INTERPRETABLE PREDICTION OF THE INFECTIOUS POTENTIAL OF NOVEL VIRUSES

Jakub M. Bartoszewicz[*†]
Hasso Plattner Institute, Digital Engineering Faculty
University of Potsdam
14482 Potsdam, Germany
jakub.bartoszewicz@hpi.de

Anja Seidel[†‡]
MF1 Bioinformatics
Department of Methodology and Research Infrastructure
Robert Koch Institute
13353 Berlin, Germany
anjaseidel1@gmx.net

Bernhard Y. Renard[*]
Hasso Plattner Institute, Digital Engineering Faculty
University of Potsdam
14482 Potsdam, Germany
bernhard.renard@hpi.de

ABSTRACT

Viruses evolve extremely quickly, so reliable methods for viral host prediction are necessary to safeguard biosecurity and biosafety alike. Here, we predict whether a virus can infect humans directly from next-generation sequencing reads. We show that deep neural architectures significantly outperform both shallow machine learning and standard, homology-based algorithms, cutting the error rates in half and generalizing to taxonomic units distant from those presented during training. We propose a new approach for convolutional filter visualization to disentangle the information content of each nucleotide from its contribution to the final classification decision. Nucleotide-resolution maps of the learned associations between pathogen genomes and the infectious phenotype can be used to detect virulence-related genes in novel agents like the SARS-CoV-2 coronavirus. Recently published version of the paper is available at https://doi.org/10.1093/nargab/lqab004.

1 INTRODUCTION

Within a globally interconnected and densely populated world, pathogens can spread more easily than they ever had before. More unknown agents are yet to be discovered, given their extremely fast-paced evolution, unexplored biodiversity and increasing human exposure (Vouga & Greub, 2016; Trappe et al., 2016). Genome sequencing is the state-of-the-art approach for the open-view pathogen detection (Lecuit & Eloit, 2014; Calistri & Palù, 2015), but clinical samples are dominated by host reads and contaminants, with often less than a hundred reads of the pathogenic virus (Andrusch et al., 2018). Metagenomic assembly is challenging, especially in time-critical applications.

[*]Alternative address: MF1 Bioinformatics, Department of Methodology and Research Infrastructure, Robert Koch Institute, 13353 Berlin, Germany.

[†]Alternative address: Department of Mathematics and Computer Science, Free University of Berlin, 14195 Berlin, Germany

[‡]Current address: Central Research Institute of Ambulatory Health Care, 10587 Berlin, Germany.

Two recent studies employ k-NN classifiers (Li & Sun, 2018) and deep learning (Mock et al., 2019) to predict host range for three well-studied species directly from viral sequences. While those approaches are limited to those particular species and do not scale to viral host-range prediction in general, Gałan et al. (2019) use logistic regression and SVMs to predict if a novel virus infects bacteria, plants, vertebrates or arthropods. However, it requires sequences of at least 3kb. This is incompatible with metagenomic sequencing workflows, where the reads are at least 10-20 times shorter. Another study used gradient boosting machines to predict reservoir hosts and transmission via arthropod vectors for known human viruses (Babayan et al., 2018). Zhang et al. (2019) developed a k-NN classifier explicitly predicting whether a new virus can potentially infect humans for sequences as short as 150bp.

To foster interpretability of CNNs for genomics, learned filters can be visualized by forward-passing multiple sequences and extracting the most-activating subsequences (Alipanahi et al., 2015) to create a position weight matrix (PWM). Visualization via direct optimization is problematic, as it results in generating a dense matrix even though the input sequences are one-hot encoded (Lanchantin et al., 2016; 2017). This can be alleviated with Integrated Gradients (Sundararajan et al., 2016; Jha et al., 2019) or DeepLIFT, which propagates activation differences relative to a selected reference back to the input, reducing the overhead of obtaining accurate gradients (Shrikumar et al., 2017).

