REAL-TIME PATHOGENICITY PREDICTION DURING GENOME SEQUENCING OF NOVEL VIRUSES AND BACTERIA

Anonymous authors
Paper under double-blind review

ABSTRACT

Novel pathogens evolve quickly and may emerge rapidly, causing dangerous outbreaks or even global pandemics. Next-generation sequencing is the state-of-the art in open-view pathogen detection, and one of the few methods available at the earliest stages of an epidemic, even when the biological threat is unknown. Analyzing the samples as the sequencer is running can greatly reduce the turnaround time, but existing tools rely on close matches to lists of known pathogens and perform poorly on novel species. We train ResNets invariant to the reverse-complementarity of DNA, classify raw, incomplete Illumina and Nanopore reads and integrate our models with HiLive2, a real-time Illumina mapper. Our approach outperforms alternatives based on machine learning and sequence alignment on simulated and real data, including SARS-CoV-2 sequencing runs. After just 10% of the Illumina run is completed, we increase the true positive rate 80-fold compared to the live-mapping approach. The first 250bp of Nanopore reads, corresponding to 0.5s of sequencing time, are enough to yield predictions more accurate than mapping the finished long reads.

1 INTRODUCTION

If an outbreak involves a new, unknown pathogen, targeted diagnostic panels are not available at first. Open-view approaches must be used and next-generation sequencing of the pathogen's DNA or RNA is the method of choice (Lecuit & Eloit, 2014; Calistri & Palù, 2015). A swift response is crucial, as the number of cases and associated deaths rises exponentially. Analyzing the samples during the sequencing run, as the reads are produced, enables greatly improved turnaround times. Read mappers are a class of algorithms and tools designed explicitly to match the obtained DNA reads with the known genomes they most probably originate from. However, as they are designed for fast and precise sequence alignment, they are expected to miss most of the reads originating from genomes highly divergent from the available references. Therefore, even though existing real-time mappers, HiLive (Lindner et al., 2017) and HiLive2 (Loka et al., 2019), do cover standard read-based pathogen detection workflows, their performance on novel agents is limited by their dependence on databases of known species. In this work, we use ResNets invariant to DNA reverse-complementarity to predict if a read originates from a human pathogen and show that they are a promising alternative to read mapping if the correct reference genome is not yet known.

Deneke et al. (2017) have shown that taxonomy-dependent methods like read-mapping (with optional additional filtering steps), BLAST (Altschul et al., 1990; Camacho et al., 2009) or Kraken (Wood & Salzberg, 2014), which try to assign target sequences to their closest taxonomic matches, fail to yield any predictions for a significant fraction of reads originating from novel pathogens. BLAST was the best of those approaches, missing the least reads and achieving the highest accuracy at a relatively high computational cost. Deneke et al. (2017) showed that in the context of detecting novel bacterial pathogens, a random forest approach performs much better. Zhang et al. (2019) used a k-NN classifier to develop a similar method for detection of human-infecting viruses. An analogous approach based on CNNs and LSTMs, DeePaC, outperforms the traditional machine learning algorithms on both novel bacteria (Bartoszewicz et al., 2020) and viruses (DeePaC-vir, Bartoszewicz et al. (2021)). However, previously available models were optimized for a relatively long

read length of 250bp (with one viral model trained for 150bp reads). We suspected that they would underperform in a real-time analysis scenario, where much shorter reads are analyzed.

2 METHODS

Throughout this paper, we will use the term *subread* in a special sense: the first k nucleotides of a given sequencing read (in other words, a *prefix* of a read). We used the DeePaC and DeePaC-vir datasets of bacterial and viral reads originating from mixtures of pathogens and non-pathogens (Bartoszewicz et al., 2020; 2021). The training, validation and test sets contain different bacteria or viruses, so that the generalization error is estimated for novel agents. We used the test datasets to generate corresponding subread datasets with k between 25 and 250, explicitly modeling new information incoming during a sequencing run, as each subread length k corresponds to the kth cycle. To generalize over a large spectrum of possible subread lengths, we built mixed-length training and validation sets by randomly choosing a different k between 25 and 250 for every read in the training set. As Bartoszewicz et al. (2021) have previously presented both 250bp and 150bp-trained CNN classifier for viruses, we also generated an analogous 150bp subread bacterial dataset, and used it to train a corresponding CNN.

