Towards Interpreting Zoonotic Potential of Betacoronavirus Sequences With Attention

Kahini Wadhawan[1,*], **Payel Das**[2,*], **Barbara A. Han**[3], **Ilya R. Fischhoff**[3], **Adrian C. Castellanos**[3], **Arvind Varsani**[4], and **Kush R. Varshney**[2]

[1]IBM Research, New Delhi, India
[2]IBM Research, Yorktown Heights, NY, USA
[3]Cary Institute of Ecosystem Studies, NY, USA
[4]The Biodesign Institute, Arizona State University, USA
[*]kawadhaw@in.ibm.com and daspa@us.ibm.com

Abstract

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 across betacoronaviruses. 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.

1 Introduction

The majority of viruses emerging in humans arise from animal hosts. For example, SARS-CoV-2, the virus causing the COVID-19 pandemic, belongs to the betacoronavirus(β-CoVs) family from which numerous other viruses have been described from multiple mammal species around the world. The typical method for surveiling for such viruses is to identify a conserved region of the genome. For instance, the RdRp (RNA-dependent RNA polymerase) gene is a highly conserved sequence that is commonly used to ascertain β-CoVs presence from sampled wildlife species. While this conserved nature makes RdRp sequences a reliable target for accurately identifying novel β-CoVs from field surveillance, it does not enable predictions about whether the virus poses a zoonotic threat to humans. This is because the RdRp sequences encode the polymerase responsible for replicating the RNA genome once inside the cell, which is not involved in host cell entry or the infection process. 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. Numerous closely-related β-CoVs have been identified from field surveillance over the past several years, but no clear and cost-effective way to triage the zoonotic risk posed by these newly discovered viruses. Beyond quantifying the zoonotic risk posed by a novel virus, preventing their emergence also relies on differentiating the key elements that enable some viruses to successfully infect humans. In other words, explaining what viral structures enable high zoonotic potential will be critical for informing what downstream actions will be effective for preventing spillover and viral emergence.

Recent studies have shown stronger performance of machine learning based approaches for zoonotic transmission prediction task, when compared to homology based approaches, both at the protein sequence and gene sequence level (Eng et al., 2017; Xu et al., 2017; Li & Sun, 2018; Zhang et al., 2019; Mock et al., 2019; Wang et al., 2013; Bartoszewicz et al., 2021). There is a consensus that deep models outperform shallow classifiers. Previous works have often used either feature vectors reflecting compositional and physicochemical properties of protein sequences or on word vectors. In a few studies, input features were augmented with structural details, e.g. (Fischhoff et al., 2021). In this study, we develop a deep neural net classifier for predicting zoonotic potential of RdRp se-

quences. For this purpose, we employ an attention mechanism enhanced long short-term memory (LSTM) model (Hochreiter & Schmidhuber, 1997) trained on RdRp sequences, as attention Bahdanau et al. (2014) is known to better cope with longer length sequences than a simple LSTM model, as is the case here. The final model yields a 94% prediction accuracy.

For wide-spread adoption of deep learning models for high-stake tasks, such as zoonotic transmission prediction, it is important that such models are interpretable and explainable. Towards this direction, we intend to interpret the attention signals of the intermediate layers of deep classifier by visualizing the attention on the amino acid characters of the sample sequences. We also investigate average attention difference between two classes as a function of the amino acid type. Finally, we visualize the learned attentions on the 3D RdRp structure, as the 3D structure determines its function. Analysis of these internal workings of the classifier model provides important insights into likely mechanism of zoonotic transmission, as class-level attention differences map to the key protein-protein interaction sites. To our knowledge, this is the first machine learning work that provides sequence and structure-level understanding of zoonotic transmission potential of protein sequences.

2 DATA: RdRp SEQUENCE DATA AND LABEL ANNOTATION

We assembled a dataset of published RdRp gene sequences for all β-CoVs, freely available in Gen-Bank (Benson et al., 2013). For each published sequence, we added a binary label to designate zoonotic status based on the results of published research. The potential zoonotic sequences included ones similar or identical to sequences that caused MERS coronavirus (MERS-CoV) SARS. In the NCBI records, "SARS" was often used to describe these viruses; here we refer to SARS-CoV-1 to distinguish from SARS-CoV-2 causing the COVID-19 pandemic. MERS-CoV or SARS-CoV-1 sequences found in humans were categorized as zoonotic. If no host species was identified for a sequence, we did not categorize zoonotic status for that sequence. Prior to modeling, all sequences for which we were unable to categorize zoonotic status were removed from the dataset. For sequences passaged or adapted through cell culture of a different species than the original host, we made no determination about zoonotic status, as passaging may have changed zoonotic capacity. These included viruses described as involving Vero cell culture, from African green monkey (*Cercopithecus aethiops*) cells (Ammerman et al., 2008; Tsui et al., 2003); mouse-adapted (Lu et al., 2009); recombinant; or LLC-MK2, cultured in rhesus monkey (*Macaca mulatta*) cells (e.g., Kaye et al. (2006)). Information about these forms of cell culture was often found in the host field in NCBI; or as part of "passage details" in the notes field. If a sequence originated in one person and was cultured in cells from a different human cell line prior to sequencing, we considered it to remain zoonotic as we did not expect differences among people to change the zoonotic status.

