Preprint
doi.10.1093/xxx/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

A deep learning framework for real-time detection of novel pathogens during sequencing

Jakub M. Bartoszewicz [1-4]*, Ulrich Genske [1-5] and Bernhard Y. Renard [1-3]*

[1]Data Analytics and Computational Statistics, Hasso Plattner Institute, 14482 Potsdam, Germany
[2]Digital Engineering Faculty, University of Postdam, 14482 Potsdam, Germany
[3]Bioinformatics Unit (MF1), Robert Koch Institute, 13353 Berlin, Germany and
[4]Department of Mathematics and Computer Science, Free University of Berlin, 14195 Berlin, Germany.
[5]Currently at: Department of Radiology, Charité – Universitätsmedizin Berlin, Freie Universität Berlin, Humboldt-Universität zu Berlin, and Berlin Institute of Health, 10117 Berlin, Germany.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: 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. Machine learning approaches can predict if single reads originate from more distant, unknown pathogens, but require relatively long input sequences and processed data from a finished sequencing run.

Results: We present DeePaC-Live, a Python package for real-time pathogenic potential prediction directly from incomplete sequencing reads. We train deep neural networks to classify Illumina and Nanopore reads and integrate our models with HiLive2, a real-time Illumina mapper. DeePaC-Live outperforms alternatives based on machine learning and sequence alignment on simulated and real data, including SARS-CoV-2 sequencing runs. After just 50 Illumina cycles, 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. Our approach could also be used for screening synthetic sequences against biosecurity threats.

Availability: The code is available at: https://gitlab.com/dacs-hpi/deepac-live and https://gitlab.com/dacs-hpi/deepac. The package can be installed with Bioconda, Docker or pip.

Contact: Jakub.Bartoszewicz@hpi.de, Bernhard.Renard@hpi.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

1.1 Background

The SARS-CoV-2 coronavirus emerged in December 2019 causing an outbreak of COVID-19, a severe respiratory disease, which quickly spiraled out of control into a global pandemic of 2020. This virus of probable zoonotic origin (Zhou *et al.*, 2020) is a terrifying example of how easily new agents can spread. What is more, many more novel pathogens are expected to emerge. They evolve extremely quickly due to high mutation rates or horizontal gene transfer, while human exposure to the vast majority of unexplored microbial biodiversity is rapidly growing (Vouga and Greub, 2016; Trappe *et al.*, 2016). New biosafety threats may come as novel bacterial agents like the Shiga-toxigenic *Escherichia coli* strain that caused a deadly epidemic in 2011 (Frank *et al.*, 2011). Outbreaks of previously unknown viruses can be even more severe. This is not limited to coronaviruses like SARS-CoV-1, MERS and SARS-CoV-2. Novel strains of the Influenza A virus caused four pandemics in less than a hundred years (from the "Spanish flu" of 1918 until the "swine flu" of 2009), killing millions of people. Even known pathogens may be difficult to control, as proven by the outbreaks of Zika and Ebola in the 2010s (Calvignac-Spencer *et al.*, 2014). Importantly, many viruses can switch between more than one host or evolve silently in an animal reservoir before infecting humans. This happened before to HIV, Ebola, many dangerous strains of the Influenza A virus and the coronaviruses mentioned above.

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 is the method of choice (Lecuit and Eloit, 2014; Calistri and Palù, 2015). A swift response is crucial, as the

© The Author 2021.

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. This can be achieved by design using long-read sequencing like Oxford Nanopore (ONT). However, lower throughput and high error rates of those technologies impede their adoption for pathogen detection. Scalability, cost-efficiency and accuracy of Illumina sequencing still make it a gold standard, although it may change in the future with the establishment of improved ONT protocols and computational methods (Loka *et al.*, 2019).

1.2 Real-time analysis of Illumina sequencing data

Analyzing Illumina reads during the sequencing run poses unique technical and algorithmic challenges. The DRAGEN system relies on field-programmable gate arrays (FPGAs) to speed up the computation and can be combined with specialized protocols to detect clinically relevant variants in the human genome but depends on finished reads (Miller *et al.*, 2015). An alternative approach is to use the general purpose computational infrastructure and optimize the algorithms for fast and accurate analysis of incomplete reads as they are produced, during the sequencing run. HiLive (Lindner *et al.*, 2017) and HiLive2 (Loka *et al.*, 2019) are real-time mappers, performing on par with the traditional mappers like Bowtie2 (Langmead and Salzberg, 2012) and BWA (Li and Durbin, 2010) with no live-analysis capabilities. However, as read mappers 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 live-analysis tools and associated pipelines do cover standard read-based pathogen detection workflows, their performance on novel agents is limited by their dependence on databases of known species. The same problem applies to sequence alignment and taxonomic classification in general, also outside of the real-time analysis context (National Research Council, 2010). In this work, we show that using deep learning to predict if a read originates from a human pathogen is a promising alternative to mapping the reads to known references if the correct reference genome is not yet known or unavailable.

