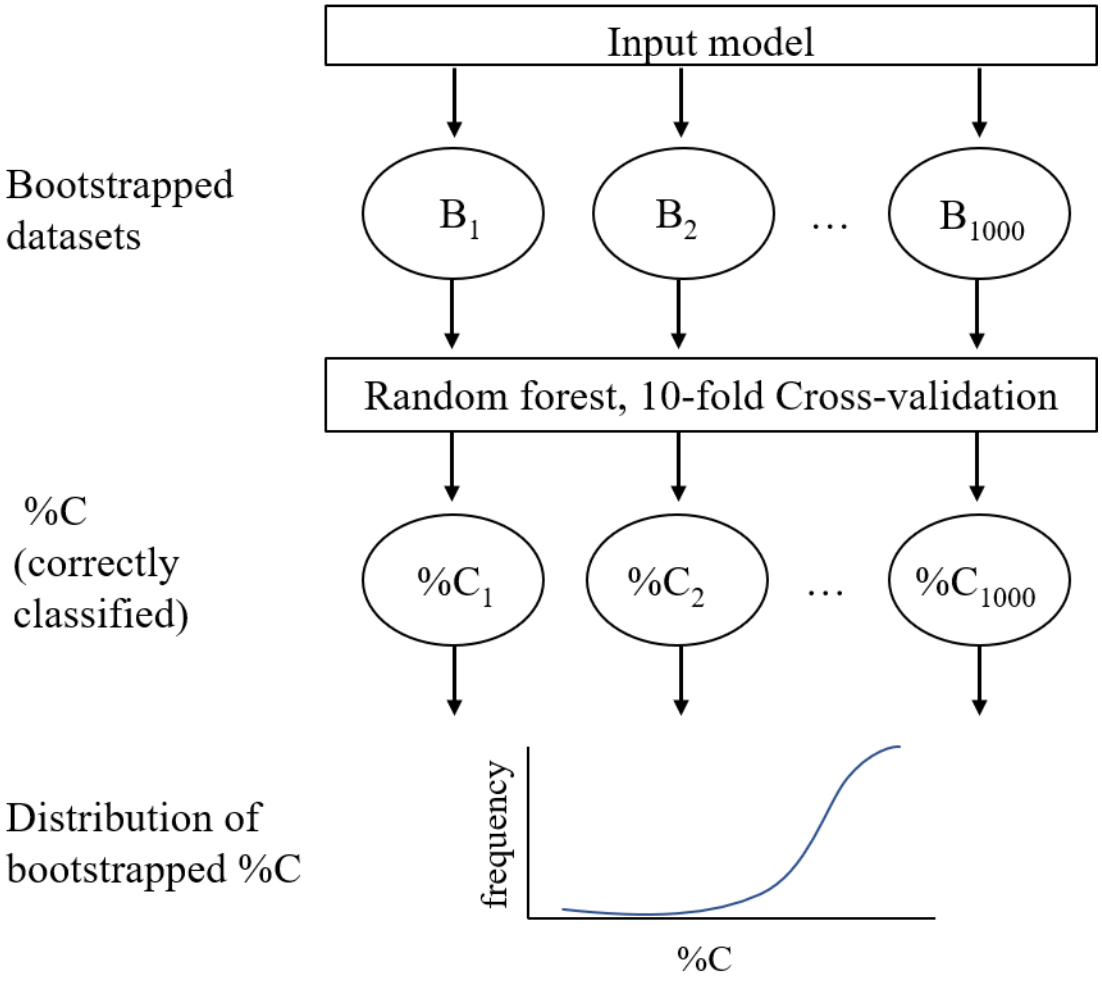


FIG 4. Schematic of bootstrapping process for cross validation of selected models. In this case, each of the 1000 bootstrapped datasets contains feature vectors for 316 protein sequences.

Supplemental Material Figure Legends

FIG S1: Calculation of secondary structure elements. A flowchart showing the process for calculating the secondary structure elements with Jpred4.

Table S1. All ML Models Examined

Model #	Viral Family	Features	ZeroR Score	Correctly Classified	Fisher's Exact Test
1, **C.1**	Coronaviridae	total N-sites, density, %N, %M, %S	86.793%	100.000%	0.006
2	Coronaviridae	total N-sites, density, %M, %S	86.793%	100.000%	0.006
3	Coronaviridae	total N-sites, density	86.793%	100.000%	0.006
4, **D.1**	Coronaviridae	total N-sites, density, %M, %S, %sheet, %helix, %longest sheet, %longest helix	86.793%	100.000%	0.006
5, **B.1**	Coronaviridae	total N-sites, density, %sheet, %helix, %longest helix	86.793%	100.000%	0.006
6, **A.1**	Coronaviridae	%sheet, %helix, %longest sheet, %longest helix	86.793%	96.226%	0.065
7, **A.1**	Paramyxoviridae	%sheet, %helix, %longest sheet, %longest helix	87.097%	100.000%	0.056
8, **A.1**	Pneumoviridae	%sheet, %helix, %longest sheet, %longest helix	88.462%	98.077%	0.004
9, **A.1**	Adenorividae	%sheet, %helix, %longest sheet, %longest helix	87.500%	98.438%	0.015
10, **A.1**	Orthomyxoviridae	%sheet, %helix, %longest sheet, %longest helix	87.500%	96.875	0.039
11, **B.1**	Paramyxoviridae	total N-sites, density, %sheet, %helix, %longest helix	87.097%	93.548%	0.238
12, **B.1**	Pneumoviridae	total N-sites, density, %sheet, %helix, %longest helix	86.462%	98.077%	0.004
13, **B.1**	Adenorividae	total N-sites, density, %sheet, %helix, %longest helix	87.500%	98.438%	0.015
14, **B.1**	Orthomyxoviridae	total N-sites, density, %sheet, %helix, %longest helix	87.500%	96.875%	0.039
15, **C**	Coronaviridae + Paramyxoviridae + Pneumoviridae +	total N-sites, density, %N, %M, %S	87.658%	92.721%	0.011

