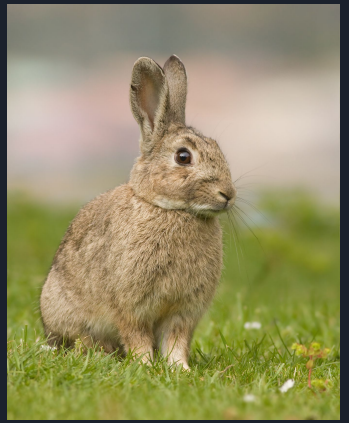
The Role of Machine Learning in SARS-CoV-2 Susceptibility Classification

Zak White, 2022

SARS-CoV-2

COVID-19 is caused by the virus SARS-CoV-2

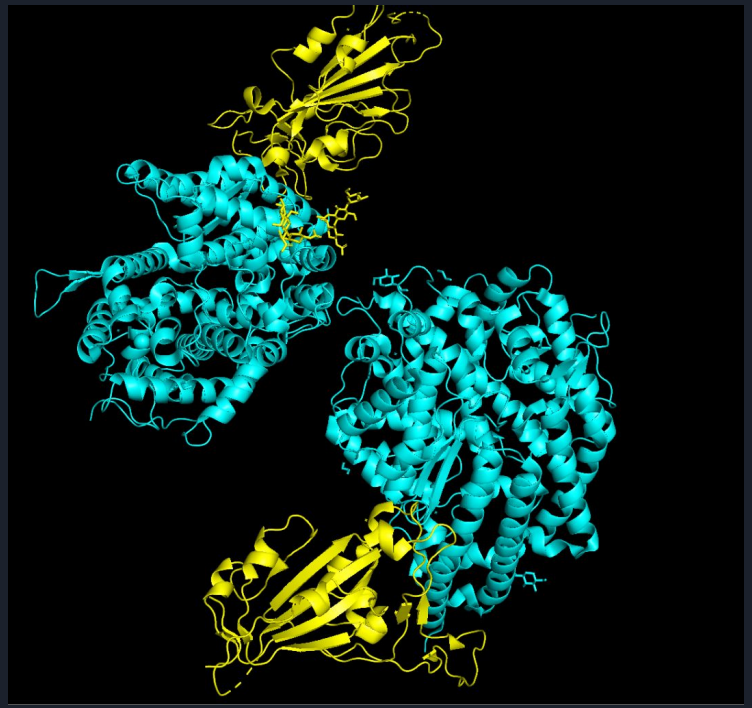
- Likely of zoonotic origin
- Some non-human animals are known to be susceptible; some are not



ACE2

The SARS-CoV-2 spike protein's target in potential hosts is the ACE2 protein (angiotensin-converting enzyme 2)

- Differences in ACE2 sequences can result in different interaction with the spike protein, or possibly inhibit interaction altogether



Sequence Classification

Proteins are biomolecules made up of amino acids

Sequence of acids can be mapped to a feature vector for use in classification problems

- Features are categorical and nonordinal
- This problem is the basis of this study

```
MSSSSWLLLSLVAVTAAQSTIEEQAKTFLDKFNHEAEDLFYQSSLASWNYNTNITEENVQNMNNAGDKWS
AFLKEQSTLAQMYPLQEIQNLTVKLQLQALQQNGSSVLSEDKSKRLNTILNTMSTIYSTGKVCNPDNPQE
CLLLEPGLNEIMANSLDYNERLWAWESWRSEVGKQLRPLYEEYVVLKNEMARANHYEDYGDYWRGDYEVN
GVDGYDYSRGQLIEDVEHTFEEIKPLYEHLHAYVRAKLMNAYPSYISPIGCLPAHLLGDMWGRFWTNLYS
LTVPFGQKPNIDVTDAMVDQAWDAQRIFKEAEKFFVSVGLPNMTQGFWENSMLTDPGNVQKAVCHPTAWD
LGKGDFRILMCTKVTMDDFLTAHHEMGHIQYDMAYAAQPFLLRNGANEGFHEAVGEIMSLSAATPKHLKS
IGLLSPDFQEDNETEINFLLKQALTIVGTLPFTYMLEKWRWMVFKGEIPKDQWMKKWWEMKREIVGVVEP
VPHDETYCDPASLFHVSNDYSFIRYYTRTLYQFQFQEALCQAAKHEGPLHKCDISNSTEAGQKLFNMLRL
GKSEPWTLALENVVGAKNMNVRPLLNYFEPLFTWLKDQNKNSFVGWSTDWSPYADQSIKVRISLKSALGD
RAYEWNDNEMYLFRSSVAYAMRQYFLKVKNQMILFGEEDVRVANLKPRISFNFFVTAPKNVSDIIPRTEV
EKAIRMSRSRINDAFRLNDNSLEFLGIQPTLGPPNQPPVSIWLIVFGVVMGVIVVGIVILIFTGIRDRKK
KNKARSGENPYASIDISKGENNPGFQNTDDVQTSF
```

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

Objective

Objective:

- Compare the performance of ML-based classification models of susceptibility with the performance of a biology-driven AI model

Hypothesis:

- Expect the biology model to outperform the ML models

Methods

41 ACE2 sequences were collected from the NCBI protein database, each identified as susceptible or immune (14 immune and 27 susceptible)

Three binary classification models were designed:

- A baseline model created using a feature selection algorithm which scored the acids at each index as a feature
- An eliminative model designed using an iterative feature selection algorithm that isolated mutations that appeared exclusively in immune sequences
- A structural model that was supported by the findings of a manual investigation of the ACE2 structure and its interaction with the SARS-CoV-2 spike protein

Results

The adjacent table highlights which residues were used in each model, and compares them with the UniProt database's ACE2 mutagenesis overview.

- It is clear the structural model was more true to the findings outlined on UniProt, and the baseline model was least similar

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

Results (2)

Classification confusion matrices were made for each model based on testing subsets. The results were used to compute accuracy, sensitivity (true positive rate), and specificity (true negative rate)

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

The structural model had the highest classification accuracy and sensitivity, and the eliminative model had the highest specificity.