We also considered sequences found in animals to be zoonotic if the authors of the sequence considered it to be functionally identical to sequences found in people. For MERS-CoV, we emphasized virus taxonomic clade position and locality, which is related to clade position, as for MERS-CoV this information is more readily available than information about sequence similarity to human MERS-CoV. MERS-CoV in clades A and B include known human sequences and are known to cross from camels (*Camelus dromedarius*) to humans (Lau et al., 2016; Chan et al., 2014). For ambiguous MERS-CoV sequences, we categorized as zoonotic if there were human cases found in the same geographic area. MERS-CoV in Africa, in clade C, are similar but with no known human cases Chu et al. (2018). Clade C has lower titer and virulence in lab experiments, compared to clade A or B . We assumed that clade C MERS-CoV to be non-zoonotic. For SARS-CoV-1 viruses found in animals, if insufficient information was available to confirm that the virus was zoonotic then it was labeled non-zoonotic. For example, SARS-CoV-1 viruses found in bats were categorized as non-zoonotic if no study was published associated with the sequence. SARS-CoV-1 reported in palm civet (*Paguma larvata*) were considered zoonotic if the authors considered it to be identical to sequences found in people Wang et al. (2005), and otherwise categorized as non-zoonotic (Yi, 2005). We considered as zoonotic a sequence isolated from an animal that also replicated in human cell lines, *e.g.* a SARS-CoV-1 from *Rhinolophus sinicus* bats that replicated in human cell culture (Ge et al., 2013). If the sequence was found in an animal species not known to be a bridge species to people, *e.g.* a ferret, we categorized it as non-zoonotic (Town et al., 2012).
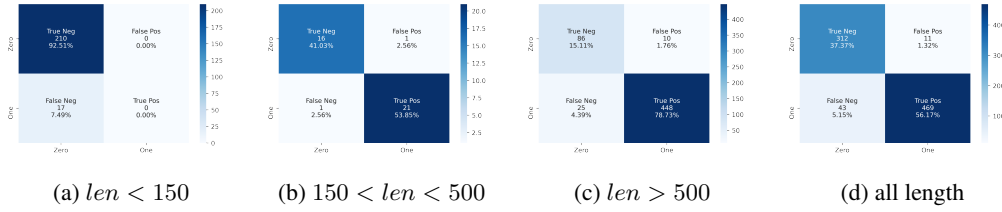
| (a) $len < 150$ | (b) $150 < len < 500$ | (c) $len > 500$ | (d) all length |

Figure 1: Model performance confusion matrix computed over range of sequence lengths.

(a) Length distribution   (b) Avg. attention difference vs residue number   (c) Feature importance (avg. attention over amino acids)
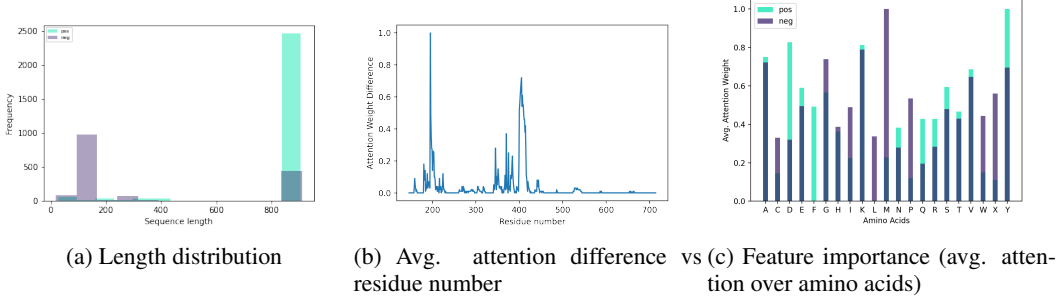
Figure 2: Data statistics and attention at 3D and 1D level plots.

