TOWARDS INTERPRETING ZOONOTIC POTENTIAL OF BETACORONAVIRUS SEQUENCES WITH ATTENTION

Kahini Wadhawan, Payel Das, Barbara A. Han, Ilya R. Fischhoff, Adrian C. Castellanos, Arvind Varsani, Kush R. Varshney; kawadhaw@in.ibm.com and daspa@us.ibm.com

Current methods for viral discovery target evolutionarily conserved proteins that accurately identify virus families but remain unable to distinguish the zoonotic potential of newly discovered viruses. Here, we apply an attention-enhanced long-short-term memory (LSTM) deep neural net classifier to a highly conserved viral protein target to predict zoonotic potential of betacoronavirus sequences. The classifier performs with a 94% accuracy. Analysis and visualization of attention at the sequence and structure-level features indicate possible association between important protein-protein interactions governing viral replication in zoonotic betacoronaviruses and zoonotic transmission.

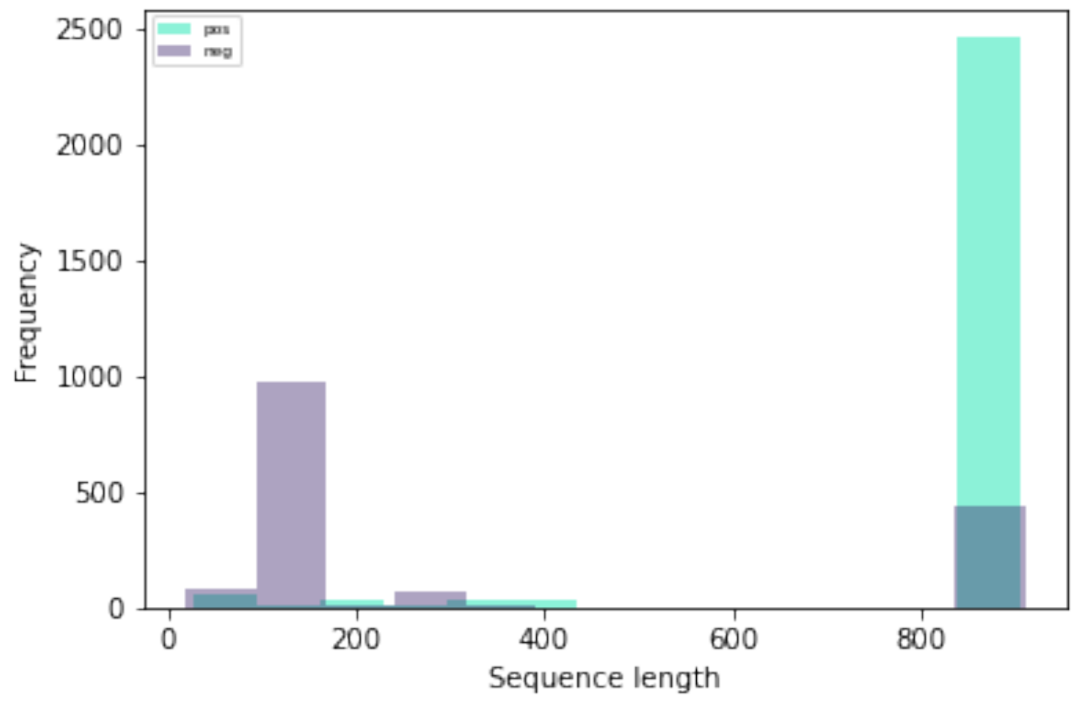
Introduction

- Majority of viruses emerging in humans arise from animal hosts. SARS-CoV-2 causing COVID-19 belongs to betacoronavirus (β-CoV) family.
- Method of surveilling involves identifying a conserved region of the genome. For instance, RdRp (RNA-dependent RNA polymerase) gene is a highly conserved sequence that is commonly used to ascertain β-CoVs presence from sampled wildlife species. It does not enable predictions about whether the virus poses a zoonotic threat to humans.
- The disconnect between broad viral surveillance and accurate predictions about zoonotic potential of newly discovered viruses remains a major research frontier, which has been highlighted by the emergence of SARS-CoV-2.