First, we trained models based on two relatively shallow architectures shown previously to perform well in the pathogenicity or host-range prediction task – a wide reverse-complement CNN consisting of 2 convolutional layers and 2 fully-connected layers and a reverse-complement bidirectional LSTM. For more design details and the description of the reverse-complement variants of convolutional and LSTM layers, we refer the reader to Bartoszewicz et al. (2020; 2021). Those architectures guarantee identical predictions for sequences in their forward and reverse-complement orientations in a single forward pass. Previous work has established them as the method of choice in the read-based pathogenic or infectious potential prediction task. However, as short subreads convey less information, we expected the subread classification problem to be more challenging than in the case of relatively long 250bp reads. Therefore, we implemented a new architecture – a reverse-complement ResNet extending the previous work with skip connections (He et al., 2016) while maintaining invariance to reverse-complementarity. We considered 18- and 34-layer ResNet variants where all convolutional layers of a standard ResNet (including size-1 convolutions in skip connections) are replaced with reverse-complement convolutions (Bartoszewicz et al., 2020). After hyperparameter tuning (Appendix, Section A.1), we selected the 18-layer ResNets. We then combined HiLive2 with ResNets to extract reads mappable to known references and predict the phenotype for the unmapped reads. This enables identification of the closest relatives of the new pathogen, while still predicting labels for reads missed by the mapper. We benchmark our approach against previously published deep learning (Bartoszewicz et al., 2020; 2021), random forest (Deneke et al., 2017), kNN (Zhang et al., 2019) and alignment-based (Camacho et al., 2009; Li, 2018) methods, retraining our models on Nanopore data where necessary (Appendix, Section A.2).

3 RESULTS

Table 1 presents average performance over the whole sequencing run (all cycles for both mates) for the bacterial dataset. The highest accuracy is achieved by the ResNet-based hybrid classifier. High recall of DeePaC (CNN) is actually an artifact – its predictions for shorter subreads are extremely imprecise (precision for 25bp is is 50.6%), suggesting that the network simply classifies an overwhelming majority of short subreads as positive regardless of their actual sequence. This effect does not occur for our hybrid classifier, suggesting that although it achieves the second-highest true positive rate overall, it is likely the most sensitive method useful in practice.

For the viral dataset, a deep learning approach performs slightly better than HiLive2 even on the reads that HiLive2 is able to map. If only the mapped reads are considered, the ResNet correctly labels 90.7% of them, compared to 89.8% for HiLive2 itself. The effect is especially strong for reads 50bp and longer where the average accuracy of the ResNet rises to 91.8%, while HiLive2's stays the same. This is most probably the reason behind the better accuracy of the pure ResNet classifier in comparison to the hybrid classifier also when unmapped reads are considered, as presented in Table 2. Cycle-by-cycle accuracy comparisons are presented in the Appendix (Fig 1 and Fig 2).

Table 1: Average performance on reads from novel bacterial species across the whole sequencing run. The hybrid classifier combining HiLive2 and ResNet achieves the highest accuracy. Recall of DeePaC (CNN) is inflated by its unreliable predictions for short subreads. HiLive2 is the most precise method.

	Accuracy	Precision	Recall
HiLive2+ResNet (ours)	**83.6**	80.0	89.8
ResNet (ours)	82.1	79.2	86.7
DeePaC (CNN)	79.3	75.6	**91.4**
DeePaC (LSTM)	79.7	75.2	87.0
PaPrBaG	74.5	72.7	78.1
BLAST	60.8	84.8	76.1
HiLive2	22.2	**97.3**	36.3

Table 2: Average performance on reads from novel viruses across the whole sequencing run. ResNet alone achieves the highest accuracy and is the most sensitive method overall, while HiLive2 is the most precise.

	Accuracy	Precision	Recall
HiLive2+ResNet (ours)	85.9	93.1	77.6
ResNet (ours)	**86.5**	90.5	**81.5**
DeePaC (CNN-150)	84.8	92.3	75.6
DeePaC (CNN)	62.5	48.0	26.7
DeePaC (LSTM)	79.0	90.0	66.0
kNN	60.7	78.0	54.1
BLAST	73.2	97.2	73.7
HiLive2	51.1	**99.2**	50.3