3 METHODS

The resulting dataset contained a total of 4259 sequences of length up to 910 characters, with vocabulary of 21 amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, X, Y). We use 60/20/20 ratio to create train/val/test split, resulting in 2548, 876 and 835 sequences respectively, maintaining the positive to negative ratio in each split. A LSTM model enhanced with a single head attention mechanism was trained on this RdRp sequence dataset to predict their zoonotic potential (as a binary classification task). The model architecture was comprised of a single encoded layer with single head attention Bahdanau et al. (2014). The model accepts in the sequence of amino acids $x = (x_1, x_2, ..., x_L)$ and learns an embedding $z = (z_1, z_2, ..., z_L)$, which is passed to a dropout layer followed by a RELU and a softmax layer. Best test accuracy of 94% was achieved after hyper-parameter search. More details are in SI section A.1.

4 RESULTS

Zoonotic classification accuracy. In this study, we use an attention-enhanced classifier, as attention has been proved to be helpful for dealing with longer sequences (Bahdanau et al., 2016). In our dataset, the max sequence length is 910. Figure 2a shows the color coded length distribution of sequences for both positive and negative classes. The dataset is very length skewed. We see that the majority of zoonotic sequences are of length > 800, while the shorter length < 200 are mostly non-zoonotic. Despite the classification challenges presented by this skew, the best performing model achieved an overall accuracy of 94%. Figure 1 shows confusion matrices for different length ranges. For shorter length sequences ($len < 150$), the classifier correctly predicted > 92% of the true negatives, but missed a small number of true positives. Improved predictions of true positives was achieved with increase in sequence length. F1-scores are reported in SI Table 1 and section A.1.3.

Interpretation using Attention Visualization. Mapping predictions back to a target sequence has been used both as a means of investigating a given model's performance and as a method of sequence analysis. For example, convolutional filter visualization has been used for genome sequence classification (Bartoszewicz et al., 2021). Protein sequence position importance was investigated by an attention function in subcellular localization classification task (Jurtz et al., 2017). Recently, 3D visualization of attention of a Transformer protein sequence model revealed accurate capture of structural and binding properties (Vig et al., 2020). Similarly, in this study, we visualize and identify
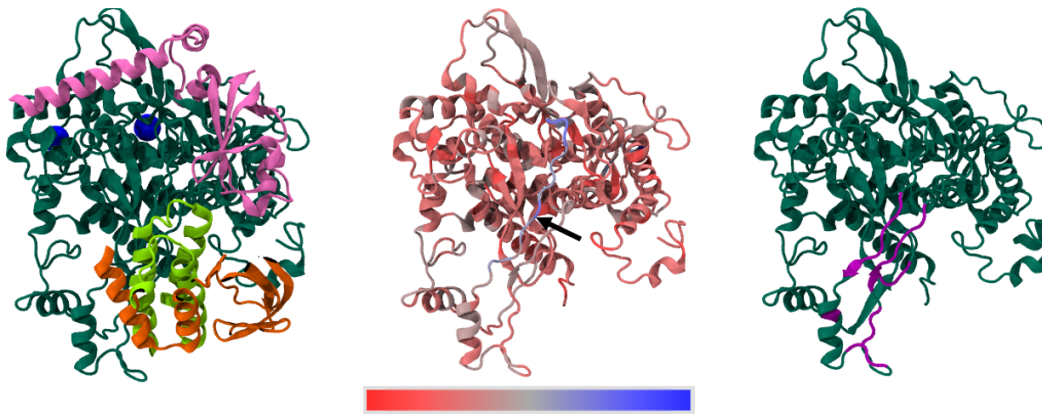
Figure 3: $Left$: Structure of SARS-CoV nsp12 RdRp (dark green) bound to nsp7 (light green) and nsp8 (pink) co-factors (pdb id:6NUR). Metal ions shown as blue spheres. $Middle$: Average attention differences between two classes, mapped to the SARS-CoV nsp12 structure (6NUR.pdb chain A residue 146 to 714). Red-gray-blue (low to high) colorscale is used to visualize the attention heatmap. $Right$: Regions of RdRp interacting (cutoff: 6.5 Å) with nsp7 highlighted in magenta.

regions of 1D protein sequence and 3D protein structure that are important for the prediction of zoonotic potential. Figure 2c shows averaged normalized attentions for each amino acid character, estimated by averaging over all sequences within the same class. At the amino acid level, significant differences in attention are evident between two classes. For example, relatively higher attention in positive sequences is found in the amino acids D, Y, R, F, and Q.