1.3 Read-based detection of novel pathogens

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 and 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. In contrast, taxonomy-agnostic methods try to reduce their database dependency by assigning putative phenotypes directly to analysed sequences, deliberately omitting the taxonomic classification step. For example, NBC (Rosen *et al.*, 2011) is a naïve Bayes classifier based on k-mer frequency features that can be trained to classify reads directly into arbitrary classes. However, in the context of detecting novel bacterial pathogens, a random forest approach of PaPrBaG (Deneke *et al.*, 2017) performs much better. Zhang *et al.* (2019) used a kNN classifier to develop a similar method for detection of human-infecting viruses. An analogous deep learning approach, DeePaC, outperforms the traditional machine learning algorithms on both novel bacteria (Bartoszewicz *et al.*, 2020) and viruses (DeePaC-vir, Bartoszewicz *et al.* (2021)), offering an additional level of interpretability on nucleotide, read and genome levels. A similar method has been developed by Mock *et al.* (2020). However, it focuses on detailed predictions for a small set of three viral species and cannot be used in an open-view setting. Preliminary work shown independently by Guo *et al.* (2020) supports it, but the code, trained models or installables are not available yet, so the method cannot be reused. What is more, the authors did not guarantee "novelty" of the viruses in the test set – it could contain genomes of viruses present in the training set, as long as they were resequenced after 2018. Performance on truly novel agents is difficult to assess. While pathogenicity prediction methods using whole genomes or protein sets as input also exist, this work focuses on read-based classification to offer real-time predictions and avoid delays necessitated by assembly pipelines. However, read-based methods have been shown to perform well and also for full genomes and assembled contigs, achieving similar or better performance than alignment-based approaches (Deneke *et al.*, 2017; Bartoszewicz *et al.*, 2020, 2021).

Methods developed for pathogen detection in the context of real-time sequencing could also be used to improve screening workflows for safe and secure synthetic biology. Host-range of viral pathogens can be deliberately modified (Herfst *et al.*, 2012; Imai *et al.*, 2012), and a virus similar to the Variola virus (the cause of smallpox and a bioweapon) was synthesized (Noyce *et al.*, 2018; Thiel, 2018). Lipsitch and Inglesby (2014) speculated on modifications increasing the pathogenicity of coronaviruses. On the other hand, a report by the National Academies of Sciences, Engineering, and Medicine (2018) sees virulence-enhancing manipulation of existing bacteria as the issue of the highest concern. Computational screening of ordered sequences is a standard, but challenging precaution measure used by the DNA synthesis industry; evaluation of novel sequences requires significant computational resources and expert analysts. As it depends on sequence alignment against databases of known threats, it suffers from the same problems as other taxonomy-dependent pathogen detection methods. False positive rates are high, especially for oligonucleotide building blocks below 200bp (Diggans and Leproust, 2019).

2 Methods

2.1 Design

To capture Illumina reads as they are generated by a sequencer, we use HiLive2's BCL file conversion and real-time mapping capabilities (Lindner *et al.*, 2017; Loka *et al.*, 2019). Our tool, DeePaC-Live, consists of three asynchronously callable modules. The sender module watches the HiLive2 output directory, detecting BAM files with both mapped and unmapped reads. By default, it selects only the unmapped reads for further analysis, but this can be adjusted by the user to focus on either mapped reads, or all sequenced reads. The output of the sender module may be automatically sent over to a remote server (e.g. a GPU-equipped machine) using the SFTP protocol. Data privacy issues should be kept in mind.

The receiver module may operate on the remote or local machine depending on the available infrastructure. It captures the sender's output and uses a selected deep neural network to predict pathogenic potentials (standard sigmoid output scores between 0 and 1) for all the selected reads. Then, it filters them according to a predefined decision threshold (typically 0.5), outputting separate files for reads associated with a pathogenic and nonpathogenic phenotype. Finally, the optional refiltering module allows reanalyzing the prediction with an alternative threshold (e.g. to select only the highest-confidence predictions) and averaging the outputs of multiple receiver modules to create a simple ensemble classifier.

