THE ROLE OF MACHINE LEARNING IN SARS-CoV-2 SUSCEPTIBILITY CLASSIFICATION

by

Zak White
University of Victoria

A graduating thesis submitted in partial fulfilment of the requirements for the Honours degree of

BACHELOR OF SCIENCE

in the Department of Computer Science

We acknowledge with respect the Lekwungen peoples on whose traditional territory the university stands and the Songhees, Esquimalt, and WSANEC peoples whose historical relationships with the land continue to this day.

THE ROLE OF MACHINE LEARNING IN
SARS-CoV-2 SUSCEPTIBILITY CLASSIFICATION

by

Zak White
University of Victoria

Abstract

SARS-CoV-2 is a contagious virus established to affect not only humans, but other mammal species. Studies over the last two years have revealed certain species are distinctly immune to the virus, which can be attributed to differences in the ACE2 protein, the virus' target protein, in various hosts. This study applies machine learning methods to classify hosts as susceptible or immune to SARS-CoV-2 based on their ACE2 sequences.

Machine learning has faced criticism within the field of biology for its uninterpretable logic; it's imperative that biologists and medical professionals can have confidence in the tools they use, and this isn't always possible with machine learning. This is an investigation into the importance of explainable machine learning within bioinformatics that involves the comparison of three models with varying degrees of biological considerations.

This study validated the hypothesis that biology-driven machine learning applications outperform pure machine learning models, and produced other interesting findings on how mutations in the ACE2 protein can affect susceptibility to SARS-CoV-2.

Contents

List of Tables and Figures

Acknowledgments

I would like to thank:

My thesis supervisor **Dr. Hosna Jabbari** for offering me the opportunity to perform this research, inspiring the project topic, and for her guidance through the process. Further, the members of the **Computational Biology Research and Analytics (COBRA)** group at the University of Victoria for their resources and feedback.

My supportive friends.

Most profoundly, **my parents and family** for their endless unconditional support and love at every step of my life leading to this academic achievement.

Dedication

To Isla.

Everything in this world is yours.

1 Introduction

For over two years, the COVID-19 pandemic has been a headline research topic as it inescapably invades our world and affects our lives in unparalleled ways. COVID-19 is a disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a virus believed have zoonotic origins [1].

Since its discovery, SARS-CoV-2 studies have identified a range of animal hosts — including some nonhuman primates, mustelids, bats, and felines — that can be infected by and can transmit the virus [2–8]. Conversely, certain research has supported the conclusion that other species (e.g. the domestic pig) are explicitly immune to the virus [2, 7, 9, 10]. This disjunction promotes the question: *what makes some species susceptible to the virus and others immune?* Investigating this question has been an ongoing focus in bioinformatics to mitigate interspecies transmission.

The answer might correspond to how the the virus interacts with its target protein in different hosts. The SARS-COV-2 spike protein's target is the angiotensin-converting enzyme 2 (ACE2). Differences in the ACE2 sequence can affect the spike protein's binding affinity, thus affecting the host's susceptibility to the virus.

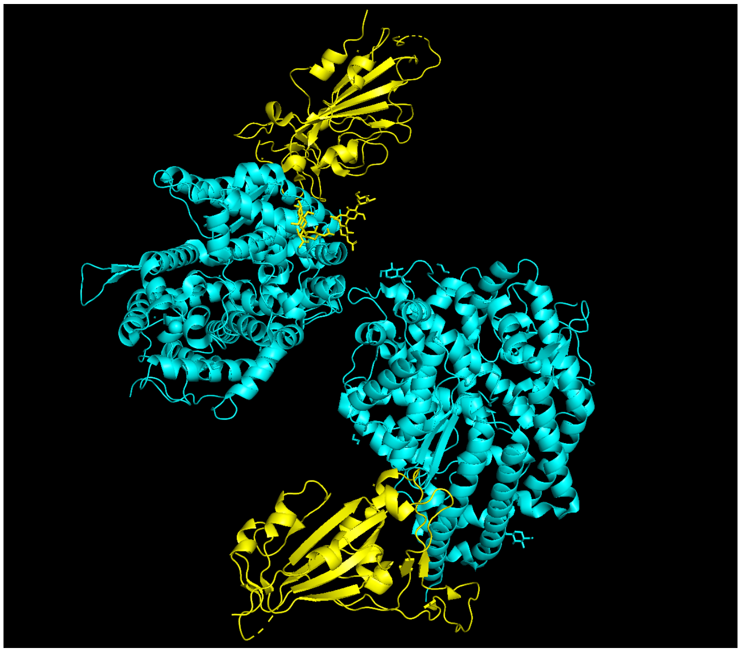

Figure 1.1: Structure of SARS-CoV-2 spike protein (yellow) complexed with its target ACE2 (cyan)

1.1 Machine Learning

Machine learning is a branch of artificial intelligence used to build models for classification or decision-making based on data, without explicit programming. Nowadays, machine learning is used across industrial fields to plan, make correct decisions, and increase the efficiency of the processes [11].

Within the field of biology, machine learning has outperformed human analysis in certain tasks. For example, in 2018, Kakadiaris *et al.* designed a support-vector machine model for classifying risk of cardiovascular disease that produced an accuracy of over 85% and missed fewer disease events than the US guidelines' manual risk calculator.

Despite its success, the use of machine learning within biology has been met with criticism. Models, especially black-box models (where the internal working is not revealed), have been condemned for their uninterpretable logic, and for not being paired with a biological explanation [13]. Bioinformatic problems often represent high-stakes decisions (e.g. diagnoses), so it is imperative that models used to address the problems can be trusted from a biological standpoint.