We further evaluated our approach on data from a real SARS-CoV-2 sequencing run. Note that the training database did not contain a SARS-CoV-2 reference genome, mimicking the pre-pandemic state of knowledge. In this setting, we used BLAST as an example follow-up analysis of the reads filtered with our hybrid classifier or the pure deep learning approach after just 50 cycles (Table 3). As in the case of the open-view viral dataset, the neural network itself performs better than the hybrid classifier, being more accurate even on the reads mappable with HiLive2. Here, HiLive2 suffers from a high false negative rate of 96.3% even when only the mapped reads are considered, which is probably because it was designed for mapping against known references. Using the ResNet alone results in better performance. Notably, even the spurious non-pathogenic identifications of HiLive2 can be useful – 99.7% of the mapped reads are identified as originating from coronaviruses, and 90.5% are identified as bat coronaviruses, including the *Rhinolophus* (horseshoe bat) coronaviruses and bat SARS-like viruses, which are probably closely related to SARS-CoV-2.

Similar identifications can be made with BLAST on the much larger set of ResNet-filtered reads. As the deep learning models consistently outperform BLAST, we use them as predictors to extract reads of interest. A BLAST follow-up analysis annotates the selected reads with their closest taxonomic matches (which may include non-pathogens) wherever a match is found. Our results suggest that predictions of the ResNet are reliable, while offering a recall rate 80 times higher than HiLive2. However, further analysis steps (with BLAST or other approaches, e.g. taxonomic classifiers) are required to gain more fine-grained insights into the origin of ResNet-filtered reads.

Finally, we evaluated Nanopore-trained models to investigate possible applications to noisier long-read sequencing technologies (Table 4). As expected, mapping with minimap2 (Li, 2018) is the most precise method and Illumina-trained neural networks underperform in this context. Noisy reads are especially challenging for Illumina-trained LSTMs. Their precision and true positive rates become unstable, resulting in relatively low accuracy. Using Nanopore error models for training promotes more robust models. Strikingly, first 250bp of a read (corresponding to ca. 0.5s of sequencing time) are enough for our models to noticeably outperform minimap2, even when it uses whole reads. This holds for real data as well. When the correct reference is not yet available (as before the pandemic), Minimap2 recalls only 9.9% of full SARS-CoV-2 reads, compared to 52.7% for our ResNet.

Table 3: Reads identified as pathogenic from the SARS-CoV-2 sequencing run. The ResNet is able to identify the most reads, but cannot annotate them with matches to the closest known references. Combining HiLive2 (HL) or BLAST with the ResNet identifies taxonomic signals while extracting more reads than pure mapping. BLAST output can be used to indicate the closest taxonomic match.

	Reads	Recall	Annotations	Annot. rate
HL	3236	0.6%	3236	0.6%
HL+ResNet	211411	41.0%	3236	0.6%
ResNet	**264646**	**51.3%**	0	0.0%
HL+ResNet+BLAST	211411	41.0%	142794	27.7%
ResNet+BLAST	**264646**	**51.3%**	**195634**	**37.9%**

Table 4: Performance on Nanopore data. Minimap2 was evaluated on both full reads and 250bp subreads. ResNets were trained on Nanopore data with identical species composition as the Illumina data used for DeePaC CNNs and LSTMs and evaluated on 250bp subreads. Minimap2 yields no matches for between 13% (viruses, full length) and 69% (bacteria, 250bp) of the reads. Acc. – accuracy, Prec. – precision, Rec. – recall.

	Bacteria			Viruses		
	Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
ResNet (ours)	**78.5**	73.4	89.3	**88.5**	90.3	**86.4**
DeePaC (CNN)	77.8	73.1	87.8	81.4	84.0	77.6
DeePaC (LSTM)	73.7	66.5	**95.7**	78.5	92.3	62.1
minimap2 (250bp)	30.5	**97.2**	46.6	59.3	**99.2**	61.8
minimap2 (full)	66.7	91.4	79.7	80.2	98.9	82.4

4 Discussion

In this work, we present a new approach for real-time prediction of pathogenic potential of novel bacteria and viruses, accessing the intermediate files of an Illumina sequencer. We develop new deep learning models specialized in inference from incomplete short- and long-read sequencing data and show that they outperform the previous state-of-the-art on both simulated and real reads. The limitations of the previously described read-based methods of predicting pathogenic potentials with machine learning (Deneke et al., 2017; Zhang et al., 2019; Bartoszewicz et al., 2020; 2021) apply to this study too. Any assumptions and biases affecting the labels will be reflected by the trained classifier. As Nanopore-trained models perform relatively well despite higher sequencing noise, we imagine incorporating pathogenicity prediction into real-time selective sequencing workflows (Loose et al., 2016). Since 250bp subreads are enough to make predictions more accurate than possible with mapping even fully sequenced reads, it would be possible to terminate sequencing of some reads quickly to focus on sequencing those originating from pathogens. The code is available at https://tinyurl.com/2u8vjypn (real-time inference and HiLive2 integration) and https://tinyurl.com/ffht8kw (training and data preprocessing).