	Adenorividae + Orthomyxoviridae				
16, **D**	Coronaviridae + Paramyxoviridae + Pneumoviridae + Adenorividae + Orthomyxoviridae	total N-sites, density, %M, %S, %sheet, %helix, %longest sheet, %longest helix	87.658%	98.101%	<0.001
17, **A**	Coronaviridae + Paramyxoviridae + Pneumoviridae + Adenorividae + Orthomyxoviridae	%sheet, %helix, %longest sheet, %longest helix	87.658%	97.785%	<0.001
18, **B**	Coronaviridae + Paramyxoviridae + Pneumoviridae + Adenorividae + Orthomyxoviridae	total N-sites, density, %sheet, %helix, %longest helix	87.658%	98.101%	<0.001

Table S2. ML Models Successfully Differentiate Spike from Non-Spike [a, b, c] for a 50-50 balanced bootstrapping dataset [d]

Model*	Features	Bootstrapping Mean/Upper CI (95%)
A-50-50	%sheet, %helix, %longest sheet, %longest helix	98.89% / 98.92%
B-50-50	total N-sites, density, %sheet, %helix, %longest helix	99.68%/99.84%

[a] For 5 virus families, 39 viral sequences, 316 individual proteins
[b] Random forest (RF) scores calculated in Weka with 10-fold cross-validation (CV)
[c] 1000 resampled with replacement bootstrapping datasets as input into Weka RF tests with 10-fold CV yielding 10,000 values of "% Correctly Classified"
[d] Each model has 158 spike and 158 non-spike proteins, randomly sampled with replacement

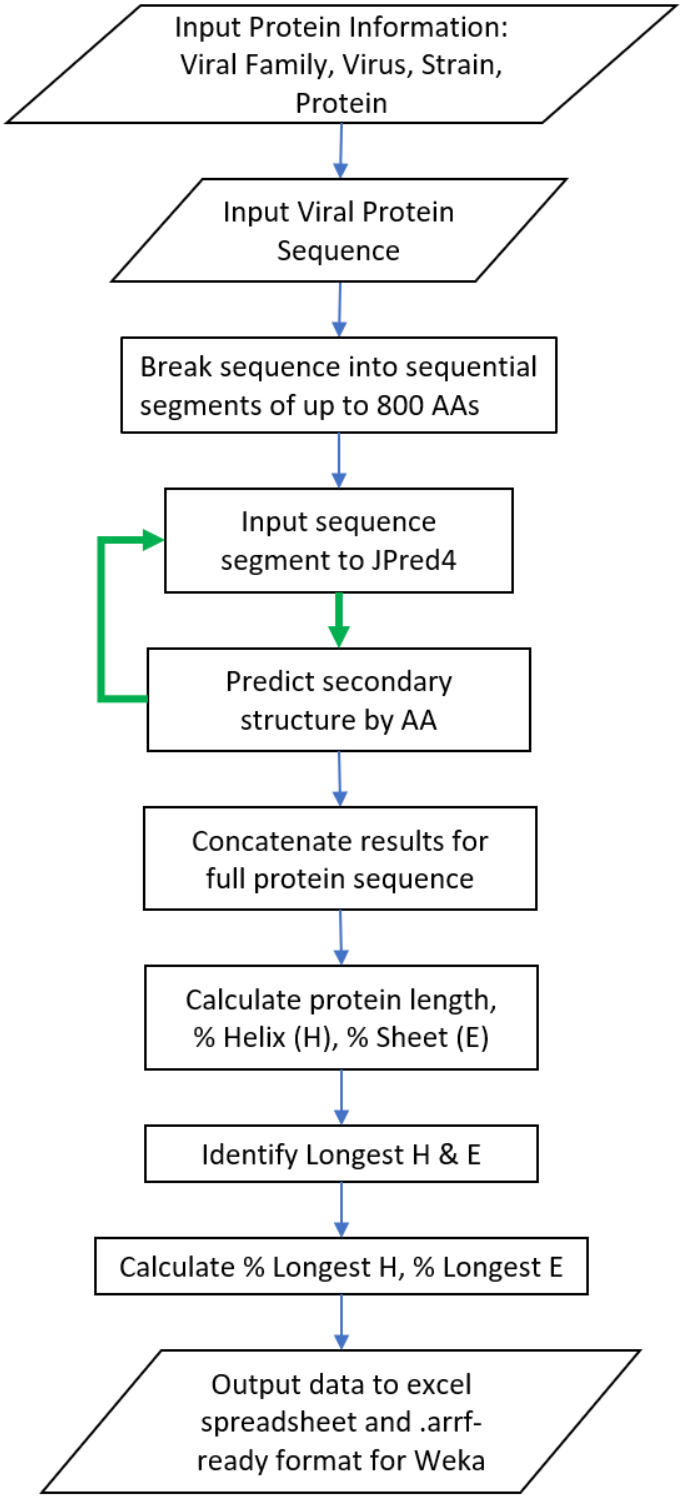


FIG. S1: Calculation of secondary structure elements. A flowchart showing the process for calculating the secondary structure elements with Jpred4.