Any custom Keras model can be used for predictions, including arbitrary binary classifiers for tasks other than described here. We also support seamless integration with the built-in DeePaC (Bartoszewicz *et al.*, 2020) and DeePaC-vir (Bartoszewicz *et al.*, 2021) models. 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. Therefore, we trained new models, aiming to achieve high performance for both the intermediate cycles and

the final output of the sequencer. As the prediction functions of DeePaC were not optimized for fast inference, we added a possibility to adjust the inference batch size to fully utilize computing power of a given GPU. We set the batch size to 1536, being the highest multiple of 512 that would not cause out-of-memory errors for any of the tested models. Note that while the batch size could be further increased for the CNN and ResNet-based networks used in this study, this did not speed up inference any more.

2.2 Data preparation

The original DeePaC dataset consists of 250bp simulated Illumina reads in fastq format. The training, validation and test sets contain reads originating from different species of pathogenic (including opportunistic pathogens) or commensal bacteria labelled using the IMG database (Chen *et al.*, 2019). The DeePaC-vir dataset is built in an analogous way using different viruses mined from the Virus-Host Database (Mihara *et al.*, 2016). Three alternative versions of the viral dataset are available, differing in the negative class definition. We used the fully open-view "All" dataset, containing all viruses available in VHDB. In all cases, the training set contained 20 million single reads, the validation set contained 2.5 million single reads and the test set – 2.5 million paired-end reads. This setup allows training models correctly handling single, isolated reads, but also testing their performance on read pairs. All sets were balanced with regard to the class distribution and contained a mixture of reads originating from multiple different species. Most importantly, the training, validation and held-out test sets contain different viruses or bacterial species, so that generalization to "novel" agents (i.e. unseen in training) can be explicitly evaluated. For more details regarding the dataset generation, we refer the reader to the corresponding publications (Bartoszewicz *et al.*, 2020, 2021).

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 original DeePaC and DeePac-vir test datasets to generate corresponding subread datasets with subread lengths between 25 and 250 (full read) with a step of 25. Every subread of every subread set had a corresponding subread in all the other sets. Therefore, we could explicitly model 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 for every read in the set. In this setup, all integer values of k between 25 and 250 were allowed. As Bartoszewicz *et al.* (2021) have previously presented both 250bp and 150bp-trained reverse-complement CNN classifier for viruses, we also generated an analogous 150bp subread bacterial dataset, and used it to train a corresponding CNN.

2.3 ResNets and hybrid classifiers

We investigated two relatively shallow architectures shown previously to perform well in the pathogenicity or host-range prediction task – a 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 subread sequences convey less information, we expected the subread classification problem to be more challenging than in the case of relatively long 250bp reads. We suspected that a deeper, more expressive network could perform better. Therefore, we implemented a new architecture – a reverse-complement ResNet extending the previous work with skip connections (He *et al.*, 2016) while satisfying the reverse-complementarity constraint.

More specifically, 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). We trained them for a maximum of 30 epochs, using early stopping with a patience of 10 epochs (see Table S1 and Fig. S1 for architecture details). 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.

Finally, we combine HiLive2 with DeePaC to extract reads mappable to known references and predict the phenotype for the unmapped reads. This enables identification of the closest relatives of the analyzed pathogen, while still predicting labels for reads missed by the mapper. The reads associated with the pathogenic or infectious phenotype may then be extracted and used in downstream analysis.

2.4 Benchmarking

2.4.1 Bacteria

We compare our neural networks and hybrid classifiers 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).

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 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. 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.

2.4.2 Viruses

We use an analogous approach to evaluate the classification performance on reads originating from novel viruses. However, since PaPrBaG is a method developed for bacterial genomes, we benchmark our model 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. As kNN yields binary predictions, we integrate them using the same approach we use for BLAST. Finally, we compare our models to the original DeePaC-vir models.