Acknowledgements

We thank Tobias P. Loka and Melania Nowicka for multiple valuable discussions and comments.

References

Steven F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.

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, 01 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz541. URL https://doi.org/10.1093/bioinformatics/btz541.

Jakub M. Bartoszewicz, Anja Seidel, and Bernhard Y. Renard. Interpretable detection of novel human viruses from genome sequencing data. *NAR Genomics and Bioinformatics*, 3(lqab004), February 2021. ISSN 2631-9268. doi: 10.1093/nargab/lqab004. URL https://doi.org/10.1093/nargab/lqab004.

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.

Carlus Deneke, Robert Rentzsch, and Bernhard Y. Renard. PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. *Scientific Reports*, 7:39194, 2017. ISSN 2045-2322. doi: 10.1038/srep39194. URL https://www.nature.com/articles/srep39194.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

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.

Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty191. URL https://doi.org/10.1093/bioinformatics/bty191. _eprint: https://academic.oup.com/bioinformatics/article-pdf/34/18/3094/25731859/bty191.pdf.

Yu Li, Sheng Wang, Chongwei Bi, Zhaowen Qiu, Mo Li, and Xin Gao. DeepSimulator1. 5: a more powerful, quicker and lighter simulator for Nanopore sequencing. *Bioinformatics*, 36(8): 2578–2580, 2020. doi: 10.1093/bioinformatics/btz963.

Martin S. Lindner, Benjamin Strauch, Jakob M. Schulze, Simon H. Tausch, Piotr W. Dabrowski, Andreas Nitsche, and Bernhard Y. Renard. HiLive: real-time mapping of illumina reads while sequencing. *Bioinformatics*, 33(6):917–319, 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw659. URL https://academic.oup.com/bioinformatics/article/33/6/917/2567469.

Tobias P. Loka, Simon H. Tausch, and Bernhard Y. Renard. Reliable variant calling during runtime of Illumina sequencing. *Scientific Reports*, 9(1):1–8, November 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-52991-z. URL https://www.nature.com/articles/s41598-019-52991-z.

Matthew Loose, Sunir Malla, and Michael Stout. Real-time selective sequencing using nanopore technology. *Nature Methods*, 13(9):751–754, September 2016. ISSN 1548-7105. doi: 10.1038/nmeth.3930. URL https://www.nature.com/articles/nmeth.3930. Number: 9 Publisher: Nature Publishing Group.

Franka J. Rang, Wigard P. Kloosterman, and Jeroen de Ridder. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, 19(1):90, July 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1462-9. URL https://doi.org/10.1186/s13059-018-1462-9.

Derrick E. Wood and Steven L. Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, 15(3):R46, 2014. ISSN 1474-760X. doi: 10.1186/gb-2014-15-3-r46. URL https://doi.org/10.1186/gb-2014-15-3-r46.

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

A.1 ResNets and hybrid classifiers

We trained 18- and 34-layer ResNets invariant to the reverse-complementarity of the inputs (see Table 5 for architecture details) for a maximum of 30 epochs, using early stopping with a patience of 10 epochs. For all models, we used input dropout, which may be understood as switching a random fraction of the input nucleotides to Ns. As generating subreads already discards some sequence information, we retuned the input dropout rate for the bacterial models, testing the values of 0.2 and 0.25. For the viral models, it was already shown that the dropout rate of 0.25 works better even in the case of 150bp subreads; we therefore only considered the higher value. We compared the CNN, LSTM and ResNet models trained on the mixed-length datasets; in addition to that we also considered the bacterial CNN trained on 150bp subreads analogous to the viral $\text{CNN}_{\text{All-150}}$ from Bartoszewicz et al. (2021). The ResNet-18 trained with an input dropout rate of 0.25 achieved the highest accuracy on the bacterial mixed-length validation set and was selected for further evaluation. For viruses, the ResNet models were the best as well – although the ResNet-34 was the most accurate in absolute terms, the error rate improvement over the 18-layer variant was negligible (<0.5%) while the computational cost (measured in wall-clock time of both training and inference) was roughly twice as high. Since inference speed is crucial for the application presented here, we decided to select the equally accurate but faster and more efficient ResNet-18.