Since the 3D structure of the protein determines its function, we project the average attention heatmap differences between the two classes on the publicly available SARS-CoV nsp12 RdRp structure bound to the partnering nsp7 and nsp12 cofactor proteins (Kirchdoerfer & Ward, 2019)(Figure 3). Figure 2b show these attention differences as a function of residue number of nsp12 template. For selection criteria refer SI section A.2.1. Figure 5 further highlights these differences mapped on to the homology model (using SARS-CoV nsp12 as template) of individual sequences from the positive and negative class. These visualizations clearly show higher attentions within the positive sequences map to the N-terminal structural regions that correspond to the nsp12-nsp7 interaction surface. This result is interesting in light of prior work showing that the assembly of RdRp with cofactor proteins is crucial for polymerase activity that governs viral replication (Yin et al., 2020; Subissi et al., 2014). In fact, protein-protein interaction design has been shown as a proven approach for RdRp inhibition in other related viruses (Subissi et al., 2014).

5 DISCUSSION AND FUTURE WORK

We present an attention-enhanced LSTM deep neural net classifier that predicts zoonotic transmission in β-CoVs. Crucially, this classification was achieved using widely available RdRp sequences. Accurate prediction of zoonotic potential from RdRp sequences in β-CoVs (or similarly used reliable primer sequences for other viral families) has the potential to alleviate a critical bottleneck in viral surveillance, which has so far been unable to infer zoonotic potential of novel viruses discovered as part of routine surveillance. Further investigation of attention maps at both the sequence and the structure level provided important interpretability of the "black-box" model. Our results indicate mapping of class-level attention differences at protein-protein interaction sites of the RdRp structure. Those sites are of significant functional importance for binding with cofactors crucial for polymerase activity involved in viral replication. Thus, while 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 will explore more sophisticated attention mechanism; multi-head (e.g. Transformers (Devlin et al., 2019) (Vaswani et al., 2017)) for better interpretability and prediction. We also plan extend the current study beyond β-CoVs to investigate the robustness in capturing of biological context via model features.

REFERENCES

S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

Nicole C Ammerman, Magda Beier-Sexton, and Abdu F Azad. Growth and maintenance of Vero cell lines. *Current Protocols in Microbiology*, 11(1):A.4E.1–A.4E.7, 2008.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.

Jakub M Bartoszewicz, Anja Seidel, and Bernhard Y Renard. Interpretable detection of novel human viruses from genome sequencing data. *NAR genomics and bioinformatics*, 3(1):lqab004, 2021.

Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic Acids Res.*, 41(Database issue):D36–42, January 2013.

Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000. ISSN 0305-1048. doi: 10.1093/nar/28.1.235. URL https://doi.org/10.1093/nar/28.1.235.

Renee WY Chan, Maged G Hemida, Ghazi Kayali, Daniel KW Chu, Leo LM Poon, Abdelmohsen Alnaeem, Mohamed A Ali, Kin P Tao, Hoi Y Ng, Michael CW Chan, Yi Guan, John M Nicholls, and JS Malik Peiris. Tropism and replication of Middle East respiratory syndrome coronavirus from dromedary camels in the human respiratory tract: an in-vitro and ex-vivo study. *The Lancet Respiratory Medicine*, 2(10):813–822, 2014.

Daniel K. W. Chu, Kenrie P. Y. Hui, Ranawaka A. P. M. Perera, Eve Miguel, Daniela Niemeyer, Jincun Zhao, Rudragouda Channappanavar, Gytis Dudas, Jamiu O. Oladipo, Amadou Traore, Ouafaa Fassi-Fihri, Abraham Ali, Getnet F. Demissie, Doreen Muth, Michael C. W. Chan, John M. Nicholls, David K. Meyerholz, Sulyman A. Kuranga, Gezahegne Mamo, Ziqi Zhou, Ray T. Y. So, Maged G. Hemida, Richard J. Webby, Francois Roger, Andrew Rambaut, Leo L. M. Poon, Stanley Perlman, Christian Drosten, Veronique Chevalier, and Malik Peiris. MERS coronaviruses from camels in Africa exhibit region-dependent genetic diversity. *Proceedings of the National Academy of Sciences*, 115(12):3144–3149, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Christine LP Eng, Joo Chuan Tong, and Tin Wee Tan. Predicting zoonotic risk of influenza a viruses from host tropism protein signature using random forest. *International journal of molecular sciences*, 18(6):1135, 2017.

Ilya R Fischhoff, Adrian A Castellanos, Joao PGLM Rodrigues, Arvind Varsani, and Barbara A Han. Predicting the zoonotic capacity of mammal species for sars-cov-2. *bioRxiv*, 2021.

Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–18, 2020. ISSN 2162-2388. doi: 10.1109/tnnls.2020.3019893. URL http://dx.doi.org/10.1109/TNNLS.2020.3019893.

Xing-Yi Ge, Jia-Lu Li, Xing-Lou Yang, Aleksei A. Chmura, Guangjian Zhu, Jonathan H. Epstein, Jonna K. Mazet, Ben Hu, Wei Zhang, Cheng Peng, Yu-Ji Zhang, Chu-Ming Luo, Bing Tan, Ning Wang, Yan Zhu, Gary Crameri, Shu-Yi Zhang, Lin-Fa Wang, Peter Daszak, and Zheng-Li Shi. Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*, 503(7477):535–538, 2013.

S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer and J. F. Kolen (eds.), *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Vanessa Isabell Jurtz, Alexander Rosenberg Johansen, Morten Nielsen, Jose Juan Almagro Armenteros, Henrik Nielsen, Casper Kaae Sønderby, Ole Winther, and Søren Kaae Sønderby. An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics*, 33(22):3685–3690, 2017.

Matthew Kaye, Julian Druce, Thomas Tran, Renata Kostecki, Doris Chibo, Jessica Morris, Mike Catton, and Chris Birch. SARS-associated coronavirus replication in cell lines. *Emerging Infectious Diseases*, 12(1):128–133, 2006.

Robert N Kirchdoerfer and Andrew B Ward. Structure of the sars-cov nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nature communications*, 10(1):1–9, 2019.

Susanna KP Lau, Renate Wernery, Emily YM Wong, Sunitha Joseph, Alan KL Tsang, Nissy Annie Georgy Patteril, Shyna K Elizabeth, Kwok-Hung Chan, Rubeena Muhammed, Jöerg Kinne, Kwok-Yung Yuen, Ulrich Wernery, and Patrick CY Woo. Polyphyletic origin of MERS coronaviruses and isolation of a novel clade A strain from dromedary camels in the United Arab Emirates. *Emerging Microbes and Infections*, 5(12):e128, 2016.

Han Li and Fengzhu Sun. Comparative studies of alignment, alignment-free and svm based approaches for predicting the hosts of viruses based on viral sequences. *Scientific reports*, 8(1):1–9, 2018.

Baojing Lu, Ling Tao, Ting Wang, Zhenhua Zheng, Bao Li, Ze Chen, Yi Huang, Qinxue Hu, and Hanzhong Wang. Humoral and cellular immune responses induced by 3a DNA vaccines against severe acute respiratory syndrome (SARS) or SARS-like coronavirus in mice. *Clinical and Vaccine Immunology*, 16(1):73–77, 2009.

Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation, 2015.

Florian Mock, Adrian Viehweger, Emanuel Barth, and Manja Marz. Viral host prediction with deep learning. *bioRxiv*, pp. 575571, 2019.

Lorenzo Subissi, Clara C Posthuma, Axelle Collet, Jessika C Zevenhoven-Dobbe, Alexander E Gorbalenya, Etienne Decroly, Eric J Snijder, Bruno Canard, and Isabelle Imbert. One severe acute respiratory syndrome coronavirus protein complex integrates processive rna polymerase and exonuclease activities. *Proceedings of the National Academy of Sciences*, 111(37):E3900–E3909, 2014.

C.D. Town, R.A. Halpin, A. Ransier, N. Fedorova, L. Overton, T. Tsitrin, T. Stockwell, P. Amedeo, B. Bishop, H. Chen, P. Edworthy, N. Gupta, D. Katzel, K. Li, S. Schobel, S. Shrivastava, V. Thovarai, S. Wang, N. Powers, S. Grimes, B. Yamamoto, A. Robb, D.E. Wentworth, and D.L. Noah. SARS coronavirus Tor2 isolate Tor2/FP1-10851, complete genome, 2012. `https://www.ncbi.nlm.nih.gov/nuccore/JX163924.1` [Accessed: 29 April 2020].

Stephen KW Tsui, Stephen SC Chim, and YM Dennis Lo. Coronavirus genomic-sequence variations and the epidemiology of the severe acute respiratory syndrome. *The New England Journal of Medicine*, 349(2):187–188, 2003.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

Jia Wang, Chuang Ma, Zheng Kou, Yan-Hong Zhou, and Huai-Lan Liu. Predicting transmission of avian influenza a viruses from avian to human by using informative physicochemical properties. *International journal of data mining and bioinformatics*, 7(2):166–179, 2013.