1.2 Sequence Classification

Proteins are biomolecules that comprise chains of organic compound units known as amino acids. There are twenty amino acids, each of which can be represented using a letter character. By defining a protein as a sequence of characters corresponding to the amino acid chain, the sequence can be used as a feature vector for classification. Take the human ACE2 sequence, built of a chain of 805 acid residues:

$$MSSSSW...DVQTSF$$

where each letter corresponds to a distinct acid (methionine as M, serine as S, etc.). The sequence can be converted into a feature vector:

$$x = (M, S, S, S, W, ..., D, V, Q, T, S, F)$$

which can be used for classification. The features are categorical (limited to a fixed set of values) and nonordinal (there is no logical sorting order).

The basis of this study is using the amino acid feature vectors of a protein to classify sequences as susceptible or immune to SARS-CoV-2.

1.3 Objective and Contribution

This study is an investigation into the validity of machine learning for classifying sequence susceptibility to SARS-CoV-2, to gain insight into the validity of machine learning as a bioinformatic tool at a higher level.

Three classification models are designed:

- a **baseline** machine learning model;

- an **eliminative** machine learning model; and
- a **structural** model that incorporates the findings of a protein structure analysis

The models are tested against a set of host protein sequences to evaluate their performances. It is hypothesized that the structural model will outperform its machine learning counterparts as the more accurate classifier.

1.4 Related Work

Machine learning has been growing in popularity for sequence classification [14], but an ongoing barrier is the magnitude of the feature vectors: proteins can be built of hundreds or thousands of amino acids, so reducing the problem to a manageable number of features can be challenging [15]. Bioinformaticians including Iqbal *et al.* (2014) and Saidi *et al.* (2010) have developed methods for feature selection specifically designed for reducing size of the feature vector of protein sequences, each leading to models that outperform standard machine learning tools.

Patel (2016) reported linear discriminant analysis was a more accurate method than support vector machines, naive Bayes, and logistic regression when classifying amino acid sequences as secretory or not. Discriminant analysis involves applying statistical methods to find a set of features that best separates the classes [17]. The principals of discriminant analysis were implemented in the machine learning models of this study: training sequences were analyzed to identify which acids were most influential to susceptibility, and those acids were used as indicators for susceptibility in the models.

Many studies have included structural investigations into how changes in the ACE2 sequence affects binding with the SARS-CoV-2 spike protein:

- In an overview of ACE2 and its function, Li and Qin (2021) reported major interaction sites including positions 24-42, 79-83, and 353-357, emphasizing positions 38, 41, 42, 82, 83, 353, and 355.
- Zhao *et al.* (2020) investigated the effect of mutations in 23 critical sites, highlighting the effects of positions 24, 27, 30, 31, 34, 41, 79, 82, 83, 325, 329, and 353.
- Luan *et al.* (2020) consolidated the results of existing structural analyses and used their findings to predict the host range of the virus (the set of species the virus is capable of infecting), focusing on the acids in positions 31, 35, 38, 82, and 353.

2 Method

2.1 Data

Fourty-one ACE2 sequences from various animal hosts were collected from the NCBI protein database[1]. Each host has been identified as either susceptible or insusceptible to the virus, based on:

- investigations into non-human primates as models for human infection;
- research into potential origin species (bats, pangolins);
- lab studies that assessed binding affinities between the SARS-CoV-2 spike protein and ACE2 in various hosts;
- studies focusing on transmission between humans and livestock, house pets, and city animals; and
- reports on city and zoo animals contracting the virus.

Additional sequences were collected to classify hosts whose susceptibilities are unknown. Each sequence was aligned with the human ACE2 sequence using the Needleman-Wunsch algorithm (via the EMBOSS Needle pairwise sequence alignment tool[2]) to produce an optimal global alignment. This preprocessing step ensured that indexing of each sequence correlated to that of the human ortholog.

For the structural analysis, three protein models were retrieved from the AlphaFold Protein Structure Database[3], and one structure modelling the interaction between the SARS-CoV-2 spike protein and human ACE2 was retrieved from the RCSB Protein Data Bank[4].

Exhaustive lists of sequences are attached in Appendix A.

2.2 Baseline Model

First, a baseline classification model was created using a feature importance algorithm which scored the acids at each index as a feature. This step was implemented using the scikit-learn[5] machine learning package's univariate feature selection tools, with the features scored using chi-squared tests with a 97.5% confidence threshold ($\alpha = 0.025$).

This algorithm identified which acids in the training data had the greatest effect on susceptibility, and the results were used to create a categorical binary decision tree.

[1] https://www.ncbi.nlm.nih.gov/protein

[2] https://www.ebi.ac.uk/Tools/psa/emboss_needle/

[3] https://alphafold.ebi.ac.uk/

[4] https://www.rcsb.org/

[5] https://scikit-learn.org/

2.3 Eliminative Model

A second model was designed by extending on the results of an existing study (White, 2021) that focused on mutations that eliminate binding affinity between ACE2 and the SARS-CoV-2 spike protein. The original analysis involved an iterative algorithm that isolated mutations that appeared exclusively in immune sequences (Algorithm 2.1).