2 METHODS

We accessed the Virus-Host Database (Mihara et al., 2016) on July 31, 2019 and downloaded all available data. In total, the dataset contained 9,496 viruses with curated host information, including 1,309 human viruses. We generated a training set containing 80% of the genomes, and validation and test sets with 10% of the genomes each. We simulated 250bp long Illumina reads following a previously described protocol (Bartoszewicz et al., 2019) using Mason (Holtgrewe, 2010), yielding 20 million reads for training, 2.5 million for validation and 2.5 million paired reads as the held-out test set, with a balanced number of reads per class.

We investigate modified CNN and LSTM architectures invariant to DNA reverse-complementarity. They guarantee identical predictions for both forward and reverse-complement orientations of any given nucleotide sequence and have been previously shown to accurately predict bacterial pathogenicity (Bartoszewicz et al., 2019). As two mates in a read pair should originate from the same virus, predictions obtained for them can be averaged for a boost in performance. If a contig or genome is available, averaging predictions for constituting reads yields a prediction for the whole sequence. We compare our networks to Zhang et al. (2019), the only other approach tested on raw NGS reads and detecting human viruses in a fully open view setting. We use the human blood DNA virome dataset that they used (Moustafa et al., 2017) for an unbiased comparison. As it consists of real 150bp reads, our CNN was retrained on the first 150bp of our training reads. We also tested the performance of using the standard approaches, dc-megablast and NGS mappers, to search against an indexed database of labeled training genomes. Both alignment and k-NN can yield conflicting predictions for the individual mates in a read pair, or no prediction at all. Similarly to Bartoszewicz et al. (2019), at least one match is needed to predict a label, and conflicting predictions are treated as if no match was found. Missing predictions lower both true positive and true negative rates.

To visualize the learned convolutional filters, we downsample a matching test set to 125,000 reads and pass it through the network. We used the DeepSHAP implementation (Lundberg & Lee, 2017) of DeepLIFT (Shrikumar et al., 2017) with an all-N (all-zero) reference to extract score-weighted subsequences with the highest contribution score per filter. We calculate average filter contributions to obtain a crude ranking of feature importance. To disentangle a nucleotide's information content (IC) from its contribution, we introduce partial Shapley values. For any given feature x_i, intermediate neuron y_j and the output z, we aim to measure how x_i contributes to z while regarding only the fraction of the total contribution of x_i that influences how y_j contributes to z. This differs from recently introduced contribution weight matrices (Avsec et al., 2021), where feature attributions are used as a representation of an identified transcription factor binding site irreducible to a given intermediate neuron. Using the formalism of DeepLIFT's multipliers and their reinterpretation in SHAP (Lundberg & Lee, 2017), we backpropagate the activation differences only along the paths "passing through" y_j. We define partial multipliers $m_{x_i z}^{(y_j)}$ (Eq. 1) and express them in terms of Shapley values and activation differences w.r.t. the expected activation values (reference activation). Calculating

Table 1: Performance on read pairs. Bowtie2 (Langmead & Salzberg, 2012), BWA-MEM (Li & Durbin, 2009) and BLAST (Camacho et al., 2009) yield no predictions for over 35%, 19% and 10% of the samples, respectively.

	ACC.	PRECISION	RECALL
k-NN	57.1	86.7	52.1
BOWTIE2	58.6	99.2	59.2
BWA-MEM	72.8	98.9	73.9
BLAST	80.6	98.4	79.1
CNN (OURS)	89.9	93.9	85.4
LSTM (OURS)	86.4	89.0	83.0

Table 2: Performance on whole available genomes. Negative class is the majority class. BLAST (reads) and our networks use read-wise majority vote or output averaging, k-NN (genome) and BLAST (genome) use contig-wise majority vote.