M Wang, MY Yan, HF Xu, W Liang, B Kan, BJ Zheng, HL Chen, H Zheng, YM Xu, EM Zhang, HX Wang, JR Ye, GC Li, MC Li, ZG Cui, YF Liu, RT Guo, XN Liu, LH Zhan, DH Zhou, AL Zhao, R Hai, DZ Yu, Y Guan, and JG Xu. SARS-CoV infection in a restaurant from palm civet. *Emerging Infectious Diseases*, 11(12):1860, 2005.

Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, Rosalba Lepore, and Torsten Schwede. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research*, 46(W1):W296–W303, 05 2018. ISSN 0305-1048. doi: 10.1093/nar/gky427. URL https://doi.org/10.1093/nar/gky427.

Beibei Xu, Zhiying Tan, Kenli Li, Taijiao Jiang, and Yousong Peng. Predicting the host of influenza viruses based on the word vector. *PeerJ*, 5:e3579, 2017.

G. Yi. SARS coronavirus HC/SZ/61/03, complete genome, 2005. https://www.ncbi.nlm.nih.gov/nuccore/AY515512.1 [Accessed: 29 April 2020].

Wanchao Yin, Chunyou Mao, Xiaodong Luan, Dan-Dan Shen, Qingya Shen, Haixia Su, Xiaoxi Wang, Fulai Zhou, Wenfeng Zhao, Minqi Gao, et al. Structural basis for inhibition of the rna-dependent rna polymerase from sars-cov-2 by remdesivir. *Science*, 368(6498):1499–1504, 2020.

Zheng Zhang, Zena Cai, Zhiying Tan, Congyu Lu, Taijiao Jiang, Gaihua Zhang, and Yousong Peng. Rapid identification of human-infecting viruses. *Transboundary and emerging diseases*, 66(6): 2517–2522, 2019.

A APPENDIX

A.1 MODEL DETAILS

A.1.1 THE MODEL

For sequential data, Recurrent Neural Networks were introduced but as the sequence grows, the RNNs, weights could grow beyond control or vanish. Long Short-Term Memory (Hochreiter & Schmidhuber, 1997) was proposed to deal with vanishing gradient problem (Hochreiter et al., 2001) and learn long-term dependency among longer time period. Attention further learns these dependencies in a smarter and relevant way (Luong et al., 2015; Galassi et al., 2020; Bahdanau et al., 2014).

LSTM estimates conditional probability $p(y_1, ..., y_T | x_1, ..., x_L)$. Our task was binary classification, so the input was sequence of characters $(x_1, ..., x_L)$ and output was binary probability vector among two classes, positive and negative such that $\sum_{t=1}^{2} p(y_t) = 1$. The LSTM computes this conditional probability by first obtaining the fixed-dimensional representation h of the input sequence and given by the last hidden state of the LSTM. We enhanced our model with attention so, a final hidden attentional state h_t' was computed which was passed to a dropout layer, RELU and followed by softmax to obtain predictive distribution over classes.

For the attention part, we used multiplicative attention (Luong et al., 2015). All the hidden states of the encoder LSTM h_t were used to compute context vector c_t. A variable length alignment vector was computed using current target hidden state h_t with each source hidden state h_s

$$\alpha_t(s) = align(h_t, h_s) = \frac{\exp(score(h_t, h_s))}{\sum_{s'} \exp(score(h_t, h_{s'}))} \tag{1}$$

Here, $score$ is content-based function for which we used dot product, refer (Luong et al., 2015) for other alternatives.

(a) $len < 150$ (b) $150 < len < 500$ (c) $len > 500$ (d) all length
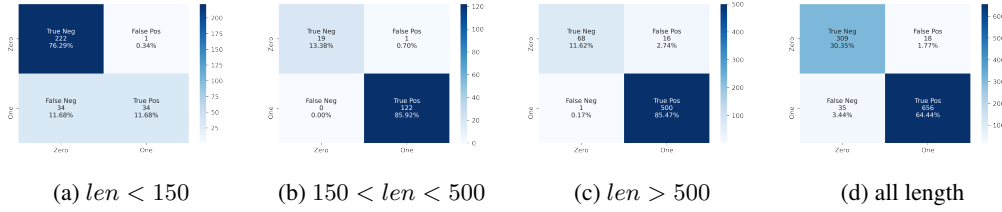
Figure 4: Model performance confusion matrix computed over range of sequence lengths for balanced dataset refer SI section A.1.3.

$$score(h_t, h_s) = h_t^T h_s \qquad (2)$$

Given the alignment vector as weights, the context vector c_t is computed as the weighted average over all the source hidden states.(Luong et al., 2015).