Algorithm 2.1: Finding potential influential mutations

```
for each immune sequence $S^-$ do
    for each $i$ <length of $S^-$ do
        $isInfluentialIndex \leftarrow$ True;
        for each susceptible sequence $S^+$ do
            if $i^{th}$ acid in $S^-$ == $i^{th}$ acid in $S^+$ then
                $isInfluentialIndex \leftarrow$ False;
            end
        end
        if isInfluentialIndex then
            // The mutation may be influential
            Record $i$, $i^{th}$ acid in $S^-$
        end
    end
end
```

The mutations were scored based on frequency in immune sequences and inversely based on the total number of mutations in the sequence. The mutations with scores that exceeded a defined threshold ($\alpha = 0.5$) were used as the basis for another binary decision tree.

2.4 Structural Model

A reduced model was created by pairing the findings of the eliminative model with a structural analysis, with the goal of removing extraneous results. For each mutation identified in the original work used for the previous model, a manual structural investigation into the mutation was conducted using the molecular visualization tool PyMOL[6], involving:

- comparing the structure of the mutated sequence with the structure of the human ACE2 sequence.
- looking at a model of the spike protein binding with human ACE2 to understand if the position may be a binding site.

[6] `https://pymol.org/`

2.5 Model Comparison

The classification models were evaluated using testing sets of different lengths: the baseline model was tested against seven sequences (80/20 train/test split), the eliminative model was tested against 26 sequences (the subset of sequences that were not used in the original study), and the structural model was tested using the full set of 41 sequences.

The testing sets were used to produce confusion metrics for each model, from which performance sensitivity, specificity, and accuracy were derived. These performance metrics were used to compare the models.

Two types of pairwise statistical hypothesis tests were performed to determine if there was statistical difference between models:

- Two-sample tests for equality of proportions were used for sensitivity and specificity.
- McNemar's tests were used for error rate.

3 Results

3.1 Preliminary Results

The feature selection algorithm for the baseline model found six sites that best discriminate the susceptibility of the training sequences (Table 3.1). Each position has specific set of amino acids that increase likelihood of a sequence being immune.

Pos	Acids that impede susceptibility
83	Phenylalanine (F)
211	Aspartate (D)
212	Alanine (A), Aspartate (D), Serine (S)
246	Arginine (R)
251	Valine (V)
687	Not Alanine (A) or Serine (S)

Table 3.1: Summary of feature selection algorithm results

The eliminative algorithm based on the original study identified eight mutations that only appeared in immune sequences, shown in Table 3.2.

Pos	Mutation		
31	Lysine (K)	→	Aspartate (D)
41	Tyrosine (Y)	→	Alanine (A)
66	Glycine (G) Arginine (R)	⇒	Alanine (A)
83	Tyrosine (Y)	→	Phenylalanine (F)
113	Serine (S) Arginine (R)	⇒	Asparagine (N)
353	Lysine (K)	→	Histidine (H)
426	Proline (P)	→	Serine (S)
679	Isoleucine (I)	→	Valine (V)

Table 3.2: Mutations that may abolish susceptibility (eliminative)

3.2 Structural Analysis

By complementing the findings of the preliminary findings with a structural analysis, the feature set used for the structural model was reduced to five sites.

As an sample of the findings of the investigation, Figure 3.1 depicts the change caused by the mutation at position 353 from lysine (K) to histidine (H). The human ortholog is coloured pink, and the rat ortholog containing the

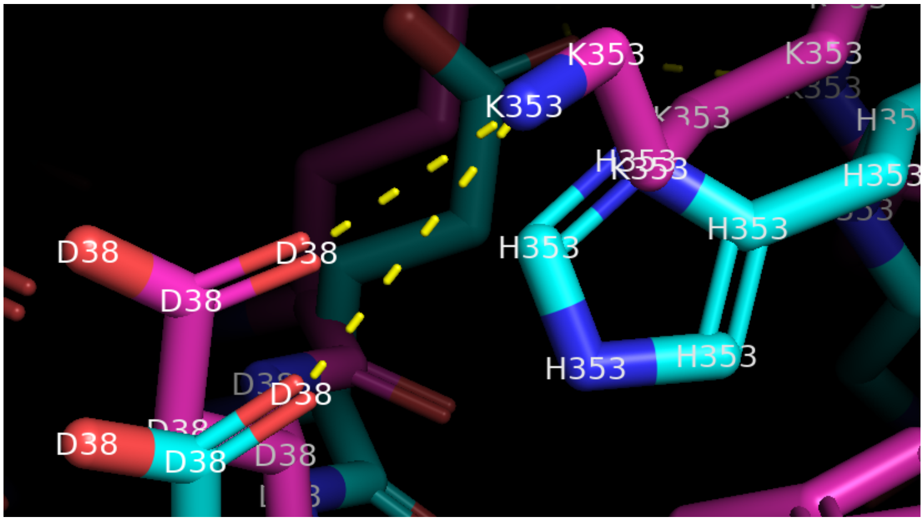

Figure 3.1: The structural effect of mutation H353

mutation is coloured cyan. This mutation abolishes a polar bond between the residue at position 353 and the aspartate (D) at position 38. Further, the lysine at position 353 in the human sequence displayed potential to bond with the virus spike protein, suggesting it may be a binding site. Since the mutation had an affect on the structure of the protein and there is evidence it may be a binding site, the mutation was included in the structural model.

Pos	**Mutation**		
31	Lysine (K)	→	Aspartate (D)
83	Tyrosine (Y)	→	Phenylalanine (F)
113	Serine (S) Arginine (R)	⇒	Asparagine (N)
353	Lysine (K)	→	Histidine (H)
426	Proline (P)	→	Serine (S)

Table 3.3: Mutations that may abolish susceptibility (structural)

The list of mutations used for the structural model is depicted in Table 3.3, and a more detailed overview of the structural analysis involved is included in Appendix B.