	AUPR	RECALL	SPECIFICITY	BALANCED ACC.
BLAST (READS)	N/A	85.5	95.1	90.3
CNN (OURS)	91.2	89.3	94.2	91.7
LSTM (OURS)	85.8	96.2	76.4	86.3
k-NN (GENOME)	N/A	93.9	71.6	82.8
BLAST (GENOME)	N/A	86.3	94.6	90.5

partial multipliers is equivalent to zeroing out the multipliers $m_{y_k z}$ for all $k \neq j$ before backpropagating $m_{y_j z}$ further. We define partial Shapley values $\phi_i^{(y_j)}(z; x)$ analogously to how Shapley values can be approximated by a product of multipliers and input differences w.r.t. the reference (Eq. 2):

$$\phi_{x_i z}^{(y_j)} = m_{x_i y_j} m_{y_j z} = \frac{\phi_i(y_j; x)\ \phi_j(z; y)}{(x_i - \mathrm{E}[x_i])(y_j - \mathrm{E}[y_j])} \tag{1}$$

$$\phi_i^{(y_j)}(z; x) = \phi_{x_i z}^{(y_j)}(x_i - \mathrm{E}[x_i]) = \frac{\phi_i(y_j; x)\ \phi_j(z; y)}{y_j - \mathrm{E}[y_j]} \tag{2}$$

For the first convolutional layer, $\phi_i^{(y_j)}(z; x)$ can be efficiently calculated given $\phi_j(z; y)$ and an all-zero reference. In this case we do not traverse any non-linearities and can directly use the linear rule (Shrikumar et al., 2017) to calculate $\phi_i(y_j; x)$ as a product of the weight w_i and the one-hot encoded input x_i. To analyse which parts of a genome are associated with the infectious phenotype, we scramble it into overlapping, 250bp windows and predict the average output at each position of the genome. We also calculate average contributions of each nucleotide. We rank genes by the average score and scan the genome with the learned filters to find genes enriched in activating "motifs".

3 RESULTS

Alignment-based approaches struggle with analysing novel pathogens (Table 1). k-NN classifier (Zhang et al., 2019) yields confident predictions most of the time, achieving higher accuracy for single reads (75.5%; our best: 87.8%) than for read pairs. Our method can also be used on assembled contigs and full genomes if they are available, as well as on read sets from pure, single-virus samples (Table 2). In the human blood virome dataset the positive class massively outnumbers the negative class, so all models achieve over 99% precision (Table 3), but significantly differ in balanced accuracy.

The average information content of our motifs is strongly correlated nucleotide-wise with IC of max-activation logos of Alipanahi et al. (2015) (Spearman's $r_s > 0.95$ for all but one contributing filter pair). The average IC is 0.04 bit higher (Wilcoxon test, $p < 10^{-15}$). Therefore, our contribution logos represent analogous "motifs", while extracting additional, nucleotide-level interpretations (Fig. 1). Some nucleotides consistently appear in the activating subsequences, but the sign of their contributions is opposite to the filter's (low-IC nucleotides of a different color, Fig. 1c). Those

Table 3: Performance on the human blood virome dataset. Positive class is the majority class.

	AUPR	RECALL	SPECIFICITY	BALANCED ACC.
k-NN	99.5	80.9	85.4	83.1
CNN (OURS)	>99.9	97.3	96.2	96.8
LSTM (OURS)	>99.9	88.2	95.5	91.8

"counter-contributions" may arise if a nucleotide with a negative weight forms a frequent motif with others with positive weights strong enough to activate the lter. Some lters seem to learn gapped motifs resembling a codon structure (Fig. 1c). To showcase the more general applicability of our approach beyond the viral host prediction domain, we extracted this lter from a previously published network predicting bacterial pathogenicity (Bartoszewicz et al., 2019). We analyzed a *Staphylococcus aureus* strain used by the authors as a case study. We discovered that the learned motif is indeed signi cantly enriched in coding sequences and detects amino acid repeats crucial for the function of a virulence factor sraP. A top-scoring viral lter presented in Fig. 1a is overall enriched in coding sequences in both Taï Forest ebolavirus (top hit: VP35 gene suppressing innate immune signaling) and SARS-CoV-2 (top hit: nucleocapsid gene).