2.4.3 Staphylococcus aureus

To test the performance of the bacterial models on real sequencing data, we analyze reads coming from a real sequencing run of a pathogenic bacterium *Staphylococcus aureus*. This species was not present in the training set (it had been randomly placed in the validation set), so models a "novel" pathogen without a known reference genome. The same species was used previously by Bartoszewicz *et al.* (2020) to assess the original version of DeePaC; here, we focus on analyzing the sequences as they are generated by the sequencer as opposed to predicting for full reads after the sequencing run is finished. To this end, we download an SRA archive of 251bp-long paired-end reads (accession number SRR5110368) sequenced with an Illumina MiSeq device (Manara *et al.*, 2018). We use untrimmed reads with the quality information to generate BCL files as they would be internally generated by the sequencer. We then run HiLive2 on the BCL data to map the reads to our training reference database; HiLive2 output is then parsed and passed to our models. However, we ignore the last cycle of each mate when generating HiLive2 output and subsequent analyses, as the bad quality of this last nucleotide makes it generally unreliable. We select the predictor that achieves the highest average accuracy on the DeePaC dataset and compare it to the standard, mapping-based real-time analysis with HiLive2 alone.

2.4.4 SARS-CoV-2

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 providing 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 subreads of interest while discarding the potentially non-informative ones.

2.4.5 Nanopore

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 (omitting the ResNet-34) 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 two real SRA datasets: a SARS-CoV-2 isolate (SRR11140745, collected on 14 Feb 2020) and a clinical *S. aureus* sample (SRR8776887, Dilthey *et al.* (2020)).

3 Results

3.1 Subread models

In Illumina paired-end protocols, the barcodes are sequenced after the first mate, making live-demultiplexing possible but problematic. Changing the barcode sequencing order is not trivial, as initial clustering requires sufficient sequence diversity in the first several cycles. A possible workaround uses asynchronous paired-end sequencing protocols (Loka *et al.*, 2019), sacrificing the first read's length for faster demultiplexing and relying on the second mate to compensate for the lost information. We tested our models in 100 settings corresponding to different lengths of the first mate (modelling different length-time trade-offs) and the second mate (modelling incoming information after demultiplexing). In the Fig. S2 we present a comparison of the DeePaC CNNs (Bartoszewicz *et al.*, 2020, 2021) with their subread-optimized ResNet counterparts from this study. The previous state-of-the-art for the bacterial dataset is outperformed by the ResNet trained on mixed-length subreads (Fig. S2a-b) across most of the spectrum of read length combinations. For the longest read pairs (200bp and above) accuracy is slightly lower, which could be related to the old model being explicitly optimized for 250bp-long sequences. Performance of the DeePac-vir's 250bp-trained CNN collapses for viral reads shorter than 200bp, while our model trained on the mixed-length reads maintains accuracy higher than 80% for reads as short as 50bp (Fig. S2c-d). What is more, our model slightly outperforms the previous state-of-the-art also on full-length reads, with accuracy over 90% for pairs of 225bp or more. The accuracy matrices are almost symmetric (with negligible deviations), proving that the models' performance is on average identical for the first and the second mate.

3.2 Hybrid models

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 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. PaPrBaG underperforms even though it was retrained specifically for a subread classification scenario. The mapping approach, represented by HiLive2, is the most precise. Low accuracy of both BLAST and HiLive2 reflects a high missing prediction rate (crossing 80% for HiLive2 at cycles 225-250), although BLAST performs better due to its less strict and more sensitive alignment criteria. A comparison of accuracy values at each cycle is presented in the Fig. S3.

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

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	61.7	54.1
BLAST	73.2	97.2	73.7
HiLive2	51.1	**99.2**	50.3

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. It seems to stem mainly from increased sensitivity of the deep learning approach (89.3% for reads over 50bp, compared to 81.8% for HiLive) at some cost in precision (95.3% and 99.2%, respectively).

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. It is also the most sensitive prediction method overall. However, the hybrid classifiers offer a good trade-off between HiLive2's precision and the performance of the pure deep learning approach. The kNN classifier performs worse than BLAST even for the first mate. Its performance collapses as the second mate is introduced due to many conflicting predictions for both mates, resulting in missing predictions for many read pairs. A cycle-by-cycle accuracy comparison is presented in the Fig. S4.

Real-time sequencing analyses are very precise from as early as cycle 30 until the end of the sequencing run. This holds for both taxonomic read classification with LiveKraken (Tausch *et al.*, 2018a) and variant calling based on HiLive2's mapping results (Loka *et al.*, 2019). As shown in Table 2, this is also the case in the infectious potential prediction task. We compared HiLive2's stable performance to the viral hybrid classifier and the ResNet alone, which achieved precision comparable to alignment-based approaches (Fig. S5). The hybrid classifier crosses a 90% threshold at cycle 75 (90.1%), while never plunging below 80% even for the earliest cycles. What is more, all of the HiLive2-mapped reads are included in the hybrid classifier's predictions, so no information is lost by employing the extended approach. The high precision resulting from combining the real-time mapper with the deep learning classifier suggests that the associations of reads and a pathogenic phenotype are trustworthy even at the early stages of the sequencing run, getting even more reliable as more information is gained.