3.3 Model Design

With the set of mutations that abolish susceptibility identified for each model, three categorical binary decision trees were made for classification. Table 3.4

outlines the subset of positions selected for use in each model.

	Model			
Pos	**B**	**E**	**S**	**UniProt**
31		×	×	×
41		×		×
66		×		
83	×	×	×	×
113		×	×	
211	×			
212	×			
246	×			
251	×			
353		×	×	×
426		×	×	×
679		×		
687	×			

Table 3.4: Positions of focus for each model

Each marked cell indicates the position was used as a feature in the model of the column, where B, E, and S represent the baseline, eliminative, and structural models, respectively. The UniProt column is marked for each position where the UniProt database's mutagenesis overview identified the position as influential to susceptibility [20]. The UniProt database consolidates the findings of many studies to summarize how mutations affect susceptibility; this column represents a standard of comparison for the feature set of each model.

It is clear from this table that the structural model was most true to the findings outlined on UniProt, while the baseline model was not similar at all.

3.4 Model Performance

Each model was evaluated using its corresponding test set to produce confusion matrices (Table 3.5). The rows of the matrices represent the true class of the sequences, and the columns represent the model's predicted class.

Table 3.5: Classification confusion matrices for each model

	+	-
+	5	1
-	1	1

(a) Baseline Model

	+	-
+	14	3
-	3	6

(b) Eliminative Model

	+	-
+	24	3
-	5	9

(c) Structural Model

The sensitivity (true positive rate), specificity (true negative rate), and accuracy for each model were computed based on the confusion matrices (Table 3.6). The structural model had the highest accuracy and sensitivity, meaning it was more successful than the other models at correctly identifying sequences that are susceptible, and more successful at classifying sequences overall [21]. The eliminative model had a marginally higher specificity.

Model	**Sensitivity**	**Specificity**	**Accuracy**
Baseline	0.83	0.50	0.75
Eliminative	0.82	0.67	0.77
Structural	0.89	0.64	0.80

Table 3.6: Accuracy metrics for each model

Pairwise two-sample tests for equality of proportions used to compare sensitivities and specificities revealed no statistically significant difference between the performances of the models. Further, McNemar's tests used to compare error rates found no statistical difference between the misclassification rates of models [21–23].

Formulae and equations used for computing performance metrics are included in Appendix C, and a summary of statistical tests is included in Appendix D.

4 Discussion

Revisiting the positions selected as features for each model (Table 3.4), the structural model was almost completely in agreement with the UniProt findings — the exceptions were positions 41 and 113. Position 41 was excluded from the model arbitrarily: among the data used for this study, no sequences featured mutations at position 41. Position 113 was included in the eliminative model, and kept in the structural model because of the change in structure caused by the mutation to asaparagine (N). This disconnect between the positions in the model and the UniProt findings could mean this is a novel mutation that affects susceptibility, or it could be an extraneous finding attributed to an imperfect structural analysis.

The structural model's positions of focus were most similar to the UniProt findings, but the eliminative model also shared a considerable overlap: five of eight positions in the feature set appeared on the UniProt overview, and four appeared in this study's related work [2, 9, 18]. This is surprising, especially given the limited size of training data provided for the eliminative algorithm, and suggests the algorithm has some validity.

A well-trained and -tested model could be used to predict susceptibility in hosts where it is not already known. Table 4.1 shows the classifications of new sequences for each model.

Species	**B**	**E**	**S**
Chinchilla (*chinchilla lanigera*)	×		
Grizzly bear (*ursus arctos horribilis*)	×	×	×
Black flying fox (*pteropus alecto*)		×	×
Sumatran orangutan (*pongo abelii*)	×	×	×
Tasmanian devil (*sarcophilus harrisii*)			
Orca (*orcinus orca*)	×	×	×
Turkey (*meleagris gallopavo*)			
Leatherback sea turtle (*dermochelys coriacea*)			

Table 4.1: Predicted classification of additional sequences

The eliminative and structural models made the same predictions for each sequence. Focusing on where the models classified differently, it can be argued the eliminative and structural models' classifications were more intuitive based on existing knowledge:

- based on other rodents being immune (mice, rats), it would make sense for the chinchilla to be;
- based on other bats being susceptible, it would make sense for the black flying fox to be.

The structural model outperformed the other models in terms of classification accuracy, but the eliminative model was a close second. This further validates the algorithm used to identify influential mutations for features in the eliminative model.

Although the eliminative model did not involve any biological considerations, it performed similarly to the improved structural model which did; this suggests there is some rationality to the use of machine learning for this problem. The underachievement of the baseline model suggests the univariate feature selection strategy is not as suitable, at least given the training size.

4.1 Limitations

The greatest restriction of this study was the limited amount of data available surrounding species' susceptibilities to the virus; SARS-CoV-2 is ever-topical and more information about spike protein and ACE2 interaction is still surfacing. The consequence of this limitation is a high risk of overfitting, less accurate models, and less confidence in performance metrics.

Another limitation is the categorical and nonordinal nature of the features: the acids cannot be represented as numbers, nor can they be logically sorted. This severely restricts the variety of machine learning methods that can be used for the problem.

4.2 Future Direction

There are variations on the original problem that could be interesting for future work, specifically:

- A **multinomial classification** model that categorizes sequences into degrees of susceptibility (i.e., immune, moderately susceptible, very susceptible). The UniProt mutagenesis summary for ACE2 highlights that some mutations abolish interaction with the spike protein, while others inhibit or increase interaction to various extents [20]. In the structural model, the hamster sequences was misclassified as immune by its serine at position 426, which has been identified as a mutation that only slightly inhibits interaction with the spike protein [20, 24], not abolishing interaction as the model assumes.