(a) (b) (c)

Figure 1: Nucleotide contribution logos of example lters. 1a, 1b: High (red) and low (blue) mean contribution score. 1c: Gaps resembling a codon structure, bacteria (Bartoszewicz et al., 2019).

Figure 2: Taï Forest ebolavirus genome. Top and middle: scores predicted by LSTM and CNN. Heatmap: nucleotide contributions of CNN. Genes that can be detected are highlighted in black.

Most Taï Forest ebolavirus genes (6 out of 7) can be detected with visual inspection by nding peaks of elevated infectious potential (Fig. 2). While a viral genome contains usually only a handful of genes, by using the bacterial CNN (Bartoszewicz et al., 2019) and compiling a ranking of 870 annotated genes of the analyzed *S. aureus* strain we could test if the high-ranking regions are associated with pathogenicity. Indeed, all top three genes with known names and functions (sarR, sspB and hupB) are either directly engaged in virulence or are known regulators of virulence-involved genes. Finally, we analyzed the SARS-CoV-2 coronavirus. It is predicted to infect humans, even though the data was collected at least 5 months before its emergence. E and N genes were scored the highest (apart from an uncon rmed ORF10 of just 38aa downstream of N) by the CNN and the LSTM, respectively (see Appendix, Fig. 3a). In the receptor-binding domain of the spike protein, we nd a peak of high infectious potential (Appendix Fig. 3b and Fig. 4), elevated in comparison to RaTG13, a close, bat-infecting relative of SARS-CoV-2.

4 DISCUSSION

Our results are only as good as the training data used; high quality labels and sequences are needed to develop trustworthy models. This is especially important as the method assumes no mechanistic link between an input sequence and the phenotype of interest, and the input sequence constitutes only a small fraction of the target genome without a wider biological context. Joint visualization of information content and contributions identifies complex filter structures, genome-wide plots of phenotype predictions highlight erroneous and correct predictions at gene level, and nucleotide-resolution contribution maps help track those regions down to individual residues. However, aggregation and interpretation of those results beyond case studies is non-trivial, and a promising avenue for further research. Ultimately, experimental work and traditional sequence analysis are required to distinguish true hits from false positives. The code is available at https://tinyurl.com/ffht8kw.

ACKNOWLEDGEMENTS

We thank Melania Nowicka, Lothar H. Wieler and Yong-Zhen Zhang for their input.

REFERENCES

Babak Alipanahi, Andrew Delong, Matthew T. Weirauch, and Brendan J. Frey. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838, 2015. ISSN 1546-1696. doi: 10.1038/nbt.3300. URL https://www.nature.com/articles/nbt.3300.

Andreas Andrusch, Piotr W. Dabrowski, Jeanette Klenner, Simon H. Tausch, Claudia Kohl, Abdalla A. Osman, Bernhard Y. Renard, and Andreas Nitsche. PAIPline: pathogen identification in metagenomic and clinical next generation sequencing samples. *Bioinformatics*, 34 (17):i715–i721, 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty595. URL https://academic.oup.com/bioinformatics/article/34/17/i715/5093217.

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, pp. 1–13, February 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00782-6. URL https://www.nature.com/articles/s41588-021-00782-6.

Simon A. Babayan, Richard J. Orton, and Daniel G. Streicker. Predicting reservoir hosts and arthropod vectors from evolutionary signatures in RNA virus genomes. *Science*, 362(6414): 577–580, November 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aap9072. URL https://science.sciencemag.org/content/362/6414/577.

Jakub M Bartoszewicz, Anja Seidel, Robert Rentzsch, and Bernhard Y Renard. DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics*, 36(1):81–89, 07 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz541. URL https://doi.org/10.1093/bioinformatics/btz541.

Arianna Calistri and Giorgio Palù. Editorial commentary: Unbiased next-generation sequencing and new pathogen discovery: undeniable advantages and still-existing drawbacks. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, 60(6): 889–891, 2015. ISSN 1537-6591. doi: 10.1093/cid/ciu913.

Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, December 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-421. URL https://doi.org/10.1186/1471-2105-10-421.

Wojciech Gałan, Maciej Bąk, and Małgorzata Jakubowska. Host taxon predictor - a tool for predicting taxon of the host of a newly discovered virus. *Scientific Reports*, 9(1):3436, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-39847-2. URL https://www.nature.com/articles/s41598-019-39847-2.

Manuel Holtgrewe. Mason – a read simulator for second generation sequencing data. Technical Report FU Berlin, 2010. URL http://publications.imp.fu-berlin.de/962/ .

Anupama Jha, Joseph K. Aicher, Deependra Singh, and Yoseph Barash. Improving interpretability of deep learning models: splicing codes as a case study. bioRxiv, 2019. doi: 10.1101/700096. URL https://www.biorxiv.org/content/early/2019/07/14/700096 .

Jack Lanchantin, Ritambhara Singh, Zeming Lin, and Yanjun Qi. Deep Motif: Visualizing Genomic Sequence Classifications. CoRR, abs/1605.01133, 2016. URL http://arxiv.org/abs/1605.01133 .

Jack Lanchantin, Ritambhara Singh, Beilun Wang, and Yanjun Qi. Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 22:254–265, 2017. ISSN 2335-6936. doi: 10.1142/9789813207813_0025.

Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. Nature methods, 9(4):357–359, 2012. ISSN 1548-7091. doi: 10.1038/nmeth.1923. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3322381/ .

Marc Lecuit and Marc Eloit. The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. Frontiers in Cellular and Infection Microbiology, 4:25, 2014. ISSN 2235-2988. doi: 10.3389/fcimb.2014.00025.

Fang Li. Structure, Function, and Evolution of Coronavirus Spike Proteins. Annual Review of Virology, 3(1):237–261, 2016. doi: 10.1146/annurev-virology-110615-042301. URL https://doi.org/10.1146/annurev-virology-110615-042301 .

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):10032, 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-28308-x. URL https://www.nature.com/articles/s41598-018-28308-x .

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics, 25(14):1754–1760, 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp324. URL https://academic.oup.com/bioinformatics/article/25/14/1754/225615 .

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), Advances in Neural Information Processing Systems 30, pp. 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf .

Aron Marchler-Bauer, Yu Bo, Lianyi Han, Jane He, Christopher J. Lanczycki, Shennan Lu, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, David I. Hurwitz, Fu Lu, Gabriele H. Marchler, James S. Song, Narmada Thanki, Zhouxi Wang, Roxanne A. Yamashita, Dachuan Zhang, Chanjuan Zheng, Lewis Y. Geer, and Stephen H. Bryant. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Research, 45(D1):D200–D203, 2017. ISSN 1362-4962. doi: 10.1093/nar/gkw1129.

Tomoko Mihara, Yosuke Nishimura, Yugo Shimizu, Hiroki Nishiyama, Genki Yoshikawa, Hideya Uehara, Pascal Hingamp, Susumu Goto, and Hiroyuki Ogata. Linking virus genomes with host taxonomy. Viruses, 8(3):66, 2016. ISSN 1999-4915. doi: 10.3390/v8030066.

Florian Mock, Adrian Viehweger, Emanuel Barth, and Manja Marz. Viral host prediction with deep learning. bioRxiv, pp. 575571, 2019. doi: 10.1101/575571. URL https://www.biorxiv.org/content/10.1101/575571v1 .

Ahmed Moustafa, Chao Xie, Ewen Kirkness, William Biggs, Emily Wong, Yaron Turpaz, Kenneth Bloom, Eric Delwart, Karen E. Nelson, J. Craig Venter, and Amalio Telenti. The blood DNA virome in 8,000 humans. PLOS Pathogens, 13(3):e1006292, March 2017. ISSN 1553-7374. doi: 10.1371/journal.ppat.1006292. URL https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1006292 .