We took the dot product of the last hidden state as the target h_t and context vector c_t produce attentional hidden state. This attentional vector h_t' was then passed to softmax layer get predictive distribution over the positive and negative class:

$$p(y_t | y < t, x) = softmax(W_s, h_t') \qquad (3)$$

A.1.2 TRAINING DETAILS

We used a unidirectional LSTM with single layer and single head attention with 1000 cells (max sequence length) and 200 dimensional word embeddings (amino acid characters in our case) with a vocabulary of 21 amino acids. Hyper-parameter search was conducted and model giving best accuracy was found with the configuration of input embedding layer of size 200, lstm hidden layer of size 200, dropout 0.2, trained for 20 epochs. The baseline LSTM model with no attention gave 85% accuracy.

A.1.3 MODEL PERFORMANCE IMPROVEMENTS ON BALANCED DATASET

Due to the imbalanced class distribution as a function of sequence length, model performs poorly in lower sequence length regime len < 150. It seems to overfit to negative class and hence we saw 0 True positives in len $len < 150$ fig 1. Also, relatively poor performance with mid-length range. To overcome this problem, we adopted techniques to balance our dataset especially in lower length regime by oversampling the positive samples. In the imbalanced dataset we had around 195 positive sequences with len < 400 compared to 1100 negative class sequences (refer figure 2a). So, we made 5 copies of each positive sample with length < 400 which balances the dataset in lower length regime. Model was retrained on balanced data using configuration reported in SI section A.1.2

This improved the F1 score significantly (from 0 to 0.660) in the lower length regime. Improvements in recall and F1-scores of lower and mid-length ranges can be seen in table 1. The overall F1-score of the model for all length also gets improved from 0.946 to 0.961. Figure 4 shows confusion matrices of model performance trained on balanced dataset. In the future work, we will explore more techniques to balance the dataset.

A.2 ATTENTION ANALYSIS METHODS

For interpreting the model, we carried different analysis on the attention vectors. Visualization of average attention vectors looked into the overall trends distributed into positive and negative classes at the structure level. In addition, we also investigated attention at each sample/individual level.

8

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**

Table 1: Model performance report using accuracy, precision, recall and F1-score for both imbalanced skewed dataset and balanced (oversampled positive datapoints) dataset over all sequence length ranges shows improvement in model performance by improvement in F1-scores in balanced dataset. Lower length regime ($len < 150$) shows significant increase in F1-score. Interesting significant improvements are highlighted by bold text. Precision is not defined(**NaN**) in len< 150 for imbalanced data and False negatives > 0 in this case so we take F1 to be 0 as model failed to identify any True positives when it could have done.

A.2.1 SEQUENCE SELECTION CRITERIA FOR AVERAGE ATTENTION ANALYSIS

We used attention vectors of the trained LSTM model and carried out detailed analysis on them. We computed average attention (shown in figure 3) of filtered positive (zoonotic) and negative (non-zoonotic) classes, to have an insight into what model learns about structural differences between the two classes. For mapping to a template structure, positive and negative sequences that are highly similar to SARS-CoV nsp12 RdRp structure bound to nsp7 and nsp8 co-factors (pdbid:6NUR) were selected.

Sequence similarity between RdRP sequences and fixed target 6NUR was computed using NCBI BLAST similarity search (Altschul et al., 1990). Command used is 'blastp' using an evalue of 10, word size of 3, the BLOSUM62 subsitution matrix, a gap open penalty of 9, a gap extension penalty of 1, threshold of 16, comp based stats set to 0 and window size of 15. Sequences with $alignment$ $len > 900$, $identities > 90\%$, $gaps = 0$ and predicted correctly (true positives and true negatives) by our classifier model are selected. 277 positive and 26 negative sequences are chosen for this analysis shown in figures 3 and 2b.

A.2.2 HOMOLOGY MODELING FOR INDIVIDUAL LEVEL ATTENTION ANALYSIS

For visualizing attention at the structural (3D) level of individual positive and negative sequences (shown in figure 5), their structure prediction was required as the structures for these surveilled RdRp sequences were not available in the Protein Data Bank(PDB (Berman et al., 2000)). For the same, homology modeling was conducted using SWISS-MODEL server (Waterhouse et al., 2018) using NSP12 Chain A as the template. Figure 3, 6, 7 and 8 display attention visualization on structure and sequence level for pair of true positive and true negative long length and short length sequences.