- A **multi-label classification** model that classifies hosts as susceptible, symptomatic, zoonotically potential (transmittable to humans), and/or intraspecies transmittable. Among the set of animals known to be susceptible to SARS-CoV-2, some have been identified as asymptomatic, and some unlikely to transmit the virus to humans or other hosts of the same species [4, 25].

- A model with a **multidimensional feature space**. Each of the twenty amino acids can be categorized based on polarity, charge, and molecular size. Rather than focusing on the specific acid at each position, the model could classify the sequence based on these characteristics.

A more involved structural analysis could improve the performance of the structural model, specifically its specificity if more mutations that affected structure were identified. The analysis used for this model was based on the findings of the existing study used for the eliminative model, but extending beyond that set of mutations could reveal more potential influential positions. The analysis could be further strengthened by investigating the effect of mutations at consecutive positions; existing studies have reported consecutive mutations at sites 82-84 and 425-427 that each inhibit interaction with the spike protein [18, 20, 24].

5 Conclusion

When facing high-stakes decision — a reality of biology — having confidence in the tools that influence the decision is essential. When using black-box models, the logic is hidden and inherently less trustworthy; but even white-box models, where the logic is revealed, are not immediately trustworthy when not paired by an intuitive explanation.

The problem of using machine learning to classify sequences as susceptible or immune to SARS-CoV-2 is severely limited by the availability of data. Regardless, this analysis showed that it is possible to use existing data to classify sequences somewhat reliably: the eliminative model achieved an accuracy of 77%, comparable to the maximum accuracy of 80% of the structural model.

The hypothesis of this study was validated: the structural model marginally outperformed its counterparts. Equally significantly, this study is an example of how machine learning can be used as a starting point for classification models, and that pairing it with biological reasoning can strengthen their performances.

Data and Code Availability

The repository for this project is on GitHub (`https://github.com/zakwht/sars-cov-2`), within which is contained:

- the Python code used for all machine learning aspects of the project;
- the R code used for statistical analysis;
- some of the PyMOL script used for the structural analysis; and
- the processed data used for the project.

The set of data used for this analysis is listed in Appendix A.

References

[1] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, "The proximal origin of SARS-CoV-2," eng, *Nature medicine*, vol. 26, no. 4, pp. 450–452, Apr. 2020, ISSN: 1546-170X. DOI: 10.1038/s41591-020-0820-9. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32284615.

[2] J. Luan, Y. Lu, X. Jin, and L. Zhang, "Spike protein recognition of mammalian ACE2 predicts the host range and an optimized ACE2 for SARS-CoV-2 infection," *Biochemical and biophysical research communications*, vol. 526, no. 1, pp. 165–169, May 2020, ISSN: 1090-2104. DOI: 10.1016/j.bbrc.2020.03.047. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32201080.

[3] C. Woolsey *et al.*, "Establishment of an African green monkey model for COVID-19," *bioRxiv*, 2020. DOI: 10.1101/2020.05.17.100289. [Online]. Available: https://www.biorxiv.org/content/early/2020/05/17/2020.05.17.100289.

[4] W. O. for Animal Health. "COVID-19 - Events in Animals." (Mar. 2022), [Online]. Available: https://www.oie.int/en/what-we-offer/emergency-and-resilience/covid-19/#ui-id-3.

[5] B. B. Oude Munnink *et al.*, "Transmission of SARS-CoV-2 on mink farms between humans and mink and back to humans," *Science (New York, N.Y.)*, vol. 371, no. 6525, pp. 172–177, Jan. 2021, ISSN: 1095-9203. DOI: 10.1126/science.abe5901. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33172935.

[6] K. Schlottau *et al.*, "SARS-CoV-2 in fruit bats, ferrets, pigs, and chickens: an experimental transmission study," eng, *The Lancet. Microbe*, vol. 1, no. 5, e218–e225, Sep. 2020, ISSN: 2666-5247. DOI: 10.1016/S2666-5247(20)30089-6. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32838346.

[7] C. C. Sreenivasan, M. Thomas, D. Wang, and F. Li, "Susceptibility of livestock and companion animals to COVID-19," *Journal of Medical Virology*, vol. 93, no. 3, pp. 1351–1360, Oct. 2020, ISSN: 1090-2104. DOI: 10.1002/jmv.26621.

[8] E. M. Leroy, M. Ar Gouilh, and J. Brugère-Picoux, "The risk of SARS-CoV-2 transmission to pets and other wild and domestic animals strongly mandates a one-health strategy to control the COVID-19 pandemic," *One health (Amsterdam, Netherlands)*, vol. 10, pp. 100 133–100 133, Dec. 2020, ISSN: 2352-7714. DOI: 10.1016/j.onehlt.2020.100133. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32363229.

[9] X. Zhao *et al.*, "Broad and Differential Animal Angiotensin-Converting Enzyme 2 Receptor Usage by SARS-CoV-2," *Journal of Virology*, vol. 94, no. 18, e00940–20, 2020. DOI: 10.1128/JVI.00940-20.

[10] J. Shi *et al.*, “Susceptibility of ferrets, cats, dogs, and other domesticated animals to sars-coronavirus 2,” eng, *Science (New York, N.Y.)*, vol. 368, no. 6494, pp. 1016–1020, May 2020, ISSN: 1095-9203. DOI: 10.1126/science.abb7015. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/32269068.