Dora Pinto, Young-Jun Park, Martina Beltramello, Alexandra C. Walls, M. Alejandra Tortorici, Siro Bianchi, Stefano Jaconi, Katja Culap, Fabrizia Zatta, Anna De Marco, Alessia Peter, Barbara Guarino, Roberto Spreafico, Elisabetta Cameroni, James Brett Case, Rita E. Chen, Colin Havenar-Daughton, Gyorgy Snell, Amalio Telenti, Herbert W. Virgin, Antonio Lanzavecchia, Michael S. Diamond, Katja Fink, David Veesler, and Davide Corti. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature*, pp. 1–10, May 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2349-y. URL https://www.nature.com/articles/s41586-020-2349-y. Publisher: Nature Publishing Group.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3145–3153, International Convention Centre, Sydney, Australia, August 2017. PMLR. URL http://proceedings.mlr.press/v70/shrikumar17a.html.

Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Katharine HD Crawford, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, David Veesler, and Jesse D. Bloom. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. *bioRxiv*, pp. 2020.06.17.157982, June 2020. doi: 10.1101/2020.06.17.157982. URL https://www.biorxiv.org/content/10.1101/2020.06.17.157982v1. Publisher: Cold Spring Harbor Laboratory Section: New Results.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Gradients of Counterfactuals. *CoRR*, abs/1611.02639, 2016. URL http://arxiv.org/abs/1611.02639.

Kathrin Trappe, Tobias Marschall, and Bernhard Y. Renard. Detecting horizontal gene transfer by mapping sequencing reads across species boundaries. *Bioinformatics*, 32(17):i595–i604, September 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw423. URL https://academic.oup.com/bioinformatics/article/32/17/i595/2450753.

Manon Vouga and Gilbert Greub. Emerging bacterial pathogens: the past and beyond. *Clinical Microbiology and Infection*, 22(1):12–21, January 2016. ISSN 1198-743X. doi: 10.1016/j.cmi.2015.10.010. URL https://www.clinicalmicrobiologyandinfection.com/article/S1198-743X(15)00909-X/abstract.

Daniel Wrapp, Nianshuang Wang, Kizzmekia S. Corbett, Jory A. Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S. Graham, and Jason S. McLellan. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367(6483):1260–1263, March 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abb2507. URL https://science.sciencemag.org/content/367/6483/1260. Publisher: American Association for the Advancement of Science Section: Report.

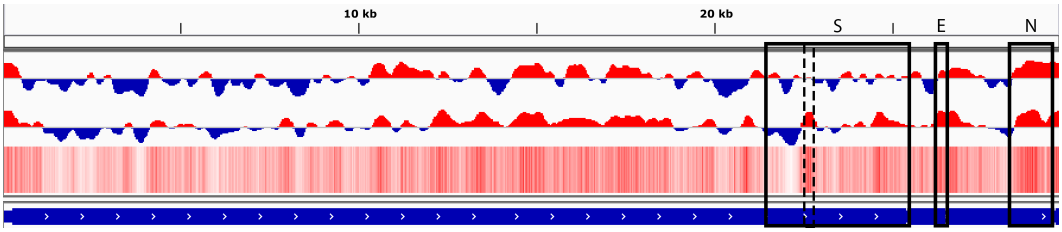
Meng Yuan, Nicholas C. Wu, Xueyong Zhu, Chang-Chun D. Lee, Ray T. Y. So, Huibin Lv, Chris K. P. Mok, and Ian A. Wilson. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. *Science*, 368(6491):630–633, May 2020. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abb7269. URL https://science.sciencemag.org/content/368/6491/630. Publisher: American Association for the Advancement of Science Section: Report.

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. doi: 10.1111/tbed.13314. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/tbed.13314.