Highly homologous sequences were filtered using $identities > 70\%$ and $gaps = 0$. Alignment length > 900 gave longer length selections and for shorter length we set alignment length between 200 to 300. We employed homology modeling of these highly similar sequences using Swiss-model server.
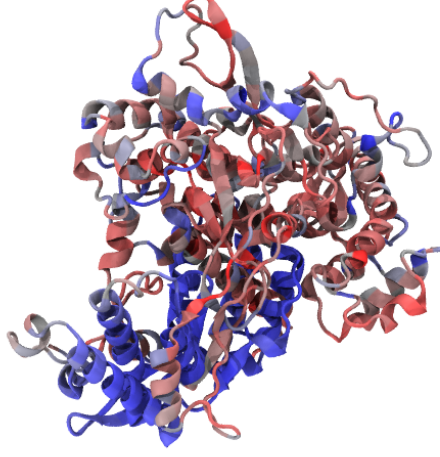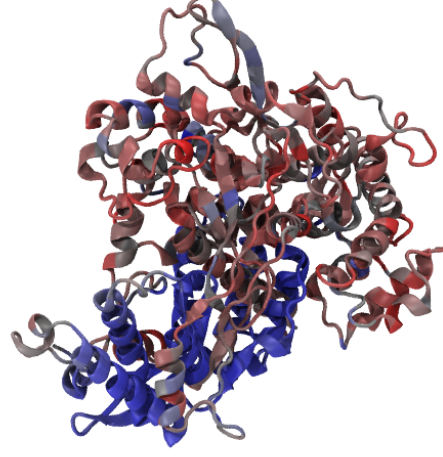
(a) Positive (zoonotic) sequence   (b) Negative (non-zoonotic) sequence

Figure 5: Attention heatmap of one positive and one negative sequences with length > 900 mapped on to their modeled structure obtained using homology modeling (6NUR.pdb chain A used as template). Colorscale used: red-gray-blue (low to high)

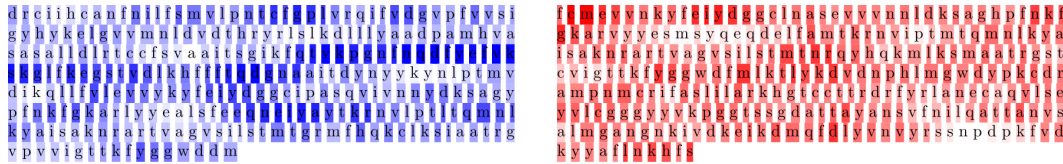(a) Positive(zoonotic) sequence, blast identities> 70%, classifier confidence score> 90%

(b) Negative(non-zoonotic) sequence, blast identities> 70%, classifier confidence score> 90%

Figure 6: Attention heatmap of one positive and one negative sequences with length > 900 at sequence 1D level. Positive with blue and negative with red.
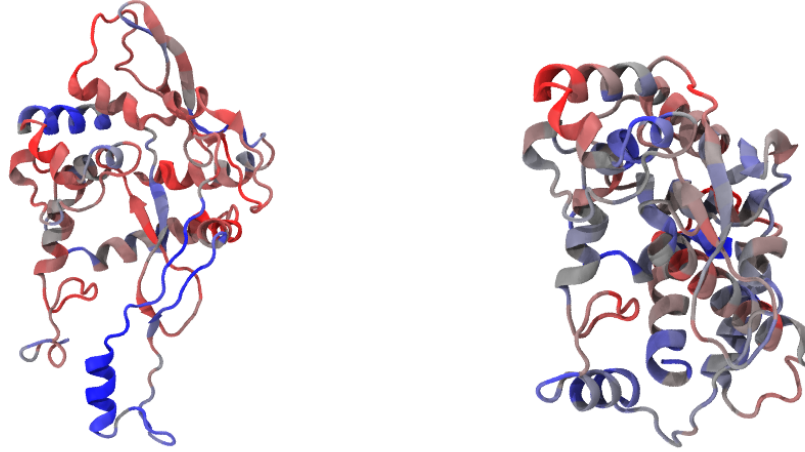
(a) Positive(zoonotic) sequence len 300, blast identities> 70%, classifier confidence score> 90%

(b) Negative(non-zoonotic) sequence len 300, blast identities> 70%, classifier confidence score> 90%

Figure 7: 3D attention heatmap, a pair of shorter length sequences that have high similarity (identities > 70%) with NSP12 chain A. For shorter length, used relaxed identities criteria as the overall length reduced so no sequences available with higher 90% identities.

(a) Positive(zoonotic) sequence len 300, blast identities> 70%, classifier confidence score> 97%

(b) Negative(non-zoonotic) sequence len 300, blast identities> 70%, classifier confidence score> 97%

Figure 8: Attention heatmap of one positive and one negative sequences with length 300 at sequence 1D level. Positive with blue and negative with red.