[11] M. S. Mottaqi, F. Mohammadipanah, and H. Sajedi, “Contribution of machine learning approaches in response to SARS-CoV-2 infection,” eng, *Informatics in medicine unlocked*, vol. 23, pp. 100 526–100 526, 2021, ISSN: 2352-9148. DOI: 10.1016/j.imu.2021.100526. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33869730.

[12] I. A. Kakadiaris, M. Vrigkas, A. A. Yen, T. Kuznetsova, M. Budoff, and M. Naghavi, “Machine Learning Outperforms ACC/AHA CVD Risk Calculator in MESA,” *Journal of the American Heart Association*, vol. 7, no. 22, e009476, 2018. DOI: 10.1161/JAHA.118.009476. eprint: https://www.ahajournals.org/doi/pdf/10.1161/JAHA.118.009476.

[13] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, no. 5, 2019, ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x.

[14] M. J. Iqbal, I. Faye, B. B. Samir, and A. Md Said, “Efficient Feature Selection and Classification of Protein Sequence Data in Bioinformatics,” *The Scientific World Journal*, vol. 2014, p. 173 869, Jun. 2014, ISSN: 2356-6140. DOI: 10.1155/2014/173869.

[15] R. Saidi, M. Maddouri, and E. Mephu Nguifo, “Protein sequences classification by means of feature extraction with substitution matrices,” *BMC Bioinformatics*, vol. 11, no. 1, p. 175, Apr. 2010, ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-175.

[16] P. Patel, “Binary classification of protein sequence using machine learning,” May 2016. [Online]. Available: http://ir.ahduni.edu.in/xmlui/handle/123456789/385.

[17] H. Abdi, “Discriminant Correspondence Analysis,” *Encyclopedia of Measurement and Statistic*, Jan. 2007. DOI: 10.1186/1471-2105-11-175. [Online]. Available: https://www.researchgate.net/publication/242390704_Discriminant_Correspondence_Analysis.

[18] R. Li and C. Qin, “Expression pattern and function of SARS-CoV-2 receptor ACE2,” Aug. 2021, ISSN: 2590-0536. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590053621000860.

[19] Z. White, “Identifying Mutations in ACE2 That Influence Susceptibility to SARS-CoV-2,” CSC482B: Computational Biology Algorithms, University of Victoria, Nov. 2021.

[20] UniProt, *ACE2 - Angiotensin-converting enzyme 2 precursor - Homo sapiens (Human) - ACE2 gene & protein.* [Online]. Available: https://www.uniprot.org/uniprot/Q9BYF1 (visited on 02/21/2022).

[21] J. Vu and D. Harrington, *Introductory Statistics for the Life and Biomedical Sciences*. OpenIntro Statistics, 2015.

[22] W. Patrick Walters, "Comparing classification models—a practical tutorial," *Journal of Computer-Aided Molecular Design*, Sep. 2021, ISSN: 1573-4951. DOI: `10.1007/s10822-021-00417-2`. [Online]. Available: `https://doi.org/10.1007/s10822-021-00417-2`.

[23] M. J. Crawley, *Statistics: An Introduction Using R*. Wiley, 2015.

[24] W. Li *et al.*, "Receptor and viral determinants of sars-coronavirus adaptation to human ace2," eng, *The EMBO journal*, vol. 24, no. 8, pp. 1634–1643, Apr. 2005, ISSN: 0261-4189. DOI: `10.1038/sj.emboj.7600640`. [Online]. Available: `https://pubmed.ncbi.nlm.nih.gov/15791205`.

[25] S. B. King and M. Singh, "Comparative genomic analysis reveals varying levels of mammalian adaptation to coronavirus infections," *PLOS Computational Biology*, vol. 17, no. 11, pp. 1–15, Nov. 2021. DOI: `10.1371/journal.pcbi.1009560`. [Online]. Available: `https://doi.org/10.1371/journal.pcbi.1009560`.

[26] M. Richard *et al.*, "SARS-CoV-2 is transmitted via contact and via the air between ferrets," eng, *Nature communications*, vol. 11, no. 1, pp. 3496–3496, Jul. 2020, ISSN: 2041-1723. DOI: `10.1038/s41467-020-17367-2`. [Online]. Available: `https://pubmed.ncbi.nlm.nih.gov/32641684`.

[27] M. V. Palmer *et al.*, "Susceptibility of white-tailed deer (Odocoileus virginianus) to SARS-CoV-2," *Journal of virology*, vol. 95, no. 11, e00083–21, Mar. 2021, ISSN: 1098-5514. DOI: `10.1128/JVI.00083-21`. [Online]. Available: `https://pubmed.ncbi.nlm.nih.gov/33692203`.

[28] J. F.-W. Chan *et al.*, "Simulation of the Clinical and Pathological Manifestations of Coronavirus Disease 2019 (COVID-19) in a Golden Syrian Hamster Model: Implications for Disease Pathogenesis and Transmissibility," *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, vol. 71, no. 9, pp. 2428–2446, Dec. 2020, ISSN: 1537-6591. DOI: `10.1093/cid/ciaa325`. [Online]. Available: `https://pubmed.ncbi.nlm.nih.gov/32215622`.

[29] S. Lu *et al.*, "Comparison of nonhuman primates identified the suitable model for COVID-19," eng, *Signal transduction and targeted therapy*, vol. 5, no. 1, pp. 157–157, Oct. 2020, ISSN: 2059-3635. DOI: `10.1038/s41392-020-00269-6`. [Online]. Available: `https://pubmed.ncbi.nlm.nih.gov/32814760`.