3.3 Runtime

To estimate the sample size that can be analyzed in real-time after parsing or mapping with HiLive2, we measured how many reads per second could be processed by the pathogenicity prediction methods compared in this study. Then, we calculated the number of predictions feasible in a time-frame corresponding to 25 cycles (with wall-time per cycle as in (Loka *et al.*, 2019)). Note that this is an inherently difficult comparison, as inference with deep learning models can be accelerated with GPUs, while other methods cannot. We used a desktop computer equipped with a consumer-grade GPU to benchmark the throughput of our models, and a 128-core machine with 500 GB RAM for the alternative approaches. In the Table S2 we present how many reads can be analyzed with no delays if the output is produced every 25 cycles, together with more detailed information on the GPUs and CPUs used and the effect of adjusting the inference batch size. Our 18-layer ResNet is faster than the original DeePaC models and only marginally slower than optimized CNNs and LSTMs, analyzing over 56 million reads in the given time-frame (5656 reads/s). This is over 6 times faster than the best non-deep learning based method for both bacteria and viruses, at a much lower cost in terms of the computational resources. This prediction speed is enough to guarantee real-time predictions for Illumina iSeq 100, MiniSeq and MiSeq devices, with a maximum of 4 and 25 million reads per run. For sequencers with even higher maximum throughput, like the obsolesced HiSeq and newer NovaSeq 550 machines (with up to 400 million reads per run), further speed-up is possible by distributing the computation across more than one GPU. Multiple receiver instances can be easily assigned to dedicated GPUs to handle different barcodes, cycles or both in parallel.

3.4 Real sequencing data

We benchmarked the best bacterial model on data from a real *S. aureus* sequencing run (see Section 2.4.3). As this dataset contains reads from just one "novel" pathogen (a species which was not present in the training database), the true positive rate (recall) and accuracy are equivalent. The hybrid HiLive2+ResNet classifier crosses the 90% threshold just after 75 cycles (when the true positive rate equals 89.8%) and reaches 98.8% in the last analyzed cycle (Fig. S6). HiLive2 is only able to identify 5.8% of the reads at its best cycle, which drops down to 2.0% at the end of the sequencing run, when longer sequences are analyzed.

We further evaluated our approach on data from a real SARS-CoV-2 sequencing run (see Section 2.4.4). 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 (Section 3.2), 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. Omitting the mapping functionality of HiLive2 (and using it just for parsing the BCL files generated by the sequencer) 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 subreads. As the deep learning models consistently

Table 3. Reads identified as pathogenic from the SARS-CoV-2 sequencing run. The ResNet alone 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 only (all), or to form a consensus predictor (cons.) by selecting subreads assigned to the pathogenic class by both the ResNet and BLAST.

	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 (all)	211411	41.0%	142794	27.7%
ResNet+BLAST (all)	**264646**	**51.3%**	**195634**	**37.9%**
HL+ResNet+BLAST (cons.)	35163	6.8%	35163	6.8%
ResNet+BLAST (cons.)	48192	9.3%	48192	9.3%

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

outperform BLAST, we use them as predictors to extract subreads of interest. A BLAST follow-up analysis annotates the selected 50bp subreads with their closest taxonomic matches (which may include non-pathogens) wherever a match is found. Alternatively, we can create a consensus predictor by treating BLAST as a confirmatory analysis to focus only on subreads which are predicted to originate from the positive class *and* have a positive BLAST match. In the latter case, we observe a significant enrichment in sequences more similar to the pathogenic SARS-CoV-1 virus. 99.3% of the subreads identified as "pathogenic" by the consensus ResNet+BLAST workflow are matched with human SARS viruses present in the training database, while the number of identified subreads is almost 15 times higher than for HiLive2 alone. Our results suggest that predictions of the ResNet, even without a BLAST follow-up, are also 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.

3.5 Nanopore reads

Finally, we evaluated our Nanopore models to investigate possible applications to noisier long-read sequencing technologies (Table 4). The Nanopore-trained ResNets achieved higher validation accuracy than the CNNs and LSTMs trained on the same data and were selected for further evaluation. As expected, mapping with minimap2 is the most precise method and Illumina-trained neural networks of DeePaC 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 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 66.9% of full-length *S. aureus* reads and only 9.9% of full SARS-CoV-2 reads, compared to 94.7% and 52.7% for our ResNets respectively (Table S3).