Data Collection

- We assembled a dataset of published RdRp gene sequences for all β-CoVs, freely available in Gen-Bank and added binary label to designate zoonotic status. The potential zoonotic sequences included ones similar or identical to sequences that caused MERS coronavirus (MERS-CoV) SARS.
- The resulting dataset contained a total of 4259 sequences of length up to 910, with a vocabulary of 21 amino acids.

Dataset Imbalance

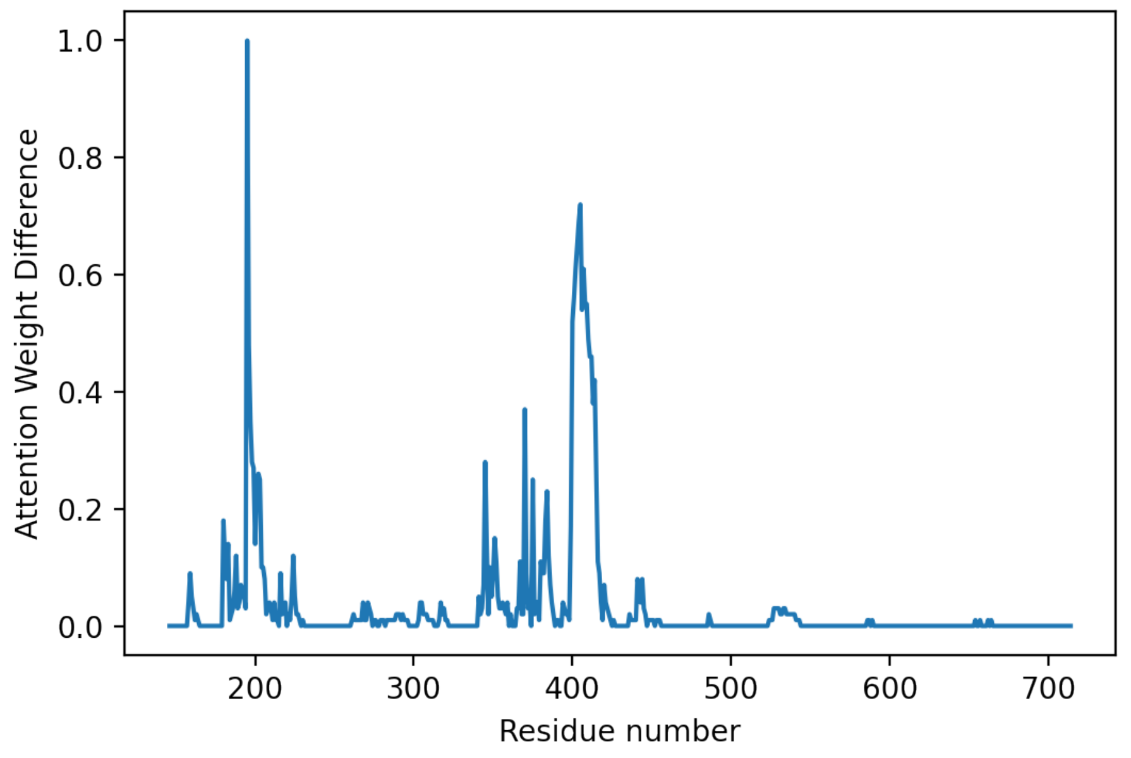


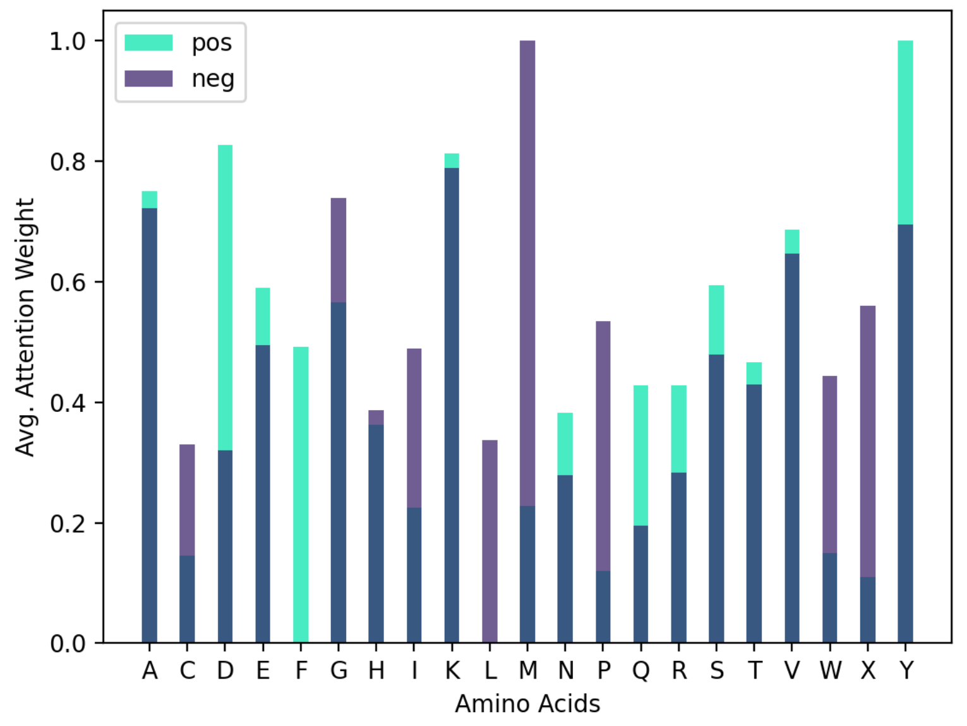
Dataset imbalance shown as a function of sequence length