Appendix A: Sequence Lists

Species	**Accession**	**Susceptible**
Dromedary (*camelus dromedarius*)	XP_031301717.1	No [2]
Raccoon (*procyon lotor*)	BAE72462.1	No [2]
Greater horseshoe bat (*rhinolophus ferrumequinum*)	XP_032963186.1	No [2]
Brown rat (*rattus norvegicus*)	NP_001012006.1	No [2][9]
House mouse (*mus musculus*)	NP_001123985.1	No [2][9]
Platypus (*ornithorhynchus anatinus*)	XP_001515597.2	No [2]
African bush elephant (*loxodonta africana*)	XP_023410960.1	No [2]
European hedgehog (*erinaceus europaeus*)	XP_007538670.1	No [2]
Raccoon dog (*nyctereutes procyonoides*)	ABW16956.1	No [2]
Meerkat (*suricata suricatta*)	XP_029786256.1	No [2]
Kangaroo rat (*dipodomys ordii*)	XP_012887573.1	No [2]
Guinea pig (*cavia porcellus*)	XP_023417808.1	No [2]
Domestic pig (*sus domesticus*)	ASK12083.1	No [7][10]
Mallard (*anas platyrhynchos*)	XP_012949915.3	No [7][10]
Human (*homo sapiens*)	BAB40370.1	Yes
Rhesus macaque (*macaca mulatta*)	XP_028697658.1	Yes [2][9]
European rabbit (*oryctolagus cuniculus*)	QHX39726.1	Yes [9]
House cat (*felis catus*)	XP_044906242.1	Yes [2][4][8][9]
Domestic dog (*canis familiaris*)	NP_001158732.1	Yes [2][4][8][9]
Siberian tiger (*panthera tigris*)	XP_042830022.1	Yes [9]
Golden snub-nosed monkey (*rhinopithecus roxellana*)	XP_010364367.2	Yes [2]
Olive baboon (*papio anubis*)	XP_021788733.1	Yes [2]
Chimpanzee (*pan troglodytes*)	XP_016798468.1	Yes [2]
Orangutan (*pongo abelii*)	XP_024096013.1	Yes [2]
Ferret (*mustela putorius furo*)	NP_001297119.1	Yes [4][5][10][26]
Mink (*mustela lutreola biedermanni*)	QNC68911.1	Yes [4][5]
White-tailed deer (*odocoileus virginianus texanus*)	XP_020768965.1	Yes [4][27]
Golden hamster (*mesocricetus auratus*)	XP_005074266.1	Yes [4][28]
Canadian lynx (*lynx canadensis*)	XP_030160839	Yes [4]
North american river otter (*lontra canadensis*)	XP_032736029.1	Yes [4]
Cougar (*puma concolor*)	XP_025790417.1	Yes [4]
Western lowland gorilla (*gorilla gorilla gorilla*)	XP_018874749.1	Yes [4]
Spotted hyena (*crocuta crocuta*)	KAF0878287.1	Yes [4]
Amur leopard (*panthera pardus orientalis*)	XP_019273509.1	Yes [4]
Pangolin (*manis pentadactyla*)	QLH93383.1	Yes [2][7]
Big-eared horseshoe bat (*rhinolophus macrotis*)	ADN93471.1	Yes [2][7]
Leschenault's rousette (*rousettus leschenaultii*)	BAF50705.1	Yes [2][7]
Common marmoset (*callithrix jacchus*)	XP_008987241.1	Yes [5][29]
Cynomolgus macacque (*macaca fascicularis*)	XP_005593094.1	Yes [5][29]
Greater short-nosed fruit bat (*cynopterus sphinx*)	QKE49997.1	Yes [5][6]
Green monkey (*chlorocebus sabaeus*)	XP_037842285.1	Yes [3][5]

Table A.1: Species with confirmed accuracies

Table A.1 lists the full set of sequences used for testing and training the models, and Table A.2 lists sequences used for susceptibility prediction, each with accession IDs from the NCBI database. ABC!

Species	**Accession**
Chinchilla (*chinchilla lanigera*)	XP_013362429.1
Grizzly bear (*ursus arctos horribilis*)	XP_026333865.1
Black flying fox (*pteropus alecto*)	XP_006911709.1
Sumatran orangutan (*pongo abelii*)	XP_024096013.1
Tasmanian devil (*sarcophilus harrisii*)	XP_031814825.1
Orca (*orcinus orca*)	XP_033283817.1
Turkey (*meleagris gallopavo*)	XP_019467554.1
Leatherback sea turtle (*dermochelys coriacea*)	XP_043360132.1

Table A.2: Orthologs used for susceptibility prediction

Table A.3 lists the protein structure models used for the structural analysis, from the AlphaFold Protein Structure Database. Structure 6VW1 was also used to model the binding between the virus spike protein and ACE2, retrieved from the RCSB Protein Data Bank.

Species	**Accession**
Human (*homo sapiens*)	Q9BYF1
House mouse (*mus musculus*)	Q8R0I0
Brown rat (*rattus norvegicus*)	Q5EGZ1

Table A.3: Orthologs used for structure analysis

Appendix B: Structural Analysis

Table B.1 summarizes the findings of the structural analysis, noting whether or not each position appeared as a potential binding site, and whether or not the mutation affected the protein's structure.