Our results suggest that our classifiers could find applications in selective sequencing workflows, enabling targeted analysis of reads originating from novel pathogens while discarding potentially less-interesting non-pathogen reads. Although a given read could contain sequences matching to pathogen references located after the initial 250bp, the risk of premature termination seems to be mitigated by the classifier's superior performance, especially in the case of novel viruses. This risk can be further adjusted to the user's needs by selecting an alternative classification threshold, manipulating the expected sensitivity, precision and false positive rates as shown by the ROC and PR curves (Fig. S7).

4 Discussion

All 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. The models presented here assign probability-like scores to DNA and RNA sequences without establishing a mechanistic link between a given sequence and the predicted phenotype. This is, however, an advantage of the proposed approach as well. By resigning from speculation on that link (e.g. by sequence alignment to known relatives), models trained on carefully selected data outperform the traditional methods in terms of both prediction speed and accuracy, yielding predictions for all sequences in the sample. Any assumptions and biases affecting the labels will be reflected by the trained classifier. While this is also the case for read alignment and k-mer based classification, they do not require retraining after a database update. On the other hand, as our classifiers generalize well to sequences absent from the training database, they may be updated less frequently while maintaining the desired performance.

A separate question is whether the captured signal has any underlying functional meaning, or is purely taxonomic in nature. BLAST, as a sensitive method of homology detection, is a gold-standard for finding taxonomic relationships. Therefore, it could be assumed that outperforming BLAST is a sign of learning more than just evolutionary distances. On the other hand, the very nature of read-based predictions renders any broader biological context inaccessible – the analyzed sequences are simply too short to contain reliable peptide-level features (Deneke *et al.*, 2017; Bartoszewicz *et al.*, 2020), let alone information about the structural or functional characteristics of the encoded proteins (or intergenic regions). However, intepretability workflows like Genome-Wide Phenotype Potential Analysis (Bartoszewicz *et al.*, 2021) have shown that read-based pathogenicity prediction models assign high pathogenic potentials to reads originating from genes engaged in virulence, also specifically in the case of *S. aureus*. What is more, they show that regions of higher pathogenic potential are non-uniformly distributed in both bacterial and viral genomes with "peaks" of elevated potential aligning with relevant genes. This suggests that even though detecting pathogenicity islands or virulence factors directly is not possible, reads associated with the pathogenic phenotype do originate from important regions of interest. On the other hand, one can expect similar sequences to yield similar predictions. This makes distinguishing between closely related pathogens and non-pathogens challenging, as shown previously for SARS-CoV-2

and its relative, RaTG13 (Bartoszewicz *et al.*, 2021). The problem can be explicitly modelled by including related viruses infecting different hosts (as in the DeePaC-vir dataset used in this study) or by training classifiers targeting novel strains of known bacterial species (Bartoszewicz *et al.*, 2020). Nevertheless, occasional misclassifications of similar sequences are a real possibility that has to be kept in mind. This also applies to alignment-based and k-mer based approaches.

The accuracy achieved by our models clearly shows that predicting if a read comes from a bacterial pathogen or a human-infecting virus is indeed possible, even if there is no reference genome available. This may actually be a form of texture bias – Brendel and Bethge (2019) have shown that CNNs can correctly classify images based on fragments of as little as 17x17 pixels. Here, a local DNA or RNA pattern is often predictive of the phenotype label assigned to the genome. However, our models do not return any information on the closest possible match, which is generally necessary in any pathogen detection task. In this study, we proposed to solve this problem by combining the deep learning approach with an alignment-based one. Alternative methods of taxonomic classification could be used instead – based on either k-mers or machine learning. Kraken (Wood and Salzberg, 2014), a k-mer approach, was outperformed by both BLAST and PaPrBaG in the study by Deneke *et al.* (2017). Nevertheless, we can imagine that using a well-trained taxonomic classifier, preferably one yielding at least putative species-level predictions for every read, would be a very useful tool for follow-up analyses of reads prefiltered with our models. On the other hand, since the reads associated with pathogenicity are often co-localized within relevant genomic features (Bartoszewicz *et al.*, 2021), assembly of the filtered reads could recover longer contigs corresponding to genes or perhaps even gene clusters.