Table 5: ResNet architecture details. Conv1 and first layers of stages Conv3-Conv5 use a stride of 2, and all other layer use the stride of 1. Stages 2-5 consist of multiple layers with the same filter width and number of filters. Batch normalization is used after all hidden layers. After the convolutions, we use global average pooling and a fully-connected output layer.

stage	ResNet-18	ResNet-34
conv1	filter width:7, filters:64	filter width:7, filters:64
conv2	[filter width:5, filters:64] x 4	[filter width:5, filters:64] x 6
conv3	[filter width:5, filters:128] x 4	[filter width:5, filters:128] x 8
conv4	[filter width:5, filters:256] x 4	[filter width:5, filters:256] x 12
conv5	[filter width:5, filters:512] x 4	[filter width:5, filters:512] x 6
pool	global av. pooling	global av. pooling
out	1-unit fully-connected	1-unit fully-connected

A.2 Benchmarking

We compare our neural networks and hybrid classifiers (Fig. 1) to the original DeePaC models, as well as an alternative random forest approach, PaPrBaG (Deneke et al., 2017). We trained a DNA-only PaPrBaG forest (Bartoszewicz et al., 2020) on the mixed-length bacterial dataset. For both machine learning approaches, we average the predictions for both mates of a read pair for a boost in accuracy (Bartoszewicz et al., 2020). We benchmark the viral model (Fig. 2) against a k-nearest neighbors (kNN) virus host classifier by Zhang et al. (2019). We train the kNN as described by the authors, using non-overlapping 500bp long "contigs" generated from the source genomes. Training based on simulated reads was not possible due to high computational cost, but Zhang et al. (2019) showed that a model trained this way can be used to predict pathogenic potentials of short NGS reads.

In addition to that, we evaluate two alignment-based methods – HiLive2 in the "very-accurate" mode (Lindner et al., 2017; Loka et al., 2019) and dc-Megablast (Camacho et al., 2009) with an E-value cutoff of 10 and the default parameters. A successful match to a pathogen reference genome is treated as a positive prediction; a match to a nonpathogen is a negative. In case of multiple matches, the top hit is selected. We build the HiLive2 FM-index and the BLAST database using

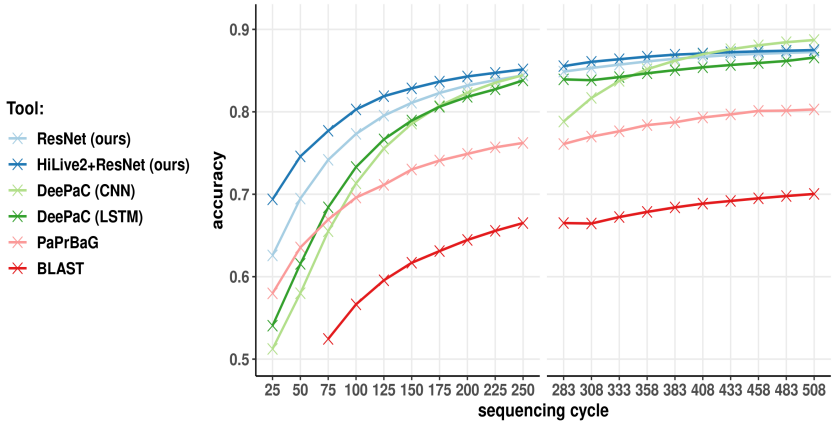

Figure 1: Accuracy for the bacterial dataset. Only the tools achieving more than 60% average accuracy are shown. For BLAST, accuracy values below 50% are not shown. Combining HiLive2 with our ResNet results in the best average accuracy overall. Cycle numbers for the second mate are shifted by 8 due to an 8nt-long simulated barcode present in the BCL files, which is removed at mapping and prediction time.

all the genomes used for training read set generation. If BLAST aligns two mates of a read pair to genomes with conflicting labels (i.e. one pathogen and one non-pathogen), we treat them both as missing predictions. As kNN yields binary predictions, we integrate them using the same approach. For HiLive2, we treat them separately, as high precision of HiLive2 warrants considering all the obtained matches as relevant. If only one mate has a match, we propagate the match to the other mate. We calculate the performance measures taking all the reads in the sample into account. Hence, missing predictions affect both true positive and true negative rates.