Pos	Results of analysis
31	Identified as a potential binding site of the spike protein.
41	No sequences in this study include a mutation at this position: rejected from the model (no effect).
66	Does not appear to be a binding site, mutation has no effect on the structure: rejected from model.
83	Identified as a potential binding site of the spike protein, mutation has no immediate effect on structure.
113	Affects bonds with serine at position 105.
353	Identified as a potential binding site of the spike protein, mutation affects bond with aspartate at position 38.
426	Identified as a potential binding site of the spike protein, mutation affects bonds with acids at positions 425 and 427.
679	Does not appear to be a binding site, mutation has no effect on the structure: rejected from model.

Table B.1: Summary of structural analysis

Appendix C: Calculations

Below are the formulas for sensitivity (true positive rate, TPR), specificity (true negative rate, TNR), and classification accuracy (ACC).

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{1}$$

$$TNR = \frac{TN}{N} = \frac{TN}{TN + FP} \tag{2}$$

$$ACC = \frac{TP + TN}{P + N} = \frac{TP + TN}{TP + FN + TN + FP} \tag{3}$$

Where:

- P is the total number of pos. sequences
- N is the total number of neg. sequences
- TP is the number of pos. sequences classified as pos. (true positive)
- TN is the number of neg. sequences classified as neg. (true negative)
- FP is the number of neg. sequences classified as pos. (false positive)
- FN is the number of pos. sequences classified as neg. (false negative)

Baseline model calculations

$$TPR = \frac{5}{6} = 0.8333$$

$$TNR = \frac{1}{2} = 0.5000$$

$$ACC = \frac{5+1}{6+2} = \frac{6}{8} = 0.7500$$

Eliminative model calculations

$$TPR = \frac{14}{17} = 0.8235$$

$$TNR = \frac{6}{9} = 0.6667$$

$$ACC = \frac{14+6}{17+9} = \frac{20}{26} = 0.7692$$

Structural model calculations

$$TPR = \frac{24}{27} = 0.8889$$

$$TNR = \frac{9}{14} = 0.6429$$

$$ACC = \frac{24+9}{27+14} = \frac{33}{41} = 0.8049$$

Appendix D: Statistical Analysis

Table D.1 summarizes the results of the nine statistical tests performed, including the chi-square test statistic (χ^2), degrees of freedom (DF) and the p-value. No null hypotheses were rejected.

McNemar's Test				
H_0: models have equal error rates				
Models		χ^2	DF	p-value
B	E	0	1	1
B	S	0.125	1	0.7237
E	S	0	1	1
Proportion Test: Sensitivity				
H_0: models have equal sensitivities				
Models		χ^2	DF	p-value
B	E	0	1	1
B	S	0	1	1
E	S	0.0269	1	0.8697
Proportion Test: Specificity				
H_0: models have equal specificities				
Models		χ^2	DF	p-value
B	E	0	1	1
B	S	0	1	1
E	S	0	1	1

Table D.1: Summary of performed hypothesis tests

The following listings are a selection of the R output of the tests, to convey more information including contingency tables for the McNemar tests and confidence intervals for the proportion tests.

```
> mcnemar.print("baseline", "eliminative", c(23,4,5,9))
        eliminative
baseline  + -
        + 23 5
        -  4 9

  McNemar's Chi-squared test with continuity correction
McNemar's chi-squared = 0, df = 1, p-value = 1

> mcnemar.print("baseline", "structural", c(24,3,5,9))
        structural
baseline  + -
        + 24 5
        -  3 9
  McNemar’s Chi-squared test with continuity correction
McNemar’s chi-squared = 0.125, df = 1, p-value = 0.7237

> mcnemar.print("eliminative", "structural", c(28,0,1,12))
            structural
eliminative  +  -
          + 28  1
          -  0 12

  McNemar’s Chi-squared test with continuity correction
McNemar’s chi-squared = 0, df = 1, p-value = 1
```

Listing D.1: R output of McNemar tests

```
> sens.test(baseline, eliminative)
baseline sensitivity: 0.833333333333333
eliminative sensitivity: 0.823529411764706
    2-sample test for equality of proportions
X-squared = 2.2428e-31, df = 1, p-value = 1
95 percent confidence interval: -0.3489446  0.3685524

> sens.test(baseline, structural)
baseline sensitivity: 0.833333333333333
structural sensitivity: 0.888888888888889
    2-sample test for equality of proportions
X-squared = 3.7369e-31, df = 1, p-value = 1
95 percent confidence interval: -0.4320077  0.3208966

> sens.test(eliminative, structural)
eliminative sensitivity: 0.823529411764706
structural sensitivity: 0.888888888888889
    2-sample test for equality of proportions
X-squared = 0.026908, df = 1, p-value = 0.8697
95 percent confidence interval: -0.3298345  0.1991156

> spec.test(baseline, eliminative)
baseline specificity: 0.5
eliminative specificity: 0.666666666666667
    2-sample test for equality of proportions
X-squared = 3.504e-32, df = 1, p-value = 1
95 percent confidence interval: -1.0000000  0.7583094
```

```
> spec.test(baseline, structural)
baseline specificity: 0.5
structural specificity: 0.642857142857143
    2-sample test for equality of proportions
X-squared = 3.0052e-32, df = 1, p-value = 1
95 percent confidence interval: -1.0000000  0.7370075

> spec.test(eliminative, structural)
eliminative specificity: 0.666666666666667
structural specificity: 0.642857142857143
    2-sample test for equality of proportions
X-squared = 1.5498e-32, df = 1, p-value = 1
95 percent confidence interval: -0.3973015  0.4449206
```

Listing D.2: R output of proportion tests