DeePaC-Live could also form a part of more complex real-time pathogen detection workflows like PathoLive (Tausch *et al.*, 2018b). For example, PAIPline (Andrusch *et al.*, 2018) identifies pathogens in metagenomic and clinical samples via mapping and a BLAST follow-up analysis, but can only start the analysis after the sequencing is finished. Exchanging Bowtie2 (Langmead and Salzberg, 2012) for HiLive2 with a DeePaC-Live hybrid classifier directing potentially informative reads to the BLAST confirmatory step could serve as a backbone of an extended, real-time version of the pipeline. Alternatively, PathoLive could be extended with DeePaC-Live and BLAST follow-up steps. To fully handle metagenomic samples, our classifiers would have to be retrained in a multi-class setting encompassing a broader spectrum of clinically relevant pathogen groups. On the other hand, if there are reasons to believe that the disease-causing agent is a virus or a bacterium, the models presented here may suffice.

As DeePaC-Live relies on either BAM or fasta input, it is not necessarily dependent on HiLive2 and can be used in combination with alternative approaches to accelerated sequencing analysis, for example the DRAGEN system. Nanopore-trained models perform relatively well despite higher sequencing noise, and 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.

Finally, our models could be used in applications beyond the sequencing context, e.g. screening against biosecurity threats at the DNA synthesis facilities. Evaluating sequences shorter than 200bp is usually not feasible due to high false positive rates and the computational burden; a PhD-level proficiency in bioinformatics is required to both implement the pipelines and analyse the results (Diggans and Leproust, 2019). Taken together, those challenges warrant investigating deep learning alternatives to traditional workflows. Our models deliver high accuracy and precision for sequences well below the established 200bp limit, and their false positive rates can be lowered even more if a decision threshold higher than the default 0.5 is used. Given the inference speed of our classifiers, we envision a system where the suspicious sequences are filtered with DeePaC-Live and piped into a follow-up analysis akin to BLAST, lowering the computational burden of sequence alignment and improving the performance.

5 Conclusion

We present a new tool 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 classifiers can also be used for sequence-based tasks beyond NGS analysis, for example as a screening system for synthetic DNA sequences difficult to evaluate before. The package can be easily installed with Bioconda (Grüning *et al.*, 2018), Docker or pip. The code and installation instructions are available at https://gitlab.com/dacs-hpi/deepac-live (real-time inference and HiLive2 integration) and https://gitlab.com/dacs-hpi/deepac (ResNet training and data preprocessing). The datasets are hosted at https://doi.org/10.5281/zenodo.4456857 along the trained models (https://doi.org/10.5281/zenodo.4456008).

Acknowledgements

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

Funding

This work was supported by the German Academic Scholarship Foundation (JMB), the BMBF-funded Computational Life Science initiative (project DeePath, to BYR) and the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A537B, 031A533A, 031A538A, 031A533B, 031A535A, 031A537C, 031A534A, 031A532B).

References

Altschul, S. F. *et al.* (1990). Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3), 403–410.

Andrusch, A. *et al.* (2018). PAIPline: pathogen identification in metagenomic and clinical next generation sequencing samples. *Bioinformatics*, **34**(17), i715–i721.

Bartoszewicz, J. M. *et al.* (2020). DeePaC: predicting pathogenic potential of novel DNA with reverse-complement neural networks. *Bioinformatics*, **36**(1), 81–89.

Bartoszewicz, J. M. *et al.* (2021). Interpretable detection of novel human viruses from genome sequencing data. *NAR Genomics and Bioinformatics (in press)*. Preprint retrieved from: https://doi.org/10.1101/2020.01.29.925354.

Brendel, W. and Bethge, M. (2019). Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet. In *International Conference on Learning Representations*.

Calistri, A. and Palù, G. (2015). 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.

Calvignac-Spencer, S. *et al.* (2014). Clock rooting further demonstrates that guinea 2014 ebov is a member of the zaïre lineage. *PLoS currents*, **6**.

Camacho, C. *et al.* (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, **10**(1), 421.

Chen, I.-M. A. *et al.* (2019). IMG/m v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Research*, **47**, D666–D677.

Deneke, C. *et al.* (2017). PaPrBaG: A machine learning approach for the detection of novel pathogens from NGS data. *Scientific Reports*, **7**, 39194.

Diggans, J. and Leproust, E. (2019). Next Steps for Access to Safe, Secure DNA Synthesis. *Frontiers in Bioengineering and Biotechnology*, **7**.

Dilthey, A. T. *et al.* (2020). Ultraplexing: increasing the efficiency of long-read sequencing for hybrid assembly with k-mer-based multiplexing. *Genome Biology*, **21**(1), 68.