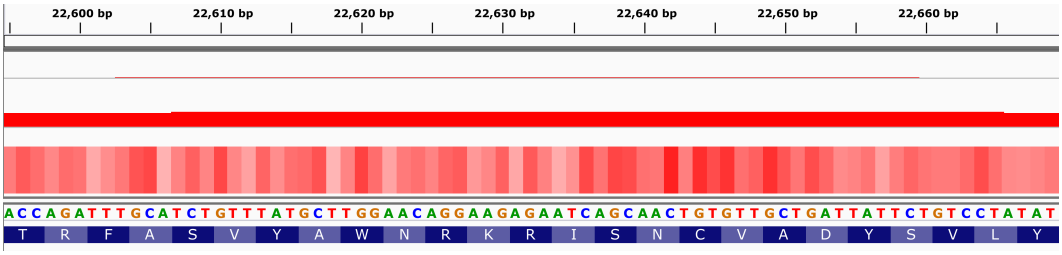
A Appendix

Fig. 3a presents a GWPA plot for the whole genome of the SARS-CoV-2 coronavirus, successfully predicted to infect humans, even though the data was collected at least 5 months before its emergence. Interestingly, its mean infectious potential (0.57 as scored by the CNN) is relatively close to

the decision threshold, while its closest known relative, a bat-infecting SARSr-CoV RaTG13, is actually falsely classified as a human virus with a slightly lower mean infectious potential (0.55). What is more, the gene encoding the spike protein, which plays a significant role in host entry Li (2016), has a mean score slightly above the threshold for SARS-CoV-2 (0.52) and below the threshold for RaTG13 (0.49). As shown in the GWPA plot (Fig. 3a), regions that the network has learned to associate with the infectious phenotype are distributed non-uniformly and tend to cluster together. This suggests that low-confidence mean prediction for those viruses is not a result of random guessing, but genuine ambiguity present in the data – and the misclassification of RaTG13 could be indicative of a general zoonotic potential of SARS-related coronaviruses. In the Fig. 3a, we highlighted the score peaks aligning the spike protein gene (S), as well as the E and N genes, which were scored the highest (apart from an unconfirmed ORF10 of just 38aa) by the CNN and the LSTM. Correlation between the CNN and LSTM outputs is significant, but species-dependent and moderate (0.28 for Ebola, 0.48 for SARS-CoV-2), which suggests they capture complementary signals.



(a)



(b)

Figure 3: SARS-CoV-2 coronavirus genome. Top: score predicted by the LSTM. Middle: score predicted by the CNN. Heatmap: nucleotide contributions of the CNN. Bottom, in blue: reference sequence. 3a: Whole genome and sequences encoding the spike protein (S), envelope protein (E) and nucleocapsid protein (N). 3b: Spike protein gene, a small peak (positions 22,595-22,669, dashed line in Fig. 3a) within the receptor-binding domain (predicted by CD-search, positions 22,517-23,185). Binding to the receptor is crucial for entry to the host cell. Local host adaptation could help switch hosts between the animal reservoir and humans.

Fig. 3b shows the nucleotide-level contributions in a small peak within the receptor-binding domain (RBD) of the S protein, crucial for recognizing the host cell. The domain location was predicted with CD-search (Marchler-Bauer et al., 2017) using the default parameters. The maximum score of this peak is noticeably higher for SARS-CoV-2 (0.87) than for its analog in RaTG13 (0.67). Fig. 4 presents the RBD in the structural context of the whole S protein (PDB ID: 6VSB, (Wrapp et al., 2020)), as well as in complex with a SARS-neutralizing antibody CR3022 (PDB ID: 6W41, (Yuan et al., 2020)). The high score peak roughly corresponds to one of the regions associated with reduced expression of the RBD (Starr et al., 2020), located in the core-RBD subdomain. It covers over 71% of the CR3022 epitope, as well as the neighbouring site of the N343 glycan. The latter is present in the epitope of another core-RBD targeting antibody, S309 (Pinto et al., 2020). All the per-residue average contributions in the region are positive, even in the regions of lower pathogenicity score, in accordance with the results presented in Fig. 3b.