Result and Analysis of Attention-enhanced LSTM Model for Binary Classification

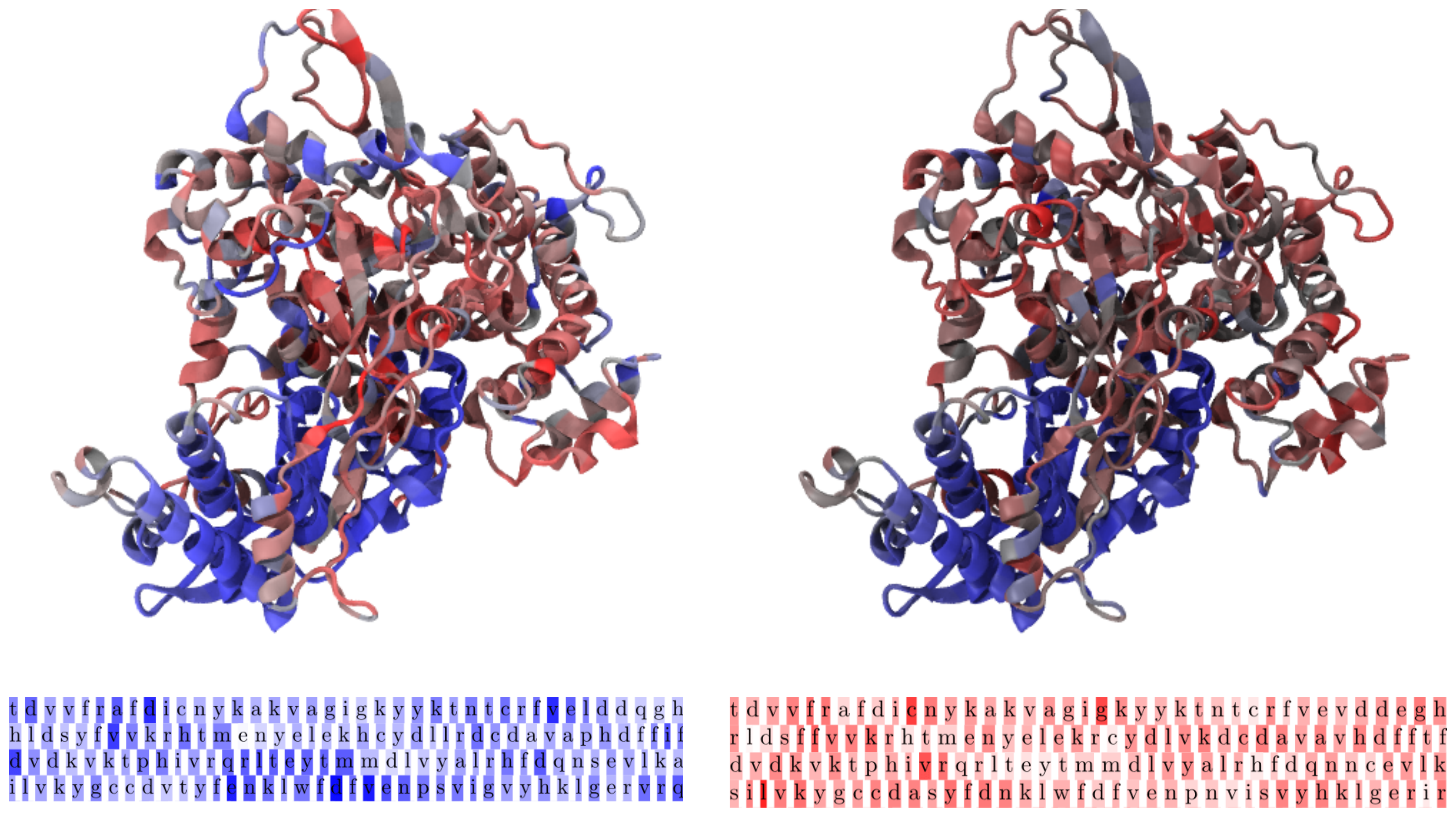
Sequence length ranges	Accuracy	Precision	Recall	F1-score
len< 150 (***imbalanced data***)	0.925	**NaN(0/0)**	**0.0**	**0.0**
150 <len< 500	0.949	0.955	0.955	0.955
len> 500	0.938	0.978	0.947	0.962
all length	0.935	0.977	0.916	**0.946**
len< 150 (***balanced data***)	0.880	**0.971**	**0.500**	**0.660**
150 <len< 500	0.993	0.992	1.00	0.996
len> 500	0.971	0.969	0.998	0.983
all length	0.948	0.973	0.949	**0.961**

Comparison of model performance on balanced and imbalanced dataset. F1-scores reported in different length regimes.