Frank, C. *et al.* (2011). Epidemic profile of shiga-toxin-producing escherichia coli o104:h4 outbreak in germany. *The New England Journal of Medicine*, **365**(19), 1771–1780.

Grüning, B. *et al.* (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, **15**(7), 475–476. Number: 7 Publisher: Nature Publishing Group.

Guo, Q. *et al.* (2020). Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm. *bioRxiv*, page 2020.01.21.914044. Publisher: Cold Spring Harbor Laboratory.

He, K. *et al.* (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Herfst, S. *et al.* (2012). Airborne Transmission of Influenza A/H5n1 Virus Between Ferrets. *Science*, **336**(6088), 1534–1541.

Imai, M. *et al.* (2012). Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1n1 virus in ferrets. *Nature*, **486**(7403), 420–428.

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, **9**(4), 357–359.

Lecuit, M. and Eloit, M. (2014). The diagnosis of infectious diseases by whole genome next generation sequencing: a new era is opening. *Frontiers in Cellular and Infection Microbiology*, **4**, 25.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**(18), 3094–3100. _eprint: https://academic.oup.com/bioinformatics/article-pdf/34/18/3094/25731859/bty191.pdf.

Li, H. and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, **26**(5), 589–595.

Li, Y. *et al.* (2020). DeepSimulator1. 5: a more powerful, quicker and lighter simulator for Nanopore sequencing. *Bioinformatics*, **36**(8), 2578–2580.

Lindner, M. S. *et al.* (2017). HiLive: real-time mapping of illumina reads while sequencing. *Bioinformatics*, **33**(6), 917–319.

Lipsitch, M. and Inglesby, T. V. (2014). Moratorium on Research Intended To Create Novel Potential Pandemic Pathogens. *mBio*, **5**(6).

Loka, T. P. *et al.* (2019). Reliable variant calling during runtime of Illumina sequencing. *Scientific Reports*, **9**(1), 1–8.

Loose, M. *et al.* (2016). Real-time selective sequencing using nanopore technology. *Nature Methods*, **13**(9), 751–754. Number: 9 Publisher: Nature Publishing Group.

Manara, S. *et al.* (2018). Whole-genome epidemiology, characterisation, and phylogenetic reconstruction of staphylococcus aureus strains in a paediatric hospital. *Genome Medicine*, **10**(1), 82.

Mihara, T. *et al.* (2016). Linking virus genomes with host taxonomy. *Viruses*, **8**(3), 66.

Miller, N. A. *et al.* (2015). A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Medicine*, **7**(1), 100.

Mock, F. *et al.* (2020). VIDHOP, viral host prediction with deep learning. *Bioinformatics*. btaa705.

National Academies of Sciences, Engineering, and Medicine (2018). *Biodefense in the Age of Synthetic Biology*. The National Academies Press, Washington, DC.

National Research Council (2010). *Sequence-Based Classification of Select Agents: A Brighter Line*. The National Academies Press, Washington, DC.

Noyce, R. S. *et al.* (2018). Construction of an infectious horsepox virus vaccine from chemically synthesized DNA fragments. *PLOS ONE*, **13**(1), e0188453.

Rang, F. J. *et al.* (2018). From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. *Genome Biology*, **19**(1), 90.

Rosen, G. L. *et al.* (2011). NBC: the naïve bayes classification tool webserver for taxonomic classification of metagenomic reads. *Bioinformatics*, **27**(1), 127–129.

Tausch, S. H. *et al.* (2018a). LiveKraken—real-time metagenomic classification of illumina data. *Bioinformatics*, **34**(21), 3750–3752.

Tausch, S. H. *et al.* (2018b). PathoLive - Real time pathogen identification from metagenomic Illumina datasets. *bioRxiv*, page 402370.

Thiel, V. (2018). Synthetic viruses-Anything new? *PLoS pathogens*, **14**(10), e1007019.

Trappe, K. *et al.* (2016). Detecting horizontal gene transfer by mapping sequencing reads across species boundaries. *Bioinformatics*, **32**(17), i595–i604.

Vouga, M. and Greub, G. (2016). Emerging bacterial pathogens: the past and beyond. *Clinical Microbiology and Infection*, **22**(1), 12–21.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, **15**(3), R46.

Zhang, Z. *et al.* (2019). Rapid identification of human-infecting viruses. *Transboundary and Emerging Diseases*, **66**(6), 2517–2522.

Zhou, P. *et al.* (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, **579**(7798), 270–273.