We also evaluate our methods on real data from a SARS-CoV-2 novel coronavirus sequencing run. The virus was not present in the training database, as it had not yet been discovered when the DeePaC-vir datasets were compiled. We downloaded an archive of 151bp-long paired-end reads originating from a COVID-19 positive human from San Diego county (SRR11314339). To showcase how our methods can be used for rapid detection of novel biological threats, we evaluate the performance of our classifiers after just 50 sequencing cycles. As the predictions of the deep learning approaches do not offer any information about the closest known relative of a novel pathogen, we extend the workflow with using BLAST on reads prefiltered by our models. This enables a drastic increase in the pathogen read identification rate while also generating more insight into their biological meaning. Using BLAST on full NGS datasets is usually not feasible because of the computational cost. What is more, is has been previously shown (Deneke et al., 2017; Bartoszewicz et al., 2020; 2021) that machine learning approaches perform better in pathogenic potential prediction tasks. Therefore, we see the combination of a filtering step with a BLAST follow-up as an in-depth analysis of the reads of interest while discarding the potentially non-informative ones.

Finally, we predict infectious potentials from more noisy subreads of Nanopore long reads. To this end, we resimulated the bacterial and viral datasets using the exact same genomes and the context-independent model of DeepSimulator 1.5 (Li et al., 2020). We set the target average read length to 8kb and discarded reads shorter than 250bp. Then, we extracted 250bp-long subreads for training and evaluation of our classifiers, but kept full reads for benchmarking against minimap2 (Li, 2018), a popular Nanopore mapper. We chose 250bp as this allows fair comparison with other models and corresponds to information available after ca. 0.5s (Rang et al., 2018). Successful predictions after such a short time could be used together with real-time selective sequencing (Loose et al., 2016) to enrich the samples in reads originating from pathogens and save resources. We trained new models for the bacterial and viral Nanopore datasets and compared them with minimap2 and models trained on 250bp Illumina reads. Evaluation of minimap2 was performed analogously to HiLive2's, selecting the representative alignment if a chimeric match was found. In addition to the simulated data prepared as explained above, we also used a dataset of real reads originating from a SARS-CoV-2 isolate (SRR11140745, collected on 14 Feb 2020).

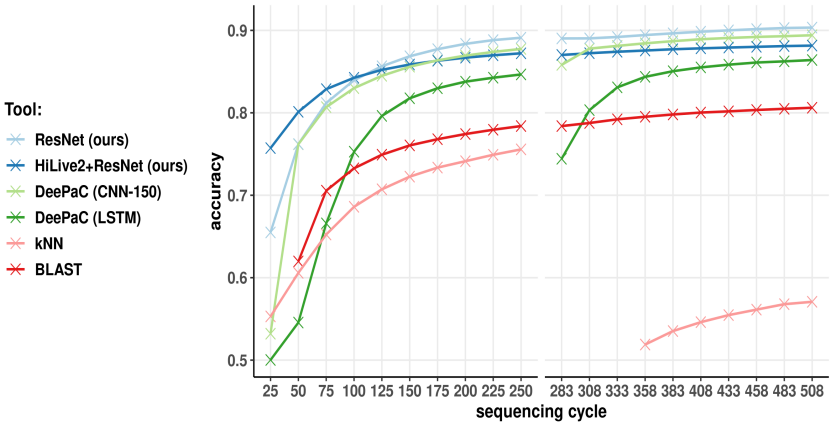

Figure 2: Accuracy for the viral dataset. Only the tools achieving more than 60% average accuracy are shown. For BLAST and kNN, accuracy values below 50% are not shown. The ResNet has the best average accuracy overall. DeePaC (CNN-150) corresponds to DeePaC-vir's $\text{CNN}_{\text{All-150}}$. DeePaC (CNN) trained on 250bp reads is omitted due to its subpar performance on reads shorter than 200bp (see Table 2 and Fig. S1c). The sudden drop in kNN performance is caused by frequent conflicting predictions for both mates, which leads to many missing predictions. Cycle numbers for the second mate are shifted by 8 due to an 8nt-long simulated barcode present in the BCL files, which is removed at mapping and prediction time.