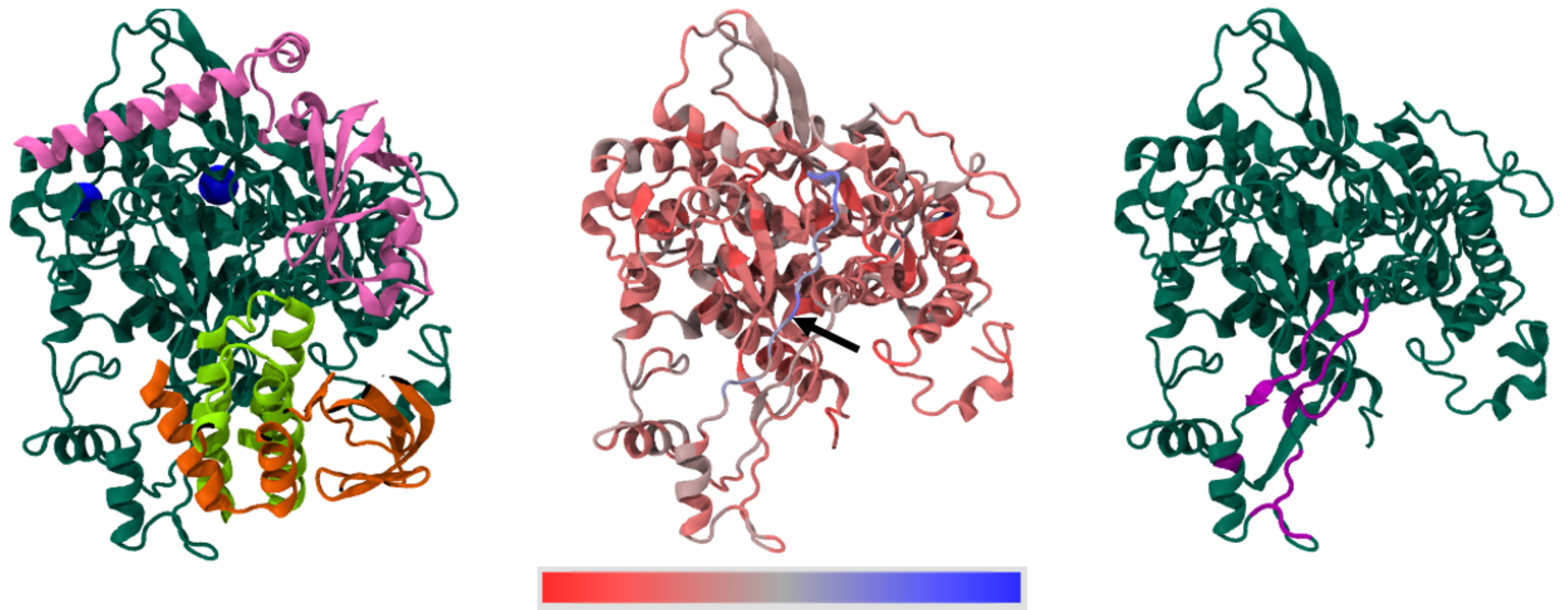


Left: Average attention difference vs residue number. Right: Feature importance plot for both positive and negative classes overlapped. (average attention over amino acids)


tdvvfrafdicnykakvagigkyyktntcrfvelddqgh
hldsyfvvkrhtmenyelekhcydllrdcdavaphdffif
dvdkvktphivrqrlteytmmdlvyalrhfdqnsevlka
ilvkygccdvtyfenklwfdfvenpsvigvyhklgervrq
tdvvfrafdicnykakvagigkyykintcrfvevddegh
rldsffvvkrhtmenyelekrcydlvkdcdavavhdfftf
dvdkvktphivrqrlteytmmdlvyalrhfdqnncevlk
silvkygccdasyfdnklwfdfvenpnvisvyhklgerir

Left: attention heatmap of single positive sample at 3D and 1D level for sequence length>900 mapped on to their modelled structure obtained using homology modelling (6NUR.pdb chain A used as template). Colorscale used: red-gray-blue (low to high). Right: attention heatmap of single negative sample. 1D maps are cropped to smaller length here.

Average attention maps



Left: Structure of SARS-CoV nsp12 RdRp (dark green) bound to nsp7 (light green) and nsp8 (pink) co-factors (pdb id:6NUR). Metal ions shown as blue spheres. Middle: Average attention differences between two classes, mapped to the SARS-CoV nsp12 structure (6NUR.pdbchain A residue 146 to 714). Red-gray-blue (low to high) color scale is used to visualize the attention heatmap. Right: Regions of RdRp interacting (cutoff: 6.5˚A) with nsp7 highlighted in magenta.

Key Results

- We present an attention-enhanced LSTM classifier that predicts zoonotic transmission in β-CoVs using RdRp sequences with 93-94% accuracy.
- We investigated attention maps at both the sequence and the structure level to get interpretability of the “black-box” model.
- Our results indicate mapping of class-level attention differences at protein-protein interaction sites of the RdRp structure.
- These sites are of significant functional importance for binding with cofactors crucial for polymerase activity involved in viral replication. Though RdRp itself is not directly involved in the infection process, physiochemical interactions with neighbouring cofactors appear to influence the function of zoonotic β-CoVs compared to non-zoonotic counterparts.

Future Work

- Explore multi-head attention mechanism for better interpretability and prediction.
- Extend the current study beyond β-CoVs to investigate the robustness in capturing of biological context via model features.